

# ESSAYS ON NEWS CONSUMPTION AND DYNAMIC PROGRAMMING

by

Jenő Pál

Sumbitted in partial fulfillment of the requirements for  
the degree of Doctor of Philosophy at  
Central European University

Supervisor: Miklós Koren

Budapest, Hungary

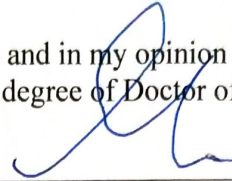
© Copyright by Jenő Pál, 2017.  
All rights reserved.

CENTRAL EUROPEAN UNIVERSITY  
DEPARTMENT OF ECONOMICS AND BUSINESS

The undersigned hereby certify that they have read and recommend to the Department of Economics for acceptance a thesis entitled **"Essays on News Consumption and Dynamic Programming"** by **Jeno Pal**.

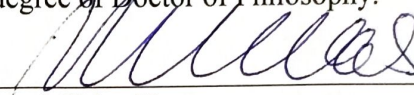
Dated: December 4, 2017

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.  
Chair of the Thesis Committee:



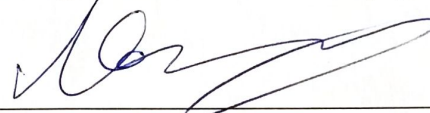
Laszlo Matyas

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.  
Advisor:



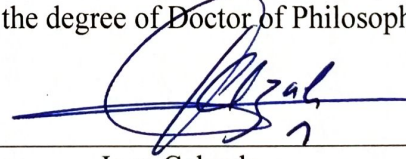
Miklos Koren

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.  
Internal Examiner:



Sergey Lychagin

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.  
External Examiner:



Joan Calzada

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.  
External Member:



Balazs Murakozy

CENTRAL EUROPEAN UNIVERSITY  
DEPARTMENT OF ECONOMICS AND BUSINESS

Author: Jenő Pál  
Title: Essays on News Consumption and Dynamic Programming  
Degree: Ph.D.  
Dated: December 4, 2017

Hereby I testify that this thesis contains no material accepted for any other degree in any other institution and that it contains no material previously written and/or published by another person except where appropriate acknowledgement is made.

Signature of the author:

  
.....  
Jenő Pál

## **Disclosure of coauthor contribution**

### **The Impact of Aggregators on Internet News Consumption**

Co-authors: Susan Athey and Markus Mobius

Susan Athey and Markus Mobius came up with the idea to analyze the shutdown of Google News in Spain. Lead by Susan Athey and Markus Mobius, all three authors contributed to elaborating the empirical strategy. Markus Mobius implemented the topic modelling of news articles and Jenő Pál implemented the empirical analysis.

### **Fitted Value Function Iteration with Probability One Contractions**

Co-author: John Stachurski

John Stachurski came up with the idea for the paper and John Stachurski and Jenő Pál both contributed substantially to developing the precise theoretical arguments and to the implementation of the numerical examples of the paper.

## Abstract

The thesis consists of three chapters: the first, single-authored chapter and the second chapter, co-authored with Susan Athey and Markus Mobius, analyze online news consumption based on browsing behavior data, while the third chapter (co-authored with John Stachurski) is about a problem in numerical dynamic programming.

## Chapter 1

Using a linked dataset of frontpages of The New York Times observed multiple times a day and browsing data, I measure the effect of the positions of news articles on the popularity of news articles. In a simple multinomial logit framework with article fixed effects I am able to identify position effects from within article differences of positions across frontpages. Focusing on the vertical positioning, I estimate a negative and decreasing strength effect of article position. The magnitude of the estimated effects is large: an article positioned in the top versus the second positions all else being equal results in 26 to 44 percent higher view share for the upper position. In a counterfactual exercise, I find a conservative mean increase of 5 percent on click-through rate resulting from solely reordering some articles on the frontpage. Furthermore, the actual ordering observed in the data is closer to a random ordering than to the counterfactual optimum. These findings point to the possibly enormous influence of editorial decisions on what people read in newspapers.

## Chapter 2

A policy debate centers around the question whether news aggregators such as Google News decrease or increase traffic to online news sites. One side of the debate, typically espoused by publishers, views aggregators as *substitutes* for traditional news consumption because aggregators' landing pages provide snippets of news stories and therefore reduce the incentive to click on the linked articles. Defendants of aggregators, on the other hand, view aggregators as *complements* because

they make it easier to discover news and therefore drive traffic to publishers. This debate has received particular attention in the European Union where two countries, Germany and Spain, enacted copyright reforms that allow newspapers to charge aggregators for linking to news snippets. In this paper, we use Spain as a natural experiment because Google News shut down all together in response to the reform in December 2014. We compare the news consumption of a large number of Google News users with a synthetic control group of similar non-Google News users. We find that the shutdown of Google News reduces overall news consumption by about 20% for treatment users, and it reduces page views on publishers other than Google News by 10%. This decrease is concentrated around small publishers while large publishers do not see significant changes in their overall traffic. We further find that when Google News shuts down, its users are able to replace some but not all of the types of news they previously read. Post-shutdown, they read less breaking news, hard news, and news that is not well covered on their favorite news publishers. These news categories explain most of the overall reduction in news consumption, and shed light on the mechanisms through which aggregators interact with traditional publishers.

### Chapter 3

This paper studies a value function iteration algorithm based on nonexpansive function approximation and Monte Carlo integration that can be applied to almost all stationary dynamic programming problems. The method can be represented using a randomized fitted Bellman operator and a corresponding algorithm that is shown to be globally convergent with probability one. When additional restrictions are imposed, an  $O_P(n^{-1/2})$  rate of convergence for Monte Carlo error is obtained. This paper has been already published (Jenő Pál and John Stachurski: *Fitted Value Function Iteration with Probability One Contractions*. Journal of Economic Dynamics and Control, 37 (2013) 251-264).

## Acknowledgements

It would have been impossible for me to write this thesis without the support of many people. I am going to be quite repetitive which is a great feeling: having so much more people standing with me in the years of my PhD than all the synonyms of "support" and "friendship" is something unique.

I am grateful to my advisor, Miklós Koren for constant support and guidance he offered me throughout the long journey of my studies. After each of our meetings I could regain belief that my work eventually will turn out to be fruitful.

Writing these lines would be impossible without Markus Mobius. He put trust in me even before we met and he continued to do so for many years since. His encouragement, support and friendship was the most important thing that kept me on the course of doing research.

Working with John Stachurski was an experience that inspired me to pursue graduate studies. In fact, it was much more than that: the confidence, friendship and the feeling of being treated as equals at a young age was a truly life-changing experience and I only hope that I can be so generous to someone else in the future.

I am also extremely lucky to have had the chance to work with Susan Athey. It has been an exceptionally motivating experience and have contributed vastly to pushing my professional limits.

I am very grateful to the examiners of my thesis, Sergey Lychagin and Joan Calzada who gave lots of constructive comments that helped me improve the quality of my papers.

The Central European University provided an academic environment in which I could freely work on whatever I felt was interesting and gave all the support I needed. I feel very lucky that I could have been part of this community.

I worked with Ádám Szeidl for years and his professionalism without compromises was always very stimulating. He was also very encouraging and supportive and his door was always open for me to share my ideas or concerns.

From my fellow PhD students I want to thank Bálint Menyhért, Kinga Marczell, Dzsamila Vonnák, Anna Adamecz-Völgyi, Péter Zsohár, Márta Bisztray, Ildikó Magyari, Ran Shorrer and János Divényi for all the ideas, discussions and friendship we could share together.

I am grateful to Veronika Orosz, Márta Jombach and Katalin Szimler, staff members at the Department of Economics and Business, who have always been very helpful and their professionalism was incredibly comforting. I am also thankful to Eszter Tímár at the Central for Academic Writing for reading my papers and for being full of encouragement.

I spent two summers as intern at Microsoft Research New England. I am grateful for their hospitality and for the extremely inspiring research environment I was part of there.

The community of the Department of Mathematics at the Corvinus University of Budapest was an environment in which I felt at home during my undergraduate studies. I have met professors who selflessly devoted a lot of time on me and provided me with very high level education. In particular I want to express my gratitude towards Gyula Magyarkuti, Zoltán Kánnai, Csaba Puskás, Imre Szabó and Péter Medvegyev.

The love of my family was essential for me to complete this thesis. My parents, grandparents and my brother stood behind me at all times for which I am ever grateful.

The most important compaion in this journey has been my wife. Her unconditional love helped me through all the depths that I faced. She always trusted my decisions and her endless patience and understanding gave me freedom to do whatever I felt was right.



# Contents

<b>List of Tables</b>	<b>x</b>
-----------------------	----------

<b>List of Figures</b>	<b>xii</b>
------------------------	------------

<b>1 Position effects in online news reading</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Related literature . . . . .	3
1.3 Data . . . . .	6
1.3.1 Browsing data . . . . .	6
1.3.2 Frontpage data . . . . .	6
1.3.3 Linking the two datasets . . . . .	7
1.4 Empirical strategy . . . . .	7
1.4.1 Linear panel estimation . . . . .	8
1.4.2 Time trend of interest . . . . .	9
1.5 Data preparation . . . . .	10
1.5.1 Choice of the newspaper: The New York Times . . .	10
1.5.2 Coding of position variables . . . . .	11
1.5.3 Time variables . . . . .	13
1.5.4 The choice situation . . . . .	15
1.6 Results . . . . .	15
1.6.1 Estimation results . . . . .	15

1.6.2	Magnitude of the effects . . . . .	19
1.6.3	Article specific constants . . . . .	22
1.7	Choosing multiple articles . . . . .	26
1.7.1	Binary logit . . . . .	27
1.7.2	Model-implied number of clicks . . . . .	28
1.8	Robustness checks . . . . .	30
1.9	Conclusion . . . . .	34
<b>2</b>	<b>The Impact of Aggregators on Internet News Consumption: <i>joint with Susan Athey and Markus Mobius</i></b>	<b>36</b>
2.1	Introduction . . . . .	36
2.2	Empirical Model and Data Description . . . . .	41
2.2.1	Theory . . . . .	41
2.2.2	Empirical Implementation and User Matching . . .	44
2.3	Google News and Overall News Consumption . . . . .	48
2.3.1	Estimating the Volume Effect . . . . .	49
2.3.2	Volume Effect by Outlet Size . . . . .	52
2.4	Decomposing the Volume Effect . . . . .	55
2.4.1	News Characteristics . . . . .	56
2.4.2	Differences Between Treatment and Control Users .	59
2.4.3	Referrals from Google News versus other Referral Modes, by Characteristic . . . . .	60
2.4.4	Volume Effect By News Characteristics . . . . .	63
2.4.5	Decomposition . . . . .	64
2.5	Conclusion . . . . .	67

<b>3</b>	<b>Fitted Value Function Iteration with Probability One Contractions: <i>joint with John Stachurski</i></b>	<b>70</b>
3.1	Introduction . . . . .	70
3.1.1	Methodology . . . . .	72
3.1.2	Related Literature . . . . .	76
3.1.3	Outline . . . . .	78
3.2	Preliminaries . . . . .	78
3.3	Set Up . . . . .	79
3.3.1	The Model . . . . .	80
3.3.2	Value Function Iteration . . . . .	81
3.4	Random Fitted VFI . . . . .	82
3.5	Analysis . . . . .	83
3.6	Rates of Convergence . . . . .	85
3.6.1	Donsker Classes . . . . .	85
3.6.2	The Lipschitz Case . . . . .	86
3.6.3	The Monotone Case . . . . .	87
3.7	Applications . . . . .	88
3.8	Conclusion . . . . .	92
3.9	Proofs . . . . .	92
	<b>Bibliography</b>	<b>99</b>
<b>A</b>	<b>Appendix for Chapter 2</b>	<b>105</b>
A.1	Additional tables . . . . .	105
A.2	Browsing Data Processing . . . . .	105
A.2.1	Active Users . . . . .	105
A.2.2	Browsing Stream . . . . .	108
A.2.3	Canonical URLs . . . . .	108

A.2.4	Landing and Article Pages . . . . .	109
A.2.5	News Minisessions and Referral Modes . . . . .	109
A.3	Topic Classification . . . . .	110
A.3.1	Data . . . . .	110
A.3.2	Pre-processing . . . . .	110
A.3.3	Newspaper Article to Wikipedia Concept matching	111
A.3.4	Pure Macro Topics . . . . .	112
A.3.5	Topic Augmentation . . . . .	113
A.3.6	Supervised cleaning . . . . .	114

# List of Tables

1.1	Descriptive statistics on choice situations . . . . .	16
1.2	Regression results with the default settings. . . . .	17
1.3	Summary statistics on the distribution of vertical position and choice share . . . . .	20
1.4	Summary statistics on the distribution of vertical position and choice share . . . . .	21
1.5	Mean and median of article-specific constants . . . . .	24
1.6	Counterfactual distribution of click-through rates . . . . .	25
1.7	Distribution of number of articles read per person per front- page. . . . .	27
1.8	Regression results from the binary logit specification . . . .	28
1.9	Effect sizes for the binary logit specification . . . . .	29
1.10	Distribution of number of articles read per person per front- page, for the observed and simulated data. . . . .	30
1.11	Robustness checks: continuous position model . . . . .	31
1.12	Robustness checks: within sections model . . . . .	32
1.13	Descriptive statistics about time spent on the frontpage before making a choice . . . . .	34
2.1	Differences between matched treatment and control users .	46
2.2	Volume difference measures for news types according to referral shares and pageview types (number of pageviews). 50	
2.3	Volume difference measures for news types according to referral shares and pageview types (dwelltime). . . . .	51

2.4	Volume effect calculations (number of pageviews). . . . .	52
2.5	Mean of user-level Herfindahl indices for outlets within a topic (pre-period, treatment users). . . . .	54
2.6	Volume drop calculations (dwelltime). . . . .	55
2.7	Differences between treatment and control users on news types (top 50 topics) . . . . .	60
2.8	Share of page views originating from different referral modes by user-scarcity . . . . .	62
2.9	Estimating the log-linear structural model for volume drop by news characteristic . . . . .	65
3.1	Distance between successive iterates of fitted value iteration	90
3.2	Approximate and analytical second moments after 5 iter- ations . . . . .	91
A.1	Differences between treatment and control users on news types (top 50 topics), $2^5 = 32$ news categories . . . . .	106
A.2	Share of pageviews in top topics by referral mode, treat- ment users. . . . .	107

# List of Figures

1.1	Coding of frontpage regions . . . . .	11
1.2	Map of the frontpage to $(x, y)$ coordinates . . . . .	12
1.3	Bulleted sub-feature . . . . .	13
1.4	Article readership through time . . . . .	14
1.5	Estimated effect of vertical position . . . . .	18
1.6	Interest decay function . . . . .	19
1.7	Distribution of article constants by number of frontpage appearances . . . . .	23
1.8	Distribution of vertical position for 1st, 2nd and 3rd choices	26
2.1	Timeline for matching Spanish control and treatment users	44
2.2	News consumption of the treatment and control groups over time . . . . .	48
2.3	Referral shares, total news consumption. . . . .	49
2.4	Referral shares and news reading volume as a function of time after publication . . . . .	61
2.5	Cumulative reading shares of top topics for referral modes direct navigation and Google News. . . . .	62
2.6	No-substitution counterfactual decrease and actual vol- ume decrease by breaking news and user scarcity. . . . .	64
2.7	No substitution counterfactual decrease and actual vol- ume decrease by popularity and broad topic. . . . .	66

3.1	Nonexpansive approximation with a radial basis kernel smoother . . . . .	74
3.2	Expansive approximation with Chebyshev polynomials . .	75
3.3	Instability in fitted VFI with Chebyshev polynomials . . .	75
3.4	25 elements of the sequence $(AR_n)^k w$ . . . . .	89



# Chapter 1

## Position effects in online news reading

### 1.1 Introduction

How much influence do newspaper editors have on what people read? The answer seems easy: as much as they want, since if they do not write on a certain topic, then people do not read about it at all in the newspaper. However, choosing the articles is not the only choice editors face. They also have to decide how to position news articles on a frontpage. This paper sets out to measure how much positioning matters in terms of how large a readership a certain article attracts.

The design of the frontpage may serve multiple strategic purposes for a newspaper. The frontpage visually suggest to readers what the newspaper considers important: frontpage sets an agenda for readers. It also plays an important role in differentiating the newspaper from other publications<sup>1</sup>, furthermore, forces of competition can also influence frontpage design (Kenney and Lacy (1987)). Also, the frontpage is an important place for online newspapers to sell display advertising, thus a well-designed frontpage that helps keeping readers engaged may affect advertising revenues.

In the age of print newspapers article positions had to be decided upon once a day. With online news frontpages are rearranged multiple times, day and night: articles already present on the frontpage are rearranged,

---

<sup>1</sup> An example of how newspapers on different political sides choose to highlight or hide certain stories can be found at [http://krisztinaszucs.com/my-product/parallelreality\\_en/](http://krisztinaszucs.com/my-product/parallelreality_en/). Similar visualizations are found in Costanza-Chock and Rey-Mazón (2016).

new articles appear and older ones are taken down from the frontpage. Frontpage arrangement, just like story selection can help reach certain goals of the newspaper, let it be increasing readership, optimizing pricing of advertisements or promoting some issues. Therefore it is worthwhile to quantify the impact of physical positioning.

This paper works with a dataset that concerns only online news consumed through desktop computers. However, these days more and more people read news through mobile devices<sup>2</sup>. As mobile screens are much smaller than those used with laptop or desktop devices, fewer news pieces fit to the visible area of a homepage. Thus, the decision of which articles to put in the first positions seems even more crucial. Because of this motivation and also because there is larger variation to exploit, in my analysis I focus on the vertical dimension of positioning.

The main contribution of the paper is to bring empirical evidence from the field using a novel identification strategy. I use consecutive frontpages of The New York Times where most of the articles overlap but change in terms of their relative and absolute positions. This variation lets me dissect the effect of positioning from other factors that play a role in article choice. I combine the frontpage data with browsing information to get a measure of article popularity over time. That is, I can observe how many people choose to read articles in parallel with how their positions change in the frontpage. To my knowledge, this is the first paper to bring direct information on article-level news popularity to the aid of analyzing effects of newspaper layout.

An important endogeneity concern that requires such a dataset to identify position effects is that typically newspapers put the best articles to the uppermost positions of their frontpages. Observing changing positions at a high frequency and using article-specific constants in a multinomial logit framework lets me separate inherent perceived article quality (which may entail its topic, wording of the title, etc.) from the effect of positions.

The key finding of the paper is that positions significantly matter for how large a readership an article attracts, both in a statistical and in an

---

<sup>2</sup>Pew Research Center (2016) reports that the share of US adults who often read news on mobile devices rose from 21% in 2013 to 36% in 2016 while the same for desktop/laptop was 35% in 2013 and 33% in 2016.

economic sense. In line with intuition, all other things being equal, the lower an article is featured on the frontpage, the less readers it attracts. This effect is non-linear: the same position difference matters more for the upper parts of the frontpage than in the lower parts. Imagining the situation when we have two otherwise identical articles in the first two positions, the first one is predicted to attract 26 to 44% more readers than the second one depending on the precise location.

The estimated empirical model also lets me perform counterfactual calculations that show how well newspapers use positioning as a way to get visitors to read news articles. Assuming an outside option of not choosing any articles from those offered on the frontpage, newspapers could increase the click-through rate by 5% on average by rearranging articles to different positions. Another calculation shows that the observed article positions result in a click-through rate that is much closer to that of a random ordering than to the optimal counterfactual ordering.

The outline of the paper is the following. After reviewing related literature in Section 1.2, in Section 1.3, I discuss the two main data sources I rely on and how to combine them. Section 1.4 presents the empirical strategy that I choose to analyze position effects. I detail data preparation steps in Section 1.5, and Section 1.6 presents estimation results and counterfactuals. Section 1.7 analyzes the issue of multiple chosen articles; other robustness checks are conducted in Section 1.8. Section 1.9 concludes. In all regression outputs, \*, \*\*, \*\*\* mean significance at the 10, 5 and 1 percent levels respectively.

## 1.2 Related literature

This paper is related to a number of different literatures on news reading, ranging from economics, finance and marketing to pattern recognition and visual studies.

Position effects are analyzed empirically by Narayanan and Kalyanam (2015) in the context of search advertising. They use a regression discontinuity approach by comparing advertisements of the same quality differing only in their position of the advertisement listing. According

to their measurements, higher positions lead to increased click-through rates, although only in certain positions. The effects are not only statistically but economically significant as well. Their methodology is not readily applicable in our context as new articles appearing in subsequent frontpages make it hard to argue that the only source of discontinuity is the changing position of articles already present before.

In the economics literature the study by Alaoui and Germano (2016) develops a theoretical model of sequential news reading where different news outlets compete for the attention of readers. In their model the order in which news articles are presented is of first order importance regarding what people read. They show that in equilibrium readers do not typically read news articles that are optimal for them. This is a consequence of news outlets positioning news articles in a way that is optimal for themselves in a competitive environment.

Fedyk (2017) analyzes effects of news positioning on asset prices. Using quasi-random variation of news positioning in the top part ("frontpage") of the Bloomberg terminal's news page she finds enormous effects (280% on asset prices and 180% on trading volume) in the first 10 minutes after news publication for securities featured in frontpage news. The effect is also present in the longer run: information from non-frontpage news gets reflected in asset prices much slower (days later) than those from frontpage news. A further important finding is that news positioning has a stronger effect on asset prices than news importance, the latter being measured by editorial judgment.

Similarly to my paper, but with the focus on news aggregators, Dellarcas et al. (2016) measure factors that contribute to whether a reader clicks on a news article or not. Using field experiments they find that the presence of an image or a longer snippet from the article decreases click-through rates. My paper focuses more on positioning and uses other elements of presentation as controls.

Two papers use data on news popularity as measured by social sharing activity (both using data on The New York Times just like this paper). Berger and Milkman (2012) focus on the question of what makes some articles more viral than others. The study finds that controlling for external factors like positioning, articles evoking positive and/or high-arousal emotions are more likely to get viral. Position effects are intro-

duced as dummy variables for regions of the frontpage; I loosely base my categorization on theirs with adding continuous measures of positioning. I note that their categories also reflect the idea that it is vertical positioning rather than horizontal matters more for attention.

The outcome variable in Berger and Milkman (2012) is whether or not an article makes it to the "most emailed list" of The New York Times which is a measure of social sharing. Instead, in my paper I apply a direct measure of readership of articles. Articles that people share or talk about in social media are not necessarily those that people read the most; some topics are naturally more fit for social sharing than others.

This is underlined by Zubiaga (2013), who analyzes The New York Times to see how much alignment there is between editors' choices to put articles into prominent positions and reception of articles in the social media. He finds that "soft news" is more likely to be shared than "hard news" that typically occupies the best positions of the newspaper frontpages.

The topic of news popularity is also analyzed in a study from the pattern recognition literature by Hensinger et al. (2013). They examine factors that contribute to a news article becoming popular, their focus being on textual information (title and introductory descriptions). In my analysis, I implicitly control for this using article-specific fixed effects. As they retrieve data via RSS, they do not have any information on positioning - in fact, the authors mark position information as an avenue for further research.

The visual studies literature also offers interesting evidence on how readers focus their attention on online news reading. A comprehensive study by Leckner (2012) reports that readers typically look at the top parts of the page, especially the upper-left article. This is in line with this paper's findings. Using eye-tracking information, Bucher and Schumacher (2006) found that visual cues and layout guide readers' attention, highlighting the importance of how news is positioned on frontpages. Murphy et al. (2006) reports on the primacy effect in ordered link choice as well as about a recency effect (choosing the last available option); this paper corroborates the presence of a primacy effect.

## 1.3 Data

There are two datasets whose linking provides the data suitable for the purposes of the analysis. The first is browsing history data comprised of a selection of US internet users. The second dataset consists of frontpages downloaded about every 2-3 hours.

### 1.3.1 Browsing data

The browsing data comes from users who use Microsoft products for internet browsing and have consented to allowing their data to be used for research purposes<sup>3</sup>. The time period under analysis is January 2016.

From the browsing logs, I construct sessions of news reading in the same way as is detailed in Athey et al. (2017). This allows me to cleanly identify news browsing events that have the following pattern: a user arrives directly at the main frontpage of a newspaper and selects one or more articles to read. These browsing sessions have timestamps at the level of seconds. Although there is some noise, below I present evidence that it is a pretty precise measure of the actual time of accessing news articles (see Figure 1.4).

### 1.3.2 Frontpage data

A scraper that downloaded data from multiple newspapers was running about 8-10 times every day in January 2016. The goal of the scraper was to record what the frontpages of the newspapers looked like and how they changed over time. There were occasions when the scraper failed for some or all the newspapers but the success rate of the exercise is very high.

Upon downloading a frontpage, it was rendered locally by simulating a browser in order to reconstruct the layout and to extract links from the frontpage. This resulted in a structured dataset that showed the vertical

---

<sup>3</sup>The data is subject to stringent privacy restrictions and at all times resides only on secure servers, and only aggregate statistics and the output of statistical models can be reported. However, we are able to construct the variables for analysis using the fully disaggregated data.

and horizontal coordinates of a link in a page, as well as the fontsize of the displayed text. As many times multiple links exist to an article in close proximity, we can also infer the presence of images.

A caveat of the data recorded is that the exact time of saving the scrapes was not recorded, only a timestamp that indicates the hour when the scraper started. This presents a challenge for linking the frontpage data with the browsing data.

### 1.3.3 Linking the two datasets

There are two aspects of the datasets that require care in order to assign browsing events to articles placed on frontpages.

First, the links of the articles appearing in the browsing data have to be matched to those extracted from frontpages. This can be done relatively easily by using a method developed in Athey et al. (2017): URLs are stemmed and are brought to a *canonical* form that can be used for matching.

Second, browsing events have to be tied to frontpage versions. This is crucial as identification of position effects ultimately comes from people browsing at different times seeing different frontpage layouts. The goal is to determine time windows in which a certain frontpage version was the one people actually saw and made a choice from.

In theory, frontpages can change continuously over time. I could only observe exact changes if I had access to very frequent scrapes. The situation is further complicated by having only an approximate time of the scrape recorded. Therefore, I am relying on rules (some rule-of-thumb, some data-driven) to identify the necessary time windows. The strategy is to show that even though I have to rely on approximations to construct the time windows, the conclusions of the analysis are excessively dependent on the particulars of the method I am using.

## 1.4 Empirical strategy

I model the decision of a reader arriving at the frontpage as a situation with discrete choices. The articles presented on the frontpage constitute

the choice set. I use the multinomial logit framework for calculations.

Choice here is modeled conditional on choosing an article, that is, without explicitly taking into account the possibility of choosing an outside option of not reading any articles. The independence of irrelevant alternatives (IIA) property of the logit framework allows me to do so without hurting identification of position effects (see, for example, Train (2009)). In essence, only relative differences between properties of articles matter for choice: ignoring or including the outside option does not change the relative choice probabilities of news articles present in frontpages. The outside option choice is relevant to the model when we want to measure the effect of counterfactual changes on the total number of people making a choice; we will return to it in Section 1.6.3.

The utility that consumer  $n$  is obtaining from article  $j$  in the frontpage version  $m$  is

$$U_{njm} = \alpha_j + \beta X_{jm} + \varepsilon_{njm}$$

where  $\varepsilon_{njm}$  have Type-I extreme value distribution and are iid. Here  $\alpha_j$  is an article-specific constant and serves to represent attributes of the article that are constant across frontpages. Importantly this constant is capturing inherent (perceived) quality and general (time-invariant) interest in the article.  $X_{jm}$  represents attributes that vary across frontpages: most importantly it embeds position information of articles on frontpages.

Choice probabilities have the well-known form of

$$P_{jm} = \frac{\exp(\alpha_j + \beta X_{jm})}{\sum_{k \in A_m} \exp(\alpha_k + \beta X_{km})} \quad (1.1)$$

in a certain frontpage  $m$  where  $A_m$  denotes the set of article (indices) that are present in frontpage  $m$ .

### 1.4.1 Linear panel estimation

Estimating the model we introduced is usually carried out via maximum likelihood methods. However, when we have many variables this can be computationally burdensome. In our case the article-specific constants are numerous, hence it is helpful to look for some practical methods that can alleviate this problem.



The simple model we introduced implies that in expectation market shares ( $s_{jm}$ , choice shares of articles on frontpages) satisfy the following relation:

$$\log(s_{jm}) = \alpha_j + \beta X_{jm} + \underbrace{\log \left( \sum_{k \in A_m} \exp(\alpha_k + \beta X_{km}) \right)}_{\tilde{\zeta}_m} \quad (1.2)$$

Note that the final term on the right is frontpage-specific. This implies a linear panel model where “individuals” are articles and “time” are the various frontpages. The last term implies that both “individual” and “time” fixed effects have to be included to the model. The effect of positions is identified by within-article position variation across frontpages and their effects on view share of articles. I apply weighted regression using the number of users making a choice per frontpage.

#### 1.4.2 Time trend of interest

While the article fixed effect takes care of a certain form of endogeneity (position and inherent quality of articles are probably correlated), there is another natural candidate for possible omitted variable bias. Beyond the time-invariant quality, articles also differ in how fresh the story is that they write about. Even if someone has not yet read the newspaper on a certain day, she may have heard news from other sources (other newspapers, radio, etc) and may no longer have additional motivation to read about a relatively older story.

To control for this we can extend the model introduced before as follows:

$$U_{njt} = \alpha_j + \tilde{\zeta}_m + g(t - t_j) + \beta X_{jt} + \varepsilon_{njt}$$

where  $t$  denotes absolute time and  $t_j$  is the publication time of the article (in some common unit).  $g$  is a function that captures the phenomena of *interest decay*. Typically we can think of it as being decreasing, however, it is ultimately an empirical question to decide.

The decay function  $g$  is assumed to be the same for all articles. An extension of the model could make it article specific (similarly to the constant  $\alpha_j$ ) or at least to be specific to the topic of the article.

Time is considered here to be continuous whereas the frontpage model formulated before looked at frontpages as discrete units. We can use an approximation of the continuous model by assuming that for the time period when the frontpage was present the decay term is constant. Take frontpage  $m$  which is valid from time  $t_1^m$  to  $t_2^m$ . By default we simply assume that the interest decay is  $g((t_2^m - t_1^m)/2 - t_j)$ . If the article was published during the time of the frontpage interval ( $t_j \in [t_1^m, t_2^m]$ ), I take the decay to be  $g(0)$ .

Substituting this to the utility we arrive at

$$\log(s_{jm}) = \alpha_j + \xi_m + g((t_2^m - t_1^m)/2 - t_j) + \beta X_{jm}$$

In practice I use a fifth-order polynomial to approximate the unknown function  $g$ . With this choice the model is still linear in parameters so standard fixed effects estimation routines can be used.

## 1.5 Data preparation

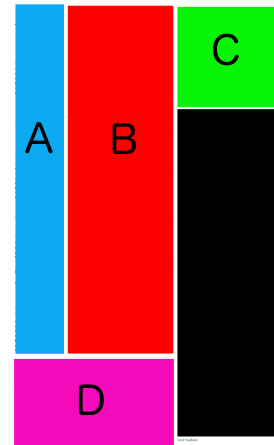
### 1.5.1 Choice of the newspaper: The New York Times

I choose The New York Times as the newspaper to be analyzed for the following reasons. First, The New York Times has a large enough readership so that there are enough readers for the various frontpage versions in my data. Second, the scraper worked well for The New York Times: based on manual inspection, the structured data resulting from the scraper very nicely mirrors the positions of the articles in the downloaded screenshot and there are practically no articles in the main sections that have not been recorded to the structured data.

Furthermore, unlike for some news sites that use different technology, personalization does not interfere with the empirical exercise. For example, on Yahoo News users can indicate their preferences, based on which the frontpage becomes individual-specific. Such personalization makes it very hard to argue that the scraped frontpages are the ones that people are actually seeing when making choices of what to read. This is not an issue for The New York Times.



(a) A frontpage



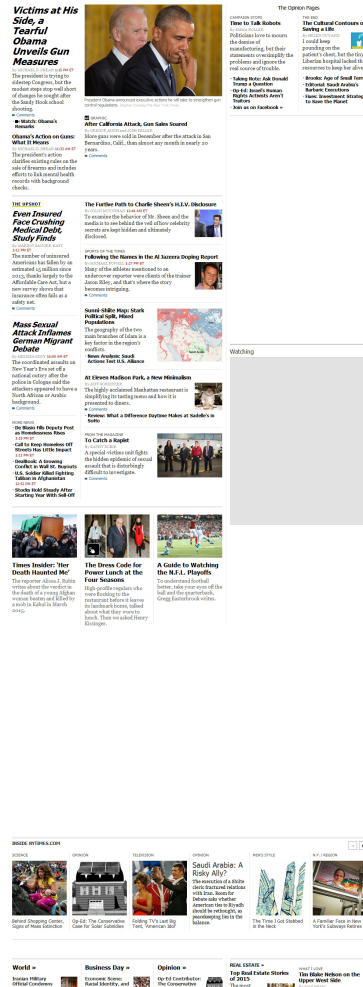
(b) Frontpage regions

Figure 1.1: Coding of frontpage regions. A: left panel, B: middle panel, C: opinion panel, D: lower panel, E: lowermost stripe. Black regions are not considered for the analysis. The white region in the lower-middle part is for videos and is also excluded from the analysis.

## 1.5.2 Coding of position variables

Frontpages of newspapers are rather long and may contain multiple links to the same articles. However, parts of the frontpage are not looked at equally frequently; typically, lower parts of the frontpage are browsed less often. In order to get meaningful position effects, I focus on parts of the frontpage that are in the upper region of the whole frontpage.

I divide the frontpage into larger, well-defined regions, loosely following Berger and Milkman (2012). An example is shown in Figure 1.1. The black region in the bottom of the page shown represents the lower part of the frontpage that I do not consider for analysis. Another black region is a live feed-type section on the right; the scraper could not reliably extract links from this part of the frontpage. Articles shown in the non-black regions constitute the subject of the analysis. In particular, the main focus will be on regions A and B, the left and the middle panels.



(a) A frontpage from January 5th, 2016

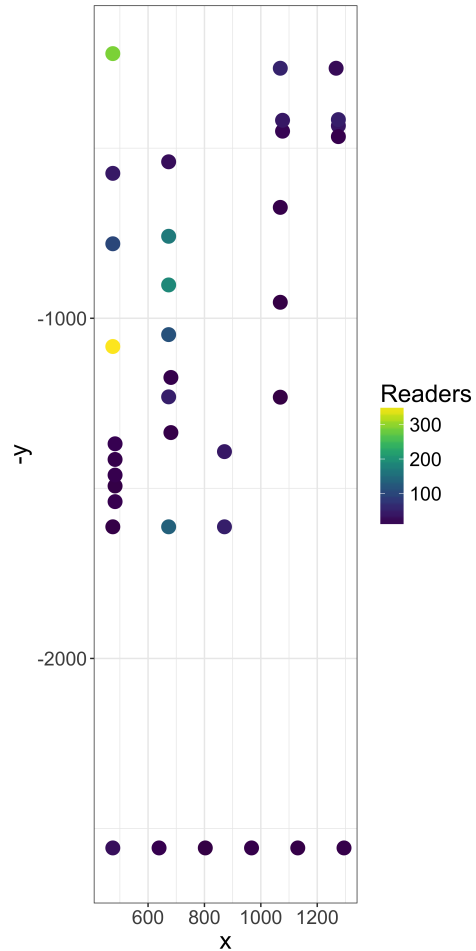
(b) The frontpage coded to  $(x, y)$  coordinates

Figure 1.2: Map of the frontpage to  $(x, y)$  coordinates (not normalized). The color expresses how many users accessed the article in the time window pertaining to the frontpage.

The most direct measure of position I use are the horizontal ( $x$ ) and vertical  $y$  coordinates of an article on the frontpage. A visual example of how articles on the frontpage relate to the coded positions is shown in Figure 1.2. I normalize these measures linearly to a 0-1 scale in the region of the frontpage I consider for analysis.

In addition to the position measure, I also code other elements of the visual presentation:

1. **fontsize**: the size in which the main title of the article is presented
2. **images**: whether or not the article is presented with an image besides having a title and a summary



Figure 1.3: The article whose link is highlighted in yellow is coded as a “bulleted sub-feature”

3. “bulleted sub-features”: there are some articles that are displayed as related to other ones, supplementing them or elaborating on related themes. These, unlike “regular” articles, are only presented with one sentence as a bullet point to another article. An example is shown in Figure 1.3.

### 1.5.3 Time variables

**Publication time of articles.** The exact time when articles were published on the homepage are not recorded. One way to extract such information would be to go to the article online and read the timestamp of publication. However, this is not reliable due to minor updates and edits happening after publication.

I choose a data-driven way to infer publication times of articles. For each quarter-hour I count the number of unique users who access the article. While generally the time of pageview recorded in the raw data is of very good quality, there is some noise that should be filtered. Were this not the case, I could simply use the time of the first pageview for the article as the publication date. Taking noise into account, publication time is defined as the first quarter hour when the number of unique readers is above a certain threshold.

In Figure 1.4, I show the number of unique readers per quarter hour over time for two articles. First, the daily cyclical of news reading is nicely visible: news reading activity is concentrated to daytime hours. Second, there is a sharp jump in both figures: the jump roughly marks the time of appearance on the frontpage, whereas pageviews before the jump can be considered as noise.

As for more popular articles, the noise can be larger in absolute terms, I choose different thresholds for articles of different popularity. By de-

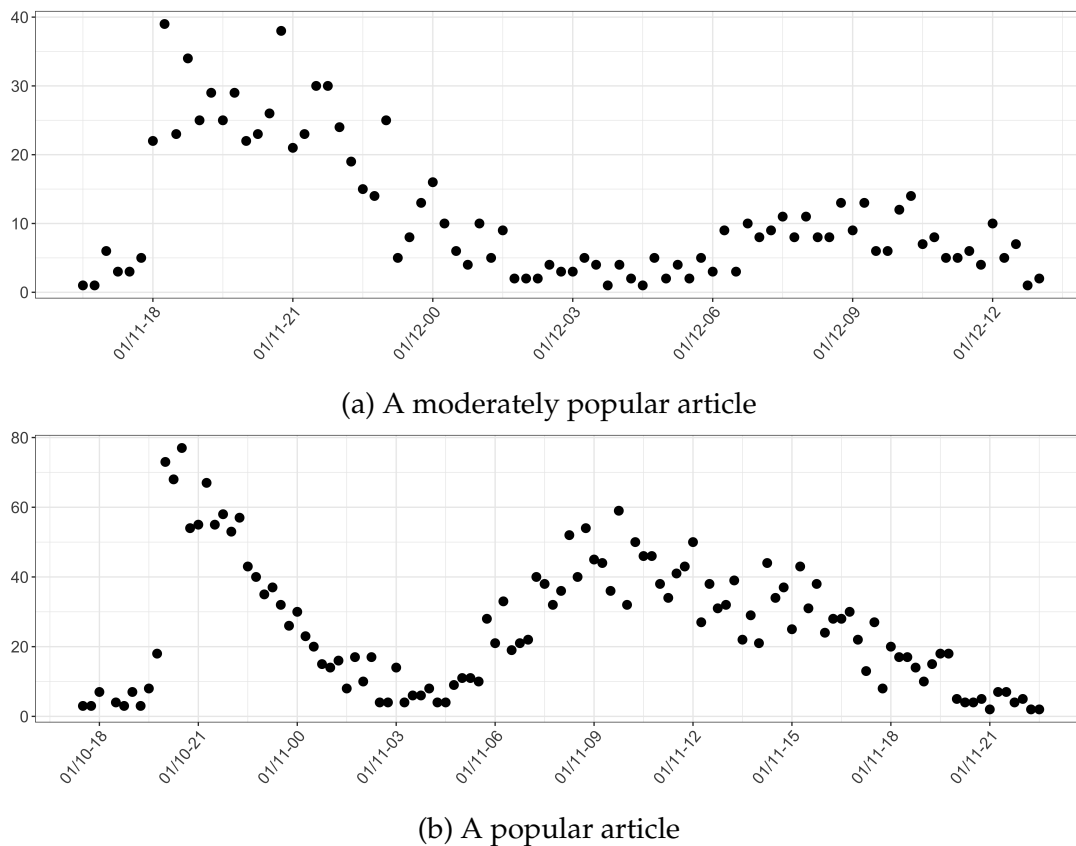


Figure 1.4: Unique people reading articles per quarter hour in time.

fault I require 5 readers for articles below 200 total readers and 10 above. Robustness to this parametrization is examined later.

**Time windows for frontpages.** As noted above, the exact times between a recorded frontpage was the one readers saw when having arrived at the homepage of The New York Times is not known. I only have an approximate time when the frontpage was downloaded, which is only precise to the hour level. To overcome the lack of data, I rely on a simple rule to determine time windows for frontpages: I take the timestamp and put a  $[-45 \text{ minutes}, +75 \text{ minutes}]$  bound around it to get a time window. For example, the frontpage with timestamp „2016010513” (standing for 2016 January 5th, 1 PM) is assumed to have been valid from 12:15 PM to 2:15 PM on January 5th, 2016.

This method is admittedly a crude approximation and calls for robustness analysis, which I perform in Section 1.8.

### 1.5.4 The choice situation

Users may read more than one article when arriving at the frontpage of a newspaper and they may do so multiple times per day. In order to simplify the analysis, I only use the first of the following events per user per day: the user goes directly to the frontpage of the newspaper and chooses an article. That is, by default I neither use the second, third, etc. articles read after opening the frontpage, nor subsequent "open frontpage, choose an article" events later in a day. Later, I check for robustness of results: I am considering allowing for more than one choice per frontpage and also choice on multiple frontpages per day (see Sections 1.7 and 1.8).

I filter the data I am using for estimation in two further ways. First, I exclude data from January 23rd to 25th. During this time a heavy snowstorm hit New York City; news was dominated by reports on this event and the structure of the frontpage also changed somewhat due to this. Second, I only use articles that were chosen by at least 0.5% of people arriving at the particular frontpage. Robustness to this threshold is analyzed in Section 1.8.

After applying these data cleaning steps we arrive at a sample of 232 frontpages with a total of 1,275 articles. The median number of choices happening on a frontpage is 809; in total, 196,001 choices happen altogether on the frontpages. The median number of articles on a frontpage is 26 and a typical article appears on 4 frontpages while the longer-lasting ones appear on as many as 9 pages. These descriptive statistics along with other ones can be seen in Table 1.1.

## 1.6 Results

### 1.6.1 Estimation results

I present results from estimating the model using two different codings of positions. In the first versions, I use the continuous measure of  $x$  and  $y$  positions of articles. Then I use dummy variables for the various regions. Finally, I combine these two and assess the effect of vertical positioning within the two main sections (left and middle panes).

Number of articles	1,275
Number of frontpages	232
Number of all choices	196,001

	10%	25%	50%	75%	90%
Articles per frontpage	21	23	26	28	29
Readers per frontpage	240.1	547.5	809	1,153.5	1,410.5
Frontpage appearances per article	1	2	4	6	9

Table 1.1: Descriptive statistics on choice situations. In the second table, percentiles of the distributions are shown.

The main focus of the analysis will be the vertical dimension of positions. I make this choice for a number of reasons. First, there is much larger variation in this dimension than in the horizontal, especially within sections. Also, vertical positioning gets more and more important with the increased use of mobile devices; while I do not have access to mobile news reading data, I believe my findings on the vertical dimension have important implications for that domain as well.

Estimation results can be seen in Table 1.2. In specification (1) I only use the vertical and horizontal coordinates as explanatory variables together with their squared values. The coefficients of interest are that of  $y$  and  $y^2$ : the first order term has a negative coefficient that is larger in absolute value than that of the second order term. As  $y$  is normalized to be between 0 and 1, this means that there is a negative effect of vertical positions news popularity. This effect is decreasing in absolute value when we go lower and lower on the frontpage. I highlight this finding since it carries through to most of the different models and robustness checks that I examine in the sequel. See Figure 1.5 (a) for what the effect looks like as  $y$  varies.

In the second specification, I add other explanatory variables: fontsize, dummies for whether an article has an image, whether it is a bulleted subtitle and for the decaying interest in time after the article was published. I use a 5th-order polynomial of time elapsing after publication to control for the time trend.

It is visible that the coefficients of interest decrease in absolute value af-



Table 1.2: Regression results with the default settings.

	<i>Dependent variable:</i>			
	log(share)			
	(1)	(2)	(3)	(4)
x	−0.08 (0.11)	0.28*** (0.11)		
y	−4.08*** (0.16)	−2.88*** (0.16)		
$x^2$	−0.39*** (0.11)	−0.50*** (0.10)		
$y^2$	2.58*** (0.17)	1.53*** (0.16)		
Fontsize		0.02*** (0.004)	0.03*** (0.01)	0.02*** (0.005)
Has image		0.10*** (0.04)	0.17*** (0.04)	0.13*** (0.04)
Bulleted title		−0.44*** (0.04)	−0.49*** (0.04)	−0.49*** (0.04)
Lower panel			0.27*** (0.04)	0.27*** (0.04)
Opinion panel			0.97*** (0.05)	1.05*** (0.05)
Middle panel (top)			1.21*** (0.06)	
Middle panel (non-top)			0.71*** (0.04)	
Left panel (top)			1.07*** (0.09)	
Left panel (non-top)			0.59*** (0.06)	
Left panel				1.25*** (0.09)
Middle panel				1.49*** (0.06)
Left panel * $y$				−3.16*** (0.42)
Left panel * $y^2$				3.07*** (0.80)
Middle panel * $y$				−3.50*** (0.44)
Middle panel * $y^2$				2.44*** (0.88)
Interest decay	No	Yes	Yes	Yes
Observations	5,945	5,945	5,945	5,945
R <sup>2</sup>	0.78	0.82	0.81	0.82

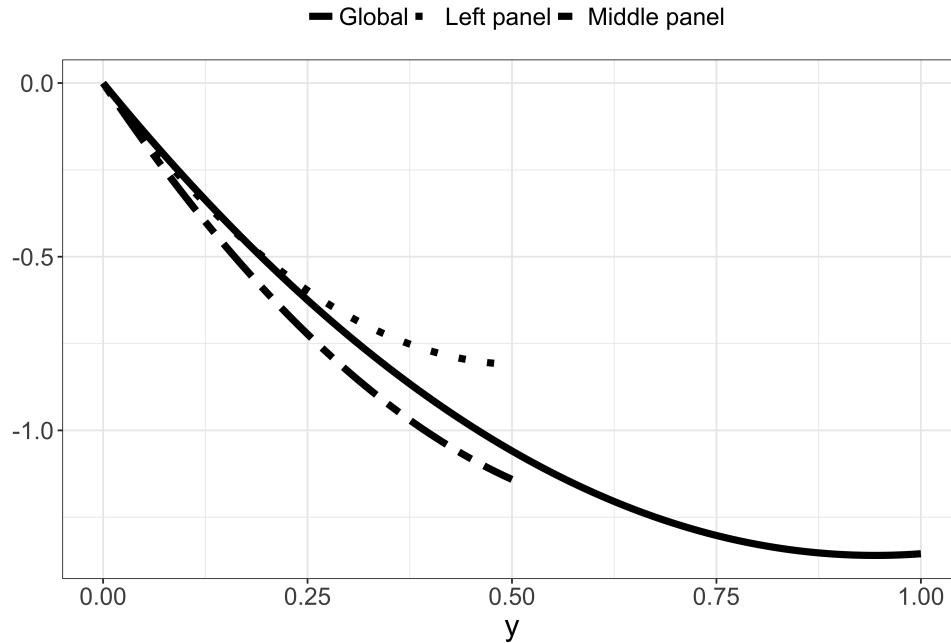


Figure 1.5: Estimated effect of vertical position: "Global" is from specification (2) of Table 1.2, "Left panel" and "Middle panel" are from specification (4).

ter these controls are added. Coefficients of the controls are significant at high significance levels and model fit is increased as well: these all confirm that these aspects of news popularity are important to consider. Nevertheless, the main finding of the negative, decreasing strength effect of vertical positioning holds up.

Coefficients of the other visual elements are significant and have intuitive signs. Images positively influence choice: having one beside the article improves choice probability by 10.5% all else being equal. Font size has a similar effect: for example, if a 16pt font is used instead of a 12pt for the headline, *ceteris paribus* the choice probability is bigger by 8.3%. Articles put into bulleted subtitles have quite low predicted choice probabilities; they are also typically found in lower sections of the page.

Finally, the estimated interest decay function in Figure 1.6 indeed fulfills its intuitive role: even though it was not imposed in the estimation, it is estimated to be a decreasing function.

Specification (3) uses dummy variables for the discrete position regions I defined above (the base category being the lowermost stripe of articles with images, see Figure 1.1 (a)). Separate dummy variables are added for the top position of the left and middle panes. Dummies have again intuitive ordering: top positions dominate lower ones within both main

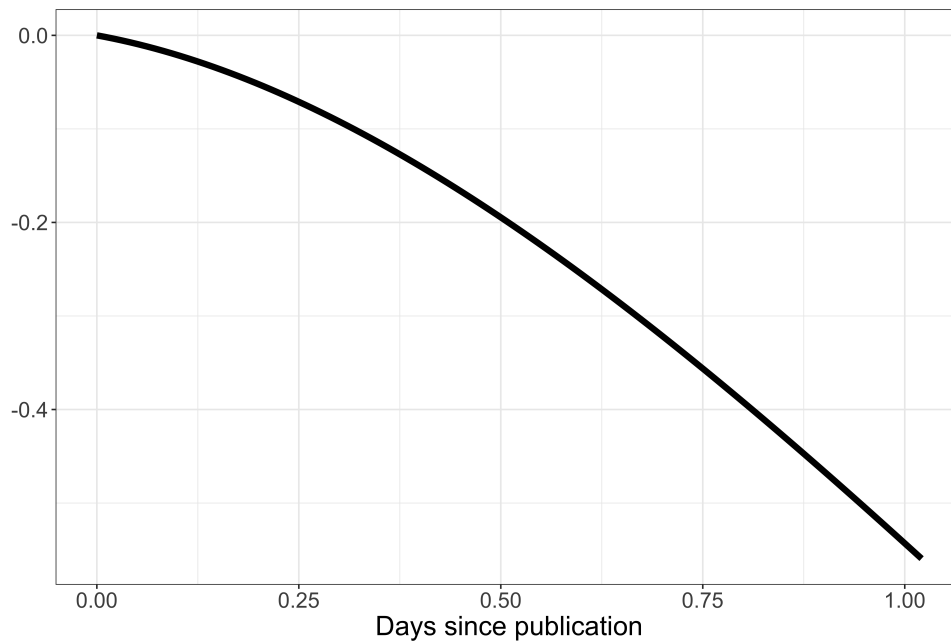


Figure 1.6: Interest decay function from specification (2) of Table 1.2, a 5th order polynomial. Plotted until the 95th percentile of time elapsed since publication for all articles on all frontpages.

panes, and generally lower position regions have smaller effects on view share.

Finally, column (4) of Table 1.2 focuses on the two main vertically positioned regions, the left and middle panes (taking both the top and the other articles into account). I examine the effect of vertical positioning within these sections: the effects are displayed in Figure 1.5. The estimated effect in the middle panel is somewhat stronger than in the left panel and the two surround the global estimate from specification (2). Remarkably, the general finding of a negative and decreasing strength effect holds up within both of the main panels.

### 1.6.2 Magnitude of the effects

We have seen that position effects are statistically significant, however, their economic significance is yet to be established. As we use a nonlinear function of the vertical position in the logit model, the magnitude of the effects depends on which position we examine.

The most straightforward way to assess the magnitudes of effects is to answer the following question: *Assume two articles are identical except that*

rank	p25	median	p75	rank	p25	median	p75
1	0.065	0.098	0.145	1	0.028	0.058	0.100
2	0.015	0.031	0.064	2	0.023	0.041	0.077
3	0.015	0.031	0.056	3	0.019	0.035	0.055
4	0.014	0.027	0.047	4	0.015	0.026	0.049
5	0.012	0.020	0.046	5	0.013	0.023	0.038

(a) Left panel, choice share

rank	p25	median	p75	rank	p25	median	p75
1	0.000	0.000	0.007	1	0.011	0.081	0.098
2	0.118	0.133	0.161	2	0.136	0.159	0.178
3	0.154	0.201	0.248	3	0.181	0.210	0.237
4	0.221	0.272	0.335	4	0.222	0.260	0.290
5	0.277	0.330	0.384	5	0.268	0.307	0.343

(b) Middle panel, choice share

(c) Left panel, vertical position

(d) Middle panel, vertical position

Table 1.3: Summary statistics on the distribution of vertical position and choice share for first five articles in left and middle panels.

*they are positioned differently on a frontpage. How does their relative popularity compare?*

In order to guide which positions to use to calculate effects, I took the top 5 articles from all frontpages in the left and middle panes. For each of the positions (1st to 5th) I calculated the 25th, 50th and 75th percentile of the distribution of their vertical positions and choice shares across all frontpages, see Table 1.3. I use the median values as the basis of assessing the magnitude of the position effects.

First, let us calculate the relative popularity of two hypothetical articles that are identical in all aspects except that one (A) is displayed in the first position of the left pane while the other (B) is in the second. Denote the quadratic effect of the vertical position with  $\hat{f}(y)$ . Then the relative popularity of the two articles is

$$\frac{P_A}{P_B} = \frac{\exp(\hat{f}(0))}{\exp(\hat{f}(0.133))} = 1.442 \quad (1.3)$$

That is, putting two articles with identical characteristics to the first two

ranks	popularity ratio	ranks	popularity ratio
1 and 2	1.442	1 and 2	1.256
2 and 3	1.155	2 and 3	1.142
3 and 4	1.130	3 and 4	1.127
4 and 5	1.078	4 and 5	1.105

(a) Left panel

(b) Middle panel

Table 1.4: Vertical position effects for hypothetical identical articles. A row shows the model-implied ratio of view share of two identical articles positioned in two consecutive positions, separately for the left and middle panels. Vertical  $y$  positions for the various ranks are median values taken from Table 1.3.

positions of the left column results in 44% more pageviews for the one placed to the first position relative to the second.

In a similar manner, we can calculate the difference for all positions in the left and middle panes. Table 1.4 shows the results for all consecutive positions in the left and middle panes. The effect is decreasing consistently with the previously seen decreasing effect of vertical positions. Magnitudes are fairly large: starting from 44 and 26 percent for the first two positions, even at the 4th to 5th positions they amount to about 8 to 11 percent.

There is quite large difference between the estimated effects for the first two positions in the two main panel and it warrants some discussion. Table 1.3 gives hints why this may be the case. First, the typical first article of the left panel is positioned higher than that of the middle panel's (an example for this can be seen in Figure 1.2). Therefore, the first two positions of the two panels are not directly comparable. Second, the raw ratio of the median view share of a typical first and second articles is around 3 for the left panel and is only 1.4 in the middle panel.

To sum up, the effects I identify are economically significant and we can conclude that positioning influences popularity of news articles to a large extent.

### 1.6.3 Article specific constants

Article specific constants are used to represent time-invariant quality of news articles in the estimated model. When relying on the linear fixed effect estimation of the logit model, the value of these constants does not play a role. However, getting an estimate of the value of these article quality measures is desirable so that interesting counterfactual exercises can be carried out.

**Estimation of the constants.** In standard logit settings estimating choice-specific constants is usually done by using an iterative procedure known as *the contraction* (see, for example, Train (2009)). The method can be both used as a post-estimation procedure or as an integral part of estimating the model. A key element of the procedure is that each alternative in each choice situation has a different constant. A property of the logit model is that in this case there is a unique set of constants that equates predicted and actual choice shares of alternatives in each choice situation.

Our model is different from the standard setting in that I assume a single constant for each article common to choice situations in order to capture inherent quality of articles. This means that unless the data was generated by the model exactly, we are not going to be able to perfectly equate observed and predicted choice shares.

Using the idea from the standard situation I define a simple optimization problem where I aim to minimize the sum of squared deviations of observed and predicted choice shares:

$$\min_{\{\alpha_j\}} \sum_{j,m} w_m \left( s_{jm} - \frac{\exp(\alpha_j + \hat{\beta} X_{jm})}{\sum_{k \in A_m} \exp(\alpha_k + \hat{\beta} X_{km})} \right)^2$$

Here  $w_m$  are the number of users for each frontpage and  $\hat{\beta}$  is the vector of estimated parameters other than the article constants (from specification (4) in Table 1.2). That is, I look for the article constants in a post-estimation procedure in order to bring observed and predicted choice shares close.

I relied on numerical optimization procedures to solve the minimization problem. Although there are more than a thousand constants to be esti-

mated, the minimization arrives at a solution in only a few minutes on a standard laptop<sup>4</sup>. Correlation of the predicted and actual choice shares is above 0.9.

**Explorative analysis.** I interpret the estimated article-specific constants as a proxy for time-invariant article quality. It is plausible to assume that newspapers both have a good sense of which articles are good and which ones less so and that they also monitor popularity of articles in real time. If this is the case we can expect that better articles spend more time on frontpages. In order to explore this I categorize articles based on the number of different frontpages in which they appear and calculate summary statistics of the article constants. Figure 1.7 shows that this hypothesis is correct: articles featured in more frontpages typically have higher estimated quality.

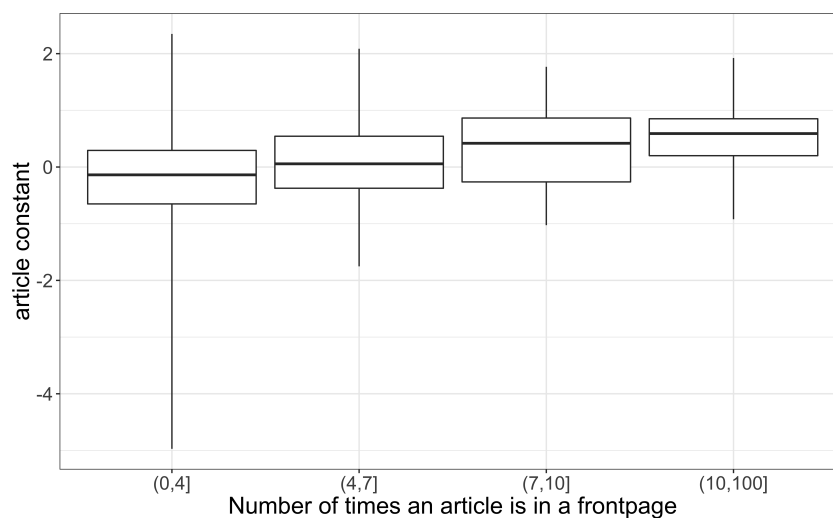


Figure 1.7: Distribution of article constants by number of frontpage appearances. 10th, 25th, 50th, 75th, 90th percentiles are shown.

Is it true that the best articles are the ones to be put to top positions? Table 1.5 shows the mean and median of article-specific constants for three groups of articles: 1) those that ever made it to the top position of the left or middle panes 2) other articles that were present in the left or middle panes 3) all others. As expected, articles in the top positions are

<sup>4</sup>I also created a simulation exercise to see how the procedure performs. With about 1.5 times more articles than in the actual data, assuming that the data was generated by the empirical model the minimization exercise successfully found the article-specific constants (with correlation with the originals more than 0.99). I relied on the NLOpt library's Sbplx algorithm to carry out the optimization (Johnson (n.d.), Rowan (1990)).

of very good quality and typically the two main sections contain articles of better quality.

	mean	median
left or middle panel articles	0.032	0.016
top position articles	0.215	0.187
other articles	-0.134	-0.122

Table 1.5: Mean and median of article-specific constants of articles that were ever positioned in the top of left or middle panes versus all others.

**A counterfactual exercise: rearrangement of articles.** Assume that there is an outside option of not choosing any of the articles and that the newspaper is interested in maximizing click-through. Under the logit model positive assortative arrangement is optimal: the newspaper maximizes the chance of any articles being chosen by ordering articles decreasingly compared to their utility without vertical position effects<sup>5</sup>.

From this point of view, do we observe an optimal positioning of articles by The New York Times? To get at this question, I calculate predicted utility of articles (using both the article-specific constants and estimates of the other parameters of the utility function) *without* the vertical position effect. I focus on the left and middle panels<sup>6</sup> and as a counterfactual exercise I reorder articles so that they follow the optimal order based on their calculated utilities without the vertical position effects.

Based on aggregate calculations, around 60% of people who arrive to the frontpage do not click on any articles. Thus, I assume that the share of outside option choice is 60% for each frontpage and ask: *If we changed the order of articles in the left and middle panels to the optimal order, by how much could we increase the 40% click-through rate?*

The simplicity of the logit framework allows me to answer this question easily. Denote the observed part of the utility of the outside option for

<sup>5</sup>The reason for this is that if  $X_1 > X_2$  and  $B > C$ , then  $\exp(X_1 + B) + \exp(X_2 + C) > \exp(X_1 + C) + \exp(X_2 + B)$ .

<sup>6</sup>The upper-right corner of the page where opinion articles are displayed is also a very important and popular part of the newspaper. However, there is less variation in vertical positioning there and the types of articles are also quite different from the two main panels.



frontpage  $m$  with  $V_{0m}$ . From the logit formula we can back out  $V_{0m}$ :

$$\frac{\exp(V_{0m})}{\exp(V_{0m}) + \sum_{k \in A_m} \exp(\hat{V}_{km})} = 0.6 \Rightarrow V_{0m} = \log \left( \frac{0.6}{1 - 0.6} \sum_{k \in A_m} \exp(\hat{V}_{km}) \right)$$

where  $\hat{V}_{jm}$  is the estimated utility of article  $j$  on frontpage  $m$ . Then, we can rearrange the articles to calculate counterfactual utility values  $\tilde{V}_{km}$  ( $k \in A_m$ ) and calculate the implied outside option choice share from the same formula.

As shown in Table 1.6, the mean and median click-through rate improvement is about 5 percent from the baseline. As this rearrangement only concerns the left and middle panes it can be regarded as a lower bound for gains from rearrangement. As all other visual elements and other properties (e.g., when articles are put to the frontpage and how long they are staying) are the same, this effect is quite substantial.

10%	25%	50%	mean	75%	90%
0.405	0.410	0.418	0.419	0.426	0.435

Table 1.6: Distribution of click-through rates across frontpages as a result of rearranging articles in the left and middle panes. The baseline share is 0.4.

To get another benchmark for the magnitude of this effect, imagine that the newspaper orders articles randomly in the left and middle panes. How large would click-through rate be in this case? I simulated 100 times this scenario and in each case I computed the mean click-through rate across frontpages. Then, in turn, I calculated the mean of these which is 39.3%. This means that the baseline click-through rate of 40% is much closer to a random ordering than to the 41.9% of the optimally ranked case. In light of this finding, the 5% click-through rate increase resulting from the optimal counterfactual reordering seems even larger.

I note that with other behavioral models like Alaoui and Germano (2016) it may not be always optimal to put the best article to the first place; the sequential search behavioral model there is different from the static choice model assumed here.

## 1.7 Choosing multiple articles

A strong assumption of the analysis presented is that only the first article chosen by an individual on a certain day is taken into consideration when creating the sample. In many cases there are other articles that are read as well.

This can undermine the identification of position effects. Consider the following model of news reading: the reader chooses all articles that she wants to read and reads them in the order of their appearance in the frontpage. In this case considering only the first one mechanically induces position effects even if in reality position has zero effect on what readers choose to read. To put it differently, we want to be confident that the position does not simply affect click order but rather preferences and choice.

A descriptive evidence that gives a warning sign about the issue is the following. If we take people who choose more than one articles and look at the distribution of vertical position according to the order of reading, we see that first articles' position is generally higher than that of the second and the third (see Figure 1.8). Ultimately it is a quantitative question to decide how strong a role this effect plays in the identified position effects.

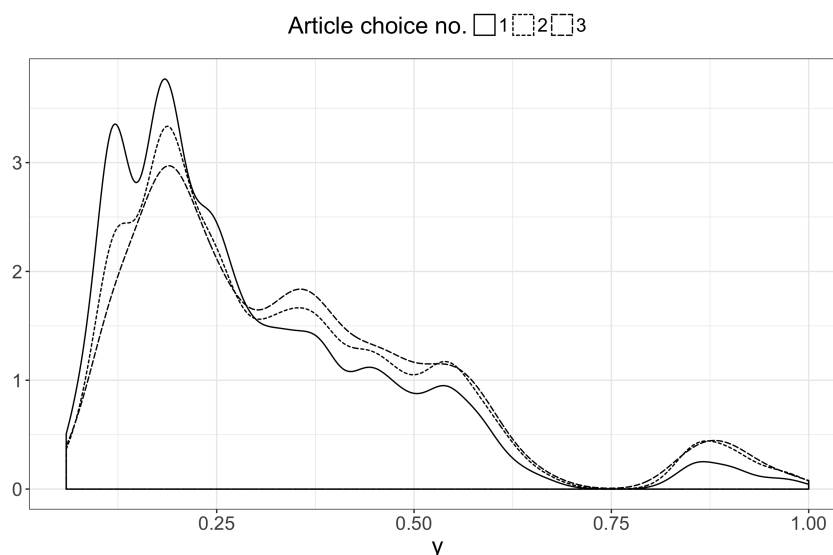


Figure 1.8: Distribution of vertical position for 1st, 2nd and 3rd choices of people who choose at least 2 articles to read.

I present two approaches that can be used to shed light on this issue.

### 1.7.1 Binary logit

One test for this is to re-define the choice situation as follows. Each individual arriving to the frontpage decides about each article whether to read it or not. Choice is modeled via binary logit, otherwise the behavioral model is the same as before (considering the same controls and functional forms). If position has no effect on preferences but only on click order, this would imply that position has no effect on choice probability in the current setup.

I still use only the first frontpage view event per person per day but now I allow decision makers to decide on each of the articles in the frontpage. The majority of people in the sample read only one article but there are some reading more (for the distribution see Table 1.7).

1	2	3	4+
0.87	0.10	0.02	0.01

Table 1.7: Distribution of number of articles read per person per frontpage.

The model of article choice can be summarized by the implied choice probabilities:

$$P_{jm} = \frac{1}{1 + \exp(-(\alpha_j + \beta X_{jm}))}$$

This results in a linear functional form for the log-odds ratio in expectation which, again, allows us to use the same linear fixed effects estimation procedure as in the case of the baseline model<sup>7</sup>:

$$\log \left( \frac{s_{jm}}{1 - s_{jm}} \right) = \alpha_j + \xi_m + \beta X_{jt} \quad (1.4)$$

Coefficients of interest from model specifications (2) and (4) are shown in Table 1.8. The main result of the negative and decreasing strength effect of vertical positions holds up. For easier evaluation, I also calculated the model-implied relative choice probabilities (Table 1.9) similarly to

<sup>7</sup> I apply frontpage-specific constants  $\xi_m$  just like in the multinomial case. There they served as an integral part of the model (they soak up the effects of varying denominators in the choice probabilities, see 1.2), while here they are used for consistency.

	<i>Dependent variable:</i>	
	log(share / (1 - share))	
	(2)	(4)
$y$	−2.39*** (0.12)	
$y^2$	1.10*** (0.10)	
Left panel * $y$		−2.44*** (0.28)
Left panel * $y^2$		1.86*** (0.39)
Middle panel * $y$		−3.03*** (0.27)
Middle panel * $y^2$		2.00*** (0.40)
Observations	5,941	5,941
R <sup>2</sup>	0.85	0.85

Table 1.8: Regression results from the binary logit specification. For comparison with the baseline model, see Table 1.2.

Table 1.4: effects of vertical positions are similar but somewhat milder compared to those obtained from the multinomial case. Nevertheless, I conclude that based on this exercise taking only the first article per frontpage does not lead to spuriously identified position effects.

### 1.7.2 Model-implied number of clicks

If the hypothesis that position affects click order but not preferences is true, the main multinomial specification relying on the first chosen articles would underpredict the number of subsequent articles chosen. The number of second, third, etc. clicks per reader per frontpage is a moment of the data that is not used for estimation. As such it is meaningful to explore how well the empirical model matches this quantity.

The steps I take to perform this exercise are the following:

1. Take the estimated coefficients (and article constants) and simulate

ranks	popularity ratio	ranks	popularity ratio
1 and 2	1.306	1 and 2	1.208
2 and 3	1.126	2 and 3	1.119
3 and 4	1.115	3 and 4	1.108
4 and 5	1.076	4 and 5	1.091

(a) Left panel

(b) Middle panel

Table 1.9: Effect sizes for the binary logit specification for identical articles differing only in their positions. For a comparison with the baseline results, see Table 1.4.

for each frontpage choices of decision makers.

2. Using the observed share of outside option choice of 0.6 I calibrate an outside option value that reproduces this share in the simulated data.
3. Calculate the number of articles for each frontpage and decision maker that exceeds the value of the outside option.
4. Compare this distribution to the observed distribution of the number of articles chosen by a person in a frontpage.

The distributional properties of the number of chosen articles coming from the simulation are shown in Table 1.10 along with the distribution observed in the data (repeated from Table 1.7). The two distributions are reasonably close to each other, if anything, on the contrary to the implication of the "pure click order" hypothesis, the simulation slightly overpredicts the number of multiple article choices. A reason for the thin tail of the observed distribution can be that people read articles not only by choosing from the frontpage but also by following links in articles. For example, if a reader reads an article A that links to article B and she reads it through the link, when she goes back to the frontpage she will presumably not choose B again, even though B might be of high utility to her. In the data I am only working with choices made directly from the frontpage that results in losing pageviews in the data that would require modeling of reading through article links explicitly.

Nevertheless, I conclude that the observed and simulated moments of

multiple article choices are reasonably close and confirm that position effects indeed affect choice, not only click order.

	1	2	3	4+
observed	0.87	0.10	0.02	0.01
simulated	0.77	0.19	0.03	0.004

Table 1.10: Distribution of number of articles read per person per frontpage, for the observed and simulated data.

## 1.8 Robustness checks

The analysis presented so far relies on a number of assumptions that I made during the data cleaning and processing phase. Since some of them may look somewhat arbitrary, I argue that none of these significantly influences the conclusions. I extract coefficients of interest in Tables 1.11 and 1.12 from model specifications (2) and (4) above respectively for all the robustness checks I conduct.

**Time window for frontpages** As the exact time interval when a frontpage was the actual one viewers saw is not known, I rely on approximations. As it is crucial to match pageview events to frontpages seen as closely as possible, I analyze robustness of the results to the heuristics in use.

In particular, instead of the default [-45 minutes, +75 minutes] bound put on the frontpage timestamp, I examine how coefficients of interest change when this bound is a) shorter: [-15 minutes, + 45 minutes], b) shifted: [-15 minutes, +105 minutes]. These can be read in Tables 1.11 and 1.12, columns (2) and (3) respectively. The qualitative properties of coefficients remain the same in both cases.

**Article publication time** Finding the publication times of articles is based on the browsing data rather than any external source. While it is desirable to be data-driven, the functional form used to determine the quarter-hour of publication is somewhat arbitrary. Thus, I examine the sensitivity of the estimates to a change in this function.

Table 1.11: Robustness checks: continuous position model

<i>Dependent variable:</i>									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log(share)								
y	-2.88*** (0.16)	-2.79*** (0.17)	-3.02*** (0.14)	-2.93*** (0.16)	-3.03*** (0.15)	-2.54*** (0.15)	-2.47*** (0.15)	-1.88*** (0.15)	-2.78*** (0.15)
y <sup>2</sup>	1.53*** (0.16)	1.52*** (0.17)	1.68*** (0.14)	1.61*** (0.16)	1.69*** (0.16)	1.55*** (0.15)	1.00*** (0.13)	0.79*** (0.15)	1.39*** (0.16)
Observations	5,945	5,789	5,989	6,079	6,006	3,755	8,488	6,069	5,945
R <sup>2</sup>	0.82	0.80	0.86	0.83	0.83	0.83	0.81	0.81	0.82

Table 1.12: Robustness checks: within sections model

	<i>Dependent variable:</i>								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Left panel * $y$	-3.16*** (0.42)	-2.82*** (0.44)	-3.67*** (0.38)	-3.22*** (0.42)	-3.33*** (0.42)	-2.70*** (0.31)	-3.74*** (0.38)	-1.74*** (0.42)	-3.16*** (0.42)
Left panel * $y^2$	3.07*** (0.80)	2.43*** (0.80)	3.71*** (0.71)	3.32*** (0.80)	3.01*** (0.77)	1.95*** (0.55)	3.08*** (0.59)	1.52*** (0.77)	3.08*** (0.79)
Middle panel * $y$	-3.50*** (0.44)	-3.98*** (0.46)	-4.33*** (0.39)	-3.68*** (0.44)	-3.72*** (0.43)	-2.67*** (0.26)	-2.48*** (0.29)	-2.12*** (0.42)	-3.46*** (0.44)
Middle panel * $y^2$	2.44*** (0.88)	3.85*** (0.89)	4.33*** (0.76)	2.97*** (0.89)	2.99*** (0.85)	1.71*** (0.43)	1.59*** (0.40)	1.02 (0.83)	2.40*** (0.88)
Observations	5,945	5,789	5,989	6,079	6,006	3,755	8,488	6,069	5,945
R <sup>2</sup>	0.82	0.80	0.86	0.83	0.83	0.83	0.81	0.81	0.82



In column (4), moving away from the default parametrization I require 3 readers for articles below 100 total readers, 7 readers between 100 and 250 and 10 readers above 250. The first quarter hour crossing these thresholds constitutes the publication time of an article. Estimated coefficients are practically the same as in the default case.

**Definition of the choice situation.** The choice situation is defined as the first choice from the frontpage per user per day. The goal of defining choice this way is to guard against dependency of previously read stories that would call for modeling sequential choice sets at the user level.

For completeness, I examine how the main results change with a weaker definition, allowing for possibly multiple choices per day. A choice enters calculations if it is the first one for a certain user for a particular frontpage. Results can be seen in column (5) of Tables 1.11 and 1.12: coefficients are virtually unchanged.

**View share threshold.** I applied a threshold of minimum 0.5% view share to consider an article on a frontpage to be in the sample. I mainly do this to exclude the influence of very marginal articles whose pageview counts might be affected more by noise. In theory logit models can be estimated on subsamples of the choice sets in an unbiased fashion. However, sensitivity to this arbitrary threshold should be examined.

In column (6), I increase the threshold to 2% while in column (7) I decrease it to 0.1%. As can be seen from the results, the qualitative patterns are unchanged.

**Time spent before choice.** Logit as a behavioral model assumes that the decision maker knows the utility of each alternative and chooses the one with the maximum level. Taking it seriously may call for some care: choices before which we cannot plausibly believe that the reader made her choice considering all articles may not be suitable for the model.

Dwelltime on the frontpage before selecting an article and the position of the selected article are correlated: the less time is spent before the choice the higher is the position of the article (linear correlation coefficient being 0.18). Distribution of dwelltimes is displayed in Table 1.13: in many cases time before choice is very short<sup>8</sup>.

<sup>8</sup>Based on previous experience with this kind of data in Athey et al. (2017) it is known that dwell-

10%	25%	50%	75%	90%
13	25	48	90	156

Table 1.13: Descriptive statistics about time spent (in seconds) on the frontpage before making a choice.

I took the 25th percentile of dwelltimes on frontpages and used only those pageviews that exceeded this limit. Column (8) shows results in line with intuition: measured position effects are milder than in the baseline case. However, the main qualitative pattern of a negative and decreasing effect of position on choice is still present.

This said, I believe taking into account all choices regardless of time spent before making a decision makes the exercise closer to what we really want to measure. After all, a very important mechanism through which position affects choice is that editors lower the search costs for some articles that they position prominently. I think modelling news choice accounting for such detailed mechanisms is a potentially fruitful avenue for future research, however, the logit approach presented here is a good first approximation.

**Controlling for time after publication.** In the default specification I used a polynomial of the time elapsing after publication to control for a decay in interest in articles. Instead of using time in a continuous manner we can use the chronological order of article appearance. In order to implement it I use dummy variables for the newest 3 articles, then for 4th-6th, 7th-10th, 11th-20th.

Regression results using these discrete controls can be found in column (9). Results are identical to those of the main specification.

## 1.9 Conclusion

This paper set out to determine the importance of how newspapers position news articles on their frontpages for how large readership an article

---

time is measured with noise: zeros and very large values are relatively frequently present in the data. However, non-extreme percentiles of the distribution can be used reliably.

attracts. Linking article position information on subsequent frontpages to browsing data I identified large effects of vertical positioning on news popularity. Furthermore, a counterfactual exercise pointed towards possible gains in click-through rates resulting from ordering the articles differently.

The identified effects highlight the importance of editorial decisions of newspapers: through position choices, they can directly influence what people are likely to read. Presumably newspapers differ in the topics they choose to promote using positioning on their frontpages. An especially interesting aspect of this is politics: the current analysis could be taken further to analyze how much newspapers influence the political readings of their audience and whether or not this contributes to rising political polarization (see, e.g., Gentzkow (2016)). Repeating the same estimation for newspapers on the opposite side of the political spectrum than The New York Times and comparing the results would certainly be a valuable line of research.

I would like to highlight two other possible steps to take further the investigation to. First, this study only covers desktop users. However, one of the main motivations of the paper is that news reading moves more and more to mobile screens where vertical position effects are expected to play an even bigger role. The empirical assessment of this hypothesis is of great importance. Second, getting even more frequent information on frontpage layout and a longer time period would enable researchers to identify more precise position effects.

## Chapter 2

# The Impact of Aggregators on Internet News Consumption

*joint with Susan Athey and Markus Mobius*

### 2.1 Introduction

A recent policy debate concerns the impact of the internet on the news media. Many authors have noted a series of stylized facts about the industry that suggest the impact of the internet has been quite negative: for example, the Newspaper Association of America reports that from 2000 to 2009, newspaper advertising revenue declined by 57% in real terms, and circulation fell by 18%. Digital has become quite important for news publishers: Pew Research reports that by 2015, a quarter of newspaper advertising revenue comes from digital, but digital revenue has not replaced lost revenue from traditional advertising. In addition to many widely publicized bankruptcies, investment in journalism has been reported to decline; for example, newsroom employment declined 40% between 1994 and 2014.<sup>1</sup> The issues cut across large and small publishers: a 2015 survey of U.S. digital publishers focusing on local news found that fewer than half were profitable.<sup>2</sup> At the same time, there has been popular discussion about how publishers and individual journalists have responded to the incentives created by the digital environment,

---

<sup>1</sup>See Barthel (June 15, 2016; accessed November 17, 2016).

<sup>2</sup>See Lu and Holcomb (June 15, 2016; accessed November 17, 2016)

for example, by optimizing the writing and headlines for search engines, aggregators, and social media.<sup>3</sup>

One particularly contentious point in this debate is the role of news aggregators. Pure aggregators such as Google News do not produce any original content but rather curate content created by other outlets through using a combination of human editorial judgement and computer algorithms. The results are presented with a few sentences and perhaps photos from the original article; to read the full article, users can click through and go to the web site of the original content creator. Thus, news aggregators act in dual roles: their front pages look very similar to news outlets who produce original content, and thus may be a substitute for them; yet they also aggregate a wide range of sources, and may be an effective mechanism for search and discovery, which places it in the role of an upstream complement to the outlets who produce news. The magnitudes of these two effects, as well as the answer to more detailed questions about how aggregators affect different types of outlets and readership of different types of news, determine whether aggregators increase or decrease the returns to investment in news reporting.

This debate has received particular attention in the European Union where two countries, Germany and Spain, enacted copyright reforms that allow newspapers to charge aggregators for linking to news snippets. The German law came into effect of August 1, 2013 and allowed newspapers to provide a free license to aggregators. Members of the main German newspaper trade association VG Media provided Google News with a free license - hence, the introduction of the law had no impact on Google News in Germany. However, no other aggregator received such a free license and since August 2014 smaller German aggregators such as GMX, Web.de and T-Online have scaled down or discontinued their services. The Spanish law came into effect in January 2015 and did not provide for a free license – Google News therefore decided to shut down its news aggregator on December 16, 2014 and other aggregators such as Yahoo and Bing News have followed suit.

In this paper, we use the shutdown of Google News in Spain as a natural experiment to evaluate how news aggregators affect news consumption.

---

<sup>3</sup>Numerous how-to guides and advice for reporters to modify articles and headlines have become over time; for an older one, see Smith (April 4, 2008; accessed November 17, 2016), while a more recent discussion can be found at Jafri (January 27, 2014; accessed November 17, 2016).

Our dataset is a sample of all browsing events for more than 100,000 users in Spain who use Microsoft products for browsing the internet and have opted in to allowing their data to be used for research purposes.<sup>4</sup> We use this dataset to construct quasi-experimental treatment and control groups. Our treatment users are all Google News users. We match these users with a synthetic control group of non-Google News users who have the same news consumption patterns as the corresponding Google News users after the shutdown of Google News. Our matching procedure therefore selects control users who make the same consumption decisions in the absence of Google News (when both groups have access to the same news discovery technology) and therefore have the same underlying preferences.

We then first compare the *overall* news consumption of treatment and control users pre-shutdown across all news categories. This allows us to evaluate the impact on total traffic to newspapers if Google News is shut down – this quantity is of fundamental interest to publishers and policy makers who want to know whether aggregators steal traffic from news creators (substitutes view) or increase traffic (complements view). We find that the removal of Google News reduces overall news consumption (including consumption of the Google News homepage) by about 20% for treatment users, while visits to news publishers decline by about 10%. This decrease is concentrated around small publishers, while large publishers do not see significant changes in their overall traffic (but see an increase in their own home page views, offset by a decrease in views of articles).

These results highlight the potential impact of intermediaries on industry structure: they make it easier for consumers to search and consume products from small firms, increasing competition across publishers for consumer attention. We have seen similar effects in other technology-enabled intermediaries, such as eBay, Uber, AirBnb, and travel and price comparison sites, where the technology platform reduces search costs and enables smaller firms that may lack name recognition or reputation to be discovered by consumers. Whether this is good or bad for consumer welfare depends on details of how investment is spread across

---

<sup>4</sup>The data is subject to stringent privacy restrictions and at all times resides only on secure servers, and only aggregate statistics and the output of statistical models can be reported. However, we are able to construct the variables for analysis using the fully disaggregated data.

these firms, as well as how investment impacts market share. In the case of news, the welfare effects depend on whether the investments that increase visibility on aggregators are welfare-enhancing (unique investigative journalism) or wasteful (misleading headlines), as well as whether smaller firms add to the diversity of alternative perspectives rather than reproduce news where the investments have been made by others.

We further find that when Google News shuts down, its users are able to replace some but not all of the types of news they previously read. Post-shutdown, they read less breaking news, hard news, and news that is not well covered on their favorite news publishers. These news categories explain most of the overall reduction in news consumption. The result about breaking news highlights an advantage for aggregators: they can always be “first to market” with the latest news, since they can link to articles as soon as they appear on publisher home pages. They can also offer more breadth than any individual publisher, allowing readers to access topics not covered by their favorite outlets, as well as allowing readers to read more in-depth coverage on particular topics. Finally, the Google News homepage focuses more on hard news than the typical publisher home page.

Despite the intrinsic policy importance of the news industry and the close attention this issue has received from regulators, there is very little existing empirical evidence on the impact of aggregators. The paper closest to this one is Calzada and Gil (2016), which independently studies the same event using a different data source. Their paper finds an almost identical magnitude (an 11% reduction in visits to news outlets due to the Google News shutdown). Their paper uses aggregate data about site visits, while our paper relies on individual-level browsing data. As such, we are able to explore how individual consumers substitute the missing Google news consumption, and how the content of their consumption changes. The difference in data sources also affects the identification strategy: we are able to use Spanish users who were not previously Google News users as a control group, while Calzada and Gil (2016) relies on France and Germany to control for time trends in news viewing. A limitation of our study is that it only includes users of Microsoft products, who account for less than half of PC news browsing; thus, Calzada and Gil (2016) provides confirmation that their behavior

is broadly representative.

Another closely related paper is Chiou and Tucker (2015). They study a “natural experiment” where Google News had a dispute with the Associated Press, and as a result, did not show Associated Press content for about seven weeks. The paper has aggregate data about page views to Google News as well as the sites visited immediately after Google news. They use views to Yahoo! News as a control. The paper finds that Google News is a complement to news outlets: taking the Associated Press content away from Google News leads to fewer visits to news outlets (where Associated Press articles are featured). Our paper is complementary to theirs: our main finding (complementarity) is consistent with theirs – however, our individual level data allows us to (a) dis-aggregate the effects by outlet size and (b) analyze the types of news consumption that see the biggest drops, which allows us a more nuanced analysis. Though less directly related to our work, a literature on the network structure of information flows on the web finds that “hubs” may improve information flows.<sup>5</sup>

The balance of the paper is organized as follows. In Section 2.2 we introduce a simple empirical model of news consumption and describe a matching algorithm that allows us to construct synthetic control and treatment groups. We then use this framework in Section 2.3 to document the drop in overall news consumption after the shutdown of Google News in Spain on December 16, 2014. We decompose this overall volume drop in Section 2.4 and show that it is predominantly driven by a reduction in the consumption of scarce and breaking news.

---

<sup>5</sup>The prevailing vein of this work was pioneered by Kleinberg (1999), who developed the hub-authority information flow model and the Hyperlink-Induced Topic Search (HITS) algorithm (see also Kleinberg and Lawrence, 2001). Weber and Monge (2011) introduce a third category of nodes, sources, to Kleinberg’s model and use their expanded model to study news information flow. Using random graph models to study hyperlink structure, they find that hubs very rarely exist in pure form. Rather, there is often a degree of reciprocity such that hubs function to some degree as distribution nodes for information through the network.



## 2.2 Empirical Model and Data Description

### 2.2.1 Theory

In this section we present a stylized model of news consumption that helps motivate our empirical strategy for estimating the effects of aggregators.

#### *Preferences for News*

A consumer  $i$  has one unit of time that she can allocate between reading news stories and other activities. Every news story has a vector  $c$  of characteristics, referred to as its “type,” with components indexed by  $d = 1, \dots, D$ . For example, a particular characteristic might indicate whether a news story is breaking news or not, or whether the story is about a popular topic. The characteristics space is a finite product set  $\mathcal{C} = \prod_{d=1, \dots, D} \mathcal{C}_d$ . We denote user  $i$ ’s consumption of type  $c$  news at time  $t$  by  $N_{i,c}^t$ , and the vector of overall consumption is denoted  $N_i^t$ .

The total consumption of other leisure activities is denoted with  $L_i^t$ . The user’s utility from consuming the bundle  $(N_i^t, L_i^t)$  at time  $t$  is described by a nested Cobb-Douglas utility function:

$$U_{it} = \left[ \prod_{c \in \mathcal{C}} (N_{i,c}^t)^{\alpha_{i,c}} \right]^{\tau^t \tau_i} L_i^{1 - \tau^t \tau_i} \quad (2.1)$$

The key implications of this functional form are described as follows. First, a user’s share of time spent reading news can be decomposed into a date effect and a person effect:  $\tau^t$  captures weekday and seasonal effects in preferences for reading news, while  $\tau_i$  captures the importance that individual  $i$  attaches to news reading. Second, a utility maximizing individual will consume a constant share of news of a particular type, so long as the costs of finding different types of news do not change. We will formally check the assumption that preferences are stable over time below.

#### *Discovery of News*

Consumers do not directly know what articles are available to read on a given day. In order to discover articles, they must visit home pages of news outlets, view social media, use search, or use aggregators.

Home pages or “landing pages” of news outlets merit special discussion. These pages have a list of article headlines and snippets from the stories, as well as links to the stories. Home pages play a dual role for consumers: they are a mechanism to learn about articles that are available, and users also consume news directly by reading the headline and the beginning of a news story. For simplicity, we will assume that the utility derived from a single landing page is equivalent to reading a fraction of each member of a set of articles (where in principle the fraction could be greater than one).

First consider a case without aggregators. For each news type  $c$ , the user can utilize direct navigation to outlet home pages, search, and social, and translate 1 unit of time into consumption of a set of articles, denoted  $\pi_c^D$ ,  $\pi_c^{Se}$ , and  $\pi_c^{So}$ , respectively. As discussed above, when the user visits outlet home pages, the visit results in consumption of news  $\pi_c^H$  (in units equivalent to articles). Thus,  $\pi_c^H$  describes the quantity of news of each type consumed directly on the home pages, and  $\pi_c^D$  describes the quantity of articles consumed by clicking on links on the home pages. Let  $\pi_c = \pi_c^D + \pi_c^{Se} + \pi_c^{So} + \pi_c^H$  denote the sum of these. So far, we have not placed any restrictions on the discovery process; we have simply introduced notation describing its output in terms of the objects that create utility for the consumer. For example, if the consumer selects investments of time in visiting home pages, search engines, and social media to maximize expected utility, our notation can be interpreted as summarizing the number of articles consumed when the consumer follows an optimal time investment policy.

Our next step is to introduce a key simplifying assumption, one that leads to a tractable empirical model as well: we assume that the discovery process has constant returns to scale, so that allocating more or less time results in a proportional increase or decrease in the news discovered of each type from each source. In a more detailed microeconomic model of search, the constant returns assumption would imply restrictions on the process by which users engage in search and discovery, the availability of content and home pages, as well as the technology available to search.

Now consider the case where aggregators are available. We will assume there is a single aggregator, Google News, and that the availability of

Google News changes the amount of news a consumer can consume per unit of time, both by changing the number of articles of each type that can be discovered, and by introducing a new home page that includes partial news stories from a large variety of publishers. We denote the consumption of news through direct navigation, search, social, and outlet home pages when Google News is also available as  $(\tilde{\pi}_c^D, \tilde{\pi}_c^{Se}, \tilde{\pi}_c^{So}, \tilde{\pi}_c^H)$  with sum  $\tilde{\pi}_c$ . We denote the news articles accessed through Google News by  $\tilde{\pi}_c^{GN}$ , and the consumption of news through reading the Google News home page as  $\tilde{\pi}_c^{HGN}$ . We assume that the aggregator-augmented technology is at least as productive as the general-purpose technology at finding articles:  $\tilde{\pi}_c + \tilde{\pi}_c^{GN} + \tilde{\pi}_c^{HGN} > \pi_c$  for all  $c$ .

The next definition formally captures the notion of aggregators complementing or substituting for traditional news-reading.

**Definition 2.2.1.** Aggregators are substitutes (complements) to traditional news-reading technology for news of type  $c$  if  $\tilde{\pi}_c + \tilde{\pi}_c^{GN} < (\geq) \pi_c$ .

Note, that even though aggregator-augmented technology always weakly increases total news consumption it might decrease the number of news stories that are directly consumed on the publishers' websites if aggregators are substitutes and consumers can effectively consume news directly on aggregators' home pages ( $\tilde{\pi}_c^{HGN}$ ).

We can refine the definition further by saying that Google News is a substitute (complement) for reading articles if  $\tilde{\pi}_c^D + \tilde{\pi}_c^{Se} + \tilde{\pi}_c^{So} + \tilde{\pi}_c^{GN} < (\geq) \pi_c^D + \pi_c^{Se} + \pi_c^{So}$ . Analogously, Google News is a substitute for publisher home pages if the consumption of those publisher home pages is lower in a regime with Google News. Note, we have not directly defined notation for the number of page views of publisher home pages, because  $\pi^H$  expresses consumption of home pages in units of articles.

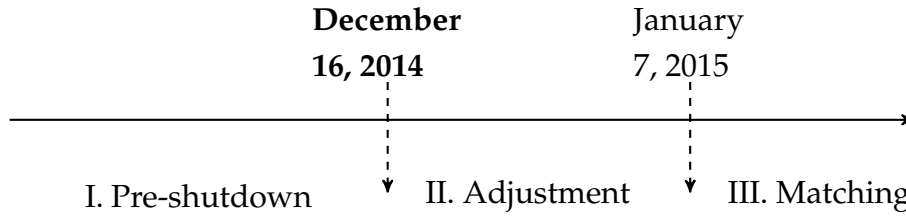
We can now derive consumer demand for news stories of type  $c$  with both technologies. Utility maximization implies that when aggregators are not available, a user will consume

$$N_{i,c}^t = \tau^t \tau^i \alpha_{i,c} \pi_c \quad (2.2)$$

while a user of the aggregator-augmented technology will consume:

$$N_{i,c}^t = \tau^t \tau^i \alpha_{i,c} (\tilde{\pi}_c + \tilde{\pi}_c^{GN} + \tilde{\pi}_c^{HGN}) \quad (2.3)$$

Figure 2.1: Timeline for matching Spanish control and treatment users



Since we will measure news consumption by the number of pages visited and the dwell time spent on publisher's websites we denote the news consumed directly on publishers' websites when using aggregator-augmented technology with  $\tilde{N}_{i,c}^t$ :

$$\tilde{N}_{i,c}^t = \tau^t \tau^i \alpha_{i,c} (\tilde{\pi}_c + \tilde{\pi}_c^{GN}) \quad (2.4)$$

For a particular type of news on a given day, the ratio of news consumed on publisher websites using the two technologies is directly proportional to the relative productivity of the two different discovery processes:

$$\frac{\tilde{N}_{i,c}^t}{N_{i,c}^t} = \frac{\tilde{\pi}_c + \tilde{\pi}_c^{GN}}{\pi_c} \quad (2.5)$$

### 2.2.2 Empirical Implementation and User Matching

We next take this model to our data. We distinguish between three main periods shown in Figure 2.1 – the pre-shut period I (before December 16, 2014), a 3-week adjustment period (until January 7, 2016) and a matching period (after January 7, 2016). We define a set of “treatment users” as those who have used Google News before the shutdown. For each such treatment user  $i$  we find a control user  $\hat{i}$  who *does not* use Google News in the pre-shutdown period but consumes news in the same way as the treatment user during the matching period. We will then assume that treatment and control users have the same preferences (e.g.  $\tau^i = \tau^{\hat{i}}$  and  $\alpha_{i,c} = \alpha_{\hat{i},c}$  for all news types  $c$ ) because they make the same consumption decisions when having access to the same technology.

We will now describe the details of the matching algorithm.

*Active users.* Section A.2 of the supplementary appendix describes in detail how we construct our user-level data set from the browser logs. Our dataset consists of a sample of desktop users and we focus on users who are active 90% of the weeks between October 1, 2014 and March 17, 2015 and who read news for at least 10 weeks. The universe of *active users* in Spain comprises about 158,000 people.

Browsing events are linked through referrer URLs: if a page load has a referrer it indicates that the user clicked on a link on that referring page to visit the present page. The browsing stream can be partitioned into a set of trees where the root of each tree either has an empty referrer or refers to a non-publisher page (such as a Google search page, a social media page or a news aggregator). We refer to these trees as “news mini sessions” (or NMS). The root of each NMS defines the referral mode: we distinguish direct navigation (an empty referrer), search (Google, Bing or Yahoo search referral), social (Facebook or Twitter referral) and other referrals.

We identify a “treatment user” as any user who has used Google News at least once during the pre-shutdown period. The remaining users are potential control users.

*Topics.* Section A.3 of the supplementary appendix describes in detail how we construct topics. We classify all URLs into landing pages and article pages based on frequency distribution of page loads across the observation period (articles tend to have most page loads concentrated within a few days after publication while landing pages are visited at a fairly constant rate throughout).

We then scrape all 110 million article pages and focus on the 61% of articles that contain more than 100 words (the remainder are often slide shows or video pages). We compare the content of each such article to all the Wikipedia articles published on the Spanish-language Wikipedia. We exploit the fact that most major news events receive a Wikipedia entry within days or weeks. Wikipedia provides us with a convenient and stable set of topics. We construct a set of nested topics consisting of a “super-topic” (such as *Health*) and a sub-topic (such as *Spain Ebola crisis* which hit Spain in 2014/2015). We manage to classify 53% of all articles into about 300 topics. The top 50 topics cover 75% of all page loads.

We hired 7 student evaluators to rate the quality of our super-topic and sub-topic assignment for 500 articles and found that 95% of super-topic and 85% of sub-topic assignments were deemed correct.

Table 2.1: Differences between matched treatment and control users

statistic	is treated?	total	news	article	search	breaking	"hard news"
min	0	173	101	0	0	0	0
min	1	202	64	0	0	0	0
median	0	3,635	293	122	61	33	13
median	1	3,609	292	124	60	32	13
mean	0	5,250.921	396.243	170.920	99.752	50.982	25.011
mean	1	5,339.146	397.523	174.262	103.655	49.960	24.403
sd	0	5,301.106	294.181	144.934	124.428	54.563	36.368
sd	1	6,139.647	296.541	148.091	136.504	52.349	32.624
max	0	61,397	1,368	1,143	1,178	793	462
max	1	140,589	1,372	1,140	1,149	766	295
N	0	2,317	2,317	2,317	2,317	2,317	2,317
N	1	2,317	2,317	2,317	2,317	2,317	2,317

*Matching algorithm.* For each treatment user we rank all potential control users according to a proximity score that is based on the components listed in Table 2.1: (1) total overall page views, news page views (landing pages and articles) and news article page views, (2) news page views accessed through search and (3) breaking news page views and page views on hard news (excluding celebrity news and sports among top 50 topics).

Formally, we calculate for each variable the empirical distribution among all active users. Then for each user we calculate its percentile rank in the distribution. For each treatment user we calculate the difference from each potential control user in these percentile scores and weigh these scores equally to calculate a single proximity score. This procedure ranks potential control users for each treatment user. We then use the random serial dictatorship algorithm to assign a unique control user to each treatment user. The results of the matching are displayed in Table 2.1<sup>6</sup>. Overall, the treatment and control group differ by less than 5% along each of the matched dimensions (where means are compared

<sup>6</sup>After applying the matching procedure we focus on treatment-control pairs with at least 100 news pageviews in the matching period. However, all our qualitative results hold up even when using the full sample.

using data only from the matching period).

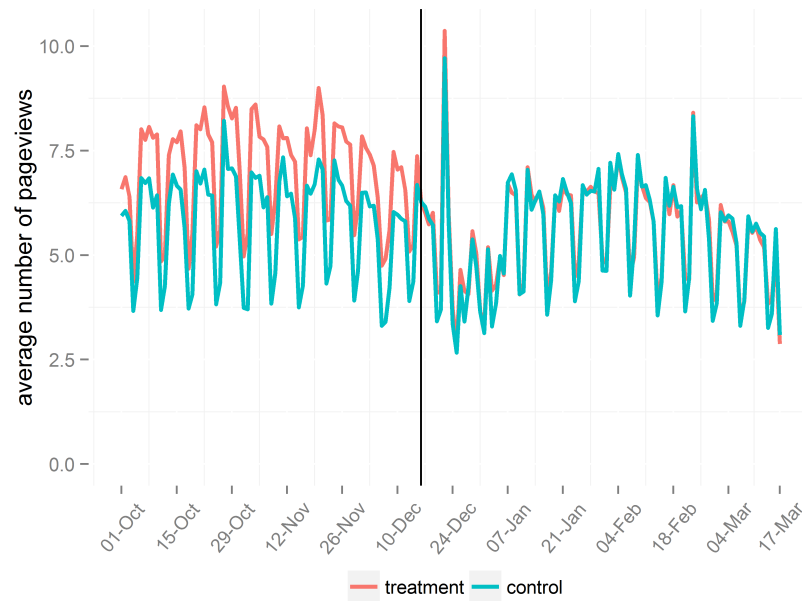
*Empirical model.* Our matched samples of treatment and control users allows us to estimate the effect of the Google News shutdown by comparing the news consumption of treatment and control users during the pre-shutdown period. Since we have constructed the control group such that (by revealed preference, under the assumptions of our stylized model) both groups of users have identical preferences, we can interpret the consumption behavior of matched control user as the predicted (counterfactual) behavior for the corresponding treatment user if Google News was not available.

Our approach of matching on behavior *after* the “treatment” (taking away Google News) has been applied is somewhat non-standard at first glance. A superficially more natural approach might be to match users based on their behavior in the pre-shutdown period. However, on closer examination, it would be hard to justify an empirical strategy based on such an approach. Since in the pre-shutdown period, Google News users are accessing news through a different search and discovery technology, there is no reason to believe that a Google News user and a non-Google News user who have the same preferences would read the same news—in general we would not expect that. On the other hand, two users with the same preferences should read the same amounts of different types of news in the post-shutdown period.

Our analysis would be more straightforward if Google News was introduced as an option rather than taken away. Our approach needs to rely on a few additional assumptions, namely that Google News does not have persistent effects on reader behavior even after it disappears. As described above, to partially deal with the possibility of persistence, we leave out an “adjustment period” after Google News shuts down. However, we do not see evidence in the data that the treatment and control users have differences in behavior that change over time, reducing the potential for concern.

As a robustness check, we also considered an empirical approach based on matching on pre-shutdown news consumption and measuring differences in the post-shutdown period; this approach yielded qualitatively similar results, as we discuss further below.

Figure 2.2: News consumption of the treatment and control groups over time



## 2.3 Google News and Overall News Consumption

We now analyze the impact of Google News on *overall* news consumption across all news categories. The magnitude of this combined effect is fundamental for publishers and policy makers who have to decide whether aggregators steal traffic from news creators (substitutes view) or increase traffic (complements view). In order to gain some intuition, we start with a graphical analysis in Figure 2.2 which shows the average daily news reading (in term of number of pageviews) over time for both control and treatment users.

As a result of our matching procedure the two groups have very similar news consumption levels after the shutdown (marked with the vertical line). In fact, the consumption levels of both groups are nearly indistinguishable even during the adjustment period. Moreover, control users have an overall stable consumption level before and after the shutdown (other than the holiday period and a spike corresponding to the indictment of Spain's Princess Cristina on corruption charges) which is to be expected since the Google News shutdown did not affect them. The stability also implies that the details of how we model time trends will not have a big impact on our results.

In contrast, Google News users have much higher news consumption

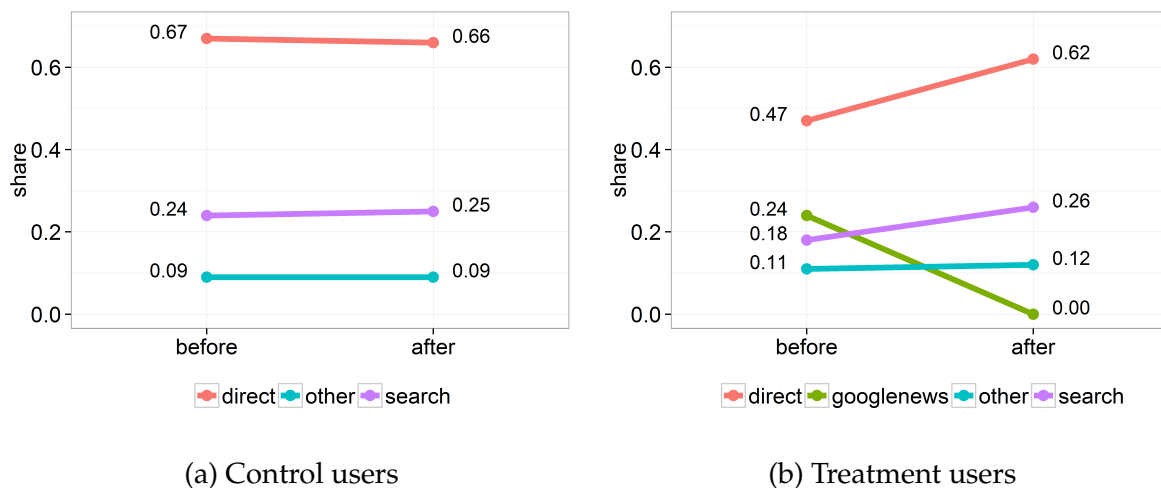


compared to control users before the shutdown. Consequently, there is a sharp and pronounced drop in news volume once Google News becomes unavailable. We call this phenomenon the “volume effect.”

In order to gain some insight how Google News affects news discovery we show the referral modes for total news reading for both groups in Figure 2.3. The referral modes are classified as direct navigation, search, Google News and other (which includes social media, emails, forums, etc).

For control users, referral shares are constant in time, as expected. During the post-period, treatment users have very similar referral shares compared to control users. This is again expected since we are matching both on total news consumption and consumption via search. Before the shutdown, reading through Google News makes up 24% of users’ consumption. Direct navigation and search take up this share after the shutdown, while the referral share through other modes stays roughly constant.

Figure 2.3: Referral shares, total news consumption.



### 2.3.1 Estimating the Volume Effect

We now estimate the effect of removing Google News by comparing the consumption of treatment and control users pre-shutdown. Using our model, we can express the total news consumption of all treatment users

during the pre-shutdown period by  $\tilde{N}^I$ :

$$\tilde{N}^I = \tau^I \sum_{\hat{i}, c \in \mathcal{C}} \tau^{\hat{i}} \alpha_{\hat{i}, c} (\tilde{\pi}_c + \tilde{\pi}_c^{GN}) \quad (2.6)$$

The corresponding consumption of all control user is  $N^I$ :

$$N^I = \tau^I \sum_{i, c \in \mathcal{C}} \tau^i \alpha_{i, c} \pi_c \quad (2.7)$$

Therefore, the ratio of total news consumption of treatment and control users equals:

$$\frac{\tilde{N}^I}{N^I} = \frac{\sum_{\hat{i}, c \in \mathcal{C}} \tau^{\hat{i}} \alpha_{\hat{i}, c} (\tilde{\pi}_c + \tilde{\pi}_c^{GN})}{\sum_{i, c \in \mathcal{C}} \tau^i \alpha_{i, c} \pi_c} \quad (2.8)$$

Note, that we are exploiting our matched sample here which allows us to replace  $\tau^i$  with  $\tau^{\hat{i}}$  and  $\alpha_{i, c}$  with  $\alpha_{\hat{i}, c}$ . This ratio measures the overall impact of Google News on treatment users, while factoring out seasonal and day effects.

Table 2.2: Volume difference measures for news types according to referral shares and pageview types (number of pageviews).

Referral mode	Total	Total (excl. GN)	Article	Landing page	Landing page (excl. GN)	Other content
Total	0.197*** (0.019)	0.096*** (0.018)	0.288*** (0.021)	0.158*** (0.024)	-0.085*** (0.023)	-0.076 (0.058)
Direct	-0.172*** (0.025)	-0.172*** (0.025)	-0.223*** (0.030)	-0.117*** (0.026)	-0.117*** (0.026)	-0.268*** (0.077)
Search	-0.056* (0.032)	-0.073** (0.032)	-0.023 (0.032)	-0.102*** (0.042)	-0.150*** (0.041)	-0.050 (0.097)
Other	0.343*** (0.058)	0.340*** (0.058)	0.420*** (0.053)	0.166 (0.105)	0.148 (0.105)	0.158* (0.095)

Notes: Bootstrapped standard errors in parenthesis.

Taking the logarithm of both sides of (2.8) suggests an empirical approach to estimating the percentage change in news consumption when Google News is removed. Table 2.2 shows the results of this calculation, where we average the logarithm of the left-hand side of (2.8) using data from the pre-period, where we use the number of pageviews as our measure of news consumption. The top row shows the volume effect across

Table 2.3: Volume difference measures for news types according to referral shares and pageview types (dwelltime).

Referral mode	Total	Total (excl. GN)	Article	Landing page	Landing page (excl. GN)	Other content
Total	0.307*** (0.022)	0.170*** (0.019)	0.371*** (0.024)	0.258*** (0.032)	-0.044* (0.026)	0.105 (0.071)
Direct	-0.136*** (0.027)	-0.136*** (0.027)	-0.231*** (0.033)	-0.078*** (0.030)	-0.078*** (0.030)	-0.142* (0.081)
Search	-0.004 (0.036)	-0.028 (0.035)	0.029 (0.040)	-0.053 (0.047)	-0.118*** (0.046)	-0.018 (0.124)
Other	0.429*** (0.060)	0.426*** (0.060)	0.460*** (0.060)	0.300*** (0.117)	0.281*** (0.117)	0.424*** (0.145)

Notes: Bootstrapped standard errors in parenthesis.

all navigation modes for total news as well as the sub-categories of only articles, landing pages and other content (slide shows and videos).<sup>7</sup> Column (5) explicitly excludes Google News landing pages and therefore only captures publishers' landing pages. Rows 2 to 4 capture the change just for direct navigation, search and other modes of accessing news. We calculate standard errors using the bootstrap, sampling treated-control pairs as a unit.

Treatment users have 19.7 percent higher consumption in the pre-shutdown period compared to control users, including their consumption of the Google News home page. As a result of our matching procedure and the exogenous nature of the shutdown we interpret this as the causal effect of the presence of Google News on news consumption volume. This volume change comes from two sources: (1) Google News users consume 28.8 percent more articles but 8.5 percent fewer landing pages (omitting the Google News landing page). Hence, Google News is a complement to overall news reading and articles but a substitute to landing pages. The landing page result is intuitive since Google News directly links to publishers' articles and bypasses the landing pages. However, this decline is more than compensated for by increased traffic to articles. (It should be noted, however, that a large share of advertising revenue comes from the publishers' landing pages.)

<sup>7</sup>Slideshowes and videos are not classified as articles if the word count is less than 100 words.

Google News has only a small effect on news accessed through search. It complements news accessed through “other” navigation modes (such as social media). However, as we saw before, these make up only a small share of total news referrals.

A question that might arise is whether Google News referrals might be of lower quality than others; perhaps when using Google News, readers will click on many articles, glance at them, and then quickly return to the Google News home page. To assess this hypothesis, Table 2.3 shows the same calculations using dwelltime (time spent on the page) as a measure of news consumption.<sup>8</sup> The numbers corroborate our previous conclusions using pageviews.

### 2.3.2 Volume Effect by Outlet Size

Table 2.4: Volume effect calculations (number of pageviews).

Outlet type	Referral mode	Total	Article	Landing page	Other content
Top 20	Total	0.022 (0.023)	0.201*** (0.025)	-0.137*** (0.026)	-0.109* (0.064)
Top 20	Direct	-0.218*** (0.031)	-0.261*** (0.037)	-0.169*** (0.030)	-0.282*** (0.085)
Top 20	Search	-0.120*** (0.038)	-0.076** (0.036)	-0.193*** (0.048)	-0.045 (0.092)
Below top 20	Total	0.263*** (0.037)	0.446*** (0.035)	0.046 (0.052)	0.106 (0.100)
Below top 20	Direct	-0.042 (0.051)	-0.113** (0.054)	0.012 (0.055)	-0.111 (0.124)
Below top 20	Search	0.028 (0.059)	0.061 (0.056)	-0.021 (0.082)	-0.044 (0.132)

Notes: Bootstrapped standard errors in parenthesis.

We now break out the volume effect by outlet size. This is an important exercise because the Spanish copyright reform was primarily advocated by bigger newspapers, and also because theory suggests that the effect might be different, with the search and discovery function playing

<sup>8</sup>Section A.2.2 of the Data Appendix explains how the dwelltime variable is logged in our data.

a more important role for smaller outlets. For smaller outlets, the users may not expect the benefit of accessing the landing pages to outweigh the time to access it, and further, the users may not be aware of some smaller publishers. Using the whole sample period and all active users, we order news outlets based on their size measured in terms of total pageviews. Then we compare the top 20 publishers to the smaller ones. Results are shown in Table 2.4.

A striking pattern can be observed when we compare the total news consumption of treatment and control users: while the effect of Google News on the top 20 publishers is not statistically significantly different from zero (with estimates small in magnitude), smaller outlets gain as much as 26.3 percent from the presence of Google News. When we dissect this effect by page view type, we see that the availability of Google News changes the page view mix of towards articles and away from landing pages for larger outlets, but the two effects cancel in the aggregate. For smaller outlets, the landing page traffic is unaffected by Google News but article pageviews increase by 44.6 percent. Consistent with the promotion of smaller outlets, Google News also makes it easier for users to facilitate multiple news sources: Table 2.5 compares the user-level Herfindahl indices for outlets for the top 20 topics by referral mode. For each topic, the Google News concentration index is lower than for direct navigation.

Therefore, the evidence supports the hypothesis that Google News is neither a complement or substitute for bigger publishers, but is a substitute for the high-revenue landing pages of those outlets. We also see that Google News is a substitute for articles accessed through direct navigation—highlighting that large outlets lose some of their curation role. If part of the long-term incentive for news outlets to maintain their brand comes from the way they curate news, through the selection of articles to highlight prominently on their landing pages, then the fact that Google News is in effect selecting what articles from each outlet to highlight on the Google News home page may decrease the incentive of publishers to invest in the quality of their curation and thus their brand. This is an example of a broader concern publishers articulate surrounding aggregators and social media: they worry that they are being

Table 2.5: Mean of user-level Herfindahl indices for outlets within a topic (pre-period, treatment users).

Topic	direct	googlenews	search
Business/Finance: Macroeconomics and Economic Policy	0.793	0.603	0.789
Politics: Independence of Catalonia	0.797	0.531	0.793
Spain: Government	0.770	0.562	0.800
Health: Spain Ebola Crisis	0.825	0.558	0.785
Sports: Atltico de Madrid	0.834	0.721	0.863
Entertainment: TV Show Gran Hermano VIP	0.900	0.741	0.806
Spain: Corruption Operation Pnica	0.839	0.676	0.856
Celebrities: Duchess of Alba dies	0.874	0.708	0.808
Sports: FC Barcelona	0.865	0.720	0.891
Sports: Real Madrid	0.884	0.705	0.873
Sports: Soccer Players	0.886	0.726	0.876
International News: Central and South America	0.855	0.712	0.898
Sports: Formula 1	0.882	0.734	0.822
Entertainment: Movies and Actors	0.900	0.811	0.889
Entertainment: Television	0.893	0.784	0.881

“disintermediated” and “commoditized,”<sup>9</sup> consistent with a decreased ability to differentiate their products in the eyes of consumers, as their content is accessed instead through an intermediary.

On the other hand, our evidence supports the hypothesis that Google News is a strong complement for small outlets. This implies that bigger outlets gained compared to smaller ones due to the shutdown of Google News and users visited a less diverse set of outlets. The social welfare implications of this finding are unclear; there are potentially competing effects. If small outlets produce unique content with alternative viewpoints or reporting, then Google News supports “media diversity” and helps smaller publishers get viewership and thus returns on their investments in journalism. On the other hand, if smaller outlets mostly copy news from larger outlets, without investing in reporting, then Google News might decrease the returns to investment for the primary sources of investigative journalism. This paper does not provide evidence that speaks directly to these welfare tradeoffs. In principle, the investments

<sup>9</sup>See, e.g., Sterling (April 7, 2009; accessed November 17, 2016), Filloux (July 27, 2016; accessed November 17, 2016), Weiner and Group (May 11, 2015; accessed November 17, 2016)

in journalism can be measured using data from newspaper employment and expenditures, and diversity of viewpoints as well as the extent of original reporting can be assessed from textual analysis. This represents an interesting avenue for future work.

Table 2.6 repeats the same exercise using dwelltime as the consumption measure. Overall, the evidence supports the findings from our analysis of pageviews. The main difference is that total dwelltime increases slightly even for bigger outlets under the presence of Google News: this reflects the fact that users tend to spend more time on article pages compared to landing pages. Dwell time on landing pages declines for large outlets, in line with the results for page views. In the current advertising environment, advertising is typically sold on the basis of page views, not dwell time, so the newspapers may not see revenue increases that correspond to the dwell time increase.

Table 2.6: Volume drop calculations (dwelltime).

Outlet type	Referral mode	Total	Article	Landing page	Other content
Top 20	Total	0.103*** (0.024)	0.308*** (0.029)	-0.106*** (0.031)	0.100 (0.081)
Top 20	Direct	-0.189*** (0.032)	-0.279*** (0.040)	-0.137*** (0.036)	-0.157* (0.089)
Top 20	Search	-0.086** (0.042)	-0.019 (0.046)	-0.195*** (0.054)	0.004 (0.135)
Below top 20	Total	0.324*** (0.038)	0.489*** (0.040)	0.123*** (0.053)	0.132 (0.124)
Below top 20	Direct	0.003 (0.053)	-0.111* (0.065)	0.075 (0.056)	-0.053 (0.178)
Below top 20	Search	0.108 (0.069)	0.113* (0.067)	0.121 (0.104)	-0.214 (0.181)

Notes: Bootstrapped standard errors in parenthesis.

## 2.4 Decomposing the Volume Effect

In Section 2.3 we analyze the overall drop in news consumption when Google News becomes unavailable. In order to better understand *why*

consumption decreases we examine how reading changes as a function of news characteristics. Based on anecdotal evidence, we focus on five characteristics: recency of news/breaking news, the extent to which different news stories are widely covered by newspapers (global supply scarcity), the extent to which news stories are well covered by readers' favorite news outlets (relative supply scarcity), popularity of news stories, and whether a story is about "hard" or "soft" news. As motivated in the introduction, understanding what kinds of news are particularly affected by Google News helps explain how Google News competes with or complements existing outlets, and also sheds light on how aggregators affect the incentives of news publishers to produce different types of news.

Formally, we define each news category by whether the category does or does not have each of the  $D$  "characteristics" (in our empirical work, the  $D = 5$  characteristics just described), so that  $c$  is a vector in  $\{0, 1\}^D$ . For each of the  $d = 1, \dots, D$  conditions, there is a corresponding component of  $c$ , denoted  $c_d \in \{0, 1\}$ , that indicates whether news type  $c$  satisfies condition  $d$ . For example,  $d$  might represent the condition "the news is breaking news" or "the news is hard news," and  $c_d = 1$  if category  $c$  satisfies the condition.

### 2.4.1 News Characteristics

We now define the characteristics we use in our empirical analysis.

**Breaking News.** Google News is an algorithmic service that is capable of collecting the most recent stories from multiple news sources. In fact, as of 2016 there is a "See realtime coverage" button on Google News that provides users with the latest relevant news for a particular event. One might therefore suspect that aggregator-augmented browsing lowers the cost for consuming breaking news compared to traditional news browsing. Most traditional readers rely only on a handful of newspapers that (1) may happen to miss the latest events and (2) do not provide links to other news sources covering the same events. As a result, we hypothesize that the shutdown of Google News leads to a disproportionate drop in the consumption of *breaking news* stories. We proxy breaking news by determining the hour of publication of each news ar-



title and measuring the time elapsing between the publication and the consumption of the news story.

**Supply Scarcity.** Some news topics are covered by a large number newspapers while others receive less attention. This mismatch can be caused by newspapers miscalculating readers' interest or by a lack of reporters who can write such stories. For example, a terrorist attack in a neighboring country might be of great interest to local readers, but newspapers might lack correspondents in that country. We distinguish between two types of scarcity: *relative supply scarcity* captures the idea that users read a limited number of publications and these publications might poorly cover certain news topics compared to other newspapers. In contrast, *global supply scarcity* captures poor coverage of a certain topic across all newspapers.

Google News can potentially mitigate both types of scarcity. First of all, personalization allows aggregators to adapt to user preferences. For example, if a user's favorite newspapers poorly cover financial news then Google News can adapt accordingly and prioritize financial news. Moreover, Google News provides typically multiple references for each news topic which makes it easier for users to find related coverage for globally scarce news topics.

We next formally define both types of scarcity.

*Relative Supply Scarcity.* Let us index users by  $i$ , topics by  $d$ , newspapers by  $n$  and days by  $t$ . The daily share of articles in newspaper  $n$  on topic  $d$  on day  $t$  is denoted with  $s_{ndt}$ .<sup>10</sup> The total pageviews consumed by user  $i$  for newspaper  $n$  during the matching period is denoted with  $y_{in}$  – it measures the intensity with which user  $i$  reads newspaper  $n$ . We define  $y_i = \sum_n y_{in}$  as the total pageviews of user  $i$  across all newspapers during the matching period and  $y_n = \sum_i y_{in}$  as the total pageviews for newspaper  $n$  across all users during the matching period.<sup>11</sup>

We then define *aggregate supply* for topic  $d$  on day  $t$ :

$$x_{dt} = \frac{\sum_n y_n s_{ndt}}{\sum_n y_n} \quad (2.9)$$

Intuitively, aggregate supply measures the average coverage for news topic  $d$  across all newspapers weighted by their popularity.

<sup>10</sup>Note, that  $\sum_d s_{ndt} = 1$ .

<sup>11</sup>Note: we focus on direct page-views, all pages (both landing and article pages).

We then define the *user-level supply* of news faced by user  $i$  for topic  $d$  on day  $t$ :

$$x_{idt} = \frac{\sum_n y_{in} s_{ndt}}{\sum_n y_{in}} \quad (2.10)$$

Intuitively, take a random pageview for the user - with probability  $y_{in}/y_i$  the user selects newspaper  $n$  and then reads a random article which is going to be on topic  $d$  with probability  $s_{ndt}$ . This is the reader's individual-level supply of articles on topic  $d$  on day  $t$ .

This allows to then define *relative supply*  $r_{idt}$  as follows:

$$r_{idt} = x_{idt} - x_{dt} \quad (2.11)$$

Relative supply has two important aggregation properties. First of all, for any specific user, relative supply across topics sum up to 0:

$$\sum_d r_{idt} = 0 \quad (2.12)$$

Second, for any specific topic, relative supply across users sum up to 0:<sup>12</sup>

$$\sum_i w_i r_{idt} = 0 \quad \text{where } w_i = \frac{y_i}{\sum_i y_i} \quad (2.14)$$

Therefore, topics are neither scarce nor abundant *on average* but they are only so *within* users.

*Global Supply Scarcity.* We identify globally scarce topics by comparing aggregate demand to aggregate supply  $x_{dt}$ . Formally, we define aggregate demand  $q_{dt}$  as the pageview share across all topics:

$$q_{dt} = \frac{y_{dt}}{\sum_d y_{dt}}$$

Relatively globally scarce topics are defined as those topics  $d$  for which

$$q_{dt} - x_{dt} > 0$$

---

<sup>12</sup> First of all, we know that  $\sum_i w_i = 1$ . Therefore, we have  $\sum_i w_i x_{dt} = x_{dt}$ . Moreover, we have:

$$\begin{aligned} \sum_i w_i x_{idt} &= \sum_i \frac{y_i}{\sum_i y_i} \frac{\sum_n y_{in} s_{ndt}}{\sum_n y_{in}} = \sum_i \frac{\sum_n y_{in} s_{ndt}}{\sum_i y_i} \\ &= \sum_n \frac{\sum_i y_{in} s_{ndt}}{\sum_n y_n} = \sum_n \frac{y_n s_{ndt}}{\sum_n y_n} = x_{dt} \end{aligned} \quad (2.13)$$

Intuitively, these topics have excessively high readership compared to the number of articles (information) that are available through traditional news browsing (direct navigation).

**Popularity and Hard News.** Google News might also promote popular news topics. We say that a topic is popular if it among the top 5 topics measured by pageviews during the pre-shutdown period.

Google News might be particularly effective in lowering the cost of finding hard news and niche topics which are insufficiently covered by most news outlets. Therefore, we might expect a larger volume effect for hard news topic when Google News is no longer available. Formally, we say that a topic covers hard news if it is neither a celebrity or sports topic.

#### 2.4.2 Differences Between Treatment and Control Users

Our matching algorithm does not explicitly match treatment and control users' consumption along all characteristics; the only two incorporated in the matching are breaking news and hard news. Thus, testing whether news consumption differs in other characteristics helps evaluate whether there are important residual sources of heterogeneity between the two groups. However, Table 2.7 demonstrates that treatment and control users look very similar along the five basic characteristics in the matching period; in the Appendix, we test equality all  $2^5 = 32$  categories defined by the vector of characteristics  $c$ . To account for multiple testing, we report whether the differences are significant using the Benjamini-Hochberg (BH) procedure (where we apply the procedure to the set of 5 characteristics, and then separately to the set of 32 categories, in each case reporting whether we can reject equality of the two groups at the 10% level).

Results for the 5 characteristics are reported in Table 2.7, while results for the 32 categories are in Appendix Table A.1. Among the largest differences we find, we see that globally scarce news consumption is 6.6% higher for control users, and the difference is statistically significant (when the hypothesis test is considered in isolation) at the 5% level. We also see differences two of the 32 categories, namely categories  $0 : 1 : 0 : 0 : 1$  and  $0 : 0 : 0 : 0 : 1$  (globally scarce and breaking, and only breaking, respectively), where we reject equality of the news browsing

volumes even after applying the (BH) correction. Although breaking news was included as a feature for matching, other characteristics were also considered and so we did not attain perfect matching in practice. The category that only has the globally scarce characteristic also shows fairly large discrepancies between treated and control, although the test for equality is not significant after applying the BH correction. The other categories are similar in magnitude and differences are not significant.

Table 2.7: Differences between treatment and control users on news types (top 50 topics). To deal with issues of multiple testing, the Benjamini-Hochberg procedure was used to control for the false discovery rate at the 10% level.

characteristic	treatment	control	p-value	BH corrected
Relative scarcity	15.691	15.506	0.694	not rejected
Global scarcity	24.625	26.258	0.031	not rejected
Popularity	15.387	16.255	0.129	not rejected
Hard news	24.275	24.897	0.482	not rejected
Breaking news	12.294	13.119	0.062	not rejected

### 2.4.3 Referrals from Google News versus other Referral Modes, by Characteristic

Next, we show how news consumption obtained through Google News differs from other referrals modes along our five characteristics.

First, Figure 2.4 shows the number of pageviews and share of news consumed through the four main referral modes for treatment and control users. As expected, control users' news consumption patterns are similar before and after the shutdown. For control users (as well as post-shutdown, for treatment users), most news articles are browsed at 3 hours after publication of the articles. Treatment users, however, access news faster after their publication when Google News is available (pre-shutdown): the peak in news consumption occurs at 0 and 1 hours after publication. The Figure also illustrates that Google News is the main referral source for very recently breaking news in the pre-shutdown period. After the shutdown, the peak is very close to that of control users, and it is occurring later than before the shutdown, illustrating the

Figure 2.4: Referral shares and news reading volume as a function of time after publication

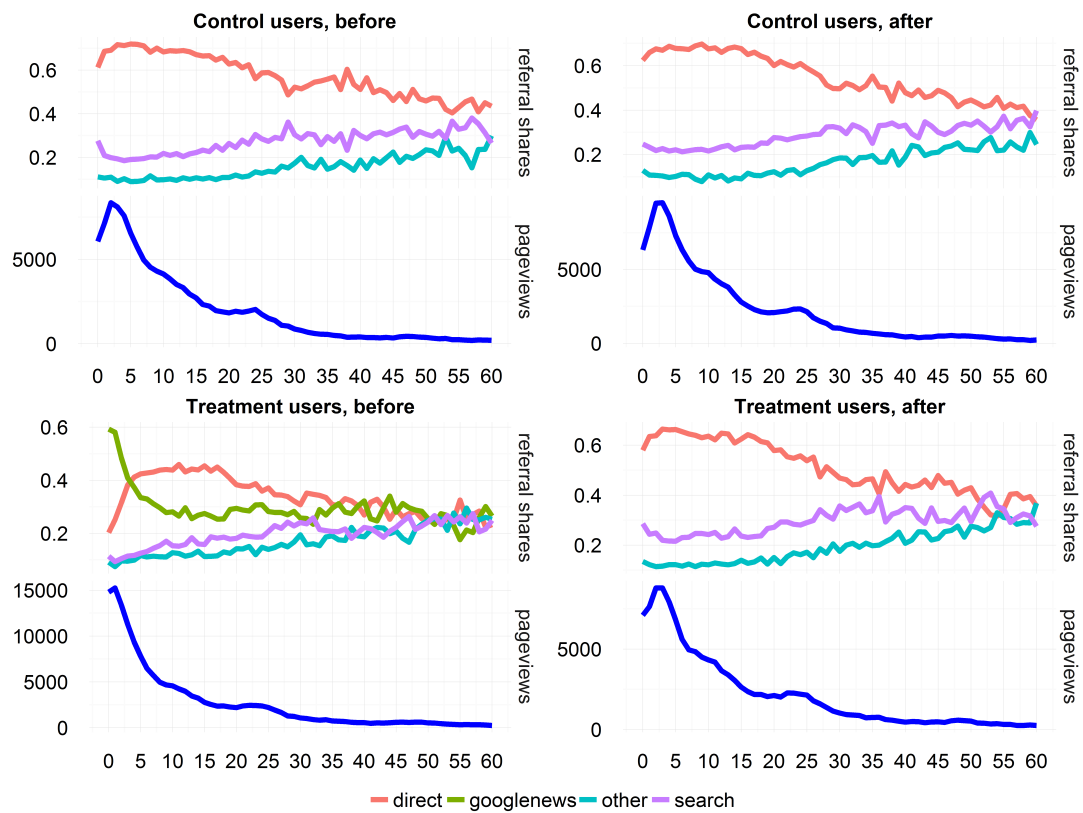
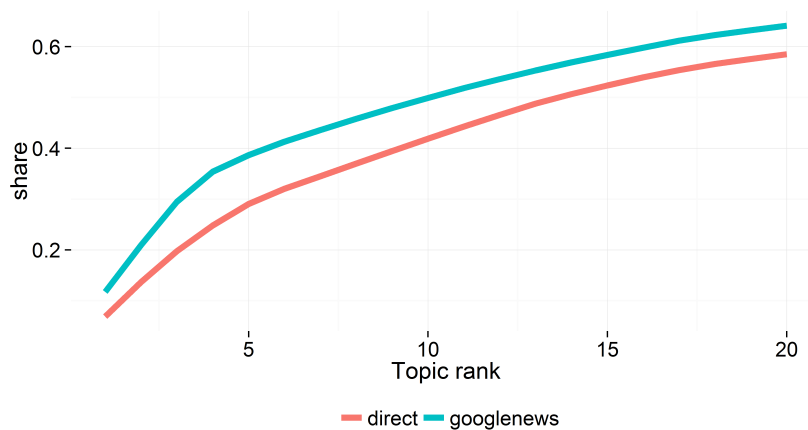


Table 2.8: Share of page views originating from different referral modes by user-scarcity for treatment users pre-shutdown (top 50 topics)

referral mode	non-scarce	scarce
direct	0.397	0.222
googlenews	0.315	0.436
other	0.130	0.158
search	0.157	0.183

Figure 2.5: Cumulative reading shares of top topics for referral modes direct navigation and Google News.



Google News users do not find an alternative source for late-breaking news.

Next, Table 2.8 presents evidence that Google News is an important source of user-scarce news: its share is much higher for scarce than for non-scarce topics (44% versus 32%). We also verified that this result holds for both light and heavy Google News users, using various cut-offs to define light and heavy.

Treatment users also read more popular topics through Google News compared to direct navigation. Figure 2.5 shows the cumulative reading shares for top topics by referral mode (see Tables A.2a and A.2b in the Appendix for a detailed breakdown).

#### 2.4.4 Volume Effect By News Characteristics

We now compare treatment and control users in the pre-shutdown period. Recall, that we matched for each Google News user a corresponding control user who has the same news consumption preferences (through matching in the matching period when both groups have access to the same news browsing technology). Therefore, we can simply compare the consumption of treatment and control users in the pre-shutdown period to simulate the effect of a removal of Google News.

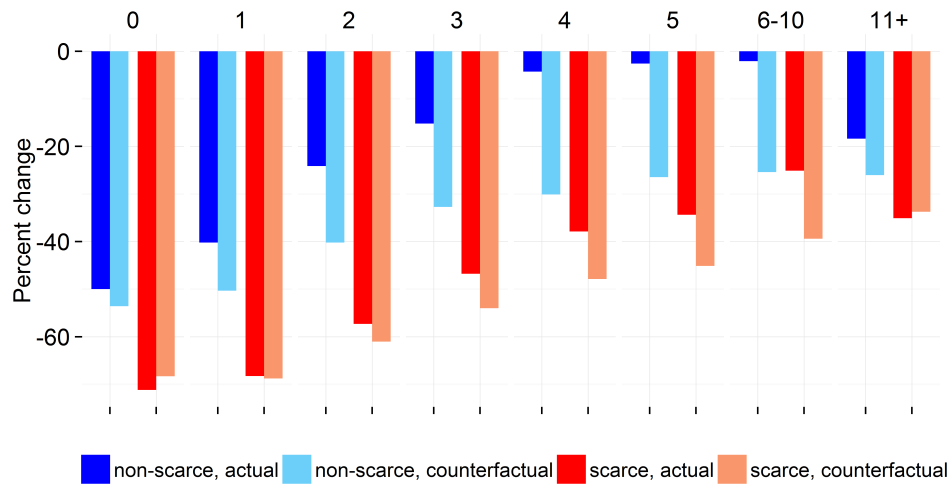
In Figure 2.6 we depict the change in news consumption between treatment and control users by hour of publication (capturing the breaking news effect) and by scarce and non-scarce topics (red versus blue). The lightly shaded bars (light red and light blue) indicate the decrease in news consumption that would occur if treatment users would simply stop using Google News and not substitute to other publishers; this is calculated by simply removing all news viewing that is referred by Google News from the treatment users' aggregate viewing. We refer to this as the "no-substitution" counterfactual. The darkly shaded decreases (dark red and dark blue) indicate the actual decreases.

First, we observe that the overall (actual) decrease in news consumption that we observed in Section 2.3 is reflected in every single combination of hour of publication and scarcity. However, the decrease is particularly stark for breaking news and scarce news: scarce breaking news falls by almost 70% while non-scarce, non-breaking news falls by only less than 20%.

Second, comparing the actual decrease to the no-substitution counterfactual, we find that there is very little user substitution for user-scarce news in the first 2 hours of publication: the overall decrease in news consumptions closely tracks the Google News predicted decrease. This finding highlights an important source of differentiation for Google News, as we see that the users are not able to find a good substitute for this news after the shutdown. In general, the figure illustrates that substitution is stronger for non-scarce news. This is consistent with users switching from Google News to direct navigation, since it is easier to find non-scarce than scarce news in a direct way.

Similarly, Figure 2.7 compares the change in news consumption for pop-

Figure 2.6: No-substitution counterfactual decrease and actual volume decrease by breaking news and user scarcity.



ular (blue) and hard news (red). The contrast between popular and unpopular topics for the actual volume drop is not large in magnitude, while the decreases for hard news are larger than those for celebrity and sports news. Except for popular, celebrity and sports news, where there appears to be little substitution, there is moderate substitution for Google News in the other categories (but by no means full substitution, which would lead to decreases of zero).

All of the differences between counterfactual and actual news decreases in the two figures are statistically significant at the 5% level (even after applying a BH correction for multiple testing), except for the differences corresponding to non-scarce news read in 4-10 hours after publication.

### 2.4.5 Decomposition

In this section we decompose the volume drop of news types (recalling that a category is a vector of binary indicators for news characteristics) into the impact of each of the underlying characteristics. Our empirical strategy is based on comparing the relative consumption of treatment and control users within a news category:

$$\frac{\tilde{N}_c^I}{N_c^I} = \frac{\tilde{\pi}_c + \tilde{\pi}_c^{GN}}{\pi_c} \quad (2.15)$$

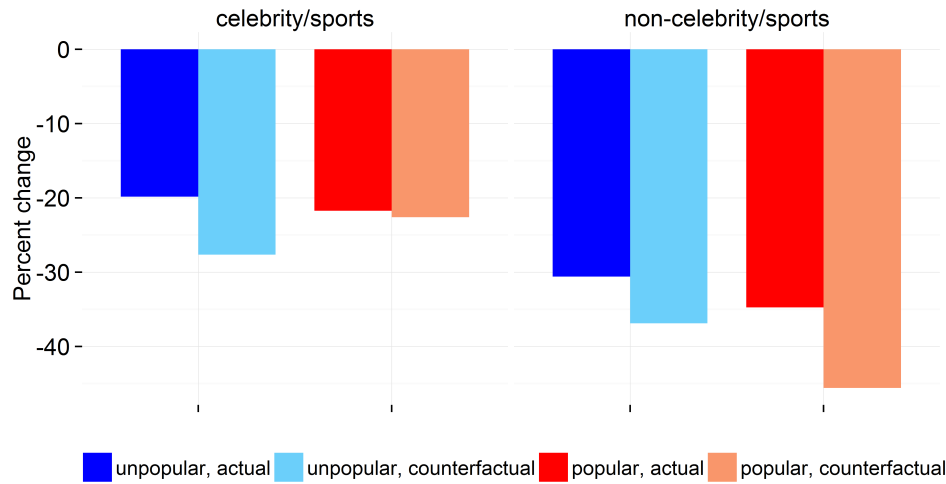


Table 2.9: Estimating the log-linear structural model for volume drop by news characteristic

	<i>Dependent variable:</i>
	log(control / treatment)
Relative scarcity	0.302*** (0.035)
Global scarcity	0.013 (0.020)
Popularity	0.057** (0.026)
Hard news	0.138*** (0.038)
Breaking news	0.222*** (0.022)
constant	0.005 (0.041)

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Figure 2.7: No substitution counterfactual decrease and actual volume decrease by popularity and broad topic.



The right-hand side of this equation measures the relative productivity of traditional and aggregator-augmented news browsing when reading news of type  $c$ .

In order to decompose the right-hand side of (2.15), we assume the following functional form:

$$\frac{\tilde{\pi}_c + \tilde{\pi}_c^{GN}}{\pi_c} = \exp(\gamma_0 + \sum_d \gamma_d \cdot c_d + \epsilon_c) \quad (2.16)$$

Hence,  $\gamma_d$  captures the efficiency loss from losing aggregator access for news dimension  $d$ . The constant  $\gamma_0$  captures the efficiency loss that is not captured by any of the other characteristics.

We estimate the relative contribution of each news characteristic based on the structural model given in (2.16):

$$\log \left( \frac{N_c^I}{\tilde{N}_c^I} \right) = \gamma_0 + \sum_d \gamma_d \cdot c_d + \epsilon_c \quad (2.17)$$

Since we have defined  $D = 5$  binary news characteristics, there are  $2^5 = 32$  news types. We can use the volume drops for those 32 news types to estimate these 6 coefficients.

The estimation procedure we use is a minimum distance procedure. For each news type, we use the sample analogue of equation 2.17 as a mo-

ment condition; details of the method are described in Cameron and Trivedi (2005, p. 202).

We report the results in Table 2.9. As observed in the previous Section, breaking news and user scarcity are the most important factors in explaining the volume drop – they decrease page consumption by 22.2% and 30.2%, respectively. Hard news and popularity also contribute (13.8% and 5.7%), while global scarcity has a small coefficient which is not significantly different from zero. Interestingly, the constant term is small and not statistically significant which implies that the five characteristics that we have included in our simple additive structural model can fully account for the volume drop.

## 2.5 Conclusion

A general theme of innovation since the advent of the internet, starting with eBay and extending to firms like Airbnb and Uber, is that digital intermediaries can reduce search costs and increase the ability of small, flexible sellers to access consumers. The consumer benefits enabled by these intermediaries are clearly very large in magnitude: consumers can access a much larger diversity of products, potentially at lower prices, and small sellers can be found by consumers, potentially creating large welfare gains when consumers find the perfect match for them. These benefits have been the subject of a literature on the welfare benefits of giving consumers access to the “long tail” of products.

Incumbents and regulators have noted a number of challenges when considering regulation of digital intermediaries. The new entrants often have different business models than incumbents, and may appear in some respects to be in the same product market as incumbents (substitutes), while in other ways they appear to be upstream (complements). For example, eBay can compete with traditional retailers, but it can also help small retailers acquire additional customers or sell unsold inventory; Uber competes directly with taxis and limosines, but can also refer customers to existing limosine businesses who have spare capacity; and AirBnb competes with hotels but can help small lodging providers find consumers. In many cases, the new entrants require less fixed cost investment and expenditures in developing the end consumer product,

instead investing research and development to provide sophisticated search and discovery technology. Of course, the value of search and discovery is capped at the value of the products that are available to discover, and so the ultimate welfare evaluation of intermediaries must take into account the long-term quality, assortment, and pricing of final products for consumers. However, economists have long understood that changes in short-term competitive conditions have ambiguous effects on long-term investments in research and development, product quality, and industry structure. Thus, every industry requires its own evaluation.

In the case of News Aggregators, the news industry has called for regulation, resulting in a number of policy interventions across a variety of countries, including the regulatory action in Spain that was the focus of this paper. Most recently, in September of 2016, the European Union has proposed new regulations for news aggregators, as they consider requiring internet companies to pay for news.<sup>13</sup> The empirical evidence we have presented in this paper speaks to some, but not all, of the issues at stake.

Our analysis documents a large, positive effect of Google News on small outlets, as well as on the ability of consumers to access certain types of news, such as breaking news or news that is not well covered on their favorite outlets. These findings highlight the large potential for welfare benefits from improved search and discovery, the “upstream” or complementary role for an intermediary. At the same time, our findings also highlight that while large publishers may not see an effect in overall page views as a result of aggregators, they may lose traffic to their home pages, as well as their role in curating news, as readers read articles referred by Google News at the expense of articles referred by their own home pages (where newspapers monetize the home pages much better than articles). If readers do not pay attention to the identity of the publisher when they read articles on Google News, then the large publishers may lose their incentives to maintain a reputation for quality, and consumers may be less willing to subscribe to the publisher or use the publisher’s mobile application.

Further research is required to assess the ultimate welfare costs and ben-

<sup>13</sup>See, e.g., Schechner and Woo (September 14, 2016; accessed November 17, 2016)

efits. More broadly, our analysis covered only a few weeks before and after the change; ideally, we would want to understand the long-term response of both readers and publishers to changes in policy surrounding aggregators.

## Chapter 3

# Fitted Value Function Iteration with Probability One Contractions

*joint with John Stachurski*

### 3.1 Introduction

Many economic models contain stochastic dynamic programs (SDPs), either as representations of competitive equilibria, or, more commonly, as sub-problems defining the behavior of firms, households, or other individual agents. When solving these SDPs, computational constraints remain a major bottleneck. The difficulty is particularly acute in settings where the SDP must be solved at a large number of different parameterizations, either to compute equilibria (as in Bewley models and dynamic games), or to estimate structural econometric models with unknown parameters in the primitives of the SDP.

In recent years, many specialist algorithms have been proposed. These specialist algorithms take advantages of certain features of a given application in order to obtain fast convergence rates. In most of these studies, global (or even local) convergence of the algorithm to the optimum is not proved. Instead, the authors test the algorithm against a special case possessing analytical solutions, or compare their convergence rates against competing algorithms in a specific example. Needless to say, this provides no guarantee of convergence for problems other than the examples treated in the study.

There are good reasons to be concerned about whether common dynamic programming routines do in fact converge to optimal value functions and policies. This is particularly the case in continuous state space settings, where iterative techniques involve some form of function approximation. The interplay between function approximation and dynamic programming routines is known to be relatively delicate. For example, in value function iteration, many function approximation techniques fail to preserve the global contraction properties of the Bellman operator, and several authors have already demonstrated how adding standard function approximation steps can lead to cycles and failure of convergence (Boyan and Moore, 1995; Baird, 1995; Tsitsiklis and Van Roy, 1996).

While specialist algorithms known to converge quickly in particular settings are certainly of value, in this paper our aim is to study a simulation-based value iteration algorithm that has guaranteed convergence properties across a very wide variety of applications. Our set up includes a function approximation step, admits continuous state-action spaces, and makes no use of densities. We provide general conditions under which the fixed point of our random fitted Bellman operator converges uniformly to the value function with probability one. Under additional regularity conditions, we show that the supremum norm deviation is  $O_P(n^{-1/2})$ .

Our techniques provide a natural alternative to discretized value function iteration, a method which also has very broad applicability, and remains a popular benchmark in economic applications. In discretized value function iteration, a continuous state/action problem is replaced by a “nearby” discrete problem. Relative to the method we study here, discretization has several disadvantages. First, while the discretized algorithm always locates the solution to the *discrete* problem, the deviation between this discrete solution and the solution to the original problem is not easily obtained. To the best of our knowledge, no global convergence results are available in a setting as general as the one that we treat here.<sup>1</sup> Second, in terms of finite time properties, discrete representation

<sup>1</sup>One issue is that discretization errors for continuous curves tend to be bounded in terms of derivatives, which fail to exist or cannot be bounded in many economic settings. Optimal growth models often have unbounded derivatives as a result of Inada conditions. Derivatives can fail to exist when models include discrete choices, binding constraints, non-convexities and so on.

of continuous curves is costly relative to continuous parametric representations and inherently subject to the curse of dimensionality.<sup>2</sup>

### 3.1.1 Methodology

Successful fitted value function iteration in a continuous state setting requires careful choice of both function approximation scheme and of numerical integration method. In the numerical dynamic programming literature, most authors use standard approximation theory to guide their choices. For example, Chebyshev polynomials are popular for function approximation because classical approximation theory shows they have desirable properties in terms of fitting a certain class of functions. However, the final objective with numerical dynamic program is to produce an estimate of the value or policy function, and to this end the criteria for “good” approximation methods should include not only their approximation properties at each step, but also their interplay with the overall dynamic programming algorithm.

As discussed above, failure to consider this interplay may result in iterative techniques that break the global contraction property of the theoretical Bellman operator, causing instability. Without contractiveness, small errors can be compounded at each iteration, and analysis of the algorithm is highly problematic. For this reason, we focus on approximation methods that preserve contractiveness. On the function approximation side, we use nonexpansive approximation operators, as pioneered by Judd and Solnick (1994) and Gordon (1995).<sup>3</sup> On the numerical integration side we use Monte Carlo. We prove below that this combination preserves contractiveness with probability one. (The “probability one” statement means that the approximate Bellman operator is a contraction

---

<sup>2</sup>The number of data points needed to represent function in  $\mathbb{R}^d$  parametrically may be polynomial in  $d$ , while discrete representations are always exponential. The intractability of discrete representations in moderate to high dimensions has led practitioners in fields such as engineering and computer science to *reverse* the discretization process, replacing inherently discrete dynamic programs with continuous ones. This idea dates back to Bellman (Bellman et al., 1963).

<sup>3</sup>The focus in Judd and Solnick (1994) is on shape-preserving approximations. It turns out that some of these approximation methods are nonexpansive (see, for example, the proof of Theorem 4). Gordon (1995) is concerned exclusively with nonexpansive approximation. Other closely related studies in terms of approximation methods include Drummond (1996), Guestrin et al. (2001), Santos and Vigo-Aguiar (1998) and Stachurski (2008).



for every realization of the Monte Carlo sample.) The contractiveness of the approximate Bellman operator is central to all of our error analysis.

Regarding the use of Monte Carlo for numerical integration, a potential disadvantage is that Monte Carlo integration is often less efficient than deterministic routines in low dimensional settings with smooth functions. On the other hand, Monte Carlo methods tend to perform well in higher dimensional settings, or when the target function is less smooth. In the present setting, there is an additional efficiency: The Monte Carlo sample is drawn once-off to form the approximate Bellman operator, and re-used many times to evaluate expectations.

Regarding the function approximation step, the idea behind nonexpansive approximation operators is straightforward: Let  $A$  be an operator, mapping real-valued function  $w$  into another real-valued function  $Aw$  on the same domain. The image  $Aw$  is interpreted as an approximation of  $w$ . The operator  $A$  is called nonexpansive if, for any two functions  $v$  and  $w$ , the images  $Av$  and  $Aw$  are no further apart than the preimages  $v$  and  $w$ . (Here distance is measured in the supremum norm.)

Examples of nonexpansive approximation schemes include continuous piecewise linear interpolation,  $k$ -nearest neighbors, Schoenberg variation diminishing splines and the class of kernel smoothers. With kernel smoothers, the approximation  $Aw$  of  $w$  has the form

$$(Aw)(x) = \sum_{i=1}^m w(x_i) \psi \left\{ \frac{\|x - x_i\|}{h} \right\} \eta(x). \quad (3.1)$$

The scalar  $h$  is a smoothing parameter,  $\{x_i\}_{i=1}^m$  is a set of grid points,  $\psi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is continuous and monotone decreasing, and  $\eta(x)$  is a normalization term defined as  $\eta(x) := 1/[\sum_{j=1}^m \psi(\|x - x_j\|/h)]$ . Essentially,  $(Aw)(x)$  is a weighted average of the function values  $\{w(x_i)\}$  on the grid, with higher weight for grid points close to  $x$ . A common choice for  $\psi$ , particularly in higher dimensions, is  $\psi(t) = \exp(-t^2)$ . In this case  $A$  is called a (Gaussian) radial basis function kernel smoother. Nonexpansiveness of  $A$  is shown in Gordon (1995, Theorem 3.2).

An illustration of nonexpansiveness using a radial basis kernel is given in Figure 3.1. The figure shows two functions  $w$  and  $v$  as dashed lines, and their approximations  $Aw$  and  $Av$  as unbroken lines. (The approximation is deliberately coarse, so that the difference between the func-

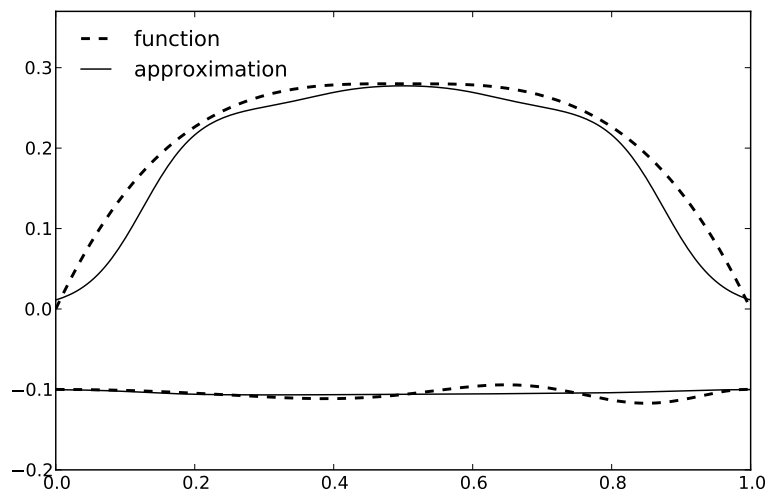


Figure 3.1: Nonexpansive approximation with a radial basis kernel smoother

tions and their approximations is clearly visible.) It can be seen from the figure that the sup norm (i.e., maximal) deviation between the approximations is no larger than the maximal deviation between the original functions. For comparison, see Figure 3.2, where the same functions are approximated with Chebyshev polynomials. In this case the approximating polynomial overshoots the higher function in the middle of the interval, and the maximal deviation (occurring near 0.5 on the  $x$ -axis) is increased by approximation.<sup>4</sup>

Figure 3.3 shows an example of instability in fitted value function iteration, when Chebyshev polynomials are used for approximation in the standard optimal growth model with log utility and Cobb-Douglas production  $f(k) = k^\alpha$ . (See Example 3.6.1 for a description of the model.) Here the discount factor is 0.95,  $\alpha = 0.33$  and the state space is from  $10^{-5}$  to 1. The shock  $z$  is multiplicative and lognormal, with parameters  $\mu = 0$  and  $\sigma = 0.25$ . The initial condition for the iteration is the utility function. The Chebyshev polynomials are of order 10 with 150 nodes, and integration is Gaussian quadrature of order 5. In the plot, the true value function is the dashed line. The iterates of the approximate Bellman operator diverge downwards (successive iterates are plotted in darker grey) and away from the true value function.<sup>5</sup>

In this example, non-convergence is relatively extreme. (After all, fail-

<sup>4</sup>Both approximations use 5 grid points. The kernel smoother uses the radial kernel  $\psi(t) = \exp(-t^2)$

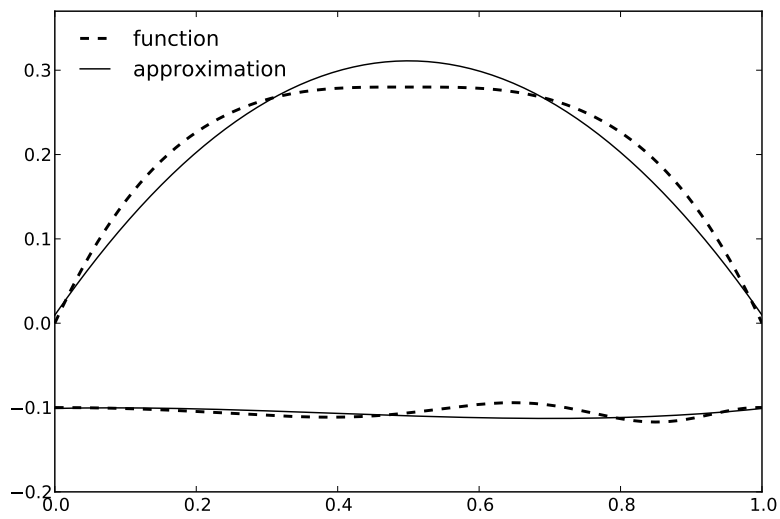


Figure 3.2: Expansive approximation with Chebyshev polynomials

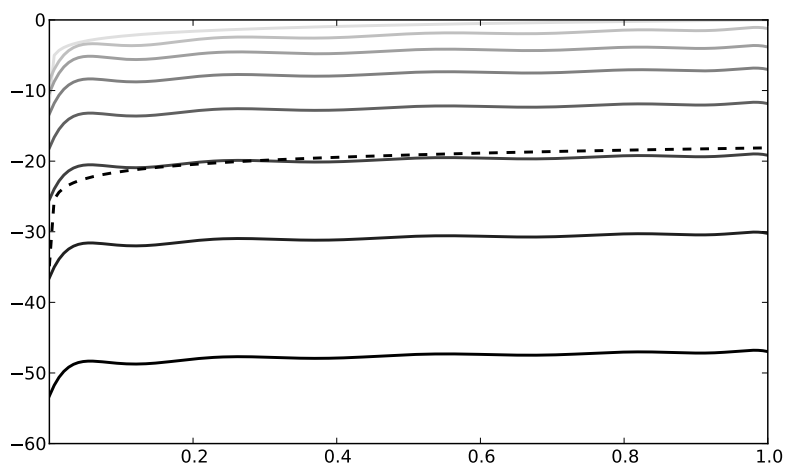


Figure 3.3: Instability in fitted VFI with Chebyshev polynomials

ure of convergence does not necessarily entail divergence. For example, initial convergence followed by cycling is a less extreme failure of convergence that may be more likely in practice.) However, the results are informative because they suggest that convergence results for this particular class of polynomials cannot be obtained without restrictions that may not hold even for very basic economic models. (Judd and Solnick (1994) also mention finding examples where “wild divergence” occurs while conducting fitted value iteration with polynomials, although details are not given in the paper.)

Finally, it should also be added that the probability one global stability that is guaranteed in our algorithm is not necessarily important in and of itself, since a stable algorithm can in principle converge to an object that has no relationship to the object one wishes to compute. Rather, stability is a first step that allows us to pursue a general consistency result.

### 3.1.2 Related Literature

The emphasis in the paper is on global convergence in a general setting. Numerical methods for more specific models with additional structure can be found in many papers. The literature is too large to survey here. Useful introductions can be found in Marimon and Scott (2001) or Aruoba et al. (2006). Both of these references include discussion of methods for solving smooth SDPs where optimal policies satisfy Euler equations. (We assume no such smoothness here.) An iterative method for concave SDPs is analyzed in the recent paper of Fukushima and Waki (2013). Some well-known algorithms based on Monte Carlo include those found in Keane and Wolpin (1994), Rust (1997), Pakes and McGuire (2001) and Longstaff and Schwartz (2001).

Several authors have published studies on finite-time bounds for fitted value function iteration. For example, Rust (1997) proposed an inge-

---

with  $h = 8$ . The Chebyshev polynomial approximation uses a third order polynomial.

<sup>5</sup>The growth model is the one presented in Section 3.6.3. For the distance between iterates see the last column of Table 3.1 (Section 3.7). In the exercise, computation of Chebyshev polynomials was standard (see, e.g., Judd, 1998, Chapter 6). We repeated the experiment with the extended Chebyshev array method and obtained similar results. In Section 3.7 below, we re-run the same experiment, but this time using various approximation methods that satisfy our conditions. In all cases, the sequence of iterates converge (all but last column of Table 3.1). The code for all our experiments can be found at <https://sites.google.com/site/fviprobone/>.

nious function approximation step that can be implemented when the one-step transition probabilities for the dynamic programming problem are absolutely continuous with respect to Lebesgue measure (i.e., the distribution for the next period state given current state and action can be represented by a density). He proved that for decision problems satisfying certain Lipschitz conditions, his algorithm breaks the curse of dimensionality, in the sense that worst-case computational complexity is polynomial in the dimension of the state space. Further important developments for models satisfying similar restrictions were reported in Munos and Szepesvári (2008). Finally, Stachurski (2008) provided finite time bounds for fitted value iteration, although the problem of numerical integration was not considered.

The contribution of this paper is different. The generality of our setting precludes us from developing tight finite-time bounds. Instead, our focus is on obtaining consistency under very weak assumptions. For example, in our consistency result, we make no smoothness or differentiability assumptions. This allows us to include models with discrete choices, occasionally binding constraints, non-convexities and so on. In addition, we do not assume that the one-step transition probabilities are absolutely continuous, an assumption that was central to Rust's (1997) algorithm. This additional generality is important in economics, since many applications have one-step transition probabilities that fail to be absolutely continuous. To give an example, consider a benchmark macroeconomic model, where next period capital stock is given by

$$k_{t+1} = (1 - \delta)k_t + f(k_t, z_t) - c_t.$$

Here  $\delta$  is a depreciation rate,  $f$  is a production function,  $c_t$  is consumption and  $\{z_t\}$  is an exogenous shock process, typically Markovian. Observe that as soon as the current state  $(k_t, z_t)$  and the current action  $c_t$  is given, next period capital is deterministic. As a result, the one-step transition probability fails to be absolutely continuous, and cannot be represented by a density.

In this example, the problem is caused by stochastic rank deficiency—the shock space has lower dimension than the state space. While the example is simplistic, it is also representative of the growth and macroeconomic literature—see for example the standard formulation of Stokey and Lucas (1989, p. 290)—and illustrates the fact that many models in

these fields cannot be treated with density-based approaches unless modifications are imposed.

In addition to stochastic rank deficiency, failure of absolute continuity can be caused by discrete shocks (e.g., labor productivity shocks following discrete Markov chains), constraints and other common features. Representative dynamic programming problems where the transition probability fails to be absolutely continuous include those found in Kydland and Prescott (1982, p. 1354), Aiyagari (1994), Huggett (1997) and Clementi and Hopenhayn (2006).

### 3.1.3 Outline

Section 3.2 of the paper provides background concepts and notation. Section 3.3 defines the model, and Section 3.4 introduces the algorithm. Section 3.5 provides convergence results, Section 3.6 discusses rates of convergence, Section 3.7 gives applications, and Section 3.8 concludes. Proofs can be found in Section 3.9.

## 3.2 Preliminaries

We begin by introducing notation. For topological space  $\mathbb{T}$ , the symbol  $\mathcal{C}(\mathbb{T})$  denotes the collection of continuous, bounded, real-valued functions on  $\mathbb{T}$ , while  $\|\cdot\|$  is the supremum norm on  $\mathcal{C}(\mathbb{T})$ . Operator  $S: \mathcal{C}(\mathbb{T}) \rightarrow \mathcal{C}(\mathbb{T})$  is called a contraction of modulus  $\rho$  if  $0 \leq \rho < 1$  and

$$\|Sv - Sw\| \leq \rho \|v - w\| \text{ for all pairs } v, w \in \mathcal{C}(\mathbb{T}). \quad (3.2)$$

$S$  is called nonexpansive if (3.2) holds with  $\rho = 1$ . By Banach's contraction mapping theorem, every contraction  $S$  of modulus  $\rho$  on  $\mathcal{C}(\mathbb{T})$  has a unique fixed point  $W \in \mathcal{C}(\mathbb{T})$ , and, moreover,  $\|S^n w - W\| = O(\rho^n)$  for each  $w \in \mathcal{C}(\mathbb{T})$ .

**Lemma 3.2.1.** *Let  $S$  and  $S'$  be operators from  $\mathcal{C}(\mathbb{T})$  to itself.*

1. *If  $S$  is nonexpansive and  $S'$  is a contraction of modulus  $\rho$ , then the composition  $S \circ S'$  is a contraction of modulus  $\rho$ .*

2. If  $S$  and  $S'$  are both contractions of modulus  $\rho$  with fixed points  $W$  and  $W'$  respectively, then  $\|W - W'\| \leq (1 - \rho)^{-1} \|SW' - W'\|$ .

Part 1 is trivial. For a proof of part 2, see, for example, Rust (1997, Lemma 2.1).

In what follows, all random variables are defined on a common probability space  $(\Omega, \mathcal{F}, \mathbf{P})$ , and  $\mathbf{E}$  is the expectation with respect to  $\mathbf{P}$ . If  $X$  is a map from  $\Omega$  into  $\mathbb{R}$  that is not necessarily measurable, then the outer expectation of  $X$  is  $\mathbf{E}^*X := \inf_Y \mathbf{E}Y$ , where the infimum is over all real random variables  $Y$  such that  $X \leq Y$  and  $\mathbf{E}Y$  exists (cf., e.g., van der Vaart, 1998, p. 258). For a sequence of possibly nonmeasurable maps  $\{U_n\}$  from  $\Omega$  into a metric space  $(\mathbb{T}, d)$  and a  $\mathbb{T}$ -valued random variable  $U$ , we say that  $U_n \rightarrow U$  holds  $\mathbf{P}^*$ -almost surely if there exists a measurable real-valued sequence  $\Delta_n$  with  $d(U_n, U) \leq \Delta_n$  and  $\mathbf{P}\{\Delta_n \rightarrow 0\} = 1$ . We say that  $U_n$  converges in distribution to  $U$  if  $\mathbf{E}^*g(U_n) \rightarrow \mathbf{E}g(U)$  for every  $g \in \mathcal{C}(\mathbb{T})$ . For the former convergence we write  $U_n \xrightarrow{a.s.*} U$ , while for the latter we write  $U_n \xrightarrow{d*} U$ .

The continuous mapping theorem continues to hold in this setting:

**Lemma 3.2.2.** *If  $\mathbb{T}'$  is another metric space and  $g: \mathbb{T} \rightarrow \mathbb{T}'$  is continuous, then*

$$U_n \xrightarrow{d*} U \implies g(U_n) \xrightarrow{d*} g(U).$$

Let  $\{X_n\}_{n \geq 1}$  be a sequence of (not necessarily measurable) maps from  $\Omega$  into  $\mathbb{R}$ . We write  $X_n = O_{P^*}(n^{-1/2})$  if there exists a sequence of real-valued random variables  $\{\Delta_n\}_{n \geq 1}$  such that  $|X_n| \leq \Delta_n$  for all  $n$  and  $\Delta_n = O_P(n^{-1/2})$ .

### 3.3 Set Up

In this section we introduce a general stochastic dynamic programming problem and describe the value function iteration algorithm.

### 3.3.1 The Model

Consider an SDP of the following form. A controller observes the state  $x \in \mathbb{X}$  of a given system, and responds with an action  $a$  from a feasible set  $\Gamma(x) \subset \mathbb{A}$ . Given this state-action pair  $(x, a)$ , the controller receives current reward  $r(x, a)$ , and the new state is determined as  $x' = F(x, a, U)$ , where  $U$  is a draw from a fixed distribution  $\phi$ . The process now repeats. The controller's objective is to maximize the sum of expected discounted rewards given discount factor  $\beta$ .

The sets  $\mathbb{X}$  and  $\mathbb{A}$  are referred to as the state and action spaces respectively, and  $\Gamma$  is called the feasible correspondence. We let

$$\mathbb{G} := \text{graph } \Gamma := \{(x, a) \in \mathbb{X} \times \mathbb{A} : a \in \Gamma(x)\}.$$

The set  $\mathbb{G}$  is called the set of feasible state-action pairs.

A feasible policy is a Borel measurable map  $\sigma: \mathbb{X} \rightarrow \mathbb{A}$  such that  $\sigma(x) \in \Gamma(x)$  for all  $x \in \mathbb{X}$ . Let  $\Sigma$  be the set of all such policies. The controller's problem is

$$\max_{\sigma \in \Sigma} \left\{ \mathbf{E} \sum_{t=0}^{\infty} \beta^t r(X_t, \sigma(X_t)) \right\}$$

subject to

$$X_{t+1} = F(X_t, \sigma(X_t), U_{t+1}) \quad \text{with } x_0 \text{ given.} \quad (3.3)$$

Almost any stationary infinite horizon dynamic program with additively separable preferences can be formulated in this way. We assume throughout the paper that

1.  $\mathbb{X}$  and  $\mathbb{A}$  are compact metric spaces.
2.  $\Gamma$  is continuous and compact-valued.
3. The shocks  $\{U_t\}_{t \geq 1}$  are IID with common distribution  $\phi$ .<sup>6</sup>

<sup>6</sup>The assumption of IID shocks is not restrictive. For example, consider the following macroeconomic model with exogenous Markov shock sequence: The state space is a product space  $K \times Z \subset \mathbb{R}^m \times \mathbb{R}^n$ , where  $k \in K$  is a vector of endogenous variables and  $z \in Z$  is a vector of exogenous variables. Technology is summarized by a feasible set  $\Theta \subset K \times Z \times K$ . The exogenous process  $\{z_t\}_{t \geq 0}$  evolves according to  $z_{t+1} = g(z_t, \epsilon_{t+1})$ , where  $\{\epsilon_t\}_{t \geq 1}$  is IID. Instantaneous rewards are given by  $v(k, z, k')$ . This formulation is a special case of our SDP. To see this, for the state take  $x := (k, z) \in K \times Z$ , and for the action take  $a := k' \in K$ . The feasible correspondence is  $\Gamma(x) := \Gamma(k, z) := \{k' \in K : (k, z, k') \in \Theta\}$ . The shock is  $u := \epsilon$ , and the transition function is  $F(x, a, u) := F(k, z, k', \epsilon) := (k', g(z, \epsilon)) \in K \times Z$ . The reward function is  $r(x, a) := r(k, z, k') := v(k, z, k')$ .



4.  $\phi$  is a Borel probability measure over metric space  $\mathbb{U}$ .
5. The reward function  $r: \mathbb{G} \rightarrow \mathbb{R}$  is continuous.
6. The function  $\mathbb{G} \ni (x, a) \mapsto F(x, a, u) \in \mathbb{X}$  is continuous for all  $u \in \mathbb{U}$ .

For  $\{X_t\}$  as given by (3.3), let  $V_\sigma(x_0) = \mathbf{E} \sum_{t=0}^{\infty} \beta^t r(X_t, \sigma(X_t))$ . Let  $T: \mathcal{C}(\mathbb{X}) \rightarrow \mathcal{C}(\mathbb{X})$  be the Bellman operator, defined at  $v \in \mathcal{C}(\mathbb{X})$  by

$$Tv(x) := \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \int v[F(x, a, u)] \phi(du) \right\} \quad (x \in \mathbb{X}). \quad (3.4)$$

For  $v \in \mathcal{C}(\mathbb{X})$ , a policy  $\sigma \in \Sigma$  is called  $v$ -greedy if  $\sigma(x)$  is a maximizer of the right-hand side of (3.4) for all  $x \in \mathbb{X}$ . The value function  $V_T$  is defined pointwise on  $\mathbb{X}$  by  $V_T(x) = \sup_{\sigma \in \Sigma} V_\sigma(x)$ . A policy  $\sigma \in \Sigma$  is called optimal if  $V_\sigma = V_T$ .

### 3.3.2 Value Function Iteration

The following result is standard:

**Theorem 3.3.1.** *Under Assumptions 1–6 above,*

1.  $T$  is a contraction of modulus  $\beta$  on  $\mathcal{C}(\mathbb{X})$ , and  $V_T$  is the unique fixed point;
2. a policy  $\sigma \in \Sigma$  is optimal if and only if it is  $V_T$ -greedy; and
3. at least one such policy exists.

In principle,  $V_T$  can be calculated by value function iteration (VFI), which involves fixing an initial  $v \in \mathcal{C}(\mathbb{X})$  and iterating with  $T$ . From Theorem 3.3.1 we have  $\|T^k v - V_T\| = O(\beta^k)$ . Using this fact and optimality of  $V_T$ -greedy policies, one can show that a  $T^k v$ -greedy policy is approximately optimal when  $k$  is sufficiently large.<sup>7</sup>

<sup>7</sup>See, e.g., Puterman (1994, Theorem 6.3.1). An appropriate  $k$  is usually chosen according to some stopping criterion that depends on the deviation between successive iterates of  $T$ .

### 3.4 Random Fitted VFI

Evaluation of the expression  $r(x, a) + \beta \int v[F(x, a, u)]\phi(du)$  on the right-hand side of (3.4) requires approximation of the integral. To do so we use Monte Carlo, which allows us to preserve the contractiveness of the Bellman operator, as discussed below. Another advantage of using Monte Carlo in our set up is that we will be able to evaluate every integral by drawing a single sample

$$U_1, \dots, U_n \stackrel{\text{iid}}{\sim} \phi \quad (3.5)$$

once off. Given this sample, we then iterate with the random Bellman operator  $R_n$  defined by

$$R_n v(x) := \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \frac{1}{n} \sum_{i=1}^n v[F(x, a, U_i)] \right\} \quad (x \in \mathbb{X}). \quad (3.6)$$

A realization of  $\omega \in \Omega$  determines a particular realization  $\{U_i(\omega)\}_{i=1}^n$  of the sample (3.5), which in turn defines a realization  $R_n(\omega)$  of  $R_n$ . Each realization  $R_n(\omega)$  is an operator from  $\mathcal{C}(\mathbb{X})$  to itself.

A second numerical issue is as follows: If  $\mathbb{X}$  is infinite, then, for arbitrary given  $w \in \mathcal{C}(\mathbb{X})$ , one cannot evaluate either  $Tw(x)$  or  $R_n w(x)$  at each  $x \in \mathbb{X}$  in finite time (or store the functions in a look-up table). Hence, we approximate  $R_n w$  using a finite parametric representation. To do so, we introduce an approximation operator  $A: \mathcal{C}(\mathbb{X}) \rightarrow \mathcal{A}(\mathbb{X}) \subset \mathcal{C}(\mathbb{X})$ , where, given function  $w \in \mathcal{C}(\mathbb{X})$ ,  $Aw$  is an approximation of  $w$ , and  $\mathcal{A}(\mathbb{X})$  is a class of functions such that each element can be represented by a finite number of parameters. In addition, we assume that  $Aw$  can be computed on the basis of a finite number of observations (i.e., by observing the value of  $w(x)$  at a finite number of  $x \in \mathbb{X}$ ). For example, the mapping  $w \mapsto Aw$  might proceed by evaluating  $w$  on a fixed and finite grid of points  $\{x_i\}_{i=1}^m$ , and then constructing  $Aw$  based on these “interpolation” points. Finally, we assume throughout that  $A$  is nonexpansive.

**Example 3.4.1.** Continuous piecewise linear interpolation in  $\mathbb{R}^d$  is a non-expansive approximation scheme.<sup>8</sup>

<sup>8</sup>To describe it formally, let  $\mathbb{X}$  be a convex subset of  $\mathbb{R}^d$ , let  $\mathbb{V}$  be a finite subset of  $\mathbb{X}$  such that the

Other nonexpansive schemes include kernel smoothers (see Section 3.1.1),  $k$ -nearest neighbors, shape-preserving Schumaker splines, and the variation diminishing splines of Schoenberg.

The complete procedure for random fitted value function iteration is given in Algorithm 1. In step 3,  $(AR_n)^k$  is the  $k$ -th composition of the operator  $AR_n := A \circ R_n$  with itself. In practice, when applying the operator  $AR_n$  to a given function  $w$ , first  $R_n w$  is evaluated on a finite grid of points  $\{x_i\}_{i=1}^m$  by solving the maximization problem in (3.6) at each  $x_i$ .  $A$  is then applied to produce the fitted function  $AR_n w$ .

---

**Algorithm 1:** Random Fitted VFI

---

- 1 generate the sample  $\{U_1, \dots, U_n\} \stackrel{\text{iid}}{\sim} \phi$  in (3.5) ;
  - 2 fix  $v \in \mathcal{C}(\mathbb{X})$  ;
  - 3 compute  $v_k := (AR_n)^k v$  by starting at  $v$  and iterating with  $AR_n$   $k$  times;
  - 4 compute a  $v_k$ -greedy policy  $\sigma$  ;
- 

### 3.5 Analysis

We begin our analysis with the following lemma, the proof of which can be found in Section 3.9.

**Lemma 3.5.1.** *The operator  $AT$  is a contraction on  $\mathcal{C}(\mathbb{X})$  of modulus  $\beta$ . The operator  $AR_n(\omega)$  is also a contraction on  $\mathcal{C}(\mathbb{X})$  of modulus  $\beta$  for all  $n \in \mathbb{N}$  and all  $\omega \in \Omega$ .*

As a consequence, there exists

1. a unique fixed point  $V_{AT} \in \mathcal{C}(\mathbb{X})$  of  $AT$
2. a unique fixed point  $V_{AR_n(\omega)} \in \mathcal{C}(\mathbb{X})$  of  $AR_n(\omega)$  for each  $\omega \in \Omega$

---

convex hull of  $\mathbb{V}$  is  $\mathbb{X}$ , and let  $T$  be a  $\mathbb{V}$ -triangularization of  $\mathbb{X}$ . (That is,  $T$  is a finite collection of non-degenerate simplexes such that the vertices of each simplex lie in  $\mathbb{V}$  and any two simplexes intersect on a common face or not at all.) Given a simplex  $\Delta \in T$  with vertices  $\zeta_1, \dots, \zeta_{d+1}$ , each  $x \in \Delta$  can be represented uniquely as  $\sum_{i=1}^{d+1} \lambda(x, i) \zeta_i$ , where  $\lambda(x, i)$  is its  $i$ -th barycentric coordinate relative to  $\Delta$ . (By definition,  $\lambda(x, i) \geq 0$  and  $\sum_{i=1}^{d+1} \lambda(x, i) = 1$ .) For  $v \in \mathcal{C}(\mathbb{X})$ , we define  $A$  by  $Av(x) = \sum_{i=1}^{d+1} \lambda(x, i) v(\zeta_i)$ . The operator  $A$  is nonexpansive (see, e.g., Stachurski, 2008).

The operator  $AT$  is the fitted Bellman operator where function approximation is included, but the integral is computed exactly. Its fixed point  $V_{AT}$  is deterministic. On the other hand,  $V_{AR_n}$  is random. In what follows, we refer to  $V_{AR_n}$  as a random function, although  $\omega \mapsto V_{AR_n}(\omega)$  may not be Borel measurable as a mapping from  $\Omega$  to  $\mathcal{C}(\mathbb{X})$ .

Our primary goal is to study the convergence of  $V_{AR_n}$  to the value function  $V_T$ .<sup>9</sup> By the triangle inequality, the error can be decomposed as

$$\|V_T - V_{AR_n}\| \leq \|V_T - V_{AT}\| + \|V_{AT} - V_{AR_n}\| \quad \forall n \in \mathbb{N}. \quad (3.7)$$

Let us consider the two terms on the right-hand side of (3.7). The first term is the function approximation error, caused by replacing  $T$  with  $AT$ . The second is the integral approximation error, caused by replacing  $AT$  with  $AR_n$ . We now consider these two errors in turn.

Consider first the function approximation error  $\|V_T - V_{AT}\|$ . Given a suitable approximation scheme, establishing convergence of the error to zero is relatively straightforward. Rates of convergence will depend on the particular function approximation scheme used in a given implementation, but sufficiently “fine” approximations will make the error arbitrarily small, provided that the range space of  $A$  becomes sufficiently rich. To justify this claim, consider the kernel smoother  $A$  in (3.1). The next result shows that the function approximation error can be made arbitrarily small for an operator  $A$  in this class without additional assumptions.

**Lemma 3.5.2.** *For any  $\epsilon > 0$ , there exists a choice of  $\{x_i\}_{i=1}^m$ ,  $\psi$  and  $h$  such that the corresponding operator  $A$  in (3.1) satisfies  $\|V_T - V_{AT}\| < \epsilon$ .*

Now we turn our attention to the integral approximation error, which is the second term on the right-hand side of (3.7). Our first major result for the paper shows probability one convergence without any additional assumptions.<sup>10</sup>

**Theorem 3.5.1.**  $\|V_{AT} - V_{AR_n}\| \xrightarrow{a.s.*} 0$  as  $n \rightarrow \infty$ .

<sup>9</sup>The relative optimality of the  $(AR_n)^k v$ -greedy policy  $\sigma$  computed by Algorithm 1 depends on the deviation between  $(AR_n)^k v$  and  $V_T$ . Using the triangle inequality, we can bound the latter by  $\|(AR_n)^k v - V_{AR_n}\| + \|V_{AR_n} - V_T\|$ . By Lemma 3.5.1, the first term is  $O(\beta^k)$  in  $k$ . Convergence of  $V_{AR_n}$  to  $V_T$  is less clear, and hence we focus on this term.

<sup>10</sup>Since Borel measurability of  $\omega \mapsto V_{AR_n}(\omega)$  is inherently problematic, the theorem uses the concept of  $\mathbf{P}^*$ -almost sure convergence.

Regarding the proof, recall that a class  $\mathcal{H}$  of bounded measurable functions mapping  $\mathbb{U}$  into  $\mathbb{R}$  is called  $\phi$ -Glivenko-Cantelli if the strong law of large numbers holds uniformly over  $\mathcal{H}$ , in the sense that if  $\{U_i\}$  is an IID sample from  $\phi$ , then

$$\lim_{n \rightarrow \infty} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(U_i) - \int h d\phi \right| = 0 \quad \mathbf{P}^*\text{-almost surely} \quad (3.8)$$

(cf., e.g., van der Vaart, 1998). In the proof of Theorem 3.5.1, a sequence of relatively straightforward manipulations shows that if we set

$$\mathcal{H} := \{\mathbb{U} \ni u \mapsto V_{AT}[F(x, a, u)] \in \mathbb{R} : (x, a) \in \mathbb{G}\}, \quad (3.9)$$

then the error  $\|V_{AT} - V_{AR_n}\|$  is bounded above by a constant times the supremum on the left-hand side of (3.8). That  $\mathcal{H}$  has the Glivenko-Cantelli property is then verified using a well-known sufficient condition. Details are in Section 3.9.

## 3.6 Rates of Convergence

The result in Theorem 3.5.1 gives no indication as to the rate of convergence. To obtain a rate, we need to give a rate for the convergence in (3.8). The  $\phi$ -Glivenko-Cantelli property used in the proof of Theorem 3.5.1 is not sufficient for rates, so further restrictions on the class  $\mathcal{H}$  are required.

### 3.6.1 Donsker Classes

Let  $\mathcal{G}$  be a class of uniformly bounded, measurable functions from  $\mathbb{U}$  into  $\mathbb{R}$ , and let  $(b\mathcal{G}, \|\cdot\|)$  be the Banach space of bounded, real-valued functionals on  $\mathcal{G}$  with the supremum norm. The class  $\mathcal{G}$  is called  $\phi$ -Donsker if

$$v_n(g) := \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n g(U_i) - \int g d\phi \right\}$$

converges in distribution to a tight Gaussian process  $\nu$  in the space  $b\mathcal{G}$  (cf., e.g., van der Vaart, 1998, p. 269). Here  $\omega \mapsto v_n(\cdot)(\omega)$  and  $\omega \mapsto$

$\nu(\cdot)(\omega)$  are maps from  $\Omega$  into  $b\mathcal{G}$ . The maps  $\omega \mapsto \nu_n(\cdot)(\omega)$  are not necessarily measurable, and convergence in distribution is to be understood in the sense of  $\nu_n \xrightarrow{d^*} \nu$ .

**Proposition 3.6.1.** *If  $\mathcal{H}$  defined in (3.9) is  $\phi$ -Donsker, then  $\|V_{AR_n} - V_{AT}\| = O_{P^*}(n^{-1/2})$ .*

In essence, Proposition 3.6.1 tells us that the  $\sqrt{n}$  rate can be obtained if the SDP has enough structure for the function class (3.9) to have the  $\phi$ -Donsker property. There are several well-known sets of sufficient conditions for a function class to be  $\phi$ -Donsker. Below, we use two of these sets to obtain rates of convergence in important but relatively specialized settings.

### 3.6.2 The Lipschitz Case

Our first result is based on a Lipschitz condition. To apply the method, we add the following assumptions:

- (i)  $\mathbb{G} \subset \mathbb{R}^d$ .
- (ii)  $Aw$  is Lipschitz continuous for every  $w \in \mathcal{C}(\mathbb{X})$ .<sup>11</sup>
- (iii) There exists a measurable function  $m_0: \mathbb{U} \rightarrow \mathbb{R}$  with  $\int m_0^2 d\phi < \infty$  and<sup>12</sup>

$$\|F(y, u) - F(y', u)\|_2 \leq m_0(u) \|y - y'\|_2 \quad \forall y, y' \in \mathbb{G}, u \in \mathbb{U}. \quad (3.10)$$

Notice that the assumptions concern only the transition function, not the reward function. Many dynamic macroeconomic models have Lipschitz transition rules. The consumer's problem in the incomplete markets models of Aiyagari (1994) and Huggett (1997) are obvious examples, and many recent variations have a similar structure (see, e.g., Pijoan-Mas, 2006, or Ábrahám and Cárceles-Poveda, 2010).

<sup>11</sup>This condition depends on the approximation architecture used in the fitted VFI routine, and is satisfied by, for example, the piecewise linear interpolation operator in Example 3.4.1.

<sup>12</sup>Here  $\|\cdot\|_2$  represents the Euclidean norm on  $\mathbb{R}^d$ .

**Proposition 3.6.2.** *If (i)–(ii) hold, then  $\|V_{AR_n} - V_{AT}\| = O_{P^*}(n^{-1/2})$ .*

An important special case of our Lipschitz assumption is models with linear transition rules. The next lemma provides details.

**Lemma 3.6.1.** *If  $\mathbb{U} \subset \mathbb{R}^k$ , and  $F$  is linear, in the sense that*

$$F(x, a, u) = Ax + Ba + Cu \quad (x \in \mathbb{X}, a \in \Gamma(x), u \in \mathbb{U}) \quad (3.11)$$

*for matrices  $A$ ,  $B$  and  $C$ , then Assumption (ii) is satisfied.*

### 3.6.3 The Monotone Case

Another way to establish the  $\phi$ -Donsker property is via monotonicity. To this end, we drop the Lipschitz assumptions of Section 3.6.2 and replace them with the following:

- (i)  $\mathbb{X} \subset \mathbb{R}^d$  and  $\mathbb{U} \subset \mathbb{R}$ .
- (ii)  $A$  maps  $i\mathcal{C}(\mathbb{X})$  to itself, where  $i\mathcal{C}(\mathbb{X})$  is the increasing functions in  $\mathcal{C}(\mathbb{X})$ .
- (iii) For all  $x, x' \in \mathbb{X}$  with  $x \leq x'$ ,  $\Gamma(x) \subset \Gamma(x')$ ;  $r(x, a) \leq r(x', a)$  for all  $a \in \Gamma(x)$ ; and  $F(x, a, u) \leq F(x', a, u)$  for all  $a \in \Gamma(x)$  and  $u \in \mathbb{U}$ .
- (iv) For all  $y \in \mathbb{G}$ ,  $F(y, u) \leq F(y, u')$  whenever  $u \leq u'$ .

Assumption (ii) depends on the approximation architecture, and is satisfied by, for example, the linear interpolation operator in Example 3.4.1. The other assumptions are satisfied by a number of standard models.

**Proposition 3.6.3.** *If (i)–(iv) hold, then  $\|V_{AR_n} - V_{AT}\| = O_{P^*}(n^{-1/2})$ .*

**Example 3.6.1.** Consider the growth model

$$\begin{aligned} \max \mathbb{E} \sum_{t=0}^{\infty} \beta^t w(c_t) \quad \text{subject to} \\ c_t \geq 0, \quad k_{t+1} \geq 0, \quad c_t + k_{t+1} \leq z_t f(k_t). \end{aligned}$$

Suppose that  $z_t$  is Markov, following transition rule  $z_{t+1} = g(z_t, U_{t+1})$ , where  $(U_t)_{t \geq 1}$  is IID. The state is  $(k, z) \in \mathbb{R}_+^2$ . To write the model in our framework, we take  $F(k, z, k', u) = (k', g(z, u))$ ,  $r(k, z, k') = w(zf(k) - k')$  and  $\Gamma(k, z) = [0, zf(k)]$ . If  $f$  and  $g$  are both increasing, then Assumptions (iii) and (iv) above are satisfied.

### 3.7 Applications

In this section we consider two numerical applications. In the first application, we revisit the instability illustrated in Figure 3.3, and re-run the experiment using Algorithm 1 and a variety of approximation operators that conform to our assumptions. As in Figure 3.3, we use a well-known special case of the growth model in Example 3.6.1, with  $w(c) = \ln c$  and  $f(k) = k^\alpha$ . The shock is IID and lognormal, with parameters  $\mu = 0$  and  $\sigma = 0.25$ . The remaining parameters are  $\beta = 0.95$  and  $\alpha = 0.33$ . The natural state space  $(0, \infty)$  is truncated to the interval  $\mathbb{X} = [10^{-5}, 1]$ . In performing value function iteration, we use 150 grid points. All of these settings are identical to Figure 3.3.

Regarding the details of the algorithm, since the shock is IID, we can reduce the two-dimensional state space in Example 3.6.1 down to one dimension by setting  $x := uf(k)$ . Here  $u$  is the current shock,  $x$  represents output, and the model is mapped into our set up via  $F(x, k, u) = uk^\alpha$ ,  $r(x, k) = \ln(x - k)$  and  $\Gamma(x) = [0, x]$ . The random Bellman operator has the form

$$R_n v(x) := \max_{0 \leq k \leq x} \left\{ \ln(x - k) + \beta \frac{1}{n} \sum_{i=1}^n v(U_i k^\alpha) \right\} \quad (x \in \mathbb{X}), \quad (3.12)$$

where the shocks  $U_1, \dots, U_n$  are IID draws from the lognormal density specified above. In all cases we take  $n = 100$ . For the approximation operator  $A$ , we use several nonexpansive operators: Continuous piecewise linear interpolation (pwise-linear),  $k$ -nearest neighbors with  $k = 1$  (nearest nbr) and the kernel smoother in (3.1) with  $h = 0.25$  (ksmooth 1),  $h = 0.5$  (ksmooth 2) and  $h = 0.75$  (ksmooth 3). Algorithm 1 is applied to these specifications, with initial condition  $v = w$  (i.e., starting at the utility function).

The columns from pwise-linear to ksmooth 3 in Table 3.1 show the sup-norm distance  $\|(AR_n)^j w - (AR_n)^{j-1} w\|$  between successive iterates of  $AR_n$ , with each column corresponding to a different approximation method. In all cases the distance is monotonically decreasing, as implied by the theory. (For pwise-linear, the first 25 iterates  $(AR_n)^k w$  are themselves plotted in Figure 3.4. The dashed line is the true value function.) For comparison purposes, the last column (chebyshev) of Table 3.1



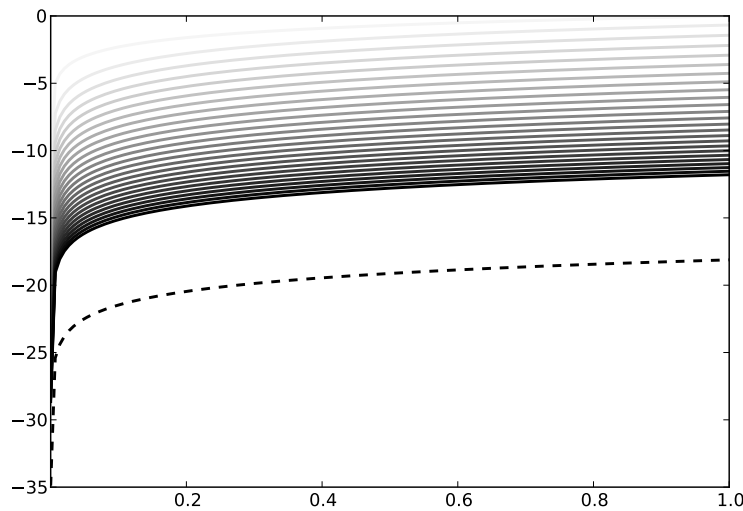


Figure 3.4: 25 elements of the sequence  $(AR_n)^{kw}$

gives distances between iterates for the Chebyshev–quadrature method with Chebyshev polynomials of order 10. As illustrated previously in Figure 3.3, the sequence of iterates diverges.

As a second numerical application, we consider the stochastic LQ problem, where

$$r(x, a) = -x'Rx - 2a'Hx - a'Qa, \quad F(x, a, u) = Mx + Na + u \quad (3.13)$$

and  $a$  is unconstrained. Here  $x$  is a  $k$ -vector,  $a$  is a  $j$ -vector,  $R$  is  $k \times k$  and positive definite,  $Q$  is  $j \times j$  and positive definite,  $H$  is  $j \times k$ ,  $M$  is  $k \times k$  and  $N$  is  $k \times j$ . As in the previous application, the LQ problem has an analytical solution against which the results of our algorithm can be compared.<sup>13</sup> We wish to investigate whether our algorithm produces accurate output with a relatively small grid and low number of iterations. (In other words, we wish to investigate the small sample properties, given that our theory already covers the asymptotics.) To be specific, we focus

<sup>13</sup>More correctly, the LQ problem can be solved by iterating on the Riccati equation, a solution method which combines analytical and numerical elements. However, the method is accurate and stable, and hence we refer to the output of the method as the “exact” or “analytical” solution.

iterate	pwise-linear	nearest nbr	ksmooth 1	ksmooth 2	ksmooth 3	chebyshev
1	4.320550	4.320550	7.429569	8.456384	8.921369	2.476053
2	1.994024	1.733695	1.122005	1.153222	1.200513	2.624954
3	1.158638	1.113960	0.975892	1.076373	1.062914	3.449942
4	0.875121	0.856361	0.885546	1.015897	1.000330	5.212505
5	0.760378	0.750628	0.827461	0.963558	0.949328	7.880053
6	0.700180	0.691972	0.781766	0.915041	0.901760	11.910691
7	0.658340	0.650938	0.741351	0.869216	0.856661	18.007254
8	0.623338	0.616740	0.703879	0.825739	0.813827	27.227489
9	0.591538	0.586429	0.668562	0.784449	0.773136	41.169071
10	0.561770	0.557302	0.635096	0.745225	0.734479	62.249318
11	0.533623	0.528925	0.603330	0.707964	0.697755	94.123510
12	0.506925	0.503980	0.573160	0.672566	0.662867	142.318592
13	0.481573	0.477532	0.544501	0.638937	0.629724	215.191524
14	0.457493	0.454445	0.517276	0.606991	0.598238	325.378374
15	0.434618	0.431318	0.491412	0.576641	0.568326	491.985391
16	0.412887	0.409204	0.466841	0.547809	0.539909	743.902007
17	0.392243	0.390251	0.443499	0.520419	0.512914	1124.810219
18	0.372630	0.369872	0.421324	0.494398	0.487268	1700.758995
19	0.353999	0.351188	0.400258	0.469678	0.462905	2571.617068
20	0.336299	0.334070	0.380245	0.446194	0.439760	3888.390044
21	0.319484	0.316834	0.361233	0.423884	0.417772	5879.404569
22	0.303510	0.301357	0.343171	0.402690	0.396883	8889.899853
23	0.288334	0.286551	0.326013	0.382555	0.377039	13441.891687
24	0.273918	0.272124	0.309712	0.363428	0.358187	20324.689266
25	0.260222	0.258491	0.294226	0.345256	0.340278	30731.760334
26	0.247211	0.245460	0.279515	0.327993	0.323264	46467.676868
27	0.234850	0.233864	0.265539	0.311594	0.307101	70261.025402
28	0.223108	0.221631	0.252262	0.296014	0.291746	106237.540229
29	0.211952	0.210856	0.239649	0.281213	0.277158	160635.500111
30	0.201355	0.201770	0.227667	0.267153	0.263300	242887.437345
31	0.191287	0.191375	0.216283	0.253795	0.250135	367255.726034
32	0.181723	0.181114	0.205469	0.241105	0.237629	555305.658369
33	0.172636	0.171922	0.195196	0.229050	0.225747	839644.836982
34	0.164005	0.163115	0.185436	0.217598	0.214460	1269577.288914
35	0.155804	0.154937	0.176164	0.206718	0.203737	1919652.716880
36	0.148014	0.147639	0.167356	0.196382	0.193550	2902593.316376
37	0.140613	0.140678	0.158988	0.186563	0.183872	4388839.651145
38	0.133583	0.133131	0.151039	0.177235	0.174679	6636104.815231
39	0.126904	0.126643	0.143487	0.168373	0.165945	10034061.533153
40	0.120558	0.120358	0.136313	0.159954	0.157648	15171910.880616

Table 3.1: Distance between successive iterates of fitted value iteration

$\beta$		0.9	0.925	0.95	0.975
Var( $w_t$ )	analytical	0.00294547	0.00294097	0.00293646	0.00293194
	approximate	0.00294443	0.00293989	0.00293534	0.00293078
Cov( $w_t, y_t$ )	analytical	0.00230056	0.00229881	0.00229704	0.00229528
	approximate	0.00230019	0.00229842	0.00229664	0.00229486

Table 3.2: Approximate and analytical second moments after 5 iterations

on the quadratic problem

$$\begin{aligned}
\max \mathbf{E} \sum_{t=0}^{\infty} \beta^t \{-(c_t - b)^2 - \gamma i_t^2\} \quad & \text{subject to} \\
w_{t+1} &= \delta w_t + i_t \\
c_t + i_t &= r w_t + y_t \\
y_{t+1} &= (1 - \rho) \bar{y} + \rho y_t + \varepsilon_{t+1}.
\end{aligned}$$

Here  $w$  denotes the stock of a commodity,  $y$  is exogenous supply of the commodity,  $c$  is consumption,  $i$  is investment,  $r$  is the net interest rate per unit of the commodity (which can be borrowed for short sales and other purposes),  $\gamma$  and  $b$  are preference parameters,  $\delta$  parameterizes depreciation,  $\rho$  is the correlation coefficient of  $\{y_t\}$ ,  $\bar{y}$  is the unconditional mean, and  $\{\varepsilon_t\}_{t \geq 0}$  is an IID shock with distribution  $N(0, \sigma^2)$ . The problem maps into the LQ setting (3.13) with state  $x = (w, y)$  and control  $a = i$ .

In our experiment, we assess the accuracy of the solution computed via Algorithm 1 by comparing it against the analytical solution. To provide a comparison with practical implications, we compare the second moments of the state variables under the approximate and analytical optimal policies.<sup>14</sup> The two (centered) moments we consider are Var( $w_t$ ) and Cov( $w_t, y_t$ ). The approximation scheme is the radial basis kernel smoother (see Section 3.1.1) with  $h = 0.00125$ .

Table 3.2 shows the results of the simulation for 4 different values of the discount parameter  $\beta$ .<sup>15</sup> In all cases, the approximate policies are computed from 5 iterations of the fitted random Bellman operator  $AR_n$

<sup>14</sup>Moments are calculated by simulating time series of length 5,000 from the approximate and analytical policies respectively, and taking sample averages over time.

<sup>15</sup>The values of the other parameters are  $b = 30$ ,  $\gamma = 1.1$ ,  $r = 1/\beta - 1$ ,  $\rho = 0.4$ ,  $\bar{y} = 29$ ,  $\delta = 0.9$  and  $\sigma = 0.1$ .

(i.e., by setting  $k = 5$  in Algorithm 1). The initial condition from which iteration begins is the constant zero function. The value of  $n$  is 100, and the number of grid points is 400 (the cartesian product of 20 grid points in each of the two dimensions). Despite the coarseness of the grid and the low number of iterations, the moments produced by the fitted value function iteration algorithm with the kernel smoother scheme are close approximations of the true values.<sup>16</sup>

### 3.8 Conclusion

We studied a Monte Carlo VFI algorithm with function approximation. We proved that the algorithm is consistent for a wide variety of models. This guaranteed convergence stands in contrast to many other numerical techniques proposed in the literature. Under additional restrictions, we established a parametric rate of convergence, independent of the dimension of the state, action and shock spaces.

Many avenues for future research exist. First, we identified only two cases where the  $\phi$ -Donsker property is satisfied (the Lipschitz and monotonicity conditions of Sections 3.6.2 and 3.6.3). Additional research should illuminate other cases. In addition, we treated only stationary, additively separable, infinite horizon SDPs, leaving open the cases of non-stationary models, optimal stopping, and general recursive utility. All of these issues are left for future study.

### 3.9 Proofs

*Proof of Lemma 3.5.1.* The contractiveness of  $AT$  follows from Lemma 3.2.1. Next we consider contractiveness of  $R_n$ . Fix  $n \in \mathbb{N}$  and  $\omega \in \Omega$ . Let  $R := R_n(\omega)$ . Fix  $w, w' \in \mathcal{C}(\mathbb{X})$  and  $x \in \mathbb{X}$ . In view of (3.15), we have

$$|Rw(x) - Rw'(x)| \leq \beta \max_{a \in \Gamma(x)} \left| \frac{1}{n} \sum_{i=1}^n w[F(x, a, U_i(\omega))] - \frac{1}{n} \sum_{i=1}^n w'[F(x, a, U_i(\omega))] \right|.$$

---

<sup>16</sup>Code for our experiments can be found at <https://sites.google.com/site/fviprobone/>.

Using the triangle inequality and the definition of  $\|\cdot\|$ , we obtain

$$|Rw(x) - Rw'(x)| \leq \beta \|w - w'\|.$$

Taking the supremum over  $x \in \mathbb{X}$  yields the desired result.

Finally, contractiveness of  $AR_n$  now follows from Lemma 3.2.1.  $\square$

*Proof of Lemma 3.5.2.* Fix  $\epsilon > 0$ . By Lemma 3.2.1, we have

$$\|V_T - V_{AT}\| \leq (1 - \beta)^{-1} \|AV_T - V_T\|. \quad (3.14)$$

Since  $\mathbb{X}$  is compact,  $V_T$  is uniformly continuous, and we select  $\delta > 0$  with  $|V_T(x) - V_T(y)| < (1 - \beta)\epsilon$  whenever  $d(x, y) < \delta$ . Using compactness again, we choose  $\{x_i\}_{i=1}^m$  such that, given any  $x \in \mathbb{X}$ , there exists at least one  $x_i$  with  $d(x, x_i) < \delta$ . Finally, we choose  $\psi$  such that  $\psi(u) = 0$  whenever  $u$  is greater than some constant  $M$ ,<sup>17</sup> and  $h$  such that  $Mh < \delta$ . Now fix any  $x \in \mathbb{X}$ . Letting  $\lambda(x, i) := \psi[d(x, x_i)/h] / \sum_j \psi[d(x, x_j)/h]$ , we can write

$$|AV_T(x) - V_T(x)| = \left| \sum_i \lambda(x, i) V_T(x_i) - V_T(x) \right| \leq \sum_i \lambda(x, i) |V_T(x_i) - V_T(x)|.$$

If  $d(x, x_i) \geq \delta$ , then  $d(x, x_i)/h \geq M$ , and hence  $\psi[d(x, x_i)/h] = \lambda(x, i) = 0$ . For the remaining terms in the sum we have  $d(x, x_i) < \delta$ , and hence  $|V_T(x_i) - V_T(x)| < (1 - \beta)\epsilon$ . Since  $x$  is arbitrary, we have  $\|AV_T - V_T\| < (1 - \beta)\epsilon$ . Combining this bound with (3.14) completes the proof of the lemma.  $\square$

*Proof of Theorem 3.5.1.* By Lemma 3.2.1 and the nonexpansiveness of  $A$ , we have

$$\begin{aligned} \|V_{AR_n} - V_{AT}\| &\leq \frac{1}{1 - \beta} \|AR_n V_{AT} - V_{AT}\| \\ &= \frac{1}{1 - \beta} \|AR_n V_{AT} - AT V_{AT}\| \\ &\leq \frac{1}{1 - \beta} \|R_n V_{AT} - TV_{AT}\|. \end{aligned}$$

<sup>17</sup>A typical example is the Epanechnikov kernel.

Hence, to prove Theorem 3.5.1, it is sufficient to prove that  $\|R_n V_{AT} - TV_{AT}\|$  converges to zero with probability one. To bound this term, we make use of the following standard inequality: If  $g, g' \in \mathcal{C}(\mathbb{Y})$  for compact set  $\mathbb{Y}$ , then

$$|\max g - \max g'| \leq \max |g - g'| =: \|g - g'\|. \quad (3.15)$$

Using (3.15), we obtain

$$\begin{aligned} & |R_n V_{AT}(x) - TV_{AT}(x)| \\ & \leq \beta \max_{a \in \Gamma(x)} \left| \frac{1}{n} \sum_{i=1}^n V_{AT}[F(x, a, U_i)] - \int V_{AT}[F(x, a, u)] \phi(du) \right|, \end{aligned}$$

where  $x \in \mathbb{X}$  is arbitrary. Taking the supremum over all  $x \in \mathbb{X}$ , we now have

$$\begin{aligned} & \|R_n V_{AT} - TV_{AT}\| \\ & \leq \beta \max_{(x,a) \in \mathbb{G}} \left| \frac{1}{n} \sum_{i=1}^n V_{AT}[F(x, a, U_i)] - \int V_{AT}[F(x, a, u)] \phi(du) \right|. \quad (3.16) \end{aligned}$$

Let  $y = (x, a)$  denote a typical element of  $\mathbb{G}$ , and let

$$h_y(u) := h_{(x,a)}(u) := V_{AT}[F(x, a, u)] := V_{AT}[F(y, u)]. \quad (3.17)$$

Also, for  $h: \mathbb{U} \rightarrow \mathbb{R}$ , let  $\phi_n(h) := \frac{1}{n} \sum_{i=1}^n h(U_i)$  and  $\phi(h) := \int h d\phi$ . Using this notation, (3.16) becomes

$$\|R_n V_{AT} - TV_{AT}\| \leq \beta \max_{y \in \mathbb{G}} |\phi_n(h_y) - \phi(h_y)|. \quad (3.18)$$

A class  $\mathcal{H}$  of bounded measurable functions mapping  $\mathbb{U}$  into  $\mathbb{R}$  is called  $\phi$ -Glivenko-Cantelli if  $\sup_{h \in \mathcal{H}} |\phi_n(h) - \phi(h)| \rightarrow 0$   $\mathbf{P}^*$ -almost surely as  $n \rightarrow \infty$ . A sufficient condition for this property<sup>18</sup> is that  $\mathcal{H}$  consists of functions  $h_\alpha: \mathbb{U} \rightarrow \mathbb{R}$  with index  $\alpha$  in metric space  $\Lambda$ , and, moreover:

1.  $\Lambda$  is compact;
2.  $\Lambda \ni \alpha \mapsto h_\alpha(u) \in \mathbb{R}$  is continuous for every  $u \in \mathbb{U}$ ; and

<sup>18</sup>See, for example, van der Vaart, 1998, p. 272.

3. there exists a measurable function  $H: \mathbb{U} \rightarrow \mathbb{R}$  such that  $\int H d\phi < \infty$  and  $|h_\alpha| \leq H$  for every  $\alpha \in \Lambda$ .

In our case, the relevant class of functions is  $\{h_y\}_{y \in \mathbb{G}}$ , where  $h_y$  is defined in (3.17). This family of functions satisfies the sufficient conditions in 1–3 above. First,  $\mathbb{G}$  is a compact metric space, due to our assumptions on  $\mathbb{X}$ ,  $\mathbb{A}$  and  $\Gamma$ . Second,  $\mathbb{G} \ni y \mapsto h_y(u) := V_{AT}[F(y, u)] \in \mathbb{R}$  is continuous for every  $u \in \mathbb{U}$ , due to continuity of  $V_{AT}$  and  $F$ . Third,  $|h_y(u)|$  is bounded above by the finite constant  $\|V_{AT}\|$  for all  $y \in \mathbb{G}$  and  $u \in \mathbb{U}$ . Hence,  $\{h_y\}_{y \in \mathbb{G}}$  is  $\phi$ -Glivenko-Cantelli. This concludes the proof.  $\square$

*Proof of Proposition 3.6.1.* We need some preliminary results and additional notation. Throughout the proof,  $\mathcal{H} = \{h_y\}_{y \in \mathbb{G}}$  is the function class defined in both (3.9) and (3.17). In addition, let

$$G_n(y) := v_n(h_y) := \sqrt{n}(\phi_n(h_y) - \phi(h_y)) \quad (n \in \mathbb{N}, y \in \mathbb{G}).$$

$G_n$  can be understood as a real-valued stochastic process indexed by  $y \in \mathbb{G}$ :

$$G_n(y)(\omega) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n h_y(U_i(\omega)) - \int h_y(u) \phi(du) \right) \in \mathbb{R}.$$

Regarding measurability, we have the following result, which is proved immediately below the current proof:

**Lemma 3.9.1.** *For each  $n \in \mathbb{N}$ , the following measurability results hold:*

1.  $\omega \mapsto G_n(\cdot)(\omega)$  is a  $\mathcal{C}(\mathbb{G})$ -valued random variable, and
2.  $\omega \mapsto \|G_n(\cdot)(\omega)\| = \sup_{y \in \mathbb{G}} |G_n(y)(\omega)|$  is a real-valued random variable.

In view of (3.18), we have

$$\|V_{AR_n} - V_{AT}\| \leq \frac{\beta}{1 - \beta} n^{-1/2} \sup_{y \in \mathbb{G}} |G_n(y)|.$$

Since  $\mathcal{H} := \{h_y\}_{y \in \mathbb{G}}$  is  $\phi$ -Donsker, we have  $\nu_n \xrightarrow{d^*} \nu$ , where  $\nu$  is a Gaussian process on  $\mathcal{H}$ . By Lemma 3.2.2 and continuity of the norm  $\|\cdot\|$  on  $b\mathcal{H}$ , we then have  $\|\nu_n\| \xrightarrow{d^*} \|\nu\|$  in  $\mathbb{R}$ . Observe that

$$\|\nu_n\| = \sup_{h \in \mathcal{H}} |\nu_n(h)| = \sup_{y \in \mathbb{G}} |\nu_n(h_y)| = \sup_{y \in \mathbb{G}} |G_n(y)|,$$

and hence  $\sup_{y \in \mathbb{G}} |G_n(y)| \xrightarrow{d^*} \|\nu\|$ . By part 2 of Lemma 3.9.1, this is convergence in distribution in the regular sense, and, as a consequence, we have  $\sup_{y \in \mathbb{G}} |G_n(y)| = O_P(1)$ . Therefore

$$\|V_{AR_n} - V_{AT}\| \leq \frac{\beta}{1 - \beta} n^{-1/2} O_P(1) = O_P(n^{-1/2}).$$

This concludes the proof of Proposition 3.6.1. □

*Proof of Lemma 3.9.1.* We begin by proving measurability of  $\omega \mapsto H(\cdot)(\omega)$ , where

$$H(y)(\omega) = h_y(U(\omega)) = V_{AT}[F(y, U(\omega))].$$

Since  $\mathbb{G}$  is compact in the product topology, the Stone–Weierstrass theorem implies that  $\mathcal{C}(\mathbb{G})$  is separable. Hence, by the Pettis measurability theorem, we need only show that  $\omega \mapsto \ell(H(\cdot)(\omega))$  is measurable for each  $\ell$  in the dual space  $\mathcal{C}(\mathbb{G})^*$  of  $\mathcal{C}(\mathbb{G})$ . By the Riesz representation theorem,  $\mathcal{C}(\mathbb{G})^*$  can be identified with  $\mathcal{M}(\mathbb{G})$ , the space of finite signed Borel measures on  $\mathbb{G}$ . Thus, it remains to show that

$$\Omega \ni \omega \mapsto \int H(y)(\omega) \gamma(dy) \in \mathbb{R} \text{ is measurable} \quad \forall \gamma \in \mathcal{M}(\mathbb{G})$$

To this end it is sufficient to show that  $H(y)(\omega) = V_{AT}[F(y, U(\omega))]$  is measurable with respect to the product  $\sigma$ -algebra  $\mathcal{B}_{\mathbb{G}} \otimes \mathcal{F}$ , where  $\mathcal{B}_{\mathbb{G}}$  is the Borel  $\sigma$ -algebra on  $\mathbb{G}$ . Since  $H$  is continuous with respect to  $y$  and measurable with respect to  $\omega$ ,  $H$  is a Carathéodory function (Aliprantis and Border, 2006, Definition 4.50). As  $\mathbb{G}$  is separable, measurability with respect to  $\mathcal{B}_{\mathbb{G}} \otimes \mathcal{F}$  is established (Aliprantis and Border, 2006, Lemma 4.51).



Given measurability of  $\omega \mapsto H(\cdot)(\omega)$ , measurability of  $\omega \mapsto G_n(\cdot)(\omega)$  follows from the fact that linear combinations of measurable random elements of a separable Banach space are themselves measurable.

Regarding the second claim in the lemma, measurability of  $\omega \mapsto \|G_n(\cdot)(\omega)\|$  follows from measurability of  $\omega \mapsto G_n(\cdot)(\omega)$ , continuity of the norm as a map from  $\mathcal{C}(\mathbb{G})$  to  $\mathbb{R}$ , and the fact that continuous transformations of measurable mappings are measurable.  $\square$

*Proof of Proposition 3.6.2.* By Proposition 3.6.1, it suffices to show that the class  $\{h_y\}_{y \in \mathbb{G}}$  is  $\phi$ -Donsker when (i)–(iii) hold. A sufficient condition for  $\{h_y\}_{y \in \mathbb{G}}$  to be  $\phi$ -Donsker is the existence of a measurable function  $m: \mathbb{U} \rightarrow \mathbb{R}$  such that  $\int m^2 d\phi < \infty$  and

$$|h_y(u) - h_{y'}(u)| \leq m(u) \|y - y'\|_2 \quad \forall y, y' \in \mathbb{G}, u \in \mathbb{U} \quad (3.19)$$

(see, e.g., van der Vaart, 1998, p. 271). To find such an  $m$ , observe that  $V_{AT}$  is Lipschitz, as follows from (ii) and the relation  $V_{AT} = ATV_{AT}$ . As a consequence, there exists a  $K < \infty$  such that, for any  $y, y' \in \mathbb{G}$  and  $u \in \mathbb{U}$ , we have

$$\begin{aligned} |h_y(u) - h_{y'}(u)| &:= |V_{AT}[F(y, u)] - V_{AT}[F(y', u)]| \\ &\leq K \|F(y, u) - F(y', u)\|_2 \leq Km_0(u) \|y - y'\|_2, \end{aligned}$$

where  $m_0$  is the function in (iii). Letting  $m := Km_0$ , we see that  $\int m^2 d\phi = K^2 \int m_0^2 d\phi < \infty$ . All the conditions are now verified, and hence  $\{h_y\}_{y \in \mathbb{G}}$  is  $\phi$ -Donsker.  $\square$

*Proof of Lemma 3.6.1.* To see this, observe that for any  $y = (x, a) \in \mathbb{G}$ ,  $y' = (x', a') \in \mathbb{G}$ , and  $u \in \mathbb{U}$ ,

$$\begin{aligned} \|Ax + Ba + Cu - Ax' - Ba' - Cu\|_2 \\ = \|A(x - x') + B(a - a')\|_2 \leq \gamma(\|x - x'\|_2 + \|a - a'\|_2), \end{aligned}$$

where  $\gamma$  is the maximum of the operator norms of  $A$  and  $B$ . Since  $y = (x, a) \mapsto \|x\|_2 + \|a\|_2 \in \mathbb{R}$  defines a norm on  $\mathbb{R}^d$ , and since all norms on  $\mathbb{R}^d$  are equivalent, we obtain

$$\|F(y, u) - F(y', u)\|_2 \leq M\gamma \|y - y'\|_2 \quad \forall y, y' \in \mathbb{G}, u \in \mathbb{U}$$

for some  $M < \infty$ . This verifies (ii).  $\square$

*Proof of Proposition 3.6.3.* From van der Vaart (1998, p. 273), it suffices to show that the class  $\{h_y\}_{y \in \mathbb{G}}$  is uniformly bounded on  $\mathbb{U}$ , and that each element  $h_y$  is monotone increasing on  $\mathbb{U}$ . Since  $h_y(u) = V_{AT}[F(y, u)]$ , uniform boundedness will hold if  $V_{AT}$  is bounded on  $\mathbb{X}$ . That this is the case follows from the fact that  $\mathbb{X}$  is compact and  $V_{AT} \in \mathcal{C}(\mathbb{X})$ .

Regarding monotonicity, we begin by showing that  $V_{AT}$  is monotone increasing. To see that this is the case, observe that  $V_{AT}$  is the fixed point of  $AT$  in  $\mathcal{C}(\mathbb{X})$ . Since  $i\mathcal{C}(\mathbb{X})$  is a closed subset of  $\mathcal{C}(\mathbb{X})$ , we need only show that  $AT$  maps  $i\mathcal{C}(\mathbb{X})$  into itself. Since  $A: i\mathcal{C}(\mathbb{X}) \rightarrow i\mathcal{C}(\mathbb{X})$  by assumption, it remains to verify that  $T$  also has this property. For a proof of this fact, see Stachurski (2009, Theorem 12.1.2). As a result,  $V_{AT}$  is increasing, and the claim in the proposition now follows from Assumption (iv) above.  $\square$

# Bibliography

- Ábrahám, Árpád and Eva Cárceles-Poveda**, “Endogenous trading constraints with incomplete asset markets,” *Journal of Economic Theory*, 2010, 145 (3), 974–1004.
- Aiyagari, S. Rao**, “Uninsured Idiosyncratic Risk and Aggregate Saving,” *The Quarterly Journal of Economics*, 1994, 109 (3), 659–684.
- Alaoui, Larbi and Fabrizio Germano**, “Time Scarcity and the Market for News,” 2016.
- Aliprantis, Charalambos. D. and Kim. C. Border**, *Infinite Dimensional Analysis*, 3rd ed., New York: Springer-Verlag, 2006.
- Aruoba, S. Boraan, Jesús Fernández-Villaverde, and Juan F. Rubio-Ramírez**, “Comparing solution methods for dynamic equilibrium economies,” *Journal of Economic Dynamics and Control*, 2006, 30 (12), 2477–2508.
- Athey, Susan, Markus Mobius, and Jenő Pál**, “The Impact of Aggregators on Internet News Consumption,” 2017.
- Baird, Leemon**, “Residual Algorithms: Reinforcement Learning with Function Approximation,” in “Machine Learning: Proceedings of the Twelfth International Conference on Machine Learning,” Morgan Kaufman, 1995, pp. 30–37.
- Barthel, Michale**, *Newspapers: Fact Sheet*, Pew Research Center, June 15, 2016; accessed November 17, 2016. <http://www.journalism.org/2016/06/15/newspapers-fact-sheet/>.
- Bellman, Richard, Robert Kalaba, and Bella Kotkin**, “Polynomial Approximation: A New Computational Technique in Dynamic Programming,” *Mathematics of Computation*, 1963, 17 (82), 155–161.

**Berger, Jonah and Katherine L. Milkman**, "What Makes Online Content Viral?," *Journal of Marketing Research*, 2012, 49 (2), 192–205.

**Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre**, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, 2008, 2008 (10), P10008.

**Boyan, Justin A. and Andrew W. Moore**, "Generalization in Reinforcement Learning: Safely Approximating the Value Function," in Gerald Tesauro, David S. Touretzky, and Todd K. Leen, eds., *Advances in Neural Information Processing Systems 7*, 1995, pp. 369–376.

**Bucher, Hans-Jürgen and Peter Schumacher**, "The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print and online media," *Communications*, 2006, 31 (3).

**Calzada, Joan and Ricard Gil**, "What do News Aggregators Do? Evidence from Google News in Spain and Germany," September 2016. SSRN working paper, available at <https://ssrn.com/abstract=2837553>.

**Cameron, A. Colin and Pravin K. Trivedi**, *Microeconometric: Methods and Applications*, Cambridge University Press, May 2005.

**Chiou, Lesley and Catherine Tucker**, "Content aggregation by platforms: The case of the news media," 2015. NBER working paper 21404.

**Clementi, Gian Luca and Hugo A. Hopenhayn**, "A Theory of Financing Constraints and Firm Dynamics," *The Quarterly Journal of Economics*, 2006, 121 (1), 229–265.

**Costanza-Chock, Sasha and Pablo Rey-Mazón**, "PageOneX : New Approaches to Newspaper Front Page Analysis," *International Journal of Communication*, 2016, 10, 2318–2345.

**Dellarocas, Chrysanthos, Juliana Sutanto, Mihai Calin, and Elia Palme**, "Attention Allocation in Information-Rich Environments: The Case of News Aggregators," *Management Science*, sep 2016, 62 (9), 2543–2562.

**Drummond, Chris**, “Preventing Overshoot of Splines with Application to Reinforcement Learning,” 1996.

**Fedyk, Anastassia**, “Front Page News: The Effect of News Positioning on Financial Markets,” 2017.

**Filloux, Frederic**, *News Publishers’ Facebook Problem*, Monday Note, July 27, 2016; accessed November 17, 2016. <https://mondaynote.com/news-publishers-facebook-problem-6752f1c35037>.

**Fukushima, Kenichi and Yuichiro Waki**, “A polyhedral approximation approach to concave numerical dynamic programming,” *Journal of Economic Dynamics and Control*, 2013, 37 (11), 2322–2335.

**Gentzkow, Matthew**, “Polarization in 2016,” *Toulouse Network for Information Technology Whitepaper*, 2016.

**Gordon, Geoffrey J.**, “Stable Function Approximation in Dynamic Programming,” in “Proceedings of the 12th International Conference on Machine Learning” 1995.

**Guestrin, Carlos, Daphne Koller, and Ronald Parr**, “Max-Norm Projections for Factored MDPs,” in “International Joint Conference on Artificial Intelligence Vol 1.” 2001, pp. 673–680.

**Hensinger, Elena, Ilias Flaounas, and Nello Cristianini**, “Modelling and predicting news popularity,” *Pattern Analysis and Applications*, 2013, 16 (4), 623–635.

**Huggett, Mark**, “The one-sector growth model with idiosyncratic shocks: Steady states and dynamics,” *Journal of Monetary Economics*, 1997, 39 (3), 385–403.

**Jafri, Salma**, *How to Write Headlines Google Will Love & You and I Will Click, Read, and Share*, Search Engine Watch, January 27, 2014; accessed November 17, 2016. <https://searchenginewatch.com/sew/how-to/2325076/how-to-write-headlines-google-will-love-you-and-i-will-click-read-and-share>.

**Johnson, Steven G.**, “The NLOpt nonlinear-optimization package.”

**Judd, Kenneth L.**, *Numerical Methods in Economics*, MIT Press, 1998.

- **and Andrew Solnick**, “Numerical Dynamic Programming with Shape-Preserving Splines,” 1994.
- Keane, Michael P. and Kenneth I. Wolpin**, “The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence,” *The Review of Economics and Statistics*, 1994, 76 (4), 648–672.
- Kenney, Keith and Stephen Lacy**, “Economic Forces behind Newspapers’ Increasing Use of Color and Graphics,” *Newspaper Research Journal*, mar 1987, 8 (3), 33–41.
- Kleinberg, Jon and Steve Lawrence**, “The structure of the Web,” *Science*, 2001, 294 (5548), 1849–1850.
- Kleinberg, Jon M.**, “Authoritative sources in a hyperlinked environment,” *Journal of the ACM (JACM)*, 1999, 46 (5), 604–632.
- Kydland, Finn E. and Edward C. Prescott**, “Time to Build and Aggregate Fluctuations,” *Econometrica*, 1982, 50 (6), 1345–1370.
- Leckner, Sara**, “Presentation factors affecting reading behaviour in readers of newspaper media: an eye-tracking perspective,” *Visual Communication*, 2012, 11 (2), 163–184.
- Longstaff, Francis A. and Eduardo S. Schwartz**, “Valuing American Options by Simulation: A Simple Least-Squares Approach,” *Review of Financial Studies*, 2001, 14 (1), 113–147.
- Lu, Kristine and Jesse Holcomb**, *Digital News Revenue: Fact Sheet*, Pew Resarch Center, June 15, 2016; accessed November 17, 2016. <http://www.journalism.org/2016/06/15/digital-news-revenue-fact-sheet/>.
- Marimon, Ramon and Andrew Scott, eds**, *Computational Methods for the Study of Dynamic Economies*, Oxford University Press, 2001.
- Munos, Remi and Csaba Szepesváry**, “Finite-Time Bounds for Fitted Value Iteration,” *Journal of Machine Learning Research*, 2008, 1, 815–857.
- Murphy, Jamie, Charles Hofacker, and Richard Mizerski**, “Primacy and Recency Effects on Clicking Behavior,” *Journal of Computer-Mediated Communication*, 2006, 11 (2), 522–535.

- Narayanan, Sridhar and Kirthi Kalyanam**, "Position Effects in Search Advertising and their Moderators: A Regression Discontinuity Approach," *Marketing Science*, 2015, 34 (3), 388–407.
- Pakes, Ariel and Paul McGuire**, "Stochastic Algorithms, Symmetric Markov Perfect Equilibrium, and the 'curse' of Dimensionality," *Econometrica*, 2001, 69 (5), 1261–1281.
- Pew Research Center**, "The Modern News Consumer," 2016.
- Pijoan-Mas, Josep**, "Precautionary savings or working longer hours?," *Review of Economic Dynamics*, 2006, 9 (2), 326–352.
- Puterman, Martin**, *Markov Decision Processes* Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., 1994.
- Rincon-Zapatero, Juan Pablo and Carlos Rodriguez-Palmero**, "Existence and Uniqueness of Solutions to the Bellman Equation in the Unbounded Case," *Econometrica*, 2003, 71 (5), 1519–1555.
- Rowan, Thomas Harvey**, "Functional Stability Analysis of Numerical Algorithms." PhD dissertation, University of Texas at Austin 1990.
- Rust, John**, "Using Randomization to Break the Curse of Dimensionality," *Econometrica*, 1997, 65 (3), 487–516.
- Santos, Manuel S. and Jesus Vigo-Aguiar**, "Analysis of a Numerical Dynamic Programming Algorithm Applied to Economic Models," *Econometrica*, 1998, 66 (2), 409–426.
- Schechner, Sam and Stu Woo**, *EU Executive Proposes New Copyright, Communications Laws*, Wall Street Journal, September 14, 2016; accessed November 17, 2016. <http://www.wsj.com/articles/eu-leaders-propose-new-copyright-communications-laws-1473850691>.
- Smith, Shawn**, *Headline writing: How to write web headlines that catch search engine spiders*, New Media Bytes, April 4, 2008; accessed November 17, 2016). <http://www.newmediabytes.com/2008/04/10/how-to-write-headlines-for-search-engines/>.

**Stachurski, John**, “Continuous State Dynamic Programming via Non-expansive Approximation,” *Computational Economics*, 2008, 31 (2), 141–160.

–, *Economic Dynamics: Theory and Computation*, Cambridge, MA: The MIT Press, 2009.

**Sterling, Greg**, *Amid Tensions Googles Eric Schmidt Addresses Newspaper Conference*, Search Engine Land, April 7, 2009; accessed November 17, 2016. <http://searchengineland.com/amid-tensions-googles-eric-schmidt-addresses-newspaper-conference-17237>.

**Stokey, Nancy L., Robert E. Lucas, and Edward C. Prescott**, *Recursive Methods in Economic Dynamics*, Massachusetts: Harvard University Press, 1989.

**Train, Kenneth E.**, *Discrete Choice Methods with Simulation*, 2nd ed., Cambridge University Press, 2009.

**Tsitsiklis, John N. and Benjamin Van Roy**, “Feature-Based Methods for Large Scale Dynamic Programming,” *Machine Learning*, 1996, 22 (1-3), 59–94.

**van der Vaart, Aad W.**, *Asymptotic Statistics*, Cambridge: Cambridge University Press, 1998.

**Weber, Matthew S. and Peter Monge**, “The flow of digital news in a network of sources, authorities, and hubs,” *Journal of Communication*, 2011, 61 (6), 1062–1081.

**Weiner, Corey and Jun Group**, *Facebooks new “Instant Articles” program poses dilemma for publishers*, VentureBeat, May 11, 2015; accessed November 17, 2016. <http://venturebeat.com/2015/05/11/facebook-new-instant-articles-program-poses-dilemma-for-publishers/>.

**Zubiaga, Arkaitz**, “Newspaper editors vs the crowd,” in “Proceedings of the 22nd International Conference on World Wide Web - WWW ’13 Companion” ACM Press New York, New York, USA 2013, pp. 879–880.



# Appendix A

## Appendix for Chapter 2

### A.1 Additional tables

We present two tables that are too large for the main text. Table A.1 gives a detailed picture on differences between treatment and control users in the matching period. Tables A.2a and A.2b show share of top topics for treatment users via direct navigation and Google News.

### A.2 Browsing Data Processing

In this Section, we describe how we construct a clean dataset from the raw browser log.

#### A.2.1 Active Users

We only observe desktop users. We restrict attention to users whom we consistently observe over our observation period. User attrition occurs for various reasons such as change in default browser or a browser update that deletes the anonymous machine ID that links all browsing events for the same user across time. We label a user as *active* if both of the following criteria are met: (a) the user has some browsing activity in at least 90 percent of all weeks between October 1, 2014 and March

Table A.1: Differences between treatment and control users on news types (top 50 topics),  $2^5 = 32$  news categories. The types are 1 if a property applies to it and 0 if not, in the order of: relative scarcity, global scarcity, popularity, hard news, breaking news. We report whether we can reject equality of the two groups at the 10% level after applying the Benjamini-Hochberg correction.

value	treatment	control	p-value	BH corrected
0:1:0:0:1	1.272	1.513	0.003	rejected
0:0:0:0:1	1.125	1.314	0.005	rejected
0:1:0:0:0	2.909	3.307	0.011	not rejected
0:1:1:1:0	3.141	3.495	0.068	not rejected
0:1:1:0:1	0.393	0.482	0.084	not rejected
0:1:1:1:1	1.483	1.661	0.111	not rejected
1:0:1:0:0	0.680	0.578	0.119	not rejected
1:0:1:1:1	0.293	0.269	0.132	not rejected
0:0:1:1:1	0.666	0.726	0.144	not rejected
1:0:0:1:0	1.768	1.660	0.154	not rejected
0:0:1:0:1	0.689	0.603	0.189	not rejected
0:0:0:1:0	2.177	2.076	0.253	not rejected
0:1:1:0:0	1.038	1.204	0.257	not rejected
1:0:0:0:1	0.441	0.469	0.260	not rejected
1:0:0:0:0	1.464	1.391	0.282	not rejected
1:1:1:0:1	0.213	0.240	0.325	not rejected
0:0:1:1:0	1.404	1.479	0.341	not rejected
1:1:1:1:0	1.649	1.716	0.369	not rejected
1:0:1:0:1	0.273	0.255	0.382	not rejected
0:1:0:1:1	1.695	1.766	0.402	not rejected
1:0:0:1:1	0.547	0.567	0.484	not rejected
1:1:0:1:1	0.725	0.748	0.494	not rejected
1:1:1:0:0	0.823	0.879	0.499	not rejected
1:1:0:1:0	2.791	2.737	0.603	not rejected
1:1:0:0:1	0.516	0.530	0.612	not rejected
1:1:0:0:0	2.133	2.092	0.634	not rejected
0:1:0:1:0	3.223	3.265	0.771	not rejected
0:0:0:1:1	1.340	1.354	0.801	not rejected
0:0:1:0:0	1.268	1.293	0.827	not rejected
0:0:0:0:0	2.198	2.209	0.933	not rejected
1:0:1:1:0	0.752	0.753	0.976	not rejected
1:1:1:1:1	0.624	0.623	0.991	not rejected

Table A.2: Share of pageviews in top topics by referral mode, treatment users.

## (a) direct navigation

Topic	Share	Cum. share
Politics: Independence of Catalonia	0.069	0.069
Business/Finance: Macroeconomics and Economic Policy	0.068	0.137
Spain: Government	0.061	0.197
Sports: Atltico de Madrid	0.051	0.248
Health: Spain Ebola Crisis	0.042	0.290
Spain: Corruption Operation Pnica	0.030	0.320
Celebrities: Duchess of Alba dies	0.025	0.345
International News: Central and South America	0.025	0.370
Sports: Real Madrid	0.025	0.394
Sports: FC Barcelona	0.025	0.419
Sports: Soccer Players	0.024	0.443
Sports: Formula 1	0.023	0.466
Entertainment: TV Show Gran Hermano VIP	0.022	0.488
Spain: ETA	0.019	0.507
Business/Finance: Spain Banks	0.017	0.524

## (b) Google News

Topic	Share	Cum. share
Politics: Independence of Catalonia	0.117	0.117
Spain: Government	0.093	0.210
Health: Spain Ebola Crisis	0.085	0.295
Business/Finance: Macroeconomics and Economic Policy	0.060	0.354
Spain: Corruption Operation Pnica	0.033	0.387
Business/Finance: Spain Banks	0.026	0.413
Celebrities: Duchess of Alba dies	0.023	0.436
Sports: Atltico de Madrid	0.022	0.458
Sports: Real Madrid	0.021	0.479
International News: Central and South America	0.020	0.499
Sports: FC Barcelona	0.019	0.518
Conflict: War against Islamic State	0.018	0.536
Sports: Soccer Players	0.017	0.553
Spain: ETA	0.016	0.569
Politics: Spain	0.015	0.584

17, 2015; and (b) there is news reading activity in any 10-week long sub-period (for any of the top 177 Spanish publications).<sup>1</sup> There are 158,575 users who satisfy this definition of activity and they constitute the user base that we use for analysis.

### A.2.2 Browsing Stream

For each active user, we observe a set of *browsing sessions* which have a discrete beginning and end time (such as logging into and out of the computer). When there is a gap of more than one hour between browsing events, we define a new session starting after the gap ends. Within a session, we observe a time-stamped sequence of page loads which are linked through a referrer URL. For example, if a user navigates first to the NYT homepage and then clicks on an article on the home page, the browsing stream will show two page loads and the referral source for the second load will be the index page.

An absent referrer field indicates that the user either used a bookmark or typed the URL directly into the browser. An important special case are search queries that result in a visit to a landing page (such as <http://www.nytimes.com>). Often, the user first searches for a phrase such as “NYT” and then clicks on the link that appears on the search page. Even though the referrer field in this case indicates that the page load originated from a search we treat the referrer field as empty (as if the user had clicked a bookmark).

The browsing stream also provides us with a measure of how long the user spent on a certain page (“dwell time”). This data is capped at 300 seconds since larger values typically indicate inactivity (the user has walked away from the computer but left the browser open).

### A.2.3 Canonical URLs

URLs are often not unique even if they load the same article because publishers frequently include session and navigational parameters into

---

<sup>1</sup>We constructed this list by looking at all the unique domains that receive referrals from Google News Spain up to December 16, 2014. We then manually double-checked that this list contains the major Spanish newspapers by circulation, top weekly magazines and top radio stations and news portals.

the URL.<sup>2</sup> We therefore “stem” all URLs into a canonical format that removes most query parameters except for certain cases such as `?id=` which are sometimes used to index pages.

#### A.2.4 Landing and Article Pages

It is important for us to distinguish between article pages and landing pages (such as `http://www.nytimes.com` or `http://www.nytimes.com/business`). For each canonical URL, we count the number of page loads on every single day across all users and order days in reverse order according to page count. We then analyze how many days it takes to generate 80% of all page loads – if it takes more than 15 days then we regard a URL as landing page and otherwise an article. Intuitively, we exploit the fact that articles tend to appear only a small number of days and therefore generate the bulk of page views within that time frame. The NYT homepage, on the other hand, is loaded at a fairly constant rate on every single day.

When evaluated by human auditing, we found that this simple algorithm successfully classifies about 95% of all canonical URLs correctly.

#### A.2.5 News Minisessions and Referral Modes

Most users spend only a fraction of time browsing news during a browsing session - the remainder is spent on Facebook, email etc. Thanks to the referrer URLs, we can think of news browsing within a browsing session as a set of “trees”: each tree has a single root (a URL without a news referrer) and a number of branches that each have a (news) referrer URL that indirectly links to the root.

We call such a tree a “news mini session” (or NMS). We assign a referral mode to each NMS depending on whether the root page load was the result of direct navigation (such as a bookmark), a search on Google, Bing or Yahoo, navigation from Twitter or Facebook (such as clicking a news article on these platforms), a referral from Google News or some other origin (such as clicking on an email).

<sup>2</sup>Such as `https://www.nytimes.com?origin=google`, for example.

## A.3 Topic Classification

We create topics for news articles by mapping unique news article URLs (excluding landing pages) to Spanish language Wikipedia articles. We exploit the fact that Wikipedia covers most major news events within a short period of time. This provides us with a stable taxonomy for classifying a corpus of news articles. After automatically generating the clusters of articles, we manually name topics by their “super-topics” (high-level topics) and the “sub-topics” that in general correspond to the specific subject of matched Wikipedia articles.

### A.3.1 Data

*Scraping.* We scrape all unique URLs and extract the text using Boilerpipe which is an open-source program that can detect text within an html page (optimized for newspaper articles). Within each publisher we run a cleaning algorithm which compares the article text of all articles and identifies repeated sentences across articles. This algorithm removes boilerplate text such as privacy statements which otherwise pollute the article text. Sentences that occur in distinct URLs on more than 3 days a week over a one month period are removed using this heuristic.

*Wikipedia.* Wikimedia holds an up-to-date online repository of all Wikipedia articles for several languages. We extract all Wikipedia articles with more than 100 words. We also extract outlinks (connections between Wikipedia articles).

### A.3.2 Pre-processing

We build a dictionary of stemmed words using a standard Spanish language stemmer from all Wikipedia concepts (a concept is really an article on Wikipedia - we refer to Wikipedia articles as “concepts” in order to avoid confusion). We ignore rare words (words that appear in fewer than 3 wikipedia articles) and Spanish stop words (such as articles and frequent words such as the Spanish equivalents of “and” and “or”). We calculate the IDF (inverse document frequency) score for each dictionary

word defined as:

$$\log \left( \frac{\# \text{ all words}}{\# \text{ Wikipedia concepts in which specific word appears}} \right)$$

We do the same for each bigram on Wikipedia.

For each newspaper article we find all words and bigrams. We count within each newspaper article (including title) the number of occurrences of each word and bigram and calculate the TFIDF score as the number of occurrences times IDF score.<sup>3</sup> We then order all words and bigrams within the article according to this score in reverse order and take the top 30 words and top 30 bigrams. For each article, we call those top 30 keywords and top 30 bigrams the *article fingerprint*.

### A.3.3 Newspaper Article to Wikipedia Concept matching

In this step we match Wikipedia concepts to each newspaper article.

We calculate a word-based newspaper article/Wikipedia concept similarity score by taking the top 25 keywords of the article (out of 30). We then iterate through all Wikipedia concepts. For each Wikipedia concept we (i) check which of these 25 words appear in the Wikipedia concept; (ii) construct the sum of TFIDF scores for those words; (iii) divide this sum by the sum of all 25 TFIDF scores. This provides us a score between 0 and 1 that captures the similarity between the newspaper article and the Wikipedia concept based on words. We calculate an analogous bigram-based similarity score.

Each of these two scores lies between 0 and 1. We calculate the weighted average for each news article/Wikipedia article pair by placing weight 0.6 on the word score. We create a list of the top 10 Wikipedia concepts according to this weighted similarity score.<sup>4</sup>

<sup>3</sup>TFIDF scores are a standard scoring rule that is commonly used for information retrieval systems.

<sup>4</sup>We determined this weight as well as the number of keywords and bigrams on which the weight is based by hiring student raters who evaluated the resulting Wikipedia article assignment for a sample of newspaper articles. We then determined the three parameters through a grid search that maximized the overall evaluation score.

### A.3.4 Pure Macro Topics

Many Wikipedia concepts refer to very similar ideas. For example, the English language Wikipedia has several articles related to the passage of the Affordable Care Act. In order to define clean topics it is necessary to cluster closely related concepts into what we call a “pure topic.” In order to accomplish this clustering we construct a graph that connects Wikipedia concepts. This graph is based on two subgraphs – the *co-occurrence graph* and the *common outlink graph*.

The co-occurrence graph is constructed by assigning an edge weight to each pair of Wikipedia concepts  $i$  and  $j$  as follows: we find the number of newspaper articles that (i) list both these concepts among the top 3 matches and (ii) assign scores to both concepts of at least 0.5. Intuitively, the co-occurrence graph assigns a large weight to pairs of substitutable concepts.

For the outlink graph the weight between two concepts  $i$  and  $j$  is defined as follows: we count the number of concepts that both  $i$  and  $j$  either link to or are linked from (common neighbors). We then normalize this score by dividing by the size of the union of all articles that link to or are linked from  $i$  or  $j$ .

Finally, we merge both graphs by multiplying the edge weights (where some weights might go to zero as a result of this multiplication). Intuitively, Wikipedia concepts that are close in this merged graph tend to both have a high co-occurrence count and a high share of common outlinks. We use the fast Louvain algorithm (which maximizes modularity) to detect clusters (Blondel et al., 2008). We call these resulting clusters the “pure topics.”

Going back to the articles, every newspaper article whose top-scoring Wikipedia concept has at least weighted score 0.5 and is part of such a pure topic is then assigned to this pure topic. Articles whose top-scoring concept has a score of less than 0.5 or that do not connect to any concept in a pure topic are left unassigned at this stage.



### A.3.5 Topic Augmentation

In this step, we assign some of the so far unassigned newspaper articles to pure topics. We again use a clustering algorithm which we first overview before providing details. The basic idea is to construct a super-network of all articles (both assigned and unassigned). We create strong links between assigned articles in this super-network that belong to the same pure macro topic. This ensures that these articles will be assigned to the same clusters when finding communities in the super-network. We also add links between unassigned and assigned articles based on semantic and word similarity.

In an ideal world, we would construct the super-network using all articles. The problem is that with  $N$  million articles, calculating edge weights between all pairs of articles requires  $N^2$  calculations which is infeasible given the number of unique articles. To reduce the computation time, we use a short-cut that iteratively adds nodes to the network and groups them into communities.

We take the set of pure topics such that the page view count makes up 90% of assigned articles. For each pure topic we take 20 articles which have a top weighted Wikipedia concept score above 0.9, and we sample by page view weight (these tend to be representative articles that clearly belong to the respective pure macro topic). We construct a base network of “islands” where each island is a complete graph that connects all 20 base article belonging to a particular macro topic (weight 1 on each edge). Hence, there are as many islands as pure topics. The total number of base articles  $M$  is equal to 20 times the number of pure topics.

We then take batches of  $M/2$  unassigned articles and embed them in this base network. We connect them to each of the  $M$  base articles by using a weighted average of the following two sub-weight.

1. The first sub-weight  $w_{ij}$  between an unassigned article  $i$  and base article  $j$  is determined by semantic similarity (distance between the associated Wikipedia concepts). We take the top 5 associated Wikipedia concepts for both  $i$  and  $j$ , and then take the intersection of both concepts sets. Then we sum  $i$ 's scores for these overlapping concepts and divide by the sum of  $i$ 's top 5 scores to get a number between 0 and 1. Analogously, we calculate the corresponding

score for  $j$ . The average of these two numbers defines our first sub-weight  $w_{ij}$ . Intuitively, this weight is large if both newspaper articles are associated with similar Wikipedia concepts and more important concepts receive higher weight. We then calculate the top 1 percentile cutoff over all these sub-weights and set all the weights below this cutoff to 0. This “pruning” reduces noise.

2. The second sub-weight  $\tilde{w}_{ij}$  measures word and bigram similarity between the two articles. We consider word similarity first. We take the top 8 keywords for articles  $i$  and  $j$  and the associated TFIDF scores from both fingerprints. We then use the same procedure as for semantic similarity. We then repeat the procedure again for the top 5 bigrams. We average the resulting scores by putting weight 0.6 on the word score. This provides us with the second sub-weight  $\tilde{w}_{ij}$ . Intuitively, this weight is large if both articles are associated with similar keywords and bigrams. We use the same 1 percentile pruning as applied to the semantic network.

We add both sub-weights and divide by half. This gives us a  $[0, 1]$  weight that defines the graph links between the base network and the unassigned batch articles. We again run the Louvain community detection on each resulting super-network, such that we always preserve the original pure macro topics but “augment” them with the previously unassigned articles.

The algorithm scales linearly with the dataset.

### A.3.6 Supervised cleaning

The last step is the only supervised step that requires manual intervention. Recall, that each macro topic is defined by the set of associated Wikipedia concepts. We have student evaluators define English-language nested labels for these 320 labels such as *Health: Spain Ebola crisis* for the macro topic that describes the 2015 Ebola crisis in Spain. Our nesting has two levels: super-topic (*Health*) and sub-topic (*Spain Ebola crisis*). This makes it easy to evaluate the final topic assignment by separately rating the super-topic and sub-topic assignments.