COMPUTATIONAL PERSPECTIVES ON INDIVIDUAL AND COLLECTIVE DECISION-MAKING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kiran Tomlinson

August 2024

COMPUTATIONAL PERSPECTIVES ON INDIVIDUAL AND COLLECTIVE DECISION-MAKING

Kiran Tomlinson, Ph.D.

Cornell University 2024

Individual decisions determine the success of companies (Pepsi or Coke?) and the structure of our social networks (Alice or Bob?), while collective decisions determine the composition of our governments and the outcomes of criminal trials, among countless other facets of our lives. As such, understanding the factors that contribute to these decisions is crucial, both for predicting future decisions and for designing interventions. In this dissertation, we use computational techniques to address two core questions towards this end. First, can we learn about how people make choices from individual decision-making data? Second, how do we aggregate group preferences in collective decision-making and what are the consequences of different aggregation mechanisms?

After a brief introduction in Part I, Part II describes several methods to learn the effects of social and contextual factors on preferences in individual discrete choice settings, synthesizing tools from interpretable machine learning, causal inference, and graph learning. In Part III, we turn to collective decisions, focusing on theoretically characterizing the behavior of two commonly used voting systems, plurality and instant runoff voting (IRV). In particular, we explore what happens under IRV when voters are forced to submit top-truncated preferences, prove that IRV favors moderate candidates in a way plurality does not, and examine the dynamics of candidate policies under a boundedly-rational imitative model. We conclude in Part IV with closing thoughts and directions for future work.

BIOGRAPHICAL SKETCH

Kiran Tomlinson is a computer scientist specializing in the study of human preferences and the algorithms that interact with them, including voting systems and recommender systems. He was born in Monaco and raised in Levallois-Perret, France and Austin, Texas. In 2019, he received a BA in Computer Science and Mathematics from Carleton College, *summa cum laude.* While an undergraduate, he interned twice at the NASA Johnson Space Center, where he worked on support software for Orion. After Carleton, he began his PhD in Computer Science at Cornell, advised first by Austin Benson and later by Jon Kleinberg. During his PhD, he interned twice at Microsoft, in the Office of Applied Research and in Microsoft Research, and spent one semester as a Visiting Instructor of Computer Science at Carleton College. After completing his PhD, he will join Microsoft Research as a Senior Researcher in the Augmented Learning and Reasoning group.

To my mother. Merci, thank you, and obrigado.

ACKNOWLEDGEMENTS

The list of people who made my PhD possible and enjoyable is long.

First, I am so lucky to have been advised by someone as thoughtful, kind, and insightful as Jon Kleinberg. Thank you for working with me when I was in a tricky spot, for guiding me in interesting and exciting directions, and for asking the best questions. I'm very grateful for your support and mentorship.

I also want to thank my first advisor, Austin Benson, whose guidance helped me grow into a confident researcher. I really appreciate everything you taught me about the research process, especially helping me become more comfortable with ambiguity and developing my skills in paper-writing, plot-making, and slide-craft. I would have been proud to graduate with you as my advisor had you wanted to stay in academia.

I want to give a huge thank you to Johan Ugander, who has been like a second (or third?) advisor to me. You made the advisor switch so much easier and are a role model in your curiosity, excitement, and pretty much everything.

Another thank you goes to my two additional committee members, Bobby Kleinberg and Thorsten Joachims; you both helped make me feel welcome at Cornell and provided guidance as I started my PhD.

Looking farther back, I wouldn't be at Cornell without the guidance of David Liben-Nowell, who has inspired me since my first visit to Carleton. Thank you for your continuous support, for starting my research path in college, and for making me fall in love with CS theory. I'm grateful to Layla Oesper for showing me what full-time research looks like, teaching me how to write a paper, and sending me to my first conference. Working with you convinced me to go to grad school. More broadly, I am grateful to the Carleton College Computer Science Department for creating such a vibrant and friendly atmosphere, both as a student and as a visiting instructor. In particular, thank you to Jeff Ondich, Amy Csizmar Dalal, Anna Rafferty, Eric Alexander, Dave Musicant, and Sneha Narayan. I also want to thank Eric Egge and Deanna Haunsperger in the Math Department for showing me the beauty of a great proof.

Outside of my undergraduate and graduate degrees, I've been lucky to work with great mentors at Microsoft, NASA, and UT Austin. Thank you to Longqi Yang, Mengting Wan, and Jennifer Neville, for letting me work on amazing problems and being advocates for my work. Thank you to Nathan Uitenbroek for your faith in my abilities. Thank you to Calvin Lin for generously giving your time to a high-school student who knew nothing about prefetching, in many ways starting the chain of dominos that led me here.

In addition to those listed above, I want to thank my other coauthors: Ben Aoki-Sherwood, Catherine Bregou, Patty Commins, Zach DiNardo, Brent Hecht, Tina Liu, Cao Lu, Tanvi Namjoshi, Anna Ritz, Jaime Teevan, and Thomas Zeng. It was a pleasure to work with each and every one of you. In particular, I want to give a special thank you to Tanvi, who ran many simulations for Chapter 8.

My time in Ithaca was made vastly better by my wonderful peers in the CS department. They are too many to name in full, but I particularly want to thank Rachit Nigam, Sloan Nietert, Marios Papachristou, and Greg Yauney for the bike rides and long talks; Jerry Chee, Griffin Berlstein, and Kangbo Li for being great cubicle-mates; Kate Donahue and Katy Blumer for being inspirations in research and in positivity; Eric Campbell for making Ithaca more musical; and Stanley Celestin and Spencer Peters for the Lake Street cooking extravaganzas.

I am deeply grateful to Sophia Franco for making my PhD so joyful.

Finally, I'm grateful to my family for their support through all the ups and downs and moves. Thank you to my dear vovó, Patricia Maazel; to my father, Peter Tomlinson; to my cheerleaders Daria, Lionel, and Owen Steketee; and to my grandfather, Lorin Maazel, who made my education possible. Finally and most of all, thank you to my mother, Anjali Maazel, for giving me so much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Part I

Introduction

CHAPTER 1

INTRODUCTION

Decision-making is the fundamental human task that shapes the trajectories of our lives and of our societies; as such, it has been the focus of intense study for decades (and even centuries) in fields as varied as mathematics, psychology, economics, political science, and—most recently—computer science. While approaches from behavioral economics and psychology can experimentally reveal the underlying factors that govern our choices, computational tools provide a powerful means for extracting useful insights from large-scale decision-making datasets, designing efficient and flexible predictive models, designing optimal decision-making interventions, and examining decision-making procedures as algorithms. By better understanding decision-making through these tools, we can design more effective public policies, create more successful products and media, and identify biases.

This dissertation addresses two types of decision-making: at the individual and collective level. First, we design interpretable models to reveal how contextual and social factors influence individual discrete choices, where people choose from a set of available items. We show how these models provide improved prediction by accounting for contextual and social effects on preferences. We also examine the causal inference challenges in learning such factors from observational data. In the second part of the dissertation, we focus on collective decision-making, where a large number of individuals combine their preferences through voting to make a decision for the group. We focus on the two single-winner voting systems in most common use (plurality and instant runoff) and analyze them as algorithms, understanding their worst- and average-case behavior. To bridge the two parts, we also explore optimal interventions in group decisions using individual-level models of choice.

We begin by providing the reader with some background on the areas of discrete choice and voting, which will be useful in understanding how our work is positioned in the literature. We also provide a summary of our contributions in both individual and collective decision-making.

1.1 Discrete choice background

Discrete choice describes settings where individuals are faced with a set of options and select one of them (Train, 2009). Such settings occur constantly in our lives, from the mundane to the highly consequential. For instance, will you drive, walk, or take the bus to work? Will you buy Chobani, Fage, or Oikos yogurt? Which graduate school will you attend? Which city will you move to? The answers to these questions (even the apparently mundane ones) have significant impacts on urban transit systems and yogurt companies, as well as university enrollments and the economic fates of cities—indeed, it is hard to conceive of any aspect of our lives which is not affected by discrete choices. Given the ubiquity of discrete choice, it is not surprising that there has been a substantial quantity of research in this area in the last century.

The core goal of the study of discrete choice is to provide an explanatory model for the choices people make. Ideally, such a model can tell us the factors that contribute to a choice and can also provide predictions for future choices. These models can then be used for downstream tasks like designing new products that align with people's preferences (Wassenaar et al., 2005) or deciding what to stock in a store to maximize revenue (Rusmevichientong et al., 2014). The conventional wisdom is that choice models need to be stochastic, since we cannot hope to capture every single factor that might determine an action: sometimes you might buy Fage and sometimes Oikos, and since we should never expect to know all the micro-level factors that go into this variation, it is often useful to model them as random. On the other hand, the classical approach in economics is to posit that individuals are rational utility maximizers. The theory of *random utility models* (RUMs) (Thurstone, 1927; Marschak, 1959; Block and Marschak, 1960; Manski, 1977) bridges these two ideas, positing that individuals sample random utilities for each item and then select the one with highest observed utility (these random utilities may differ in different instances, making choice stochastic). Much of the early work in discrete choice is based on another rationality assumption known as the *independence of irrelevant alternatives* (IIA) or *Luce's choice axiom* (Luce, 1959), which informally states that the utility of an item should be independent of the set of alternatives in which it appears. Interestingly, there is a unique RUM satisfying IIA, where the random utility of each item is equal to its mean utility plus i.i.d. Gumbel noise (Luce and Suppes, 1965; McFadden, 1974; Yellott Jr, 1977). In a seminal paper, McFadden (1974) showed that this Gumbel-noise RUM, which he called the *conditional logit*, can be estimated from population-level choice data and that utilities can be parameterized by covariates of the items or choosers to understand how those features influence preferences. This work eventually led to the 2000 Nobel Prize in Economics, awarded to McFadden (and shared with James Heckman) for "his development of theory and methods for analyzing discrete choice" (NobelPrize.org, 2000).

After this early work on rational choice, the study of discrete choice took a major turn in the '80s and '90s—along with much of behavioral economics and psychology—as increasing doubt was cast on traditional rationality assumptions. In particular, laboratory experiments found that people consistently violate assumptions like IIA, changing their apparent preferences depending on the context of the choice (Tversky and Kahneman, 1981; Huber et al., 1982; Simonson and Tversky, 1992; Shafir et al., 1993). For instance, people tend to choose a compromise among the available items (Simonson, 1989), such as the middle-priced wine on a menu (different menus then result in violations of IIA). Such violations are called *context effects*, and much of the work on discrete choice in the last several decades has focused on detecting and modeling context effects.

Most recently, the increasing availability of large online choice datasets has led to significant interest in discrete choice in the computer science community, including in models of context effects (Chen and Joachims, 2016a; Benson et al., 2016, 2018b; Seshadri et al., 2019; Bower and Balzano, 2020; Rosenfeld et al., 2020) and applications of discrete choice to social network growth (Overgoor et al., 2019, 2020; Gupta and Porter, 2022; Ma et al., 2022). This line of work is solidly based in the random utility theory pioneered by McFadden and Manski, but also draws on Tversky and Simonson's findings about the context-dependence of preferences. Meanwhile, these papers bring in tools from computer science and machine learning, including neural networks (Rosenfeld et al., 2020), sample complexity analysis (Seshadri et al., 2019), and algorithm design (Benson et al., 2016). These computational techniques have helped make discrete choice modeling more efficient and powerful in the age of big data. It is in this computational discrete choice literature that Chapters 2 to 5 of this dissertation reside.

1.2 Voting background

Now that we have established some context for the individual decision-making portion of this dissertation, we turn to collective decision-making. In order to make society-level decisions, such as electing a president or deciding on city policies, democracies need to combine the preferences of a large number of individuals into a single choice. Such a voting system is comprised of two components: an elicitation mechanism (how do we ask people about their preferences?) and an aggregation mechanism (how do we combine the elicited preferences into a final decision?). Discrete choice models tell us about how people might reply to the elicitation mechanism, but the aggregation mechanism is something quite different: we must impose a particular algorithm to combine preferences into a societal choice.

A huge number of different voting systems have been proposed, dating back hundreds and even thousands of years; ancient Athens used majority voting in the 5th century BCE (Tridimas, 2019), Borda count was proposed in 1435 by Nicholas of Cusa for electing the Holy Roman Emperor (Emerson, 2013), and approval voting was used to elect the Doge of Venice from 1268 to 1789 (Lines, 1986). Each of these systems can be understood as an algorithm whose input is the set of preferences of the population and whose output is the winner. Analyzing these algorithms (or designing new ones) is the purview of *computational social choice*, where computational techniques have proven very useful for better understanding voting systems. Some voting systems are even NP-hard to evaluate; one notable example is Dodgson's method (Bartholdi et al., 1989; Ratliff, 2001), invented by Charles Dodgson in 1876 (also known by his pen name, Lewis Carroll), although he did not know of its NP-hardness at the time.

One of the central themes in the mathematical theory of voting is that a perfect voting system is fundamentally impossible. Indeed, this is one of the reasons for the huge proliferation of different systems. The first impossibility in voting was discovered by the Marquis de Condorcet in 1785 (de Condorcet, 1785): when preferences are expressed as rankings over candidates, it is possible that a majority of voters prefers A to B, a majority prefers B to C, and a majority prefers C to A—creating a *Condorcet cycle*. This is known as the *Condorcet paradox*, and can arise even with only three voters with preference rankings ABC, CAB, and BCA. If there does exist some candidate which is preferred by a majority of voters to every other candidate, they are known as the *Condorcet winner*. A voting system that always elects the Condorcet winner when one exists is called *Condorcet consistent*. In some sense, the Condorcet paradox shows that there are some profiles (as collections of rankings are termed in voting theory) in which no "good" winner can be chosen. Another famous impossibility result in voting is due to Arrow, who showed that no ranking-based voting system can simultaneously satisfy three seemingly desirable properties: non-dictatorship, Pareto efficiency, and the independence of irrelevant alternatives[1] (Arrow, 1950). Yet another impossibility result is the Gibbard–Satterthwaite theorem, which essentially states that any voting system can be strategically manipulated (Gibbard, 1973; Satterthwaite, 1975). In sum, we cannot hope to identify a perfect voting system. Rather, we are left to choose between a number of suboptimal options—although some exhibit more desirable properties than others.

There are many properties which might be considered desirable in a voting system, such as the likelihood of selecting the Condorcet winner (if one exists), the simplicity and explainability of the system, and a myriad of theoretical properties devised by voting theorists (monotonicity, unanimity, and participation, to name

[1] This IIA property is different from Luce's IIA axiom in discrete choice, although they are similar in spirit. The precise definitions of Arrow's properties are not important for our purposes.

just a few; see Nurmi (1987) for a discussion of different voting systems and their properties). Despite the extensive research into various voting systems, real-world elections often use very simple systems like plurality. As Niemi and Riker observe, "in the real world, however, the adoption of a voting system is determined largely by considerations of expense, convenience and familiarity" (1976). This is especially apparent in the United States, which primarily uses plurality voting, despite the fact that voting experts widely agree that plurality is among the worst voting systems used in practice (for instance, a survey of experts ranked plurality last among nine possible voting systems (Bowler et al., 2005)). However, there has been a recent push to adopt *instant runoff voting* (IRV, commonly called *ranked-choice voting*) in the United States as a better alternative to plurality. IRV is already used by two states, Maine and Alaska, as well as a number of local municipalities including Minneapolis, San Francisco, and New York City. It is this context that frames the second part of this dissertation, which seeks to better understand how IRV and plurality behave through mathematical analysis and simulation, with the goal of informing the ongoing debate between the two systems.

A crucial tool in the analysis of voting systems that we will use extensively is the concept of *preference restrictions* (Elkind et al., 2016, 2022). As discussed earlier, general preference profiles can be very complex and messy, exhibiting features like Condorcet cycles. However, if we assume that voter preferences have some additional structure, then much of this complexity is resolved. Indeed, real-world preferences do appear to have such structure; many of the pathological worst cases occur rarely in practice. Perhaps the most famous preference restriction is *single-peakedness* (Black, 1948; Arrow, 1951), where candidates lie on a one-dimensional axis and voters each have an ideal point along the axis, preferring candidates closer to this point. With single-peaked preferences, a Condorcet winner always exists and is the candidate closest to the median voter. We will use this and other preference restrictions in Chapters 6 to 8.

1.3 Overview of individual decision-making contributions

Given this brief overview of discrete choice and voting, the two areas covered in this dissertation, we provide a summary of our contributions, beginning with individual decision-making.

In Chapter 2, we introduce a novel choice model for contextual preferences which is easy to fit to data and interpret. The model, which we call the *linear context logit* (LCL) is an extension of the conditional logit (McFadden, 1974) that captures how different features of available items influence each others' valuations. For instance, how do people's preferences differ when choosing from a set of expensive restaurants compared to a set of cheap restaurants? We might hypothesize that service speed is more important when choosing between cheap restaurants; this is exactly the type of effect that the LCL captures. We show that the LCL has a negative log-likelihood, so it is easy to perform maximum likelihood estimation by gradient descent methods. We also prove a necessary and sufficient identifiability condition, which is typically satisfied in practice. We fit the LCL to a variety of real-world choice datasets, including preferences over cars, hotel bookings on Expedia, and sushi types. We find a number of statistically significant context effects (using a likelihood ratio test against conditional logit), including that people paid for more expensive hotels on Expedia when many options had high star ratings, but less expensive hotels when many options had high review scores. Additionally, we apply the LCL to a variety of social network link-formation datasets, where the discrete choice is which node to connect with (Overgoor et al., 2019). We find a variety of shared effects among datasets of similar kinds; for instance, in several online commenting datasets, we find that individuals are more likely to prioritize shared connections when choosing from a popular group with high in-degree.

Next, in Chapter 3, we turn to a major challenge in identifying context effects from observational choice data: the possibility of confounding factors. Without additional knowledge, we cannot conclude that the effects we observe using the LCL and other models of context effects are capturing actual causal influences on preferences rather than coincidental associations in the data. We formalize this issue as *choice set confounding* and establish the conditions when we can be confident that apparent context effects are in fact causal. We also demonstrate that the San Fransisco transportation datasets (Koppelman and Bhat, 2006) commonly used as testbeds for context effect models (Koppelman and Bhat, 2006; Ragain and Ugander, 2016; Benson et al., 2016; Seshadri et al., 2019) very likely have choice set confounding rather than true context effects. Then, we adapt tools from causal inference to the discrete choice setting, allowing us to recover true context effects using regression controls and inverse probability weighting (given certain independence assumptions, as is standard in causal inference). We find in the Expedia hotel choice data that much of the apparent deviation from IIA we observed in Chapter 2 is due to choice set confounding, although LCL continues to fit the data better than conditional logit even after adding controls. These causal inference methods rely on having access to features of the choosers, but we also develop a clustering-based method which allows us to mitigate confounding even without chooser features.

We conclude the first part of this dissertation by addressing another factor that influences preferences: social ties. There are several ways in which social networks can inform choice prediction, as preferences are shaped by social context through word-of-mouth and effects like conformity (Feinberg et al., 2020; Axsen and Kurani, 2012). Additionally, people with similar preferences are more likely to be friends (McPherson et al., 2001), making social structures even more informative about preference correlations. In Chapter 4, we adapt methods from graph learning for use in discrete choice, taking advantage of this relationship between social ties and similarity in preferences. We apply Laplacian regularization to discrete choice models, and adapt label propagation and graph neural networks for choice prediction. We find significant improvement in choice prediction in an app-downloading dataset, where individuals in close contact are more likely to download the same apps. We also use geographic adjacency networks to improve county-level election models using our graph-based discrete choice approach, which is especially useful in a semi-supervised setting where data is only available in a subset of counties.

1.4 Overview of collective decision-making contributions

Chapter 5 provides a bridge between the two sections of the dissertation, as it uses discrete choice models in a collective decision-making scenario. In that chapter, we consider a setting where a group of agents makes a choice over a shared set of options, such as friends deciding where to go for dinner or a hiring committee deciding over a set of candidates. Each agent has their own preferences and chooses according to a discrete choice model. The primary goal we consider is to help the decision-makers agree on the available options by suggesting new alternatives. This is a combinatorial optimization problem, as introducing different sets of alternatives has different effects on the consensus of the group. We prove that this optimization problem is NP-hard and develop a fully polynomial time approximation scheme (FPTAS) with an additive approximation guarantee. We also consider two other objectives, promoting a particular item and minimizing rather than maximizing agreement, which we also show are NP-hard and admit the same approximation algorithm. These results hold across multiple discrete choice models, including the logit and several models of context effects. Interestingly, we show that natural preference restrictions can make promoting an item easy but leave optimizing agreement NP-hard. In this sense, we show that promotion is easier than consensus (when suggesting additional items). We demonstrate the effectiveness and efficiency of our approximation algorithm in three real-world choice datasets.

We then turn our attention squarely to voting for the remainder of the dissertation. In Chapter 6, we ask how the length of a ballot (i.e., the number of available ranking slots) can influence the winner of an IRV election. This is particularly relevant to the ongoing debate around IRV and plurality, as plurality can be viewed as IRV with only a single ranking slot (ballot length 1), while different municipalities using IRV use ballot lengths ranging from 3 (as in San Francisco) to the total number of candidates running (as in Maine). The extent to which ballot length might influence the outcome of an election was previously unknown. We establish that in a k-candidate election with fixed voter preferences, there can be up to $k-1$ different winners depending on which ballot length is chosen. Moreover, the sequences of winners at different ballot lengths can be nearly arbitrary, subject only to a simple feasibility constraint, and we provide explicit constructions achieving any feasible sequence. For instance, we can construct a profile such that candidate A wins at prime ballot lengths and candidate B wins at composite ballot lengths. We show that these pathological profiles need only a quadratic number of voters (in

k, the number of candidates) and do not rely on tie-breaking. We also investigate the extent to which preference restrictions eliminate these pathologies; we can still achieve $O(\sqrt{k})$ different winners across ballot lengths with single-peaked profiles. In a large collection of 168 real-world election datasets, we find different winners across ballot lengths in 25% of them; but we only a find a single instance with more than two different winners. Simulations confirm this finding: ballot length often has some impact on the winner (particularly going from ranking only top 1 or 2 to 5 or more), but pathological cases are very rare. These results help highlight ballot length as a consequential degree of freedom in designing IRV elections. We have received some feedback that these results may be seen as critical of IRV as a voting system, but we remind the reader that plurality is IRV with only top-1 rankings and that well-behaved voting systems are provably impossible; IRV is indeed imperfect, but we still see it as an improvement to plurality.

Next, in Chapter 7, we address a common point of contention between supporters and critics of IRV: whether it benefits moderates or extreme candidates. To do this, we use a one-dimension model of political ideology, from the political left to the right, adding a Euclidean metric to single-peaked preferences. In this setting, we prove a very surprising yet simple new result: when voters are uniformly distributed along the political spectrum, IRV always elects a moderate candidate from the middle two-thirds of the spectrum (when one is available). In contrast, plurality can elect arbitrarily extreme candidates. Note that in this setting, any Condorcet consistent voting system would elect the most moderate candidate, but neither IRV nor plurality are Condorcet consistent. We show that IRV's moderating effect generalizes to other symmetric voter distributions. In addition, we use a stick-breaking argument to derive the asymptotic winning vote share under plurality with uniform voters and candidates. This allows us to prove that the position of the plurality winner approaches uniform over the interval as the number of uniformly distributed candidates grows. Finally, we also describe a geometric approach of deriving the exact plurality winner distribution with uniform voters and candidates and apply it for the case of $k = 3$ candidates. Overall, the results of this chapter demonstrate that IRV exhibits a preference for moderate candidates in a way that plurality does not.

In this dissertation's final chapter, Chapter 8, we continue using a one-dimensional model of ideology, but begin to consider strategic behavior by candidates. The previous two chapters dealt only with the mechanics of the voting system, setting aside the possibility of candidate strategy. Of course, real candidates are motivated to choose policies in such a way that benefits their chances of election. Previous work in this area often assumes candidates play strategic games and analyzes Nash equilibria of these games. However, real-world elections are so complex that we argue candidates are unable to play optimally. We therefore explore a boundedly-rational heuristic that candidates might follow: imitate the policy of a previous winner. This results in a discrete-time evolutionary replicator dynamics model of candidate positioning. We prove that this model has two different asymptotic behaviors depending on how many candidates run in each election. When there are fewer than five candidates per election, we prove that imitating previous winner policies results in all candidates converging to the median of a uniformly distributed voting population. In contrast, we prove that the candidate distribution does not converge to the center with $k \geq 5$ candidates per election. When all candidates are in the middle half of the interval, we also show the stronger statement that the candidate density in an interval around the center goes to zero. We tie these replicator dynamics convergence results in with prior strategic work and demonstrate our theoretical results in simulation.

1.5 Publications

The chapters of this dissertation have been published previously as follows.

Chapter 2:

Kiran Tomlinson and Austin R Benson. Learning interpretable feature context effects in discrete choice. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1582–1592, 2021

Chapter 3:

Kiran Tomlinson, Johan Ugander, and Austin R. Benson. Choice set confounding in discrete choice. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1571–1581, 2021

Chapter 4:

Kiran Tomlinson and Austin R. Benson. Graph-based methods for discrete choice. *Network Science*, 12(1):21–40, 2024

Chapter 5:

Kiran Tomlinson and Austin R Benson. Choice set optimization under discrete choice models of group decisions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9514–9525. 2020

Chapter 6:

Kiran Tomlinson, Johan Ugander, and Jon Kleinberg. Ballot length in instant runoff voting. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 5841–5849, 2023

Chapter 7:

Kiran Tomlinson, Johan Ugander, and Jon Kleinberg. The moderating effect of instant runoff voting. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 9909–9917, 2024b

Chapter 8 (preprint):

Kiran Tomlinson, Tanvi Namjoshi, Johan Ugander, and Jon Kleinberg. Replicating electoral success. *arXiv:2402.17109*, 2024a

Part II

Individual Decision-Making

CHAPTER 2

LEARNING INTERPRETABLE FEATURE CONTEXT EFFECTS IN DISCRETE CHOICE

In a *discrete choice* setting, an individual chooses between a finite set of available items called a *choice set*. This general framework describes a host of important scenarios, including purchasing (Anderson et al., 1992), transportation decisions (Train, 2009), voting (Dow and Endersby, 2004), and the formation of new social connections (Overgoor et al., 2019; Gupta and Porter, 2022). Discovering and understanding the factors that contribute to the choices people make has broad applications in, e.g., recommender systems (Yang et al., 2011; Ruiz et al., 2020), Web search (Ieong et al., 2012), online dating platforms (Bruch et al., 2016), and policy design (Brownstone et al., 1996).

The *conditional logit* (McFadden, 1974) (also called the *multinomial logit*) is the most famous and widely used discrete choice model. This model obeys the axiom of *independence of irrelevant alternatives* (IIA) (Luce, 1959), that relative preferences between items are unaffected by the choice set—if someone prefers x to y, they should still do so when z is also an option. However, experiments on human decisions (Huber et al., 1982; Simonson and Tversky, 1992; Shafir et al., 1993; Trueblood et al., 2013) as well as direct measurement on choice data (Small and Hsiao, 1985; Benson et al., 2016; Seshadri et al., 2019) have found that this assumption often does not hold in practice. These "IIA violations" are termed *context effects* (Prelec et al., 1997; Rooderkerk et al., 2011). Examples include the *attraction effect* (Huber et al., 1982), where including an inferior item makes a better option more attractive, and the *similarity effect* (Tversky, 1972), where similar items split the preferences of the chooser.

The ubiquity of context effects has driven the development of more nuanced models capable of capturing them. In machine learning, the goal is typically to design models for better predictions via learned context effects (Chen and Joachims, 2016a,b; Seshadri et al., 2019; Pfannschmidt et al., 2022; Rosenfeld et al., 2020; Bower and Balzano, 2020; Ruiz et al., 2020). However, the effects accounted for by models using neural networks or item embeddings (Pfannschmidt et al., 2022; Chen and Joachims, 2016b; Rosenfeld et al., 2020) are difficult to interpret. Other models learn context effects at the level of individual items (Chen and Joachims, 2016a; Seshadri et al., 2019; Natenzon, 2019; Ruiz et al., 2020), preventing generalization to items not in the training set and making it difficult to discover context effects coming from item features (e.g., price). Within behavioral economics, context effect models tend to be engineered for specific effects and are often only applied to controlled special-purpose datasets (Rooderkerk et al., 2011; Tversky and Simonson, 1993; Masatlioglu et al., 2012).

Here, we provide methods for learning a wide class of context effects from large, pre-existing choice datasets in a variety of domains. The key advantage of our approach is that we can take a choice dataset collected in any domain (possibly collected passively), efficiently train a model, and directly interpret the learned parameters as intuitive context effects. For example, we find in a hotel booking dataset that users presented with more hotels on sale showed increased willingness to pay. This lets us hypothesize that "on sale" tags on hotels exerts a context effect on the user, making them feel better about selecting a more expensive option. Context effects extracted by our methods could then motivate further experimental work such as A/B testing or choice set design to steer behavior. We focus on the setting where items are described by a set of features (e.g., for hotels: price, star rating) and where the utility of each item is a function of its features. This setup has two major benefits, as it enables (i) making predictions about new items not observed in training data, and (ii) learning generalizable and testable effects that can inform marketing, advertising, or recommendation.

We define *feature context effects* that describe changes in importance of features when determining choice as a function of features of the choice set. For instance, suppose a diner has two choice sets on two different occasions, one consisting of fast food chains and the other of high-end restaurants. In the choice set with lower prices, the diner could value service speed relatively more, and in the choice set with higher prices, the diner might place more weight on wine selection. We introduce two models, the linear context logit (LCL) and decomposed linear context logit (DLCL), to learn these types of feature context effects directly from choice data.

We perform an extensive analysis of choice datasets using our models, showing that statistically significant feature context effects occur in empirical data and recovering intuitive effects. For example, we find evidence that people pick more expensive hotels when their choice sets have high star ratings, that people offered more oily sushi show more aversion to oiliness (a possible similarity effect), and that when deciding whose Facebook wall to post on, people care more about mutual connections when choosing from popular friends.[1] Accounting for feature context effects also improves prediction accuracy in many datasets, although our primary focus is learning interpretable context effects. Additionally, we show to statistically test for effects and how sparsity-encouraging regularization can identify the most influential context effects.

Our empirical study is split into two parts. First, we examine datasets collected to understand preferences, covering a variety of choices including sushi,

[1]These are all correlative rather than causal claims; Chapter 3 deals at length with the challenge of causality in modeling context effects.

hotel bookings, and cars. Second, we apply our methods to social network analysis, where we demonstrate context effects in competing theories of triadic closure, the of new friendships to form among friends-of-friends (Easley and Kleinberg, 2010; Granovetter, 1973). Discrete choice models have recently found compelling use in analyzing social network dynamics (Overgoor et al., 2019, 2020; Gupta and Porter, 2022; Feinberg et al., 2020). Here, we find new insights by incorporating context effects.

Additional related work

Within machine learning, our LCL model is similar in spirit to the context-dependent random utility model (CDM) (Seshadri et al., 2019) in that we consider pairwise contextual interactions, but with the important distinction that our model operates on features rather than on items, allowing for the discovery of general, non-item-specific effects. Our framework for context-dependent utilities is related to set-dependent weights (Rosenfeld et al., 2020) and FETA (Pfannschmidt et al., 2022); these methods are optimized for prediction accuracy and are difficult to interpret. Other models for context effects include the blade-chest model (Chen and Joachims, 2016a,b) for pairwise comparisons and the salient features model (Bower and Balzano, 2020), which considers different subsets of features in each choice set.

Recent research has framed network growth (the formation of new connections in, e.g., communication or friendship networks) as discrete choice and have suggested context effects as a means for more flexible modeling (Overgoor et al., 2019). The models we introduce are a step in this direction, and we find context effects useful for both improved predictions and gaining new social insights. Other research has explored mixed (Gupta and Porter, 2022) and de-mixed (Overgoor et al., 2020) choice models for network growth, but these approaches do not reveal if or how the features of items in each choice set affect the preferences of choosers.

2.1 Discrete choice background

We briefly review the discrete choice modeling framework (see the book by Train (Train, 2009) for a thorough treatment). In a discrete choice setting, an individual selects one item from a set of available items, the choice set. We use $\mathcal{U}$ to denote the universe of all items and $C \subseteq \mathcal{U}$ the choice set in a particular choice instance. A choice dataset $\mathcal{D}$ is a set of n pairs (i, C), where $i \in C$ is the item selected. Each item i is described by a vector of d features $y_i \in \mathbb{R}^d$ that determine the preferences of the chooser.

A popular class of discrete choice models, random utility models (Marschak, 1959) (RUMs), are based on the idea that individuals try to maximize their utility, but can only do so noisily. In a RUM, an individual draws a random utility for each item (where each item has its own utility distribution) and selects the item with maximum observed utility. The workhorse RUM using item features is the conditional logit (CL) (McFadden, 1974), which has interpretable parameters that are readily estimated from data.[2] In the CL model, the observed utility of each item i is the random quantity $\theta^T y_i + \epsilon$, where the latent parameter $\theta \in \mathbb{R}^d$ (the *preference vector*) stores the relative importance of each feature (the *preference coefficients*) and the random noise term ϵ follows a standard Gumbel distribution with CDF $e^{-e^{-x}}$. This noise distribution is chosen so that the CL choice probabilities have a simple closed form (Train, 2009): a softmax over the utilities. Under a CL, the

[2]Many sources call this model the multinomial logit (e.g., Hausman and McFadden, 1984).

probability that i is chosen from the choice set C, denoted $\Pr(i \mid C)$, is

$$\Pr(i \mid C) = \frac{\exp(\theta^T y_i)}{\sum_{j \in C} \exp(\theta^T y_j)}. \tag{2.1}$$

The CL model famously obeys the axiom of *independence of irrelevant alternatives* (IIA) (Luce, 1959), stating that relative choice probabilities are unaffected by the choice set. Formally, a model satisfies IIA if for any two choice sets C, C' and items $i, j \in C \cap C'$,

$$\frac{\Pr(i \mid C)}{\Pr(j \mid C)} = \frac{\Pr(i \mid C')}{\Pr(j \mid C')}.$$

As we have discussed, this assumption is often violated in practice due to context effects. One model that can account for (some) context effects is the *mixed logit*. The DLCL model that we will introduce is related to a mixed logit, so we briefly describe it here. In a (discrete) mixed logit, there are M populations, each of which has its own preference vector θ_m. The mixing parameters $\pi_1, \dots, \pi_M$, with $\sum_{m=1}^{M} \pi_m = 1$, describe the relative sizes of the populations. This results in choice probabilities

$$\Pr(i \mid C) = \sum_{m=1}^{M} \pi_m \frac{\exp(\theta_m^T y_i)}{\sum_{j \in C} \exp(\theta_m^T y_j)}. \tag{2.2}$$

While mixed logit can produce IIA violations, it does so by hypothesizing populations each with their own context-effect-free preferences, meaning that context effects only appear in the aggregate data. In contrast, our models are designed to identify context effects in individual preferences.

2.2 Models of feature context effects

In order to capture context effects at the individual level, the choice set itself needs to influence the preferences of a chooser. In the most general extension of the CL, we could replace θ with $\theta + F(C)$, where $F\colon \mathcal{P}(\mathcal{U}) \to \mathbb{R}^d$ is an arbitrary function of the choice set (this is analogous to the set-dependent weights model (Rosenfeld et al., 2020), but framed as a RUM). This allows each feature to exert an arbitrary influence on the base preference coefficient of each other feature. We say that a *feature context effect* occurs when $F(C) \neq 0$.

We make two simplifying assumptions on the choice set effect function $F(C)$ that will aid interpretability. The first is that the effect of a choice set additively decomposes into effects of its items, i.e., $F(C)$ is proportional to $\sum_{j \in C} f(y_j)$ for some function $f\colon \mathcal{U} \to \mathbb{R}^d$. While in principle higher-order interactions are possible, the number of such interactions is exponential in the size of the choice set. This makes it difficult to extract such effects from typical choice datasets that do not contain observations from every possible choice set; moreover, higher-order interactions are usually sparse (Batsell and Polking, 1985). Second, we assume that the effect of each item is diluted in large choice sets and we model this with a proportionality constant of $1/|C|$ so that $F(C) = 1/|C| \sum_{j \in C} f(y_j)$.

2.2.1 Linear context logit

In principle, features could exert arbitrary influences on each other, but we focus on the case when context effects are linear, which makes inference tractable and, crucially, preserves interpretability. We use $y_C = 1/|C| \sum_{j \in C} y_j$ to denote the mean feature vector of the choice set C. For f linear, we can write $f(y_j) = Ay_j$ for some matrix $A \in \mathbb{R}^{d \times d}$, and the choice set context function F is

$$F(C) = \frac{1}{|C|} \sum_{j \in C} f(y_j) = \frac{1}{|C|} \sum_{j \in C} Ay_j = Ay_C.$$

We call this model the *linear context logit (LCL)*, and it produces choice probabilities

$$\Pr(i \mid C) = \frac{\exp([\theta + Ay_C]^T y_i)}{\sum_{j \in C} \exp([\theta + Ay_C]^T y_j)}. \tag{2.3}$$

In the LCL, A_{pq} specifies the effect of feature q on the coefficient of feature p. If A_{pq} is positive (resp. negative), then higher values of q in the choice set result in a higher (resp. lower) preference coefficient for p. If $A = 0$, then the LCL reduces to CL.

When analyzing data in Section 2.5, we often see large diagonal entries of A. The signs of the diagonal entries of A can be explained by known context effects. The case of $A_{pp} > 0$ relates to the attraction effect (high values of a feature amplify fine-grained differences along that dimension), and the case of $A_{pp} < 0$ is consistent with the similarity effect (high values of a feature devalue it),

Just as in the CL, we can derive the closed form in (2.3) if choosers have random utilities $[\theta + Ay_C]^T y_i + \epsilon$, where ϵ follows a standard Gumbel distribution and the random variable samples are i.i.d. If we want a more parsimonious model, we can impose sparsity on A through L_1 regularization (we do this in our empirical analysis) or we could use a low-rank approximation of A. A constant-rank approximation makes the number of parameters linear in d.

2.2.2 Decomposed linear context logit

The LCL implicitly assumes that the intercepts of all linear context effects exerted by one feature are the same (we have d^2 slopes in A, but only d intercepts in θ). Motivated by varying intercepts in empirical data (Figure 2.1), we develop a second model that decomposes the LCL into context effects exerted by each feature, which we call the *decomposed linear context logit (DLCL)*. In the language of choice set effect functions, we now have d context effect functions $F_1, \ldots, F_d$ where F_k only depends on the values of feature k. We also replace θ with d base preference vectors $B_1, \ldots, B_d$ (which we combine into a $d \times d$ matrix B; subscripts index columns) that provide varying intercepts. This gives us d contextual utilities $B_1 + F_1(C), \ldots, B_d + F_d(C)$ that we combine in a mixture model.

Making the same assumptions as for the LCL, we decompose each choice set effect function $F_k(C) = \frac{1}{|C|} \sum_{j \in C} f_k((y_j)_k)$ (here, f_k is a function of only the kth feature, $(y_j)_k$). Assuming linearity (and storing context effects exerted by feature k in the kth column of A), we arrive at

$$F_k(C) = \frac{1}{|C|} \sum_{j \in C} f_k((y_j)_k) = \frac{1}{|C|} \sum_{j \in C} A_k (y_j)_k = A_k (y_C)_k.$$

We use mixture weights $\pi_1, \ldots, \pi_d$ with $\sum_{k=1}^d \pi_k = 1$ to describe the relative strengths of effects exerted by each feature. The DLCL is then a mixture of d logits, where each component captures the context effects from a single feature. The choice probabilities are

$$\Pr(i \mid C) = \sum_{k=1}^{d} \pi_k \frac{\exp\left([B_k + A_k (y_C)_k]^T y_i\right)}{\sum_{j \in C} \exp\left([B_k + A_k (y_C)_k]^T y_j\right)}. \tag{2.4}$$

Each component corresponds to an LCL with the constraint that all columns of A except the kth are zero. The matrix A has the same interpretation as in the LCL, while B_{pq} represents the importance of feature p when feature q is zero (i.e., the intercept of the linear context effect exerted on p by q).

2.3 Identifiability of the LCL

Identifiability is a key feature of models that ensures we can uniquely learn parameters and thus interpret them meaningfully. We provide three results characterizing the identifiability of the LCL. Most significantly, we prove a necessary and sufficient condition that exactly determines when the model is identifiable (Theorem 1). However, the condition is somewhat hard to reason about, so we also prove a simple necessary condition (Proposition 1) and a simple sufficient condition (Proposition 2). These results give further insight into the main theorem. Proofs are in Appendix A.

Following Seshadri et al. (2019), we use $\mathcal{C}_\mathcal{D}$ to denote the set of unique choice sets appearing in the dataset $\mathcal{D}$, and we say that an LCL is *identifiable* from a dataset if there do not exist two distinct sets of parameters (θ, A) and (θ', A') that produce identical probability distributions over every choice set $C \in \mathcal{C}_\mathcal{D}$. In the following, $\otimes$ denotes the Kronecker product.

Theorem 1. *A d-feature linear context logit is identifiable from a dataset $\mathcal{D}$ if and only if*

$$\operatorname{span}\left\{ \begin{bmatrix} y_C \\ 1 \end{bmatrix} \otimes (y_i - y_C) \mid C \in \mathcal{C}_\mathcal{D}, i \in C \right\} = \mathbb{R}^{d^2+d}. \tag{2.5}$$

Theorem 1 says that identification requires enough choice sets with sufficiently different mean features containing enough sufficiently different items (with coupling between the two requirements). The condition of Theorem 1 is often satisfied in practice if there are no redundant features (18 out of 22 that we analyze uniquely identify the LCL).

To better understand the span condition, we provide a simple necessary condition for indentifiability. Recall that a set of vectors $\{y_0, \ldots, y_d\} \subset \mathbb{R}^d$ is *affinely independent* if the set of vectors $\{y_1 - y_0, \ldots, y_d - y_0\}$ is linearly independent.

Propostion 1. *If a d-feature linear context logit is uniquely identifiable from a dataset $\mathcal{D}$, then the dataset must contain $d+1$ choice sets with affinely independent mean feature vectors.*

This necessary condition stems from formulating item utility as the affine transformation $\theta + Ay_C$, which requires $d+1$ points to be identified. The span condition in Theorem 1 is more difficult to reason about because of the coupling between individual feature vectors y_i and mean feature vectors y_C. We therefore provide a simple sufficient condition for identifiability that decouples these requirements and is optimal in the number of distinct choice sets.

Propostion 2. *If a dataset $\mathcal{D}$ contains $d+1$ distinct choice sets $C_0, \ldots, C_d$ such that*

i. the set of mean feature vectors $\{y_{C_0}, \ldots, y_{C_d}\}$ is affinely independent (the necessary condition from Proposition 1) and

ii. in each choice set C_i, there is some set of $d+1$ items with affinely independent features,

then we can uniquely identify a d-feature LCL.

We leave characterization of DLCL identifiability for future work, as even mixed logits have notoriously complex identifiability conditions (Grün and Leisch, 2008; Zhao et al.; Chierichetti et al., 2018b).

2.4 Estimation

Given a dataset $\mathcal{D}$ consisting of observations (i, C), where i was selected from the choice set C, we wish to recover the parameters of a model that best describe the dataset. In this section, we describe estimation procedures for the LCL and DLCL. First, we show that the likelihood function of the LCL is log-concave and simple to optimize. On the other hand, the DLCL does not have a log-concave likelihood, but we derive an expectation-maximization algorithm that only requires optimizing convex subproblems.

We wish to find parameters that minimize the negative log-liklihood (NLL) of a model, which is equivalent to maximizing the likelihood. The NLL of the linear context logit is

$$-\ell(\theta, A; \mathcal{D}) = -\sum_{(i,C)\in\mathcal{D}} \log \frac{\exp([\theta + Ay_C]^T y_i)}{\sum_{j\in C}\exp([\theta + Ay_C]^T y_j)} \tag{2.6}$$

$$= \sum_{(i,C)\in\mathcal{D}} -(\theta + Ay_C)^T y_i + \log \sum_{j\in C} \exp([\theta + Ay_C]^T y_j). \tag{2.7}$$

This function is convex in θ and A (equivalently, the likelihood is log-concave). To see this, notice that the first term in the summand of (2.7) is a linear combination of entries of θ and A, so it is jointly convex in θ and A. Meanwhile, log-sum-exp is convex and monotonically increasing, so its composition with the linear functions $[\theta + Ay_C]^T y_j$ is also convex. We then have that $-\ell(\theta, A; \mathcal{D})$ is convex, as the sum of convex functions is convex. Moreover, the second partial derivatives of the NLL function are all bounded (by a constant depending on the dataset), so its gradient is Lipschitz continuous. We can therefore use gradient descent to efficiently find a global optimum of $-\ell(\theta, A; \mathcal{D})$.

On the other hand, the NLL of the DLCL (like that of the mixed logit) is not convex, so we can only hope to find a local optimum with gradient descent. To address this challenge, we develop an expectation-maximization (EM) algorithm for DLCL estimation. The algorithm mirrors the EM algorithm for estimating a mixed logit (Train, 2009), except that the M step updates estimates for A and B. (see Appendix A.2). An advantage of EM for DLCL is that it only requires optimizing convex functions with Lipschitz-continuous gradients, and EM is guaranteed to improve the log-likelihood at each step. While EM may still arrive at a local optimum, we find that for most of our datasets, it finds better model parameters than stochastic gradient descent on the likelihood.

2.5 Data analysis

We apply our LCL and DLCL models to two collections of empirical choice datasets. First, we examine datasets specifically collected to understand preference in various domains, such as car purchasing and hotel booking. The features describing items naturally differ in these datasets. The second collection of datasets comes from a particular choice process in social networks, namely the formation of new connections. Here, we use graph properties as features (such as in-degree, a proxy for popularity (Moody et al., 2011)), allowing us to compare social dynamics across email, SMS, trust, and comment networks. In both dataset collections, we first establish that context effects occur and that our models better describe the data than traditional context-effect-free models, CL, and mixed logit. We then show how the learned models can be interpreted to recover intuitive feature context effects. Our code, results, and links to documented versions of every dataset are available at `https://github.com/tomlinsonk/feature-context-effects`.

2.5.1 Estimation details

For prediction experiments, we use 60% of samples for training, 20% for validation, and 20% for testing. When testing model fit with likelihood-ratio tests, we estimate models from the entire dataset. We use PyTorch's Adam optimizer for maximum likelihood estimation, with batch size 128 and the `amsgrad` flag. We run the optimizer for 500 epochs or 1 hour, whichever comes first. For the whole-data fits, we use weight decay 0.001 and search over learning rates of $0.0005, 0.001, 0.005, 0.01, 0.05, 0.1$, selecting the one that results in the highest likelihood. For our prediction experiments, we perform a grid search over the weight decays $0, 0.0001, 0.0005, 0.001, 0.005, 0.01$ and the same learning rates as above, selecting the pair with the best likelihood on the validation set. Predictions are evaluated on the held-out test set. We use d (the number of features) components for mixed logit to provide a fair comparison against DLCL (which always uses d components).

2.5.2 General choice datasets

We analyze six choice datasets from online and survey data (Table 2.1) previously used in discrete choice research: SUSHI (Kamishima, 2003); EXPEDIA (Kaggle, 2013a); DISTRICT and DISTRICT-SMART (Kaufman et al., 2017; Bower and Balzano, 2020); CAR-A and CAR-B (Abbasnejad et al., 2013); and CAR-ALT (Brownstone et al., 1996; McFadden and Train, 2000). In all datasets, we standardize the features to have zero mean and unit variance, which allows us to more meaningfully compare learned parameters across datasets. The LCL is identifiable in DISTRICT-SMART, EXPEDIA, and SUSHI, but not the others. However,

Table 2.1: General choice datasets summary.

Dataset	Choices	Features	Largest Choice Set
DISTRICT	5376	27	2
DISTRICT-SMART	5376	6	2
SUSHI	5000	6	10
EXPEDIA	276593	5	38
CAR-A	2675	4	2
CAR-B	2206	5	2
CAR-ALT	4654	21	6

the L_2 regularization we apply (via weight decay) identifies the model in all cases.

2.5.3 Network datasets

Recent work cast many network growth models in terms of discrete choice (Overgoor et al., 2019). In a directed graph, the formation of the edge $u \to v$ can be thought of as a choice by the node u to initiate new contact with v (the graph might be a citation, communication, or friendship network). The set from which u chooses can vary, including all nodes in the graph or only a subset of "close" nodes. We focus specifically on directed *triadic closure* (Granovetter, 1973; Easley and Kleinberg, 2010), where the node u closes a triangle $u \to v \to w$ by adding the edge $u \to w$. This phenomenon is used in many influential network growth models (Jin et al., 2001; Holme and Kim, 2002; Vázquez, 2003) and real-world networks show evidence of triadic closure in the form of high clustering coefficients (Easley and Kleinberg, 2010) and closure coefficients (Yin et al., 2020).

Choices from temporal network data

Our network analysis assumes that the graphs grow according to a multi-mode model that combines triadic closure with a method of global edge formation. In particular, we assume that at each step, an initiating node decides to either form an edge to any node in the graph with probability r or close a triangle with probability $1 - r$. This setup, also used by the Jackson–Rogers model (Jackson and Rogers, 2007) and the (r, p)-model (Overgoor et al., 2019), singles out instances of triadic closure to study separately from global edge formation. When a node u chooses to close a triangle, we assume u first picks one of its neighbors v uniformly at random before choosing one of v's neighbors as a new connection.

Each time we observe a new edge $u \to w$ closing a previously unclosed triangle, we select a hypothesized intermediate v uniformly at random ($u \to w$ can close multiple triangles at once through different intermediates). We consider the choice set for the closure to be the out-neighbors of v that are not out-neighbors of u.

Node features

The features of each node in the choice set are computed at the instant before the edge is closed (the features evolve as the network grows). In our datasets, we have timestamps on each edge and an edge may be observed many times (e.g., in an email network, u may send w many emails). The number of times an edge is observed is its *weight*; an edge not in the graph has weight 0. We use six features to describe each node w that could be selected by the chooser u: (1) *in-degree*: the number of edges entering the target node w; (2) *shared neighbors*: the number of in- or out-neighbors of u that are also in- or out-neighbors of w; (3) *reciprocal*

Table 2.2: Network datasets summary.

Dataset	Nodes	Edges	Triangle closures
SYNTHETIC-CL	1000	391294	50000
SYNTHETIC-LCL	1000	380584	50000
EMAIL-ENRON	18592	53477	19900
EMAIL-EU	986	24929	19603
EMAIL-W3C	20082	33409	3271
SMS-A	44430	68834	6311
SMS-B	72146	100974	9376
SMS-C	14433	23285	2732
BITCOIN-ALPHA	3783	24186	8823
BITCOIN-OTC	5881	35592	12750
REDDIT-HYPERLINK	23499	91946	37115
WIKI-TALK	22067	81125	27505
FACEBOOK-WALL	46952	274086	68776
MATHOVERFLOW	24818	239978	137455
COLLEGE-MSG	1899	20296	6267

weight: the weight of the reverse edge $u \leftarrow w$; (4) *send recency*: the number of seconds since w initiated any outgoing edge; (5) *receive recency*: the number of seconds since w received any incoming edge; (6) *reciprocal recency*: the number of seconds since the reverse edge $u \leftarrow w$ was last observed.

Following Overgoor et al. (2020), we log-transform features 1 and 2. We take $\log(1 + \text{feature } 3)$ to handle weight 0 (in-degree and shared neighbors are never 0, since v is always a shared neighbor of u and w). Lastly, we transform the temporal features with $\log^{-1}(2 + \text{feature})$ and set them to 0 if the event has never occurred. This ensures that (1) we can handle 0 seconds since the last event, (2) higher values mean more recency, and (3) "no occurrence" results in the lowest possible value of the transformed feature.

Network datasets

We examine 13 network datasets: three email datasets (EMAIL-ENRON (Benson et al., 2018a), EMAIL-EU (Leskovec et al., 2007; Yin et al., 2017), EMAIL-W3C (Craswell et al., 2005; Benson and Kleinberg, 2018)); three SMS datasets (SMS-A, SMS-B, and SMS-C (Wu et al., 2010)), two Bitcoin trust datasets (BITCOIN-ALPHA and BITCOIN-OTC (Kumar et al., 2016, 2018b)), an online messaging dataset (COLLEGE-MSG (Panzarasa et al., 2009)), a hyperlink dataset (REDDIT-HYPERLINK (Kumar et al., 2018a)), and three online forum datasets (FACEBOOK-WALL (Viswanath et al., 2009), MATHOVERFLOW (Paranjape et al., 2017), and WIKI-TALK (Leskovec et al., 2010b,a)). In addition, we generate two synthetic networks, SYNTHETIC-CL and SYNTHETIC-LCL. Specifically, we begin with 1000 isolated nodes. At each step, we add an edge uniformly at random with probability 0.9. With probability 0.1, we close a triangle by selecting a node u and one of its neighbors v uniformly at random. We then use either a CL (for SYNTHETIC-CL) or LCL (for SYNTHETIC-LCL) to choose which triangle $u \to v \to ?$ to close (if there are no triangles for u to close, we add a random edge). We use the same features as in the empirical datasets, with Poisson-distributed simulated timestamp gaps between successive edges until 50000 triangles are closed.

Table 2.2 summarizes the network data. Using Theorem 1, we find that the LCL is uniquely identifiable in every network dataset. Whereas we split the general choice datasets into training, validation, and testing sets uniformly at random, we instead split the network datasets temporally so that future edges are predicted based on parameters estimated from past edges.

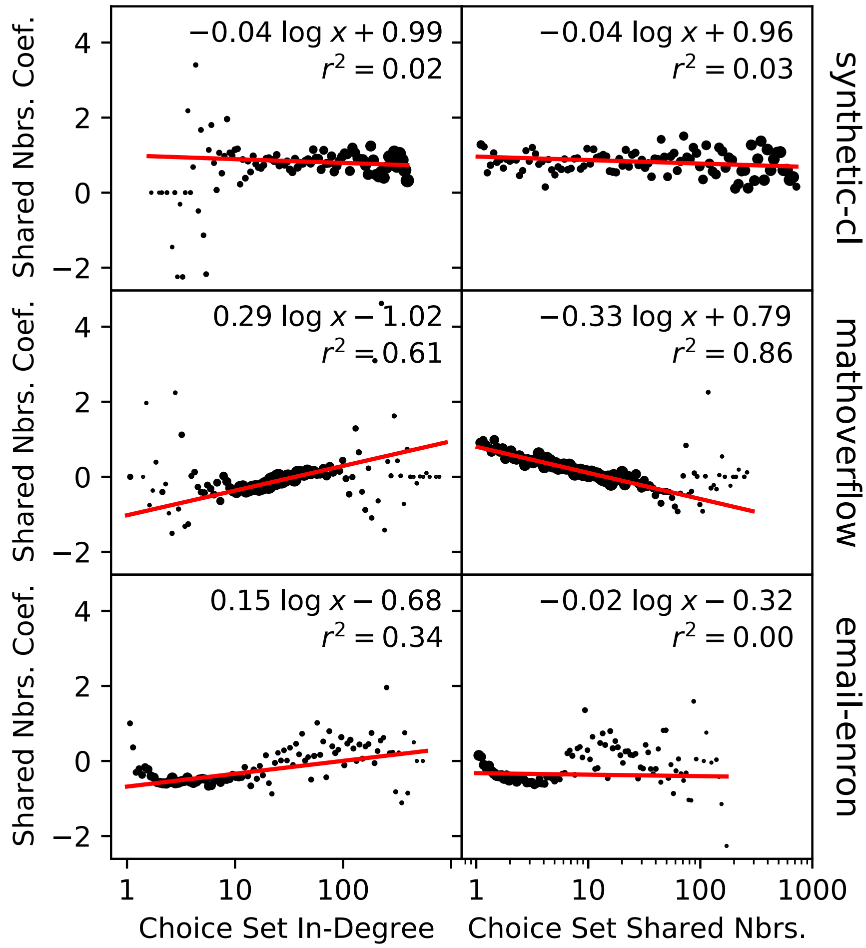


Figure 2.1: Learned preference coefficients of CLs trained on samples binned by mean choice set feature. Each point shows the preference coefficient of the shared neighbors feature in choice sets with varying mean in-degree (left column) and shared neighbor counts (right column). These coefficients were found by splitting observations into 100 bins according to their mean feature values and learning a CL for each bin separately. The area of each point is proportional to the square root of the number of observations in its bin. The red lines are weighted least squares fits.

2.5.4 Results

Our analysis focuses on two issues: whether significant linear feature context effects appear in practice and if so, how we can identify and interpret them using our models.

Binned CLs for visualizing feature context effects

As a first step towards identifying whether linear context effects occur, we bin the samples of each dataset according to the mean feature values in the choice set. We then fit CLs within each bin, examining whether the preference coefficients of features vary with the mean choice set features. Figure 2.1 shows two clear linear (with the respect to the log-transformed feature) context effects in MATHOVERFLOW: (1) as the mean in-degree of the choice set increases, so does the shared neighbors preference coefficient and (2) the shared neighbors coefficient decreases in choice sets with higher mean shared neighbors. Colloquially, (1) close ties are a stronger predictor of new connections when selecting between a set of popular individuals and (2) common connections matter less when choosing from a closely connected group. The different intercepts of these two effects in MATHOVERFLOW also motivate decomposing the LCL to the DLCL. The figure also shows some evidence of non-linear context effects in EMAIL-ENRON, which is of interest for future work.

Evaluating model fit

With our evidence that context effects are worth capturing, we compare our LCL and DLCL models to the traditional choice models they subsume (CL and mixed logit) with likelihood-ratio tests. To correct for multiple hypotheses, we use $p < 0.001$ as our significance threshold.

Table 2.3 shows the total NLL of every dataset under the four models, along with markers indicating the significant likelihood-ratio tests. In the empirical network datasets, all likelihood-ratio tests are significant (all with $p < 10^{-9}$),

Table 2.3: Dataset negative log-likelihoods. Bolded entries indicate the highest likelihood for a dataset.

	CL	LCL	Mixed logit	DLCL
DISTRICT	3313	**3130**	3258	3206
DISTRICT-SMART	3426	**3278***	3351	3303†
EXPEDIA	839505	837649*	839055	**837569**†
SUSHI	9821	9773*	9793	**9764**
CAR-A	1702	1694	1696	**1692**
CAR-B	1305	1295	1297	**1284**
CAR-ALT	7393	**6733***	7301	7011†
SYNTHETIC-CL	**210473**	210486	210503	210504
SYNTHETIC-LCL	140279	**137232***	139539	137937†
WIKI-TALK	99608	97748*	95761	**95134**†
REDDIT-HYPERLINK	135108	132880*	133766	**132473**†
BITCOIN-ALPHA	19675	19190*	19093	**18877**†
BITCOIN-OTC	26968	26101*	25768	**25348**†
SMS-A	8252	**8056***	8239	8154†
SMS-B	13153	**12823***	13147	12975†
SMS-C	4988	4880*	4928	**4871**†
EMAIL-ENRON	73015	70061*	71450	**69254**†
EMAIL-EU	53025	51822*	51988	**51431**†
EMAIL-W3C	11012	10677*	9898	**9758**†
FACEBOOK-WALL	118208	**116062***	117210	116328†
COLLEGE-MSG	14575	14120*	13849	**13712**†
MATHOVERFLOW	500537	479999*	440482	**435932**†

*Significant likelihood-ratio test vs. CL ($p < 0.001$)
†Significant likelihood-ratio test vs. mixed logit ($p < 0.001$)

indicating that feature context effects are occurring. In the general choice datasets, EXPEDIA ($p < 10^{-16}$), DISTRICT-SMART ($p < 10^{-16}$), SUSHI ($p = 1.6 \times 10^{-7}$), and CAR-ALT ($p < 10^{-16}$) have significant tests for the LCL.

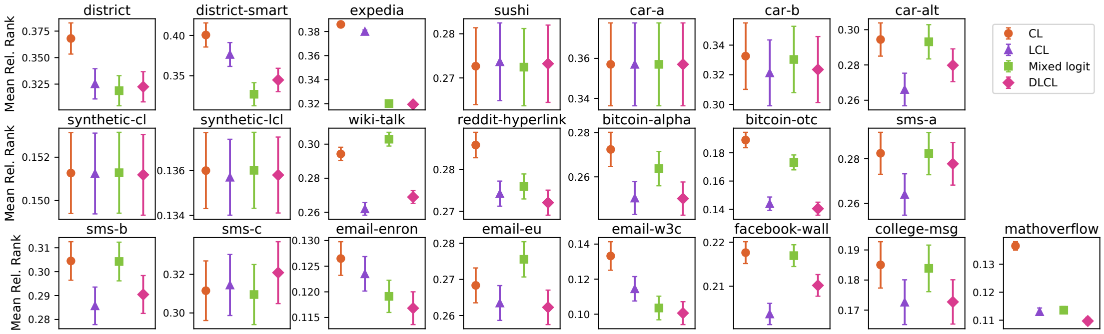


Figure 2.2: Mean relative rank of predictions on held-out test data (lower is better). Error bars show standard error of the mean.

Evaluating predictive power

The likelihood-ratio tests provide strong evidence for the presence of feature context effects. A related question is whether our methods improve out-of-sample predictions. To address this question, we measure the mean relative rank of the true selected item in the output ranking of each method. We define the *relative rank* of an item i to be its index when the choice set C is sorted in descending probability order (with ties resolved by taking the mean of all possible indices), divided by $|C| - 1$. The mean relative rank is a measure of how good the model's predictions are, from 0 (best) to 1 (worst). We use this rather than mean reciprocal rank because the choice sets have variable sizes (Fuhr, 2018). Figure 2.2 shows that LCL and DLCL make better predictions than CL and mixed logit across many datasets. In some cases, the improvement are quite large; for example, in BITCOIN-OTC, the mean relative rank is 24% better in LCL than in CL.

Interpreting learned models on general choice datasets

The previous analyses of model fit and predictive power indicate that linear context effects are indeed a significant factor. We now investigate what these effects are

Table 2.4: Five largest context effects in SUSHI.

Effect (q **on** p)	A_{pq} (std. err.)	p-value
popularity on *popularity*	−0.28 (0.15)	0.066
availability on *is maki*	0.24 (0.14)	0.087
oiliness on *oiliness*	−0.20 (0.08)	0.0089
popularity on *availability*	0.19 (0.14)	0.16
availability on *oiliness*	−0.18 (0.10)	0.064

and show how our models can be interpreted to discover choice behaviors. We focus on the LCL because of its simpler structure and convex objective. For the general choice datasets, we select two datasets for detailed examination: EXPEDIA and SUSHI. The five context effects with largest magnitude in each dataset are shown in Tables 2.4 and 2.5. Note that features are all standardized, so picking the largest entries of A is meaningful.

Using the asymptotic normality of the maximum likelihood estimator (Wasserman, 2013), we can compute standard errors for the parameter estimates and p-values for the null hypothesis that a particular context effect is zero. This procedure is inexpensive: we need a single pass over the dataset after training to estimate the Fisher information matrix, from which standard errors can be directly computed (this is a standard procedure in statistical inference (Wasserman, 2013)).[3]

First, we examine SUSHI, which has randomly chosen choice sets. The most significant effect is that respondents given more oily sushi options showed more aversion to oily sushi (Table 2.4). The randomization of choice sets allows us to hypothesize that this is causal: too much oiliness on the menu makes oily foods less appealing, which could be an example of the similarity effect. The other context

[3]Another useful (but more computationally expensive) approach is to constrain A to zero in all but one entry of interest. This preserves NLL convexity, still allows for likelihood ratio tests, and can be used to determine the effect size of a context effect.

Table 2.5: Five largest context effects in EXPEDIA.

Effect (q on p)	A_{pq} (std. err.)	p-value
location score on *price*	-0.47 (0.05)	$< 10^{-16}$
on promotion on *price*	0.27 (0.03)	$< 10^{-16}$
review score on *price*	-0.19 (0.03)	1.4×10^{-9}
star rating on *price*	0.15 (0.04)	6.7×10^{-5}
price on *star rating*	0.10 (0.00)	$< 10^{-16}$

effects with largest magnitude in A are not significant.

In EXPEDIA, all five of the largest-magnitude effects are statistically significant (Table 2.5). The largest effect in the full model is a decrease in willingness to pay (i.e., cheaper options are more preferred) when the mean location score of the choice set is high. Additionally, if many of the options are marked as "on promotion," people seem more willing to book higher priced hotels. Interestingly, when the available hotels tend to be well-reviewed by other Expedia users, people are more price-averse, but they are less price-averse when the available hotels tend to have high star ratings. This may be because people searching for five-star hotels are not looking for the cheapest options, whereas people searching for well-reviewed hotels are looking for good deals. (The dataset does not include this information, only the location, length of stay, booking window, adult/children count, and room count of the search.) Finally, people choosing between more expensive hotels placed more weight on high star rating. When interpreting these effects, it is important to keep in mind that the choice sets in EXPEDIA may be influenced by user preferences to begin with, so we cannot determine whether the effects are causal. Nonetheless, the learned LCL model could motivate a randomized controlled trial aimed at determining causal effects. It also illustrates an important point to keep in mind when using choice data from recommender systems: choice sets are not necessarily independent from preferences.

Interpreting learned models on network growth datasets

We take a different approach to examine context effects in the network datasets, showcasing another useful application of the LCL. To visualize what context effects influence choice in the network datasets, we apply L_1 regularization of varying strength to the LCL matrix A during training, which encourages sparsity. Figure 2.3 visualizes the learned A matrices. Recall that a column of A corresponds to the feature exerting an effect and a row to the influenced feature.

Figure 2.3 reveals several effects shared by multiple datasets. For example, in MATHOVERFLOW, FACEBOOK-WALL, SMS-A, SMS-B, and REDDIT-HYPERLINK, feature 1 (in-degree) has a positive effect on feature 2's coefficient (shared neighbors). This suggests that close connections matter more when choosing from a popular group. And in EMAIL-ENRON and EMAIL-W3C, there is a negative effect of feature 6 (reciprocal recency) on feature 1 (in-degree): high-volume email recipients are less likely to be targeted when the sender's inbox has recent messages from other potential targets.

In both of these examples, when increasing regularization causes those entries of A to go to 0, we see a jump in the likelihood, indicating that these are important effects to capture (note that we plot NLL, so lower is better). Additionally, we see in the top row how a dataset with no context effects (SYNTHETIC-CL) behaves: A immediately goes to 0 when any L_1 regularization is applied, without any worsening of the likelihood.

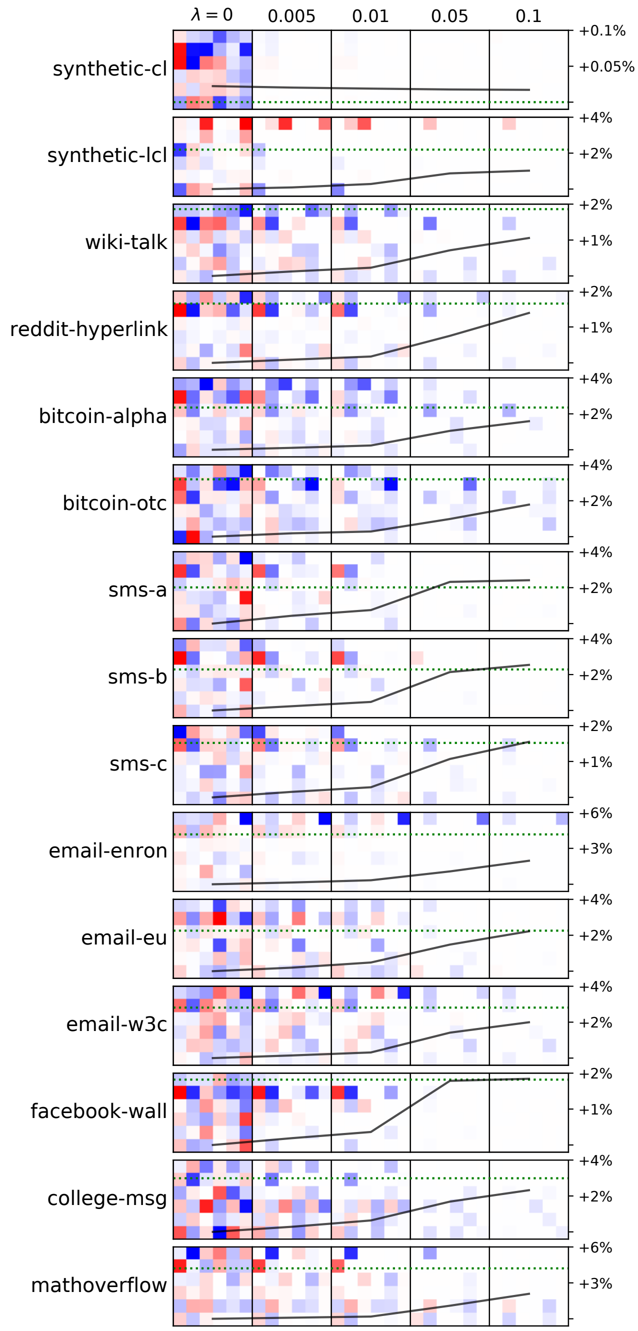


Figure 2.3: Effect of L_1 regularization on the LCL context effect matrix. The parameter λ (increasing left to right) controls the strength of regularization. Each box visualizes the learned matrix A (blue = negative, red = positive, white = zero; consistent color scales within but not between rows) at the given λ value. Features in A are in the order from Section 2.5.3, top-down and left-right. The black line tracks the total NLL of the LCL (the % on the y-axes is relative to the NLL of the best model plotted for that dataset). The dotted green line is the significance threshold of a likelihood-ratio test against a CL ($p < 0.001$; black line below the threshold means the LCL is a significantly better fit than CL).

2.6 Discussion

Discovering intuitive context effects from choice data using our models has a number of potential applications. In recommender systems, insight into context effects could inform the set of options suggested to the user. Our models can also produce hypotheses for more controlled investigation in economics or psychology. A key contribution is showing how intuitive and general context effects can be automatically recovered from observed choices and tested for significance. While we focused on linear context effects for simplicity, some datasets (e.g., EMAIL-ENRON in Figure 2.1) show evidence of non-linearity. Capturing these more complex effects while retaining ease of training and interpretation would be valuable.

Our network analysis revealed several context effects in network growth, which can aid modeling within network science and social network analysis. We focused on triadic closure, where context effects can be observed in small choice sets. Incorporating context effects in other modes of network growth (such as connections with unrelated nodes) is an interesting avenue for future research. A challenge is that global modes of edge formation have large choice sets, requiring negative sampling for effective estimation (Overgoor et al., 2020), which seems difficult to adapt for models with context effects.

A limitation of our approach is that the generalizability of identified effects is constrained by correlations in the data. For example, choice sets arising from recommender systems (such as EXPEDIA) are correlated with the preferences of their users by design. This makes it difficult to distinguish between how a user's preferences are affected by the choice set and how the user's preferences influence the choice set. In the following chapter, we adapt causal inference methods to the discrete choice setting to address this issue.

In other situations, we might have random choice sets (as in SUSHI) or we might have no information about how choice sets are determined. In the latter case, our approach could also be used to find evidence of choice sets targeted at chooser preferences: if we observe many positive self-effects (i.e., preference for star-rating is higher in sets with high star-rating), this could mean that choice sets are being catered to people's preferences. In some cases, this could be undesirable (e.g., if the party presenting individuals with options is supposed to be impartial), and our methods could provide a mechanism for identifying unwanted interventions.

Another challenging direction for future work would be a method of discovering more complex relational context effects from choice data. The feature context effects we study describe the influence of one feature on another, but some of the traditional context effects studied in economics and psychology (e.g., the compromise effect) are based on the relationship between the features of several items. These effects are typically studied with targeted models that are hand-crafted to capture the desired effect. A general method of encoding and learning relational context effects could enable the discovery of new complex effects not yet envisioned by choice theorists, but that nonetheless appear in choice data.

CHAPTER 3

CHOICE SET CONFOUNDING IN DISCRETE CHOICE

In this chapter, we turn to the issue of identifying whether context effects are *causal* rather than merely *correlational*. For instance, does observing well-reviewed hotels in the EXPEDIA dataset actually cause choosers to prefer less expensive hotels, or is there another explanation for the significant context effect we observed in Table 2.5?

As discussed in the previous chapter, machine learning approaches have enabled more accurate choice modeling and prediction (Seshadri et al., 2019; Rosenfeld et al., 2020; Bower and Balzano, 2020). However, observational choice data analysis has thus far overlooked a crucial fact: the *choice set assignment mechanism* underlying a dataset can have a significant impact on the generalization of learned choice models, in particular their validity on *counterfactuals*. Understanding how new choice sets affect preferences in such counterfactuals is key to many applications, such as determining which alternative-fuel vehicles to subsidize or which movies to recommend. In particular, chooser-dependent choice set assignment coupled with heterogeneous preferences can severely mislead choice models, as they do not model the influence of preferences on choice set assignment. Recommender systems are one extreme case, where items are selected specifically to appeal to a user. Such situations also arise in transportation decisions, online shopping, and personalized Web search, resulting in widespread (but often invisible) error in choice models learned from this data.

Drawing on connections with causal inference (Imbens and Rubin, 2015), we term the issue of chooser-dependent choice set assignment *choice set confounding*. Choice set confounding is a major issue for recent machine learning methods whose success is due to capturing context effects. While context effects are widespread and worth capturing (e.g., Huber et al., 1982; Simonson, 1989), choice set confounding can result in spurious effects and over-fitting, and it is unclear if recent machine learning models are learning true effects or simply being misled by chooser-dependent choice set assignment.

In this chapter, we formalize when choice set confounding is an issue and show that it can result in arbitrary systems of choice probabilities, even if choosers are rational utility-maximizers (in contrast, tractable choice models only describe a tiny fraction of possible choice systems). We also provide strong evidence of choice set confounding in two transportation datasets commonly used to demonstrate the presence of context effects and to test new models (Koppelman and Bhat, 2006; Seshadri et al., 2019; Ragain and Ugander, 2016; Benson et al., 2016). Then, to manage choice set confounding, we first adapt two causal inference methods—inverse probability weighting (IPW) and regression controls—to train choice models in the presence of confounding. These methods require chooser covariates satisfying certain assumptions that differ from the traditional causal inference setting. For instance, given access to the same covariates used by a recommender system to construct choice sets, we can reweight the dataset to learn a choice model as if choice sets had been user-independent. Alternatively, we can incorporate covariates into the choice model itself, recovering individual preferences as long as those covariates capture preference heterogeneity.

We also show how to manage choice set confounding without such covariates, as many observational datasets have little information about the individuals making choices. We demonstrate a link between models accounting for context effects and models for choice systems induced by choice set confounding. For example, we derive the context-dependent random utility model (CDM) (Seshadri et al., 2020) from the perspective of choice set confounding, by treating the choice set as a vector of substitute covariates (e.g., "someone who is offered item i").

We develop spectral clustering methods typically used for co-clustering (Dhillon, 2001) that exploit choice set assignment as a signal for chooser preferences, as a way to improve counterfactual predictions for observed choosers. To show why and when this can work, we frame the problem of finding sufficient chooser covariates as a problem of recovering latent cluster membership in a stochastic block model (SBM) of the bipartite graph that connects choosers to the items in their choice sets.

In addition to theoretical analysis, we demonstrate the efficacy of our methods on real-world choice data. We provide evidence that IPW reduces confounding when modeling hotel booking data, making the choice system more consistent with utility-maximization and making inferred parameters more plausible. For example, the confounded data overweights the importance of price, since many users are shown hotels matching their preferences and select the cheapest one. Factors such as star rating would play a more important role in counterfactuals. We also evaluate our clustering approach on online shopping data. By training separate models for different chooser clusters, we outperform a mixture model that attempts to discover preference heterogeneity from choices alone, ignoring the signal from choice set assignment.

All of the code, results, and links to the data used in this chapter are available at `https://github.com/tomlinsonk/choice-set-confounding`.

Additional related work

This research is inspired by recent computational advances in learning context-dependent preferences (Seshadri et al., 2019; Rosenfeld et al., 2020; Bower and Balzano, 2020; Tomlinson and Benson, 2021; Pfannschmidt et al., 2022), including the LCL. These methods exhibit strong gains by exploiting context effects but are often evaluated on data with possible choice set confounding. Similar confounding issues are well-studied in rating and ranking data within recommender systems (Marlin et al., 2007; Schnabel et al., 2016; Wang et al., 2020, 2019), but these approaches do not directly apply to choice data. The causal inference ideas that we develop are based on long-standing methods (Imbens, 2004; Imbens and Rubin, 2015); the challenge we address is how to adapt them for discrete choice data.

The role of choice set assignment does occasionally appear in the choice literature. For instance, Manski (1977) used choice set assignment probabilities to derive random utility models (Manski, 1977). More often, traditional choice theory has focused on latent *consideration sets*, which are subsets of alternatives that are actually considered by choosers (Ben-Akiva and Boccara, 1995; Bierlaire et al., 2010) where non-uniform choice set probabilities play a key role. In another setting, Manski and Lerman (1977) used an approach similar to our inverse probability weighting. They were concerned with "choice-based samples," where we first sample an item and then get an observation of a chooser who selected that item (usually, we sample a chooser and then observe their choice).

The use of regression controls in discrete choice (i.e., including chooser covariates in the utility function) is standard in econometrics (Stratton et al., 2008; Bhat and Gossen, 2004; Train, 2009). However, in these settings, regression aims to understand how the attributes of an individual affect decision-making, which can unknowingly and accidentally help with confounding. This may explain why choice set confounding has not been widely recognized (additionally, in an interview, Manksi discusses that choice set generation has been under-explored (Tamer, 2019)). We formalize when and how regression adjusts for choice set confounding.

3.1 Discrete choice background

We will need some additional notation and definitions beyond what we already introduced in Chapter 2. For clarity, we also remind the reader of the basics. Let $\mathcal{U}$ denote a universe of n items and $\mathcal{A}$ a population of individuals. In a discrete choice setting, a *chooser* $a \in \mathcal{A}$ is presented a nonempty *choice set* $C \subseteq \mathcal{U}$ and they choose one item $i \in C$. Specifically, a is sampled with probability $\Pr(a)$, then C is presented to a with probability $\Pr(C \mid a)$, and finally a selects i with probability $\Pr(i \mid a, C)$. Most discrete choice analysis focuses only on $\Pr(i \mid a, C)$ or $\Pr(i \mid C)$—for instance, in Chapter 2, we only dealt with $\Pr(i \mid C)$—here we consider this entire process. A discrete choice dataset $\mathcal{D}$ is a collection of tuples (C, i) generated by this process. We use $\mathcal{C}_\mathcal{D}$ to denote the set of unique choice sets in $\mathcal{D}$.

Discrete choice models posit a parametric form for choice probabilities, with parameters learned from data. The *universal logit* (McFadden et al., 1977) can express any system of choice probabilities (called a *choice system*). Under a universal logit, each chooser a has a scalar utility $u_i(C, a)$ for item i in choice set C. Choice probabilities are then a softmax of these utilities:

$$\Pr(i \mid a, C) = \frac{\exp(u_i(C, a))}{\sum_{j \in C} \exp(u_j(C, a))}.$$

As we discussed in the context of conditional logit, the softmax arises from a notion of rational utility-maximization (Train, 2009): these are the choice probabilities if a observes random utilities $u_i(C, a) + \epsilon$ (where the ϵ are i.i.d. Gumbel-distributed for each item and choice) and selects the item with maximum observed utility. The above model has too many degrees of freedom to be practical (e.g., it has entirely separate parameters for every chooser a), and typically one assumes utilities are fixed across sets and individuals. This is what we will call the *logit* model (McFadden, 1974), where $u_i(C, a) = u_i, \forall C, \forall a$.

Other discrete choice models come from different assumptions on $u_i(C, a)$, trading off descriptive power for ease of inference and interpretation. For example, we may have access to a vector of covariates $\boldsymbol{x_a} \in \mathbb{R}^{d_x}$ for person a. Similarly, an item i may be described by a vector of features $\boldsymbol{y_i} \in \mathbb{R}^{d_y}$. We can write $u_i(C, a)$ as a function of $\boldsymbol{x_a}, \boldsymbol{y_i}$, or both, yielding several choice models (Table 3.1), which we will refer to as the multinomial logit (MNL), conditional logit (CL), and conditional multinomial logit (CML).[1] All of these models obey a common assumption, the *independence of irrelevant alternatives* (IIA) (Train, 2009). IIA states that relative choice probabilities are conserved across choice sets:

$$\frac{\Pr(i \mid a, C)}{\Pr(j \mid a, C)} = \frac{\Pr(i \mid a, C')}{\Pr(j \mid a, C')}.$$

To be precise, this is individual-level rather than group-level IIA, which was what we stated in Chapter 2. Among the models in Table 3.1, the latter is only obeyed by the logit and conditional logit. In general, models obey individual-level IIA

[1]"MNL" and "CL" are both sometimes used as blanket terms encompassing all of these models, sometimes even by the same author in different papers (McFadden, 1974; Hausman and McFadden, 1984). As it will be useful to easily distinguish between them, we follow the convention (Hoffman and Duncan, 1988) that "multinomial" means chooser covariates are used and "conditional" means item features are used. Additionally, for CML, we assume $\boldsymbol{\gamma_i} = B^T \boldsymbol{x_a}$, which reduces the number of parameters from $d_y + nd_x$ to $d_y(d_x + 1)$, allowing us to use the model when the number of items is prohibitively large.

Table 3.1: Discrete choice models. The item and chooser feature vectors $\boldsymbol{y_i}$ and $\boldsymbol{x_a}$ are part of the dataset, while $u_i \in \mathbb{R}, \boldsymbol{\theta} \in \mathbb{R}^{d_y}, \boldsymbol{\gamma_i} \in \mathbb{R}^{d_x}$, and $B \in \mathbb{R}^{d_y \times d_x}$ are learned parameters.

Model	$u_i(C, a)$	# **Parameters**
logit	u_i	n
multinomial logit (MNL)	$u_i + \boldsymbol{x_a}^T \boldsymbol{\gamma_i}$	$n(d_x + 1)$
conditional logit (CL)	$\boldsymbol{y_i}^T \boldsymbol{\theta}$	d_y
conditional multinomial logit (CML)	$\boldsymbol{y_i}^T(\boldsymbol{\theta} + B\boldsymbol{x_a})$	$d_y(d_x + 1)$

if utility is independent of C, i.e., $u_i(C, a) = u_i(a)$ and obey group-level IIA if $u_i(C, a)$ is independent of both C and a.

While the IIA assumption is convenient, it is commonly violated through context effects (Huber et al., 1982; Simonson and Tversky, 1992; Benson et al., 2016). Due to the ubiquity of context effects, models incorporating information from the choice set have become increasingly popular and have shown considerable success (Seshadri et al., 2019; Rosenfeld et al., 2020; Bower and Balzano, 2020; Tomlinson and Benson, 2021). Other models allow IIA violations without explicitly modeling effects of the choice set (Ragain and Ugander, 2016; McFadden and Train, 2000; Benson et al., 2016).

We briefly introduce one of these context effect models, the *context-dependent random utility model* (CDM) (Seshadri et al., 2019), and review the LCL from Chapter 2. In the CDM, each item in the choice set exerts a pull on the utility of every other item: $u_i(C, a) = \sum_{j \in C \setminus i} p_{ij}$. The CDM can be derived as a second-order approximation to universal logit (where plain logit is the first-order approximation) (Seshadri et al., 2019). The LCL instead operates in settings with item features, adjusting the conditional logit parameter $\boldsymbol{\theta}$ according to a linear transformation of the choice set's mean feature vector: $u_i(C, a) = \boldsymbol{y_i}^T(\boldsymbol{\theta} + A\boldsymbol{y_C})$, where $\boldsymbol{y_C} = 1/|C| \sum_{j \in C} \boldsymbol{y_j}$. To incorporate chooser covariates, we define multino-

Table 3.2: Context effect models. $p_{ij} \in \mathbb{R}, \boldsymbol{\gamma_i} \in \mathbb{R}^{d_x}, \boldsymbol{\theta} \in \mathbb{R}^{d_y}, A \in \mathbb{R}^{d_y \times d_y}, B \in \mathbb{R}^{d_y \times d_x}$ are learned parameters.

Model	$u_i(C, a)$	**# Parameters**
CDM (Seshadri et al., 2019)	$\sum_{j \in C \setminus i} p_{ij}$	$n(n-1)$
mult. CDM (MCDM)	$\sum_{j \in C \setminus i} p_{ij} + \boldsymbol{x_a}^T \boldsymbol{\gamma_i}$	$n(n + d_x)$
LCL (Tomlinson and Benson, 2021)	$\boldsymbol{y_i}^T(\boldsymbol{\theta} + A\boldsymbol{y_C})$	$d_y(d_y + 1)$
mult. LCL (MLCL)	$\boldsymbol{y_i}^T(\boldsymbol{\theta} + A\boldsymbol{y_C} + B\boldsymbol{x_a})$	$d_y(d_y + d_x + 1)$

mial versions of these models (Table 3.2). For this chapter, LCL and CDM should be thought of as the simplest context effect models with and without item features.

In contrast, *mixed logit* (McFadden and Train, 2000) accounts for group-level rather than individual-level IIA violations. Recall from Chapter 2 that a (discrete) mixed logit is a mixture of K logits with mixing proportions $\pi_1, \dots, \pi_K$ such that $\sum_{k=1}^K \pi_k = 1$. With $u_i(a_k)$ denoting the utility of the kth component for item i, a mixed logit has choice probabilities

$$\Pr(i \mid C) = \sum_{k=1}^{K} \pi_k \frac{\exp(u_i(a_k))}{\sum_{j \in C} \exp(u_j(a_k))}. \tag{3.1}$$

This can result in a choice system violating IIA but not because any individual chooser experiences context effects. Rather, the aggregation of several choosers each obeying IIA can result in IIA violations.

3.2 Choice set confounding

The traditional approach to choice modeling is to learn a single model for $\Pr(i \mid C)$ (such as a logit or an LCL) and assume it represents overall choice behavior, namely, that the model accurately reflects average choice probabilities $\mathrm{E}_a[\Pr(i \mid a, C)]$. However, $\Pr(i \mid C)$ need not represent average choice behavior at all, as this is only guaranteed under restrictive independence assumptions.

Observation 1. If, for all $a \in \mathcal{A}, C \in \mathcal{C}_\mathcal{D}, i \in C$, at least one of

1. $\Pr(C) = \Pr(C \mid a)$ (chooser-independent choice sets) or
2. $\Pr(i \mid a, C) = \Pr(i \mid C)$ (chooser-independent preferences)

holds, then $\Pr(i \mid C) = \mathrm{E}_a[\Pr(i \mid a, C)$. If both conditions are violated, then this equality can fail.

Proof. Conditioning over choosers yields $\Pr(i \mid C) = \sum_a \Pr(i \mid a, C) \Pr(a \mid C)$. Meanwhile, $\mathrm{E}_a[\Pr(i \mid a, C)] = \sum_a \Pr(i \mid a, C) \Pr(a)$. These are equal if condition (1) holds (since independence also implies $\Pr(a) = \Pr(a \mid C)$). If condition (2) holds, then we directly have $\mathrm{E}_a[\Pr(i \mid a, C)] = \Pr(i \mid C)$. In Example 1 below, we will see an instance where this equality fails when neither (1) nor (2) hold. $\square$

When we have both chooser-dependent sets and preferences, observed choice probabilities $\Pr(i \mid C)$ can differ significantly from true aggregate choice probabilities $\mathrm{E}_a[\Pr(i \mid a, C)]$. We call this phenomenon *choice set confounding*, and provide the following toy example as an illustration.

Example 1. Let $\mathcal{U} = \{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$. Choosers are either cat people or dog people choosing a pet, with choice probabilities

	$\{\mathsf{cat}, \mathsf{dog}\}$	$\{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$
cat person	$3/4$, $1/4$	$3/4$, $1/4$, 0
dog person	$1/4$, $3/4$	$1/4$, $3/4$, 0

Note that the preferences of cat and dog people do not change when fish are included in the choice set. Choice sets are assigned non-independently: cat people see $\{\mathsf{cat}, \mathsf{dog}\}$ w.p. $^3/_4$ and $\{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$ w.p. $^1/_4$ (vice-versa for dog people). Let the population consist of $^1/_4$ cat people and $^3/_4$ dog people. If we only observe samples (C, i) without knowing who is a cat person and who is a dog person,

$$\Pr(\mathsf{dog} \mid \{\mathsf{cat}, \mathsf{dog}\}) = {}^1/_2 \cdot {}^1/_4 + {}^1/_2 \cdot {}^3/_4 = {}^1/_2$$

$$\Pr(\mathsf{dog} \mid \{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}) = {}^1/_{10} \cdot {}^1/_4 + {}^9/_{10} \cdot {}^3/_4 = {}^7/_{10}.$$

However,

$$\mathrm{E}_a[\Pr(\mathsf{dog} \mid a, \{\mathsf{cat}, \mathsf{dog}\})] = {}^1/_4 \cdot {}^1/_4 + {}^3/_4 \cdot {}^3/_4 = {}^5/_8$$

$$\mathrm{E}_a[\Pr(\mathsf{dog} \mid a, \{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}) = {}^1/_4 \cdot {}^1/_4 + {}^3/_4 \cdot {}^3/_4 = {}^5/_8.$$

This mismatch is especially problematic for models that use choice-set dependent utilities $u_i(C)$, such as those designed to account for context effects. From the above data, we might conclude that the presence of a fish causes a dog to become a more appealing option. This spurious context effect would be seized upon by context-based models and even result in improved predictive performance on test data drawn from the same distribution. However, these models would make biased predictions on counterfactual examples where sets are chosen from a different distribution.

In reality, no one's choice would be affected by adding fish to their choice set—it's a red herring. This is a *causal inference* problem. We want to know the cause of a choice, but we are being misled as to whether the change in preferences between the $\{\mathsf{cat}, \mathsf{dog}\}$ and $\{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$ choice sets is due to the presence of fish or to a hidden confounder: the underlying preferences of cat and dog people, coupled with chooser-dependent choice set assignment.

Extending this idea, the equality in Observation 1 can fail dramatically. If the population consists of individuals each of whom obeys IIA (i.e., chooses according to a logit), then $\mathrm{E}_a[\Pr(i \mid a, C)]$ is exactly the mixed logit choice probability. On the other hand, $\Pr(i \mid C)$ can express an arbitrary choice system with choice set confounding.

Theorem 2. *Mixed logit with chooser-dependent choice sets is powerful enough to express any system of choice probabilities.*

Proof. First, notice that any universal logit choice probabilities aggregated over a population can be expressed by set-dependent utilities $u_i^*(C)$ for each $C \subseteq \mathcal{U}, i \in C$. For every choice set $C \subseteq \mathcal{U}$, construct a chooser a_C with fixed utilities $u_i(a_C) = u_i^*(C)$. Let $\Pr(C \mid a_C) = 1$ and $\Pr(C' \mid a_C) = 0$ for all other $C' \neq C$. The choice probabilities of this mixture with chooser-dependent sets is the same as in the original system, and the mixture has finitely many $(2^{|\mathcal{U}|} - 1)$ components, one for each nonempty choice set C. $\square$

Arbitrary choice systems are much more powerful than mixed logit (even ones with continuous mixtures). For example, it is impossible for mixed logit to violate *regularity*, the condition that $\Pr(i \mid C) \geq \Pr(i \mid C \cup \{j\})$ for all $C \subseteq \mathcal{U}, i \in C, j \in \mathcal{U}$, as choice probabilities for i can only go down in each mixture component when we include j. On the other hand, even Example 1 has a regularity violation (picking a dog is more likely when a fish is available), despite there being only two types of choosers, both adhering to IIA.

We have shown that choice set confounding is an issue in theory, and we now demonstrate it to be a problem in practice. We present evidence of choice set confounding in two transportation choice datasets, SF-WORK and SF-SHOP (Koppelman and Bhat, 2006). These datasets consist of San Francisco (SF) resident surveys for preferred transportation mode to work or shopping, where the choice set is the set of modes available to a respondent. The SF datasets are common testbeds for choice models violating IIA (Koppelman and Bhat, 2006; Seshadri et al., 2019; Ragain and Ugander, 2016; Benson et al., 2016) and in choice applications (Tomlinson and Benson, 2020; Agarwal et al., 2018).

Table 3.3: Regularity violations in SF-WORK and SF-SHOP, impossible under mixed logit. Including additional item(s) appears to increase the probability that DA or DA/SR is chosen. The differences are significant according to Fisher's exact test (SF-WORK: $p = 6.5 \times 10^{-9}$, SF-SHOP: $p = 0.005$).

SF-WORK **Choice set (C)**	Pr(DA | C)	N
{DA, SR 2, SR 3+, Transit}	0.72	1661
{DA, SR 2, SR 3+, Transit, Bike}	0.83	829

SF-SHOP **Choice set (C)**	Pr(DA/SR | C)	N
{DA, DA/SR, SR 2, SR 3+, SR 2/SR 3+, Transit}	0.17	534
{DA, DA/SR, SR 2, SR 3+, SR 2/SR 3+, Transit, Bike, Walk}	0.23	1315

DA: drive alone. SR: shared ride, number indicates car occupancy.
Slashes indicate different mode used for outbound and inbound trips.

The SF data have regularity violations (see Table 3.3), ruling out the possibility that the IIA violations in these datasets are just due to mixtures of choosers obeying IIA. Thus, these datasets either have (1) true context effects or (2) choice set confounding. So far, the literature has focused on (1), but we argue that (2) is more likely. We compare the likelihoods of logit, MNL, CDM, and MCDM (recall Tables 3.1 and 3.2; MCDM is a CDM with the MNL chooser covariates term) on these datasets through likelihood-ratio tests (Table 3.4). MNL and MCDM both account for chooser-dependent preferences through covariates, while CDM and MCDM both account for context effects. With true context effects, we would expect CDM to be significantly more likely than logit and MCDM to be significantly

Table 3.4: Likelihood gains in SF-WORK, SF-SHOP, and EXPEDIA from covariates and context with likelihood ratio test (LRT) p-values. $\Delta\ell$ denotes improvement in log-likelihood.

Comparison	**Testing**	**Controlling**	$\Delta\ell$	**LRT** p
SF-WORK				
Logit to MNL	covariates	—	883	$< 10^{-10}$
Logit to CDM	context	—	85	$< 10^{-10}$
CDM to MCDM	covariates	context	819	$< 10^{-10}$
MNL to MCDM	context	covariates	20	0.08
SF-SHOP				
Logit to MNL	covariates	—	343	$< 10^{-10}$
Logit to CDM	context	—	96	$< 10^{-10}$
CDM to MCDM	covariates	context	276	$< 10^{-10}$
MNL to MCDM	context	covariates	29	0.36
EXPEDIA				
CL to CML	covariates	—	1218	$< 10^{-10}$
CL to LCL	context	—	2345	$< 10^{-10}$
LCL to MLCL	covariates	context	1167	$< 10^{-10}$
CML to MLCL	context	covariates	2294	$< 10^{-10}$

more likely than MNL. However, this is not the case. While CDM is significantly more likely than logit, MCDM is not significantly more likely than MNL in both SF datasets. Thus, context effects only appear significant before controlling for preference heterogeneity through covariates. This is exactly what we would expect if the IIA violations in these datasets are due to choice set confounding rather than context effects. In contrast, we see significant context effects in the EXPEDIA hotel-booking dataset (Kaggle, 2013a) even after controlling for covariates (this dataset uses item features, hence the different models in Table 3.4), so context effects are likely. This dataset consists of search results (choice sets) and hotel bookings (choices), and we explore it further in Section 3.3.4.

The choice set confounding leads to a key question: how were choice sets constructed in SF-WORK and SF-SHOP? According to Koppelman and Bhat (2006), choice sets were imputed based on chooser covariates from the survey. For instance, walking was included as an option if a respondent's distance to the destination was < 4 miles and driving was included if they had a driver's license and at least one car in their household (Koppelman and Bhat, 2006). This choice set assignment is highly chooser-dependent, resulting in strong choice set confounding.

Example 1 and the SF datasets highlight how confounding can lead to spurious context effects and incorrect average choice probabilities. Next, in Section 3.3, we adapt methods from causal inference so that chooser covariates can correct choice probability estimates. And in Section 3.4, we address what can be done without covariates if we want to (1) make predictions under chooser-dependent choice set assignment mechanisms or (2) make counterfactual predictions for previously observed choosers.

3.3 Causal inference methods

In traditional causal inference (Rubin, 1974; Imbens, 2004; Imbens and Rubin, 2015), we wish to estimate the causal effect of an intervention (e.g., a medical treatment) from observational data. However, we cannot simply compare the outcomes of the treated and untreated cohorts if treatment was not randomly assigned—confounders might affect both whether someone was treated and their outcome. There are many methods to debias treatment effect estimation, including matching (Rubin, 1974; Rosenbaum and Rubin, 1983), inverse probability weighting (IPW) (Hirano et al., 2003), and regression (Rubin, 1977). One can also combine methods, such as IPW and regression, which is the basis for *doubly robust* estimators (Bang and Robins, 2005).

Here, we adapt causal inference methods to estimate unbiased discrete choice models from data with choice set confounding. First, we adapt IPW to learn unbiased models that do not use chooser covariates in the utility function. After, we show an equivalence between incorporating chooser covariates in the utility function and regression for causal inference. Finally, we combine these methods for doubly robust choice model estimation. For discrete choice, these methods require new assumptions and have different guarantees. We first provide a brief introduction to causal inference terminology in the binary treatment setting, such as an observational medical study (in contrast, we will think of choice sets as treatments).

In potential outcomes notation (Rubin, 2005), each person i has covariates X_i and is either treated ($T_i = 1$) or untreated ($T_i = 0$). At some point after treatment, we measure the *outcome* $Y_i(T_i)$. A typical goal of the causal inference methods above is to estimate the *average treatment effect* $\mathrm{E}_i[Y_i(1) - Y_i(0)]$. All of these methods rely on untestable assumptions; in particular, they rely on *strong ignorability* (Rosenbaum and Rubin, 1983; Imbens, 2004) (also called *unconfoundedness* or *no unmeasured confounders*), which requires that the treatment is independent from the outcome, conditioned on observed covariates: $\Pr(T_i \mid X_i, Y_i) = \Pr(T_i \mid X_i), \forall i$.

3.3.1 Inverse probability weighting

IPW estimation commonly requires estimating *propensity scores* describing the probability of each treatment assignment given individual covariates. The true probabilities $\Pr(T_i \mid X_i)$ are unknown, so estimated "propensities" $\widehat{\Pr}(T_i \mid X_i)$ are learned from observed data, typically via logistic regression (Austin, 2011).

Propensities can then be used to estimate average treatment effects or, as in our case, to re-weight a model's training data (Freedman and Berk, 2008). By weighting each sample by the inverse of its propensity, we effectively construct a *pseudo-population* where treatment is assigned independently from covariates. In addition to ignorability, IPW requires *positivity*, the assumption that all propensities satisfy $0 < \Pr(T_i \mid X_i) < 1$.

In the discrete choice setting, we think of choice sets as treatments. By Observation 1, we need chooser-independent choice sets in order to learn an unbiased choice model. Our idea of IPW for discrete choice is to create a pseudo-dataset in which this is true and to learn a choice model over that pseudo-dataset. To do this, we model choice set assignment probabilities $\Pr(C \mid a)$. We can then replace each sample (i, a, C) with $1/[|\mathcal{C}_\mathcal{D}| \Pr(C \mid a)]$ copies, creating a pseudo-dataset $\tilde{\mathcal{D}}$ with uniformly random choice sets (note that we allow "fractional samples," since we don't explicity construct $\tilde{\mathcal{D}}$). However, we cannot hope to learn $\Pr(C \mid a)$ in datasets with only a single observation per chooser (which is very often the case). We instead need to rely on observed covariates $\boldsymbol{x_a}$. We thus learn $\Pr(C \mid \boldsymbol{x_a})$ and use these propensities to construct $\tilde{\mathcal{D}}$. For the analysis, we assume we know the true propensities, but a correctly specified choice set assignment model learned from data is sufficient.

To learn a choice model from $\tilde{\mathcal{D}}$, we can simply add weights to the model's log-likelihood function, resulting in

$$\ell(\theta; \tilde{\mathcal{D}}) = \sum_{(i,C,a)\in\mathcal{D}} \frac{\log \Pr_\theta(i \mid C)}{|\mathcal{C}_\mathcal{D}| \Pr(C \mid \boldsymbol{x_a})}. \tag{3.2}$$

In order for $\Pr(C \mid \boldsymbol{x_a})$ to be an effective stand-in for $\Pr(C \mid a)$, we need the following assumption (see Figure 3.1).

Definition 1. *Choice set ignorability* is satisfied if choice sets are independent of choosers, conditioned on chooser covariates: $\Pr(C \mid a, \boldsymbol{x_a}) = \Pr(C \mid \boldsymbol{x_a})$.

Just as in standard IPW, we also need positivity (of choice set propensities). Under these assumptions, IPW guarantees that empirical choice probabilities in the pseudo-dataset $\tilde{\mathcal{D}}$ reflect aggregate choice probabilities in the true population. To formalize this, we introduce $\mathcal{D}^*$, an idealized dataset with uniformly random choice set assignment for every chooser (of the same size as $\mathcal{D}$). $\mathcal{D}^*$ consists of $|\mathcal{D}|$ independent samples (a, C, i) each occuring with probability $\Pr(a)\frac{1}{|\mathcal{C}_\mathcal{D}|}\Pr(i \mid a, C)$. We now show that the IPW-weighted log-likelihood (eq. (3.2)) is, in expectation, the same as the log-likelihood function over $\mathcal{D}^*$. Since $\mathcal{D}^*$ has chooser-independent choice sets, we can train a model for $\Pr(i \mid C)$ using eq. (3.2) and expect it to capture unbiased aggregate choice probabilities (by Observation 1).

Theorem 3. *If, for all $a \in \mathcal{A}, C \in \mathcal{C}_\mathcal{D}$,*

1. $0 < \Pr(C \mid \boldsymbol{x_a}) < 1$ *(positivity), and*
2. $\Pr(C \mid a, \boldsymbol{x_a}) = \Pr(C \mid \boldsymbol{x_a})$ *(choice set ignorability),*

then $E_\mathcal{D}[\ell(\theta; \tilde{\mathcal{D}})] = E_{\mathcal{D}^*}[\ell(\theta; \mathcal{D}^*)]$.

Proofs for this and subsequent results can be found in Appendix B. Choice set ignorability is crucial to the success of IPW, so we should assess when this assumption is reasonable. If choice sets are generated by an exogenous process (such as a recommender system, as in the EXPEDIA dataset), then as long as we have access to the same covariates as that process, choice set ignorability holds, although learning the propensities may still be a challenge. However, in other datasets, choice sets are formed through self-directed browsing (e.g., clicking around an online shop,

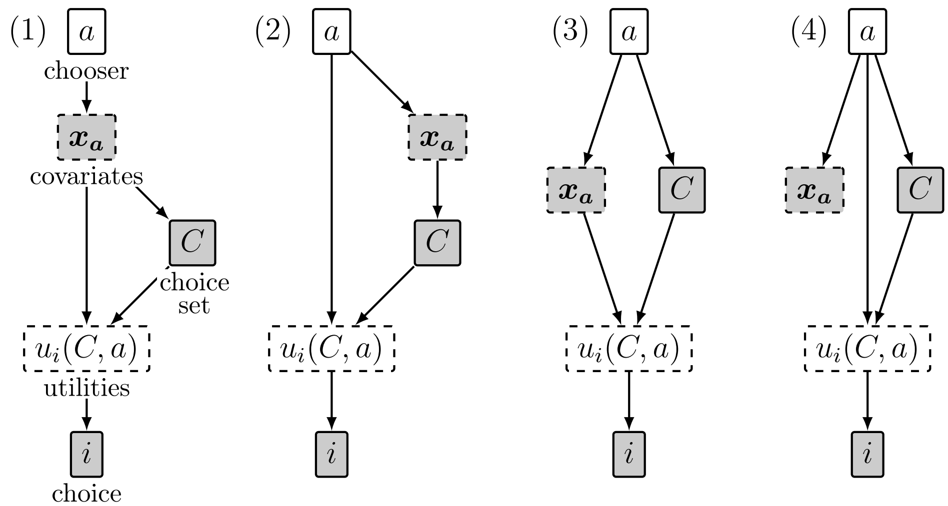


Figure 3.1: Graphical representations of chooser covariate assumptions: (1) ignorability; (2) choice set ignorability; (3) preference ignorability; (4) no ignorability. Shaded nodes are observed, dashed nodes are deterministic.

as in the YOOCHOOSE dataset we examine later). In those cases, basic covariates (age, gender, etc.) are unlikely to fully capture choice set generation, since sets result from the complexities of human behavior rather than the simpler algorithmic behavior of a recommender system. As in traditional causal inference, the validity of choice set ignorability must be determined by the practitioner applying the method.

3.3.2 Regression

An alternative to using chooser covariates to learn choice set propensities is to incorporate covariates directly into the utility formulation, as in the multinomial or conditional multinomial logit models. If chooser covariates fully capture their preferences and the choice model is correctly specified, then the model that we learn is consistent. We formalize the first condition as follows (see Figure 3.1).

Definition 2. *Preference ignorability* is satisfied if choice probabilities are independent of choosers, conditioned on chooser covariates: $\Pr(i \mid a, \boldsymbol{x_a}, C) = \Pr(i \mid \boldsymbol{x_a}, C)$.

Given correct specification and preference ignorability, the choice model will be consistent in terms of aggregate choice probabilities and result in accurate individual choice probability estimates.

Theorem 4. *If $\Pr(i \mid a, \boldsymbol{x_a}, C) = \Pr(i \mid \boldsymbol{x_a}, C)$ for all $a \in \mathcal{A}, C \in \mathcal{C}_\mathcal{D}, i \in C$ (preference ignorability), then the MLE of a correctly specified (and well-behaved, in the standard MLE sense (Wasserman, 2013, Theorem 9.13)) choice model that incorporates chooser covariates $\boldsymbol{x_a}$ is consistent: $\lim_{|\mathcal{D}|\to\infty} \widehat{\Pr}(i \mid \boldsymbol{x_a}, C) = \Pr(i \mid a, C)$.*

While the guarantee of regression is stronger than IPW, preference ignorability is more challenging to satisfy in practice. Instead of needing all covariates used to generate choice sets, we need covariates to fully describe choice behavior.

3.3.3 Doubly robust estimation

A constraint of both IPW and regression is correct model specification, either of the choice set propensity model or of the choice model. In traditional causal inference, one can combine both methods to provide guarantees if either model is correctly specified, producing *doubly robust* estimators (Bang and Robins, 2005; Funk et al., 2011). In the same way, we can combine IPW and regression for choice models and achieve their respective guarantees if their respective conditions are satisfied. In other words, the two methods do not interfere with each other. However, this increases the variance of estimates, so it may be advisable to only use one method if we are confident in one of the assumptions.

3.3.4 Empirical analysis of IPW and regression

We begin by evaluating regression and IPW adjustments in synthetic data, and then apply our methods to the EXPEDIA dataset (training details in Appendix B.3).

Counterfactual evaluation in synthetic data

We generate synthetic data with heterogeneous preferences, CDM-style context effects, and choice set confounding. Specifically, we use 20 items with embeddings $\boldsymbol{y_i} \in \mathbb{R}^2$ sampled uniformly from the unit circle. We also generate embeddings $\boldsymbol{x_a}$ in the same way for each chooser a. Each chooser a picks items according to an MCDM, where the utility for i is a sum of $\boldsymbol{x_a^T y_i}$ plus a CDM term shared by all choosers, with each "push/pull" term $p_{ij} \sim \text{Uniform}(-1, 1)$. To generate a choice set for a, we sample a uniformly random set with probability 0.25 (to satisfy positivity) and otherwise include each item with probability $^1/_{(1+e^{-c\boldsymbol{x_a^T y_i}})}$, where c is the *confounding strength* (we condition on having at least two items in the choice set). Higher confounding strength results in sets containing items more preferred by a. Each trial consists of 10000 samples. Item embeddings are unobserved, but chooser embeddings are used as covariates. We train models on a confounded portion of the data and measure prediction quality on a held-out confounded subset as well as a counterfactual portion with uniformly random choice sets. For IPW, we estimate choice set propensities via per-item logistic regression, multiplying item propensities to get set propensities.

To measure prediction quality, we use the mean relative position of the true choice in the list of predictions sorted in descending probability order. A value of 1 says that the true choices were all predicted as most likely. As confounding

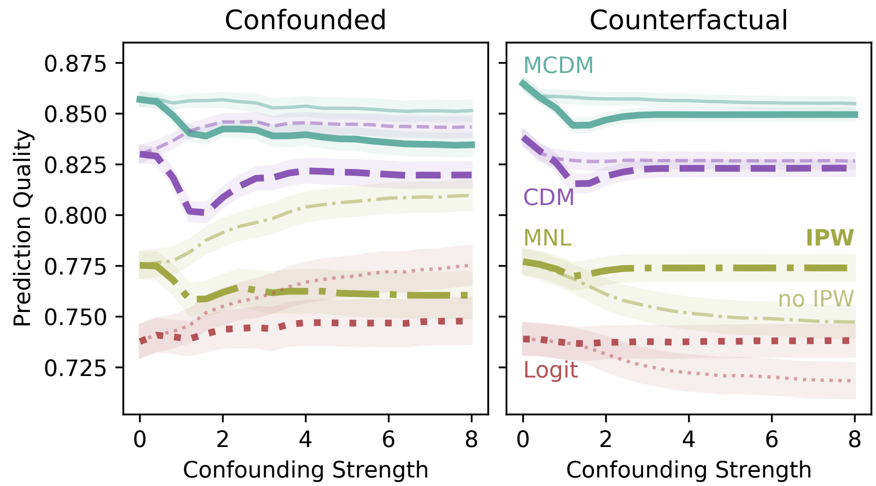


Figure 3.2: Mean prediction quality of models on synthetic data with both context effects and choice set confounding, with IPW (bold) and without IPW (light). Left: out-of-sample predictions on data with confounding. Right: counterfactual predictions of models trained on confounded data. Shaded regions show standard error over 16 trials.

strength increases, prediction quality increases in the confounded data for logit, MNL, and CDM, while decreasing on counterfactual data (Figure 3.2). For logit and MNL, IPW leads to models that generalize better to counterfactual data. For CDM, IPW correctly prevents the illusion of increased performance with more confounding (although variance caused by IPW appears to result in a small dip in performance at low confounding). Since preference ignorability is satisfied, IPW is unnecessary for MCDM: regression with the correctly specified model successfully generalizes despite confounding.

Empirical data with chooser covariates

We now consider the EXPEDIA hotel choice dataset (Kaggle, 2013a) from Section 3.2, using five hotel features: star rating, review score, location score, price, and promotion status. This allows us to use feature-based choice models (CL, CML, LCL, and MLCL; Tables 3.1 and 3.2). The dataset includes information about chooser searches, such as the number of adults and children in their party, which likely have strong effects on choice sets (i.e., search results). This is an excellent testbed for IPW since these covariates are likely informative about choice sets, making choice set ignorability more reasonable than preference ignorability.

We do not have counterfactual choices for the EXPEDIA data, but we still consider several types of analysis. First, we recall the results from Table 3.4 to see if apparent context effects are accounted for by chooser covariates. There, in contrast to the SF datasets, context effects still appear significant after controlling for covariates. In fact, context effects provide a larger likelihood boost than the chooser covariates. Thus, either (1) there are true context effects or (2) the chooser covariates in EXPEDIA do not satisfy preference ignorability (or both). Based on the nature of the covariates, (2) seems very likely: the number of children in the chooser's party and the length of their stay are unlikely to fully describe hotel preferences.

Since regression is inconclusive, we also apply IPW. To learn choice set propensities, we use a probabilistic model of the mean feature vectors of choice sets. We assume these vectors follow a multivariate Gaussian conditioned on chooser covariates, with mean $W\boldsymbol{x_a} + \boldsymbol{z}$ for some $W \in \mathbb{R}^{d_y \times d_x}, \boldsymbol{z} \in \mathbb{R}^{d_y}$. Given observed mean choice set vectors $\boldsymbol{y_C}$ and corresponding chooser covariates $\boldsymbol{x_a}$, we compute the maximum-likelihood W, $\boldsymbol{z}$, and covariance matrix (see Appendix B.2). This model gives us propensities for any (a, C) pair.

Using IPW with these propensities dramatically decreases the negative impact of high price in all four models (Figure 3.3). After adjusting for confounding, the models indicate that users are more willing to book more expensive hotels. This

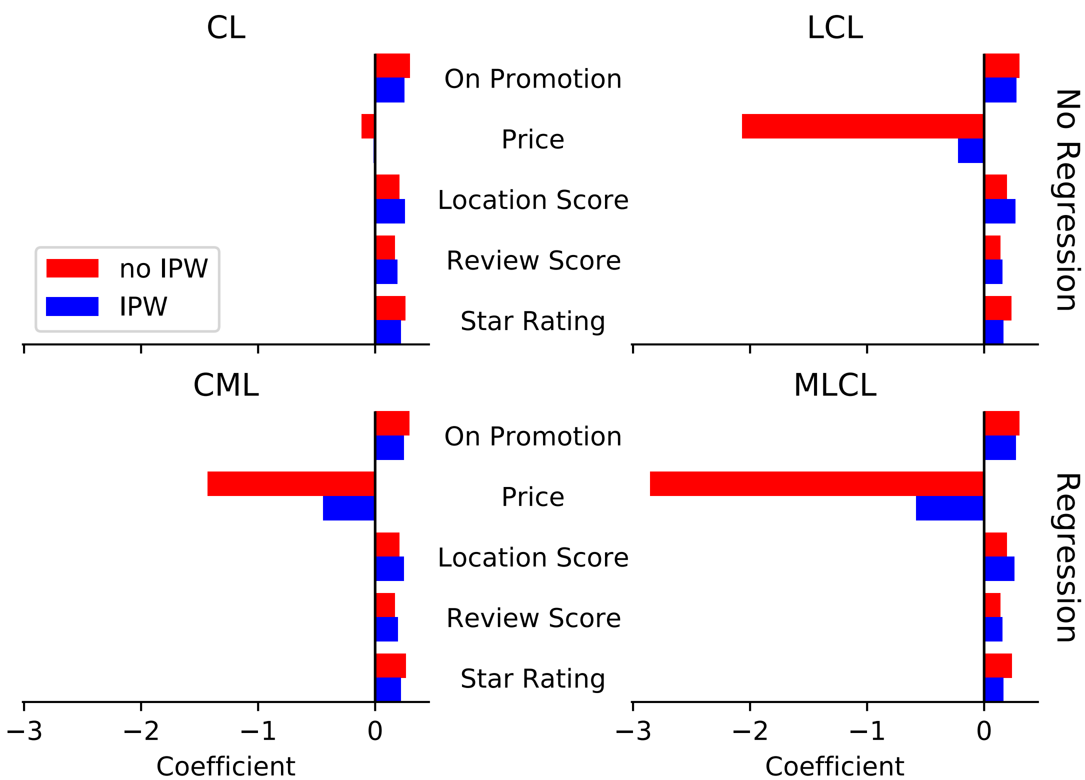


Figure 3.3: Preference coefficients $\boldsymbol{\theta}$ in EXPEDIA for CL and LCL (top row, no regression); and CML and MLCL (bottom row, with regression), with and without IPW. A higher coefficient means choosers prefer higher values of the feature.

makes sense if Expedia is recommending relevant hotels: among a set of hotels matching a user's desired characteristics (such as location and star rating), we would expect them to select the cheapest option. On the other hand, if we presented users with a set of random hotels, location and star rating might play a stronger role in determining their choice, since a random set might have many cheap hotels that are undesirable for other reasons. In addition to the preference coefficients, IPW affects the context effect matrix A in the LCL and MLCL (Figure 3.4). In both models, IPW decreases (but does not entirely eliminate) the strong price context effects. This is evidence that some of the apparent context effects in the dataset are due to choice set confounding.

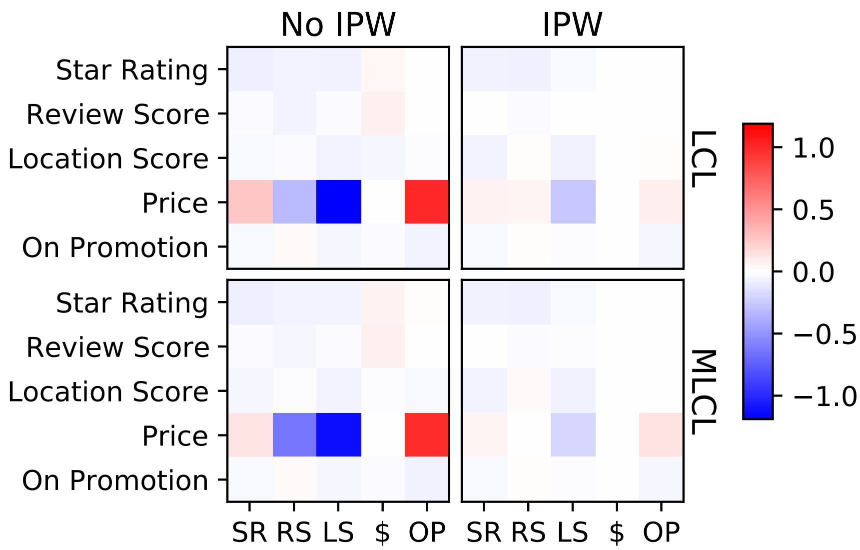


Figure 3.4: LCL and MLCL context effect matrix A in EXPEDIA with and without IPW. A higher value means choosers prefer a row feature more in a set where the mean column feature (abbreviated) is high; 0 indicates no context effect.

Table 3.5: Log-likelihoods and estimated random-set log-likelihoods with IPW on EXPEDIA. After adjusting for confounding, the data is far easier to explain.

Model	**Confounded**	**IPW-adjusted**
CL	−839499	−786653
CML	−838281	−785753
LCL	−837154	−784770
MLCL	−835986	−783928

Finally, the estimated likelihoods of the models under IPW are significantly better than without IPW (Table 3.5). We normalize the IPW-weighted log-likelihood by the sum of the IPW weights, which provides an estimate of what the IPW-trained model's log-likelihood would be given random sets. The gap between likelihood with no IPW and estimated likelihood with IPW dwarfs the gaps between different choice models, indicating that accounting for choice set confounding makes the data much more consistent with the random utility maximization principle underlying all four models. (By Theorem 2, choice set confounding can result in choice systems far from rational behavior, even when choosers are rational.)

3.4 Managing without covariates

So far, we have used chooser covariates to correct for choice set confounding. However, in some choice data, there are no covariates available, or we are not willing to make ignorability assumptions. Here, we show what can be done in this setting.

3.4.1 Within-distribution prediction

Unfortunately, by Theorem 2, it is impossible to determine whether IIA violations are caused by choice set confounding or true context effects in the absence of chooser information. Nonetheless, we can still exploit IIA violations—whatever their origin—to improve prediction, as long as we are careful not to make counterfactual predictions. This is essentially what researchers developing context effect models (Ragain and Ugander, 2016; Seshadri et al., 2019; Bower and Balzano, 2020; Tomlinson and Benson, 2021; Rosenfeld et al., 2020) have been doing (without a framework for understanding the possibility of choice set confounding and the associated risks for counterfactual prediction). Beyond emphasizing a need for caution, we also establish a duality between models accounting for context effects and models accounting for choice set confounding; specifically, we show that a model equivalent to the CDM—which was designed with context effects in mind—can be derived purely from the perspective of choice set confounding.

In a multinomial logit (MNL), we learn a latent parameter vector $\boldsymbol{\gamma_i}$ for each item $i \in \mathcal{U}$ and model utilies as $u_i(a) = \boldsymbol{x_a}^T \boldsymbol{\gamma_i}$ (omitting the intercept term). Suppose we don't have any chooser covariates, but we know choice set assignment depends on choosers. We could then use the choice set itself as a surrogate for user covariates (e.g., one covariate could be "someone who is offered item i"). Let

$\mathbf{1}_{\boldsymbol{C_a}}$ be a binary encoding of the choice set C_a of a chooser a (a length $|\mathcal{U}|$ vector with a 1 in position i if $i \in C_a$). Consider treating $\mathbf{1}_{\boldsymbol{C_a}}$ as a substitute for the user covariates $\boldsymbol{x_a}$. Then the MNL model is

$$\Pr(i \mid C_a) = \frac{\exp(\mathbf{1}_{\boldsymbol{C_a}}^T \boldsymbol{\gamma_i})}{\sum_{j \in C_a} \exp(\mathbf{1}_{\boldsymbol{C_a}}^T \boldsymbol{\gamma_j})}.$$

The utility of i in set C_a is $\sum_{j \in C_a} \gamma_{ij}$, which is exactly the CDM (with self-pulls, since the sum is over C_a rather than $C_a \setminus i$), a model designed to capture choice-set-dependent utilities. Thus, the CDM can either be thought of as accounting for pairwise interactions between items or using the choice set as a stand-in for user covariates.

One natural question this duality raises is how the set of choice systems expressible by CDM (or other context-effect models) compares to the choice systems induced by mixed populations of IIA choosers with choice set confounding, which take the form

$$\Pr(i \mid C) = \sum_{a \in \mathcal{A}} \Pr(a \mid C) \frac{\exp(u_i(a))}{\sum_{j \in C} \exp(u_j(a))}. \tag{3.3}$$

Mixtures of logits such as eq. (3.3) are notoriously hard to analyze (even in two-component case (Chierichetti et al., 2018b)), so no simple equivalence between a context-effect model and such a mixture is likely. In fact, eq. (3.3) is even trickier than standard mixed logit (eq. (3.1)), since the mixture weights depend on the choice set.

Nonetheless, some progress in this direction is possible. Here, we provide an instance where the LCL approximates a choice system induced by choice set confounding (of the form of eq. (3.3)). Recall that the LCL has utilities $u_i(a, C) = (\boldsymbol{\theta} + A\boldsymbol{y_C})^T \boldsymbol{y_i}$, where $\boldsymbol{y_C}$ is the mean feature vector over the choice set. If we make Gaussian assumptions on the distribution of features and on choice set assignment, and if chooser utilities are inner products of chooser and item vectors, then the LCL is a mean-field approximation to the induced choice system. In particular, we assume choice sets are generated to be similar to items the chooser a would like (as in a recommender system) by sampling items from a Gaussian with mean $\boldsymbol{x_a}$.

Theorem 5. *Let items and choosers both be represented by vectors in $\mathbb{R}^d$. Suppose chooser covariates $\boldsymbol{x_a}$ are distributed in the population according to a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma_0)$, and a choice set for chooser a is constructed by sampling k items from the multivariate Gaussian $\mathcal{N}(\boldsymbol{x_a}, \Sigma)$. Additionally, assume choosers have the utility function $u_i(a, C) = \boldsymbol{x_a}^T \boldsymbol{y_i}$. Then the expected chooser given a choice set C, $\boldsymbol{x_a^*} = \mathrm{E}[\boldsymbol{x_a} \mid C]$, has LCL choice probabilities, with*

$$\boldsymbol{\theta} = \tfrac{1}{k}\Sigma(\Sigma_0 + \tfrac{1}{k}\Sigma)^{-1}\boldsymbol{\mu}, \quad A = \Sigma_0(\Sigma_0 + \tfrac{1}{k}\Sigma)^{-1}.$$

Thus, the LCL can either be thought of as a context effect model, or as an approximation to the choice system induced by recommender-style preferred item overrepresentation.

3.4.2 Counterfactuals for known choosers

To make counterfactual predictions without chooser covariates or insufficiently descriptive covariates (preventing us from applying IPW or regression), we develop a clustering method for the challenge of choice set confounding. Suppose a recommender system suggests two sets of movies to two users: $\{\textsf{Romance A}, \textsf{Romance B}\}$ to a_1 and $\{\textsf{Drama A}, \textsf{Drama B}\}$ to a_2. While we know nothing about a_1 or a_2, we might be inclined to think a_1 is likely to pick Romance A from $\{\textsf{Romance A}, \textsf{Drama A}\}$, while a_2 is likely to pick Drama A from the same choice set. Similar to the CDM derivation in the previous section, the choice set is a signal for chooser preferences. We can also apply collaborative filtering principles, with the distinction that instead of thinking that similar users like similar items, we assume similar choosers are shown similar choice sets. There is a limitation, though, as this approach only lets us make predictions for choosers who appear in the original dataset. While there are many ways of using information from choice set assignment, we highlight an approach for the case where we have corresponding types of choosers and items (e.g., "romance fans" for "romance movies").

Suppose that choosers are more likely to have an item in their choice set if it matches their type. Define the $m \times n$ matrix R, where $R_{ij} = 1$ if the ith choice set includes item j and $R_{ij} = 0$ otherwise. We can think of R as the upper right block of the adjacency matrix of a bipartite graph between choosers and items, in which an edge (a, i) means that i is in a's choice set. With fixed choice set inclusion probabilities for each type, clustering choosers into types based on their choice sets is then an instance of the bipartite stochastic block model (SBM) recovery problem (Larremore et al., 2014; Abbe, 2018).

In Theorem 6, we apply a classic exact recovery result due to McSherry (2001) to show how a choice system with discrete types can be deconfounded without access to chooser covariates (i.e., knowledge of type membership), but any bipartite SBM clustering algorithm could be used (see Abbe (2018) for a survey of SBM results).

Theorem 6. *Suppose items and choosers are jointly split into k types. Let s be the smallest number of items or choosers of any type and let $n = |A| + |\mathcal{U}|$. Suppose that for each chooser $a \in A$, $i \in \mathcal{U}$ is included in a's choice set with probability p if a and i are of the same type and with probability q otherwise.*

There exists a constant C such that for large enough n, if

$$s(p-q)^2 > Ck\left({}^{n}\!/\!_{s} + \log {}^{n}\!/\!_{\delta}\right), \tag{3.4}$$

then w.p. $1-\delta$, we can efficiently learn the type of every item and every chooser given a dataset $\mathcal{D}$ with one choice from each $a \in A$.

While McSherry's algorithm has strong theoretical guarantees, a more practical implementation is spectral co-clustering (Dhillon, 2001), which performs well for our purposes. Once we recover type memberships, we train separate models for each type of chooser and use the model for a chooser's type for deconfounded counterfactual predictions.

3.4.3 Empirical data without chooser covariates

We apply our spectral co-clustering method to the YOOCHOOOSE online shopping dataset (Ben-Shimon et al., 2015). The dataset consists of all items clicked on in a session and an indicator of whether each item was purchased. We consider each purchase to be a choice from the set of all items viewed in the session. We group items by category (e.g., sports equipment) removing those with fewer than 100 purchases, leaving 29 categories.

We then perform spectral co-clustering (Dhillon, 2001) on the choice set matrix R with 2 to 10 chooser clusters and train a separate logit on each cluster. We ignore the item clusters. We compare against random clustering with the cluster sizes found by spectral clustering and mixed logit with the same number of components.

Spectral clustered logit describes the data much better than random clustering or even mixed logit (Figure 3.5). Note that the clusters are based only on choice

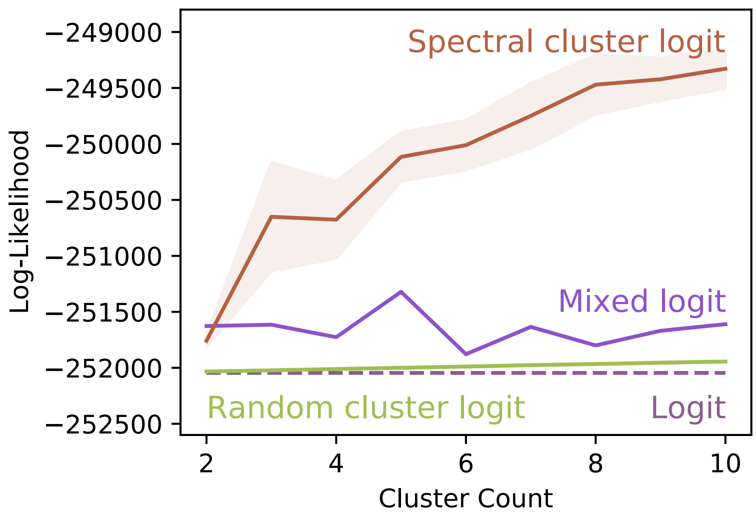


Figure 3.5: YOOCHOOSE log-likelihood comparison. Spectral and random cluster results are averaged over eight trials, with one standard deviation shaded.

set assignment, not choice behavior. In contrast, mixed logit bases its mixture components solely based on choices. The strong performance of spectral clustering indicates that choice sets are informative about preferences, and our use of this information is much easier than learning a mixture model.

3.5 Discussion

Choice set confounding is widespread and can affect choice probability estimates, alter or introduce context effects, and lead to poor generalization. Existing models ignoring chooser covariates are particularly susceptible, but plugging in covariates is not a universal solution. We saw that covariates may be more informative about choice sets than preferences, making IPW more viable than regression. An important contribution is formalizing and demonstrating choice set confounding, as it has significant implications for discrete choice modeling. For instance, initial research on the SF transportation data used extensive nested logit modeling to account for IIA violations (Koppelman and Bhat, 2006), which we can manage with choice set confounding.

Our methods are a first step in addressing confounding. A challenge was learning choice set propensities for IPW. Simple logistic regression can work for binary treatments, but estimating exponentially many choice set propensities is difficult. In EXPEDIA, we learned a distribution over mean choice set feature vectors as an approximation. Other methods for learning set assignment probabilities would be valuable. Instrumental variables are another causal inference approach (Hernán and Robins, 2006) that could be used in our setting, but identifying instruments for choice data is difficult. Alternatively, a matching approach (Imbens, 2004) could compare pairs of similar choosers with different choice sets. Other directions for future investigation include rigorous methods of detecting choice set confounding, or verifying that it has been successfully accounted for, and of testing assumptions.

CHAPTER 4

GRAPH-BASED METHODS FOR DISCRETE CHOICE

One of the crucial aspects of human decision-making is that, as fundamentally social creatures, our preferences are strongly influenced by our social context. Viral trends, conformity, word-of-mouth, and signaling all play roles in behavior, including choices (Feinberg et al., 2020; Axsen and Kurani, 2012). Additionally, people with similar preferences, beliefs, and identities are more likely to be friends in the first place, a phenomenon known as *homophily* (McPherson et al., 2001). Together, these factors indicate that social network structure could be very informative in predicting choices. In economics and sociology, there has been growing interest in incorporating social factors into discrete choice models (McFadden, 2010; Maness et al., 2015; Feinberg et al., 2020). However, the methods used so far in these fields have largely been limited to simple feature-based summaries of social influence (e.g., what fraction of someone's friends have selected an item (Goetzke and Rave, 2011)).

On the other hand, the machine learning community has developed a rich assortment of graph learning techniques that can incorporate entire social networks into predictive models (Kipf and Welling, 2017; Jia and Benson, 2022; Wu et al., 2020), such as graph neural networks and graph-based regularization. These approaches can handle longer-range interactions and are less reliant on hand-crafted features. Because of the large gulf between the discrete choice and machine learning communities, there has been almost no study of the application of graph learning methods to discrete choice, where they have the potential for major impact. Perhaps one factor hindering the use of graph learning in discrete choice is that machine learning methods are typically designed for either regression or classification.

Discrete choice has several features distinguishing it from multiclass classification (its closest analogue)—for instance, each observation can have a different set of available items. As a concrete example, any image could be labeled as a cat in a classification setting, but people choosing between doctors may have their options dictated by their insurance policy.

Motivated by this need, we adapt graph learning techniques to incorporate social network structure into discrete choice modeling. By taking advantage of phenomena like homophily and social contagion, these approaches improve the performance of choice prediction in a social context. In particular, we demonstrate how graph neural networks can be applied to discrete choice, derive Laplacian regularization for the multinomial logit model, and adapt label propagation for discrete choice. We show in synthetic data that Laplacian regularization can improve sample complexity by orders of magnitude in an idealized scenario.

To evaluate our methods, we perform experiments on real-world election data and Android app installations, with networks derived from Facebook friendships, geographic adjacency, and Bluetooth pings between phones. We find that such network structures can improve the predictions of discrete choice models in a semi-supervised learning task. For instance, Laplacian regularization improves the mean relative rank (MRR)[1] of predictions by up to a 6.8% in the Android app installation data and up to 2.6% in the 2016 US election data. In contrast with our results on app installations, we find no evidence of social influence in app usage among the same participants: social factors appear to influence the apps people get, but less so the apps they actually use. Instead, we find that app usage is dominated by personal habit. Another interesting insight provided by our discrete choice

[1]MRR measures the relative position of the true choice in the predicted ranking (Tomlinson and Benson, 2021).

models in the app installation data is the discovery of two separate groups of participants, one in which Facebook is popular, while the other prefers Myspace.[2] We further showcase the power provided by a discrete choice approach by making counterfactual predictions in the 2016 US election data with different third-party candidates on the ballot. While a common narrative is that Clinton's loss was due to spoiler effects by third-party candidates (Chalabi, 2016; Rothenberg, 2019), our results do not support this theory, although we emphasize the likelihood of confounding factors. Our tools enable us to rigorously analyze these types of questions.

4.1 Related work

There is a long line of work in sociology and network science on social behavior, including effects like contagion and herding (Centola and Macy, 2007; Easley and Kleinberg, 2010; Banerjee, 1992). More recently, there has been interest in the use of discrete choice in conjunction with network-based analysis (Feinberg et al., 2020) enabled by rich data with both social and choice components (Aharony et al., 2011). The traditional econometric approach to discrete choice modeling with social effects is to add terms to an individual's utility that depend on the actions or preferences of others (Brock and Durlauf, 2001; McFadden, 2010; Maness et al., 2015). For instance, this approach can account for an individual's desire for conformity (Bernheim, 1994). This is done by treating the choices made by a chooser's community as a feature of the chooser and applying a standard multinomial logit (Páez et al., 2008; Kim et al., 2014; Goetzke and Rave, 2011; Walker et al., 2011; Kim et al., 2018). In contrast, we focus on methods that employ the entire graph rather than derived features. This enables methods to account

[2]The dataset is from 2010, when both were popular options.

for longer-range interactions and phenomena such as network clustering without hand-crafting features. We are aware of one econometric paper that uses preference correlations over a full network in a choice model (Leung, 2019), but inference under this method requires Monte Carlo simulation. Laplacian regularization, on the other hand, allows us to find our model's maximum likelihood estimator with straightforward convex optimization. Mixture models are another way of incorporating structured preference heterogeneity into discrete choice, such as the mixed logit (McFadden and Train, 2000) and hierarchical Bayes models with mixture priors (Allenby and Rossi, 2006; Burda et al., 2008). Again, these approaches present significant challenges for inference, requiring Monte Carlo methods, variational approximations, or expectation maximization. Additionally, in positing unknown latent populations, mixture models ignore the key information provided by the structure of the network. Another large area of research in discrete choice concerns models that allow deviations from the axiom of *independence of irrelevant alternatives* (IIA) (Luce, 1959). Many of these models, such as the multinomial probit (Hausman and Wise, 1978), are very challenging to estimate. To keep our focus on incorporating network effects, we use tractable logit models obeying IIA. However, there are recent non-IIA models admitting efficient inference to which we could apply our methods (Seshadri et al., 2019; Bower and Balzano, 2020; Tomlinson and Benson, 2021); this is beyond the scope of the present work, but we expand further on this idea in the discussion.

In another direction, there are many machine learning methods that use network structure in predictive tasks; graph neural networks (GNNs) (Kipf and Welling, 2017; Xu et al., 2019; Wu et al., 2020) are a popular example. Discrete choice is related to classification tasks, but the set of available items (i.e., labels) is specific to each observation—additionally, discrete choice models are heavily in-formed by economic notions of preference and rationality (McFadden, 1974; Train, 2009). A more traditional machine learning method of exploiting network structure for classification is label propagation (Zhu and Ghahramani, 2002), which we extend to the discrete choice setting. Recent work has shown how to combine label propagation with GNNs for improved performance (Jia and Benson, 2020) and presented a unified generative model framework for label propagation, GNNs, and Laplacian regularization (Jia and Benson, 2022). The present work can be seen as an adaptation and empirical study of the methods from (Jia and Benson, 2022) for discrete choice rather than regression.

The idea of applying Laplacian regularization to discrete choice models appeared several years ago in an unpublished draft (Zhang et al., 2017). However, the draft did not provide experiments beyond binary choices (which reduces to standard semi-supervised node classification (Kipf and Welling, 2017)). In contrast, we compare Laplacian regularization with other methods of incorporating social network structure (GNNs and propagation) on real-world multi-alternative choice datasets.

There is a large body of existing research on predicting app usage and installation, including using social network structure (Baeza-Yates et al., 2015; Pan et al., 2011; Xu et al., 2013), but our use of network-based discrete choice models for this problem is novel. Our approach has the advantage of being applicable to both usage and installation with minimal differences, allowing us to compare the relative importance of social structure in these settings. Another line of related work applies discrete choice models to networks in order to model edge formation (Overgoor et al., 2019; Tomlinson and Benson, 2021; Gupta and Porter, 2022; Overgoor et al., 2020).

4.2 Preliminaries

We again review our the basic discrete choice notation, while introducing concepts specific to this chapter. In a discrete choice setting, we have a universe of *items* $\mathcal{U}$ and a set of *choosers* $\mathcal{A}$. In each choice instance, a chooser $a \in \mathcal{A}$ observes a *choice set* $C \subset \mathcal{U}$ and chooses one item $i \in C$. Each item $i \in \mathcal{U}$ may be described by a vector of features $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$. Similarly, a chooser a may have a vector of features $\boldsymbol{x}_a \in \mathbb{R}^{d_x}$. In the most general form, a choice model assigns choice probabilities for a to each item $i \in C$:

$$\Pr(i \mid a, C) = \frac{\exp(u_{\boldsymbol{\theta}}(i, C, a))}{\sum_{j \in C} \exp(u_{\boldsymbol{\theta}}(i, C, a))}, \tag{4.1}$$

where $u_{\boldsymbol{\theta}}(i, C, a)$ is the utility of item i to chooser a when seen in choice set C, a function with parameters $\boldsymbol{\theta}$. Note that since the utilities in Equation (4.1) can depend on the choice set, this general form can express choice probabilities that vary arbitrarily across choice sets (this is sometimes called the *universal logit*). When constructing more useful parsimonious models, the utilities $u_{\boldsymbol{\theta}}(i, C, a)$ can depend on $\boldsymbol{x}_a$, $\boldsymbol{y}_i$, both, or neither. In the simplest case—the traditional logit model—$u_{\boldsymbol{\theta}}(i, C, a) = u_i$ is constant over choosers and sets. This formulation is attractive from an econometric perspective, since it corresponds to a rationality assumption: if we suppose a chooser has underlying utilities $u_1, \ldots, u_k$ and observes a perturbation of their utilities $u_i + \varepsilon_i$ (where ε_i follows a Gumbel distribution) before selecting the maximum observed utility item, then their resulting choice probabilities take the form of a logit (McFadden, 1974).

When we add a linear term in chooser features to the logit model, the result is the *multinomial logit* (MNL) (Hoffman and Duncan, 1988; McFadden, 1974), with utilities $u_\theta(i, C, a) = u_i + \boldsymbol{\gamma_i}^T \boldsymbol{x_a}$, where u_i are item-specific utilities and $\boldsymbol{\gamma_i}$ is a vector of item-specific coefficients capturing interactions with the chooser features $\boldsymbol{x_a}$. Similarly, when we add a linear term in item features, the result is a *conditional logit* (CL), with utilities $u_i + \boldsymbol{\varphi}^T \boldsymbol{y_i}$. The *conditional multinomial logit* (CML) has both the chooser and item feature terms: $u_i + \boldsymbol{\varphi}^T \boldsymbol{y_i} + \boldsymbol{\gamma_i}^T \boldsymbol{x_a}$. In order to capture heterogeneous preferences among a group of choosers, one natural approach is to allow each chooser a to have different logit utilities. We call this a *per-chooser logit*, which is specified by per-chooser utilities $u_{\boldsymbol{\theta}}(i, C, a) = u_{ia}$. Similarly, a *per-chooser conditional logit* has varying item feature coefficients $\boldsymbol{\varphi}_a$ for each chooser a, with $u_{\boldsymbol{\theta}}(i, C, a) = u_{ia} + \boldsymbol{\varphi}_a^T \boldsymbol{y_i}$. More generally, we call any choice model parameter which varies across choosers a *per-chooser parameter*.

In addition to this standard discrete choice setup, our settings also have a network describing the relationships between choosers. Choosers are nodes in an undirected graph $G = (\mathcal{A}, E)$ where the presence of an edge $(a, b) \in E$ indicates a connection between a and b (e.g., a friendship). We assume G is connected. The *Laplacian* of G is $L = D - A$, where D is the diagonal degree matrix of G and A is the adjacency matrix. The Laplacian has a number of useful applications, including in graph clustering (Hagen and Kahng, 1992) and counting spanning trees (Merris, 1994). For our purposes, the key property of the Laplacian is that quadratic forms of L measure how much a node-wise vector differs across edges of the graph (we elaborate on this property below). We use $n = |\mathcal{A}|$, $m = |E|$, and $k = |U|$. Finally, I denotes the identity matrix.

4.3 Graph-based methods for discrete choice

We identify three phases in choice prediction where networks can be incorporated: networks can be used (1) to inform model parameters, (2) to learn chooser representations, or (3) to directly produce predictions. In this section, we develop representative methods in each category. We briefly describe each method before diving into more detail.

First, networks can inform inference for a model that already accounts for chooser heterogeneity. This is done by incorporating the correlations in utilities (or other choice model parameters) of individuals who are close to each other in the network; we refer to these as *preference correlations* for simplicity. Our Laplacian regularization approach (described in Section 4.3.1) does exactly this, and we show that it corresponds to a Bayesian prior on network-based preference correlations. Second, networks can be used to learn latent representations of choosers that are then used as features in a choice model like the MNL. GNNs have been extensively studied as representation-learning tools—in Section 4.3.2, we focus on how to incorporate them into choice models, using graph convolutional networks (GCNs) (Kipf and Welling, 2017) as our canonical example. Third, direct network-based methods (such as label propagation (Zhu and Ghahramani, 2002), which repeatedly averages a node's neighboring labels) can also be used as a simple baseline for choice predictions. While this approach is simple and efficient, it lacks the proper handling of choice sets of the previous probabilistic approaches. Nonetheless, we find it a useful and effective baseline, and we adapt label propagation for discrete choice in Section 4.3.3.

4.3.1 Laplacian regularization

We begin by describing how to incorporate network information in a choice model like MNL through Laplacian regularization (Ando and Zhang, 2006). Laplacian regularization encourages parameters corresponding to connected nodes to be similar through a loss term of the form $\lambda\boldsymbol{\alpha}^T L\boldsymbol{\alpha}$, where L is the graph Laplacian (as defined in Section 4.2), $\boldsymbol{\alpha}$ is the vector of parameter values for each node, and λ is the scalar regularization strength. A famous identity is that $\boldsymbol{\alpha}^T L\boldsymbol{\alpha} = \sum_{(i,j)\in E}(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j)^2$, which more clearly shows the regularization of connected nodes' parameters towards each other. This also shows that the Laplacian is positive semi-definite, since $\boldsymbol{\alpha}^T L\boldsymbol{\alpha} \geq 0$, which will be useful to preserve the convexity of the multinomial logit's (negative) log-likelihood.

The idea of using Laplacian regularization for discrete choice was proposed in (Zhang et al., 2017) (although they focused on regularizing intercept terms in binary logistic regression). We generalize the idea to be applicable to any logit-based choice model and show that it corresponds to Bayesian inference with a network correlation prior. We then specialize to the models we use in our experiments. Laplacian regularization is simple to implement, can be added to any logit-based choice model with per-chooser parameters, and only requires training one extra hyperparameter. Laplacian regularization also carries a number of advantages over another approach to accounting for structured preference heterogeneity, mixture modeling.

Theory of Laplacian-regularized choice models

Consider a general choice model, as in Equation (4.1). We split the parameters $\boldsymbol{\theta}$ into two sets $\theta_{\mathcal{A}}$ and θ_G, where parameters $\boldsymbol{\alpha} \in \theta_{\mathcal{A}}, \boldsymbol{\alpha} \in \mathbb{R}^n$ vary over choosers and parameters $\beta \in \theta_G, \beta \in \mathbb{R}$ are constant over choosers. The log-likelihood of a general choice model is:

$$\ell(\boldsymbol{\theta}; \mathcal{D}) = \sum_{(i,a,C) \in \mathcal{D}} \Big[\log(u_{\boldsymbol{\theta}}(i, C, a)) - \log \sum_{j \in C} \exp(u_{\boldsymbol{\theta}}(j, C, a)) \Big]. \tag{4.2}$$

The Laplacian- and L_2-regularized log-likelihood (with L_2 regularization strength γ) is then

$$\ell_L(\boldsymbol{\theta}; \mathcal{D}) = \ell(\boldsymbol{\theta}; \mathcal{D}) - \frac{\lambda}{2} \sum_{\boldsymbol{\alpha} \in \theta_{\mathcal{A}}} \boldsymbol{\alpha}^T L \boldsymbol{\alpha} - \frac{\gamma}{2} \sum_{\boldsymbol{\alpha} \in \theta_{\mathcal{A}}} ||\boldsymbol{\alpha}||_2^2. \tag{4.3}$$

We show that regularized maximum likelihood estimation of $\boldsymbol{\theta}$ corresponds to Bayesian inference with a prior on per-chooser parameters that encourages smoothness over the network. In contrast, existing results on priors for semi-supervised regression (Xu et al., 2010; Chin et al., 2019) typically split the nodes into observed and unobserved, fixing the observed values and only considering randomness over unobserved nodes. In choice modeling, observing choices at a node only updates our beliefs about their preferences, leaving some uncertainty. Our result also allows some parameters of the choice model to be chooser-dependent and others to be constant across choosers, allowing it to be fully general over choice models. Finally, we note that L_2 regularization can also be applied to the global parameters β, which as usual corresponds to a Gaussian prior on these parameters—however, we state the result with uniform priors to emphasize the Laplacian regularization on the per-chooser parameters $\boldsymbol{\alpha}$.

Theorem 7. *The maximizer $\boldsymbol{\theta}^*_{MLE}$ of the Laplacian- and L_2-regularized log-likelihood $\ell_L(\boldsymbol{\theta}; \mathcal{D})$ is the maximum a posteriori estimate $\boldsymbol{\theta}^*_{MAP}$ after observing $\mathcal{D}$*

under the i.i.d. priors $\boldsymbol{\alpha} \sim \mathcal{N}(0, [\lambda L + \gamma I]^{-1})$ *for each* $\boldsymbol{\alpha} \in \theta_{\mathcal{A}}$ *and i.i.d. uniform priors for each* $\beta \in \theta_G$.

Proof. First, recall that L is positive semi-definite, so $\lambda L + \gamma I$ (with $\gamma, \lambda > 0$) is positive definite and invertible. Now, using Bayes' Theorem,

$$\begin{aligned}
\boldsymbol{\theta}^*_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathcal{D}) \\
&= \arg\max_{\boldsymbol{\theta}} \frac{\Pr(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\Pr(\mathcal{D})}.
\end{aligned}$$

Since $\Pr(\mathcal{D})$ is independent of the parameters and log is monotonic and increasing,

$$\boldsymbol{\theta}^*_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \left[\log \Pr(\mathcal{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right].$$

Notice that the first term is exactly the log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D})$. Additionally, the priors of each parameter are independent, so

$$\log p(\boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha} \in \theta_{\mathcal{A}}} \log p(\boldsymbol{\alpha}) + \sum_{\beta \in \theta_G} \log p(\beta).$$

Since the priors $p(\beta)$ are uniform, they do not affect the maximizer:

$$\boldsymbol{\theta}^*_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \left[\ell(\boldsymbol{\theta}; \mathcal{D}) + \sum_{\boldsymbol{\alpha} \in \theta_{\mathcal{A}}} \log p(\boldsymbol{\alpha})\right].$$

Now consider the Gaussian priors $p(\boldsymbol{\alpha})$:

$$p(\boldsymbol{\alpha}) = (2\pi)^{n/2} \det[(\lambda L + \gamma I)^{-1}]^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^T [(\lambda L + \gamma I)^{-1}]^{-1} \boldsymbol{\alpha}\right).$$

Simplifying the term in the exp reveals the two regularization terms:

$$\begin{aligned}
-\frac{1}{2}\boldsymbol{\alpha}^T [(\lambda L + \gamma I)^{-1}]^{-1} \boldsymbol{\alpha} &= -\frac{1}{2}\boldsymbol{\alpha}^T (\lambda L + \gamma I) \boldsymbol{\alpha} \\
&= -\frac{\lambda}{2}\boldsymbol{\alpha}^T L \boldsymbol{\alpha} - \frac{\gamma}{2}\boldsymbol{\alpha}^T \boldsymbol{\alpha} \\
&= -\frac{\lambda}{2}\boldsymbol{\alpha}^T L \boldsymbol{\alpha} - \frac{\gamma}{2}||\boldsymbol{\alpha}||_2^2.
\end{aligned}$$

We thus have, for a constant c independent of $\boldsymbol{\alpha}$,

$$
\begin{aligned}
\log p(\boldsymbol{\alpha}) &= \log\left((2\pi)^{n/2}\det[(\lambda L+\gamma I)^{-1}]^{-1/2}\exp\left(-\frac{\lambda}{2}\boldsymbol{\alpha}^T L\boldsymbol{\alpha}-\frac{\gamma}{2}||\boldsymbol{\alpha}||_2^2\right)\right)\\
&= -\frac{\lambda}{2}\boldsymbol{\alpha}^T L\boldsymbol{\alpha}-\frac{\gamma}{2}||\boldsymbol{\alpha}||_2^2+c.
\end{aligned}
$$

Plugging this is in and dropping the constants not affecting the maximizer yields

$$
\begin{aligned}
\boldsymbol{\theta}^*_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}}\left[\ell(\boldsymbol{\theta};\mathcal{D})-\frac{\lambda}{2}\sum_{\boldsymbol{\alpha}\in\theta_{\mathcal{A}}}\boldsymbol{\alpha}^T L\boldsymbol{\alpha}-\frac{\gamma}{2}\sum_{\boldsymbol{\alpha}\in\theta_{\mathcal{A}}}||\boldsymbol{\alpha}||_2^2\right]\\
&= \arg\max_{\boldsymbol{\theta}}\ell_L(\boldsymbol{\theta};\mathcal{D})\\
&= \boldsymbol{\theta}^*_{\text{MLE}}.
\end{aligned}
$$

□

Notice that the Gaussian in the theorem above has precision (i.e., inverse covariance) matrix $\lambda L+\gamma I$. The partial correlation between the per-chooser parameters $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$, $i\neq j$, (controlling for all other nodes) is therefore

$$
-\frac{\lambda L_{ij}}{\sqrt{(\lambda L_{ii}+\gamma)(\lambda L_{jj}+\gamma)}}=\frac{\lambda A_{ij}}{\sqrt{(\lambda d_i+\gamma)(\lambda d_j+\gamma)}} \tag{4.4}
$$

using the standard Gaussian identity relating precision and partial correlation (Liang et al., 2015) (where d_i is the degree of i). If both $d_i, d_j > 0$ and γ is small, then we can approximate

$$
\frac{\lambda A_{ij}}{\sqrt{(\lambda d_i+\gamma)(\lambda d_j+\gamma)}}\approx\frac{\lambda A_{ij}}{\sqrt{(\lambda d_i)(\lambda d_j)}}=\frac{A_{ij}}{\sqrt{d_i d_j}}. \tag{4.5}
$$

This is easy to interpret: $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ have partial correlation 0 when i and j are unconnected ($A_{ij}=0$) and positive partial correlation when they are connected (larger when they have fewer other neighbors). That is, the Gaussian prior in the theorem assumes neighboring nodes have correlated preferences.

Laplacian-regularized logit models

To incorporate Laplacian regularization in our four logit models (logit, MNL, CL, CML), we add per-chooser utilities v_{ia} for each item i and chooser a to the utility formulations. For instance, this results in the following utility function for a per-chooser MNL: $u_\theta(i, C, a) = u_i + \boldsymbol{x_a}^T \boldsymbol{\gamma_i} + v_{ia}$. While we could get rid of the global utilities u_i, L_2 regularization enables us to learn a parsimonious model where u_i is the global baseline utility and v_{ia} represents per-chooser deviations. The per-chooser parameters of a Laplacian-regularized logit are $\theta_A = \{\boldsymbol{v_i}\}_{i\in U}$, where the vector $\boldsymbol{v_i}$ stacks the values of v_{ia} for each chooser $a \in A$. All other parameters are global. The Laplacian- and L_2-regularized log-likelihood can then be written down by combining Equations (4.2) and (4.3). Crucially, since the Laplacian is positive semi-definite, the terms $-\frac{\lambda}{2}\boldsymbol{v_i}^T L \boldsymbol{v_i}$ are concave—and since all four logit log-likelihoods are concave (as is the L_2 regularization term), their regularized negative log likelihoods (NLLs) are convex. This enables us to easily learn maximum-likelihood models with standard convex optimization methods.

4.3.2 Graph neural networks

Graph neural networks (GNNs) (Wu et al., 2020) use a graph to structure the aggregations performed by a neural network, allowing parameters for neighboring nodes to influence each other. We test the canonical GNN, a graph convolutional network (GCN) (Kipf and Welling, 2017), where node embeddings are averaged across neighbors before each neural network layer. There are many other types of GNNs (see (Wu et al., 2020) for a survey)—we emphasize that we do not claim this particular GCN approach to be optimal for discrete choice. Rather, we illustrate how GNNs can be applied to choice data and encourage further exploration.

In a depth-d GCN, each layer performs the following operation, producing a sequence of embeddings $H^{(0)}, \dots, H^{(d)}$:

$$H^{(i+1)} = \sigma(A' H^{(i)} W^{(i)}) \tag{4.6}$$

where $H^{(0)}$ is initialized using node features (if they are available—if not, $H^{(0)}$ is learned), σ is an activation function, $W^{(i)}$ are parameters, and $A' = (D + 2I)^{-1/2}(A + I)(D + 2I)^{-1/2}$ is the degree-normalized adjacency matrix (with self-loops). Self-loops are added to G to allow a node's current embedding to influence its embedding in the next layer. We can either use $H^{(d)}$ as the final embeddings or concatenate each layer's embedding into a final embedding H. In our experiments, we use a two-layer GCN (both with output dimension 16) and concatenate the layer embeddings. For simplicity, we fix the dropout rate at 0.5.

To apply GCNs to discrete choice, we can treat the final node embeddings as chooser features and apply an MNL, modeling utilities as $u_\theta(i, C, a) = u_i + H_a^T \boldsymbol{\gamma_i}$, where u_i and $\boldsymbol{\gamma_i}$ are per-item parameters (the intercept and embedding coefficients, respectively). If item features are also available, we add the conditional logit term $\boldsymbol{\theta}^T \boldsymbol{y_i}$. Thanks to automatic differentiation software such as PyTorch (Paszke et al., 2019), we can train both the GCN and MNL/CML weights end-to-end. Again, any node representation learning method could be used for the embeddings H—we use a GCN for simplicity.

In general, graph neural networks have the advantage of being highly flexible, able to capture complex interactions between the features of neighboring nodes. However some recent research has indicated that non-linearity is less helpful for classification in GNNs than in traditional neural networks tasks (Wu et al., 2019). With the additional modeling power comes significant additional difficulty in training and hyperparameter selection (for embedding dimensions, depth, dropout rate, and activation function).

4.3.3 Choice fraction propagation

We also consider a baseline method that uses the graph to directly derive choice predictions, without a probabilistic model of choice. We extend label propagation (Zhou et al., 2003; Jia and Benson, 2022) to multi-alternative discrete choice. The three features distinguishing the choice setting from standard label propagation is that we can observe multiple "labels" (i.e., choices) per chooser, each observation may have had different available labels, and that not all labels are available at inference time. Given training data of observed choices of the form (i, C, a), where chooser $a \in \mathcal{A}$ chose item $i \in C \subseteq \mathcal{U}$, we assign each chooser a a vector $\boldsymbol{z_a}$ of size $k = |U|$ with each item's *choice fraction*. That is, the ith entry of $\boldsymbol{z_a}$ stores the fraction of times a chose i in the observed data out of all opportunities they had to do so (i.e., the number of times i appeared in their choice set). We use choice fraction rather than counts to normalize by the number of observations for a chooser and not to count against an item instances when it was not available.

We then apply label propagation to the vectors $\boldsymbol{z_a}$ over G. Let $Z^{(0)}$ be the matrix whose rows are $\boldsymbol{z_a}$. As in standard label propagation, we iterate $Z^{(i+1)} \leftarrow (1-\rho)Z^{(0)} + \rho D^{-1/2} A D^{-1/2} Z^{(i)}$ until convergence, where $\rho \in [0, 1]$ is a hyperparameter that controls the strength of the smoothing. Let $Z^{(\infty)}$ denote the stationary point of the iterated map. For inference, we can use the ath row of $Z^{(\infty)}$ (in practice, we will have a matrix arbitrarily close to $Z^{(\infty)}$), denoted $\boldsymbol{z_a^{(\infty)}}$, to make predictions for chooser a. Given a choice set C, we predict a will choose the argmax of $\boldsymbol{z_a^{(\infty)}}$ restricted to items appearing in C. Note that in a semi-supervised

Table 4.1: Dataset summary. $|A|$: number of choosers (aggregated at the county/precinct for elections), $|U|$: number of items, $|C|$: choice set sizes, N: number of observed choices, d_x: number of chooser features.

Dataset	$\|A\|$	$\|U\|$	$\|C\|$	N	d_x	d_y
APP-INSTALL	139	127	51–127	4,039	—	—
APP-USAGE	104	121	2–55	20,564	—	1
US-ELECTION-2016	3,112	32	3–22	135,382,576	19	—
CA-ELECTION-2016	21,495	182	2	261,278,336*	17	—
CA-ELECTION-2020	17,282	170	2	225,606,176*	17	—

*Voters had more than one election (i.e., choice) on their ballots.

setting, we do not observe any choices from the test choosers, so their entries of $Z^{(0)}$ will be zero. The term $(1-\rho)Z^{(0)}$ then acts as a uniform prior, regularizing the test chooser entries of $Z^{(\infty)}$ towards 0. Since choice fraction propagation does not use chooser or item features, it is best suited to scenarios where neither are available.

4.4 Networked discrete choice data

We now describe several datasets in which we can leverage social network structure for improved choice prediction using the methods we develop. Table 4.1 shows a summary of our datasets, which are available at https://osf.io/egj4q/.

4.4.1 Friends and Family app data

The Friends and Family dataset (Aharony et al., 2011) follows over 400 residents of a young-family community in North America during 2010-2011. The dataset is remarkably rich, capturing many aspects of the participants' lives. For instance, they were given Android phones with purpose-made logging software that captured app installation and usage as well as Bluetooth pings between participants' phones. We use the installation and usage data to construct two separate choice datasets (APP-INSTALL and APP-USAGE) and use a network built from Bluetooth pings, as in (Aharony et al., 2011). We ignore uncommon and built-in apps (for instance, we ignore apps whose package names begin with `com.android`, `com.motorola`, `com.htc`, `com.sec`, and `com.google`), leaving a universe $\mathcal{U}$ of 127 apps in APP-INSTALL and 121 in APP-USAGE (e.g., Twitter, Facebook, and Myspace).

To construct APP-INSTALL, we use scans that checked which apps were installed on each participant's phone every 10 minutes – 3 hours. Each time a new app i appears in a scan for a participant a, we consider that a choice from the set of apps C that were not installed at the time of the last scan. We use a plain logit as the baseline model in APP-INSTALL, since no item features are readily available. To construct APP-USAGE, we use 30-second resolution scans of running apps. To separate usage into sessions, we select instances where a participant ran an app for the first time in the last hour. We consider such app runs to be a choice i from the set of all apps C installed on participant a's phone at that time. Our discrete choice approach enables us to account for these differences in app availability. In APP-USAGE, we use a conditional logit with a single instance-specific item feature: *recency*, defined as $\log^{-1}$(seconds since last use) or 0 if the user has not used the app. While it would be possible to construct more complex sets of features with additional effort (for instance categorizing different types of apps or tracking down their Android store ratings), a simple baseline suffices to demonstrate how social network structure can benefit choice modeling even in the absence of item and user features.

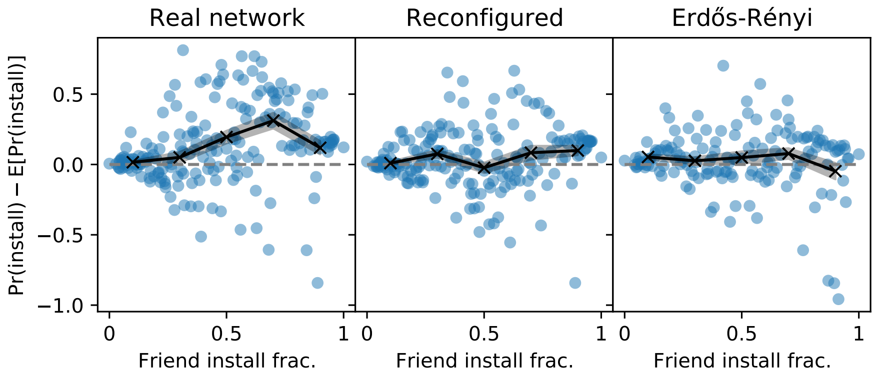


Figure 4.1: Difference between actual and expected install rates (if friendships were irrelevant). The left subplot is with the real network, while the right two are two null models. Each dot is a (participant, app) pair. The black line marks the mean over five bins, with the shaded region showing the standard error of the mean.

To form the social network G over participants for both datasets, we use Bluetooth proximity hits—like the original study (Aharony et al., 2011), we only consider hits in the month of April between 7am and midnight (to avoid coincidental hits from neighbors at night). For each participant a, we form the link (a, b) to each of their 10 most common interaction partners b (we also tested thresholds 2–9, but our methods all performed very similarly). We perform this thresholding because the Bluetooth ping network is extremely dense and contains many edges that are likely not socially meaningful (for instance, nearby phones may ping each other when two strangers shop in the same store). Prior research on this data found that social contacts were useful in predicting app installations, but did not employ a discrete choice approach (Aharony et al., 2011; Pan et al., 2011). Our discrete choice approach allows us to account for multi-hop social connections and the context of each installation (i.e., what apps were already installed).

As a warm-up data analysis, we show that people are more likely to install an app the more of their friends have it (but not if we randomize friendships). Let n

the total number of people, n_i be the number of people who installed application i, f_a the number of friends of person a, and f_{ai} the number of friends of person a who have app i. Suppose app installations are independent of friendships. If we sample some person a uniformly at random and check which of their friends have app i, then the probability that a also has app i is $(n_i - f_{ai})/(n - f_a)$ (simply the remaining fraction of people who have the app, after observing the friends of a). However, if app installations correlate across friendships, the observed probability would be higher when f_{ai}/f_a is larger. We measure the empirical probability that a person has an app at different friend-installation fractions. Specifically, we measure

$$\frac{1}{nk} \sum_{i \in U, a \in \mathcal{A}} \left(\mathbf{1}_{ai} - \frac{n_i - f_{ai}}{n - f_a} \right), \tag{4.7}$$

where $\mathbf{1}_{ai}$ is an indicator for whether person a has app i. Notice that if friendships are uncorrelated with app installations, the expectation of the summand is 0. Instead of taking the mean over all app pairs, we take the mean at each unique friend-installation fraction to see if having more friends with an app results in stronger deviations from uniform installations. This is exactly what we observe: when people have more friends with an app, they are more likely to install it (Figure 4.1). In contrast with two null models (a configuration model with the same degree distribution and an Erdős–Rényi graph with the same density), we see an increase in peoples' installation probabilities as a larger fraction of their friends have an app. This is in line with findings that the probability an individuals joins a social network community increases with the number of their friends in the community (Backstrom et al., 2006). However, it is worth emphasizing that this finding is purely correlational—we have no way of knowing whether increased installation rates are due to homophily in the social network, word-of-mouth contagion, or other confounding factors.

Figure 4.2: 2016 US presidential election vote shares for conservative independent Evan McMullin. Notice his regional popularity and the spillover from Utah to southeast Idaho. McMullin was not on the ballot in filled-in states. The lack of spillover into Colorado may be due to its crowded field (22 candidates) or because it is less conservative than Idaho.

4.4.2 County-level US presidential election data

US presidential election data is a common testbed for graph learning methods using a county-level adjacency network, but the typical approaches are to treat elections as binary classification or regression problems (predicting the vote shares of one party) (Jia and Benson, 2020; Zhou et al., 2020; Huang et al., 2021). However, this ignores the fact that voters have more than two options—moreover, different candidates can be on the ballot in different states. The universe of items $\mathcal{U}$ in our 2016 election data contains no fewer than 31 different candidates (and a "none of these candidates" option in Nevada, which received nearly 4% of the votes in one county). While third-party candidates are unlikely to win in the US, they often receive a non-trivial (and quite possibly consequential) fraction of votes. For instance, in the 2016 election, independent candidate Evan McMullin received 21.5%

of the vote in Utah, while Libertarian candidate Gary Johnson and Green Party candidate Jill Stein received 3.3% and 1.1% nationally (the gap between Clinton and Trump was only 2%). A discrete choice approach enables us to include third-party candidates and account for different ballots in different states. As a visual example, in Figure 4.2 we show the states in which McMullin appeared on the ballot as well as his per-county vote share. By accounting for ballot variation, we can make counterfactual predictions about what would happen if different candidates had been on the ballot, which is difficult without a discrete choice framework. For example, given McMullin's regional support in Utah, it is possible that he would have fared better in Nevada (Utah's western neighbor) than in an East Coast state like New York. Using the entire ballots also allows us to account for one possible reason why McMullin's vote share appears not to spilled over into Colorado, while it did into Idaho: Colorado had fully 22 candidates on the ballot, while Idaho only had 8. A discrete choice approach handles this issue cleanly, while regression on vote shares does not. We note that, due to inherent limitations of observational data, we cannot be sure of the causes of the effects we observe (Tomlinson et al., 2021)—nonetheless, a discrete choice approach enables more flexible modeling and can improve prediction performance regardless of the cause of preference correlations.

We gathered county-level 2016 presidential voting data from (Kearney, 2018) and county data from Jia and Benson (2020),[3] which includes a county adjacency network, county-level demographic data (e.g., education, income, birth rates, USDA economic typology,[4] and unemployment rates), and the Social Connectedness Index (SCI) (Bailey et al., 2018) measuring the relative frequency of Facebook

[3]One county—Oglala Lakota County, South Dakota (FIPS 46102)—was named Shannon County (FIPS 46113) until 2015, which resulted in some missing data. We manually renamed it in the data and extrapolated missing data from previous years.

[4]https://www.ers.usda.gov/data-products/county-typology-codes/

friendships between each pair of counties. We aggregate all votes at the county level, treating each county as a chooser a and using county features as $\boldsymbol{x}_a$ (modeling voting choices in aggregate is standard practice (Alvarez and Nagler, 1998)). For the graph G, we tested using both the geographic adjacency network and a network formed by connecting each county to the 10 others with which it has the highest SCI. We found almost identical results with both networks, so we only only discuss the results using the SCI network. We refer to the resulting dataset as US-ELECTION-2016.

4.4.3 California precinct-level election data

The presidential election data is particularly interesting because different ballots have different candidates, all running in the same election. For instance, this is analogous to having different regional availability of goods within a category in an online shopping service. In our next two datasets, CA-ELECTION-2016 and CA-ELECTION-2020, we highlight a different scenario: when ballots in different locations may have different *elections*. Extending the online shopping analogy, this emulates the case where different users view different recommended categories of items. Although it is beyond the scope of the present work, a discrete choice approach would enable measuring cross-election effects, such as coattail effects (Hogan, 2005; Ferejohn and Calvert, 1984) where higher-office elections increase excitement for down-ballot races.

To construct these datasets, we used data from the 2016 and 2020 California general elections from the Statewide Database.[5] This includes per-precinct registration and voting data as well as shapefiles describing the geographic boundaries

[5] https://statewidedatabase.org; 2016 and 2020 data accessed 8/20/20 and 3/22/21, resp.

of each precinct (California has over 20,000 voting precincts). The registration data contains precinct-level demographics (counts for party affiliation, sex, ethnicity, and age ranges), although such data was not available for all precincts. We restrict the data to the precincts for which all three data types were available: voting, registration, and shapefile (99.8% of votes cast are included in our processed 2016 data, and 99.0% in our 2020 data). Again, we treat each precinct as a chooser a with demographic features $\boldsymbol{x}_a$.

Our processed California data includes elections for the US Senate, US House of Representatives, California State Senate, and California ballot propositions. We set aside presidential votes due to overlap with the previous dataset and state assembly votes to keep the data size manageable. Due to California's nonpartisan top-two primary system,[6] there are two candidates running for each office — however, each voter has a different set of elections on their ballot due to differences in US congress and California state senate districts (the state has 53 congressional districts and 40 state senate districts). A discrete choice approach enables us to train a single model accounting for preferences over all types of candidates. We use the precinct adjacency network G (since SCI is not available at the finer-grained precinct level), which we constructed from the Statewide Database shapefiles using QGIS (https://qgis.org).

4.5 Empirical results

We begin by demonstrating the sample complexity benefit of using network structure through Laplacian regularization on synthetic data. We then apply all

[6] https://www.sos.ca.gov/elections/primary-elections-california

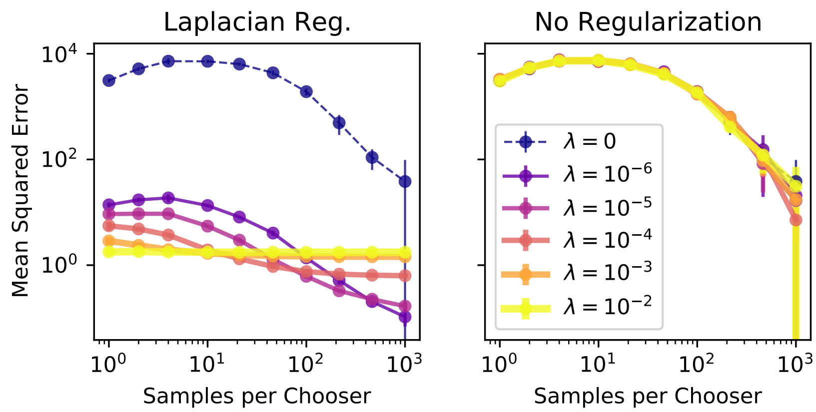


Figure 4.3: Estimation error of item utilities with (left) and without (right) Laplacian regularization on synthetic data generated according to the priors in Theorem 7, with varying homophily strength λ. Error bars (most are tiny) show standard error over 8 trials. Using Laplacian regularization can improve sample complexity by orders of magnitude.

three approaches to our datasets, compare their performance, and demonstrate the insights provided by a networked discrete choice approach. See Table 4.1 in Section 4.4 for a dataset overview. Our code and instructions for reproducing results from this chapter are available at `https://github.com/tomlinsonk/graph-based-discrete-choice/`.

4.5.1 Improved sample complexity with Laplacian regularization

By leveraging correlations between node preferences through Laplacian regularization, we need fewer samples per node in order to achieve the same inference quality. When preferences are smooth over the network, an observation of a choice by one node gives us information about the preferences of its neighbors (and its neighbors' neighbors, etc.), effectively increasing the usefulness of each observation. In Figure 4.3, we show the sample complexity benefit of Laplacian regularization in synthetic data with 100-node Erdős–Rényi graphs ($p = 0.1$) and preferences over 20 items generated according to the prior from Theorem 7. In each of 8 trials, we generate the graph, sample utilities, and then simulate a varying number of choices by each chooser. We repeat this for different homophily strengths λ. For each simulated choice, we first draw a choice set size uniformly between 2 and 20, then pick a uniformly random choice set of that size. We then measure the mean-squared error in inferred utilities of observed items (fixing the utility of the first item to 0 for identification). When applying Laplacian regularization, we use the corresponding value of λ used to generate the data (in real-world data, this needs to be selected through cross-validation). We train the models for 100 epochs.

In this best-case scenario, we need orders of magnitude fewer samples per chooser if we take advantage of preference correlations: with Laplacian regularization, estimation error with only 1 sample per chooser is lower than the estimation error with no regularization and 1000 samples per chooser. The stronger the homophily, the fewer observations are needed to achieve optimal performance, since a node's neighbor's choices are more informative.

4.5.2 Prediction performance comparison

We now evaluate our approaches on real-world choice data. In the style of semi-supervised learning, we use a subset of choosers for training and held-out choosers for validation and testing. This emulates a scenario where it is too expensive to gather data from everyone in the network or existing data is not available for all nodes (e.g., perhaps not all individuals have consented to choice data collection). We vary the fraction of training choosers from 0.1 to 0.8 in increments of 0.1, using half of the remaining choosers for validation and half for testing. We perform 8 independent sampling trials at each fraction in the election datasets and 64 in the smaller Friends and Family datasets.

As a baseline, we use standard logit models with no network information. For the election datasets, we use an MNL that uses county/precinct features to predict votes. This approach to modeling elections is common in political science (Dow and Endersby, 2004). For APP-INSTALL, we use a simple logit. For APP-USAGE, we use a conditional logit (CL) with recency (as defined in Section 4.4.1). We then compare the three graph-based methods we propose to the baseline choice model: a GCN-augmented MNL (or CML), a Laplacian-regularized logit (or CL/MNL) with per-chooser utilities, and choice fraction propagation. Aside from propagation, we train the other methods with batch Rprop (Riedmiller and Braun, 1993), as implemented in PyTorch (Paszke et al., 2019). For each dataset–model pair, we select the hyperparameters that result in the lowest validation loss in a grid search; we tested learning rates $10^{-3}, 10^{-2}, 10^{-1}$ and L_2 regularization strengths 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1} (we also tested no L_2 regularization in the two app datasets). We similarly select Laplacian λ using validation data from $10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$ in the election datasets (in addition to these, we also test $10^{0}, 10^{-1}, 10^{-6}, 10^{-7}$ in the app datasets) and propagation ρ from $0.1, 0.25, 0.5, 0.75, 1$. The smaller hyperparameter ranges in the election datasets were used due to runtime constraints. We train the likelihood-based models for 100 epochs, or until the squared gradient magnitude falls below 10^{-8}. For propagation, we perform 256 iterations, breaking if the sum of squared differences between consecutive iterates falls below 10^{-8}. We note that we did not aggressively fine-tune the GCN beyond learning rate and L_2 regularization strength, since it has many more hyperparameters than our other approaches and is more expensive to train. Our GCN results should therefore be interpreted as the performance a discrete choice practitioner should expect to

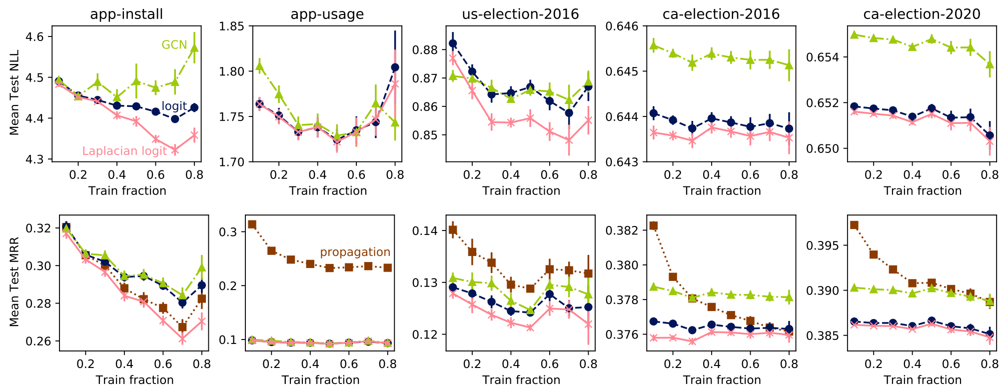


Figure 4.4: Test negative log likelihoods (NLL; top row; lower is better) and mean relative ranks (MRR; bottom row; lower is better) on the two Friends and Family datasets and three election datasets (error bars show standard error over chooser sampling). "Logit" signifies plain logit in APP-INSTALL, CL in APP-USAGE, and MNL in the election datasets. Laplacian regularization improves performance in APP-INSTALL, while no method improves on CL in APP-USAGE. In the election data, Laplacian MNL, but not GCN, outperforms MNL across train fractions. Propagation performs well on APP-INSTALL, but very poorly on APP-USAGE, as it does not utilize recency. Despite not using county/precinct features, propagation can be competitive in the election data.

Table 4.2: Runtime in seconds to train and test each model, with standard error over 4 trials.

Dataset	**CL/MNL**	**Laplacian**	**GCN**	**Propagation**
APP-INSTALL	2.5 ± 0.0	2.2 ± 0.0	9.0 ± 0.2	0.2 ± 0.0
APP-USAGE	12 ± 0	13 ± 0	41 ± 0	0.9 ± 0.0
US-ELECTION-2016	18 ± 0	19 ± 0	20 ± 0.0	1.0 ± 0.0
CA-ELECTION-2016	605 ± 6	647 ± 3	758 ± 4	63 ± 0
CA-ELECTION-2020	450 ± 79	397 ± 2	485 ± 51	38 ± 0

achieve in a reasonable amount of time using the model, which we believe is an important metric.

In Figure 4.4, we show results of all four approaches on all five datasets. We evaluate the three likelihood-based methods using their test set negative log likelihood (NLL) and use *mean relative rank* (MRR) (Tomlinson and Benson, 2021) to evaluate propagation. For one sample, MRR is defined as the relative position of the actual choice in the list of predictions in decreasing confidence order (where 0 is the beginning of the list and and 1 is the end). We then report the mean MRR over the test set. In APP-INSTALL, both Laplacian regularization and propagation improve prediction performance over the baseline logit model, and the advantage increases with the fraction of participants used for training (up to 6.8% better MRR). However, the GCN performs worse than logit in terms of likelihood and the same or worse in terms of MRR. In contrast, graph-based methods do not outperform a conditional logit in APP-INSTALL. In the three election datasets, Laplacian-regularized MNL consistently outperforms MNL (with up to 2.6% better MRR in US-ELECTION-2016; the margin in the California data is small but outside errorbars), while the GCN performs on par with MNL in US-ELECTION-2016 and worse in the California datasets.

These results yield insight into the role networks play in different choice behaviors. In APP-USAGE, we find no benefit from using social network structure using any method. Instead, the recency feature appears to dominate, with propagation (which has no access to item features) performing much worse than the three models that do incorporate recency. This indicates that app usage is driven by individual habit rather than by external social factors. On the other hand, our results show that app *installation* has a strong social component: even simple Bluetooth proximity between friends provides a signal that they will install (but not necessarily use) similar apps. This finding highlights how combining a discrete choice approach with network data can illuminate the role social networks play in different choice behaviors. In the election data, especially CA-ELECTION-2016, even simple choice propagation performs remarkably well, despite *entirely ignoring demographic features*. This reveals that many of the important predictive demographic features (such as party affiliation, age, and ethnicity) are so strongly correlated over the adjacency network that we don't need to know information about you to predict your vote: it suffices to know about your neighbors or your neighbors' neighbors.

We also compare the runtime of each method. To measure runtime, each model was run on a 50-25-25 train-validation-test split of each dataset four times. Since the hyperparameters are not crucial for runtime measurements (especially because Rprop is not sensitive to initial learning rate as an adaptive method), we fixed the learning rate at 0.01, L_2 regularization strength at 0.001, Laplace regularization strength at 0.0001, and propagation ρ at 0.5. For each trial, we trained and tested each model once, shuffling the order of models to avoid systematic bias due to caching. Laplacian regularization has very low overhead over CL/MNL, while GCN is up to 4× slower in the smaller datasets (see Table 4.2). In the larger datasets, PyTorch's built-in parallelism reduces this relative gap. Propagation is more than 10× faster than the choice models in every dataset.

4.5.3 Facebook and Myspace communities in APP-INSTALL

Given that we observed significant improvement in prediction performance in APP-INSTALL, we take a closer look at the patterns learned by the Laplacian-regularized logit compared to the plain logit. In particular, the Facebook and Myspace apps were in the top 20 most-preferred apps under both models ((see Appendix C.1 for the full lists)). Given that these were competitor apps at the time,[7] we hypothesized that they might be popular among different groups of participants. This is exactly what we observe in the learned parameters of the Laplacian-regularized logit. Facebook and Myspace are in the top 10 highest-utility apps for 70 and 27

[7]The dataset is from 2010; Facebook surpassed Myspace's popularity in the US in 2009.

Table 4.3: Edge densities within/between the groups preferring Facebook ($|F| = 70$) and Myspace ($|M| = 27$) in APP-INSTALL. Left: including the 3 choosers in $F \cap M$. Right: excluding $F \cap M$.

	F	M
F	11.2%	4.9%
M	4.9%	11.7%

	F	M
F	11.3%	5.8%
M	5.8%	12.0%

Table 4.4: Maximum likelihood 2016 election outcomes under our model under the three scenarios in Section 4.5.4. We show mean vote shares (with 95% confidence interval over trials) for the top three predicted candidates and differences in state outcomes between the counterfactual prediction and reality. C: Clinton, T: Trump, Outcome: Electoral College votes. T → C denotes that a state won by Trump goes for Clinton under the model. States abbreviated by postal code.

	Scenario 1	**Scenario 2**	**Scenario 3**
C %	47.7 ± 0.1	50.7 ± 0.1	37.5 ± 2.2
T %	46.3 ± 0.1	49.3 ± 0.1	37.0 ± 1.9
T → C	—	—	PA
C → T	ME*, MN, NV, NH	ME*, MN, NV, NH	MN, NV, NH
Other	—	—	RI ("None")
Outcome	T 326, C 205	T 326, C 205	T 304, C 223

*Maine allocates Electoral College votes proportionally—we assume a 3-1 split.

participants, respectively (out of 139 total; we refer to these sets as F and M). Intriguingly, the overlap between F and M is only 3. Moreover, looking at the Bluetooth interaction network, we find the edge densities are more than twice as high within each of F and M than between them (Table 4.3), indicating they are true communities in the social network. In short, the Laplacian-regularized logit learns about two separate subcommunities, one in which Facebook is popular and one in which Myspace is popular.

4.5.4 Counterfactuals in the 2016 US election

One of the powerful uses of discrete choice models is applying them to counterfactual scenarios to predict what might happen under different choice sets (e.g., in assortment optimization (Rusmevichientong et al., 2010)). For instance, we can use our models to make predictions about election outcomes if different candidates had been on ballots in 2016. However, we begin this exploration with a warning: making predictions from observational data is subject to *confounders*, unobserved factors that affected both who was on which ballot and how the states voted. For example, only Nevadans had the option to vote for "None of these options," and Nevada is an outlier in a number of ways that are likely to impact voting, including its reliance on tourism, high level of diversity, and lack of income tax. This makes it less likely that the preferences of Nevadans for "None of these options" will neatly generalize to voters in other states. There are causal inference methods of managing confounding in discrete choice models; for instance, our county covariates act as regression controls (Tomlinson et al., 2021). If those covariates fully described variation in county voting preferences, then the resulting choice models would be unbiased, even with confounding (Tomlinson et al., 2021). However, we do not believe the covariates fully describe voting, since we can improve prediction by using regional or social correlations not captured by the county features. Nonetheless, examining our model's counterfactual predictions is still instructive, demonstrating an application of choice models, providing insight into the model's behavior, and motivating randomized experiments to test predictions about the effect of ballot changes. We note that the MNL we use obeys IIA, preventing relative preferences for candidates changing within a particular county when choice sets change. However, since states contain many counties, they are mixtures of MNLs (which can violate IIA), so their outcomes can change under the model.

A widespread narrative of the 2016 election is that third-party candidates cost Clinton the election by disproportionately taking votes from her (Chalabi, 2016; Rothenberg, 2019). To test this hypothesis, we examine three counterfactual scenarios: (Scenario 1) all ballots have five options: Clinton, Trump, Johnson, Stein, and McMullin; (Scenario 2) ballots only list Clinton and Trump, and (Scenario 3) ballots are as they were in 2016, but "None of these candidates" is added to every ballot. For each scenario, we take the best (validation-selected) Laplacian-regularized MNL trained on 80% of counties from each of the 8 county sampling trials and average their vote count predictions. Maximum-likelihood outcomes under the model are shown in Table 4.4. We find no evidence to support the claim that third-party candidates hurt Clinton more than Trump. None of the scenarios changed the two major measures of outcome: Clinton maintained the popular vote advantage, while Trump carried the Electoral College. A few swing states change hands in the predictions. The model places more weight on "None of these candidates" than seems realistic (for instance, predicting it to be the plurality winner in Rhode Island), likely because training data is only available for this option in a single state, leading to confounding. We also note that under the true choice sets, the model's maximum likelihood state outcomes are the same as in Scenarios 1 and 2. A more complete analysis would examine the full distribution of Electoral College outcomes rather than just the maximum likelihood outcome, but we leave such analysis for future work as it is not our main focus.

4.6 Discussion

As we have seen, social and geographic network structure can be very useful in modeling the choices of a group of connected individuals, since people tend to have more similar preferences to their network neighborhood than to distant strangers. Several possible explanations are possible for this phenomenon: people may be more likely to become friends with similarly-minded individuals (homophily) or trends may spread across existing friendships (contagion). Unfortunately, determining whether homophily or contagion is responsible for similar behavior among friends is notoriously difficult (and often impossible (Shalizi and Thomas, 2011)). e We saw poor performance from the GCN relative to the logit models—as we noted, there are many hyperparameters that could be fine-tuned to possibly improve this performance, although this might not be practical for non-experts. Additionally, there are a host of other GNNs that could outperform GCNs in a choice task. Our contributions in this area are to demonstrate how GNN models can be adapted for networked choice problems and to encourage further exploration of such problems. However, our findings are consistent with several lines of recent work that show simple propagation-based methods outperforming graph neural networks (Huang et al., 2021; Wu et al., 2019; He et al., 2020).

There are several interesting avenues for future work in graph-based methods for discrete choice. As we noted, much of the recent machine learning interest in discrete choice (Seshadri et al., 2019; Bower and Balzano, 2020; Rosenfeld et al., 2020; Tomlinson and Benson, 2021) has revolved around incorporating context effects (violations of IIA). Combining our methods with such approaches could answer questions thats are to our knowledge entirely unaddressed in the literature (and possibly even unasked): Do context effects have a social component? If so, what kinds of context effects? Can we improve contextual choice prediction with social structure (in terms of accuracy or sample complexity)? Another natural extension of our work is to use a weighted Laplacian when we have a weighted social network. In another direction, choice data could be studied as an extra signal for community detection in networks, building on our identification of the Facebook and Myspace communities in the Friends and Family data.

Part III

Collective Decision-Making

CHAPTER 5

CHOICE SET OPTIMIZATION UNDER DISCRETE CHOICE MODELS OF GROUP DECISIONS

We begin our exploration of collective decision-making with an application of discrete choice models to groups of choosers. As we discussed in the first part of this dissertation, much of the computational work on choice has been devoted to designing and fitting models for predicting future choices. In addition to prediction, another area of interest is determining effective interventions to influence choice—advertising and political campaigning are prime examples. In heterogeneous groups, the goal might be to encourage consensus (Amir et al., 2015), or, for an ill-intentioned adversary, to sow discord, e.g., amongst political parties (Rosenberg et al., 2020).

One particular method of influence is introducing new alternatives or options. While early economic models assume that alternatives are irrelevant to the relative ranking of options (Luce, 1959; McFadden, 1974), experimental work has consistently found that new alternatives have strong effects on our choices (Huber et al., 1982; Simonson and Tversky, 1992; Shafir et al., 1993; Trueblood et al., 2013). As we have discussed at length, these effects are called *context effects* or *choice set effects*. Direct measurements on choice data have also revealed choice set effects in several domains (Benson et al., 2016; Seshadri et al., 2019).

Here, we pose adding new alternatives as a discrete optimization problem for influencing a collection of decision makers, such as the inhabitants of a city or the visitors to a website. To this end, we consider various models for how someone makes a choice from a given set of alternatives, where the model parameters can be readily estimated from data. In our setup, everyone has a base set of alternatives from which they make a choice, and the goal is to find a set of additional alternatives to optimize some function of the group's joint preferences on the base set. We specifically analyze three objectives: (i) *agreement* in preferences amongst the group; (ii) *disagreement* in preferences amongst the group; and (iii) *promotion* of a particular item (decision).

We use the framework of *discrete choice* (Train, 2009) to probabilistically model a person's choice from a given set of items, called the *choice set*. These models are parameterized for individual preferences, and when fitting parameters from data, preferences are commonly aggregated at the level of a sub-population of individuals. Discrete choice models such as the multinomial logit and elimination-by-aspects have played a central role in behavioral economics for several decades with diverse applications, including forest management (Hanley et al., 1998), social networks formation (Overgoor et al., 2019), and marketing campaigns (Fader and McAlister, 1990). More recently, new choice data and algorithms have spurred machine learning research on models for choice set effects (Ragain and Ugander, 2016; Chierichetti et al., 2018b; Seshadri et al., 2019; Pfannschmidt et al., 2022; Rosenfeld et al., 2020; Bower and Balzano, 2020).

We provide the relevant background on discrete choice models in Section 5.2. From this, we formally define and analyze three choice set optimization problems—Agreement, Disagreement, and Promotion—and analyze them under four discrete choice models: multinomial logit (McFadden, 1974), the context dependent random utility model (Seshadri et al., 2019), nested logit (McFadden, 1978), and elimination-by-aspects (Tversky, 1972). We first prove that the choice set optimization problems are NP-hard in general for these models. After, we identify natural restrictions of the problems under which they become tractable. These restrictions reveal a fundamental boundary: promoting a particular item within a group is easier than minimizing or maximizing consensus. More specifically, we show that restricting the choice models can make Promotion tractable while leaving Agreement and Disagreement NP-hard, indicating that the interaction between individuals introduces significant complexity to choice set optimization.

After this, we provide efficient approximation algorithms with guarantees for all three problems under several choice models, and we validate our algorithms on choice data. Model parameters are learned for different types of individuals based on features (e.g., where someone lives). From these learned models, we apply our algorithms to optimize group-level preferences. Our algorithms outperform a natural baseline on real-world data coming from transportation choices, insurance policy purchases, and online shopping.

5.1 Related work

Our work fits within recent interest from computer science and machine learning on discrete choice models in general and choice set effects in particular. For example, choice set effects abundant in online data has led to richer data models (Ieong et al., 2012; Chen and Joachims, 2016b; Ragain and Ugander, 2016; Seshadri et al., 2019; Makhijani and Ugander, 2019; Rosenfeld et al., 2020; Bower and Balzano, 2020), new methods for testing the presence of choice set effects (Benson et al., 2016; Seshadri et al., 2019; Seshadri and Ugander, 2019), and new learning algorithms (Kleinberg et al., 2017; Chierichetti et al., 2018b). More broadly, there are efforts on learning algorithms for multinomial logit mixtures (Oh and Shah, 2014; Ammar et al., 2014; Kallus and Udell, 2016; Zhao and Xia, 2019), Plackett-Luce models (Maystre and Grossglauser, 2015; Zhao et al.), and other random utility models (Oh et al., 2015; Chierichetti et al., 2018a; Benson et al., 2018b).

One of our optimization problems is maximizing group agreement by introducing new alternatives. This is motivated in part by how additional context can sway opinion on controversial topics (Munson et al., 2013; Liao and Fu, 2014; Graells-Garrido et al., 2014). There are also related algorithms for decreasing polarization in social networks (Garimella et al., 2017; Matakos et al., 2017; Chen et al., 2018; Musco et al., 2018), although we have no explicit network and adopt a choice-theoretic framework.

Our choice set optimization framework is similar to assortment optimization in operations research, where the goal is find the optimal set of products to offer in order to maximize revenue (Talluri and Van Ryzin, 2004). Discrete choice models are extensively used in this line of research, including the multinomial logit (Rusmevichientong et al., 2010, 2014) and nested logit (Gallego and Topaloglu, 2014; Davis et al., 2014) models. We instead focus our attention primarily on optimizing agreement among individuals, which has not been explored in traditional revenue-focused assortment optimization.

Finally, our problems relate to group decision-making. In psychology, introducing new shared information is critical for group decisions (Stasser and Titus, 1985; Lu et al., 2012). In computer science, the complexity of group Bayesian reasoning is a concern (Hązła et al., 2021, 2019).

5.2 Background and preliminaries

We first remind the reader of the basics of discrete choice. For consistency with the published version of this chapter, we use slightly different notation for choice probabilities and utilities than we did in Part II. In the setting we explore, one or more individuals make a (possibly random) *choice* of a single item (or alternative) from a finite set of items called a *choice set*. We use $\mathcal{U}$ to denote the universe of items and $C \subseteq \mathcal{U}$ the choice set. Thus, given C, an individual chooses some item $x \in C$.

Given C, a discrete choice *model* provides a probability for choosing each item $x \in C$. We analyze four broad discrete choice models that are all *random utility models* (RUMs), which derive from economic rationality. In a RUM, an individual observes a random utility for each item $x \in C$ and then chooses the one with the largest utility. We model each individual's choices through the same RUM but with possibly different parameters to capture preference heterogeneity. In this sense, we have a mixture model.

Choice data typically contains many observations from various choice sets. We occasionally have data specific enough to model the choices of a particular individual, but often only one choice is recorded per person, making accurate preference learning impossible at that scale. Thus, we instead model the heterogeneous preferences of sub-populations or categories of individuals. For convenience, we still use "individual" or "person" when referring to components of a mixed population, since we can treat each component as a decision-making agent with its own preferences. In contrast, we use the term "group" to refer to the entire population. We use A to denote the set of individuals (in the broad sense above), and $a \in A$ indexes model parameters.

The parameters of the RUMs we analyze can be inferred from data, and our theoretical results and algorithms assume that we have learned these parameters. Our analysis focuses on how the probability of selecting an item x from a choice set C changes as we add new alternative items from $\overline{C} = \mathcal{U} \setminus C$ to the choice set.

We let $n = |A|$, $k = |C|$, and $m = |\overline{C}|$ for notation. We mostly use $n = 2$, which is sufficient for hardness proofs.

Multinomial logit (MNL)

The *multinomial logit*[1] (MNL) model (McFadden, 1974) is the workhorse of discrete choice theory. In MNL, an individual a's preferences are encoded by a true utility $u_a(x)$ for every item $x \in \mathcal{U}$. The observations are noisy random utilities $\tilde{u}_a(x) = u_a(x) + \varepsilon$, where ε follows a Gumbel distribution. Under this model, the probability that individual a picks item x from choice set C (i.e., $x = \arg\max_{y \in C} \tilde{u}_a(y)$) is the softmax over item utilities:

$$\Pr(a \leftarrow x \mid C) = \frac{e^{u_a(x)}}{\sum_{y \in C} e^{u_a(y)}}. \tag{5.1}$$

We use the term *exp-utility* for terms like $e^{u_a(x)}$. The utility of an item is often parameterized as a function of features of the item in order to generalize to unseen data. For example, a linear function is an additive utility model (Tversky and Simonson, 1993) and looks like logistic regression. In our analysis, we work directly with the utilities.

The MNL satisfies *independence of irrelevant alternatives* (IIA) (Luce, 1959), the property that for any two choice sets C, D and two items $x, y \in C \cap D$:

[1]For consistency with the published version (Tomlinson and Benson, 2020), this chapter uses "multinomial logit" to refer to what we have simply called the "logit" in previous chapters.

$\frac{\Pr(a \leftarrow x|C)}{\Pr(a \leftarrow y|C)} = \frac{\Pr(a \leftarrow x|D)}{\Pr(a \leftarrow y|D)}$. In other words, the choice set has no effect on a's relative probability of choosing x or y.[2] Although IIA is intuitively pleasing, behavioral experiments show that it is often violated in practice (Huber et al., 1982; Simonson and Tversky, 1992). Thus, there are many models that account for IIA violations, including the other ones we analyze.

Context-dependent random utility model (CDM)

The CDM (Seshadri et al., 2019) is an extension of MNL that can model IIA violations. The core idea is to approximate choice set effects by the effect of each item's presence on the utilities of the other items. For instance, a diner's preference for a ribeye steak may decrease relative to a fish option if filet mignon is also available. Formally, each item z exerts a pull on a's utility from x, which we denote $p_a(z, x)$. The CDM then resembles the MNL with utilities $u_a(x \mid C) = u_a(x) + \sum_{z \in C} p_a(z, x)$. This leads to choice probabilities that are a softmax over the context-dependent utilities:

$$\Pr(a \leftarrow x \mid C) = \frac{e^{u_a(x|C)}}{\sum_{y \in C} e^{u_a(y|C)}}. \tag{5.2}$$

Nested logit (NL)

The nested logit (NL) model (McFadden, 1978) instead accounts for choice set effects by grouping similar items into *nests* that people choose between successively. For example, a diner may first choose between a vegetarian, fish, or steak meal and then select a particular dish. NL can be derived by introducing correlation

[2] Over $a \in A$, we have a mixed logit which does not have to satisfy IIA (McFadden and Train, 2000). Here, we are interested in the IIA property at the individual level.

between the random utility noise ε in MNL; here, we instead consider a generalized tree-based version of the model.[3]

The (generalized) NL model for an individual a consists of a tree T_a with a leaf for each item in $\mathcal{U}$, where the internal nodes represent categories of items. Rather than having a utility only on items, each person a also has utilities $u_a(v)$ on all nodes $v \in T_a$ (except the root). Given a choice set C, let $T_a(C)$ be the subtree of T_a induced by C and all ancestors of C. To choose an item from C, a starts at the root and repeatedly picks between the children of the current node according to the MNL model until reaching a leaf.

Elimination-by-aspects (EBA)

While the previous models are based on MNL, the elimination-by-aspects (EBA) model (Tversky, 1972) has a different structure. In EBA, each item x has a set of *aspects* x' representing properties of the item, and person a has a utility $u_a(\chi) > 0$ on each aspect χ. An item is chosen by repeatedly picking an aspect with probability proportional to its utility and eliminating all items that do not have that aspect until only one item remains (or, if all remaining items have the same aspects, the choice is made uniformly at random). For example, a diner may first eliminate items that are too expensive, then disregard meat options, and finally look for dishes with pasta before choosing mushroom ravioli. Formally, let $C' = \bigcup_{x \in C} x'$ be the set of aspects of items in C and let $C^0 = \bigcap_{x \in C} x'$ be the aspects shared by all items in C. Additionally, let $C_\chi = \{x \in C \mid \chi \in x'\}$. The probability that individual a picks item x from choice set C is recursively defined

[3]Certain parameter regimes in this generalized model do not correspond to RUMs (Train, 2009), but this model is easier to analyze and captures the salient structure.

as

$$\Pr(a \leftarrow x \mid C) = \frac{\sum_{\chi \in x' \setminus C^0} u_a(\chi) \Pr(a \leftarrow x \mid C_\chi)}{\sum_{\psi \in C' \setminus C^0} u_a(\psi)}. \tag{5.3}$$

If all remaining items have the same aspects ($C' = C^0$), the denominator is zero, and $\Pr(a \leftarrow x \mid C) = \frac{1}{|C|}$ in that case.

Encoding MNLs in other models

Although the three models with context effects appear quite different, they all subsume the MNL model. Thus, if we prove a problem hard under MNL, then it is hard under all four models.

Lemma 1. *The MNL model is a special case of the CDM, NL, and EBA models.*

Proof. Let $\mathcal{M}$ be an MNL model. For the CDM, use the utilities from $\mathcal{M}$ and set all pulls to 0. For NL, make all items children of T_a's root and use the utilities from $\mathcal{M}$. Lastly, for EBA, assign a unique aspect χ_x to each item $x \in \mathcal{U}$ with utility $u_a(\chi_x) = e^{u_a(x)}$. Following (5.3), $\Pr(a \leftarrow x \mid C) = \frac{u_a(\chi_x) \Pr(a \leftarrow x \mid C_{\chi_x})}{\sum_{\psi \in C' \setminus C^0} u_a(\psi)}$. Since $C_{\chi_x} = \{x\}$, $\Pr(a \leftarrow x \mid C_{\chi_x}) = 1$ and thus $\Pr(a \leftarrow x \mid C) \propto u_a(\chi_x) = e^{u_a(x)}$, matching the MNL $\mathcal{M}$. $\square$

5.3 Choice set optimization problems

By introducing new alternatives to the choice set C, we can modify the relationships amongst individual preferences, resulting in different dynamics at the collective level. Similar ideas are well-studied in voting models, e.g., introducing alternatives to change winners selected by Borda count (Easley and Kleinberg, 2010). Here, we study how to optimize choice sets for various group-level objectives, measured in terms of individual choice probabilities coming from discrete choice models.

Agreement and Disagreement

Since we are modeling the preferences of a collection of decision-makers, one important metric is the amount of disagreement (conversely, agreement) about which item to select. Given a set of alternatives $Z \subseteq \overline{C}$ we might introduce, we quantify the disagreement this would induce as the sum of all pairwise differences between individual choice probabilities over C:

$$D(Z) = \sum_{\{a,b\} \subseteq A, x \in C} |\Pr(a \leftarrow x \mid C \cup Z) - \Pr(b \leftarrow x \mid C \cup Z)|. \tag{5.4}$$

Here, we care about the disagreement on the original choice set C that results from preferences over the new choice set $C \cup Z$. In this setup, C could represent core options (e.g., two major health care policies under deliberation) and Z additional alternatives designed to sway opinions.

Concretely, we study the following problem: given $A, C, \overline{C}$, and a choice model, minimize (or maximize) $D(Z)$ over $Z \subseteq \overline{C}$. We call the minimization problem Agreement and the maximization problem Disagreement. Agreement has applications in encouraging consensus, while Disagreement yields insight into how susceptible a group may be to an adversary who wishes to increase conflict. Another potential application for Disagreement is to enrich the diversity of preferences present in a group.

Promotion

Promoting an item is another natural objective, which is of considerable interest in online advertising and content recommendation. Given $A, C, \overline{C}$, a choice model, and a target item $x^* \in C$, the PROMOTION problem is to find the set of alternatives $Z \subseteq \overline{C}$ whose introduction maximizes the number of individuals whose "favorite" item in C is x^*. Formally, this means maximizing the number of individuals $a \in A$ for whom $\Pr(a \leftarrow x^* \mid C \cup Z) > \Pr(a \leftarrow x \mid C \cup Z)$, $x \in C$, $x \neq x^*$. This also has applications in voting, where questions about the influence of new candidates constantly arise.

One of our contributions in this chapter is showing that promotion can be easier (in a computational complexity sense) than agreement or disagreement optimization.

5.4 Hardness results

We now characterize the computational complexity of AGREEMENT, DISAGREEMENT, and PROMOTION under the four discrete choice models. We first show that AGREEMENT and DISAGREEMENT are NP-hard under all four models and that PROMOTION is NP-hard under the three models with context effects. After, we prove that imposing additional restrictions on these discrete choice models can make PROMOTION tractable while leaving AGREEMENT and DISAGREEMENT NP-hard. The parameters of some choice models have extra degrees of freedom, e.g., MNL has additive-shift-invariant utilities. For inference, we use a standard form (e.g., sum of utilities equals zero). For ease of analysis, we do not use such standard forms, but the choice probabilities remain unambiguous.

5.4.1 Agreement

Although the MNL model does not have any context effects, introducing alternatives to the choice set can still affect the relative preferences of two different individuals. In particular, introducing alternatives can impact disagreement in a sufficiently complex way to make identifying the optimal set of alternatives computationally hard. Our proof of Theorem 8 uses a very simple MNL in the reduction, with only two individuals and two items in C, where the two individuals have *exactly the same utilities on alternatives*. In other words, even when individuals agree about new alternatives, encouraging them to agree over the choice set is hard.

Theorem 8. *In the MNL model, Agreement is NP-hard, even with just two items in C and two individuals that have identical utilities on items in $\overline{C}$.*

Proof. By reduction from Partition, an NP-complete problem (Karp, 1972). Let S be the set of integers we wish to partition into two subsets with equal sum. We construct an instance of Disagreement with $A = \{a, b\}$, $C = \{x, y\}$, $\overline{C} = S$ (abusing notation to identify alternatives with the Partition integers). Let $t = \frac{1}{2}\sum_{z \in S} z$. Define the utilities as: $u_a(x) = \log t$, $u_b(x) = \log 3t$, $u_a(y) = \log t$, $u_b(y) = \log 2t$, and $u_a(z) = u_b(z) = \log z$ for all $z \in \overline{C}$. The disagreement induced by a set of alternatives $Z \subseteq \overline{C}$ is characterized by its sum of exp-utility $s_Z = \sum_{z \in Z} z$:

$$D(Z) = \left|\tfrac{t}{2t+s_Z} - \tfrac{3t}{5t+s_Z}\right| + \left|\tfrac{t}{2t+s_Z} - \tfrac{2t}{5t+s_Z}\right|.$$

The total exp-utility of all items in $\overline{C}$ is $2t$. On the interval $[0, 2t]$, $D(Z)$ is minimized at $s_Z = t$ (Appendix D.1.1; Figure D.1, left). Thus, if we could efficiently find the set Z minimizing $D(Z)$, then we could efficiently solve PARTITION. $\square$

From Lemma 1, the other models we consider can all encode any MNL instance, which leads to the following corollary.

Corollary 1. *AGREEMENT is NP-hard in the CDM, NL, and EBA models.*

5.4.2 DISAGREEMENT

Using a similar strategy, we can construct an MNL instance whose disagreement is maximized rather than minimized at a particular target value (Theorem 9). The reduction requires an even simpler MNL setup.

Theorem 9. *In the MNL model, DISAGREEMENT is NP-hard, even with just one item in C and two individuals that have identical utilities on items in $\overline{C}$.*

Proof. By reduction from SUBSET SUM (Karp, 1972). Let S be a set of positive integers with target t. Let $A = \{a, b\}$, $C = \{x\}$, $\overline{C} = S$, with utilities: $u_a(x) = \log 2t$, $u_b(x) = \log t/2$, and $u_a(z) = u_b(z) = \log z$ for all $z \in \overline{C}$. Letting $s_Z = \sum_{z \in Z} z$, including $Z \subseteq \overline{C}$ makes the disagreement

$$D(Z) = \left| \tfrac{2t}{2t + s_Z} - \tfrac{t/2}{t/2 + s_Z} \right|.$$

For $s_Z \geq 0$, $D(Z)$ is maximized at $s_Z = t$ (Appendix D.1.1; Figure D.1, right). Thus, if we could efficiently maximize $D(Z)$, then we could efficiently solve SUBSET SUM. $\square$

By Lemma 1, we again have the following corollary.

Corollary 2. *Disagreement is NP-hard in the CDM, NL, and EBA models.*

5.4.3 Promotion

In choice models with no context effects, Promotion has a constant-time solution: under IIA, the presence of alternatives has no effect on an individual's relative preference for items in C. However, Promotion is more interesting with context effects, and we show that it is NP-hard for CDM, NL, and EBA. In Section 5.4.4, we will show that restrictions of these models make Promotion tractable but keep Agreement and Disagreement hard.

Theorem 10. *In the CDM model, Promotion is NP-hard, even with just one individual and three items in C.*

Proof. By reduction from Subset Sum. Let set S with target t be an instance of Subset Sum. Let $A = \{a\}$, $C = \{x^*, w, y\}$, $\overline{C} = S$. Using tuples interpreted entry-wise for brevity, suppose that we have the following utilities: $u_a(\langle x^*, w, y\rangle \mid C) = \langle 1, t, -t\rangle$, $u_a(z) = -\infty$ for all $z \in \overline{C}$, and $p_a(z, \langle x^*, w, y\rangle) = \langle z, 0, 2z\rangle$ for all $z \in \overline{C}$. We wish to promote x^*. Let $s_Z = \sum_{z \in Z} z$. When we include the alternatives in Z, x^* is the item in C most likely to be chosen if and only if $1 + s_Z > t$ and $1 + s_Z > -t + 2s_Z$. Since s_Z and t are integers, this is only possible if $s_Z = t$. Thus, if we could efficiently promote x^*, then we could efficiently solve Subset Sum. $\square$

We use the same Goldilocks strategy in our proofs for the NL and EBA models (details in Appendix D.1): by carefully defining utilities, we create choice instances where the optimal promotion solution is to pick just the right quantity of alternatives to increase preference for one item without overshooting. However, the NL model has a novel challenge compared to the CDM. With CDM, alternatives can increase the choice probability of an item in C, but in the NL, new alternatives only lower choice probabilities.

Theorem 11. *In the NL model,* Promotion *is NP-hard, even with just two individuals and two items in* C*.*

This construction relies on the two individuals having different tree structures. We will see in Section 5.4.4 that this is a necessary condition for the hardness of Promotion. Finally, we have the following hardness result for EBA.

Theorem 12. *In the EBA model,* Promotion *is NP-hard, even with just two individuals and two items in* C*.*

5.4.4 Restricted models that make promotion easier

We now show that, in some sense, Promotion is a fundamentally easier problem than Agreement or Disagreement. Specifically, there are simple restrictions on CDM, NL, and EBA that make Promotion tractable but leave Agreement and Disagreement NP-hard. Importantly, these restrictions still allow for choice set effects. In Appendix D.2, we also prove a strong restriction on the MNL model where Agreement and Disagreement are tractable, but we could not find meaningful restrictions for similar results on the other models.

2-item CDM with equal context effects

The proof of Theorem 10 shows that Promotion is hard with only a single individual and three items in C. However, if C only has two items *and* context effects are the same (i.e., $p_a(z, \cdot)$ is the same for all $z \in \overline{C}$), then Promotion is tractable. The optimal solution is to include all alternatives that increase utility for x^* more than the other item, as doing so makes strict progress on promoting x^*. If individuals have different context effects or if there are more than two items, then there can be conflicts between which items should be included (see Appendix D.1.2 for a proof that 2-item CDM with *unequal* context effects makes Promotion NP-hard). Although this restriction makes Promotion tractable, it leaves Agreement and Disagreement NP-hard: the proofs of Theorems 8 and 9 can be interpreted as 2-item and 1-item CDMs with equal (zero) context effects.

Same-tree NL

If we require that all individuals share the same NL tree structure, but still allow different utilities, then promotion becomes tractable. For each $z \in \overline{C}$, we can determine whether it reduces the relative choice probability of x^* based on its position in the tree: adding z decreases the relative choice probability of x^* if and only if z is a sibling of any ancestor of x^* (including x^*) or if it causes such a sibling to be added to $T_a(C)$. Thus, the solution to Promotion is to include all z not in those positions, which is a polynomial-time check. This restriction leaves Agreement and Disagreement NP-hard via Theorems 8 and 9 as we can still encode any MNL model in a same-tree NL using the tree in which all items are children of the root.

Algorithm 1 ε-additive approximation for AGREEMENT in the MNL model.

Input: n individuals A, k items C, m alternatives $\overline{C}$, utilities $u_a(\cdot) > 0$ for each $a \in A$. For brevity:
$e_{ax} \leftarrow e^{u_a(x)}$, $s_a \leftarrow \sum_{z \in \overline{C}} e_{az}$, $\delta \leftarrow \varepsilon/(2km\binom{n}{2})$
$L_0 \leftarrow$ empty n-dimensional array whose ath dimension has size $1 + \lfloor \log_{1+\delta} s_a \rfloor$ (each cell can store a set $Z \subseteq \overline{C}$ and its n exp-utility sums for each individual)

Initialize $L_0[0, \ldots, 0] \leftarrow (\emptyset, 0, \ldots, 0)$ (n zeros)
for $i = 1$ **to** m **do**
  $z \leftarrow \overline{C}[i-1]$, $L_i \leftarrow L_{i-1}$
  for each cell of L_{i-1} containing $(Z, t_1, \ldots, t_n)$ **do**
    $h \leftarrow n$-tuple w/ entries $\lfloor \log_{1+\delta}(t_j + e_{a_j z}) \rfloor$, $\forall j$
    if $L_i[h]$ is empty **then**
      $L_i[h] \leftarrow (Z \cup \{z\}, t_1 + e_{a_1 z}, \ldots, t_n + e_{a_n z})$
$Z_m \leftarrow$ collection of all sets Z in cells of L_m
return $\arg\min_{Z \in Z_m} D(Z)$ (see Equation (5.4))

Disjoint-aspect EBA

The following condition on aspects makes promoting x^* tractable: for all $z \in \overline{C}$, either $z' \cap x^{*\prime} = \emptyset$ or $z' \cap y' = \emptyset$ for all $y \in C$, $y \neq x^*$. That is, alternatives either share no aspects with x^* or share no aspects with other items in C. This prevents alternatives from cannibalizing from both x^* and its competitors. To promote x^*, we include all alternatives that share aspects with competitors of x^* but not x^* itself, which strictly promotes x^*. This condition is slightly weaker than requiring all items to have disjoint aspects, which reduces to MNL. However, this condition is again not sufficient to make AGREEMENT and DISAGREEMENT tractable, since any MNL model can be encoded in a disjoint-aspect EBA instance.

5.5 Approximation algorithms

Thus far, we have seen that several interesting group decision-making problems are NP-hard across standard discrete choice models. Here, we provide a positive result: we can compute arbitrarily good approximate solutions to many instances of these problems in polynomial time. We focus our analysis on Algorithm 1, which is an ε-additive approximation algorithm to AGREEMENT under MNL, with runtime polynomial in k, m, and $\frac{1}{\varepsilon}$, but exponential in n (recall that $k = |C|$, $m = |\overline{C}|$, and $n = |A|$). In contrast, brute force (testing every set of alternatives) is exponential in m and polynomial in k and n. AGREEMENT is NP-hard even with $n = 2$ (Theorem 8), so our algorithm provides a substantial efficiency improvement. We discuss how to extend this algorithm to other objectives and other choice models later in the section. Finally, we present a faster but less flexible mixed-integer programming approach for MNL AGREEMENT and DISAGREEMENT that performs very well in practice.

Algorithm 1 is based on an FPTAS for SUBSET SUM (Cormen et al., 2001, Sec. 35.5), and the first parts of our analysis follow some of the same steps. The core idea of our algorithm is that a set of items can be characterized by its exp-utility sums for each individual and that there are only polynomially many combinations of exp-utility sums that differ by more than a multiplicative factor $1 + \delta$. We can therefore compute all sets of alternatives with meaningfully different impacts and pick the best one. For the purpose of the algorithm, we assume all utilities are positive (otherwise we may access a negative index); utilities can always be shifted by a constant to satisfy this requirement.

We now provide an intuitive description of Algorithm 1. The array L_i has one dimension for each individual in A (we use a hash table in practice since L_i is typically sparse). The cells along a particular dimension discretize the exp-utility sums that the individual corresponding to that dimension could have for a

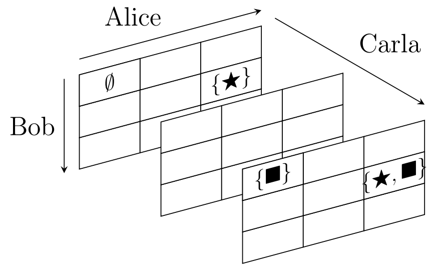


Figure 5.1: Example of the structure L_i used in Algorithm 1 for $n = 3$ individuals and $\overline{C} = \{\bigstar, \blacksquare\}$. Here, Alice has high utility for ★ and low utility for ■, Bob has medium utility for ★ and low utility for ■, and Carla has low utility for ★ and high utility for ■. The exp-utility sums stored in cells are omitted.

particular set of alternatives (Figure 5.1). In particular, if individual j has total exp-utility $t_j = \sum_{y \in Z} e^{u_j(y)}$ for a set Z, then we store Z at index $\lfloor \log_{1+\delta} t_j \rfloor$ along dimension j.

As the algorithm progresses, we place possible sets of alternatives Z in the cells of L_i according to their exp-utility sums $t_1, \ldots, t_n$ for each individual (we store $t_1, \ldots, t_n$ in the cell along with Z). We add one element at a time from $\overline{C}$ to the sets already in L_i (L_0 starts with only the empty set). If two sets have very similar exp-utility sums, they may map to the same cell, in which case only one of them is stored. If the discretization of the array is coarse enough (that is, with large enough δ), many sets of alternatives will map to the same cells, reducing the number of sets we consider and saving computational work. On the other hand, if the discretization is fine enough (δ is sufficiently small), then the best set we are left with at the end of the algorithm cannot induce a disagreement value too different from the optimal set. The proof of Theorem 13 formalizes this reasoning.

Theorem 13. *Algorithm 1 is an ε-additive approximation for* Agreement *in the MNL model.*

Proof. We will use the following lemma, which says that sets mapping to the same cell have similar exp-utility sums.

Lemma 2. *Let $\overline{C}_i$ be the first i elements processed by the outer for loop. At the end of the algorithm, for all $Z \subseteq \overline{C}_i$ with exp-utility sums t_a, there exists some $Z' \in L_i$ with exp-utility sums t'_a such that $\frac{t_a}{(1+\delta)^i} < t'_a < t_a(1+\delta)^i$, for all $a \in A$ (with δ as defined in Algorithm 1, Line 2).*

The proof is in Appendix D.3. Now let $\beta = \varepsilon/(k\binom{n}{2})$. Following our choice of δ and using Lemma 2, at the end of the algorithm, the optimal set $Z^* \subseteq \overline{C}$ (with exp-utility sums t^*_a) has some representative Z' in L_m such that

$$\frac{t^*_a}{(1+\beta/(2m))^m} < t'_a < t^*_a \left(1+\beta/(2m)\right)^m, \ \forall a \in A.$$

Since $e^x \geq (1+x/m)^m$, we have $t^*_a/e^{\frac{\beta}{2}} < t'_a < t^*_a e^{\frac{\beta}{2}}$, and since $e^x \leq 1+x+x^2$ when $x < 1$,

$$\frac{t^*_a}{1+\beta/2+\beta^2/4} < t'_a < t^*_a(1+\beta/2+\beta^2/4).$$

Finally, $\frac{t^*_a}{1+\beta} < t'_a < t^*_a(1+\beta)$ because $0 < \beta < 1$.

Now we show that $D(Z^*)$ and $D(Z')$ differ by at most ε. To do so, we first bound the difference between $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ by β. Let $c_a = \sum_{x \in C} e_{ax}$ be the total exp-utility of a on C. By the above reasoning,

$$\frac{e_{ax}}{c_a + t^*_a(1+\beta)} < \frac{e_{ax}}{c_a + t'_a} < \frac{e_{ax}}{c_a + \frac{t^*_a}{1+\beta}},$$

where the middle term is equal to $\Pr(a \leftarrow x \mid C \cup Z')$. From the lower bound, the difference between $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ could be as large as

$$\begin{aligned}&\frac{e_{ax}}{c_a + t^*_a} - \frac{e_{ax}}{c_a + t^*_a(1+\beta)} \\ &= \frac{e_{ax} t^*_a \beta}{(c_a + t^*_a)(c_a + t^*_a(1+\beta))} < \frac{e_{ax} t^*_a \beta}{2 c_a t^*_a} \leq \frac{\beta}{2}.\end{aligned}$$

From the upper bound, the difference between $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid$

$C \cup Z'$) could be as large as

$$\begin{aligned}
\frac{e_{ax}}{c_a + \frac{t_a^*}{1+\beta}} - \frac{e_{ax}}{c_a + t_a^*} &= \frac{e_{ax} t_a (1 - \frac{1}{1+\beta})}{(c_a + \frac{t_a^*}{1+\beta})(c_a + t_a^*)} \\
&= \frac{e_{ax} t_a^* \beta}{(c_a(1+\beta) + t_a^*)(c_a + t_a^*)} < \frac{e_{ax} t_a^* \beta}{2 c_a t_a^*} \leq \frac{\beta}{2}.
\end{aligned}$$

Thus, $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ differ by at most $\frac{\beta}{2}$. Using the same argument for an individual b, the disagreement between a and b about x can only increase by β with the set Z compared to the optimal set Z^*. Since there are $\binom{n}{2}$ pairs of individuals and k items in C, the total error of the algorithm is bounded by $k\binom{n}{2}\beta = \varepsilon$. □

We now show that the runtime of Algorithm 1 is $O((m+kn^2)(1+\lfloor \log_{1+\delta} s \rfloor)^n)$, where $s = \max_a s_a$ is the maximum exp-utility sum for any individual. Thus, for any fixed n, this runtime is bounded by a polynomial in k, m, and $\frac{1}{\varepsilon}$. To see this, first note that the size of L_i is bounded above by $(1 + \lfloor \log_{1+\delta} s \rfloor)^n$. For each $z \in \overline{C}$, we perform constant-time operations on each entry of L_i, for a total of $O(m(1+\lfloor \log_{1+\delta} s \rfloor)^n)$ time. Then we compute $D(Z)$ for each cell of L_m, which takes $O(kn^2)$ time per cell. The total runtime is therefore $O((m+kn^2)(1+\lfloor \log_{1+\delta} s \rfloor)^n)$, as claimed. Finally, $(1 + \lfloor \log_{1+\delta} s \rfloor)^n$ is bounded by a polynomial in m, k, and $\frac{1}{\varepsilon}$ for any fixed n (Appendix D.3.2).

Agreement is NP-hard even when individuals have equal utilities on alternatives. In this case, we only need to compute exp-utility sums for a single individual, which brings the runtime down to $O((m + kn^2) \log_{1+\delta} s)$.

Extensions to other objectives and models

Algorithm 1 can be easily extended to any objective function that is efficiently computable from utilities. For instance, Algorithm 1 can be adapted for Disagreement by replacing the arg min with an arg max on Line 12.

Algorithm 1 can also be adapted for CDM and NL. The analysis is similar and details are in Appendix D.3, although the running times and guarantees are different. With CDM, the exponent in the runtime increases to nk for Agreement and Disagreement, and the ε-additive approximation is guaranteed only if items in $\overline{C}$ exert zero pulls on each other. However, even for the general CDM, our experiments will show that the adapted algorithm remains a useful heuristic. When we adapt Algorithm 1 for NL, we retain the full approximation guarantee but the exponent in the runtime increases and has a dependence on the tree size.

Promotion is not interesting under MNL and also has a discrete rather than continuous objective, i.e., the number of people with favorite item x^* in C. For models with context effects, we can define a meaningful notion of approximation. We say that an item $y \in C \cup Z$ is an *ε-favorite* item of individual a if $\Pr(a \leftarrow y \mid C \cup Z) + \varepsilon \geq \Pr(a \leftarrow x \mid C \cup Z)$ for all $x \in C$. A solution then ε-approximates Promotion if the number of people for whom x^* is an ε-favorite item is at least the value of the optimal Promotion solution. Using this, we can adapt Algorithm 1 for Promotion under CDM and NL. Again, for CDM, the approximation has guarantees in certain parameter regimes and the NL has full approximation guarantees. Since we do not have compute $D(Z)$, the runtimes loses the kn^2 term compared to the Agreement and Disagreement versions (Appendix D.3.5).

Table 5.1: Dataset statistics: item, observation, and unique choice set counts; and percent of observations in sub-population splits.

Dataset	# items	# obs.	# sets	split %
SFWork	6	5029	12	16/84
Allstate	24	97009	2697	45/55
Yoochoose	41	90493	1567	47/53

Finally, EBA has considerably different structure than the other models. We leave algorithms for EBA to future work.

Fast exact methods for MNL

We provide another approach for solving Agreement and Disagreement in the MNL model, based on transforming the objective functions into mixed-integer bilinear programs (MIBLPs; details in Appendix D.4). MIBLPs can be solved for moderate problem sizes with high-performance branch-and-bound solvers (we use Gurobi's implementation). In practice, this approach is faster than Algorithm 1 and can optimize over larger sets $\overline{C}$. However, this approach does not easily extend to CDM, NL, or Promotion and does not have a polynomial-time runtime guarantee.

5.6 Numerical experiments

We apply our methods to three datasets (Table 5.1). The SFWork dataset (Koppelman and Bhat, 2006) comes from a survey of San Francisco residents on available (choice set) and selected (choice) transportation options to get to work. We split the respondents into two segments ($|A| = 2$) according to whether or not

Table 5.2: Sum of error over all 2-item choice sets C compared to optimal (brute force) on SFWork. Algorithm 1 is optimal.

Model	**Problem**	**Greedy**	**Algorithm 1**
MNL	Agreement	0.03	0.00
	Disagreement	0.00	0.00
rank-2 CDM	Agreement	0.14	0.00
	Disagreement	0.13	0.00
NL	Agreement	0.00	0.00
	Disagreement	0.00	0.00

they live in the "core residential district of San Fransisco or Berkeley." The Allstate dataset (Kaggle, 2013b) consists of insurance policies (items) characterized by anonymous categorical features A–G with 2 to 4 values each. Each customer views a set of policies (the choice set) before purchasing one. We reduce the number of items to 24 by considering only features A, B, and C. To model different types of individuals, we split the data into homeowners and non-homeowners (again, $|A| = 2$). The Yoochoose dataset (Ben-Shimon et al., 2015) contains online shopping data of clicks and purchases of categorized items in user browsing sessions. Choice sets are unique categories browsed in a session and the choice is the category of the purchased product (categories appearing fewer than 20 times were omitted). We split the choice data into two sub-populations by thresholding on the purchase timestamps.

For inferring maximum-likelihood models from data, we use PyTorch's Adam optimizer (Kingma and Ba, 2015; Paszke et al., 2019) with learning rate 0.05, weight decay 0.00025, batch size 128, and the `amsgrad` flag (Reddi et al., 2018). We use the low-rank (rank-2) CDM (Seshadri et al., 2019) that expresses pulls as the inner product of item embeddings. Code and data used in this chapter are available at https://github.com/tomlinsonk/choice-set-opt.

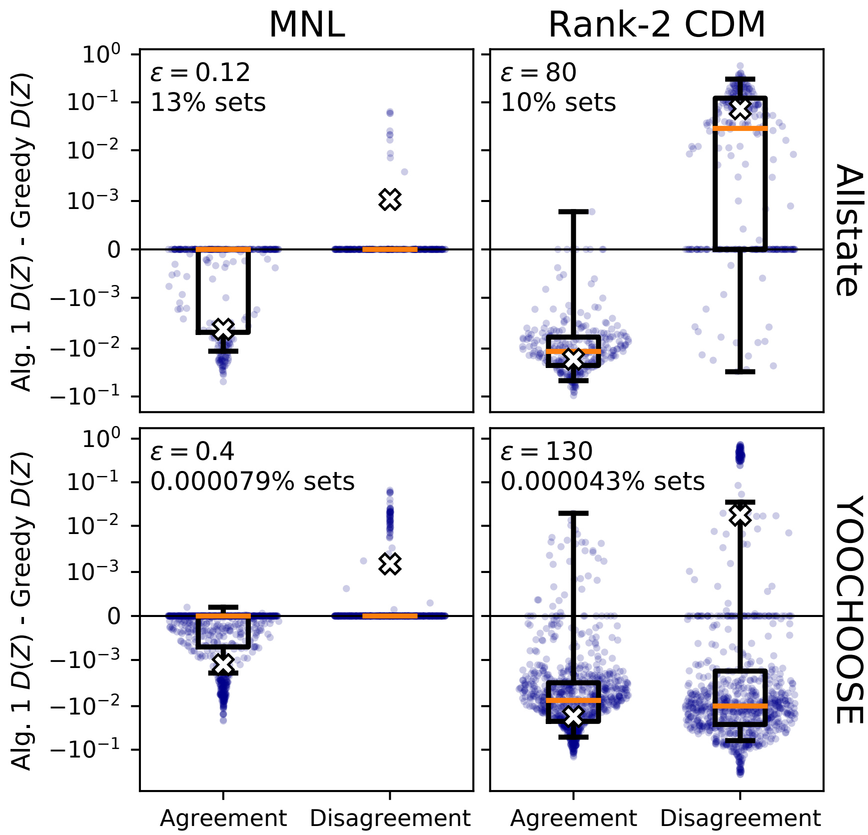


Figure 5.2: Algorithm 1 vs. Greedy performance box plots when applied to all 2-item choice sets in ALLSTATE and YOOCHOOSE under MNL and CDM (subplots also show ε and the percent of subsets of $\overline{C}$ computed by Algorithm 1, written X% sets). Each point is the difference in $D(Z)$ when Algorithm 1 and Greedy are run on a particular choice set. Horizontal spread shows approximate density and the Xs mark means. A negative (resp. positive) y-value for AGREEMENT (resp. DISAGREEMENT) indicates that Algorithm 1 outperformed Greedy. Algorithm 1 performs better in all cases except for DISAGREEMENT under CDM on YOOCHOOSE. Even in this exception, though, our approach finds a few very good solutions and Algorithm 1 has better mean performance.

For SFWORK under the MNL, CDM, and NL models, we considered all 2-item choice sets C (using all other items for $\overline{C}$) for AGREEMENT and DISAGREEMENT (for the NL model, we used the best-performing tree from Koppelman and Bhat (2006)). We compare Algorithm 1 ($\varepsilon = 0.01$) to a greedy approach (henceforth, "Greedy") that builds Z by repeatedly selecting the item from $\overline{C}$ that, when added to Z, most improves the objective, if such an item exists. This dataset was small

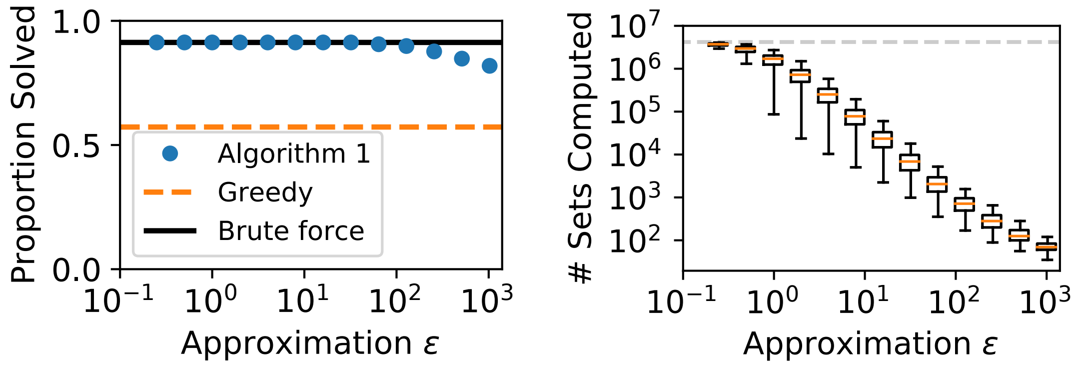


Figure 5.3: PROMOTION results on ALLSTATE 2-item choice sets. (Left) Success rate comparison; Algorithm 1 has near-optimal performance (about 9% of instances have no PROMOTION solution). (Right) Number of subsets of $\overline{C}$ computed by Algorithm 1 (dashed gray line at $2^{22} = 2^m$ for brute force computation).

enough to compare against the optimal, brute-force solution (Table 5.2). In all cases, Algorithm 1 finds the optimal solution, while Greedy is often suboptimal. However, for this value of ε, we find that Algorithm 1 searches the entire space and actually computes the brute force solution (we get the number of sets analyzed by Algorithm 1 from $|L_m|$ for a given ε and compare it to $2^{|\overline{C}|}$). Even though we have an asymptotic polynomial runtime guarantee, for small enough datasets, we might not see computational savings. Running with larger ε yielded similar results, even for $\varepsilon > 2$, when our bounds are vacuous.

The results still highlight two important points. First, even on small datasets, Greedy can be sub-optimal. For example, for AGREEMENT under CDM with $C = \{\text{drive alone}, \text{transit}\}$, Algorithm 1 found the optimal $Z = \{\text{bike, walk}\}$, inferring that both sub-populations agree on both driving less and taking transit less. However, Greedy just introduced a carpool option, which has a lower effect on discouraging driving alone or taking transit, resulting in lower agreement between city and suburban residents.

Second, our theoretical bounds can be more pessimistic than what happens in practice. Thus, we can consider larger values of ε to reduce the search space; Algorithm 1 remains a principled heuristic, and we can measure how much of the search space Algorithm 1 considers. This is the approach we take for the ALLSTATE and YOOCHOOSE data, where we find that Algorithm 1 far outperforms its theoretical worst-case bound. We again considered all 2-item choice sets C and tested our method under MNL and CDM,[4] setting ε so that the experiment took about 30 minutes to run for ALLSTATE and 2 hours for YOOCHOOSE (of that time, Greedy takes 5 seconds to run; the rest is taken up by Algorithm 1). Algorithm 1 consistently outperforms Greedy (Figure 5.2), even with $\varepsilon > 2$ for CDM. Moreover, Algorithm 1 only computes a small fraction of possible sets of alternatives, especially for YOOCHOOSE. Algorithm 1 does not perform as well with the rank-2 CDM as it does with MNL, which is to be expected as we only have approximation guarantees for CDM under particular parameter regimes (in which these data do not lie). The worse performance on CDM is due to the context effects that items from $\overline{C}$ exert on each other. Greedy does fairly well for DISAGREEMENT under CDM with YOOCHOOSE, but even in this case, Algorithm 1 performs significantly better in enough instances for the mean (but not median) performance to be better than Greedy. We repeated the experiment with 500 choice sets of size up to 5 sampled from data with similar results (Appendix D.5.3). We also ran the MIBLP approach for MNL, which performed as well as Algorithm 1 and was about 12x faster on YOOCHOOSE and 240x faster on ALLSTATE (Appendix D.5.2).

[4]In this case, we did not have available tree structures for NL, which are difficult to derive from data (Benson et al., 2016).

Promotion

We applied the CDM Promotion version of Algorithm 1 to Allstate, since this dataset is small enough to compute brute-force solutions. For each 2-item choice set C, we attempted to promote the less-popular item of the pair using brute-force, Greedy, and Algorithm 1. Algorithm 1 performed optimally up to $\varepsilon = 32$, above which it failed in only 2–26 of 252 feasible instances (Figure 5.3, left). (Here, successful promotion means that the item becomes the true favorite among C.) On the other hand, Greedy failed in 37% of the feasible instances. As in the previous experiment, our algorithm's performance in practice far exceeds the worst-case bounds. The number of sets tested by Algorithm 1 falls dramatically as ε increases (Figure 5.3, right). With more items (or a smaller range of utilities), the value of ε required to achieve the same speedup over brute force would be smaller (as with Yoochoose). In tandem, these results show that we get near-optimal Promotion performance with far fewer computations than brute force.

5.7 Discussion

Our decisions are influenced by the alternatives that are available, the choice set. In collective decision-making, altering the choice set can encourage agreement or create new conflict. We formulated this as an algorithmic question: how can we optimize the choice set for some objective?

We showed that choice set optimization is NP-hard for natural objectives under standard choice models; however, we also found that model restrictions makes promoting a choice easier than encouraging a group to agree or disagree. We developed approximation algorithms for these hard problems that are effective in practice, although there remains a gap between theoretical approximation bounds and performance on real-world data. Future work could address choice set optimization in interactive group decisions, where group members can communicate their preferences to each other or must collaborate to reach a unified decision.

CHAPTER 6

BALLOT LENGTH IN INSTANT RUNOFF VOTING

We now turn our focus from models of choice to mechanisms for aggregative preferences—i.e., voting systems. Instant runoff voting (IRV) has grown in popularity over the last two decades as an alternative to plurality voting for governmental and organizational elections. Also referred to as ranked choice voting (RCV), single transferrable vote (STV), alternative vote, preferential voting, or the Hare method, IRV allows voters to submit rankings over the candidates rather than voting for a single option. IRV determines a winner from these rankings by repeatedly eliminating the candidate who has the fewest ballots ranking them first; the ballots that listed this eliminated candidate first have their votes reallocated to the next candidate on their list. This process continues, repeatedly eliminating candidates, until only one is left—the winner.

Proponents of IRV argue that it allows voters to report their full preferences, mitigates vote-splitting when similar candidates run, encourages civility in campaigning, and saves money compared to holding separate runoff elections (FairVote, 2022; Lewyn, 2012). Many local elections in the United States use IRV, including in Minneapolis, San Fransisco, Oakland, Santa Fe, and New York City, as well as statewide elections in Maine and Alaska. IRV is also used in other countries, including Australia and Ireland.

However, IRV has vocal opponents who believe it to be too confusing for voters (Langan, 2004; Saltsman and Paxton, 2021), leading to outright bans on the use of IRV in Florida (Florida Legislature, 2022) and Tennessee (Tennessee Legislature, 2022). One particular issue critics point to is the complexity of a ballot that asks voters to rank every candidate, especially when the number of candidates is large. One official tasked with running Utah's first IRV election raised this as her primary concern after the election:

> My concerns with the current RCV law are that we would recommend the number of rankings be limited to three or five instead of an unlimited number based on the number of candidates. So although you can list as many candidates as file on the ballot, I think it is a bit confusing to voters [...] For instance, in Minneapolis they rank three. In St. Paul, they rank five. They don't usually have them rank as many candidates as there are. (Swensen, 2021, Salt Lake County Clerk)

Indeed, many municipalities have different numbers of ranking slots on their IRV ballots, what we call *ballot length*: Oakland uses three, Alaska four, and New York City five. The count goes on: ballot length six would have been mandated by the failed 2019 Ranked Choice Voting Act proposing IRV for US Congressional elections (US Congress, 2019). In Maine, voters can rank all of the candidates—even if there are 15 of them. In fact, plurality voting can be viewed as IRV with ballot length one: losing candidates are repeatedly "eliminated" (without redistribution) until the candidate with a plurality is declared the winner.

While making ballots shorter does make them simpler, it also strays from a goal of IRV: allowing voters to express their complete preferences over the candidates. Critics of IRV also raise concerns about *ballot exhaustion* during the IRV algorithm, where all candidates ranked by a voter have been eliminated and that vote no longer contributes to subsequent tallies (Burnett and Kogan, 2015).[1] Ballot length is therefore subject to competing desires: shorter ballots are easier to fill out and simpler to print, but less informative about voter preferences.

[1] In plurality, any vote not cast for the winner is "exhausted."

Despite the apparent trade-offs involved in ballot length, there has been very little investigation of how these trade-offs might work. As noted above, plurality voting can be seen as IRV with ballot length one, and so the fact that plurality and IRV can produce different outcomes already indicates that ballot length can have important consequences. But aside from early work looking at simulations and a few real-world elections (Kilgour et al., 2020; Ayadi et al., 2019) we do not have much insight into the consequences of ballot length more generally. Perhaps, for example, there are underlying structural properties to be discovered that constrain how many winners are possible as we vary the ballot length. Or perhaps "anything goes," and if we specify which candidate we'd like to see win at each possible ballot length, we can construct a fixed set of rankings that produce each desired winner at the corresponding length.

Overview of Results. In this chapter, we show that the effect of ballot length essentially behaves like the latter extreme, where almost every sequence of outcomes is possible. In particular, we prove that modulo a simple feasibility constraint, it is possible to pick any sequence of candidates (with repetitions allowed), and to have this be the sequence of winners at ballot lengths $1, 2, 3, \ldots$. For example, there are voter preferences such that one candidate wins if the election is run with odd ballot length and another wins with even ballot length. We make a central assumption that voters have fixed ideal rankings and report as long a prefix of their ideal ranking as the ballot allows. Given k candidates, we show that up to $k-1$ of them can win as the ballot length varies from $1, \ldots, k-1$ and voter preferences remain fixed. Moreover, we establish exact matching lower bounds on the number of voters required to produce $k-1$ distinct winners.

We also consider how these results are affected if we make standard modeling assumptions about voters. If we model voters abstractly as exhibiting single-peaked or single-crossing preferences, we prove that $k-1$ distinct winners across ballot lengths cannot be achieved. We also consider voters who rank candidates according to a shared one-dimensional ideological spectrum; since such voters are both single-peaked and single-crossing, there cannot be $k-1$ distinct winners in these cases. We find through simulation that in this one-dimensional case, ballot lengths above $k/2$ almost always produce the same winner as full IRV ballots.

Finally, we use data from 168 real-world elections from PrefLib (Mattei and Walsh, 2013) (most of them originally conducted using IRV), and we find that different winners across ballot lengths is a phenomenon that occurs commonly: in 25% of the PrefLib elections at least two different candidates win as the ballot length is varied by truncation. However, truly pathological cases with $k-1$ winners appear to be extremely rare: we observe at most three distinct winners across ballot lengths, and that occurs only once in the 168 PrefLib elections. But even with these real-world voter preferences, more than three winners can occur; by resampling ballots in the PrefLib elections, we observe cases with four, five, and even six different winners across ballot lengths. We note that one third of the elections initially used ballot length of at most four, where it is impossible to have more than three different winners across ballot lengths. Code and data for this chapter are available at `https://github.com/tomlinsonk/irv-ballot-length`.

6.1 Related work

There has been considerable work on what happens when individual voters choose not to rank all the candidates—a practice sometimes called *voluntary truncation*—

in contrast with *forced truncation* (i.e., ballot length restrictions) (Kilgour et al., 2020). In many voting systems including IRV, election outcomes can change dramatically as voters independently choose to rank more or fewer candidates (Saari and Van Newenhizen, 1988). This matter has been studied from a computational angle as the *possible winners* problem, which asks, given a collection of partial ballots, which candidates could become winners as those ballots are filled out (Konczak and Lang, 2005; Chevaleyre et al., 2010; Baumeister et al., 2012; Xia and Conitzer, 2011; Ayadi et al., 2019). There is also a wide array of research on how partial ballots can be used for strategic voting and campaigning (Baumeister et al., 2012; Narodytska and Walsh, 2014; Menon and Larson, 2017; Kamwa, 2022; Fishburn and Brams, 1984). On the empirical side, voluntary truncation is a concern since it can lead to ballot exhaustion (Burnett and Kogan, 2015). In political science, voluntary truncation is also referred to as *under-voting* (Neely and Cook, 2008). Several studies have asked whether different demographic groups are more likely to under-vote and how this could have a disenfranchising effect (Neely and Cook, 2008; Coll, 2021; Hoffman et al., 2021). There has also been research on "over-voting" in IRV, which refers to ranking a single candidate in more than one position (e.g., first and second), especially its correlation with underrepresented voting populations (Neely and Cook, 2008; Neely and McDaniel, 2015).

In contrast, we investigate what happens when all voter preferences are truncated as a result of ballot length. That is, we focus on a question of election design rather than on voter choice. In this direction, Ayadi et al. (2019) investigated how often IRV with short ballots produces the full-ballot winner in the Mallows model and in five PrefLib elections. However, all five PrefLib elections they studied produced the full-ballot winner at all ballot lengths—in analyzing a larger collection of 168 PrefLib elections, we find multiple winners across ballot lengths in 25% of

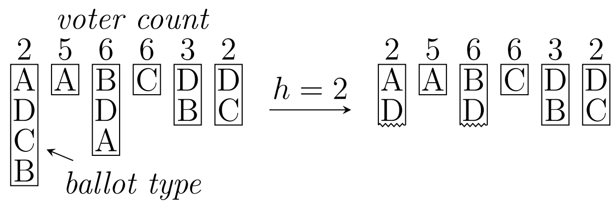


Figure 6.1: On the left, an example profile with $k = 4$ candidates A, B, C, D and $n = 24$ voters of 6 types with partial ballots. Ballots are listed top-down, with the number of voters of each type above each ballot. On the right, the profile is truncated to ballot length $h = 2$.

them. Ayadi et al. also examined several other interesting facets of IRV ballot length, including a low-communication IRV protocol (a form of online, per-voter ballot length customization) and the complexity of the possible winners problem under truncated ballots. The issue of ballot length in IRV was also touched on by Kilgour et al. (2020), who examined its effect in simulation for $k = 4, 5$, and 6 candidates, where they found up to $k-2$ distinct winners across ballot lengths. We prove that in fact $k - 1$ winners are possible *for all* $k \geq 3$. Ballot length has been considered in contexts other than IRV—for instance, research on the Boston school choice mechanism found that limiting the number of schools parents could rank to five resulted in undesirable strategic behavior (Abdulkadiroglu et al., 2006). There has also been research on ballot length in approval voting from a learning theory angle, seeking to recover a population's preferences efficiently (Garg et al., 2019).

6.2 Preliminaries

An IRV election consists of k candidates labeled $1, \ldots, k$ and n voters. Each voter j has a preference ordering over a subset of the candidates denoted by the ordered subset π_j, which we refer to as a *ballot*. At any point down the ballot, π_j can terminate, at which point the voter is indifferent over the remaining options. If

π_j includes all candidates, we call it *full*, otherwise we call it *partial*. We call a collection of ballots a *profile*. Unless otherwise specified, a profile may contain partial ballots.[2] If multiple voters have identical ballots, we say they are of the same *type*. Given a profile, IRV proceeds by eliminating the candidate with the fewest ballots ranking them first and removing them from all ballots. Ballots that have all their candidates eliminated are *exhausted*. Eliminations continue until only one candidate remains, who is declared the winner (equivalently, one can terminate when one candidate has the majority of votes from non-exhausted ballots). Ties can be broken as desired (for instance, by coin-flip), although they are unlikely in large elections.

In many real-world elections, the number of candidates a voter can rank is limited to $h < k$, which we call the *ballot length*. We assume that if the ballot length is h, voters submit the length h prefix $\pi_j(1, \ldots, h)$ of their ideal ballot π_j. Voters who would have submitted a ranking listing h or fewer candidates are unaffected. Thus, we say that ballots are *truncated* to the ballot length h. See Figure 6.1 for an example of a profile with partial ballots truncated to $h = 2$. Note that there is no difference between running IRV with ballot length k and $k - 1$, since only one candidate remains after the $(k - 1)$th elimination.

The main question we focus on is how ballot length affects an election. For instance, how many different candidates can win as the ballot length varies for a fixed profile? In order to address this question, we make some assumptions about the lack of consequential ties, since in trivial cases such as zero voters, any candidate can win depending on tie-breaks. We say that a profile is *consequential-tie-free* if tie-breaks do not affect the winner under any ballot length h. We say it is *elimination-tie-free* if a tie for last place never occurs when running IRV for any

[2] All 168 elections in the PrefLib data have partial ballots.

ballot length h. Finally, we say it is *tie-free* if no two candidates ever have tied vote counts when running IRV at any ballot length h. We note that the problem of determining if a given candidate could win under some tie-breaking sequence is known to be NP-complete (Conitzer et al., 2009).

6.3 Worst-case analysis of ballot truncation

We say a profile has c *truncation winners* if c different candidates can win depending on the ballot length. Previous simulation work found up to $k-2$ truncation winners for $k = 4, 5$, and 6 (Kilgour et al., 2020). One of our main results is that up to $k-1$ truncation winners are possible for any k. We note that it is impossible to have all k candidates win under different ballot lengths, since lengths k and $k-1$ behave the same way. All proofs omitted for readability can be found in Appendix E.3.

First, we establish an exact lower bound on the number of voters required in order to achieve $k-1$ truncation winners in consequential-tie-free profiles. Our voter lower bound is based on the observation that the winner at $h = 1$ (the plurality winner) must be eliminated second under ballot lengths ≥ 2 for $k-1$ truncation winners to occur. In order for the plurality winner to be eliminated second, the first elimination must redistribute enough votes for every other candidate to overtake the plurality winner.

Theorem 14. *For any $k > 3$, a consequential-tie-free profile requires at least $2k^2 - 2k$ voters in order to produce $k-1$ truncation winners. For $k = 3$, the lower bound is $k^2 = 9$.*

Our main theoretical result is a construction matching this lower bound, showing that $k-1$ truncation winners can occur for any $k \geq 3$. Our construction can not only produce $k-1$ truncation winners, but *any* sequence of winners over ballot lengths $1, \ldots, k-1$, provided that a candidate has not yet been eliminated.

Theorem 15. *Let there be $k > 3$ candidates, labelled $1, \ldots, k$ in their full-ballot IRV elimination order. Fix any sequence of candidates $w_1, \ldots, w_{k-1}$ such that $w_h \in \{h+1, \ldots, k\}$ for all $h \in [k-1]$. There exists a consequential-tie-free profile with $2k^2 - 2k$ partial ballots whose sequence of truncated IRV winners from $h = 1, \ldots, k-1$ is $w_1, \ldots, w_{k-1}$. For $k = 3$, such a profile exists with 9 ballots. Any sequence where $w_h \leq h$ for some $h \in [k-1]$ is impossible to realize as the sequence of truncated IRV winners for any consequential-tie-free profile.*

The idea behind the construction is to maintain a tie for second place among all candidates but two: the candidate about to be eliminated, in last, and the candidate next in the winner sequence, in first. Each elimination redistributes ballots to move the next candidates into first and last place. By carefully designing ballots, they become exhausted at just the right moment to freeze the order once we reach step h of IRV, causing the candidate currently in first to win. The example in Figure 6.1 uses this construction for $k = 4$ to achieve different winners at ballot lengths $1, 2, 3$ (namely, A, B, C). Note that the full-ballot elimination order labeling of candidates A, B, C, D is 2, 3, 4, 1, which makes the truncation winner sequence 2, 3, 4 feasible. In contrast, the sequence 2, 2, 4 would not be feasible since the candidate eliminated second under full ballots cannot win at ballot length 2. Intuitively, a winner sequence with elimination order labeling is feasible if it is element-wise at least $2, 3, \ldots, k$.

6.3.1 Restrictions on profiles

Since IRV can behave very erratically across ballot lengths for general profiles, we might hope that imposing restrictions on the space of profiles makes IRV more well-behaved. We consider three classic profile restrictions from voting theory, single-peaked (Black, 1948; Arrow, 1951), single-crossing (Gans and Smart, 1996), and 1-Euclidean preferences (see Elkind et al. (2022) for a survey of preference restrictions). A profile is *single-peaked* if there exists an order $<$ over the candidates such that, for every ballot b ranking i first, if $j < k < i$ or $i < k < j$, then j is not ranked above k in b. A profile is *single-crossing* if there exists an ordering L of the ballots such that for every ordered pair of candidates (i, j), the set of ballots ranking i above j forms an interval of L. Finally, a profile is *1-Euclidean* if there exist embeddings of the voters and candidates in $[0, 1]$ such that if voter b is closer to candidate i than to candidate j, then voter b ranks i above j.

Intuitively, single-peaked profiles arise when there is a political axis arranging candidates from left to right and voters prefer candidates closer to their ideal point on the axis (each voter can have their own ideal point). Single-crossing preferences arise when voters are arranged on an ideological axis and each candidate is most appealing to voters at a certain point on this axis. While the definitions appear similar, neither condition implies the other. 1-Euclidean profiles are both single-peaked and single-crossing—but there are profiles that are both single-peaked and single-crossing, but not 1-Euclidean (Elkind et al., 2020).

In contrast to general profiles, where $k - 1$ truncation winners can occur, we show that such cases are impossible under either single-peaked or single-crossing preferences (and therefore 1-Euclidean profiles).

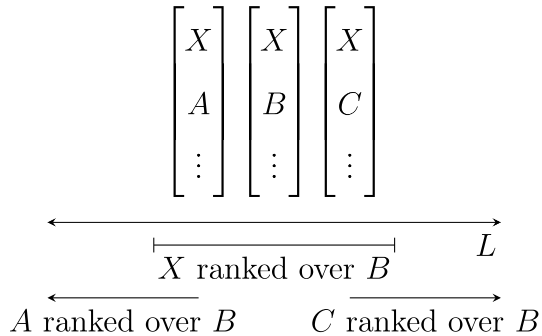
Theorem 16. *With $k \geq 5$ candidates, no consequential-tie-free single-peaked pro-*

file has $k - 1$ truncation winners.

Proof. Suppose for a contradiction that a single-peaked profile has $k-1$ truncation winners ($k \geq 5$). We know the candidate eliminated first cannot win under any ballot length. In order for the candidate eliminated second ($h \geq 2$) to win at some ballot length, it must be at $h = 1$—i.e., the plurality winner must be eliminated second under $h \geq 2$. Thus, they must be overtaken by at least three candidates (for $k \geq 5$) when the first eliminated candidate X's ballots are redistributed. But the second place on ballots listing X first can only be the candidate to the left or right of X in the single-peaked ordering, making this impossible. □

Theorem 17. *With $k \geq 5$ candidates, no consequential-tie-free single-crossing profile has $k - 1$ truncation winners.*

Proof. As in the proof of Theorem 16, we'll show that the first candidate eliminated, X, can only redistribute ballots to two candidates. Suppose for a contradiction that they redistribute ballots to at least three candidates. Call these candidates A, B, and C in the order in which they first appear as second choices in the ballots ranking X first in the single-crossing order L. By the single-crossing property, all ballots to the left of ballots starting X, A must rank A above B, since a ballot to its right ranks B above A, namely those starting X, B. Moreover, all ballots to the right of ballots starting X, C must rank C above B by symmetric reasoning. But this means B cannot have any ballots ranking them first, contradicting that X (who does have ballots ranking them first) is the first eliminated. See below for a visual depiction of this argument:



$\square$

Although the upper bound on truncation winners is strictly lower for single-peaked profiles than for general profiles, the number of achievable truncation winners still grows with k. In particular, we can show that $\Omega(\sqrt{k})$ truncation winners are possible in a consequential-tie-free single-peaked profile with $\Theta(k)$ voters.

Theorem 18. *With $k = \kappa(\kappa + 1)/2$ candidates ($\kappa \geq 3$), there is a single-peaked consequential-tie-free profile with $3\kappa(\kappa+1)/2$ partial ballots that results in κ distinct truncation winners.*

The exact upper bound on the number of truncation winners for single-peaked (and single-crossing) preferences remains an open question—it could be as large as $k - 2$. Additionally, we do not know a non-trivial lower bound on the number of achievable truncation winners for single-crossing or 1-Euclidean profiles.

6.3.2 Restrictions on ties

Since our main theorem allows ties (albeit only ties that do not affect the winners), one might be concerned that the large number of truncation winners is a byproduct of these ties. In the following results, we show that even if no vote counts are ever tied, there can still be arbitrary truncation winner sequences. We can therefore get any feasible winner sequence regardless of the tiebreaking rule. As before, we start by establishing lower bounds on the number of voters required for $k-1$ truncation winners and then provide a matching construction for tie-free profiles achieving any truncation winner sequence.

Theorem 19. *For any $k \geq 3$, an elimination-tie-free profile must contain at least $(k^3 - 3k)/2$ voters in order to produce $k-1$ truncation winners.*

Theorem 20. *For any $k \geq 3$, a tie-free profile must contain at least $(2k^3 - 5k^2 + 3k)/2$ voters in order to produce $k-1$ truncation winners.*

Note that for consequential-tie-free profiles, the lower bound on voters for $k-1$ truncation winners is $\Omega(k^2)$, but $\Omega(k^3)$ for elimination-tie-free and tie-free profiles.

Theorem 21. *Given the same setup as in Theorem 15, there exists a tie-free profile with $(2k^3 - 5k^2 + 3k)/2$ ballots whose sequence of truncated IRV winners from $h = 1, \ldots, k-1$ is $w_1, \ldots, w_{k-1}$.*

The constructions for consequential-tie-free and tie-free profiles both use $\Theta(k^2)$ *distinct* ballots. However, only $\Theta(k)$ distinct ballots are required to produce $k-1$ truncation winners. This is asymptotically tight, since each candidate who wins at some ballot length needs at least one ballot type listing them first.

Theorem 22. *Given $k > 3$ candidates, there is a tie-free profile producing $k-1$ truncation winners with $\Theta(k^3)$ voters of $\Theta(k)$ types.*

6.3.3 Full ballots

So far, all of our constructions have relied on partial ballots. For profiles with full ballots, a simple extension of our constructions using filler candidates allows us to achieve up to $k/2$ truncation winners, and in fact any feasible sequence of winners in the first half of ballot lengths.

Corollary 3. *Let $k = 2\kappa$ for some $\kappa > 3$. Label the candidates $1, \dots, 2\kappa$ in order of their elimination under full ballots. Fix any sequence $w_1, \dots w_{\kappa-1}$ such that $w_h \in \{\kappa + h + 1, \dots, 2\kappa\}$ for all $h \in [\kappa - 1]$. There exists a full-ballot consequential-tie-free profile with $2\kappa^2 - 2\kappa$ voters and a full-ballot tie-free profile with $(2\kappa^3 - 5\kappa^2 + 3\kappa)/2 + \kappa(\kappa - 1)/2$ voters whose sequences of truncation winners from $h = 1, \dots, \kappa - 1$ are $w_1, \dots, w_{\kappa-1}$.*

While we have not found a general construction with full ballots and $k-1$ truncation winners, we have found full-ballot elimination-tie-free profiles with $k - 1$ truncation winners up to $k = 10$ using a linear-programming-based search (described at the end of this section). Full ballots make intuitive constructions more challenging, but do not appear to prevent a large number of truncation winners. However, how a full ballot requirement does or doesn't change our main result remains an open question.

If instead of requiring ballots to be full, we require them to all have length at least $k/2 - c$, we can improve the above extension of our constructions and get an additional c ballot lengths at which we can specify the winner.

Corollary 4. *Let $k = 2\kappa$ for some $\kappa > 3$. Suppose we require ballots to have length at least $\kappa - c$ for $c < \kappa$. Label the candidates $1, \dots, 2\kappa$ in order of the elimination under full ballots. Fix any sequence $w_1, \dots w_{\kappa+c-1}$ such that $w_h \in$*

k	# trunc. winners	ballot types	voters	voters lower bound (Theorem 19)
4	3	7	29	26
5	4	12	55	55
6	5	23	99	99
7	6	36	161	161
8	7	57	974	244
9	8	85	1759	351
10	9	122	4855	485

Table 6.1: LP full-ballot constructions. We used different search strategies for $k \leq 7$ and $k \geq 8$, leading to profiles farther from the voter lower bound for $k \geq 8$.

$\{\kappa - c + h + 1, \ldots, 2\kappa\}$ *for all* $h \in [\kappa - 1]$. *There exists a consequential-tie-free profile with* $2\kappa^2 - 2\kappa$ *ballots and a tie-free profile with* $(2\kappa^3 - 5\kappa^2 + 3\kappa)/2 + (\kappa - c)(\kappa - c - 1)/2$ *ballots whose sequence of truncated IRV winners from* $h = 1, \ldots, \kappa + c - 1$ *is* $w_1, \ldots, w_{\kappa+c-1}$.

Given that explicit full-ballot constructions appear quite challenging, we turn to a computational approach to investigate whether full-ballot profiles can produce $k - 1$ truncation winners. Using a linear programming (LP) search, we identified elimination-tie-free profiles with full ballots and $k - 1$ truncation winners for $k = 4, 5, 6, 7, 8, 9, 10$ (the sizes of these profiles are shown in Table 6.1). Moreover, this approach was able to find instances with voter counts matching the exact lower bound in Theorem 19 for $k = 5, 6, 7$. Our approach was not able to match the voter lower bound for $k = 4$. For $k \geq 8$, we faced runtime constraints since the number of variables is exponential in the number of candidates, leading us to restrict the search space (described in further detail below). We consider elimination-tie-free profiles since they are easiest to encode as an LP, where we use constraints to enforce unambiguous eliminations.

The idea behind the search is to construct possible elimination orders across all h that could result in $k-1$ winners, express these as conditions on the sums of counts of every full ballot type in S_k (the set of permutations on k elements), and then use an LP to find a feasible real-valued solution of ballot type counts that result in that elimination order. We round these fractional ballot counts to be integers and check if the resulting profile has the desired elimination order. If not, we can try another possible elimination order or increase the gaps in the constraints so that rounding is less likely to make a solution infeasible. For $k \leq 7$, we tested all possible elimination orders, but only tested a single elimination order for $k \geq 8$ due to runtime constraints. The exact LP formulation is provided in Appendix E.2.

It remains an open question whether there exist full ballot profiles for every k that result in $k-1$ truncation winners (or any truncation winner sequence). Given the computational evidence from these LPs up to $k = 10$ candidates, we conjecture that there are such profiles.

6.3.4 Ballot length in simulation

Our theoretical results show that the winner of an IRV election can change dramatically as the ballot length varies. Here, we ask how likely these changes are through simulated profiles. Such simulation analysis was previously conducted for $k = 4, 5, 6$ (Kilgour et al., 2020). We extend these simulations up to $k = 40$ (our real-world IRV data has examples of elections with up to ≈ 30 candidates).

We simulate two different types of profiles: general profiles with rankings sampled uniformly at random and 1-Euclidean profiles with voters and candidates embedded in one dimension. For the general profiles, we fix 1000 voters. For 1-

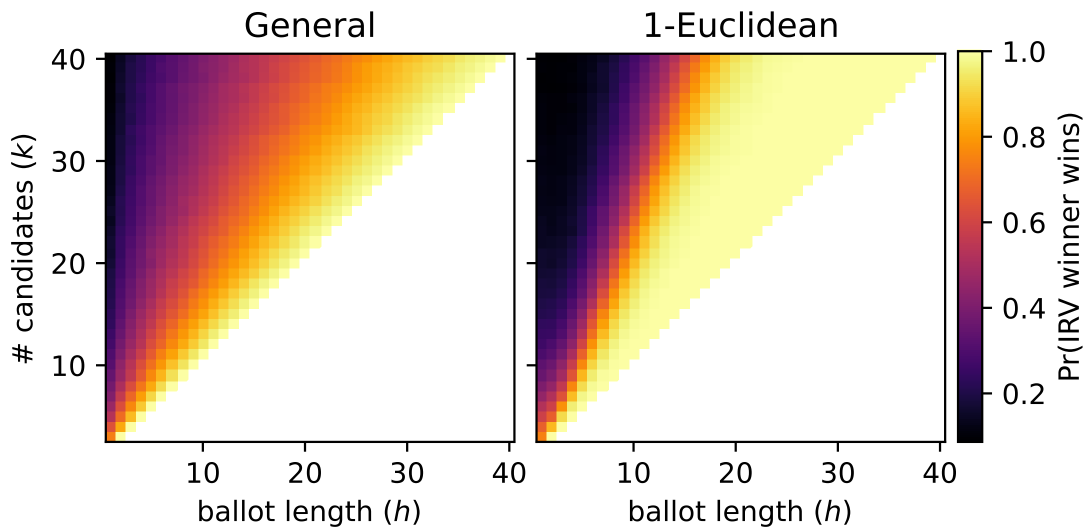


Figure 6.2: Probability that truncated ballots produce the full IRV winner for candidate counts $k = 2, \ldots, 40$ and ballot lengths $h = 1, \ldots, k-1$. (Left) For general preferences, the probability of producing the IRV winner increases smoothly with the ballot length h. (Right) For 1-Euclidean preferences, there is a sharper transition around $h = k/2$.

Euclidean profiles, we simulate an infinite voter population uniformly distributed over $[0, 1]$, where the number of first-place votes a candidate i has is the size of the interval of $[0, 1]$ containing points closer to i than any other candidate.

For both general and 1-Euclidean profiles, we simulate both full and partial ballots to gauge the effect of forced truncation with and without voluntary truncation. For general profiles with partial ballots, we independently and uniformly perform voluntary truncation on each voter's preferences before applying forced truncation in the form of ballot length. For 1-Euclidean partial ballots, we do the same with each ballot type.

In Figure 6.2, we show the probability that the full-ballot IRV winner is selected with each ballot length $1, \ldots, k-1$ for $k = 3, \ldots, 40$ with initially full ballots (the heatmaps were qualitatively identical for partial ballots; see Appendix E.4). For

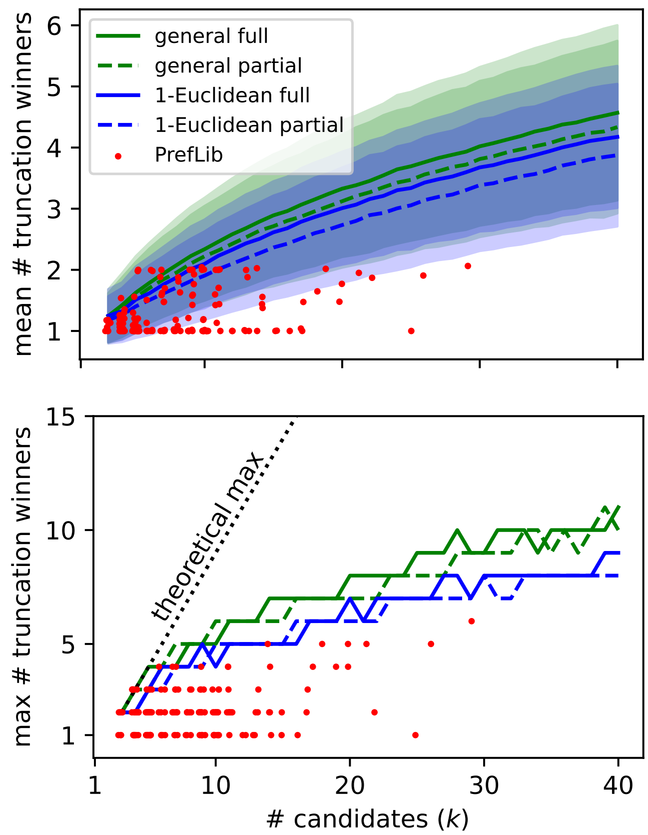


Figure 6.3: Mean (top) and maximum (bottom) number of truncation winners in 10000 synthetic ballot simulations and 10000 PrefLib resampling trials. We simulated uniform general and 1-Euclidean preferences for $k = 3, \ldots, 40$. The shaded regions show standard deviation across trials. To simulate partial ballots, each ballot is voluntarily truncated at a random length between 1 and k. While up to $k - 1$ truncation winners are possible, the mean number of truncation winners only reaches 4 around $k = 40$. 1-Euclidean profiles and profiles with partial ballots tend to produce slightly fewer truncation winners. For the PrefLib data, each point represents a single election, with horizontal jitter added for legibility. Real elections tend to produce even fewer truncation winners, although it is not rare to have more than 1.

general preferences, the probability of selecting the full-ballot IRV winner increases smoothly as ballot length increases. Additionally, for any fixed ballot length, the probability of selecting the IRV winner decreases as the number of candidates increases. For instance, for $h = 3$, the probability of selecting the IRV winner first dips below 50% at $k = 12$. For 1-Euclidean preferences, small ballot lengths are even less likely to produce the full IRV winner: for $h = 3$, the probability first drops below 50% for $k = 9$. On the other hand, the probability rapidly increases around $h = k/2$. For ballots longer than $k/2$, uniform 1-Euclidean preferences almost always produce the full IRV winner.

In Figure 6.3, we visualize the same data in a different way. We plot the mean and maximum observed numbers of truncation winners across ballot lengths (the figure also includes PrefLib winner counts described in the next section). While the difference between general and 1-Euclidean profiles was pronounced in the previous heatmaps, they result in almost the same number of truncation winners on average. Additionally, these simulated profiles tend to have a small number of truncation winners relative to the theoretical maximum. On average for $k \leq 10$, there are around two truncation winners, while the theoretical maximum is nine. Additionally, the maximum observed number of winners in 10000 simulated trials was well below the theoretical maximum, especially for larger k: we only began generating any profiles with 10 truncation winners around $k = 40$.

Intuitively, these simulation results therefore indicate that profiles with large numbers of truncation winners are very rare in the space of profiles, at least under these (uniform) measures. However, they do not appear to be significantly rarer among 1-Euclidean profiles than among general profiles, as one might have expected given the increased structure of 1-Euclidean profiles. On the other hand, profiles in which there are more than one winner across ballot lengths are very common. Thus, while truly extreme cases with $k - 1$ truncation winners might be rare, cases where ballot length has an effect occur readily in simulation.

6.4 Truncating real-world election data

Given that many truncation winners are theoretically possible, we now ask how often multiple truncation winners occur in real-world election data. To this end, we analyze voter rankings from 168 elections in PrefLib (Mattei and Walsh, 2013). This collection includes 12 American Psychological Association (APA) presidential elections (Regenwetter et al., 2007) ($h = 5$), 14 San Francisco local elections ($h = 3$), and 21 Glasgow local elections ($h = k$), among others. The number of candidates in these elections ranges from 3–29 and the number of voters from tens to hundreds of thousands. See Appendix E.5 for additional summary statistics of these elections. Some of these PrefLib datasets included a small number of ballots with multiple candidates listed at the same rank (0.5% of all ballots), which we omit.

In order to evaluate the impact of ballot length, we truncate the rankings at each possible shorter ballot length than the election actually used. We then run IRV on the truncated ballots. We assume that if ballots had been shorter, voters would have reported the same ranking, but truncated to the ballot length. It is possible that voters would express their preferences differently depending on the ballot length, so our approach should be seen as an approximation to this counterfactual scenario.

In 41/168 elections, there were two different winners across ballot lengths, and in one election, there were three different winners. Overall, 25% of the elections were sensitive to ballot length. Among the elections with ballot length $h \leq 5$, $12/85 = 14\%$ of them had two different truncation winners; for elections with $h > 5$, $29/83 = 35\%$ of elections had two or more different winners. In order to better understand the landscape of possible outcomes in each election, we also performed resampling of ballots. Given a collection of n ballots, we resample a collection of n ballots with replacement to simulate another possible election outcome with the same pool of voters. We then truncate those collections of votes to assess the impact of ballot length. In 10000 resampling trials, we observed up to six different truncation winners across the elections, but the expected number of truncation winners under resampling was between one and two for all elections (see Figure 6.3). In Figure 6.4, we also use ballot resampling to visualize the sequence of truncation winners in two PrefLib elections. The 2009 Burlington Mayoral election famously had a different plurality winner (Kurt Wright) than the elected IRV winner (Bob Kiss), but our visualization reveals that at ballot length $h = 2$, the election was a complete toss-up and could have gone either way with only a small change in ballot counts. In the right subplot, we visualize the sequence of truncation winners in the one PrefLib election that had three distinct truncation winners. Not only does this election have three truncation winners, but the sequence of winners flips back and forth, as we proved theoretically possible.

The smaller number of truncation winners in real data is likely due to the small number of front-runners in real-world elections, in contrast with the uniform preferences in our synthetic data. Our observations here are in line with the finding that ballot truncation is less likely to change the winner in the Mallows model when preferences are more tightly clustered around the central ranking (Ayadi et al., 2019).

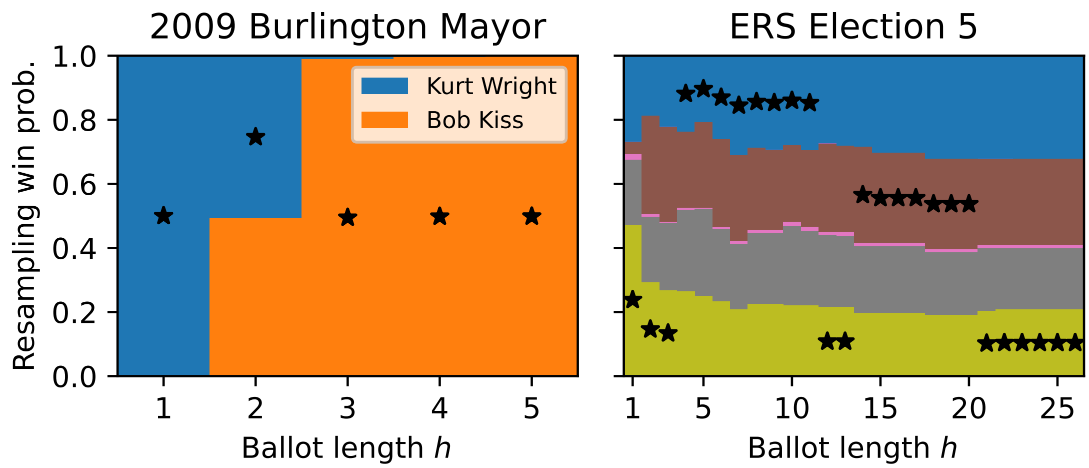


Figure 6.4: Two elections in the PrefLib data, the infamous 2009 Burlington mayoral election (left, $k = 6$, $n = 8974$) and an anonymous intra-organization election from the Electoral Reform Society (right, $k = 26$, $n = 104$). The stacked bars show the probability candidates have of winning at each ballot length under ballot resampling. Stars indicate the winners at each ballot length with actual ballot counts.

6.5 Discussion

Our theoretical results are fairly pessimistic: IRV election outcomes can change dramatically with ballot length. Our analysis of real and simulated data, on the other hand, presents a more mixed picture: ballot length regularly has an effect on the identity of the winner even in real elections, but the extreme changes between winners that are theoretically possible rarely occur, which may be cause for some degree of optimism. Nonetheless, changes in ballot length by truncation can often result in two or three different winners, even when the ballot length is short.

There are a number of open theoretical questions around ballot length. First, is it possible to achieve every feasible truncation winner sequence with complete ballots? We suspect the answer is yes, but an explicit construction has proved elusive. Second, are more than $O(\sqrt{k})$ truncation winners possible for single-peaked ballots? How many truncation winners are possible with single-crossing ballots? Similar questions could be asked for other profile restrictions, such as 1-Euclidean preferences.

Our interest in IRV is due to its increasing popularity of IRV in United States local elections, but one could also investigate the effects of ballot length in other ranking-based voting systems such as Borda count or Copeland's method. Additionally, we do not address what ballot length should be used in practice, which requires making a tradeoff between competing desires. Finally, it would be interesting to understand when elections are close enough for ballot length to affect the winner. There has been research on calculating the margin of victory for IRV (Sarwate et al., 2013; Blom et al., 2016; Magrino et al., 2011), defined as the number of votes which would need to be altered to change the winner, which is NP-hard to compute (Xia, 2012). A notion of margin of victory that relates to winners across different ballot lengths would be valuable.

CHAPTER 7

THE MODERATING EFFECT OF INSTANT RUNOFF VOTING

As discussed in the previous chapter, IRV is among the most popular alternatives to plurality for single-winner elections—but there is heated debate about whether it should be adopted more widely in the United States. Proponents of IRV claim that it encourages moderation, compromise, and civility, since candidates are incentivized to be ranked highly by as many voters as possible, including by those who do not rank them first (Dean, 2016; Diamond, 2016). Analyses of campaign communication materials and voter surveys have supported the theory that IRV increases campaign civility (Donovan et al., 2016; John and Douglas, 2017; Kropf, 2021), with extensive debate about whether this greater civility translates into winners who are also more moderate in their positions (Fraenkel and Grofman, 2006a,b; Horowitz, 2006, 2007). Analyses of potential moderating effects of IRV have primarily been based on case studies (Fraenkel and Grofman, 2004; Mitchell, 2014; Reilly, 2018) and simulation (Chamberlin and Cohen, 1978; Merrill, 1984; McGann et al., 2002), as well as empirical evidence for a moderating effect in a related voting system, two-round runoff (Bordignon et al., 2016).

In contrast, there has been almost no theoretical work on the subject; most social choice theory has focused on problems other than moderation, such as minimizing metric distortion and ensuring fairness or representation (Halpern et al., 2023; Aziz et al., 2017; Boutilier et al., 2012; Brill et al., 2022; Ebadian et al., 2022; Gkatzelis et al., 2020; Kahng et al., 2023). Two interesting specific exceptions can be found in the works of Grofman and Feld (2004) and Dellis et al. (2017). Grofman and Feld (2004) show that for single-peaked preferences and four or fewer candidates, IRV is at least as likely as plurality to elect the median candidate.

Dellis et al. (2017) show that in a citizen-candidate model, if the voter distribution is asymmetric then two-party equilibria under plurality can be more extreme than under IRV.

There is clear value in mathematical analyses that identify more general moderating tendencies. At present—beyond the noted exceptions—the arguments for IRV's moderating effects summarized above have tended to point to institutional or behavioral properties of the way candidates run their campaigns in IRV elections. A natural question, therefore, is whether this picture is complete, or whether there might be something in the definition of IRV itself that leads to outcomes with more moderate winners. Such questions are fundamental to the mathematical theory of voting more generally, where we frequently seek explanations that are rooted in the formal properties of the voting systems themselves, rather than simply the empirical regularities of how candidates and voters tend to behave in these systems. In the case of IRV, what would it mean to formalize a tendency toward moderation in the underlying structure of the voting system? To begin, we must first identify a natural set of definitions under which we can isolate such a property.

Formalizing the moderating effect of IRV. In this chapter, we propose such definitions and use them to articulate a precise sense in which IRV produces moderate winners in a way that plurality does not. We work within a standard one-dimensional model of voters and candidates: the positions of voters and candidates correspond to points drawn from distributions on the unit interval $[0, 1]$ of the real line (representing left–right ideology), and voters form preferences over candidates by ranking them in order of proximity. That is, voters favor candidates who are closer to them on the line; this is often called the *1-Euclidean* model, a common model in social choice theory (Coombs, 1964; Bogomolnaia and Laslier, 2007;

Elkind et al., 2022). We typically assume the voters and candidates are drawn from the same distribution F, but some of our results hold for fixed candidate positions. In addition to its role as one of the classical mathematical models of voter preferences, where it is sometimes called the Hotelling model (Hotelling, 1929; Downs, 1957), 1-Euclidean preferences arise naturally from higher-dimensional opinions under simple models of opinion updating (DeMarzo et al., 2003). There is wide-ranging empirical evidence suggesting that political opinions in the United States are remarkably one-dimensional (Poole and Rosenthal, 1984, 1991; Layman et al., 2006; DellaPosta et al., 2015): from a voter's views on any one of a set of issues including tax policy, immigration, climate change, gun control, and abortion, it is possible to predict the others with striking levels of confidence.

Let's consider a voting system applied to a set of k candidates and a continuum of voters in this setting: we draw k candidates independently from a given distribution F on the unit interval $[0, 1]$, and each candidate gets a vote share corresponding to the fraction of voters who are closest to them (see Figure 7.1 for examples). The use of a one-dimensional model gives a natural interpretation to the distinction between moderate and extreme candidates: a candidate is more extreme if they are closer to the endpoints of the unit interval $[0, 1]$. We take two approaches to defining a moderating effect in this model, one probabilistic (in the limit of large k) and one combinatorial (for all k). We say that a voting system has a *probabilistic moderating effect* if for some interval $I = [a, b]$ with $0 < a \leq b < 1$, the probability that the winning candidate comes from I converges to 1 as the number of candidates k goes to infinity (since we focus on symmetric voter distributions, we will typically have I symmetric about $1/2$; i.e. $b = 1 - a$). We say that a voting system has a *combinatorial moderating effect* if for all k, the presence of a candidate in I prevents any candidate outside of I from winning; i.e., a moderate candidate (inside I) is guaranteed to win as long as at least one moderate runs. (Note that a combinatorial moderating effect implies a probabilistic one, as long as the candidate distribution F places positive probability mass on I.) We call such an interval I an *exclusion zone* of the voting system, since the presence of a candidate inside this zone precludes outside candidates from winning. In this way, a voting system with a moderating effect will tend to suppress extreme candidates who lie outside a middle portion of the unit interval, while a voting system that does not have at moderating effect will allow arbitrarily extreme candidates to win with positive probability even as the number of candidates becomes large.

Using this terminology, we can state our first main result succinctly: under a uniform voter distribution, IRV has a moderating effect and plurality does not—in both the combinatorial and probabilistic senses. In particular, we prove a novel and striking fact about IRV: when voters and candidates both come from the uniform distribution on $[0, 1]$, the probability that the winning candidate produced by IRV lies outside the interval $[1/6, 5/6]$ goes to 0 as the number of candidates k goes to infinity. In sharp contrast, the distribution of the plurality winner's position converges to uniform as the number of candidates goes to infinity, allowing arbitrarily extreme candidates to win. As part of our analysis, we provide a method for deriving the distribution of plurality and IRV winner positions for finite k and perform this derivation for $k = 3$ candidates. Surprisingly, our analysis of plurality—the simpler voting system—requires much more sophisticated machinery: we establish a connection between plurality voting and a classic model in discrete probability known as the stick-breaking process and develop new asymptotic stick-breaking results for use in our analysis.

Our probabilistic result for IRV follows from a companion fact that is combi-

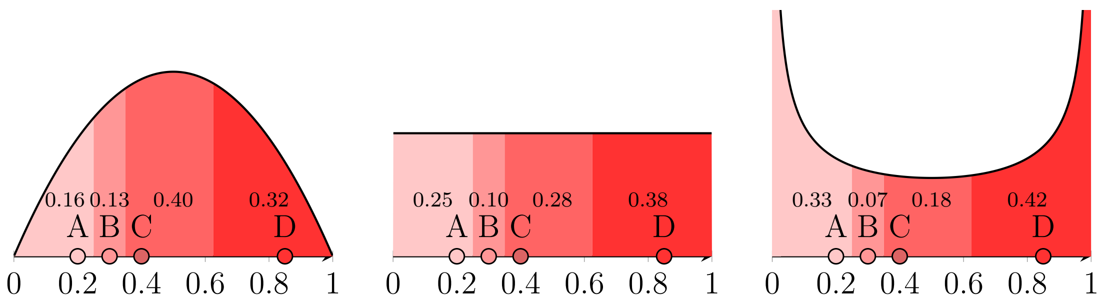


Figure 7.1: Three example voter distributions in one dimension (all Betas). Candidates A, B, C, D are positioned at 0.2, 0.3, 0.4, and 0.85. The black line shows the density function of the voter distribution. Regions are colored according to the most preferred candidate of voters in that region and annotated with the approximate vote share of that candidate. As an example, the preference ordering of a voter at 0.5 is C, B, A, D (regardless of the voter distribution). Similarly, a voter at 0.1 has preference ordering A, B, C, D. In the moderate voters example (left), C is both the plurality and IRV winner. In the uniform voters example (center), D is the plurality winner and C is the IRV winner. In the polarized voters example (right), D is the plurality winner and A is the IRV winner.

natorial in nature and comparably succinct: given any finite set of candidates in $[0, 1]$, and voters from the uniform distribution, if any of the candidates belong to the interval $[1/6, 5/6]$, then the IRV winner must come from $[1/6, 5/6]$; that is, $[1/6, 5/6]$ is an exclusion zone for IRV in the uniform case. Moreover, $[1/6, 5/6]$ is the smallest interval for which this statement is true. Again, the analogue for plurality voting with any proper sub-interval of the unit interval is false: we show that plurality has no exclusion zones.

This first main result therefore gives a precise sense in which the structure of the IRV voting system favors moderate candidates: whenever moderate candidates (in the middle two-thirds of the unit interval) are present as options, IRV will push out more extreme candidates. We then address the more challenging case of non-uniform voter distributions, where we prove that IRV continues to have a moderating effect (in the sense of our formal definitions) even for voter distributions that push probability mass out toward the extremes of the unit interval, up to a specific threshold beyond which the effects cease to hold. Thus, IRV is even able to offset a level of polarization built into the underlying distribution of voters and candidates, although it can only do so up until a certain level of polarization is reached. In contrast, we establish that plurality never has a combinatorial moderating effect for any non-pathological voter distribution.

As a final point, it is worth emphasizing what is and is not a focus of our work here. We examine IRV and plurality because of their widespread use in real-world elections and the fierce debate surrounding the adoption of IRV over plurality. We are not trying to characterize all possible voting systems that give rise to moderation (although we can show that many voting systems not in widespread use have a moderating effect, including the Coombs rule and any Condorcet method; for these systems, any symmetric interval around 0.5 is an exclusion zone). Our interest, instead, is in the following contribution to the plurality–IRV debate: there is a precise mathematical sense in which IRV has a moderating effect and plurality does not. Second, we do not analyze strategic choices by candidates about where to position themselves on the unit interval (Hotelling, 1929; Downs, 1957; Osborne, 1995), but instead derive properties of voting systems that hold for fixed candidate positions, or candidate positions drawn from a distribution. This approach produces results that are robust against the question of whether candidates are actually able to make optimal strategic positioning decisions in practice (Bendor et al., 2011); it also allows us to better understand how the voting systems themselves behave—providing a foundation for future strategic work.

7.1 Uniform voters

The previous section describes our complete model, but it is useful to review it here in the context of some more specific notation. We assume voters and candidates are both drawn from a distribution F on the unit interval $[0, 1]$, representing their ideological position on a left–right spectrum.[1] Voters prefer candidates closer to them (i.e., they have 1-Euclidean preferences). There are k candidates drawn independently from F; suppose that these draws produce candidate positions $x_1 < x_2 < \cdots < x_k$ in order. Some of our results apply regardless of the candidate distribution, relying only on the voter distribution; we will make a note of such cases.

Since we want to model the case of a large population of voters, we do not explicitly sample the voters from F, but instead think of a continuum of voters who correspond to the distribution F itself: that is, under the plurality voting rule, the fraction of voters who vote for candidate x_i is the probability mass of all voters who are closer to x_i than to any other candidate (or, equivalently, it is the probability that a voter randomly chosen according to F would be closer to x_i than any other candidate). In this section, we focus on the case where F is uniform.[2] We use $v(x_i)$ to denote the vote share for candidate x_i. Under IRV, the candidate i with the smallest $v(x_i)$ is eliminated and vote shares are recomputed without candidate i. This repeats until only one candidate remains, who is declared the winner (equivalently, elimination can terminate when a candidate achieves majority). In

[1]We will generally focus on distributions F that are symmetric around $1/2$ and represented by a density function f.

[2]To provide another perspective on the uniform voter assumption, consider the following preference assumption that also produces uniform 1-Euclidean preferences: voters are arbitrarily distributed, but rank candidates according to how many voters are between them and each candidate. That is, voters have 1-Euclidean preferences *in the voter quantile space* and are always uniformly distributed over this space by definition. All of our uniform voter results hold in that setting as well, although stated in terms of voter quantiles rather than absolute positions.

practice, voters submit a ranking over the candidates and their votes are "instantly" redistributed after each elimination.

IRV's moderating effect: A first result. With uniform 1-Euclidean voters, we now show that IRV cannot elect extreme candidates over moderates—regardless of the distribution of candidates. That is, IRV exhibits an exclusion zone in the middle of the unit interval, where the presence of moderate candidates inside the zone precludes outside extreme candidates from winning. The idea behind the proof is that as moderates get eliminated, the middle part of the interval becomes sparser, granting a higher vote share to any remaining moderates. Consider the moment when only one candidate x remains in the interval $[1/6, 5/6]$ (see Figure 7.2); extreme candidates near 0 and 1 are then too far away to "squeeze out" x. With uniform voters, the tipping point for squeezing out moderates occurs when extreme candidates are at $1/6$ and $5/6$. In the next section, we present generalizations of this result for non-uniform voter distributions.

Theorem 23. *(Combinatorial moderation for uniform IRV.) Under IRV with uniform voters over* $[0, 1]$ *and* $k \geq 3$ *candidates, if there is a candidate in* $[1/6, 5/6]$*, then the IRV winner is in* $[1/6, 5/6]$*. No smaller interval* $[c, 1 - c]$*,* $c > 1/6$*, has this property. If there are no candidates in* $[1/6, 5/6]$*, then the IRV winner is the one closest to* $1/2$*.*

Proof. Suppose first there is only one candidate $x \in [1/6, 5/6]$ and all other candidates are $< 1/6$ or $> 5/6$. Suppose without loss of generality that $x \leq 1/2$. The smallest vote share x could have occurs when there are candidates at $1/6 - \epsilon$ and $5/6 + \epsilon$. In this case, x gets vote share $(x - 1/6 + \epsilon)/2 + (5/6 + \epsilon - x)/2 = 1/3 + \epsilon$. Meanwhile, the highest vote share any candidate $< 1/6$ could have (when x is at

$1/2$) is $1/6 - \epsilon + (1/2 - 1/6 + \epsilon)/2 = 1/3 - \epsilon/2$. Thus, every candidate $< 1/6$ will be eliminated before x. At this point, x will win, since it is closer to $1/2$ than any of remaining candidates $> 5/6$ and therefore has a majority.

If there is more than one candidate in $[1/6, 5/6]$ to begin with, then as candidates are eliminated, at some point there will only be one candidate x remaining in $[1/6, 5/6]$. Either x will be the ultimate winner, or there will still be candidates $< 1/6$ or $> 5/6$, in which case x will win as argued above.

Notice that the above argument still holds if we replace $1/6$ and $5/6$ with c and $1-c$ for any $0 < c \leq 1/6$: it only reduces the vote share going towards candidates in $[0, c)$. Thus, if there is some candidate in $[c, 1-c]$, then the IRV winner is in $[c, 1-c]$. So, if there is no candidate in $[1/6/, 5/6]$, then let c be the distance between the most moderate candidate and its closest edge. This candidate must be the IRV winner, since it is the only candidate in $[c, 1-c]$ (and $c < 1/6$). In this case, the IRV winner is the most moderate candidate as claimed.

Finally, we show that no smaller interval satisfies the theorem. To do so, we describe a construction that is parametrized to handle any number of candidates $k \geq 3$. In the construction, there is one candidate at $1/2$, two candidates at c and $1-c$ for $c > 1/6$, and any remaining candidates in $(1-\epsilon, 1]$ for $\epsilon < (1/2 - c)/2$. First, all candidates right of $1-\epsilon$ will be eliminated—all of these candidates have a smaller vote share than the candidate at c. At this point, the candidate at $1/2$ has vote share less than $2(1/2 - c)/2 = 1/2 - c$. Since $c > 1/6$, this is less than $1/2 - 1/6 = 1/3$. Meanwhile, the candidates at c and $1-c$ have vote shares higher than $c + (1/2 - c)/2 = 1/4 + c/2$. Since $c > 1/6$, this is greater than $1/4 + 1/12 = 1/3$. Thus the middle candidate is eliminated. The winner is is thus outside of $[c + \delta, 1 - c - \delta]$ for all $\delta \in (0, 1/2 - c)$, despite there being a candidate

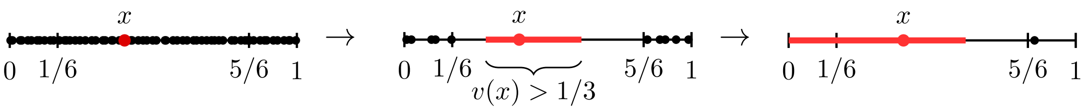


Figure 7.2: Visual depiction of the proof of Theorem 23. IRV eliminates candidates until a final candidate x remains in the exclusion zone $[1/6, 5/6]$. At this point, x gets more than $1/3$ of the vote share and cannot be eliminated next (regardless of where they are in $[1/6, 5/6]$). Candidates outside of $[1/6, 5/6]$ are thus eliminated until x wins.

in this interval. Since this construction applies for any $c > 1/6$, no interval smaller than $[1/6, 5/6]$ satisfies the theorem. □

In the language of our analysis, $[1/6, 5/6]$ is then the smallest possible exclusion zone of IRV under a uniform voter distribution. See Figure 7.2 for a visual depiction of the argument. A corollary of Theorem 23 is that if candidates are distributed uniformly at random (for instance, if voters independently and identically decide whether to run for office), then IRV elects extreme candidates with probability going to 0 as the number of candidates grows, since the probability of having no moderate candidates in $[1/6, 5/6]$ is $(1/3)^k$. In the language defined earlier, IRV thus has a probabilistic moderating effect with uniform voters and candidates.

Corollary 5. *(Probabilistic moderation for uniform IRV.) Let R_k be the position of the IRV winner with k candidates distributed uniformly at random and uniform voters.*

$$\lim_{k \to \infty} \Pr(R_k \notin [1/6, 5/6]) = 0. \tag{7.1}$$

In contrast to IRV, where the presence of candidates with moderate positions (namely, inside $[1/6, 5/6]$) precludes extreme candidates from winning, we now show that no such fact is true for plurality (excluding the extreme points 0 and 1): for any interval $I \subseteq (0, 1)$, there is some configuration of candidates such that the winner is outside of I despite having candidates in I. In other words, plurality voting does not have a combinatorial moderating effect with uniform voters.[3] Later, we generalize this result to non-uniform voter distributions. The idea behind the proof is relatively straightforward: given a set of candidates, keep adding candidates to reduce the vote share of everyone except the desired winner.

Theorem 24. *(No combinatorial moderation for uniform plurality.) Suppose voters are uniformly distributed over* $[0,1]$*. Given any set of* $\kappa \geq 1$ *distinct candidate positions* $x_1, \ldots, x_\kappa$ *with* $x_1 \notin \{0,1\}$*, there exists a configuration of* $k \geq \kappa$ *candidates (including* $x_1, \ldots, x_\kappa$*) such that the candidate at* x_1 *wins under plurality.*

Proof. We show how to add candidates to the initial set $x_1, \ldots, x_\kappa$ so that x_1 becomes the plurality winner (as long as $x_1 \notin \{0,1\}$). First, add candidates at $x_0 = 0$ and $x_{\kappa+1} = 1$ to guarantee that x_1 is between two candidates. Let x_ℓ be the candidate to the left of x_1 and let x_r be the candidate to the right of x_1. Let $v_\ell = (x_1 - x_\ell)/2$ be the vote share x_1 gets on its left and let $v_r = (x_r - x_1)/2$ be the vote share x_1 gets on its right. Add new candidates spaced by $\frac{1}{2}\min\{v_\ell, v_r\}$ in the intervals $[0, x_\ell]$ and $[x_r, 1]$. This causes every candidate in the intervals $[0, x_\ell)$ and $(x_r, 1]$ to have vote share strictly less than $\frac{1}{2}\min\{v_\ell, v_r\}$ (whether they are part of the original κ or new). Additionally, x_ℓ and x_r have vote share at most $\frac{1}{2}\min\{v_\ell, v_r\} + \max\{v_\ell, v_r\}$. Meanwhile, x_1 has vote share $v_\ell + v_r$, so x_1 is the plurality winner in the new configuration. $\square$

In addition, we prove that the asymptotic distribution of the plurality winner's position is uniform over the unit interval when voters and candidates are positioned uniformly at random. In other words, plurality does not have a probabilistic mod-

[3] An anonymous reviewer suggested an elegant construction proving this fact for symmetric intervals $I = [c, 1-c]$, which provides counterexamples for every $k \geq 3$: place candidates at $c - \epsilon, 1 - c - \epsilon$, and any others at $1 - c + \epsilon$ (for $\epsilon < c/2$). The candidate at $c - \epsilon$ wins, despite having a candidate in I.

erating effect: it does not preclude extreme candidates from winning when there are many moderate candidates to choose from. The proof is more involved, so we relegate it to Appendix F.2. Note that this result implies plurality also has no combinatorial moderation, but Theorem 24 is much easier to prove.

Theorem 25. *(No probabilistic moderation for uniform plurality.) Let P_k be the position of the plurality winner with k candidates distributed uniformly at random and uniform voters. As $k \to \infty$, P_k converges in distribution to Uniform$(0, 1)$; that is, $\lim_{k\to\infty} \Pr(P_k \leq x) = x$ for all $x \in [0, 1]$.*

The proof uses a coupling argument between plurality on the unit interval and plurality on a circle. By rotational symmetry, the plurality winner on a circle is uniformly distributed. We show that as k grows, cutting the circle to transform it into the interval does not change the winner with probability approaching 1, since cutting the circle only affects vote shares of the boundary candidates.

Thus, a key step is deriving the asymptotic distribution of the winning plurality vote share. This vote share distribution may be useful for other asymptotic analyses of plurality voting, so we describe it here. The winning plurality vote share is closely related to a category of probabilistic problems known as *stick-breaking problems*, which focus on the properties of a stick of length 1 broken into n pieces uniformly at random (Holst, 1980). Setting $n = k + 1$, these stick pieces can be viewed as the gaps between candidates (equivalently, candidates are the breakpoints of the stick). A classic result in stick-breaking is that the biggest piece will have size B_n almost exactly $\log n/n$ as n grows large (Darling, 1953; Holst, 1980) and that $nB_n - \log n$ converges to a Gumbel$(1, 0)$ distribution as $n \to \infty$. The plurality vote setting is different, since candidates get vote shares from half of the gap to their left plus half of the gap to their right (except the left- and rightmost candidates). We show that as the number of candidates grows large, the winning vote share V_k with $k = n - 1$ candidates is almost exactly $(\log n + \log\log n)/2n$ and that $nV_k - (\log n + \log\log n)/2$ also converges to Gumbel$(1, 0)$ as $k \to \infty$. Intuitively, the largest pair of adjacent gaps have size $\log n/n$ and $\log\log n/n$, and the candidate between these gaps gets vote shares from half of each gap (more correctly, the total size of the gaps is $(\log n + \log\log n)/n$). This is formalized in the following lemma used to prove Theorem 25.

Lemma 3. *Let V_k be the winning plurality vote share with k candidates distributed uniformly at random over $[0, 1]$ and uniform voters. Setting $n = k + 1$,*

$$\lim_{k\to\infty} \Pr\left(V_k \leq \frac{\log n + \log\log n + x}{2n}\right) = e^{-e^{-x}}. \tag{7.2}$$

7.1.1 Plurality and IRV winner distributions

Given these results about the asymptotic distributions of the plurality and IRV winner positions P_k and R_k, asymptotic in the number of candidates k, a natural follow-on question is whether we can say anything about these distributions for fixed values of k.

For a fixed value of k, the distributions of the plurality and IRV winner positions P_k and R_k with uniform voters and candidates have density functions f_{P_k} and f_{R_k} that are piecewise polynomial of order $k - 1$. To see this, consider a point in the k-dimensional unit hypercube, where dimension i of this point represents the position of candidate i. For every left-right order of candidates $\pi \in S_k$ (where $\pi(i)$ is the index of candidate i in left-right order and S_k is the symmetric group on k elements), we can express the region in $\mathbb{R}^k$ where candidate i wins given order π using the following collection of linear inequalities:

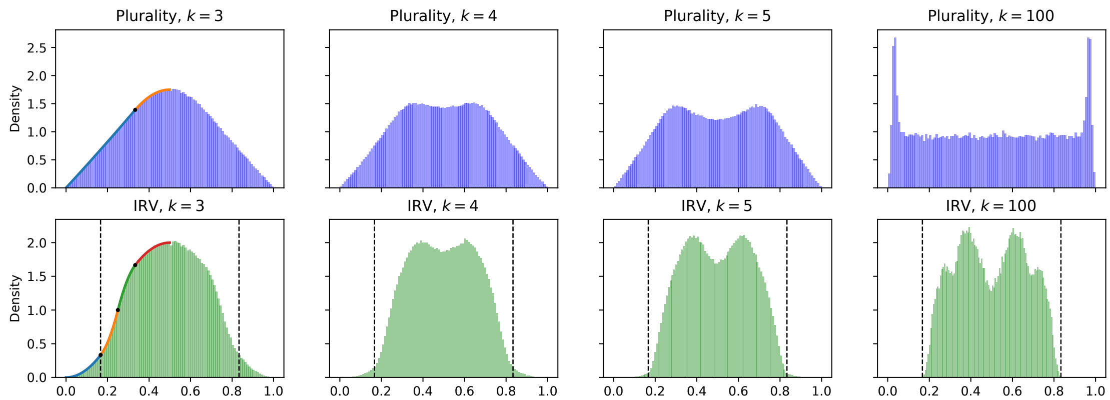


Figure 7.3: The distributions of the winning position with $k = 3, 4, 5$, and 100 candidates and continuous 1-Euclidean voters (both uniformly distributed) under plurality and IRV. The histograms are from 1 million simulation trials for $k = 3, 4, 5$ and 100000 trials for $k = 100$, while the curves plotted for $k = 3$ (shown up to $1/2$) are the exact density functions given in Propositions 3 and 4, with pieces separated by color. Note that the IRV winner is only at a position $< 1/6$ or $> 5/6$ when no candidates fall in $[1/6, 5/6]$ by Theorem 23; the dashed vertical lines outline this exclusion zone. The probabilistic moderating effect for IRV is already strong at with only $k = 5$ candidates.

$$0 \leq x_{\pi^{-1}(1)} < x_{\pi^{-1}(2)} < \cdots < x_{\pi^{-1}(k)} \leq 1,$$

$$v(x_i) = \frac{x_{r(i)} - x_{\ell(i)}}{2} > \frac{x_{r(j)} - x_{\ell(j)}}{2} = v(x_j), \qquad \text{(for all } j \neq i)$$

where $\ell(i) = \pi^{-1}(\pi(i) - 1)$ is the candidate to i's left and $r(i) = \pi^{-1}(\pi(i) + 1)$ is the candidate to i's right. The inequalities in the first line ensure the left–right candidate order matches π, while the inequalities on the second line ensure x_i has a larger vote share than any other candidate (i.e., x_i is the plurality winner). The region defined by these linear inequalities is therefore a convex polytope, as it is the intersection of a finite number of half spaces.

To find the probability that a candidate i at a particular point x wins under plurality, we can find the sum of the cross-sectional areas of these polytopes at $x_i = x$ (with one polytope for each of the $k!$ candidate orderings), integrating over the positions of the other $k-1$ candidates. This procedure produces a piecewise polynomial in x of order $k-1$, where pieces are split at the vertices of the polytopes. To convert the win probability of candidate i at position x into the winner position density at x, we scale by k to account for the symmetry in choosing i.

We can use this approach to derive f_{P_3}, the winner distribution for plurality with 3 candidates (see Figure 7.3 for a visualization). As the derivation is tedious, we present it in Appendix F.3. Additionally, Appendix F.4 includes a visualization of the winning position polyhedra for $k=3$ whose cross-sectional areas produce f_{P_3}.

Propostion 3.

$$f_{P_3}(x) = \begin{cases} x^2/2 + 4x, & x \in [0, 1/3] \\ -13x^2 + 13x - 3/2, & x \in [1/3, 1/2] \\ f_{P_3}(1-x), & x \in (1/2, 1]. \end{cases} \tag{7.3}$$

For analyzing the plurality winner distribution in this way with larger k (even after accounting for relabeling symmetry), we would need to integrate over k k-polytopes, each of which has a number of faces growing linearly with k (one face per inequality requiring that x_i beats each other x_j). Unfortunately, the number of vertices per polytope in this procedure could grow exponentially with k, potentially requiring exponentially many integrals.

The same strategy can also be used for IRV, except we no longer have only one polytope per permutation of candidates—instead, we have one polytope per combination of left-right candidate order and candidate elimination order. If we fix both, the region where candidate i wins under IRV can once again be defined by a collection of linear inequalities. We used this approach to derive the IRV winner distribution with 3 candidates, f_{R_3}. Again, see Figure 7.3 for a visualization of f_{R_3}, Appendix F.3 for the derivation, and Appendix F.4 for a visualization of the IRV polyhedra.

Propostion 4.

$$f_{R_3}(x) = \begin{cases} 12x^2, & x \in [0, 1/6] \\ 48x^2 - 12x + 1, & x \in [1/6, 1/4] \\ -48x^2 + 36x - 5, & x \in [1/4, 1/3] \\ -12x^2 + 12x - 1, & x \in [1/3, 1/2] \\ f_{R_3}(1-x), & x \in (1/2, 1]. \end{cases} \tag{7.4}$$

For IRV with general k, this analysis requires integrating over $k!$ k-polytopes, each of which has $O(k^2)$ faces: given an elimination order, we need an inequality specifying that the candidate eliminated i^{th} has a smaller vote share than each of the candidates eliminated later. Each such inequality defining a half-space can add a face to the polytope.

Note that in Proposition 4, the integral of the density $f_{R_3}(x)$ on $[0, 1/6]$ is exactly equal to half the probability that the $k-1$ losing candidates did not appear inside $[x, 1-x]$ (scaled by k to account for relabeling symmetry), since we know by Theorem 23 that a candidate can only win outside $[1/6, 5/6]$ if they are the most moderate candidate. For general $k > 3$ we can easily derive the density on $[0, 1/6]$ and $[5/6, 1]$ using the generalization of this argument: $f_{R_k}(x) = k(2x)^{k-1}$ on $[0, 1/6]$ (with the right tail being mirrored). Note that the integral of $f_{R_k}(x) = k(2x)^{k-1}$ over $[0, 1/6]$ goes to 0 as $k \to \infty$, a limit that furnishes an independent way of establishing a probabilistic moderating effect for IRV.

Having the exact winner position distributions f_{P_3} and f_{R_3} allows us to answer additional questions—for instance, how much more moderate do IRV winners tend to be for $k = 3$ with uniform voters and candidates? Using the density functions above, we can analytically compute the variances of the plurality and IRV winner distributions, $\mathrm{Var}(P_3) = 23/540$ and $\mathrm{Var}(R_3) = 25/864$. For $k = 3$, the variance of the plurality winner's position with uniform voters is thus exactly $184/125 = 1.472$ times higher than the variance of the IRV winner's position.

Connecting our results to related work, while the distribution of the winner's position is challenging to derive, the expected plurality vote share at each point is more tractable. This distribution was discovered in another context: a guessing game where the goal is to be closest to an unknown target distributed uniformly at random, against k players who guess uniformly at random (Drinen et al., 2009). The target can be thought of as a random voter and the guesses as candidate positions. The guessing game and plurality winner position distributions are similar in shape, with two prominent bumps that move outward as k grows; and both converge to uniform distributions. However, the point with the max *expected plurality vote share* (and max guessing game win probability) is not quite the same as the point with the maximum *plurality win probability*, since a candidate's position influences other candidates' vote shares.

7.2 Non-uniform voters

Given our understanding of the uniform voter case, we now broaden our scope and show that IRV exhibits exclusion zones more generally. We find that the same "squeezing" argument can be applied to any symmetric voter distribution. The generalized result hinges on a specific condition on the cumulative distribution function, Equation (7.5), which intuitively captures when, no matter where the last moderate candidate is, they cannot be squeezed out by the most moderate extremists. This condition is not always possible to satisfy non-trivially. After first giving the general statement, we present special cases where the condition is simple to state and satisfy—specifically, when the voter density is monotonic over $[0, 1/2]$. If the voter distribution is sufficiently highly polarized, the condition becomes impossible to satisfy. In this *hyper-polarized* regime, the exclusion zone of IRV actually flips, and IRV cannot elect moderate candidates over extreme ones. First, we present the general moderating effect of IRV for symmetric voter distributions.

Theorem 26. *(General combinatorial moderation for IRV.) Let f be symmetric over $[0, 1]$ with cdf F and let $c \in (0, 1/2)$. If for all $x \in [c, 1/2]$,*

$$F\left(\frac{x+1-c}{2}\right) - F\left(\frac{c+x}{2}\right) > 1/3, \tag{7.5}$$

then if there is at least one candidate in $[c, 1-c]$, the IRV winner must be in $[c, 1-c]$.

Proof. Suppose there is at least one candidate in $[0, c)$, at least one candidate in $(1-c, 1]$, and exactly one candidate x in $[c, 1-c]$ (if there no candidates in the left or right extremes, then x immediately wins by majority). Assume without loss of generality that $x \leq 1/2$. Candidate x's vote share is minimized when there are candidates at $c - \epsilon$ and $1 - c + \epsilon$. The vote share of x is then

$$v(x) = F\left(\frac{x+1-c+\epsilon}{2}\right) - F\left(\frac{c-\epsilon+x}{2}\right).$$

If Condition (7.5) is satisfied, then then we can increase the left hand side of (7.5) to find

$$v(x) = F\left(\frac{x+1-c+\epsilon}{2}\right) - F\left(\frac{c-\epsilon+x}{2}\right) > 1/3.$$

Thus x cannot be eliminated next, since there is a candidate with a smaller vote share than x. The IRV winner must therefore be in $[c, 1-c]$ by the same argument as in Theorem 23. $\square$

We now consider two cases where Condition (7.5) can be greatly simplified: when the voter distribution is moderate (f increases over $[0, 1/2]$; Theorem 27) and when voters are polarized (f decreases over $[0, 1/2]$ but $F(1/4) < 1/3$; Theorem 28). The proofs in these cases follow the same structure, but differ in where moderate candidates are easiest to squeeze out (nearer or farther from $1/2$). Proofs can be found in Appendix F.2. As another note, just as with Corollary 5, we immediately see from Theorem 26 (and the special cases below) that IRV has a probabilistic moderating effect with symmetric voter and candidate distributions (as long as they place positive mass on $[c, 1-c]$): as the number of candidates goes to infinity, the probability that the winner comes from $[c, 1-c]$ goes to 1.

Theorem 27. *(Moderate voter distribution.) Let f be symmetric over $[0, 1]$ and non-decreasing over $[0, 1/2]$. For any $c \leq F^{-1}(1/6)$, if there is a candidate in $[c, 1-c]$, then the IRV winner is in $[c, 1-c]$.*

Theorem 28. *(Polarized voter distribution.) Let f be symmetric over $[0, 1]$, non-increasing over $[0, 1/2]$, and let $F(1/4) < 1/3$. For any $c \leq 2(F^{-1}(1/3) - 1/4)$, if there is a candidate in $[c, 1-c]$, then the IRV winner is in $[c, 1-c]$.*

The uniform distribution is the unique distribution whose density function is both non-increasing and non-decreasing over $[0, 1/2]$. Indeed, for uniform $F(x) = x$, $1/6 = 2(F^{-1}(1/3) - 1/4) = F^{-1}(1/6)$. Note that for polarized voter distributions, Theorem 28 requires $F(1/4) < 1/3$ (i.e., less than $1/3$ of voters are left of $1/4$). If the population is hyper-polarized and instead $F(1/4) > 1/3$, we can prove that IRV cannot elect moderates if both extremes are represented.

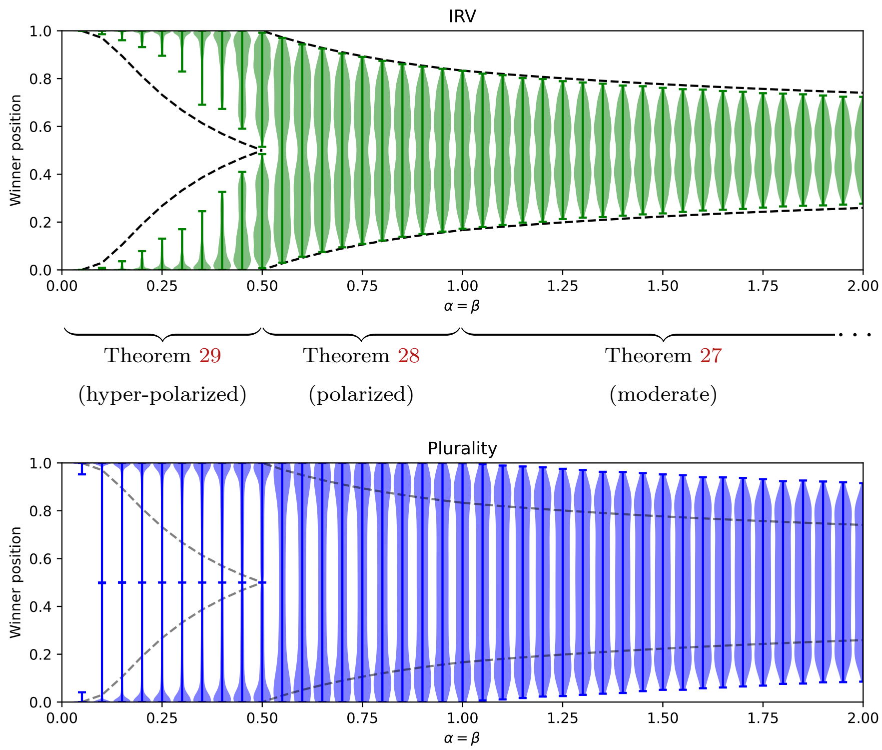


Figure 7.4: IRV (top) and plurality (bottom) winner positions with Beta(α, α)-distributed voters and candidates. The violin plots show empirical distributions from 100,000 simulation trials with $k = 30$ candidates at each α value, with whiskers marking extrema. The dashed lines show the bounds from Theorems 27 to 29 in the annotated ranges. As long as voters are not too polarized, IRV prevents extreme candidates from winning. Plurality, on the other hand, allows arbitrarily extreme candidates to win for $\alpha = 1$, when the voter distribution is uniform.

Theorem 29. *(Hyper-polarized voter distribution.) Let f be symmetric over $[0, 1]$ and let $F(1/4) > 1/3$. For any $c \geq 2F^{-1}(1/3)$, if there is at least one candidate in $[0, c]$ and at least one candidate in $[1 - c, 1]$, then the IRV winner must be in $[0, c]$ or $[1 - c, 1]$.*

We saw in Theorem 24 that plurality has no exclusion zones for uniform voters. We now show that plurality has no exclusion zones regardless of the voter distribution (given mild continuity and positivity conditions), except the points 0 and 1. The proof can be found in Appendix F.2.

Theorem 30. *(No combinatorial moderation for plurality.) Let f be continuous and strictly positive over $(0, 1)$. Given any set of $\kappa \geq 1$ distinct candidate positions $x_1, \ldots, x_\kappa$ with $x_1 \notin \{0, 1\}$, there exists a configuration of $k \geq \kappa$ candidates (including $x_1, \ldots, x_\kappa$) such that the candidate at x_1 wins under plurality. If $x_1 \in \{0, 1\}$, then there exist voter distributions where x_1 cannot win under plurality.*

Figure 7.4 provides illustrations to accompany Theorems 27 to 30, showing empirical IRV and plurality winner positions when voters (and $k = 20$ candidates) are distributed according to symmetric Beta(α, α) distributions. This family of Beta distributions is polarized for $\alpha < 1$, uniform for $\alpha = 1$, and moderate for $\alpha > 1$. Theorem 27 thus applies for $\alpha \geq 1$. The crossover point between Theorem 28 and Theorem 29 occurs at $\alpha = 1/2$ (i.e., for Beta$(1/2,\ 1/2)$, $F^{-1}(1/3) = 1/4$). Figure 7.4 also visualizes the positions of plurality winners for these voter distributions, consistent with our analysis of plurality in Theorem 30.

Finally, we revisit the existing moderating effect result of Grofman and Feld with single-peaked voters and strengthen it in the symmetric 1-Euclidean case. Recall that 1-Euclidean preferences are always single-peaked, but most sets of single-peaked preferences are not 1-Euclidean. That is, we make a stronger assumption on voter preferences and thus derive a stronger result. Grofman and Feld (2004) proved that when voters have single-peaked preferences over $k \leq 4$ candidates, if plurality elects the median candidate, so does IRV. The *median candidate* here is defined as the candidate most preferred by the median voter (with single-peaked preferences, this is the Condorcet winner (Black, 1948)). With symmetric 1-Euclidean voters, the median candidate is the candidate closest to $1/2$ (i.e., the most moderate candidate). Thus, applying the result of Grofman and Feld directly to the symmetric 1-Euclidean voter setting, we know for $k \leq 4$ that whenever plurality elects the most moderate candidate, IRV does too. In the symmetric 1-Euclidean setting, we can strengthen this theorem to consider what happens when plurality does not elect the most moderate candidate. Note that this result holds for any symmetric voter distribution.

Theorem 31. *For $k \leq 4$ with symmetric 1-Euclidean voters, the IRV winner cannot be more extreme than the plurality winner (if no ties occur). For $k \geq 5$, the IRV winner can be more extreme than the plurality winner.*

Proof. The $k = 1$ and $k = 2$ cases are trivial, since IRV and plurality are identical when $k < 3$.

For $k = 3$, suppose for a contradiction that the plurality winner P is more moderate than the IRV winner I (call the third candidate E). Under IRV, the first candidate eliminated can't be I (since they win under IRV) and can't be P (since they have the highest first-place vote share), so it must be E. In the second round of IRV, we are then left with a two-candidate plurality election between I and P. Since voters are symmetrically distributed, the more moderate of I and P thus wins under IRV, which is P. Contradiction!

For $k = 4$, suppose again for a contradiction that the IRV winner I is more extreme than the plurality winner P. As before, neither can be the first eliminated. Call the first candidate eliminated E and the fourth candidate F. Since P is more moderate than I, the final IRV round cannot be between P and I, or else P would win, contradicting that I is the IRV winner. Thus, the final round must be between I and F. P must then be the second eliminated after E. However, P has a higher vote share than both I and F in the first round. To be eliminated second, the elimination of E must cause I and F to overtake P. To redistribute votes to both I and F, E must be directly between them, with P off to one side of the I, E, F

group. Consider two cases: (1) P is adjacent to I. Since P is more moderate than I, it must get all of the vote share on the side of opposite the I, E, F group (either [1, 0.5] or [0.5, 1]), which means it has a majority—contradicting that I is the IRV winner. (2) P is adjacent to F. But then F is more moderate than I, so I cannot win in the final round—contradicting that it is the IRV winner.

For $k \geq 5$, place candidates at $\epsilon, 1/5, 1/2, 4/5$, and 1 for small ϵ (for instance $\epsilon \leq 0.01$ works; additional candidates can be packed into $[0, \epsilon]$). Note that the candidate at 1/2 is the plurality winner, with vote share 3/10. The candidates in $[0, \epsilon]$ are eliminated first under IRV, followed by the candidates at 1 and ϵ. At this point, the candidates at 1/5 and 4/5 have a higher vote share than the candidate at 1/2, who is eliminated. The IRV winner is then either at 1/5 or 4/5. □

See Figure F.1 in Appendix F.4 for simulation results demonstrating Theorem 31. All simulation code and results from this chapter are available at https://github.com/tomlinsonk/irv-moderation.

7.3 Discussion

We began by considering a contrast between IRV and plurality voting when the positions of voters and candidates are drawn from the uniform distribution on the unit interval: in this case, IRV (unlike plurality) has a moderating effect, with the probability that the winner comes from the interval $[1/6, 5/6]$ converging to 1 as the number of candidates goes to infinity. This moderating effect continues to hold (with proper sub-intervals different from $[1/6, 5/6]$) even as the distribution of voters and candidates becomes more polarized, with an increasing amount of probability mass near the endpoints of the interval, until a specific threshold of hyper-polarization is reached. Our analysis also provides methods for determining the exact distribution of winner positions in certain cases, making more fine-grained comparisons between IRV and plurality possible.

It would be interesting to consider extensions of our work in a number of directions, and here we highlight three of these. First, we did not consider strategic analyses (e.g., of Nash equilibria, as in Dellis et al. (2017)), and were instead motivated by bounded rationality (Bendor et al., 2011) and a need to better understand the underlying voting system, focusing on the non-strategic setting where candidate positions are fixed. For instance, how might candidates behave strategically given an understanding of IRV exclusion zones or the winner position distribution of IRV? Behavioral evidence for bounded rationality indicates that people tend to operate at a low strategic depth (Stahl and Wilson, 1995; Colman, 2003; Ohtsubo and Rapoport, 2006). In this framework, level-0 players act randomly, level-1 players calculate best responses to level-0 players, and so on. Our analysis therefore corresponds to level-0 strategic reasoning, and can be used as a starting point for analysis of higher-order strategy.

Second, we modeled voting populations as symmetric continuous distributions in one dimension, with preferences arising strictly from distances in this dimension. Considering higher-dimensional preference spaces would also be a natural extension of our analysis. Does IRV exhibit exclusion zones in two, three, or more dimensions? Asymmetric voter distributions would also be valuable to consider, although the notion of a *moderate* may need to be revisited in this case (perhaps based on the median voter). Using the same squeezing argument, IRV should also exhibit exclusion zones with asymmetric voter distributions, although their forms may not be as tidy as the ones we derive. Other possible extensions include non-linear voter preferences (for instance, where a voter ranks all candidates on their right before all candidates on their left, regardless of distance), probabilistic voting, and voter abstention. Practical considerations of IRV could also be taken into account; for instance, real-world elections often ask for top-truncated preferences rather than full rankings, which can the affect the outcome (Tomlinson et al., 2023). Does IRV with truncated ballots still exhibit a moderating effect?

Finally, as we noted earlier, there are voting systems that always select the most moderate candidate with symmetric 1-Euclidean voters. This is true for any system that satisfies the Condorcet criterion, selecting the Condorcet winner whenever one exists (a property that holds for the minimax, Condorcet-Hare, Copeland, and Dodgson methods, among many others (Black, 1958; Richelson, 1975; Green-Armytage et al., 2016)); it is also true for some other voting systems that do not in general satisfy the Condorcet criterion, like the Coombs rule (Coombs, 1964; Grofman and Feld, 2004). There are a variety of practical and historical reasons why these methods are not widely used for political elections. For instance, Dodgson's method is NP-hard to compute (Bartholdi et al., 1989) and the Coombs rule is very sensitive to incomplete ballots, which are common in practice. As we are motivated by ongoing debates about IRV and plurality, our attention has been restricted to these voting methods. However, a broader understanding of moderating effects of voting systems would be valuable. There has been some theoretical work on moderating effects of score-based voting systems (like Borda count and approval voting) with strategic voters and candidates (Dellis, 2009). However, it is an open question (with some computational evidence to support it (Chamberlin and Cohen, 1978)) whether other voting systems like Borda count exert a moderating effect in the setting we study, with fixed voter and candidate distributions.

CHAPTER 8

REPLICATING ELECTORAL SUCCESS

In the previous chapter, we assumed that candidate positions are drawn from a distribution or are arbitrary fixed points; this was a simplifying choice, as our focus was on the mechanism of the voting system itself. However, the real world is clearly more complex: candidates are motivated to gain office and therefore are likely to strategize about their policy position in order to maximize their chance of election. In this chapter, we explore which policies are electorally successful over time when candidates are (heuristically) strategic, a core topic in the study of elections.

The literature in this area traces its roots to Hotelling (1929) and Downs (1957). In the Hotelling–Downs model, candidates compete for election in a one-dimensional policy space. Under the assumption that voters prefer candidates closer to them in policy space, two rational office-seeking candidates will adopt the policy of the median voter, since any other position receives strictly fewer votes. Thus, the central prediction of the Hotelling–Downs model is that we should expect candidates to espouse near-identical moderate policies; in economic contexts, this is often called the *principle of minimum differentiation* (Eaton and Lipsey, 1975; De Palma et al., 1985). However, this is not what we observe in modern democracies: countries using plurality often have two dominant parties with markedly different policies (Poole and Rosenthal, 1984; Grofman, 2004; Riker, 1982). Decades of research have attempted to reconcile this observed policy divergence with the intuitive arguments of Hotelling and Downs (Grofman, 2004; Osborne, 1995), postulating additional factors like the threat of third-party entry (Palfrey, 1984) or policy- rather than office-motivated candidates (Wittman, 1983). Subsequent research has also expanded beyond two-candidate analysis to consider k-candidate elections (Cox, 1987).

The majority of this work has continued under the traditional assumption that candidates are rational and able to make strategically optimal decisions. However, the growing literature on bounded rationality (Simon, 1955, 1979) and decision-making heuristics (Tversky and Kahneman, 1974), as well as the complexity of elections, casts doubt on whether this is likely in practice. In a notable exception to the literature on rational candidate positioning, Bendor et al. (2011) argue that heuristics play a crucial role in electoral strategy:

> Campaigns are of chess like complexity—worse, probably; instead of a fixed board, campaigns are fought out on stages that can change over time, and new players can enter the game. Hence, cognitive constraints (e.g., the inability to look far down the decision tree, to anticipate your opponent's response to your response to their response to your new ad) inevitably matter. [...] Thus, political campaigns, like military ones, are filled with trial and error. A theme is tried, goes badly (or seems to), and is dropped. The staff hurries to find a new one, which seems to work initially and then weakens. A third is tried, and then a fourth. [...] *In short, there are good reasons for believing that the basic properties of experiential learning—becoming more likely to use something that has worked in the past and less likely to repeat something that has failed—hold in presidential campaigns.* (Bendor et al., 2011, emphasis ours)

Our model. In this chapter, we introduce a model of candidate positioning based on the above heuristic: candidates imitate success. We focus on plurality elections,

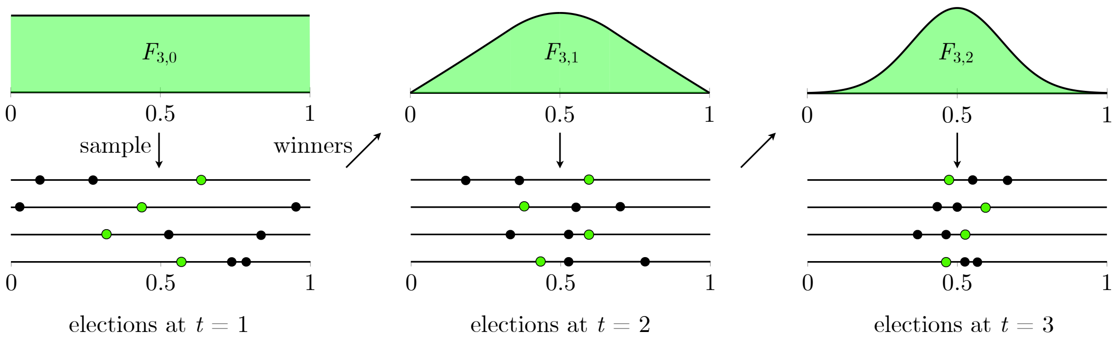


Figure 8.1: Replicator dynamics for candidate positioning with $k = 3$ candidates per election. The top row shows the winner distributions $F_{k,t}$ for each generation t, starting from a uniform distribution at $t = 0$, while the bottom row shows four example elections per generation. In each generation, candidates sample their positions from the winner distribution from the previous generation. Plurality winners (with voters uniform over $[0, 1]$) are indicated in green.

where each voter casts one vote and the candidate with the most votes wins. We assume voters have 1-Euclidean preferences (Coombs, 1950; Elkind et al., 2016), where voters and candidates occupy points in the unit interval $[0, 1]$ and voters prefer closer candidates. To represent a large voting population, our model uses a continuum of voters and continuous vote shares rather than discrete counts. Diverging from prior work, we model a large number of k-candidate elections that proceed in generations rather than an individual election or election sequence. In each generation, we assume that candidates copy the policy position of a winner from the previous generation, a simple heuristic in line with Bendor et al.'s suggestion that candidates use strategies that worked in the past. This heuristic is also supported by a wealth of political science research arguing that the imitation of policies, especially electorally successful ones, is a major feature of politics (Shipan and Volden, 2008; Böhmelt et al., 2016; Ezrow et al., 2021). As with voters, our model uses a continuous distribution of candidate positions in each generation, which can be viewed as either capturing the expected behavior of a finite number of elections or as the infinite-election limit.

This simple assumption about candidate behavior (sample a position from the distribution of winners in the last election cycle) yields a mathematical model equivalent to the well-studied *replicator dynamics* from evolutionary biology (Taylor and Jonker, 1978; Schuster and Sigmund, 1983), which have also found widespread use in economics (Safarzyńska and van den Bergh, 2010; Nelson et al., 2018). In the classic replicator dynamics, n strategies (or alleles) compete in a population, increasing in prevalence at a rate proportional to their average fitness in pairwise contests drawn from the current population. Our model arises from taking such dynamics and moving to a continuous strategy space with k-way interactions in discrete time (i.e., k-candidate elections), treating the plurality win probability as fitness; we therefore refer to it as *replicator dynamics for candidate positioning.*

In summary, then, our model operates in a sequence of generations; each generation involves a large number of identically distributed elections, and the candidates in a given generation are drawn from the distribution of winners of the previous generation's elections. Figure 8.1 provides a schematic visualization of the process with $k = 3$ candidates. While our model is phrased in terms of a large population of elections—just as the classic replicator dynamics models a population of organisms—there is a deep connection between replicator dynamics and reinforcement learning (Börgers and Sarin, 1997; Bloembergen et al., 2015), so our conclusions are likely to generalize to models of individual-level trial-and-error.

Our results. Our main technical contributions characterize the long-run behavior of the replicator dynamics for different values of k, the number of candidates per election. We find a dramatic qualitative change in the dynamics as the number of candidates k increases. For our analysis, we focus on the case in which the initial

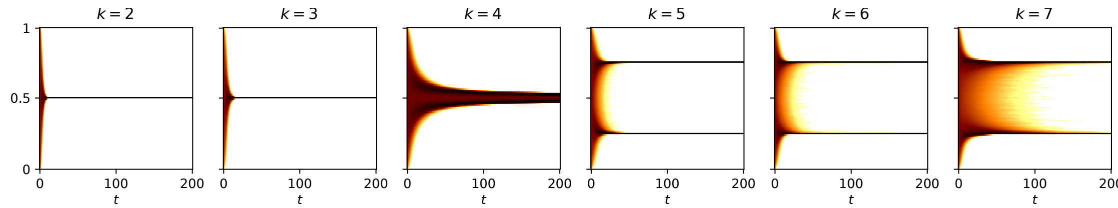


Figure 8.2: Replicator dynamics runs for $k = 2, \ldots, 7$ and 200 generations. Each plot shows 50 runs layered on top of each other, where each run simulates 100,000 elections per generation. We also use *enhanced symmetry*, a trick to keep the symmetry of the analytical model by reflecting copied points across 1/2 (discussed further in Section 8.5). Darker regions indicate higher candidate density; we use a log-scaled colormap to make low-density regions visible. As our theory establishes, the candidate distribution converges to the center for $k = 2, 3, 4$, but does not for $k \geq 5$. The convergence is very fast for $k = 2$ and 3, but much slower for $k = 4$.

distribution of candidates is symmetric and has a continuous CDF and that voters are uniformly distributed over $[0, 1]$, but we find evidence in simulation that the same patterns hold with other symmetric voter distributions. When $k = 2$, we prove that the candidate distribution converges to a point mass at 1/2 under the replicator dynamics, just like rational candidates in the Hotelling–Downs model. However, we also prove in our model that the candidate distribution converges to the center for $k = 3$ and 4, in stark contrast to three- and four-candidate extensions of the Hotelling–Downs argument (Cox, 1987). Given the behavior for $k = 2, 3$, and 4, one might be tempted to hypothesize that the replicator dynamics always cause the candidate distribution to converge to the center. Surprisingly, we prove that the pattern ends there: for any $k \geq 5$, we show that the candidate distribution does not converge to 1/2. See Figure 8.2 for simulations demonstrating the patterns that we characterize theoretically. These simulations reveal a tendency for candidate counts larger than 4 to result in two distinct clusters of policies (around 1/4 and 3/4 with uniform voters). This is strongly reminiscent of *Duverger's Law* (Duverger, 1959; Riker, 1982), the observation that plurality elections tend towards two-party systems; it is striking that it emerges here from a model that does not include any explicit reward for clustering at points away from the center or any mechanisms like the threat of third-party candidates (Bol et al., 2016).

To strengthen this characterization of the long-run replicator dynamics, we show that our convergence results are robust to noise: even if a small fraction of candidates position themselves uniformly at random, we can still show (approximate) convergence to the center for $k = 2, 3, 4$ and non-convergence for $k \geq 5$. While we are not able to theoretically derive the asymptotic distribution for $k \geq 5$ in general, we show that when the initial candidate distribution is supported only on $(1/4, 3/4)$, the candidate density in an interval around $1/2$ goes to 0. Additionally, we explore several variants of the model in simulation, including non-uniform voter distributions, noisy position-copying, memory of prior rounds of elections, and mixtures of candidate counts. Across these variants, we observe the same general pattern: convergence to the center with up to four candidates, but not with five or more. For candidate counts $k > 5$ we sometimes see complex and chaotic finite-sample effects in simulation. We conclude by relating our replicator dynamics model back to traditional analyses of Nash equilibria in the style of Hotelling and Downs. The close relationship between replicator dynamics fixed points and Nash equilibria is well-known (Hofbauer and Sigmund, 2003), but we argue that ignoring dynamics and focusing only on Nash equilibria leads to brittle conclusions. In particular, we show that different assumptions on voter behavior when candidates occupy the same points lead to dramatically different Nash equilibria than reported in prior work (Cox, 1987); in contrast, this choice has no effect on our replicator dynamics results.

To summarize, our main finding is that a simple imitation heuristic can cause candidates to either converge to the median voter or to form two distinct parties, depending on how many candidates run in each election. Intuitively, this phenomenon is driven by the bogeyman of one-dimensional plurality elections: being flanked. If a candidate is stuck between two others, they lose votes from both the left and the right. When there are too many candidates all imitating previous moderate winners, only the leftmost or rightmost of them will avoid being flanked, making more extreme candidates more successful. However, with a small enough pool of opponents, the higher vote share a moderate can receive is worth the risk of ending up stuck between two others. This emerges naturally from our dynamics, without the need for strategic forethought. The surprising fact that falls out of our mathematical analysis is that when candidates are imitators rather than optimizers, the tipping point between the Hotelling–Downs centripetal force and the centrifugal force fueled by the problem of flanking occurs between four- and five-candidate elections.

8.1 Related work

Before diving into our theoretical analysis, we briefly summarize the literature in relevant areas.

One-shot candidate positioning games. Expanding on the two-candidate Hotelling–Downs foundation, subsequent research has explored higher-dimensional spaces (Plott, 1967; Irmen and Thisse, 1998), more than two candidates (Cox, 1987), policy motivation (Wittman, 1983), uncertainty about voter positions (Calvert, 1985), and candidate valence (i.e., charisma or name recognition) (Groseclose, 2001; Bruter et al., 2010), among many other variations (see

Osborne (1995); Kurella (2017) for surveys).[1] Some models allow a third-party candidate to enter the race after the established candidates select their positions, which can lead to non-central two-party equilibria (Palfrey, 1984; Weber, 1992; Bol et al., 2016).

Dynamic models of candidate positioning. In addition to the work on one-shot games, there is also a literature on candidate positioning dynamics (Duggan and Martinelli, 2017), although in contrast to our work, the focus of this literature has been on rational two-candidate contests. One notable early paper in this line of work studies a two-party system where the party which lost the previous election is allowed to reformulate its policy to maximize votes in the next election, which can yield predictable trajectories even in higher-dimensional policy spaces (Kramer, 1977). As in the one-shot literature described above, extensions of this model of two-party dynamics have added a variety of features, including policy motivation (Wittman, 1977; Chappell and Keech, 1986), forward-looking parties (Rosenthal, 1982; Forand, 2014; Nunnari and Zápal, 2017), and—most closely related to our work—boundedly-rational candidates who are unable to exactly optimize their positions (Kollman et al., 1992, 1998; Bendor et al., 2011). Our work is set apart from this prior research on electoral dynamics with bounded rationality in our replicator dynamics approach, and our success deriving analytical results for more than two candidates. We are aware of one paper (Laslier and Ozturk Goktuna, 2016) combining a spatial model of elections and replicator dynamics, but the number of parties is fixed to two and the focus is instead on competition between office- and policy-motivated party members ("opportunists"

[1] Hotelling framed the game in terms of two shops positioning themselves along a line (or the design of two competing goods along a single axis), while Downs applied the idea to plurality elections. The two motivations yield equivalent models, so some of the papers we cite use the language of facility location or product design rather than candidate positioning.

and “militants”), where opportunists may defect to the other party.

Evolutionary game theory and replicator dynamics. Replicator dynamics (Taylor and Jonker, 1978; Schuster and Sigmund, 1983; Hofbauer and Sigmund, 2003) were introduced to study the evolution of biological populations, but have since found much broader use. Economists have used evolutionary models—including replicator dynamics—to understand investment behavior (Blume and Easley, 1992), technological innovation (Saviotti and Mani, 1995; Safarzynska and van den Bergh, 2011), and resource harvesting (Noailly et al., 2003), among many other phenomena. See Friedman (1991) for an introduction to evolutionary game theory from an economic perspective and Nelson et al. (2018); Safarzyńska and van den Bergh (2010) for surveys of evolutionary economics. Evolutionary models can even be justified without population-level evolution: models of individual-level learning can give rise to behavior equivalent to replicator dynamics (Börgers and Sarin, 1997); see Bloembergen et al. (2015) for a survey of the connection between replicator dynamics and reinforcement learning. Evolutionary models are much less common in political science than in economics, but have been used to model the corruption of elected officials (Accinelli et al., 2017), coordination by voters (Mebane Jr, 2005), and party defection (Laver and Benoit, 2003). Extensions of the classical replicator dynamics have explored the various modifications found in our model, including multi-way interactions (Gokhale and Traulsen, 2010), discrete time (Losert and Akin, 1983), and a continuous strategy space (Oechssler and Riedel, 2001; Cheung, 2016).

Elections with strategic voters. Another line of research around strategic aspects of elections focuses instead on the strategic choices made by *voters* rather than candidates (Desmedt and Elkind, 2010; Thompson et al., 2013; Obraztsova et al., 2016) (with some papers combining both voter and candidate strategy (Feddersen et al., 1990; Myerson and Weber, 1993)). Dynamics have featured prominently in the strategic voting literature (Dekel and Piccione, 2000; Callander, 2007)—in particular, under the framework of *iterative voting* (Meir et al., 2010; Lev and Rosenschein, 2012; Obraztsova et al., 2015), where voters are allowed to update their votes in successive rounds until they are satisfied. Intriguingly, this style of voter-dynamics analysis can also produce conclusions paralleling Duverger's Law, where two major candidates emerge, despite using a completely different approach to ours (Meir et al., 2014). Evolutionary dynamics have also been applied to voter behavior to explain the paradox of voting (why do people vote when their probability of affecting the outcome is near zero?) (Sieg and Schulz, 1995).

8.2 Replicator dynamics for candidate positioning

We now formally introduce our model. We consider a one-dimensional policy space represented by the unit interval $[0, 1]$. Candidates and voters reside at points in the interval. To model a large population of voters, we treat the voting population as a continuum; for our theoretical analysis, we assume voters are uniform over $[0, 1]$, but we later relax this assumption in simulation. We assume voters have 1-Euclidean preferences (Elkind et al., 2016)—that is, they vote for the closest candidate. The *vote share* of a candidate i is the fraction of voters who vote for i. With uniform voters, the vote share of a candidate is equal to half the distance between the candidates to its left and right (a candidate adjacent to a boundary gets the entire vote share on its boundary side). Under plurality voting, the candidate with the largest vote share wins; in the case of tied maximum vote shares, the tie is broken uniformly at random.

Our replicator dynamics model of candidate positioning supposes that elections proceed in generations $t = 1, 2, \ldots$, with (infinitely) many elections per generation. We assume the number of candidates in each election is fixed at k (later, we relax this assumption in simulation). The core idea of our model is that candidates in generation t chose their policy positions by copying the position of a winner from the previous generation $t-1$. More formally, let F_0 be the initial candidate distribution and let $F_{k,t}$ denote the distribution of winner positions in generation t with k candidates per election. We define $F_{k,0} = F_0$ for all k, although we typically write F_0 since the initial distribution does not depend on k. In generation t, each election consists of k candidates with positions $X_{1,t}, \ldots, X_{k,t} \sim F_{k,t-1}$. We use $F_{k,t}(x)$ to denote the CDF of the winner distribution in generation t and $f_{k,t}(x)$ to denote the PDF. Let $\text{Plurality}(X_{1,t}, \ldots, X_{k,t})$ be the position of the plurality winner given candidate positions $X_{1,t}, \ldots, X_{k,t}$ and uniformly distributed voters.

Definition 3. Given an initial candidate distribution F_0 and a candidate count k, the *replicator dynamics for candidate positioning* (under plurality with uniform 1-Euclidean voters) are, for all $t > 0$,

$$\begin{aligned} F_{k,t}(x) &= \Pr(\text{Plurality}(X_{1,t}, \ldots, X_{k,t}) \leq x), \\ X_{i,t} &\sim F_{k,t-1}, \ \forall i = 1, \ldots, k. \end{aligned} \tag{8.1}$$

Or, in terms of the PDF:

$$f_{k,t}(x) = k \cdot \Pr(\text{Plurality}(x, X_{2,t}, \ldots, X_{k,t}) = x) \cdot f_{k,t-1}(x). \tag{8.2}$$

This model can be viewed through the lens of evolutionary replicator dynamics (Taylor and Jonker, 1978; Schuster and Sigmund, 1983; Hofbauer and Sigmund, 2003), although there are several differences from the classical case. In the classic replicator dynamics, there are n discrete strategies, each of which increases in frequency proportionally to how well that strategy performs against the current population. This proportionality is exactly what Equation (8.2) captures: strategy x increases in density proportional to its plurality win rate against the current population.

The main question we study is how the candidate distribution evolves over time under the replicator dynamics. We focus on cases where F_0 is symmetric about $1/2$ and contains no point masses (i.e., the initial CDF $F_0(x)$ is continuous); we call such distributions *symmetric* and *atomless*. This ensures that the probability multiple candidates share the exact same point is 0, so we can ignore these cases for now. Since we assume F_0 is symmetric, all subsequent winner distributions are also symmetric by the symmetry of plurality with a uniform voter distribution—we lean heavily on this fact in our analysis. Some of our results require an additional assumptions on F_0. We say F_0 is *positive near 1/2* if $F_0(x) < 1/2$ for all $x < 1/2$ (equivalently, $f_0(x) > 0$ in an interval around $1/2$); the symmetry of F_0 allows us to phrase definitions like this in terms of the left half of the unit interval, and it then applies equivalently to the right half as well. We define $\mathcal{F}$ to be the set of all symmetric and atomless distributions over $[0, 1]$ and $\mathcal{F}^+ \subset \mathcal{F}$ to be the subset of such distributions which are also positive near $1/2$.

In this section, we prove our main result piece-by-piece.

Theorem 32. *Let $F_0 \in \mathcal{F}^+$. For $k \in \{2, 3, 4\}$, the candidate distribution converges to a point mass at $1/2$ under the replicator dynamics. In contrast, for $k \geq 5$, the candidate distribution does not converge to a point mass at $1/2$.*

Theorem 32 follows from Theorems 33 to 36. Our results for $k \in \{2, 3, 4\}$ give fine-grained characterizations of the dynamics, which imply convergence to the center: for $k = 2$, we derive a closed form for the CDF at generation t (Theorem 33), while for $k = 3$ and 4 we derive closed-form bounds for the CDF (Theorems 34 and 35). The negative portion of Theorem 32 offers less insight into the dynamics for $k \geq 5$, only showing non-convergence to the center (Theorem 36), but in Section 8.4 we prove a stronger result in the special case where F_0 has no extreme candidates. All proofs omitted from this section for the sake of readability can be found in Appendix G.2.1.

8.2.1 $k = 2$

Two-candidate plurality with symmetric voters is simple: whichever candidate is closer to 1/2 has the larger vote share and wins. This simplicity allows us to fully characterize the dynamics with $k = 2$. In particular, we derive a closed form for the CDF $F_{2,t}(x)$.

Theorem 33. *Let $F_0 \in \mathcal{F}$. For all $x < 1/2$ and $t \geq 0$, $F_{2,t}(x) = [2 \cdot F_0(x)]^{2^t}/2$.*

Proof. Let $x < 1/2$. Since the candidate closer to 1/2 wins with $k = 2$, $\text{Plurality}(X_{1,t}, X_{2,t}) \notin (x, 1-x)$ if and only if both $X_{1,t} \notin (x, 1-x)$ and $X_{2,t} \notin (x, 1-x)$, which occurs with probability $(2 \cdot F_{2,t-1}(x))^2$. By symmetry, we then have $F_{2,t}(x) = \Pr(\text{Plurality}(X_{1,t}, X_{2,t}) \leq x) = (2 \cdot F_{2,t-1}(x))^2/2 = 2 \cdot F_{2,t-1}(x)^2$. We can now prove the claim by induction on t. For the base case $t = 0$, $(2 \cdot F_0(x))^{2^0}/2 = F_0(x)$. For the inductive case $t \geq 1$, applying the inductive hypothesis yields:

$$F_{2,t}(x) = 2 \cdot F_{2,t-1}(x)^2 = 2\left[(2 \cdot F_0(x))^{2^{t-1}}/2\right]^2 = [2 \cdot F_0(x)]^{2^t}/2. \qquad \square$$

This result shows that for $k = 2$, the CDF at any point $x < 1/2$ with $F_{2,0}(x) < 1/2$ rapidly goes to 0—that is (apart from degenerate initial distributions) the candidate distribution converges toward a point mass at 1/2.

Corollary 6. *Let $F_0 \in \mathcal{F}^+$. For all $x < 1/2$, $\lim_{t\to\infty} F_{2,t}(x) = 0$.*

Note that by symmetry, it follows from such a statement that for all $x > 1/2$, $\lim_{t\to\infty} F_{2,t}(x) = 1$.

8.2.2 $k = 3$

For $k = 3$, plurality becomes more complex: the winner need not be the closest to 1/2 (for instance, consider candidates at positioned at $1/3, 1/2$, and $2/3$). Nonetheless, we can still show that the candidate distribution converges to the center. To do so, we find an upper bound on $F_{3,t}(x)$ which goes to 0. The idea behind the proof is to enumerate cases where a candidate in an inner interval $(x, 1-x)$ wins and add up the probability of these cases, as a function of $F_{k,t-1}(x)$. For example, if there are two candidates in $[0, x)$ and one in $(1/2, 1-x)$, then the candidate in $(1/2, 1-x)$ gets vote share greater than 1/2 and wins; this case occurs with probability $\binom{3}{2} F_{3,t-1}(x)^2 \cdot (1/2 - F_{3,t-1}(x))$. By symmetry, we can then transform this lower bound on the probability the winner is inside $(x, 1-x)$ into an upper bound on $F_{3,t}(x)$, the probability that the winner is in $[0, x]$.

Theorem 34. *Let $F_0 \in \mathcal{F}$. For all $x < 1/2$ and $t > 0$,*

$$F_{3,t}(x) \leq 3/4 \cdot F_{3,t-1}(x) + F_{3,t-1}(x)^3. \tag{8.3}$$

This can be written as a looser closed form

$$F_{3,t}(x) \leq F_0(x) \cdot \left[3/4 + F_0(x)^2\right]^t. \tag{8.4}$$

This result reveals that the candidate distribution for $k = 3$ also converges rapidly to the center. In particular, (8.4) shows that the CDF at any point x with $F_0(x) < 1/2$ decays exponentially towards 0 in t. This can also be seen by analyzing the cubic iterated map suggested by the upper bound (8.3), which converges to a stable fixed point at 0 for all initial values in $[0, 1/2)$.

Corollary 7. *Let $F_0 \in \mathcal{F}^+$. For all $x < 1/2$, $\lim_{t\to\infty} F_{3,t}(x) = 0$.*

8.2.3 $k = 4$

As with $k = 3$, we derive an upper bound on the CDF which converges to 0. However, the bound suggests convergence is much slower for $k = 4$ than for 2 or 3 (which we will see later confirmed in simulation). The proof follows the same case-enumeration strategy as $k = 3$, but simple cases only show that the CDF is non-increasing in t for $x \in (1/3, 1/2)$, giving us the following lemma.

Lemma 4. *Let $F_0 \in \mathcal{F}$. For all $x \in (1/3, 1/2)$ and $t \geq 0$, $F_{4,t}(x) \leq F_{4,0}(x)$.*

By using Lemma 4, we can strengthen the case analysis with one additional case that tips the recurrence from breaking even to shrinking exponentially towards 0. However, the base of the exponential depends very strongly on x, increasing rapidly towards 1 near 1/2.

Theorem 35. *Let $F_0 \in \mathcal{F}$. For all $x \in (1/3, 1/2)$ and $t \geq 0$,*

$$F_{4,t}(x) \leq F_0(x) \cdot \left[1 - 4(1/2 - F_0(x/3 + 1/3))^3\right]^t. \tag{8.5}$$

Note that $x/3 + 1/3$ is the point two-thirds of the way from x to 1/2. As long as $F_0(x/3 + 1/3) < 1/2$, which is true for any $x < 1/2$ if F_0 is positive near 1/2, this result shows that the CDF left of $1/2$ decays to 0 as t grows. That is, the candidate distribution converges to the center again.

Corollary 8. *Let $F_0 \in \mathcal{F}^+$. For all $x < 1/2$, $\lim_{t\to\infty} F_{4,t}(x) = 0$.*

8.2.4 $k \geq 5$

In contrast to $k = 2, 3, 4$, we now show that for any larger k, the candidate distribution does not converge to the center. The proof is based on the following observation.

Lemma 5. *For any k, if all candidates are in $(1/4, 3/4)$, then only the left- or rightmost candidate can win with uniform voters.*

Proof. Suppose all candidates are in $(1/4, 3/4)$. Any candidate between two others gets vote share less than $(1/2)/2 = 1/4$, since no two candidates are distance $1/2$ or greater apart. Meanwhile, the left- and rightmost candidates each get vote share $> 1/4$. $\square$

Intuitively, if the candidate distribution starts converging to the center, then all candidates will likely be inside $(1/4, 3/4)$, at which point only the most extreme candidates can win. When k is sufficiently large (i.e., ≥ 5), the left- and rightmost candidates are likely on opposite sides and farther from $1/2$ than the average candidate. This results in a centrifugal force preventing further progress towards the center. Formally, we prove the following theorem.

Theorem 36. *Let $F_0 \in \mathcal{F}$. For any $k \geq 5$, there exists some $x < 1/2$ such that $\lim_{t\to\infty} F_{k,t}(x) \neq 0$. That is, the candidate distribution does not converge to a point mass at $1/2$.*

More specifically, our proof assumes for a contradiction that the distribution converges to 1/2, so at some generation t^*, the probability mass left of 1/4 must be less than some small α. We then show that the CDF can never decrease at $F_{k,t^*}^{-1}(1/4)$ after generation t^*, since all candidates will likely be inside $(1/4, 3/4)$, causing only the most extreme candidates to win. This contradicts convergence to the center.

8.3 Replicator dynamics with noise

So far, we have assumed that all candidates copy winner positions from the previous generation. We now show that our results still hold in approximate forms if some of the candidates violate this behavior and instead position themselves uniformly at random. This demonstrates a way in which the convergence of the model is robust to alternative specifications.

Definition 4. Given an initial candidate distribution F_0, a candidate count k, and a noise level $\epsilon \in (0, 1]$, the *replicator dynamics for candidate positioning with ϵ-uniform noise* (under plurality with uniform 1-Euclidean voters) are, for all $t > 0$,

$$F_{k,t}^{\epsilon}(x) = \Pr(\text{Plurality}(X_{1,t}^{\epsilon}, \dots, X_{k,t}^{\epsilon}) \leq x), \tag{8.6}$$

$$F_{k,0}^{\epsilon} = F_0$$

$$X_{i,t}^{\epsilon} \sim \begin{cases} \text{Uniform}(0,1) & \text{w.p. } \epsilon, \\ F_{k,t-1}^{\epsilon} & \text{w.p. } 1-\epsilon. \end{cases}$$

As in the noiseless case, we show that the candidate distribution converges to the center under the dynamics with ϵ-uniform noise for $k = 2, 3, 4$ but do not for $k \geq 5$. However, since ϵ-uniform noise introduces non-central candidates at every

t, we need to relax the convergence requirement. The idea behind our notion of approximate convergence is that if we make the noise sufficiently small, then the distribution should get arbitrarily close to a point mass at 1/2. That is, the CDF at any point $x < 1/2$ eventually goes below any positive threshold c, for sufficiently small $\epsilon > 0$.

Definition 5. Let $F_0 \in \mathcal{F}$. The candidate distribution *approximately converges to the center* under the replicator dynamics with ϵ-uniform noise if for all $x \in [0, 1/2)$ and $c > 0$, there exists some $\epsilon_{\max} > 0$ such that if $\epsilon \in (0, \epsilon_{\max}]$, then $\limsup_{t\to\infty} F^{\epsilon}_{k,t}(x) < c$.

We now give the analogue of our main result with ϵ-uniform noise. One additional benefit of adding noise is that we no longer need to assume F_0 is positive near 1/2.

Theorem 37. *Let $F_0 \in \mathcal{F}$. For $k \in \{2, 3, 4\}$, the candidate distribution approximately converges to the center under replicator dynamics with ϵ-uniform noise. In contrast, for all $k \geq 5$, the candidate distribution does not approximately converge to the center.*

Theorem 37 follows from Theorems 38 to 41. See Appendix G.2.2 for proofs omitted from this section.

8.3.1 $k = 2$

We first show that the replicator dynamics with ϵ-uniform noise approximately converge to the center with two candidates. In fact, we can exactly characterize the limiting candidate distribution for $k = 2$. As before with $k = 2$, whichever candidate is closer to 1/2 wins, but now these candidates can either be winner-copiers or randomly positioned. The idea behind the proof is to find an iterated map for $F^{\epsilon}_{2,t}(x)$ and find the stable fixed point it converges to for $x < 1/2$, which we show is smaller than ϵ.

Theorem 38. *Let $F_0 \in \mathcal{F}$. For any $\epsilon \in (0,1)$ and $x \in [0,1/2)$ with ϵ-uniform noise,*

$$\lim_{t\to\infty} F^{\epsilon}_{2,t}(x) = \frac{1-4x\epsilon(1-\epsilon)-\sqrt{1-8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2} \le \epsilon. \tag{8.7}$$

Proof. Let $x < 1/2$ and $p = F^{\epsilon}_{2,t-1}(x)$. Each candidate in generation t is drawn from $F^{\epsilon}_{2,t-1}$ w.p. $(1-\epsilon)$ and from Uniform$(0,1)$ w.p. ϵ (call such candidates *uniform).* Uniform candidates fall outside $(x, 1-x)$ w.p. $2x$, while non-uniform candidates fall outside $(x,1-x)$ w.p. $2p$ by symmetry. A winner in generation t is not in $(x,1-x)$ if and only if both candidates fall outside this interval, which thus occurs with probability

$$\begin{aligned}\Pr(X^{\epsilon}_{1,t} \notin (x,1-x), X^{\epsilon}_{2,t} \notin (x,1-x)) &= \underbrace{(1-\epsilon)^2(2p)^2}_{\text{neither uniform}} + \underbrace{2\epsilon(1-\epsilon)(2x)(2p)}_{\text{one uniform}} + \underbrace{\epsilon^2(2x)^2}_{\text{both uniform}} \\ &= 4p^2(1-\epsilon)^2 + 8px\epsilon(1-\epsilon) + 4x^2\epsilon^2.\end{aligned}$$

By symmetry, we then have

$$\begin{aligned}F^{\epsilon}_{2,t}(x) &= \Pr(X^{\epsilon}_{1,t} \notin (x,1-x), X^{\epsilon}_{2,t} \notin (x,1-x))/2 \\ &= 2p^2(1-\epsilon)^2 + 4px\epsilon(1-\epsilon) + 2x^2\epsilon^2. \end{aligned}\tag{8.8}$$

The claim then follows from the following technical lemma, proved in Appendix G.2.2.

Lemma 6. *For all initial $p \in [0,1/2]$, $\epsilon \in (0,1)$, and $x \in [0,1/2)$, the quadratic iterated map $p' = 2p^2(1-\epsilon)^2 + 4px\epsilon(1-\epsilon) + 2x^2\epsilon^2$ converges to the fixed point $p^* = \frac{1-4x\epsilon(1-\epsilon)-\sqrt{1-8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2} \le \epsilon$.*

$\square$

This result implies approximate convergence to the center: we can simply take $\epsilon_{\max} < c$ and we then have $\lim_{t\to\infty} F^{\epsilon}_{2,t}(x) \leq \epsilon_{\max} < c$.

Corollary 9. *Let $F_0 \in \mathcal{F}$. For $k = 2$, the candidate distribution approximately converges to the center under replicator dynamics with ϵ-uniform noise.*

8.3.2 $k = 3$

We now show approximate convergence to the center for $k = 3$ with ϵ-uniform noise. As in the noiseless case, we cannot fully characterize the limiting distribution, but we are able to bound it for $\epsilon < 1/3$. The proof repeats the case analysis from the proof of Theorem 34 but with ϵ-uniform candidates, which yields a cubic iterated map. We then bound the attracting fixed point of this map, as we did for $k = 2$.

Theorem 39. *Let $F_0 \in \mathcal{F}$. For any $\epsilon \in (0, 1/3)$ and $x \in [0, 1/2)$, $\limsup_{t\to\infty} F^{\epsilon}_{3,t}(x) \leq 1.5\epsilon$.*

As with $k = 2$, this shows that for any $c > 0$, we can pick a small enough ϵ (i.e., $\epsilon < \min\{1/3, 2/3 \cdot c\}$) so that $\limsup_{t\to\infty} F^{\epsilon}_{3,t}(x) < c$.

Corollary 10. *Let $F_0 \in \mathcal{F}$. For $k = 3$, the candidate distribution approximately converges to the center under replicator dynamics with ϵ-uniform noise.*

8.3.3 $k = 4$

As with two and three candidates, we can also show approximate convergence to the center for replicator dynamics with ϵ-uniform noise and four candidates.

However, for $k = 4$, the bound on ϵ required for convergence depends on the point's distance from 1/2—just as the convergence rate did in the noiseless case. We begin with the noisy analogue of Lemma 4.

Lemma 7. *Let $F_0 \in \mathcal{F}$. With ϵ-uniform noise, for any $\epsilon \in (0, 1]$, $x \in (1/3, 1/2)$, and $t > 0$,*

$$F^{\epsilon}_{4,t}(x) \le \epsilon x + (1 - \epsilon)F^{\epsilon}_{4,t-1}(x).$$

Thus, $F^{\epsilon}_{4,t}(x) \le \max\{x, F^{\epsilon}_{4,0}(x)\}$.

We can then apply the same strategy as we did in the noiseless case and analyze the resulting iterated map as we have done for $k = 2$ and 3.

Theorem 40. *Let $F_0 \in \mathcal{F}$. For any $\epsilon \in (0, 1]$ and $x \in (1/3, 1/2)$, let $\beta = 1/2 - \epsilon(x/3+1/3) - (1-\epsilon)\max\{x/3+1/3, F_0(x/3+1/3)\}$. Then $\beta \in (0, 1/2]$ and $\limsup_{t\to\infty} F^{\epsilon}_{4,t}(x) \le \frac{1}{8\beta^3}\epsilon$.*

As long as we make ϵ sufficiently small (relative to $8\beta^3$), the CDF at $x < 1/2$ eventually goes below any desired threshold c—although the closer x is to 1/2, the smaller β becomes, and likewise the required ϵ.

Corollary 11. *Let $F_0 \in \mathcal{F}$. For $k = 4$, the candidate distribution approximately converges to the center under replicator dynamics with ϵ-uniform noise.*

8.3.4 $k \ge 5$

Now that we have seen approximate convergence to the center for $k = 2, 3, 4$, we prove that this does not happen for any higher k. The argument uses the same idea as in Theorem 36 ($k \geq 5$ without noise): that when all candidates are in $(1/4, 3/4)$, only the left- and rightmost candidates can win. As long as we make ϵ sufficiently small, the exact same approach applies, albeit with some added care to account for randomly positioned candidates.

Theorem 41. *Let $F_0 \in \mathcal{F}$. For $k \geq 5$, the candidate distribution does not approximately converge to the center under replicator dynamics with ϵ-uniform noise.*

8.4 Positive results for $k \geq 5$ with no extreme candidates

In the previous sections, our results for $k \geq 5$ have been negative, showing the candidate distribution does not converge to the center, but without indicating what the distribution converges to instead. While simulations indicate a tendency towards a two-spike equilibrium, we have not been able to theoretically characterize the limiting distribution in general for $k \geq 5$, either with or without noise. However, Lemma 5 enables us to analyze the dynamics for $k \geq 5$ (without noise) in the special case that F_0 has no extreme candidates, with support only on $(1/4, 3/4)$. In this setting, the dynamics are much simpler, as only the left- and rightmost candidates can win. The same type of argument we used before for $k \geq 5$ then provides a positive result, showing that the candidate distribution converges to one with zero mass in an interval around $1/2$. In contrast, our central convergence results for $k \in \{2, 3, 4\}$ still hold in this special case. Proofs of results in this section can be found in Appendix G.2.3.

Theorem 42. *Suppose $F_0 \in \mathcal{F}$ is supported on $(1/4, 3/4)$. Let $\ell = [1 - \sqrt{3/7}]/2 = 0.172\ldots$. For $k \geq 5$ and $x \in (F_0^{-1}(\ell), 1/2)$, $\lim_{t\to\infty} F_{k,t}(x) = 1/2$.*

When F_0 is Uniform$(1/4, 3/4)$, note that $F_0^{-1}(0.172\ldots) = 0.336\ldots$, so The-

Table 8.1: Base of the exponential from Theorem 43 for small k.

k	2	3	4	5	6
$k(1/2)^{k-2}$	2	3/2	1	5/8	3/8

orem 42 implies that as $t \to \infty$, the candidate density in $[0.34, 0.66]$ goes to 0 for $k \geq 5$. With no extreme candidates, we can also precisely characterize the density of the candidate distribution at 1/2 using a simple argument. Since only the left- or rightmost candidates can win, a candidate i at 1/2 only wins if the other candidates are all on the left or on the right. By symmetry, this occurs with probability $2 \cdot (1/2)^{k-1} = (1/2)^{k-2}$. Accounting for the k-fold symmetry in choosing candidate i and applying an inductive argument based on Equation (8.2) then gives the following result.

Theorem 43. *Suppose $F_0 \in \mathcal{F}$ is supported on $(1/4, 3/4)$. For any $k \geq 2$ and $t \geq 0$,*

$$f_{k,t}(1/2) = f_0(1/2) \cdot \left[k(1/2)^{k-2}\right]^t. \tag{8.9}$$

With support on $(1/4, 3/4)$, the behavior of the density at 1/2 therefore depends on whether $k(1/2)^{k-2}$ is smaller or larger than 1. This quantity is smaller than 1 for $k \geq 5$, larger than 1 for $k = 2, 3$ and equal to 1 for $k = 4$ (see Table 8.1). The larger k is, the more rapidly the density at 1/2 goes to 0. This simple argument reveals a mechanism driving the $k < 5$ vs $k \geq 5$ divide: $(1/2)^{k-2}$ is exactly the probability that a central candidate is not flanked. This probability decreases rapidly with k and is counterbalanced at first by the increasing number of candidates k who can be at the center—by as soon as $k \geq 5$, the exponentially low probability of being the left- or rightmost candidate at the center becomes too small.

Theorem 43 is particularly interesting for $k = 4$, since we know the distribution converges to a point mass at 1/2, but the density at 1/2 stays constant at $f_0(1/2)$

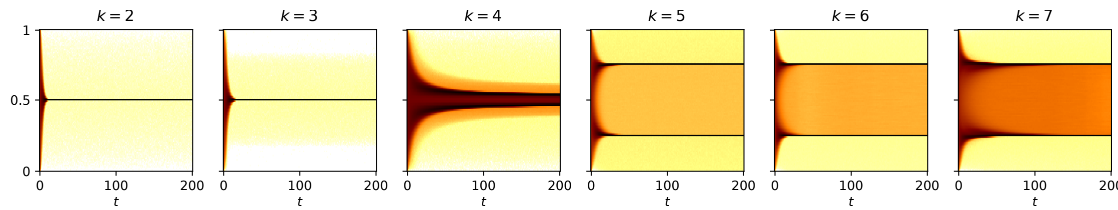


Figure 8.3: Replicator dynamics runs with 0.01-uniform noise for $k = 2, \ldots, 7$ and 200 generations, using enhanced symmetry, 50 trials per plot, and 100,000 elections per generation. The behavior is qualitatively identical to the model without noise (Figure 8.2).

when F_0 is supported on $(1/4, 3/4)$. These seemingly contradictory facts are both possible since the distribution converges by accumulating more and more mass in two spikes on each side of $1/2$ that approach the center arbitrarily closely. Theorem 43 thus highlights how $k = 4$ is a marginal tipping point which just barely converges to the center—a phenomenon also hinted at by the marginal nature of our $k = 4$ case analysis in Lemma 4 and Theorem 35: the analysis in Lemma 4 just breaks even, with the low-probability case in Theorem 35 needed to tip the scales. As we saw in Figure 8.2, this manifests in simulation as slow convergence to the center for $k = 4$.

8.5 Simulations

Having established our primary theoretical results, we demonstrate them in simulation.[2] To do so, we use Monte Carlo sampling, simulating a large number of elections per generation (100,000) and using the winners to approximate $F_{k,t}$. We initialize F_0 to be uniform. With this basic setup, we observe some effects due purely to sampling, such as oscillations due to small asymmetries in the Monte Carlo samples. In contrast, our theoretical model revolves around an evolving

[2] All of our simulation code and results are available at https://github.com/tomlinsonk/plurality-replicator-dynamics.

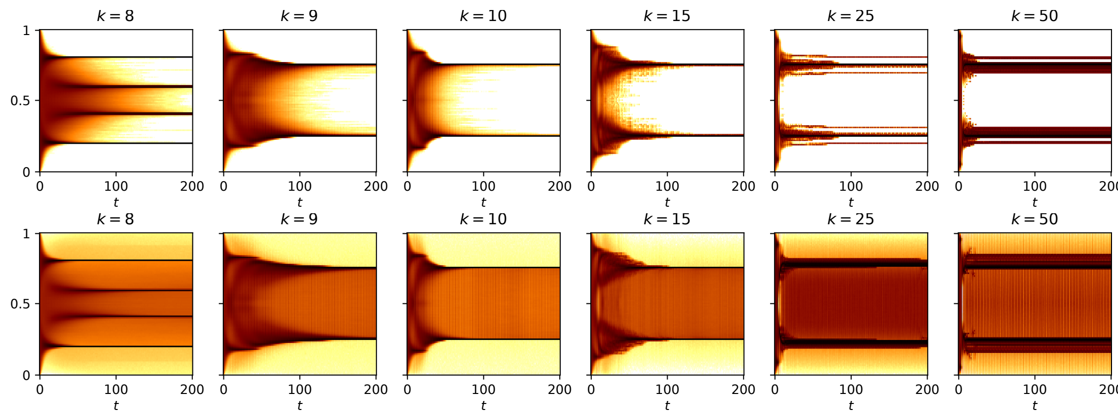


Figure 8.4: Replicator dynamics runs with no noise (top row) and 0.01-uniform noise (bottom row) for larger candidate counts k and using enhanced symmetry. Other settings are identical to Figure 8.3, with 50 runs shown in each plot. As the theory predicts, the candidate distribution does not converge to the center; but the exact behavior varies.

density which by definition is always symmetric. To preserve symmetry while maintaining the same evolving distribution, we configure our Monte Carlo sampling to use a trick we term *enhanced symmetry*, mirroring each copied position across 1/2 with probability 1/2 in every generation.

Figure 8.3 shows 50 aggregated simulation runs for $k = 2, \ldots, 7$ using enhanced symmetry and 0.01-uniform noise (recall Figure 8.2 for equivalent plots without noise). See Appendix G.1 for additional plots without enhanced symmetry and showing a single trial—these results follow the same general patterns across values of k. The candidate distributions evolve exactly as we would expect from our theory: rapid convergence to the center for $k = 2$ and 3, slow convergence for $k = 4$, and non-convergence to the center for $k \geq 5$; both with and without ϵ-uniform noise. Interestingly, $k = 5, 6$, and 7 show a tendency to converge towards two point masses at 1/4 and 3/4—but this phenomenon is sensitive to sampling asymmetries for $k = 6$ and 7 (see Figures G.1 and G.3 in Appendix G.1). In Section 8.4, we will see some theoretical justification for this two-spike behavior in a special case

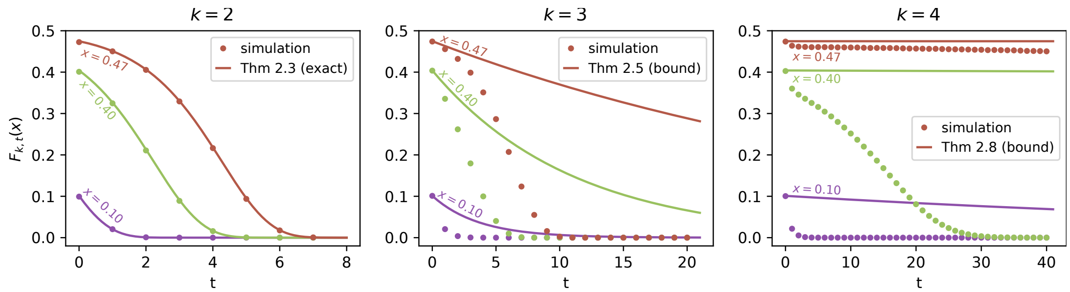


Figure 8.5: Simulations demonstrating our convergence results Theorems 33 to 35, showing the simulated candidate distribution CDF at various points x alongside the theoretical predictions. The simulations use 50 trials with 100,000 elections per generation, no noise, and enhanced symmetry. The theorems get progressively weaker: Theorem 33 provides an exact characterization of the two-candidate dynamics, while Theorems 34 and 35 give upper bounds that converge to 0.

when F_0 has no extreme candidates. In Figure 8.4, we also show simulations with larger candidate counts, from 8 to 50. In line with Theorem 36, the distributions with large k do not converge to the center. However, we see some surprising differences depending on k; for instance, the asymptotic distribution with $k = 8$ appears to have four clusters rather than two (mysteriously, all other large values of k we have tested tend towards two clusters, at least with enhanced symmetry). In Appendix G.1, we provide several additional visualizations: Figure G.4 shows simulations with only 50 elections per generation, demonstrating that our findings still hold in a small-sample setting; Figure G.5 shows large-k simulations without enhanced symmetry; finally, Figure G.6 demonstrates the no-extremes setting of Theorem 42, starting from Uniform$(1/4, 3/4)$.

In addition to confirming the picture painted by our theory, we also use simulations to explore how tight our bounds are—although our core focus is on characterizing the qualitative behavior of the model rather than achieving the tighest bounds on convergence rate. In Figure 8.5, we demonstrate the exact result from Theorem 33 and the closed-form upper bounds on the candidate distribution CDF from Theorems 34 and 35. The bounds on convergence rates for $k = 3$ and 4 are

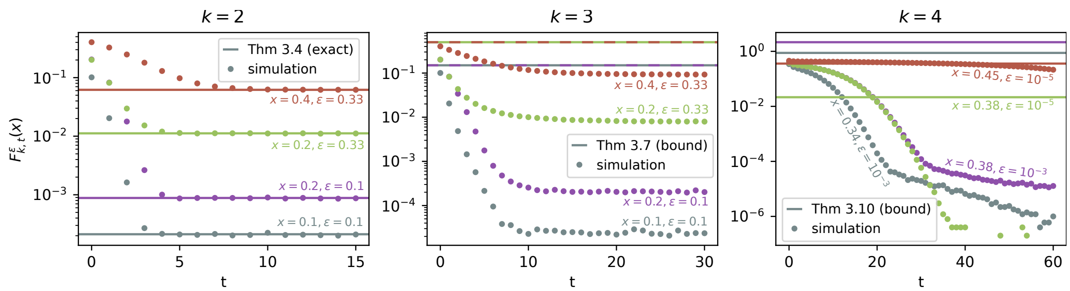


Figure 8.6: Simulations demonstrating Theorems 38 to 40, showing the simulated candidate distribution CDF at various points x and various noise levels ϵ alongside the theoretical asymptotic bounds. The simulations use 50 trials with 100,000 elections per generation and enhanced symmetry. As in the noiseless case, the theorems get progressively weaker as k increases. For $k = 3$, the asymptotic bounds depend only on ϵ, while the bounds for $k = 4$ and the exact limit for $k = 2$ depend on both ϵ and x.

indeed loose, as there are several ways that central candidates can win that are not easily captured by our case analysis. In Figure 8.6, we demonstrate Theorems 38 to 40. Note that these results are all asymptotic, characterizing or bounding the limit of the candidate distribution CDF as $t \to \infty$, whereas the results in Figure 8.5 hold for finite t. Additionally, the results with ϵ-uniform noise depend on the value of ϵ, so we experiment with several different values. Again, we see that our bounds from Theorems 39 and 40 are loose, but nonetheless hold and are non-trivial. Moreover, the exact result in Theorem 38 is nicely confirmed by simulation.

8.6 Variants of the replicator dynamics

We now demonstrate in simulation that the qualitative picture provided by our results from Theorem 32 is robust to different specifications of the model. At a high level, our model of candidate positioning consists of the following components: (1) a fixed voter distribution, (2) a subset of previous candidate positions which new candidates imitate, and (3) a rule for sampling from those previous positions. In the basic model, the voter distribution is uniform, the imitated positions are plurality winners from the previous generation, and the sampling rule chooses uniformly from those winners. Adding ϵ-uniform noise modifies the sampling rule to sometimes pick uniformly random positions. In simulation, we explore natural variations of each of these modeling components: changing the voter distribution, adding plurality winners from earlier generations or runners-up to the imitation pool, and adding copying errors to the sampling rule or sampling different numbers of candidates across elections. See Appendix G.3 for formal definitions of the variants in this section.

Non-uniform voters. We explore our replicator dynamics with symmetric unimodal and bimodal voter distributions. In Figure 8.7, we show results with three voter distributions: the unimodal distribution $\mathrm{Beta}(2, 2)$, the bimodal, extreme voter distribution $\mathrm{Beta}(0.5, 0.5)$, and a bimodal double Weibull distribution (Balakrishnan and Kocherlakota, 1985) with shape 4, location 0.5, and scale 0.3 (see Figure G.7 in Appendix G.1.1 for visualizations of these distributions). The basic pattern from Theorem 32 continues to hold with these voter distributions. However, when voters are $\mathrm{Beta}(2, 2)$-distributed, the two clusters at $k = 5$ are significantly closer to the center.

Memory. In the basic model, candidates only copy the positions of winners in the previous generation. However, real-world candidates will likely have memory of earlier winners, so in this variant, we allow candidates to sample from winner positions in any of the last m generations. In Figure 8.7, we see that adding $m = 2$ generations of memory for candidates still maintains the pattern from Theorem 32.

The results for $m = 3$ are extremely similar (see Figure G.8 in Appendix G.1).

Perturbation noise. With *perturbation noise*, each candidate slightly deviates from the position they copy, as if their imitation is imperfect. In our simulations, we add Gaussian noise with mean 0 and variance σ^2 to each copied position. Figure 8.7 shows that the candidate distribution with a small amount of perturbation noise ($\sigma^2 = 0.005$) converges to the center for $k = 2, 3, 4$ but does not for $k \geq 5$. However, with sufficient noise, higher values of k form a single central cluster; we see this in Figure 8.7 with $k = 6$ and $\sigma^2 = 0.01$. Additionally, for $k \geq 6$ the behavior varies significantly across runs without enhanced symmetry. We even observe phenomena such as party movement, divergence, and extinction, particularly for higher values of k (see Figure G.9 in Appendix G.1).

Variable candidate counts. In real-world elections, we might expect different numbers of candidates to run in different elections, but our model keeps the candidate count k constant. In this variant, we allow elections in each generation to have a mixture of several candidate counts, where candidates copy from winner positions across all k in the previous generation. We find that our results interpolate smoothly to this setting: when most elections have fewer than five candidates, we see convergence to the center, but not when most elections have $k \geq 5$ (see Figure 8.7, where we simulate an equal mixture of the listed candidate counts in each generation, with 50,000 elections per k). See Figure G.10 in Appendix G.1 for a more fine-grained experiment in which we smoothly vary the proportions of elections with $k = 3, 4, 5$.

Top-h copying. Finally, we explore a variant where candidates in generation t choose a position to copy from the pool of candidates with the top-h highest vote shares in generation $t-1$, rather than only winners (i.e., $h = 1$). In simulation, top-h copying is the only variant which strays from the dichotomy we establish in Theorem 32—perhaps unsurprisingly, given that our result is about copying winners. For $k = 3, 4$, when $h = 2$ the candidate distribution does not converge to the center and instead ends up as $k = 5$ usually does, with two clusters (see Figure 8.7, bottom right). For $h = 3$, the candidate distribution does not even appear to converge towards point masses (see Figure G.11 in Appendix G.1.1). These simulations suggest that our central finding (convergence to the center for $k < 5$) is a result of copying the positions of plurality winners specifically, and the dynamics under this heuristic.

8.7 Relationship to Nash equilibria of one-shot games

We now take a step back and examine the relationship between our dynamics and prior research on strategic positioning. As we discussed, much of the literature on candidate positioning has focused on one-shot games rather than dynamics (Osborne, 1995; Bol et al., 2016; Kurella, 2017), as in the Hotelling–Downs model. In our 1-Euclidean setting with uniform voters, the Hotelling–Downs equilibrium has both candidates positioned at 1/2—which as we showed, is also the attracting distribution of the replicator dynamics with $k = 2$. Indeed, it is well-known that Nash equilibria of one-shot games are fixed points of the corresponding replicator dynamics (Hofbauer and Sigmund, 2003), but replicator dynamics fixed points may not be Nash equilibria. We can see this intuitively in our setting by noting that a distribution F is a (symmetric, mixed-strategy) Nash equilibrium if no strategy

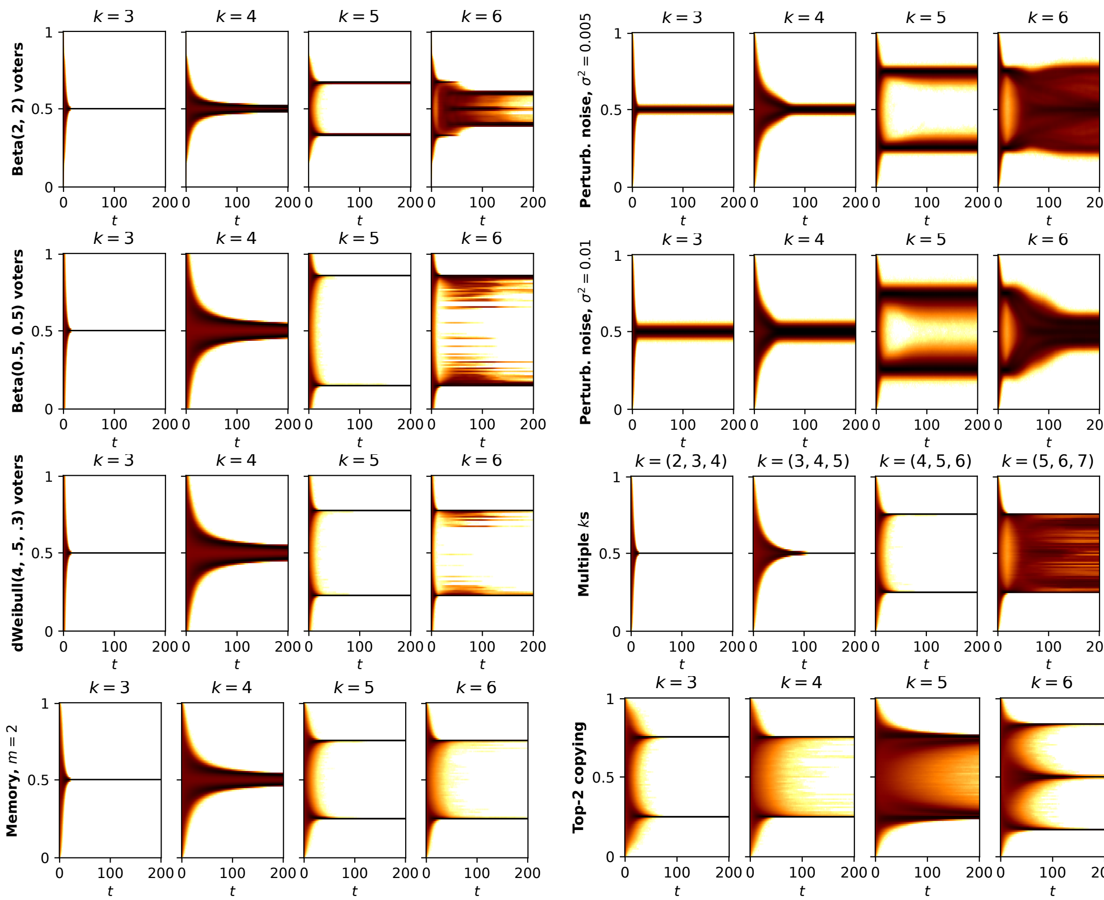


Figure 8.7: Variants of the replicator dynamics. Each plot shows 50 trials with no enhanced symmetry. Left column, top to bottom: three different voter distributions and 2 generations of memory. Right column, top to bottom: perturbation noise with $\sigma^2 = 0.005$ and 0.01, variable candidate counts, and top-2 copying. Except for top-2 copying, all of the variants converge to the center for $k < 5$. Additionally, sufficiently high perturbation noise can cause a central cluster to form for high k.

does better against F than sampling from F, while F is a fixed point of the replicator dynamics if no strategy *drawn from* F does better against F than sampling from F. For F with full support, symmetric mixed-strategy Nash equilibria and replicator dynamics fixed points thus coincide (Bauer et al., 2019). However, Nash equilibria can be unstable under the dynamics—and even if they are attractors, their basins of attraction may be negligible.

Before analyzing Nash equilibria, we first need a brief digression to address what happens when multiple candidates occupy the same point—we call these *positional ties*. Since we have so far assumed that candidate distributions are atomless, our analyses of the replicator dynamics has avoided this issue: with an atomless candidate distribution, positional ties occur with probability 0. One option for handling positional ties is to suppose that candidates fail to position themselves *exactly* at the same point and imagine that there is some infinitesimal jitter in their positions which determines a left–right order. Alternatively, we could suppose that candidates are in fact precisely at the same point, forcing voters to make an arbitrary choice between them.

Definition 6. Suppose multiple candidates occupy the same point. Under *left–right tie-breaking*, one of these candidates (chosen u.a.r.) receives the entire left vote share allocated to that point, while a different candidate (also u.a.r.) receives the entire right vote share. Under *equal split tie-breaking*, all candidates at a point share the vote share allocated to that point equally. Equivalently, voters randomly choose between equidistant candidates.

Armed with these positional tie-breaking rules, we now provide several results that demonstrate how a static analysis of Nash equilibria yields more fragile conclusions than analyzing the asymptotic behavior of the replicator dynamics. We focus on two types of equilibria: (1) symmetric mixed-strategy Nash equilibria (SMSNEs), since these relate to fixed points of the replicator dynamics; and (2) pure-strategy Nash equilibria (PSNEs), since these are the focus of classical candidate positioning analyses.

We begin by showing there are multiple SMSNEs in the one-shot candidate positioning game, but they are often unstable or have tiny basins of attraction under the dynamics; that is, they are unlikely to be relevant in practice. In contrast, as we have seen in theory and simulation, the replicator dynamics behave in qualitatively similar ways under a range of specifications. Then, we show that PSNEs are very sensitive to the choice of positional tie-breaking rule: we arrive at entirely different conclusions if we adopt left–right versus equal split tie-breaking. In contrast, the positional tie-breaking rule is irrelevant to our analysis with atomless candidate distributions.

8.7.1 Symmetric mixed-strategy Nash equilibria

Since SMSNEs are a subset of the replicator dynamics fixed points, we might hope to understand the dynamics by analyzing SMSNEs of the game where candidates seek to maximize their plurality win probability. However, we find that there are multiple SMSNEs and they can have trivial basins of attraction. For instance, every candidate at $1/2$ is a SMSNE and a replicator dynamics fixed point (with left–right tie-breaking[3]). But as we have seen, for $k \geq 5$ all symmetric atomless initial distributions do not converge to a point mass at $1/2$. On the other hand, if we allow initial distributions with point masses and the mass at $1/2$ is sufficiently high, the candidate distribution does indeed approach the all-at-$1/2$ SMSNE.

Theorem 44. *Suppose F_0 places probability mass p at $1/2$. For any $k \geq 2$, there is some $p_k^* < 1$ such that if $p > p_k^*$, the candidate distribution converges to a point mass at $1/2$ under the replicator dynamics with left–right tie-breaking. One of the fixed points of $p^k + kp^{k-1}(1-p)$ is such a p_k^*.*

See Appendix G.2.4 for proofs omitted from this section. Additionally, there

[3] In this subsection, we adopt left–right tie-breaking since it yields equilibria that more closely align with the typical behavior of the replicator dynamics. For instance, we will see in Section 8.7.2 that with equal split tie-breaking, all candidates at $1/2$ is only a SMSNE for $k = 2$—not $k = 3$ or 4, where we know the candidate distribution also converges to the center.

is another family of SMSNEs where each candidate randomly picks between the points x and $1-x$ (for $x \in (1/4, 1/2)$).

Theorem 45. *With $k \geq 4$ and left–right tie-breaking, for any $x \in (1/4, 1/2)$, the strategy where each candidate picks uniformly at random between x and $1-x$ is a SMSNE.*

Just as with the all-at-1/2 equilibrium, this SMSNE is not indicative of the typical behavior of the replicator dynamics. However, we can show as before that for non-atomless distributions, the candidate distribution can converge to this type of equilibrium.

Theorem 46. *Suppose F_0 places probability mass p at x and at $1-x$, for $1/4 < x < 1/2$. For any $k \geq 5$, there exists some $p_k^* < 1/2$ such that if $p > p_k^*$, the candidate distribution converges to point masses at x and $1-x$ under the replicator dynamics. In particular, one of the fixed points of $(2p)^k/2 + k(1-2p)((2p)^{k-1} - 2p^{k-1})/2$ is such a p_k^*.*

These results demonstrate the existence of many SMSNEs that alone do not tell us how we should expect the replicator dynamics to behave.

8.7.2 Positional tie-breaking and pure-strategy Nash equilibria

We now demonstrate how ignoring dynamics and focusing on static equilibria can yield results very sensitive to tie-breaking rules. Cox (1987) extends the Hotelling–Downs analysis to more than two candidates, characterizing PSNEs of a one-shot candidate positioning game—crucially, with equal split tie-breaking. With uniform voters and $k \geq 3$ candidates, Cox proves that there is no PSNE for odd k and that the only PSNE for even k has evenly-spaced pairs of candidates at $1/k, 3/k, \ldots, (k-1)/k$. Clearly, this analysis makes very different predictions than our replicator dynamics. However, we show that Cox's results depend strongly on equal split tie-breaking.

To state Cox's result formally, we need to fully specify the candidate objective. We focus on the objective Cox calls *complete plurality maximization*, where candidates seek first to maximize their vote margin against their strongest competitor, then second-strongest, etc. We extend this objective to allow stochastic positional tie-breaking, assuming candidates first maximize their win probability, then each of their expected vote margins. We can then state Cox's result.

Theorem 47 (Special case of Theorem 2 from Cox (1987))**.** *With uniform voters, $k \geq 3$ complete plurality maximizing candidates, and equal split tie-breaking,*

1. *if k is odd, there is no PSNE,*
2. *if k is even, then the unique PSNE has two candidates at each of the points $1/k, 3/k, \ldots, (k-1)/k$.*

If we instead use left–right tie-breaking, the picture is dramatically different. In particular, all candidates at $1/2$ is then a PSNE *for all k*: any deviant who moves from $1/2$ loses with certainty to the center candidate who captures the opposite side of the vote. Left–right tie-breaking also introduces many additional PSNEs; we list some of them in the following theorem.

Theorem 48. *The following are (some[4] of the) PSNEs with uniform voters, com-*

[4] In Appendix G.2.4, we show that for $k \leq 5$, this list of PSNEs is exhaustive (Theorem 50); for $k > 6$, there may be others.

plete plurality maximizing candidates, and left–right tie-breaking:

1. *Any* $k \geq 2$*: all* k *candidates at* $1/2$*.*
2. *Any* $k \geq 4$*: for any* $x \in (1/4, 1/2)$*,* $\lfloor k/2 \rfloor$ *candidates at* x*,* $\lfloor k/2 \rfloor$ *candidates at* $1-x$*, and the last candidate (if* k *is odd) at either* x *or* $1-x$*.*
3. *Any* $k \geq 5$*:* $\lfloor (k-1)/2 \rfloor$ *candidates at* $1/4$*,* $\lfloor (k-1)/2 \rfloor$ *candidates at* $3/4$*, one candidate at* $1/2$*, and the last candidate (if* k *is even) at either* $1/4$ *or* $3/4$*.*
4. *Even* k*: Cox's equilibrium; two candidates at each of the points* $1/k, 3/k, \ldots, (k-1)/k$*.*

Thus, the qualitative conclusions we arrive by examining Nash equilibria are very different from Cox's if we make another similarly reasonable assumption. Cox's analysis tells us we should not expect candidates converging to the center for any $k > 2$, but if we use left–right tie-breaking, we find that central configurations are equilibria for all k. The replicator dynamics reveal when these configurations are stable: only for small k. These results highlight how analyzing Nash equilibria provides a brittle picture of candidate positioning, yielding results that are sensitive to tie-breaking and do not capture iterated play. Even SMSNEs, which are closely related to replicator dynamics fixed points, fail to reveal the typical behavior of the dynamics.

8.8 Discussion

We introduced a replicator dynamics model of one-dimensional candidate positioning in plurality elections based on simple heuristic inspired by bounded rationality. Our theoretical results show that the candidates converge to the center when there are at most four candidates per election, but diverge when there are five or more candidates per election. Simulations confirm that this pattern is robust to a large range of model variations. We contrast our results to prior work that focuses on static equilibria or lacks theoretical results for more than two candidates.

Many open questions remain in the analysis of our model. The foremost is a theoretical characterization of the asymptotic candidate distribution for $k \geq 5$, although this may be challenging given the complex high-k behavior we observe in simulation. An even larger challenge is posed by expanding beyond symmetric and atomless initial candidate distributions to distributions which have points masses or are asymmetric. As we saw in Theorem 46, allowing atomless distributions means there are infinitely attracting distributions for $k \geq 5$, so the task becomes one of cataloguing all of the possible long-run candidate distributions. Theoretical results for our model variants would be interesting, such as characterizing which mixtures of candidate counts k lead to convergence to the center, or conditions on voter distributions that result in central convergence for $k \leq 4$.

While we explored several model variations in simulation, there are many more than can possibly be covered in a single chapter. Additional variations of particular interest include policy-motivated candidates, strategic voters, probabilistic voters, and higher-dimensional preferences. Another natural direction would be to explore voting systems other than plurality, like two-round runoff, instant runoff, or Borda count; Condorcet methods are considerably less interesting under our one-dimensional replicator dynamics, since the candidate closest to the median voter always wins, but might exhibit more complex behavior in higher dimensions.

Part IV

Conclusion and Future Directions

CHAPTER 9

CONCLUSION AND FUTURE DIRECTIONS

This dissertation applied a computational perspective to two areas of the study of decision-making: discrete choice, at the individual level, and voting, at the collective level. In Part II, we developed tools for recovering contextual and social factors from observational choice data, including new models and adaptations of techniques from causal inference and graph learning. In Part III, we focused on better understanding the consequences of different voting systems, with a particular focus on instant runoff voting and plurality, showing how IRV favors moderate candidates and how imitation can lead to convergent policies under plurality. Bridging the two parts, Chapter 5 applied discrete choice models to an optimal intervention problem in group decision-making. The complexity and opacity of decision-making, both at the individual level where a multitude of factors shape preferences and at the collective level where paradoxes and impossibilities abound, necessitate a variety of different approaches—here, we have seen how computational techniques can contribute.

Of course, we have only scratched the surface. Nonetheless, our explorations have revealed many possible paths for future exploration. Regarding choice models of context effects, it would be valuable to expand the scope of interpretable and learnable models to include more complex relational effects, like the compromise effect, that describe how the interaction of multiple item features influences preferences. Ideally, we should be able to recover such effects from data without baking them explicitly into a model. Some recent approaches have turned to neural networks to capture higher-order preference interactions (Rosenfeld et al., 2020; Pfannschmidt et al., 2022). While they brings the advantage of extreme flexibility, neural networks also come with baggage, namely challenges for interpretability. An ideal middle ground would be some model capable of expressing relational context effects while maintaining the ease of interpretation of a model like the LCL. How we might design and implement such a model is a challenging open question. Given our analysis of the challenges of recovering true context effects in observational choice data, collaboration with experimental discrete choice researchers would be very useful in validating models like the LCL as well as our causal inference tools. On the side of social rather than contextual influences, it would be valuable to compare the predictive and explanatory power of the full-graph social influence methods we developed with classical local-neighborhood feature-based modeling. Other future directions in the discrete choice arena include investigating interventions like assortment optimization using new contextual choice models and exploring such models in online or adaptive settings.

On the collective decision-making side, our worst-case analysis of ballot length could also be applied to other voting systems, such as Borda count, or even to the multi-winner setting, where IRV is called single-transferrable vote (STV). For voting systems that satisfy the Condorcet criterion, like Copeland or ranked pairs, it would be useful to know if some restricted form of the criterion still applies when ballots are truncated. More generally, this expands on the theme of investigating how theoretical properties of voting systems interact with practical considerations of real-world elections. Another direction relating to ballot length regards how voters actually decide on the length of their ideal ranking. Incorporating other behavioral models into our analysis (other than voters merely truncating their ideal rankings to the ballot length) would make it more robust. Regarding our work on moderation, the clear future step is expanding beyond IRV in 1-Euclidean profiles. Two different aspects can be explored: the voting system and the preference model.

One the voting system side, we could ask which non-Condorcet methods have a moderating effect in symmetric 1-Euclidean profiles (any Condorcet method would always elect the most moderate candidate). Moving to other preference models, we could ask if there are exclusion zones (for IRV or other systems) in higher-dimensional Euclidean profiles, or in other preference models like single-peaked, single-crossing, and single-peaked on a tree (Elkind et al., 2022). Our ongoing research in this direction indicates that IRV's exclusion zones disappear in two or more dimensions, but the question remains open for other voting systems. We have also been developing a theory of exclusion zones on graph-structured preferences, where candidates and voters are nodes in a graph and preferences are determined by path length. Finally, our work on replicator dynamics for candidate positioning opens up a broad range of questions. What are the dynamics with higher dimensional preferences? With voting systems other than plurality? With additional behavioral features beyond imitation and random positioning? The varied and chaotic simulation results we observed with some variations of the model indicate that there are complex and subtle phenomena at play.

Part V

Appendices

APPENDIX A

TECHNICAL DETAILS FOR CHAPTER 2: LEARNING INTERPRETABLE FEATURE CONTEXT EFFECTS IN DISCRETE CHOICE

> PREFERENCE, n. A sentiment, or frame of mind, induced by the erroneous belief that one thing is better than another.
>
> ---
>
> Ambrose Bierce, *The Devil's Dictionary*, 1906

A.1 Proofs

The proof of Theorem 1 relies on three lemmas.

Lemma 8 ((Seshadri et al., 2019), Appendix A)**.** *For any choice set C, there is a bijection between the choice probabilities $\{\Pr(i, C) \mid i \in C\}$ and the log probability ratios $\{\beta_{i,C} \mid i \in C\}$ defined by*

$$\beta_{i,C} = \log\left(\frac{\Pr(i, C)}{\left[\prod_{j\in C} \Pr(j, C)\right]^{\frac{1}{|C|}}}\right). \tag{A.1}$$

Proof. We can compute $\beta_{i,C}$ given all choice probabilities in C as defined above. To obtain probabilities given log probability ratios, take

$$\begin{aligned}
\frac{\exp(\beta_{i,C})}{\sum_{j\in C}\exp(\beta_{j,C})} &= \frac{\frac{\Pr(i,C)}{\left(\prod_{h\in C}\Pr(h,C)\right)^{\frac{1}{|C|}}}}{\sum_{j\in C}\frac{\Pr(j,C)}{\left(\prod_{h\in C}\Pr(h,C)\right)^{\frac{1}{|C|}}}}\\
&= \frac{\Pr(i, C)}{\sum_{j\in C}\Pr(j, C)}\\
&= \Pr(i, C).
\end{aligned}$$

□

This means we can prove identifiability from the βs rather than from choice probabilities. We can also get a simple form for $\beta_{i,C}$ under the LCL.

Lemma 9. *In the LCL,* $\beta_{i,C} = (\theta + Ay_C)^T(y_i - y_C)$.

Proof. Define $\theta_C = \theta + Ay_C$ for brevity.

$$\begin{aligned}
\beta_{i,C} &= \log\left(\frac{\Pr(i, C)}{\left(\prod_{h\in C}\Pr(h, C)\right)^{\frac{1}{|C|}}}\right) \\
&= \log\left(\frac{\exp\left(\theta_C^T y_i\right)}{\sum_{j\in C}\exp\left(\theta_C^T y_j\right)}\Bigg/\left(\prod_{h\in C}\frac{\exp\left(\theta_C^T y_h\right)}{\sum_{j\in C}\exp\left(\theta_C^T y_j\right)}\right)^{\frac{1}{|C|}}\right) \\
&= \log\left(\frac{\exp\left(\theta_C^T y_i\right)}{\left[\prod_{h\in C}\exp\left(\theta_C^T y_h\right)\right]^{\frac{1}{|C|}}}\right) \\
&= \theta_C^T y_i - \frac{1}{|C|}\sum_{h\in C}\theta_C^T y_j \\
&= \theta_C^T(y_i - y_C).
\end{aligned}$$

□

Let vec(A) denote the vectorization of the matrix A (the vector formed by stacking the columns of A).

Lemma 10 (Special case of the vec trick, (Roth, 1934))**.** *For any vectors* $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ *and matrix* $A \in \mathbb{R}^{m\times n}$, $x^T Ay = (y \otimes x)^T \operatorname{vec}(A)$.

Proof.

$$\begin{aligned}
x^T A y = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_i y_j &= \sum_{j=1}^{n} y_j \sum_{i=1}^{m} A_{ij} x_i \\
&= \begin{bmatrix} y_1 x \\ y_2 x \\ \vdots \\ y_n x \end{bmatrix}^T \operatorname{vec}(A) \\
&= (y \otimes x)^T \operatorname{vec}(A).
\end{aligned}$$

□

With these facts in hand, we are ready to prove Theorem 1.

Proof of Theorem 1. Consider the log probability ratio of an item i appearing in choice set C:

$$\begin{aligned}
\beta_{i,C} &= (y_i - y_C)^T (\theta + A y_C) && \text{(by Lemma 9)} \\
&= (y_i - y_C)^T \begin{bmatrix} A & \theta \end{bmatrix} \begin{bmatrix} y_C \\ 1 \end{bmatrix} \\
&= \left(\begin{bmatrix} y_C \\ 1 \end{bmatrix} \otimes (y_i - y_C) \right)^T \operatorname{vec}\left(\begin{bmatrix} A & \theta \end{bmatrix} \right). && \text{(by Lemma 10)}
\end{aligned}$$

Let $m = |\{(i, C) \mid C \in \mathcal{C}_\mathcal{D}, i \in C\}|$ be the number of distinct (item, choice set) pairs in the dataset. Index these pairs from 1 to m. We construct the following

$m \times (d^2 + d)$ linear system by stacking all the $\beta_{i,C}$ equations:

$$\begin{bmatrix} \left(\begin{bmatrix} y_{C_1} \\ 1 \end{bmatrix} \otimes (y_{i_1} - y_{C_1}) \right)^T \\ \vdots \\ \left(\begin{bmatrix} y_{C_m} \\ 1 \end{bmatrix} \otimes (y_{i_m} - y_{C_m}) \right)^T \end{bmatrix} \operatorname{vec}\left(\begin{bmatrix} A & \theta \end{bmatrix} \right) = \begin{bmatrix} \beta_{i_1,C_1} \\ \vdots \\ \beta_{i_m,C_m} \end{bmatrix}.$$

Supposing the choice probabilities are generated according to the LCL, this system is consistent (although it is highly overdetermined with a large dataset). Any solution to this system is a setting of the parameters θ, A that results in the observed log probability ratios (and therefore choice probabilities, by Lemma 8). Since we know the system is consistent, it has a unique solution (i.e., the LCL is identifiable) if and only if the rows of the matrix span $\mathbb{R}^{d^2+d}$. $\square$

Proof of Proposition 1. Suppose that $y_{C_1}, \ldots, y_{C_k}$ $(k < d+1)$ is a maximal set of affinely independent mean feature vectors appearing in the dataset $\mathcal{D}$. In each one of these choice sets C_i, the choice probabilities are determined by $\theta_{C_i} = \theta + Ay_{C_i}$. However, since $k < d+1$, there are infinitely many affine transformations $\theta + Ay_{C_i}$ that map every y_{C_i} to its corresponding θ_{C_i}. For any other choice set $C' \notin \{C_1, \ldots, C_k\}$, we can express its mean feature vector as an affine combination $y_{C'} = \sum_{i=1}^k \alpha_i y_{C_i}$, where $\sum_{i=1}^k \alpha_i = 1$. We then have $\theta_{C'} = \theta + A(\sum_{i=1}^k \alpha_i y_{C_i}) = \sum_{i=1}^k \alpha_i(\theta + Ay_{C_i}) = \sum_{i=1}^k \alpha_i \theta_{C_i}$, so any of the infinitely many affine transformations that correctly map y_{C_i} to θ_{C_i} will also map $y_{C'}$ to $\theta_{C'}$. This means there are infinitely many parameter settings θ and A that would result in the same choice probabilities, so the LCL is not identifiable. $\square$

Proof of Proposition 2. We will use differences in log probability ratios to first identify the choice set dependent utilities $\theta_C = \theta + Ay_C$ in each choice set and then combine those to determine θ and A.

To remove a dependence on mean feature vectors, consider the difference of two log probability ratios in the same choice set:

$$\begin{aligned}\beta_{i_1,C} - \beta_{i_2,C} &= \theta_C^T(y_{i_1} - y_C) - \theta_C^T(y_{i_2} - y_C) && \text{(by Lemma 9)}\\ &= \theta_C^T(y_{i_1} - y_{i_2}).\end{aligned}$$

In order to identify the vector θ_C, form the following linear system from d such differences, all in the same choice set C:

$$\begin{bmatrix}(y_{i_1} - y_{i_0})^T \\ (y_{i_2} - y_{i_0})^T \\ \vdots \\ (y_{i_d} - y_{i_0})^T\end{bmatrix} \theta_C = \begin{bmatrix}\beta_{i_1,C} - \beta_{i_0,C} \\ \beta_{i_2,C} - \beta_{i_0,C} \\ \vdots \\ \beta_{i_d,C} - \beta_{i_0,C}\end{bmatrix}$$

If the rows of the matrix are linearly independent, then we can uniquely solve this system to find θ_C. For this to be the case, we need the $d+1$ feature vectors $y_{i_0}, \ldots, y_{i_d}$ to be affinely independent.

In order to recover θ and A, we need to solve the affine system $\theta + Ay_C = \theta_C$ for θ and A given observations of y_C and θ_C. Affine transformations in d dimensions are uniquely specified by their action on a set of $d+1$ affinely independent vectors. So, if we have $d+1$ observed choice sets $C_0, \ldots, C_d$ whose mean feature vectors $y_{C_0}, \ldots, y_{C_d}$ are affinely independent (and if we know $\theta_{C_0}, \ldots, \theta_{C_d}$), then we can uniquely identify θ and A. As we have seen, we can find $\theta_{C_0}, \ldots, \theta_{C_d}$ if each of $C_0, \ldots, C_d$ has $d+1$ items with affinely independent feature vectors. $\square$

Algorithm 2 EM algorithm for estimating DLCL parameters.

Input: m observations $\mathcal{D}$, d features
$A^{(0)}, B^{(0)} \leftarrow d \times d$ randomly initialized matrices
$\pi^{(0)} \leftarrow d$-dimensional vector with all entries equal to $\frac{1}{d}$
$t \leftarrow 0$
while not converged **do**
  $p_{hk} \leftarrow \frac{\exp\left([B_k^{(t)}+A_k^{(t)}(y_C)_k]^T y_i\right)}{\sum_{j\in C}\exp\left([B_k^{(t)}+A_k^{(t)}(y_C)_k]^T y_j\right)}$
   for each $(i, C) = \mathcal{D}_h$ and $k = 1, \ldots, d$
  $r_{hk} \leftarrow \frac{\pi_k^{(t)} p_{hk}}{\sum_{g=1}^{d} \pi_g^{(t)} p_{hg}}$ for each $h = 1, \ldots, m$ and $k = 1, \ldots, d$

$$Q(A, B \mid \theta^{(t)}) \leftarrow \sum_{(i,C)=\mathcal{D}_h} \sum_{k=1}^{d} r_{hk}\Big[[B_k + A_k(y_C)_k]^T y_i - \log \sum_{j\in C} \exp\left([B_k + A_k(y_C)_k]^T y_j\right)\Big]$$

  Find a minimizer A^*, B^* of $-Q(A, B \mid \theta^{(t)})$ using gradient descent
  $A^{(t+1)} \leftarrow A^*$, $B^{(t+1)} \leftarrow B^*$
  $\pi_k^{(t+1)} \leftarrow \frac{1}{|\mathcal{D}|}\sum_{h=1}^{|\mathcal{D}|} r_{hk}$ for each $k = 1, \ldots, d$
  $t \leftarrow t + 1$
return $A^{(t)}, B^{(t)}, \pi^{(t)}$

A.2 EM algorithm for DLCL estimation

Let $\mathcal{D}_h$ denote hth observation (i, C) and $\Delta_h \in \{1, \ldots, d\}$ denote the latent mixture component that the observation $\mathcal{D}_h$ comes from (taking the view that each observation belongs to one component).

The EM algorithm (see (Hastie et al., 2009) for a general treatment) is an iterative procedure that begins with initial guesses for the parameters $\theta^{(0)} = (A^{(0)}, B^{(0)}, \pi^{(0)})$ and updates them until convergence. In the update step, we maximize the expectation of the log-likelihood $\ell(A, B; \mathcal{D}, \Delta)$ over the distribution of the unobserved variable Δ conditioned on the observations $\mathcal{D}$ and the current estimates of the parameters, denoted $\mathrm{E}_\Delta[\ell(A, B; \mathcal{D}, \Delta) \mid \mathcal{D}, \theta^{(t)}]$. The new estimates $A^{(t+1)}$ and $B^{(t+1)}$ are the maximizers of this function. The new estimate of the mixture proportions $\pi^{(t+1)}$ has a closed form based on the probability that each observation comes from each mixture component according to the current estimates of A and B. See Algorithm 2 for the complete procedure. We derive the details here, starting with a breakdown of the expectation function:

$$\begin{aligned}
&\mathrm{E}_{\Delta}[\ell(A, B; \mathcal{D}, \Delta) \mid \mathcal{D}, \theta^{(t)}] \\
&\quad = \sum_{(i,C)=\mathcal{D}_h} \sum_{k=1}^{d} \Pr(\Delta_h = k \mid i, C, \theta^{(t)}) \log \Pr(i, C \mid \Delta_h = k, A, B).
\end{aligned} \tag{A.2}$$

We can compute the first part of the summand (the *responsibilities*) using Bayes' Theorem:

$$\Pr(\Delta_h = k \mid i, C, \theta^{(t)}) = \frac{\Pr(i, C \mid \Delta_h = k, \theta^{(t)}) \Pr(\Delta_h = k \mid \theta^{(t)})}{\Pr(i, C \mid \theta^{(t)})} \tag{A.3}$$

$$= \pi_h^{(t)} \frac{\Pr(i, C \mid \Delta_h = k, \theta^{(t)})}{\Pr(i, C \mid \theta^{(t)})}. \tag{A.4}$$

The numerator of Equation (A.4) is the kth component of the DLCL choice probability (with our estimates for A and B):

$$\Pr(i, C \mid \Delta_h = k, \theta^{(t)}) = \frac{\exp\left([B_k^{(t)} + A_k^{(t)}(y_C)_k]^T y_i\right)}{\sum_{j \in C} \exp\left([B_k^{(t)} + A_k^{(t)}(y_C)_k]^T y_j\right)}. \tag{A.5}$$

Meanwhile, the denominator of Equation (A.4) is the sum of these probabilities weighted by the mixture weight estimates:

$$\Pr(i, C \mid \theta^{(t)}) = \sum_{k=1}^{d} \pi_k^{(t)} \Pr(i, C \mid \Delta_h = k, \theta^{(t)}). \tag{A.6}$$

The last term in Equation (A.2) is a function of the parameters A, B (not their estimates):

$$\log \Pr(i, C \mid \Delta_h = k, A, B) = \log\left[\frac{\exp\left([B_k + A_k(y_C)_k]^T y_i\right)}{\sum_{j \in C} \exp\left([B_k + A_k(y_C)_k]^T y_j\right)}\right] \tag{A.7}$$

$$= [B_k + A_k(y_C)_k]^T y_i - \log \sum_{j \in C} \exp\left([B_k + A_k(y_C)_k]^T y_j\right). \tag{A.8}$$

Equation (A.8) is concave by the same reasoning that the LCL's NLL (Equation (2.7)) is convex. Thus, the expectation $\mathrm{E}_{\Delta}[\ell(A, B; \mathcal{D}, \Delta) \mid \mathcal{D}, \theta^{(t)}]$, being the sum of positively scaled concave functions, is also concave. Its gradient is also Lipschitz continuous, just like the LCL's NLL. We can therefore find a global maximum using gradient ascent (in practice, we use gradient descent to minimize $-Q(A, B \mid \theta^{(t)})$).

APPENDIX B

TECHNICAL DETAILS FOR CHAPTER 3: CHOICE SET CONFOUNDING IN DISCRETE CHOICE

> Anything that happens, happens. Anything that, in happening, causes something else to happen, causes something else to happen. Anything that, in happening, causes itself to happen again, happens again. It doesn't necessarily do it in chronological order, though.
>
> Douglas Adams, *Mostly Harmless*, 1992.

B.1 Proofs

Proof of Theorem 3. We need positivity in order for IPW (and therefore $\tilde{D}$) to be well-defined. Fix i and C. Consider the coefficient of $\log \Pr_\theta(i \mid C)$ in $\ell(\theta; \mathcal{D}^*)$. In expectation, this term appears $|\mathcal{D}| \Pr(i, C)$ times. Expanding this:

$$\begin{aligned}
|\mathcal{D}| \Pr(i, C) &= |\mathcal{D}| \sum_{a \in A} \Pr(i, C \mid a) \Pr(a) \\
&= |\mathcal{D}| \sum_{a \in A} \Pr(i \mid C, a) \Pr(C \mid a) \Pr(a) \\
&= \frac{|\mathcal{D}|}{|\mathcal{C}_\mathcal{D}|} \sum_{a \in A} \Pr(i \mid C, a) \Pr(a),
\end{aligned}$$

where the last step follows from $\mathcal{D}^*$ having uniformly random choice sets. Now consider the coefficient of $\log \Pr_\theta(i \mid C)$ in $\ell(\theta; \tilde{\mathcal{D}})$. By IPW, this coefficient is

$$\sum_{\substack{(a, C', i) \in \mathcal{D} \\ i' = i, C' = C}} \frac{1}{|\mathcal{C}_\mathcal{D}| \Pr(C \mid \boldsymbol{x_a})} = \frac{1}{|\mathcal{C}_\mathcal{D}|} \sum_{a \in A} \sum_{\substack{(a', C', i') \in \mathcal{D} \\ a' = a, C' = C, i' = i}} \frac{1}{\Pr(C \mid \boldsymbol{x_a})}.$$

In expectation, the sample (a, C, i) occurs $|\mathcal{D}| \Pr(a, C, i)$ times. Additionally, by choice set ignorability, $\Pr(C \mid \boldsymbol{x_a}) = \Pr(C \mid a)$. We thus have that the expected coefficient is

$$\begin{aligned}
&\frac{1}{|\mathcal{C}_\mathcal{D}|} \sum_{a \in A} \frac{1}{\Pr(C \mid \boldsymbol{x_a})} |\mathcal{D}| \Pr(a, C, i) \\
&= \frac{|\mathcal{D}|}{|\mathcal{C}_\mathcal{D}|} \sum_{a \in A} \frac{1}{\Pr(C \mid a)} \Pr(a) \Pr(C \mid a) \Pr(i \mid C, a) \\
&= \frac{|\mathcal{D}|}{|\mathcal{C}_\mathcal{D}|} \sum_{a \in A} \Pr(a) \Pr(i \mid C, a),
\end{aligned}$$

which matches the coefficient in $\ell(\theta; \mathcal{D}^*)$. Since the expected coefficients agree for all i and C, we then have the equality. □

Proof of Theorem 4. By the consistency of the MLE, as $|\mathcal{D}| \to \infty$, parameter estimates for a correctly specified choice model converge to the true parameters. Thus, estimated choice probabilities also converge:

$$\begin{aligned}
\lim_{|\mathcal{D}| \to \infty} \hat{\Pr}(i \mid \boldsymbol{x_a}, C) &= \Pr(i \mid \boldsymbol{x_a}, C) \\
&= \Pr(i \mid a, \boldsymbol{x_a}, C) && \text{(by preference ignorability)} \\
&= \Pr(i \mid a, C).
\end{aligned}$$

□

Proof of Theorem 5. Observing the choice set gives us a noisy measurement of $\boldsymbol{x_a}$, which we can adjust using our knowledge of the distribution of $\boldsymbol{x_a}$. The posterior of a Gaussian with a Gaussian prior is also Gaussian—in particular, $\boldsymbol{x_a} \mid C$ is Gaussian, with mean

$$\mathrm{E}[\boldsymbol{x_a} \mid C] = \Sigma_0 \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{y_C} + \frac{1}{k}\Sigma \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{\mu}$$

(Duda et al., 2001, Section 3.4.3). Thus, the expected chooser $\boldsymbol{x}_{\boldsymbol{a}}^{*}$ has utilities

$$u_i(a^*, C) = \left[\Sigma_0 \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{y_C} + \frac{1}{k}\Sigma \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{\mu}\right]^T \boldsymbol{y_i}.$$

This is exactly an LCL with $\boldsymbol{\theta}$ and A as claimed. □

Proof of Theorem 6. Consider the bipartite graph whose left nodes are choosers and whose right nodes are items, each split into blocks according to their type. The choice set assignment process above defines a bipartite SBM on this graph with intra-type probabilities p and inter-type probabilities q (between chooser nodes and item nodes). Recovering types from choice sets can then be viewed as an instance of the planted partition problem (McSherry, 2001).

We can thus directly[1] apply Theorem 4 of McSherry (2001) to achieve the desired result given Equation (3.4), with the caveat that algorithm is random and succeeds with probability $1/k$.

Repeating the algorithm ck times achieves failure probability $(1 - \frac{1}{k})^{ck} \leq 1/e^c$, which is smaller than δ if $c > \log(1/\delta)$. We can thus make δ smaller by a factor of 2 (absorbing this into the constant C in eq. (3.4)) and we are left with the guarantee as stated, only increasing the running time by a factor $k \log(1/\delta)$. □

B.2 Affine-mean Gaussian choice set model

For estimating choice set propensities in EXPEDIA, we model the distribution of mean choice set features using an affine-mean Gaussian. Here, we show how this model can be easily estimated from data.

[1]Notice that $s(p-q)^2$ is a lower bound on the squared 2-norm of the columns of the SBM edge probability matrix required by (McSherry, 2001, Theorem 4). Additionally, we use the crude variance upper bound $\sigma^2 = 1$ for simplicity.

Propostion 5. *Given a dataset $\mathcal{D}$, the model $\boldsymbol{y_C} \sim \mathcal{N}(W\boldsymbol{x_a} + \boldsymbol{z}, \Sigma)$ is identifiable iff there are $m+1$ choosers in $\mathcal{D}$ with affinely independent covariates. If the model is identified, the maximum likelihood parameters $W^*, \boldsymbol{z}^*$ are the solution to the least-squares problem*

$$(W^*, \boldsymbol{z}^*) = \underset{\substack{W \in \mathbb{R}^{n\times m} \\ \boldsymbol{z} \in \mathbb{R}^n}}{\arg\min} \sum_{(a,C)\in\mathcal{D}} \|\boldsymbol{y_C} - (W\boldsymbol{x_a} + \boldsymbol{z})\|_2^2, \tag{B.1}$$

which have the closed form:

$$W^* = \left[\sum_{(a,C)\in\mathcal{D}} (\boldsymbol{y_C} - \boldsymbol{y_\mathcal{D}})\boldsymbol{x_a}^T\right] \left[\sum_{(a,C)\in\mathcal{D}} (\boldsymbol{x_a} - \boldsymbol{x_\mathcal{D}})\boldsymbol{x_a}^T\right]^{-1} \tag{B.2}$$

$$\boldsymbol{z}^* = \boldsymbol{y_\mathcal{D}} - W^*\boldsymbol{x_\mathcal{D}}, \tag{B.3}$$

where $\boldsymbol{x_\mathcal{D}} = {}^1/_{|\mathcal{D}|} \sum_{(a,C)\in\mathcal{D}} \boldsymbol{x_a}$ and $\boldsymbol{y_\mathcal{D}} = {}^1/_{|\mathcal{D}|} \sum_{(a,C)\in\mathcal{D}} \boldsymbol{y_C}$.

Additionally, the maximum likelihood covariance matrix is the sample covariance:

$$\Sigma^* = \frac{1}{|\mathcal{D}|} \sum_{(a,C)\in\mathcal{D}} (\boldsymbol{y_C} - W^*\boldsymbol{x_a} - \boldsymbol{z}^*)(\boldsymbol{y_C} - W^*\boldsymbol{x_a} - \boldsymbol{z}^*)^T. \tag{B.4}$$

Proof sketch. This can be derived following the same steps as the standard Gaussian MLE proof (with a bit of extra matrix calculus): (1) take partial derivatives of the log-likelihood with respect to W and $\boldsymbol{z}$, (2) set them to zero, (3) solve for $\boldsymbol{z}$, (4) plug this in to solve for W, (5) do the same to solve for Σ in its partial derivative. This works since the log-likelihood is still convex after adding in the affine trasformation. We omit the details as they are tedious and unenlightening. □

B.3 Experiment details

We implemented all choice models with PyTorch and (except mixed logit) train them using Rprop with no minibatching to optimize the log-likelihood for 500 epochs or until convergence (squared gradient norm $< 10^{-8}$), whichever comes first. We use ℓ_2 regularization with coefficient $\lambda = 10^{-4}$ for all models to ensure identifiability. For mixed logit, we use an expectation-maximization (EM) algorithm (Train, 2009) with a one hour timeout. Our code, results, and links to data are available at `https://github.com/tomlinsonk/choice-set-confounding`.

APPENDIX C

TECHNICAL DETAILS FOR CHAPTER 4: GRAPH-BASED METHODS FOR DISCRETE CHOICE

LYSANDER. Or else it stood upon the choice of friends—

HERMIA. O hell, to choose love by another's eyes!

William Shakespeare, *A Midsummer Night's Dream*, c. 1595

C.1 APP-INSTALL highest-utility apps

In Tables C.1 and C.2 we examine the top 20 apps recommended by logit and Laplacian logit, respectively. For logit, we look directly at utilities, while for the Laplacian-regularized logit we count the number of participants for whom the app in their top 10 by utility. For both, we compute utilities by averaging the best-performing (on validation) model at training fraction 0.8 in each trial.

Table C.1: Top 20 apps by their global logit utility. Bolded apps are deemed to be non-built-in downloadable apps by manual inspection. Built-ins escaping our filter are over-represented at the top of this list since they commonly appear in app scans.

App	**Utility**
edu.mit.media.funf.BluetoothService	2.15
android.tts	1.65
com.svox.pico	1.62
com.tmobile.selfhelp	1.52
com.noshufou.android.su	1.49
org.zenthought.android.su	1.14
com.facebook.katana	1.03
com.shazam.android	1.02
com.qo.android.moto	1.00
com.telenav.app.android.telenav	0.91
com.arcsoft.mediagallery	0.90
com.voxmobili.sync.MobileBackup	0.90
com.metago.astro	0.83
com.cooliris.media	0.78
com.pandora.android	0.78
com.gotvnetworks.client.android.altitude.activity	0.78
com.xtralogic.android.logcollector	0.77
com.orange.maps	0.74
com.myspace.android	0.70
com.telenav.app.android	0.70

Table C.2: Top 20 apps by the number of participants for whom they have top-10 utilities under the Laplacian logit. See Table C.1 for additional details.

App	**Count**
edu.mit.media.funf.BluetoothService	135
com.tmobile.selfhelp	120
android.tts	103
com.svox.pico	96
org.zenthought.android.su	95
com.qo.android.moto	82
com.noshufou.android.su	82
com.shazam.android	70
com.facebook.katana	70
com.voxmobili.sync.MobileBackup	68
com.arcsoft.mediagallery	67
com.cooliris.media	57
com.pandora.android	42
com.telenav.app.android	31
com.qo.android.gep	31
com.gotvnetworks.client.android.altitude.activity	30
com.myspace.android	27
com.telenav.app.android.telenav	26
com.vzw.vvm.androidclient	24
com.xtralogic.android.logcollector	23

APPENDIX D

TECHNICAL DETAILS FOR CHAPTER 5: CHOICE SET OPTIMIZATION UNDER DISCRETE CHOICE MODELS OF GROUP DECISIONS

That was excellently observed, say I, when I read a Passage in an Author, where his Opinion agrees with mine. When we differ, there I pronounce him to be *mistaken*.

Jonathan Swift, *Miscellanies in Prose*, 1735

D.1 Hardness proofs

D.1.1 Disagreement functions from proofs of Theorems 8 and 9

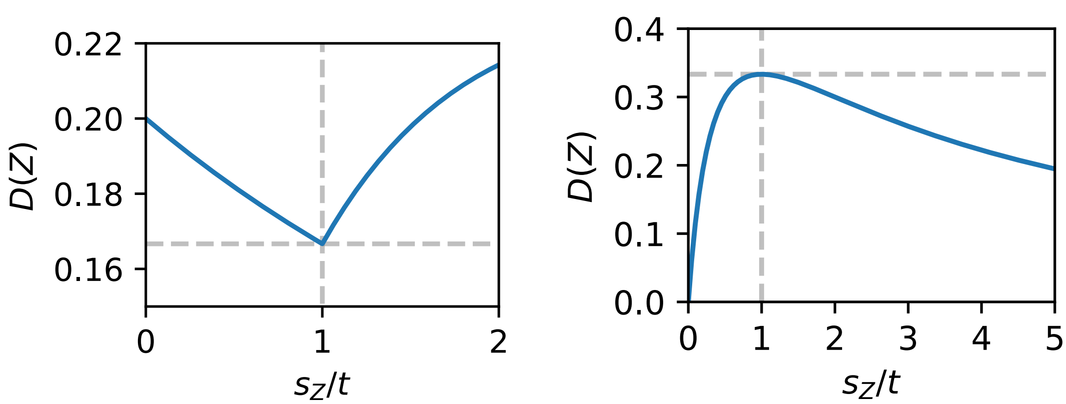


Figure D.1: (Left) Plot of $D(Z) = \left|\frac{t}{2t+s_Z} - \frac{3t}{5t+s_Z}\right| + \left|\frac{t}{2t+s_Z} - \frac{2t}{5t+s_Z}\right|$ from the proof of Theorem 8. (Right) Plot of $D(Z) = \left|\frac{2t}{2t+s_Z} - \frac{t/2}{t/2+s_Z}\right|$ from the proof of Theorem 9. Both functions are re-parameterized in terms of the ratio s_Z/t by dividing through by t and achieve local optima at $s_Z/t = 1$ (i.e. $s_Z = t$); this can be verified analytically.

D.1.2 CDM Promotion is hard with $|A| = 2, |C| = 2$

In the main text, we show CDM Promotion is NP-hard when $|A| = 1, |C| = 3$ (Theorem 10). Here, we provide an additonal proof for the case when $|A| = 2, |C| = 2$. These are the smallest hard instances of the problem ($|A| = 1, |C| = 2$ is easy to solve: introduce alternatives that increase utility for x^* for than its competitor).

Theorem 49. *In the CDM model,* Promotion *is NP-hard, even with just two individuals and two items in C.*

Proof. By reduction from Subset Sum. Let S, t be an instance of Subset Sum. Let $A = \{a, b\}$, $C = \{x, y\}$, $\overline{C} = S$. Using tuples interpreted entrywise, construct a CDM with the following parameters.

$$
\begin{aligned}
u_a(\langle x^*, y\rangle) &= \langle t + \varepsilon, 0\rangle \\
u_b(\langle x^*, y\rangle) &= \langle \varepsilon, t\rangle \\
u_a(z) = u_b(z) &= -\infty && \forall z \in \overline{C} \\
p_a(z, \langle x^*, y\rangle) &= \langle 0, z\rangle && \forall z \in \overline{C} \\
p_b(z, \langle x^*, y\rangle) &= \langle z, 0\rangle && \forall z \in \overline{C}
\end{aligned}
$$

To promote x^*, we need to add more than $t - \varepsilon$ to b's utility for x^*, but add less than $t + \varepsilon$ to a's utility for x^*. Since all pulls are integral, the only solution is a set Z whose sum of pulls is t. If we could efficiently find such a set, then we could efficiently find the Subset Sum solution. □

D.1.3 Proof of Theorem 11

Proof. By reduction from Subset Sum. Let $S = \{z_1, \ldots, z_n\}, t$ be an instance of Subset Sum. Let $A = \{a, b\}$, $C = \{x^*, y\}$, $\overline{C} = S$, and $0 < \varepsilon < 1$. The nest structures and utilities are shown in Figure D.2.

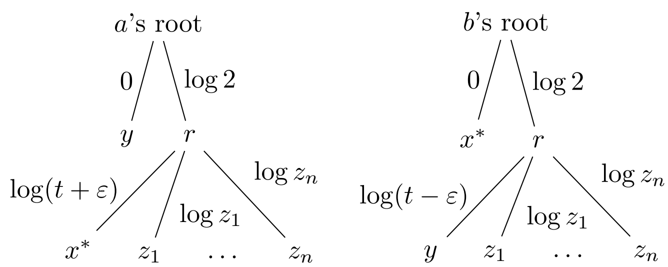


Figure D.2: NL trees used in the proof of Theorem 11. The left tree is for individual a and the right tree for individual b.

Notice that x^* and y are swapped in the two trees. We wish to promote x^*. With just the choice set C, a prefers x^* to y, but b does not. To make b prefer x^* to y, we need to cannibalize y by adding z_i items. However, this simultaneously cannibalizes x^* in a's tree, so we need to be careful not to introduce too much additional utility. To ensure a prefers x^*, we need to pick Z such that

$$
\begin{aligned}
&\Pr(a \leftarrow y \mid C \cup Z) < \Pr(a \leftarrow y \mid C \cup Z) \\
\iff\ & \frac{1}{1 + e^{\log 2}} < \frac{e^{\log 2}}{1 + e^{\log 2}} \cdot \frac{e^{\log(t+\varepsilon)}}{e^{\log(t+\varepsilon)} + \sum_{z \in Z} e^{\log z}} \\
\iff\ & \frac{1}{3} < \frac{2}{3} \cdot \frac{t + \varepsilon}{t + \varepsilon + \sum_{z \in Z} z} \\
\iff\ & \sum_{z \in Z} z < t + \varepsilon.
\end{aligned}
$$

To ensure b prefers x^*, we need

$$\begin{aligned}
\Pr(b \leftarrow x^* \mid C \cup Z) &> \Pr(b \leftarrow y \mid C \cup Z) \\
\iff \frac{1}{1+e^{\log 2}} &> \frac{e^{\log 2}}{1+e^{\log 2}} \cdot \frac{e^{\log(t-\varepsilon)}}{e^{\log(t-\varepsilon)} + \sum_{z\in Z} e^{\log z}} \\
\iff \frac{1}{3} &> \frac{2}{3} \cdot \frac{t-\varepsilon}{t-\varepsilon+\sum_{z\in Z} z} \\
\iff \sum_{z\in Z} z &> t-\varepsilon.
\end{aligned}$$

Since the z are all integers, we must then have $\sum_{z\in Z} z = t$. If we could efficiently promote x^*, we could efficiently find such a Z. □

D.1.4 Proof of Theorem 12

Proof. By reduction from SUBSET SUM. Let S, t be an instance of SUBSET SUM. Let $A = \{a, b\}$, $C = \{x^*, y\}$, $\overline{C} = S$, and $s = \sum_{z\in S} z$. Make aspects χ_z, ψ_z, γ_z for each $z \in S$ as well as two more aspects χ, ψ. The items have aspects as follows:

$$\begin{aligned}
x^{*\prime} &= \{\chi\} \cup \{\chi_z \mid z \in S\} \\
y' &= \{\psi\} \cup \{\psi_z \mid z \in S\} \\
z' &= \{\chi_z, \psi_z, \gamma_z\} && \forall z \in S
\end{aligned}$$

The individuals have the following utilities on aspects, where $0 < \varepsilon < 1$:

$$\begin{aligned}
u_a(\chi) &= 0 & u_b(\chi) &= s - t/2 + \varepsilon \\
u_a(\chi_z) &= z & u_b(\chi_z) &= 0 && \forall z \in S \\
u_a(\psi) &= s - t/2 - \varepsilon & u_b(\psi) &= 0 \\
u_a(\psi_z) &= 0 & u_b(\psi_z) &= z && \forall z \in S \\
u_a(\gamma_z) &= s - z & u_b(\gamma_z) &= s - z && \forall z \in S
\end{aligned}$$

We want to promote x^*. Notice that x^* and y have disjoint aspects. Thus the choice probabilities from C are proportional to the sum of the item's aspects:

$$\Pr(a \leftarrow x^* \mid C) \propto s$$
$$\Pr(a \leftarrow y \mid C) \propto s - \frac{t}{2} - \varepsilon$$
$$\Pr(b \leftarrow x^* \mid C) \propto s - \frac{t}{2} + \varepsilon$$
$$\Pr(b \leftarrow y \mid C) \propto s.$$

To promote x^*, we need to make b prefer x^* to y. Adding a z item cannibalizes from a's preference for x^* and b's preference for y. As in the NL proof, we want to add just enough z items to make b prefer x^* to y without making a prefer y to x^*.

First, notice that the γ_z aspects have no effect on the individuals' relative preference for x^* and y. If we introduce the alternative z, then if a picks the aspect χ_z, y will be eliminated. The remaining aspects of x^*, namely $x^{*\prime} \setminus \{\chi_z\}$, have combined utility $s - z$, as does γ_z. Therefore a will be equallly likely to pick x^* and z. Symmetric reasoning shows that if b chooses aspect ψ_z, then b will end up picking y with probability $1/2$. This means that when we include alternatives $Z \subseteq \overline{C}$, each aspect χ_z, ψ_z for $z \in Z$ effectively contributes $z/2$ to a's utility for x^* and b's utility for y rather than the full z. The optimal solution is therefore a set Z of alternatives whose sum is t, since that will cause a to have effective utility $s - t/2$ on x^*, which exceeds its utility $s - t/2 - \varepsilon$ on y. Meanwhile, b's effective utility on y will also be $s - t/2$, which is smaller than its utility $s - t/2 + \varepsilon$ on x^*. If we include less alternative weight, b will prefer y. If we include more, a will prefer y. Therefore if we could efficiently find the optimal set of alternatives to promote x^*, we could efficiently find a subset of S with sum t. $\square$

D.2 Restrictions on MNL that make Agreement and Disagreement tractable

As we saw in the proofs of Theorems 8 and 9 that Agreement and Disagreement are hard in the MNL model even when individuals have identical utilities on alternatives. This is possible because the individuals have different sums of utilities on C; one unit of utility on an alternative has a weaker effect for individuals with higher utility sums on C. To address the issue of identifiability, we assume each individual's utility sum over $\mathcal{U}$ is zero in this section. This allows us to meaningfully compare the sum of utilities of two different individuals.

Definition 7. If an individual a has $\sum_{x \in \mathcal{U}} u_a(x) = 0$, then the *stubbornness* of a is $\sigma_a = \sum_{x \in C} e^{u(x)}$.

We call this quantity "stubbornness" since it quantifies how reluctant an individual is to change its probabilities on C given a unit of utility on an alternative.

Propostion 6. *In an MNL model where all individuals are equally stubborn* and *have identical utilities on alternatives, the solution to* Agreement *is* $\overline{C}$*.*

Proof. Assume utilities are in standard form, with $\sum_{x \in \mathcal{U}} u_a(x) = 0$. Let $\sigma = \sum_{x \in C} e^{u(x)}$ be each individual's stubborness and let Z be a set of alternatives. Suppose all individuals have the same utility $u(z)$ for each alternative z. The disagreement between two individuals about a single item x in C is then:

$$\left| \frac{e^{u_a(x)}}{\sigma + \sum_{z \in Z} e^{u(z)}} - \frac{e^{u_b(x)}}{\sigma + \sum_{z \in Z} e^{u(z)}} \right| = \frac{|e^{u_a(x)} - e^{u_b(x)}|}{\sigma + \sum_{z \in Z} e^{u(z)}}.$$

Notice that this strictly decreases if $\sum_{z \in Z} e^{u(z)}$ increases, so we minimize D by including all of the alternatives. $\square$

The same reasoning also allows us to solve Disagreement in this restricted MNL model.

Corollary 12. *The solution to* Disagreement *in an equal alternative utilities, equal stubbornness MNL model is* $\emptyset$.

D.3 Approximation algorithm details and extensions

D.3.1 Proof of Lemma 2

Proof. If a set Z has total exp-utility t_a to individual a, then it is placed in L at position $\lfloor \log_{1+\delta} t_a \rfloor$ in dimension a. So, if two sets Z, Z' with exp-utility totals t_a, t'_a for individual a are mapped to the same cell of L, then for all $a \in A$, $\lfloor \log_{1+\delta} t_a \rfloor = \lfloor \log_{1+\delta} t'_a \rfloor$. We can therefore bound t'_a:

$$\log_{1+\delta} t_a - 1 < \log_{1+\delta} t'_a < \log_{1+\delta} t_a + 1.$$

Exponentiating both sides with base $1 + \delta$ and simplifying yields

$$\frac{t_a}{1+\delta} < t'_a < t_a(1+\delta). \tag{D.1}$$

With this fact in hand, we proceed by induction on i. When $i = 0$, $\overline{C}_i$ is empty and the lemma holds. Now suppose that $i > 0$ and that the lemma holds for $i - 1$. Every set in $\overline{C}_i$ was made by adding (a) zero elements or (b) one element to a set in $\overline{C}_{i-1}$. We consider these two cases separately.

(a) For any set $Z \subseteq \overline{C}_i$ that is also contained in $\overline{C}_{i-1}$, we know by the inductive hypothesis that some element in L_{i-1} satisfied the inequality. Since we never overwrite cells, the lemma also holds for Z after iteration i.

(b) Now consider sets $Z' \subseteq \overline{C}_i$ that were formed by adding the new element z to a set $Z \subseteq \overline{C}_{i-1}$. In the inner for loop, we at some point encountered the cell containing the set $Y \in L_{i-1}$ satisfying the lemma for set Z by the inductive hypothesis. Let y_a be the exp-utility totals for Y and t_a for Z. Notice that the exp-utility totals of Z' are exactly $t_a + e_{az}$. Starting with the inductive hypothesis, we see that the exp-utility totals of $Y \cup \{z\}$ satisfy

$$\frac{t_a + e_{az}}{(1+\delta)^{i-1}} < y_a + e_{az} < (t_a + e_{az})(1+\delta)^{i-1}.$$

When we go to place $Y \cup \{z\}$ in a cell, it might be unoccupied, in which case we place it in L_i and the lemma holds for Z'. If it is occupied by some other set, then by applying Equation (D.1) we find that the lemma also holds for Z'. $\square$

D.3.2 Polynomial bound on runtime of Algorithm 1

The runtime of Algorithm 1 is $O((m + kn^2)(1 + \lfloor \log_{1+\delta} s \rfloor)^n)$. We can show that the second part is bounded by a polynomial in k, m, and $\frac{1}{\varepsilon}$:

$$\begin{aligned}
(1 + \lfloor \log_{1+\delta} s \rfloor)^n &\leq \left(1 + \frac{\ln s}{\ln 1 + \delta}\right)^n \\
&\leq \left(1 + (1+\delta)\frac{\ln s}{\delta}\right)^n \qquad \text{(since } \ln(1+x) \geq \tfrac{x}{1+x} \text{ for } x > -1\text{)} \\
&= \left(1 + \frac{\ln s}{\delta} + \ln s\right)^n \\
&= \left(1 + \frac{2km\binom{n}{2}\ln s}{\varepsilon} + \ln s\right)^n
\end{aligned}$$

D.3.3 Adapting Algorithm 1 for CDM with guarantees for special cases

We can adapt Algorithm 1 for the CDM model, but we only maintain the approximation error bounds under special cases of the structure of the "pulls". Still, we can use this algorithm as a principled heuristic and it tends to work well in practice, as we saw in Figure 5.2.

As a first step, we use the alternative parametrization of the model used by Seshadri et al. (2019, Eq. (1)), which has fewer parameters. In this description of the model, utilities and context effects are merged into a single utility-adjusted pull $q_a(z, x) = p_a(z, x) - u_a(x)$, with the special case $q_a(x, x) = 0$. We then have

$$\Pr(a \leftarrow x \mid C) = \frac{\exp(\sum_{w \in C} q_a(w, x))}{\sum_{y \in C} \exp(\sum_{z \in C} q_a(z, y))}. \tag{D.2}$$

Refer to Seshadri et al. (2019, Appendix C.1) for details of the equivalence between this formulation and the one we use in the main text.

Matching the notation of the proof of Theorem 13, we use the shorthand $e_{ax} = \exp(\sum_{w \in C} q_a(w, x))$.

To adapt Algorithm 1 to the CDM, we expand L_i to have nk dimensions for each individual-item pair, increasing the runtime to $O((m+kn^2)(1+\lfloor \log_{1+\delta} s \rfloor)^{nk})$. This is only practical if nk is small, but as we have seen, Agreement, Disagreement, and Promotion are all NP-hard even with $n = 2$ and $k = 2$ or 3. Each individual-item dimension stores e_{ax}, the total exp-utility of that item to that individual given that we have included some set of alternatives. When we include an additional item from $\overline{C}$, we place the new sets in L_i with updated e_{ax} values.

This only preserves the ε-additive approximation if alternatives (items in $\overline{C}$)

have zero context effects on each other; however, they may still have context effects on items in C. Formally, we need $q_a(z, z') = 0$ for all $z, z' \in \overline{C}$ and $a \in A$. Although this is a serious restriction, it leaves AGREEMENT, DISAGREEMENT, and PROMOTION NP-hard, as the CDM we used in our proofs had this form (see also Appendix D.3.5 for how to apply Algorithm 1 to PROMOTION). If this version of the algorithm is applied to a general CDM, it may experience higher error. Nonetheless, our real-data experiments show it to be a good heuristic.

For the following analysis, we assume a CDM with zero context effects between items in $\overline{C}$. We need to verify that if every item's exp-utility is approximated to within factor $(1+\beta)^{\pm 1}$, then the total disagreement of a set is approximated to within ε as we had in the MNL case. The approximation error guarantee increases to 4ε in the restricted CDM version—to recover the ε-additive approximation, we can make δ smaller by a factor of 4 (that is, we could pick $\delta = \varepsilon/(8km\binom{n}{2})$; we instead keep the old δ for simplicity in the following analysis).

Recall that Z' is the representative in L_m of the optimal set of alternatives Z^*. For compactness, we define T_a to be the denominator of Equation (D.2), with T'_a and T^*_a referring to those denominators under the choice sets $C \cup Z'$ and $C \cup Z^*$, respectively. This is where we require zero context effects between alternatives: if alternatives interact, then storing every e_{ax} in the table (from which we can compute T_a) is not enough to determine updated choice probabilities when we add a new alternative.

The difference in the analysis begins when we bound $\Pr(a \leftarrow x \mid C \cup Z')$ on both sides using the fact that each exp-utility sum is approximated within a $1+\beta$

factor (so the probability denomiators T_a are also approximated within this factor):

$$\begin{aligned}
\frac{\frac{e^*_{ax}}{1+\beta}}{T^*_a(1+\beta)} &= \frac{1}{(1+\beta)^2}\frac{e^*_{ax}}{T^*_a} \\
&< \frac{e'_{ax}}{T'_a} = \Pr(a \leftarrow x \mid C \cup Z') \\
&< \frac{e^*_{ax}(1+\beta)}{\frac{T^*_a}{1+\beta}} = (1+\beta)^2\frac{e^*_{ax}}{T^*_a}.
\end{aligned}$$

Based on the lower bound, the difference between $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ could be as large as

$$\frac{e^*_{ax}}{T^*_a} - \frac{1}{(1+\beta)^2}\frac{e^*_{ax}}{T^*_a} \le 1 - \frac{1}{(1+\beta)^2}.$$

Now considering the upper bound, the difference between $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ could be as large as

$$(1+\beta)^2\frac{e^*_{ax}}{T^*_a} - \frac{e^*_{ax}}{T^*_a} \le (1+\beta)^2 - 1.$$

Therefore, $|\Pr(a \leftarrow x \mid C \cup Z') - \Pr(b \leftarrow x \mid C \cup Z')|$ can only exceed $|\Pr(a \leftarrow x \mid C \cup Z^*) - \Pr(b \leftarrow x \mid C \cup Z^*)|$ by at most $1 - \frac{1}{(1+\beta)^2} + (1+\beta)^2 - 1 = (1+\beta)^2 - \frac{1}{(1+\beta)^2}$. This is at most 4β:

$$\begin{aligned}
4\beta - (1+\beta)^2 + \frac{1}{(1+\beta)^2} &= \frac{\beta^2(2-\beta^2)}{(1+\beta)^2} \\
&> 0. && (\text{for } 0 < \beta < \sqrt{2})
\end{aligned}$$

So $D(Z')$ and $D(Z^*)$ are within $4\beta\binom{n}{2}k = 4\varepsilon$.

D.3.4 Adapting Algorithm 1 for NL with full guarantees

We can also adapt Algorithm 1 for the NL model, and unlike the CDM, the ε-additive approximation holds in all parameter regimes. Recall that the NL tree has two types of leaves: choice set items and alternative items. Let P_a be the set of internal nodes of individual a's tree that have at least one alternative item as a child and let $p = \max_{a \in A} |P_a|$. If we know the total exp-utility that alternatives contribute as children of each $v \in P_a$, then we can compute a's choice probabilites over items in C in polynomial time.

With this in mind, we modify Algorithm 1 by having dimensions in L for each individual for each of their nodes in P_a. This results in $\leq np$ dimensions. The algorithm then keeps track of the exp-utility sums from alternatives under each node in P_a for each individual. The exponent in the runtime increases to (at most) np, but this remains tractable for some hard instances, such as those in our hardness proofs. In some cases, we can dramatically improve the runtime of the algorithm: if the subtree under an internal node contains only alternatives as leaves in an individuals's tree, then we only need one dimension L for that individual's entire subtree, and it has only two cells: one for sets that contain at least one alternative in that subtree, and one for sets that do not. The only factor that affects the choice probabilities of items in C is whether that subtree is "active" and its root can be chosen.

We now show how the error from exp-utility sums of alternatives propagates to choice probabilities. In the NL model, $\Pr(a \leftarrow x \mid C)$ is the product of probabilities that a chooses each ancestor of x as a descends down its tree. Let $v_1, \dots, v_\ell$ be the nodes in a's tree along the path from the root to x. For compactness, we use $\Pr(x, Z)$ instead of $\Pr(a \leftarrow x \mid C \cup Z)$ in the following analysis.

Pick $\delta \leq ([\varepsilon/(2k\binom{n}{2}) + 1]^{1/\ell} - 1)/m$ and recall that $\beta = 2m\delta$. We can use the same analysis as in the proof of Theorem 13 to find that for any set $Z^* \subseteq \overline{C}$, there exists some $Z' \in L$ such that

$$\begin{aligned}
\Pr(x, Z^*) &= \Pr(v_1, Z^*) \cdot \dots \cdot \Pr(v_x, Z^*) \\
&< \left(\Pr(v_1, Z') + \frac{\beta}{2} \right) \cdot \dots \cdot \left(\Pr(v_x, Z') + \frac{\beta}{2} \right) \\
&\leq \Pr(x, Z') + \left(1 + \frac{\beta}{2} \right)^{\ell} - 1 \\
&\leq \Pr(x, Z') + \frac{\varepsilon}{2k\binom{n}{2}}.
\end{aligned}$$

Now for the lower bound, pick $\delta \leq (1 - [1 - \varepsilon/(2k\binom{n}{2})]^{1/\ell})/m$. Again from the proof of Theorem 13:

$$\begin{aligned}
\Pr(x, Z^*) &= \Pr(v_1, Z^*) \cdot \dots \cdot \Pr(v_x, Z^*) \\
&> \left(\Pr(v_1, Z') - \frac{\beta}{2} \right) \cdot \dots \cdot \left(\Pr(v_x, Z') - \frac{\beta}{2} \right) \\
&\geq \Pr(x, Z') + \left(1 - \frac{\beta}{2} \right)^{\ell} - 1 \\
&\geq \Pr(x, Z') - \frac{\varepsilon}{2k\binom{n}{2}}.
\end{aligned}$$

Let h be the maximum height of any indivdual's NL tree (so $\ell \leq h$). Then, by picking $\delta = \min\{[\varepsilon/(2k\binom{n}{2}) + 1]^{1/h} - 1, 1 - [1 - \varepsilon/(2k\binom{n}{2})]^{1/h}\}/m$, we find that $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ differ by less than $\varepsilon/(k\binom{n}{2})$ for all $x \in C$ and $a \in A$, meaning that the total disagreement between a and b cannot differ by more than ε as before.

Unfortunately, this means we need to make δ exponentially (in h) smaller in the NL model. Put another way, our error bound gets exponentially worse as h increases if we keep δ constant. However, we have seen that there are NP-hard families of NL instances in which h is a small constant (e.g., $h = 2$ in our hardness proof), so once again this algorithm is an exponential improvement over brute force. Moreover, the error bound here is often far from tight, since we use the very loose bounds $\Pr(v_i, Z') \leq 1$ in the analysis. This means the algorithm will tend to outperform the worst-case guarantee by a significant margin.

D.3.5 Adapting Algorithm 1 for Promotion

CDM Promotion with special case guarantees

Algorithm 1 can be applied to Promotion in the (restricted) CDM model with only a small modification to the CDM version described in Appendix D.3.3: at the end of the algorithm, we return the set that results in the maximum number of individuals having x^* as an ε-favorite item. Additionally, we choose $\delta = \varepsilon/(10m)$ (we don't need the factors $\binom{n}{2}$ or k since we aren't optimizing $D(Z)$).

Following the analysis in Appendix D.3.3 (with $\beta = 2m\delta = \varepsilon/5$), we find that $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ differ by at most $\max\{1 - \frac{1}{(1+\varepsilon/5)^2}, (1 + \varepsilon/5)^2 - 1\}$ for all x. On the interval $[0, 1]$, this is bounded by $\varepsilon/2$. Thus, if x^* is the favorite item for a given the optimal choice set $C \cup Z^*$, then it must be an ε-favorite of individual a given $C \cup Z'$ (as always, Z' is the representative of Z^* in L_m). This is because when we go from $C \cup Z^*$ to $C \cup Z'$, the choice probability of x^* can shrink by at most $\varepsilon/2$ and the choice probability for any other item can grow by at most $\varepsilon/2$. Thus, including Z' makes at least as many individuals have x^* as an *ε-favorite* item as including Z^* makes have x^* as a *favorite* item.

This is exactly what it means for Algorithm 1 to ε-approximate Promotion in the CDM (when items in $\overline{C}$ do not exert context effects on each other). Moreover, not having to compute $D(Z)$ makes the runtime of Algorithm 1 $O(m(1 + \lfloor \log_{1+\delta} s \rfloor)^{nk})$ when applied to Promotion in the CDM. In the general CDM, this algorithm is only a heuristic.

NL Promotion with full guarantees

A very similar idea allows us to apply the NL version of Algorithm 1 from Appendix D.3.4 to Promotion and retain an approximation guarantee. As before, use the NL version and return the set that results in the maximum number of individuals having x^* as an ε-favorite item. However, we instead use $\delta = \min\{(\varepsilon/4+1)^{1/h}-1, 1-(1-\varepsilon/4)^{1/h}\}/m$, which by the analysis in Appendix D.3.4 results in $\Pr(a \leftarrow x \mid C \cup Z^*)$ and $\Pr(a \leftarrow x \mid C \cup Z')$ differing by at most $\varepsilon/2$. As in the CDM case, this guarantees that if x^* is the favorite item for a given the optimal choice set $C \cup Z^*$, then it must be an ε-favorite of a given $C \cup Z'$. Therefore this version of Algorithm 1 ε-approximates Promotion in the NL model with runtime $O(m(1+\lfloor \log_{1+\delta} s \rfloor)^{np})$.

D.4 Mixed-integer bilinear programs for MNL agreement and disagreement optimization

D.4.1 Agreement

Let x_i be a decision variable indicating whether we add in the ith item in $\overline{C}$. Let $e_{ya} = e^{u_a(y)}$ and $e_{Ca} = \sum_{y \in C} e_{ya}$. We can write Agreement as the following 0-1 optimization problem.

$$\begin{aligned}
\min_{x} \quad & \sum_{a,b \in A} \sum_{y \in C} \left| \frac{e_{ya}}{e_{Ca} + \sum_{i \in \overline{C}} x_i e_{ia}} - \frac{e_{yb}}{e_{Cb} + \sum_{i \in \overline{C}} x_i e_{ib}} \right| \\
\text{s.t.} \quad & x_i \in \{0, 1\}
\end{aligned}$$

We can rewrite this with no absolute values by introducing new variables δ_{yab} that represent the absolute disagreement about item y between individuals a and b. We then use the standard trick for minimizing an absolute value in linear programs:

$$
\begin{aligned}
\min_{x} \quad & \sum_{a,b\in A}\sum_{y\in C}\delta_{yab} \\
\text{s.t.} \quad & \frac{e_{ya}}{e_{Ca}+\sum_{i\in\overline{C}}x_i e_{ia}} - \frac{e_{yb}}{e_{Cb}+\sum_{i\in\overline{C}}x_i e_{ib}} \le \delta_{yab} \\
& \qquad \forall y\in C, \{a,b\}\subset A \qquad\qquad\qquad , \\
& \frac{e_{yb}}{e_{Cb}+\sum_{i\in\overline{C}}x_i e_{ib}} - \frac{e_{ya}}{e_{Ca}+\sum_{i\in\overline{C}}x_i e_{ia}} \le \delta_{yab} \\
& \qquad \forall y\in C, \{a,b\}\subset A \qquad\qquad\qquad , \\
& x_i\in\{0,1\} \quad \forall i\in\overline{C}, \\
& \delta_{yab}\in\mathbb{R} \qquad \forall y\in C, \{a,b\}\subset A
\end{aligned}
$$

To get rid of the fractions, we introduce the new variables $z_a = \frac{1}{e_{Ca}+\sum_i x_i e_{ia}}$ for each individual a and add corresponding constraints enforcing the definition of z_a:

$$
\begin{aligned}
\min_{x} \quad & \sum_{a,b\in A}\sum_{y\in C}\delta_{yab} \\
\text{s.t.} \quad & \\
z_a e_{ya} - z_b e_{yb} &\le \delta_{yab} && \forall y\in C, \{a,b\}\subset A, \\
z_b e_{yb} - z_a e_{ya} &\le \delta_{yab} && \forall y\in C, \{a,b\}\subset A, \\
z_a e_{Ca} + z_a \sum_{i\in\overline{C}} x_i e_{ia} &= 1 && \forall a\in A, \\
x_i &\in \{0,1\} && \forall i\in\overline{C}, \\
\delta_{yab} &\in \mathbb{R} && \forall y\in C, \{a,b\}\subset A, \\
z_a &\in \mathbb{R} && \forall a\in A
\end{aligned}
$$

This is a mixed-integer bilinear program (MIBLP) with m binary variables, $n+k\binom{n}{2}$ real variables, $2k\binom{n}{2}$ linear constraints, and n bilinear constraints. We plug this form of the problem directly into a branch-and-bound solver (we use Gurobi).

D.4.2 Disagreement

A similar technique works for Disagreement, but maximizing an absolute value is slightly trickier than minimizing. In addition to the variables δ_{yab} that we used before, we also add new binary variables g_{yab} indicating whether each difference in choice probabilities is positive or negative. With these new variables (and following the same steps as above), Disagreement can be written as the following MIBLP:

$$
\begin{aligned}
\max_{x} \quad & \sum_{a,b \in A} \sum_{y \in C} \delta_{yab} \\
\text{s.t.} \quad & \\
& z_a e_{ya} - z_b e_{yb} \le \delta_{yab} && \forall y \in C, \{a,b\} \subset A, \\
& z_b e_{yb} - z_a e_{ya} \le \delta_{yab} && \forall y \in C, \{a,b\} \subset A, \\
& 2g_{yab} + z_a e_{ya} - z_b e_{yb} \ge \delta_{yab} && \forall y \in C, \{a,b\} \subset A, \\
& 2(1 - g_{yab}) + z_b e_{yb} - z_a e_{ya} \ge \delta_{yab} && \forall y \in C, \{a,b\} \subset A, \\
& z_a e_{Ca} + z_a \sum_{i \in \overline{C}} x_i e_{ia} = 1 && \forall a \in A, \\
& x_i \in \{0,1\} && \forall i \in \overline{C}, \\
& g_{yab} \in \{0,1\} && \forall y \in C, \{a,b\} \subset A, \\
& \delta_{yab} \in \mathbb{R} && \forall y \in C, \{a,b\} \subset A, \\
& z_a \in \mathbb{R} && \forall a \in A
\end{aligned}
$$

D.5 Additional experiment details

D.5.1 Simple example of poor performance for Greedy

As we saw in experimental data, Greedy can perform poorly even in small instances of Agreement. Below we provide an MNL instance with $n = m = k = 2$ for which the error of the greedy solution is approximately 1. With only two individuals, $0 \leq D(Z) \leq 2$, so an error of 1 is very large.

In the bad instance for greedy, $A = \{a, b\}$, $C = \{x, y\}$, $\overline{C} = \{p, q\}$, and the utilities are as follows.

$$
\begin{aligned}
u_a(x) &= 8 & \quad u_b(x) &= 8 \\
u_a(y) &= 2 & \quad u_b(y) &= 8 \\
u_a(p) &= 10 & \quad u_b(p) &= 0 \\
u_a(q) &= 0 & \quad u_b(q) &= 15
\end{aligned}
$$

In this instance of Agreement, the greedy solution is $D(\emptyset) \approx 0.9951$ (including either p or q alone increases disagreement), while the optimal solution is $D(\{p, q\}) \approx 0.0009$.

D.5.2 All-pairs agreement results for MIBLP

Figure D.3 shows the comparison in performance between Algorithm 1 and the MIBLP approach for the all-pairs Agreement and Disagreement experiment. The methods perform nearly identically on both Allstate and Yoochoose. The MIBLP approach performs marginally better in some cases of Yoochoose Agreement. As noted in the paper, the MIBLP heuristic is considerably faster

(12x and 240x on YOOCHOOSE and ALLSTATE, respectively), but provides no a priori performance guarantee and cannot be applied to CDM or NL. Nonetheless, we can see that it performs very competitively and would be a good approach to use in practice for MNL AGREEMENT and DISAGREEMENT.

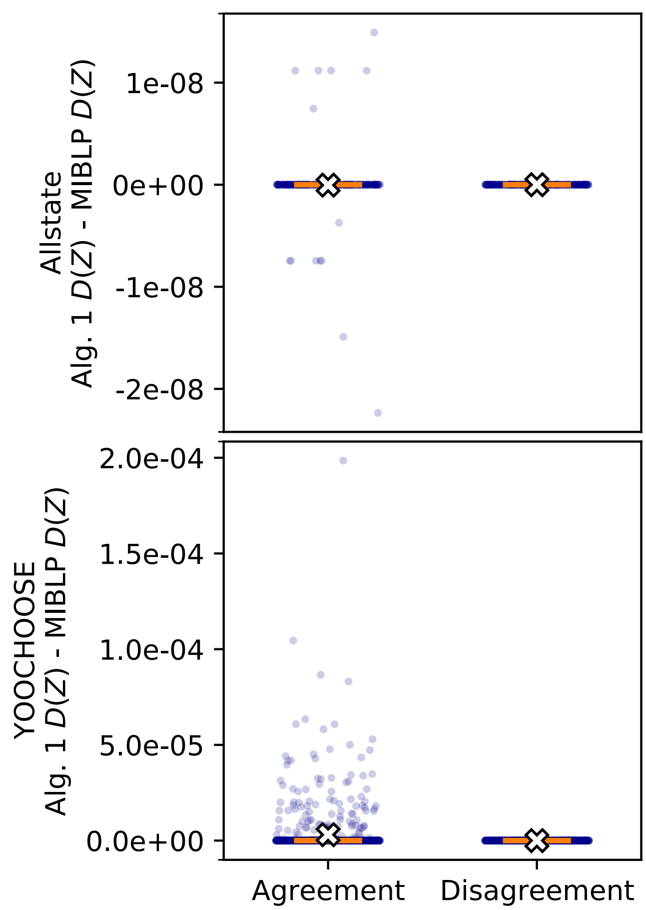


Figure D.3: MIBLP vs. Algorithm 1 performance box plots when applied to all 2-item choice sets in ALLSTATE and YOOCHOOSE under MNL. Each point is the difference in $D(Z)$ when MIBLP and Algorithm 1 are run on a choice set, and Xs mark means.

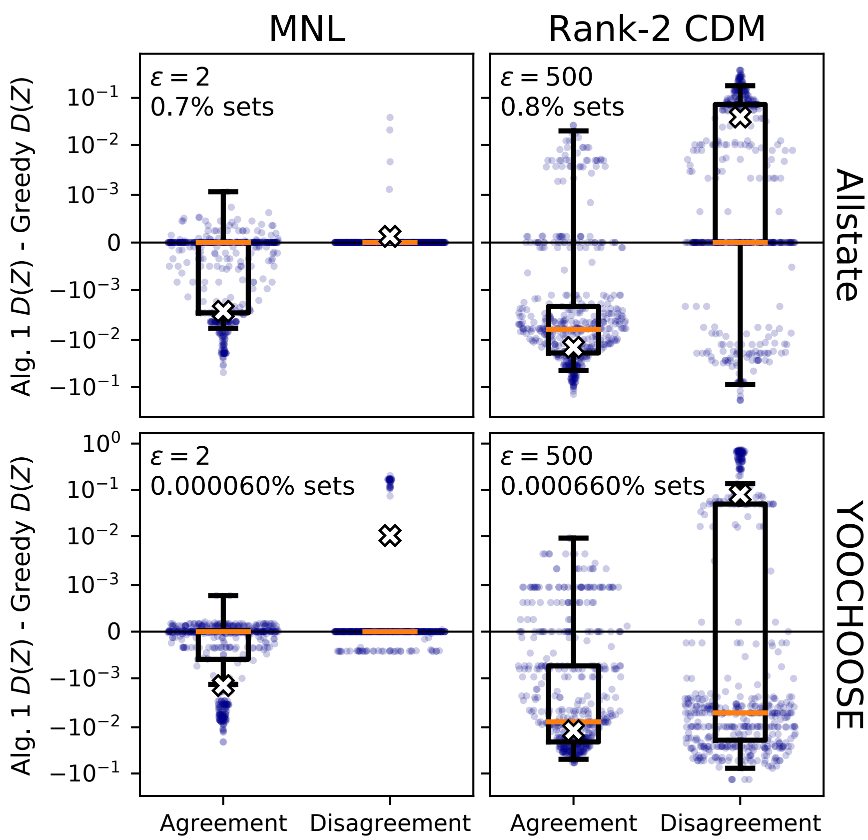


Figure D.4: Results of the agreement experiment with 500 choice sets sampled uniformly from each dataset. Compare with Figure 5.2 in the main text. Again, Algorithm 1 has better mean performance in all cases. The larger values of ε result in slightly worse performance on the margins than in Figure 5.2, but also fewer sets computed.

D.5.3 Choice sets sampled from data

We repeated the all-pairs agreement experiment with 500 choice sets of size up to 5 sampled uniformly from each dataset, allowing us to evaluate the performance of Algorithm 1 on realistic choice sets. We limited the size of sampled choice sets since the CDM version of Algorithm 1 scales poorly with $|C|$ (see Appendix D.3.3). For this version of the experiment, we fixed larger values of ε (2 for MNL, 500 for CDM) to handle larger choice sets and to keep running time down. Again, Algorithm 1 has better mean performance in every case (Figure D.4), showing that it performs well on real choice sets.

D.6 A note on ethical considerations

Influencing the preferences of decision-makers has the potential for malicious applications, so it is important to address the ethical context of this work.

Any problem with positive social applications (e.g., Agreement: encouraging consensus, Promotion: promoting environmentally-friendly transportation options, Disagreement: increasing diversity of opinions) has the potential to be used for ill. This should not prevent us from seeking methods to acheive these positive ends, but we should certainly be cognizant of the possibility of unintended applications. In a different vein, understanding when a group is susceptible to undesired interventions (or detecting such interventions) makes problems like Disagreement worth studying from an adversarial perspective. Along these lines, our hardness results are encouraging since optimal malicious interventions are difficult.

Finally, we note that all of the theoretical problems we study presuppose access to choice data from which preferences can be learned and the ability to influence choice sets. Any entity which has both of these (such as an online retailer, a government deciding transportation policy, etc.) already has significant power to influence choosers. If such an entity had malicious intent, then near-optimal Disagreement solutions would be the least of our concerns.

To summarize, these problems are worth studying because of (1) their purely theoretical value in furthering the field of discrete choice, (2) their potential for positive applications, (3) insight into the potential for harmful manipulation by an adversary, and (4) the minimal additional risk from undesired use of our methods.

APPENDIX E

TECHNICAL DETAILS FOR CHAPTER 6: BALLOT LENGTH IN INSTANT RUNOFF VOTING

> [...] chaos is found in greatest abundance wherever order is being sought. It always defeats order, because it is better organized.
>
> Terry Pratchett, *Interesting Times*, 1994

E.1 The IRV Algorithm

For each voter j, let $\pi_j^{(\ell)}$ be their ballot after step ℓ of IRV, with $\pi_j^{(0)} = \pi_j$. Let $\pi_j^{(\ell)}(h)$ denote the candidate ranked in position h by this ballot, with lower indices h corresponding to more preferred positions. A ballot $\pi_j^{(\ell)}$ at step ℓ is said to be a *vote* for candidate i if $\pi_j^{(\ell)}(1) = i$. See Algorithm 3 for a formal definition of the IRV algorithm for determining a winner given a profile $\{\pi_1, \ldots, \pi_n\}$.

Algorithm 3 Instant runoff voting.

Input: candidates $1, \ldots, k$, partial rankings π_j over the candidates for each voter j
$\pi_j^{(0)} \leftarrow \pi_j, \forall j$
$C = \{i \in \{1, \ldots, k\} \mid \exists j : i \in \pi_j\}$ {Non-eliminated candidates}
$\ell \leftarrow 0$
while $|C| > 1$ **do**
  $B = \{j \mid |\pi_j^{(\ell)}| > 0\}$ {Non-exhausted ballots}
  $i^* \leftarrow \arg\min_i \sum_{j \in B} \mathbf{1}\left[\pi_j^{(\ell)}(1) = i\right]$ {Break ties as desired}
  $\ell \leftarrow \ell + 1$
  $C \leftarrow C \setminus \{i^*\}$
  $\pi_j^{(\ell)} \leftarrow \pi_j^{(\ell-1)} \setminus \{i^*\}, \forall j$
return the winner, the last remaining candidate in C

E.2 LP for full-ballot constructions

As an example of constructing elimination orders that result in $k-1$ truncation winners, consider $k=5$. Label the candidates $1,2,3,4,5$ in order of their full-ballot IRV elimination. If we are to have 4 truncation winners, they must be $2,3,4,5$, and they must win in that order at $h=1,2,3,4$. By construction, the elimination order at $h=4$ is $1,2,3,4,5$. However, at $h=3$, the elimination order must be $1,2,3,5,4$, since 4 wins. At $h=2$, the elimination order can be $1,2,4,5,3$ or $1,2,5,4,3$. At $h=1$, it can be one of six options, corresponding to the permutations of $3,4,5$: $1,\{3,4,5\},2$. There are thus 12 possible elimination orders across ballot lengths we to consider that could result in 4 truncation winners.

Formally, let x_π denote the number of ballots with the ranking $\pi \in S_k$ over the candidates. Fix an elimination order over all ballot lengths and let the $(k-1) \times (k-1)$ matrix E store the elimination orders over ballot lengths $1,\dots k-1$. That is, E_{hi} is the index of the candidate eliminated at round i with ballot length h. Let $r(h,i)$ denote the set of remaining candidates at round i with ballot length h that are not eliminated at round i. Let $b(h,i,j) \subset S_k$ denote the set of ballot types that would be assigned to candidate j at round i with ballot length h. Note that $b(h,i,j)$ and $r(h,i)$ depend on the fixed elimination order. Let $C \geq 1$ be the elimination gap (the smallest number of votes by which eliminations are decided).

The linear program for finding $k-1$ truncation winner constructions, minimizing the number of ballots, is then:

$$\begin{aligned}
\text{minimize} \quad & \sum_{\pi \in S_k} x_\pi \\
\text{subject to} \quad & \qquad\qquad x_\pi \geq 0, \\
& C + \underbrace{\sum_{\pi \in b(h,i,E_{hi})} x_\pi}_{\#\text{ votes for } E_{hi}} \leq \underbrace{\sum_{\pi \in b(h,i,j)} x_\pi}_{\#\text{ votes for } j} .
\end{aligned}$$

We have a constraint of the first type for all $\pi \in S_k$ and constraints of the second type for $h = 1, \ldots, k-1$; $i = 1, \ldots, k-1$; and $j \in r(h,i)$. The second type of constraint encodes each elimination that takes place during IRV at each ballot length, ensuring that the eliminated candidate E_{hi} has fewer votes in that round it is eliminated than each remaining candidate $j \in r(h,i)$. We chose the objective function to find profiles with few ballots. All constructions generated by running the LP are stored in the code and data repository, which contains instructions for viewing them.

E.3 Proofs

Proof of Theorem 14. Suppose we have a consequential-tie-free profile with $k-1$ truncation winners. Then $k-1$ of the candidates each have a unique ballot length in $1, \ldots, k-1$ at which they win. Label these candidates $1, \ldots, k-1$ according to their winning ballot length. The candidate not in those $k-1$ winners, call them candidate k, must have at least 1 fewer first-place vote than any other candidate (otherwise one of the winners could be eliminated first after a tie-break, preventing them from winning at their ballot length). Now consider the winner under ballot length 1, namely candidate 1. In order for candidate 1 to be the unambiguous plurality winner, they must have at least one more vote than every other candidate.

Next, consider who is eliminated second. It has to be candidate 1: if any other candidate $i \neq 1$ can be eliminated second, then they will not be able to win at their designated ballot length $h > 1$. In order for candidate 1 to be eliminated second, they must be in unambiguous last place after candidate k's ballots are redistributed. This means at least 2 of those ballots need to go to each of candidates $2, \ldots, k-1$ (who are currently trailing candidate 1 by 1 vote). Finally, if $k > 3$, the candidate who wins at ballot length 2 (candidate 2) must be unambiguously in the lead over $3, \ldots, k-1$ after redistributing k's ballots. Either they had more initial ballots than $3, \ldots, k-1$ (but this would require at least one more ballot from candidate k to help those lower candidates overtake 1) or they got a single extra ballot from candidate k. To summarize the constraints:

1. candidates $1, \ldots, k-1$ have at least one more first-place vote than candidate k,
2. candidate 1 has at least one more first-place vote than any other candidate, and
3. candidate k has enough first-place votes to redistribute at least two each to $2, \ldots, k-1$ (plus at least one more if $k > 3$).

For $k > 3$, the total number of ballots ranking k first is thus at least $2(k-2)+1$, by constraint 3. Each of candidates $2, \ldots, k-1$ must then have at least $2(k-2)+2$ first-place ballots by constraint 1. Finally, candidate 1 must have at least $2(k-2)+3$ first-place ballots by constraint 2. The minimum number of ballots is thus $2(k-2)+1+(k-2)(2(k-2)+2)+2(k-2)+3 = 2k^2-2k$.

For $k = 3$, constraint 3 only requires $2(k-2) = 2$ first-place votes for candidate 3. Candidates 2 and 1 must then have 3 and 4 first-place votes by constraints 1 and 2, for a total of $2 + 3 + 4 = 9 = k^2$.

$\square$

Proof of Theorem 15. First, if we have a sequence with $w_h \leq h$ for some h, then this means the winner at ballot length h is eliminated hth or sooner under ballot lengths $\geq h$. This is impossible, since they would be eliminated before they win at length h.

Now suppose we have some valid sequence $w_1, \ldots, w_{k-1}$ such that $w_h \in \{h + 1, \ldots, k\}$ for $h \in [k-1]$. First, assign $2(k-2)+1$ ballots to each candidate listing them first. Give candidate w_1 an extra 2 ballots and the other candidates (except candidate 1) an extra 1 ballot each. This is a total of $2k(k-2)+k+(k-2)+2 = 2k^2 - 2k$ ballots. We now fill out the ballots initially assigned to each candidate, using S_i to denote the set of ballots ranking i first.

Except for $i = 1$, all ballots in S_i rank candidates $1, \ldots, i-1$ in positions $2, \ldots, i$. For all i, two ballots in S_i rank ℓ in position $i+1$ for each $\ell = i+2, \ldots, k$ except w_i. If $w_i \neq i+1$, one ballot in S_i ranks w_i in position $i+1$. Finally, one extra ballot in S_i ranks w_{i+1} in position $i+1$. This requires at most $2(k-2)+1$ ballots, which is covered by the $\geq 2(k-2)+1$ ballots in S_i. All ballots in S_i then terminate after their last specified entry. Notice that when i is eliminated, the effect of their redistributed votes is to put the new winner w_{i+1} in the lead and the new loser $i+1$ in last, assuming the last winner w_i was in the lead by a single vote after i is eliminated.

We now show that if ballots are truncated to length $h < k$, then candidate w_h wins under IRV. First, if we truncate ballots to length 1, candidate w_1 wins: they have 2 more first place votes than candidate 1 and 1 more than every other candidate. Thus, candidate 1 will be eliminated (with no redistribution due to the length-1 ballots), followed by the others in some order based on tie-breaking, making candidate 1 win.

Now suppose we truncate to length h ($2 \leq h < k$). Candidate 1 is eliminated first and their second place votes cause candidates $3, \ldots, k$ to overtake candidate 2, with candidate w_2 taking the lead by 1 vote. If $h = 2$, then all remaining ballots only have one candidate listed (since the second place votes for ballots assigned to candidate $\ell > 1$ are all for candidate 1, who is eliminated). Thus candidate w_2 wins after eliminating candidate 2 and then $3, \ldots, k \setminus w_2$ in some order. For $h > 2$, we'll prove inductively that for $2 \leq \ell < h$, the ℓth candidate eliminated is candidate ℓ, which causes candidate $w_{\ell+1}$ to take the lead by one vote and candidate $\ell + 1$ drop to last place by one vote.

<u>Base case</u> ($\ell = 2$): As we saw, the 2nd candidate eliminated is candidate 2. Since $h > 2$, ballots assigned to candidate 2 are not yet exhausted: two go to each of candidates $4, \ldots, k$ (except w_2); w_2 gets one if $w_2 \neq 3$ and zero otherwise; and w_3 gets one extra ballot. Since candidate w_2 was only in the lead by one vote, this causes the new leader to be candidate w_3 and candidate 3 to drop to last place, as claimed.

<u>Inductive case</u> ($2 < \ell < h$): by inductive hypothesis, candidates $2, \ldots, \ell - 1$ have been eliminated (plus candidate 1, the first to go), candidate w_ℓ is currently in the lead, and candidate ℓ is in last place. Thus, candidate ℓ is the ℓth to be eliminated. By construction, the candidates ranked in positions $2, \ldots, \ell$ on the ballots initially assigned to ℓ (namely, candidates $1, \ldots, \ell - 1$) have been eliminated. Additionally, all ballots that were redistributed to ℓ are now exhausted. Since $\ell < h$, there are still remaining places on the truncated ballot. Ballots currently assigned to

ℓ are distributed as follows: two go to each of $\ell+1,\ldots,k$ (except w_ℓ); w_ℓ gets one if $w_\ell \neq \ell+1$ and zero otherwise; and $w_{\ell+1}$ gets one extra ballot. This causes candidate $w_{\ell+1}$ to take the lead by one vote and candidate ℓ to drop to last place behind $\ell+2,\ldots,k-1$, as claimed.

Once candidate w_h is in the lead, candidates $1,\ldots,h-1$ have been eliminated, and candidate h is in last place, all the ballots only list the candidate to which they are currently assigned (since the candidates ranked up to position h on their ballots have been eliminated). Thus, h will be eliminated, followed by $h+1,\ldots,k$ (except w_h) in some order, making the winner candidate w_h, as desired. □

Proof of Theorem 18. Call candidates $1,\ldots,\kappa$ the *winners*. Each winner $i>1$ has $i-1$ *filler* candidates $f^i_1,\ldots,f^i_{i-1}$ associated with it. The single-peaked axis has winners in the order $1,\ldots,\kappa$, with i's fillers between i and $i-1$. That is, the full axis is $1, f^2_1, 2, f^3_1, f^3_2, 3, f^4_1, f^4_2, f^4_3, 4, \ldots, f^\kappa_{\kappa-1}, \kappa$. We will fill out ballots so that i wins at ballot length i, while maintaining single-peakedness.

Every winner has $\kappa+1$ ballots listing them first and winner 1 has an additional single ballot. These ballots then terminate. Each candidate's first filler f^i_1 has i ballots that list candidates $f^i_1,\ldots,f^i_{i-1}, i$ in positions $1,\ldots,i$ and then terminate. All other fillers have zero ballots listing them first.

Consider what happens at ballot length $h\leq\kappa$. If $h=1$, candidate 1 wins by one vote. For $1<h\leq\kappa$, all fillers with zero ballots are eliminated first in some order. Then, the first fillers are eliminated in the order $f^1_1, f^2_1,\ldots,f^\kappa_1$. Only fillers f^i_1 with $i\leq h$ are able to reallocate votes, since ballots for listing f^j_1 $(j>h)$ first are exhausted after f^j_1's elimination. The first-place vote counts after all fillers are eliminated are thus $\kappa+2$ for winner 1, $\kappa+1+i$ for winners $2,\ldots,h$ and $\kappa+1$ for winners $h+1,\ldots,\kappa$. With no more reallocations taking place, candidate h wins. For $h>\kappa$, candidate κ still wins. This construction therefore results in κ distinct truncation winners.

The total number of candidates in this construction is $\kappa + \sum_{i=2}^{\kappa}(i-1) = \kappa(\kappa+1)/2$. The total number of voters is $\kappa(\kappa+1)+1+\sum_{i=2}^{\kappa} i = 3\kappa(\kappa+1)/2$, as claimed. □

Proof of Theorem 19. Let $x_1 > x_2 > \cdots > x_{k-1} > x_k$ be the first place vote counts sorted in strictly descending order and index candidates in this order. Note that the inequalities must be strict so that eliminations at $h=1$ have no ties. As in the proof of Theorem 14, candidate 1 must be overtaken by candidates $2,\ldots,k-1$ when candidate k redistributes votes ($h \geq 2$). In order to make candidate 2 overtake candidate 1 after k is eliminated, k must redistribute at least two ballots to candidate 2. Similarly, candidate k must redistribute at least i ballots to each candidate $i = 2,\ldots,k-1$ for them to overtake candidate i. This requires at least $\sum_{i=2}^{k-1} i = T_{k-1}-1$ ballots listing k first, where $T_k = k(k+1)/2$ is the kth triangular number.

Candidate $k-1$ thus needs at least $T_{k-1}-1+1$ ballots listing them first since $x_{k-1} > x_k$. Similarly, candidate i needs at least $T_{k-1}-1+k-i$ ballots listing them first. Adding up these lower bounds yields the desired lower bound:

$$\begin{aligned}\sum_{i=1}^{k}(T_{k-1}-1+k-i) &= k(T_{k-1}-1)+\sum_{i=1}^{k}(k-i)\\ &= k(T_{k-1}-1)+T_{k-1}\\ &= (k+1)(T_{k-1})-k\\ &= (k+1)(k-1)k/2-k\\ &= (k^3-3k)/2.\end{aligned}$$

□

Proof of Theorem 20. The argument is almost the same as in the proof of Theorem 19, except that when candidate 1 is overtaken by candidates $2,\ldots,k-1$, the overtaking candidates cannot be tied afterwards. As before, candidate k needs to distribute at least $k-1$ ballots to candidate $k-1$ to make them overtake candidate 1. But now, they cannot merely redistribute $k-2$ to candidate $k-2$, since this could cause a tie with candidate $k-1$. In order to make all of $2,\ldots,k-1$ overtake candidate 1 and not emerge in a tie, the lowest possible totals $2,\ldots,k-1$ could have after reallocation are $x_1+1, x_1+2,\ldots,x_1+k-2$, where x_1 is the first-round vote total of candidate 1. Thus, the number of votes candidate k must reallocate is at least $\sum_{i=1}^{k-2}(x_1+i)-\sum_{i=1}^{k-2}(x_1-i)$, where the second sum is an upper bound on the number of votes candidates $2,\ldots,k-1$ have in round 1, given that they are all behind candidate 1 and not tied. This allows us to calculate the minimum number of ballots listing k first:

$$\begin{aligned}\sum_{i=1}^{k-2}(x_1+i)-\sum_{i=1}^{k-2}(x_1-i) &= 2\sum_{i=1}^{k-2}i\\ &= (k-2)(k-1)\end{aligned}$$

Candidate $k-1$ thus needs at least $(k-2)(k-1)+1$ ballots listing them first since $x_{k-1} > x_k$. Similarly, candidate i needs at least $(k-2)(k-1)+k-i$ ballots listing them first. Adding up these lower bounds yields the desired lower bound:

$$
\begin{aligned}
&\sum_{i=1}^{k}((k-2)(k-1)+k-i) \\
&= k(k-2)(k-1) + \sum_{i=1}^{k}(k-i) \\
&= k(k-2)(k-1) + k(k-1)/2 \\
&= (2k^3 - 5k^2 + 3k)/2.
\end{aligned}
$$

□

Proof of Theorem 21. The construction follows the same idea as in Theorem 15, but we no longer have the luxury of maintaining the tie for second place among all candidates who are not about to win or about to be eliminated. Instead, we will maintain gaps of a single vote between candidates, as in our lower bound proof. However, the order of candidates matters. Given a winner sequence $w_1, \ldots, w_{k-1}$, define its f-sequence as follows. Let $f_1, \ldots, f_\ell$ be the $\ell \leq k-1$ distinct truncation winners in the sequence $w_1, \ldots, w_{k-1}$ ordered by their first appearance in this sequence. Fill the remainder of the sequence $f_{\ell+1}, \ldots, f_k$ in reverse order of full-ballot elimination (i.e., $f_k = 1$), skipping candidates already in $f_1, \ldots, f_\ell$. For example, the w-sequence $4, 3, 4, 5$ for $k = 5$ would result in the f-sequence $4, 3, 5, 2, 1$ (recall that candidates are labeled in order of their full-ballot IRV elimination). Assign ballots to each candidate so that their first place vote counts result in the f-sequence, with candidate candidate f_j receiving $(k-2)(k-1)+k-j$ ballots listing them first. Call the first part of the f-sequence the *winner prefix* and the second part the *loser suffix*. We will maintain the following invariant: before step $\ell \leq h$ of IRV, the order of the remaining candidates $\ell, \ell+1, \ldots, k$ by vote count is the f-sequence of $w_\ell, w_{\ell+1}, \ldots, w_{k-1}$.

As before, let S_i denote the set of ballots listing i first. Except for $i = 1$, all ballots in S_i rank candidates $1, \ldots, i-1$ in positions $2, \ldots, i$. Next, we will fill in position $i+1$ for each S_i to maintain the f-sequence invariant.

Case (1) If $w_i = w_{i+1}$, all ballots in S_i terminate after position i.

Case (2) If $w_i = i+1$, $k-i$ ballots in S_i list each of candidates $i+2, \ldots, k$ in position $i+1$. This requires up to $(k-2)(k-1)$ ballots.

Case (3) If w_i next wins at ballot length $\ell > i+1$, then we need to insert w_i into this position in the winner prefix. Consider the sequence of winners $w_{i+1}, \ldots, w_{\ell-1}$. Let w_j be the last candidate in this sequence to make their first appearance. We will reallocate votes so that w_i is one vote behind w_j. Let c be the size of the vote gap between w_j and w_i before step i of IRV. For instance, $c = 1$ if $w_j = w_{i+1}$. For each candidate starting at w_{i+1} and going down the order of candidates by decreasing vote count before step i of IRV to w_j, $c+1$ ballots in S_i list that candidate in position $i+1$. For each candidate starting after w_j in vote count order and going down to $i+1$, c ballots in S_i list that candidate in position $i+1$. This requires at most $(k-3)(k-1) < (k-2)(k-1)$ ballots, an upper bound achieved if w_j has only one more vote than $i+1$ and $i = 1$.

Case (4) If w_i does not appear again in the sequence $w_{i+1}, \ldots, w_{k-1}$, then we will insert it into its correct position in the loser suffix. Consider the sequence of subsequent losers $i+1, \ldots, k$ and remove candidates that win at truncations lengths $i+1, \ldots, k-1$. Let j be the largest-indexed candidate in this pared-down sequence whose index is smaller than w_i (at least one such candidate exists since $i+1$ is eliminated before w_i and can't win at ballot lengths $> i$). We will insert w_i into the loser sequence so that they have one more vote than j. Let c be the size of the vote gap between w_i and j before step i of IRV. Consider the order of candidates by vote count before step i of IRV. For each candidate with more votes than j (excluding w_i), c ballots in S_i list that candidate in position $i+1$. For each candidate with fewer votes than j (including j but excluding i), $c-1$ ballots in S_i list that candidate in position $i+1$. After reallocation, w_i will then be one vote ahead of j and one vote behind the next candidate above them. This requires at most $(k-3)(k-1) < (k-2)(k-1)$ ballots, an upper bound achieved if $j = i+1$ and $i = 1$.

All ballots terminate after their last specified entry. We now prove that the truncation winner sequence of this profile is $w_1, \ldots, w_{k-1}$. We'll prove inductively that the f-sequence invariant is maintained by construction.

Base case ($\ell = 1$): By construction, the first place vote counts are exactly the f-sequence of $w_1, \ldots, w_{k-1}$.

Inductive case ($\ell \geq 2$): By inductive hypothesis, we have that after step $\ell - 1 < h$, the candidates $\ell - 1, \ldots, k$ were in their f-sequence order by decreasing vote count. We also know $\ell - 1$ must have been in last place, since they are eliminated $(\ell - 1)$st. Consider what occurs when $\ell - 1$ is eliminated. We will mirror the four cases of the construction. (1) If $w_{\ell-1} = w_\ell$, position $\ell - 1$ is empty and their ballots are all exhausted, leaving the order as is. The order of the candidates by vote count remains the f-sequence of the remaining candidates. (2) If $w_{\ell-1} = \ell$, then all candidates between $w_{\ell-1}$ and $\ell - 1$ overtake w_ℓ. The new order of candidates is again the f-sequence of the remaining candidates, since the winner prefix remains the same starting from w_ℓ and ℓ moves into last place. (3) If $w_{\ell-1}$ wins again at some $h > \ell - 1$, then our construction places it in the winner prefix exactly where it belongs: in order of first subsequent win. The loser suffix remains unchanged, leaving the correct f-sequence. (4) If $w_{\ell-1}$ does not win again at $h > \ell - 1$, then our construction inserts it into the loser suffix where it belongs: just before the highest-indexed non-subsequent-winner with a lower index than $w_{\ell-1}$. Here, the winner prefix in unaffected, leaving the correct f-sequence.

By construction, as soon as a ballot is reallocated, it becomes exhausted. Additionally, just before step h of IRV, all remaining truncated ballots are exhausted. Thus the order remains the same as trailing candidates are eliminated and w_h wins, since they were in the lead at the front of the f-sequence before step h.

Finally, this construction uses the number of ballots claimed:

$$\begin{aligned}
&\sum_{i=1}^{k} [(k-2)(k-1) + k - j] \\
&= k(k-2)(k-1) + k^2 - \sum_{i=1}^{k} j \\
&= k(k-2)(k-1) + k^2 - k(k+1)/2 \\
&= (2k^3 - 5k^2 + 3k)/2.
\end{aligned}$$

$\square$

Proof of Theorem 22. We'll construct a set of ballots such that candidate h wins at truncation $h = 1, \ldots, k-1$. Call the last candidate (the first one eliminated) k. Let $x = (2k-4)(k-2)$. Construct $x + 2(k-1)$ ballots ranking 1 first and $x + 2i$ ballots ranking candidates $i = 2, \ldots, k-2$ first. Construct $x + 3$ ballots ranking candidate $k-1$ first and x ranking candidate k first. Thus, the order of candidates from most to least first-place votes is $1, k-2, k-3, k-4, \ldots, 3, 2, k-1, k$ and the total number of ballots is $\Theta(k^3)$. We now fill in the ballots for each of the candidates.

<u>Candidate k</u>: Make $2k-4$ of the ballots ranking k first rank each of $2, \ldots, k-1$ second. These ballots then terminate. This requires $(2k-4)(k-2)$ ballots of $k-2$ types.

<u>Candidate $k-1$</u>: $2k$ of the ballots ranking $k-1$ first rank candidates $2, \ldots, k-2$ in positions $2, \ldots, k-2$, then terminate. The remaining ballots ranking $k-1$ first have length 1. Candidate $k-1$ thus uses only two ballot types.

<u>Candidates $i = 1, \ldots, k-2$</u>: Two of the ballots ranking i first rank candidates $k, 1, 2, \ldots, i-1, k-1$ in positions $2, \ldots, i+2$, then terminate—note that candidate 1's ballots of this form are $(1, k, k-1)$. The remaining ballots ranking i first have length 1. Candidate i thus uses only two ballot types.

Notice that the construction uses $O(k^2)$ ballots ranking each candidate first, for a total of $O(k^3)$ ballots. These are split among $k-2+2+2(k-2) = 3k-2 = \Theta(k)$ types. We now show that truncating ballots at length $h < k$ results in candidate h winning under IRV.

If $h = 1$, candidate 1 wins since they have the most first-place votes.

If $h = 2$, the first candidate eliminated is candidate k. Their second-place votes cause candidates $2, \ldots, k-1$ to overtake candidate 1, who is eliminated second. However, candidate 1's ballots of the form $(1, k, k-1)$—truncated to $(1, k)$—are now exhausted, so no reallocation occurs. This causes candidate $k-1$ to be eliminated third, and their $2k$ ballots ranking candidate 2 second cause candidate 2 to take the lead. The eliminations then proceed in the order $3, \ldots, k-2$, with no reallocation since those candidates' ballots are all exhausted. Candidate 2 wins.

For $h > 2$, we'll show inductively that for $\ell = 2, \ldots, h-1$ ($h \leq k-1$), the ℓth candidate eliminated is candidate $\ell - 1$, which causes candidate $k-1$ to jump one vote ahead of candidate ℓ, who falls into last place.

Base case ($\ell = 2$): The first two eliminations proceed as they did for $h = 2$. However, when candidate 1 is eliminated, their two ballots ranking $k-1$ third go to candidate $k-1$, since $h > 2$. Before this reallocation, candidate $k-1$ was in second-to-last place with $x+3+2k-4 = x+2k-1$, one vote behind candidate 2, who had $x+4+2k-4 = x+2k$ votes. The reallocation of candidate 1's ballots causes candidate $k-1$ to jump one vote ahead of candidate 2, who falls into last place.

Inductive case ($\ell > 2$): By inductive hypothesis, candidate $\ell - 2$ was last eliminated, which caused candidate $\ell - 1$ to drop into last place, one vote behind candidate $k-1$. Thus, candidate $\ell - 1$ is eliminated next. The next uneliminated candidate listed on their two ballots of length > 1 is $k-1$, since candidates $k, 1, \ldots \ell - 2$ have all been eliminated by inductive hypothesis and the cases above. When candidate $\ell - 1$ is eliminated, those two ballots cause candidate $k-1$ to jump one vote ahead of candidate ℓ, who has $x + 2\ell$ votes. Since candidate was one vote ahead of candidate $\ell - 1$ (who had $x + 2(\ell - 1)$ votes) and then gained two more, candidate $k-1$ is therefore one vote ahead of candidate ℓ after the ℓth elimination and redistribution.

We can now show that for $2 < h < k$, candidate h wins when the ballot length is

h. Consider such a ballot length h. By the inductive argument above, the $(h-1)$th candidate eliminated is candidate $h-2$, which causes candidate $h-1$ to fall into last place, one vote behind candidate $k-1$. Notice that when candidate $h-1$ is eliminated, they do not reallocate any votes to candidate $k-1$, since candidate $k-1$ appears at position $h+1$ on their two nontrivial ballots. Thus candidate $k-1$ is eliminated after candidate $h-1$ (note that if $h=k-1$, candidate $k-1$ wins after $h-1=k-2$ is eliminated). The $2k$ nontrivial ballots assigned to $k-1$ are then redistributed to candidate h, since they are the lowest-indexed non-eliminated candidate and they appear at position h on candidate $k-1$'s nontrivial ballots. These $2k$ ballots are enough to put candidate h in first place. The eliminations then proceed in the order $h+1, h+2, \ldots, k-2$, with no reallocation since those candidates' ballots only include candidates indexed lower than them (except k and $k-1$, who have been eliminated). Candidate h then wins.

□

Proof of Corollary 3. Perform the same constructions as in the proofs of Theorems 15 and 21, but with κ instead of k. Then add in another κ candidates with zero first place votes, which are always immediately eliminated (for the tie-free construction, these candidates need $0, \ldots, \kappa-1$ first-place votes to avoid a tie, resulting in the extra $\kappa(\kappa-1)/2$ ballots). Use them to fill in all the partial ballots generated by the construction up to position κ. Then fill in all ballots with the remaining candidates arbitrarily. Up to $h=\kappa$, the construction performs exactly as before, since all of the filler candidates are eliminated first regardless of ballot length, allowing the ballots to act as if they were partially filled out. No guarantees are made about the behavior under ballot lengths $h=\kappa+1, \ldots, 2\kappa$. □

Proof of Corollary 4. If we can make ballots as short as $\kappa - c$, then we can perform the same construction as above, but using c fewer fillers. Shortening ballots to $\kappa - c$ ensures that ballots will become exhausted when needed, while using fewer than $\kappa - c$ fillers. Those unused fillers are now additional candidates who could be made to win at longer ballot lengths, up to $\kappa + c - 1$. As before, we need to give the fillers $0, \ldots, k - c - 1$ first-place votes to avoid ties in the tie-free construction, resulting in the extra $(\kappa - c)(\kappa - c - 1)/2$ ballots. □

E.4 Additional figures

Figure E.1 shows two real-world IRV sample ballots with $h = 3$ and $h = 15$.

Table E.1 provides summary statistics of the PrefLib datasets we used, while Figure E.2 visualizes the distributions of k, h, and n in this data.

In Figure E.3, we show the versions of the heatmaps in Figure 6.2 with partial rather than full preferences. For general preferences, we shorted each of the 1000 voters preferences to a length uniform over $1, \ldots, k$. For 1-Euclidean voters, we uniformly shorted the preferences of each of the $\binom{k}{2} + 1$ voter types.

E.5 Experiment details

Experiments were run on a server with 144 Intel Xeon Gold 6254 CPUs and 1.5TB RAM running Ubuntu 20.04.4 LTS (Focal Fossa). All libraries used are documented in the code README, as well as detailed instructions for reproducing all experiments.

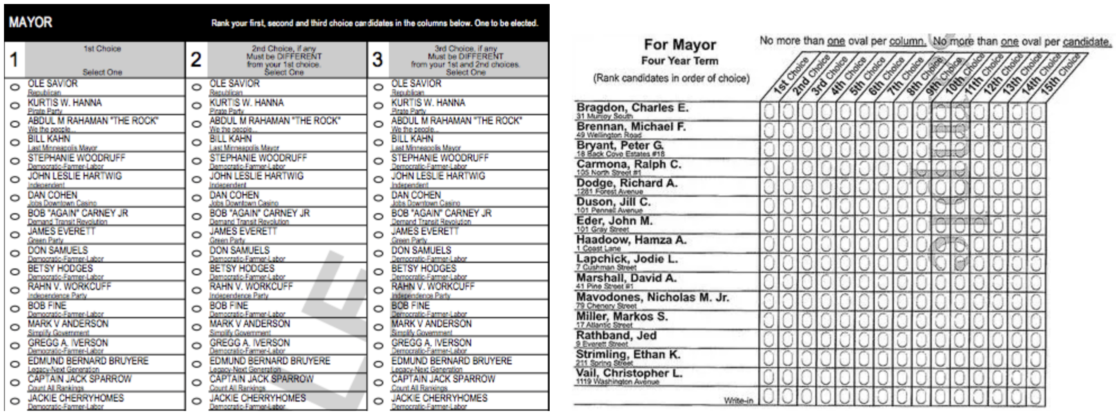


Figure E.1: Sample mayoral election ballots from Minneapolis, MN (left) and Portland, ME (right). Minneapolis ballots allow voters to rank up to three of the candidates, while Portland ballots allow voters to rank all of the candidates.

Table E.1: Dataset summary.

PrefLib id	**Election locale**	**# Elecs.**	***k***	***h***	**# Ballots**
`apa`	Am. Psych. Assn.	12	5	5	13318–20239
`aspen`	Aspen, CO	2	5–11	4–9	2468–2520
`berkley`	Berkeley, CA	1	4	3	4171
`burlington`	Burlington, VT	2	6	5	8974–9756
`debian`	Debian Project	8	4–9	4–9	143–504
`ers`	Anon. orgs.	87	3–29	3–29	9–3419
`glasgow`	Glasgow, Scotland	21	8–13	8–13	5199–12744
`irish`	Dublin, Ireland	3	9–14	9–14	29988–64081
`minneapolis`	Minneapolis, MN	2	7–9	3	32086–36655
`oakland`	Oakland, CA	7	4–11	3	11235–143860
`pierce`	Pierce County, WA	4	4–7	3	39974–298438
`sf`	San Francisco, CA	14	4–25	3	17675–193854
`sl`	San Leandro, CA	3	4–7	3	22360–25316
`takomapark`	Takoma Park, WA	1	4	4	202
`uklabor`	UK Labour Party	1	5	5	266

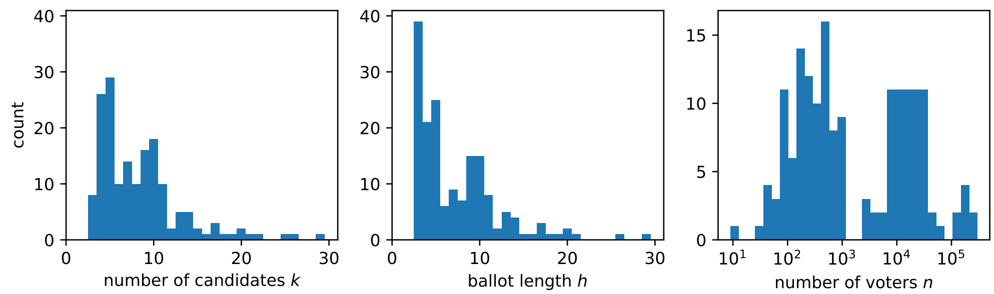


Figure E.2: Distributions of candidate counts, ballot lengths, and voter counts in the PrefLib election datasets.

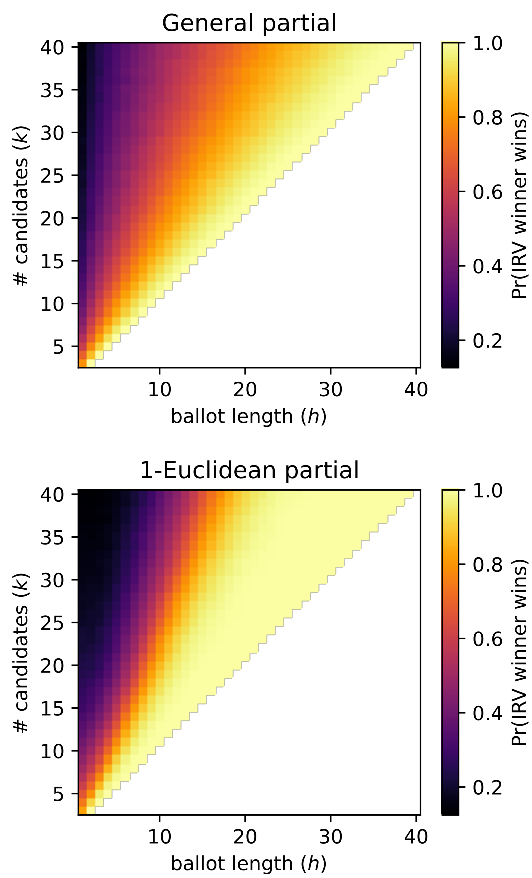


Figure E.3: Probability that truncated ballots produce the full IRV winner for general profiles (left) and 1-Euclidean profiles (right), with candidate counts $k = 2, \ldots, 40$ and ballot lengths $h = 1, \ldots, k-1$ with partial preferences (each voter's preferences are shorted uniformly at random). The results are qualitatively the same as in Figure 6.2.

APPENDIX F

TECHNICAL DETAILS FOR CHAPTER 7: THE MODERATING EFFECT OF INSTANT RUNOFF VOTING

> QUESTION 31. Whether, if three distinct points lie on a line, one of them must lie between the other two?
>
> *Objection 1.* A point has no extent; i.e., no length, breadth, or thickness. That which has no length obviously cannot lie, sit, or stand. Hence, a point cannot lie on a line.
>
> ---
> Carl E. Linderholm, *Mathematics Made Difficult*, 1972

F.1 Quotes about IRV moderation

In this section, we include some quotes indicating that moderating effects of IRV are an often-invoked argument in policy debates, suggesting the existence of a folk theory which we formalize in this work. Recall that IRV is commonly referred to as ranked-choice voting in United States.

F.2 Additional proofs

Since Theorem 25 relies on Lemma 3, we first prove Lemma 3.

Proof of Lemma 3. For notational simplicity, define $n = k + 1$ to be the number of gaps between candidates (including the leftmost and rightmost gaps bounded by 0 and 1) and let $S_1, \ldots, S_n$ be the sizes of the gaps. Additionally, let $X_1, \ldots, X_n$

be i.i.d. exponential random variables with mean 1 and let $T_n = \sum_{i=1}^{n} X_i$ be their sum. We then have $S_i = X_i/T_n$ (Holst, 1980) and T_n is independent of each S_i (the S_i are not independent, however). Since voters are uniform, the vote shares are $S_1 + S_2/2$ for the leftmost candidate and $(S_i + S_{i+1})/2$ for the ith candidate for $i = 2, \ldots, n-2$. For the rightmost candidate, we introduce an alternative indexing to avoid a subscript dependent on n. Let S_r and $S_{r'}$ be the rightmost and second-rightmost gaps, respectively, so that the rightmost candidate's vote share is $S_{r'}/2 + S_r$ (ditto for X_r and X_r'). We then have

$$\begin{aligned} S_1 + S_2/2 &= \frac{1}{T_n}(X_1 + X_2/2), \\ (S_i + S_{i+1})/2 &= \frac{1}{T_n}(X_i + X_{i+1})/2, \\ S_{r'}/2 + S_r &= \frac{1}{T_n}(X_{r'}/2 + X_r). \end{aligned}$$

Consider the asymptotic CDF of the leftmost candidate's vote share scaled by $n \geq 1$:

$$\begin{aligned} \lim_{n\to\infty} \Pr(n(S_1 + S_2/2) \leq z) &= \lim_{n\to\infty} \Pr\left(\frac{n}{T_n}(X_1 + X_2/2) \leq z\right) \\ &= \lim_{n\to\infty} \Pr\left(X_1 + X_2/2 \leq \frac{T_n}{n} z\right). \end{aligned}$$

Since it is the sum of n independent exponential RVs with mean 1, T_n has a Gamma$(n, 1)$ distribution (Holst, 1980), so it has variance n and expectation n. Thus, $\mathrm{Var}(T_n/n) = 1/n^2 \, \mathrm{Var}(T_n) = 1/n$. Since $\mathrm{Var}(T_n/n) \to 0$ as $n \to \infty$ and $\mathrm{E}[T_n/n] = 1$, we have $\lim_{n\to\infty} T_n/n = 1$. This can also be seen using the Law of Large Numbers, since $\sum_i X_i/n$ converges in probability to $\mathrm{E}[X_i] = 1$. Thus,

$$\lim_{n\to\infty} \Pr\left(X_1 + X_2/2 \leq \frac{T_n}{n} z\right) = \Pr\left(X_1 + X_2/2 \leq z\right).$$

Similarly for middle candidates:

$$
\begin{aligned}
\lim_{n\to\infty} \Pr(n(S_i + S_{i+1})/2 \le z) &= \lim_{n\to\infty} \Pr\left(\frac{n}{T_n}(X_i + X_{i+1})/2 \le z\right) \\
&= \lim_{n\to\infty} \Pr\left((X_i + X_{i+1})/2 \le \frac{T_n}{n} z\right) \\
&= \Pr\left((X_i + X_{i+1})/2 \le z\right).
\end{aligned}
$$

Likewise, for the rightmost candidate,

$$
\begin{aligned}
\lim_{n\to\infty} \Pr(n(S_{r'}/2 + S_r) \le z) &= \lim_{n\to\infty} \Pr\left(\frac{n}{T_n}(X_{r'}/2 + X_r) \le z\right) \\
&= \lim_{n\to\infty} \Pr\left(X_{r'}/2 + X_r \le \frac{T_n}{n} z\right) \\
&= \Pr\left(X_{r'}/2 + X_r \le z\right).
\end{aligned}
$$

Thus, the asymptotic distributions of n times the vote shares equal the distributions of the corresponding sums of exponentials RVs. This is the same idea used in proving the asymptotic distribution of the maximum gap size (Holst, 1980). Now consider the distribution of the maximum vote share. Let V_k be the maximum vote share with $k \ge 3$ candidates (and therefore $n = k+1$ gaps between candidates) and let M_k be the maximum corresponding exponential RV sum. As above, by LLN, $\lim_{n\to\infty} \Pr(nV_k \le z) = \Pr(M_k \le z)$. Let $L_n = X_1 + X_2/2$, $C_i = (X_i + X_{i+1})/2$, and $R_n = X_{n-1}/2 + X_n$.

$$
\begin{aligned}
\Pr(M_k \le z) &= \Pr(L_n \le z, C_2 \le z, \ldots, C_{n-2}, \le z, R_n \le z) \\
&= \Pr(L_n \le z) \Pr(R_n \le z \mid L_n \le z, C_2 \le z, \ldots, C_{n-2} \le z) \\
&\quad \cdot \prod_{i=2}^{n-2} \Pr(C_i \le z \mid L_n \le z, \ldots, C_{i-1} \le z).
\end{aligned}
$$

Then, using the facts that each X_i is independent (and thus C_i is independent of C_j for $j > i+1$ and $j < i-1$) and that each C_i is identically distributed, we can simplify the conditioning:

$$\begin{aligned}
&\Pr(M_k \le z)\\
&= \Pr(L_n \le z) \Pr(C_2 \le z \mid L_n \le z) \Pr(R_n \le z \mid C_{n-2} \le z) \prod_{i=3}^{n-2} \Pr(C_i \le z \mid C_{i-1} \le z)\\
&= \Pr(L_n \le z) \Pr(C_2 \le z \mid L_n \le z) \Pr(R_n \le z \mid C_{n-2} \le z) \Pr(C_i \le z \mid C_{i-1} \le z)^{n-4}.
\end{aligned}$$

We now only have four different probabilities to compute.

1. $\Pr(L_n \le z) = \Pr(X_1 + X_2/2 \le z)$. Note that $X_1 \sim \mathrm{Exp}(1)$, $X_2/2 \sim \mathrm{Exp}(2)$, and X_1 and $X_2/2$ are independent. Thus:

$$\begin{aligned}
\Pr(X_1 + X_2/2 \le z) &= \int_0^z e^{-t} \left(\int_0^{z-t} 2e^{-2s} ds \right) dt\\
&= 1 - e^{-2z} - 2e^{-z}.
\end{aligned}$$

2. $$\Pr(C_2 \le z \mid L_n \le z) = \frac{\Pr((X_2 + X_3)/2 \le z, X_1 + X_2/2 \le z)}{\Pr(X_1 + X_2/2 \le z)}.$$

We already know the denominator from the previous calculation. As before, $X_2/2$ and $X_3/2$ are independent and $\mathrm{Exp}(2)$ distributed. To have $(X_2 + X_3)/2 \le z$ and $X_1 + X_2/2 \le z$, we first need $X_1 \le z$, then $X_2/2 \le z - X_1$, and finally $X_3/2 \le z - X_2/2$. Thus:

$$\begin{aligned}
\Pr((X_2 + X_3)/2 \le z, X_1 + X_2/2 \le z) &= \int_0^z e^{-t} \left(\int_0^{z-t} 2e^{-2s} \left[\int_0^{z-s} 2e^{-2r}\, dr \right] ds \right) dt\\
&= 1 - 2e^{-3z} + 3e^{-2z} - 2e^{-z} - 2e^{-2z} z.
\end{aligned}$$

So,

$$\begin{aligned}
\Pr(C_2 \le z \mid L_n \le z) &= \frac{1 - 2e^{-3z} + 3e^{-2z} - 2e^{-z} - 2e^{-2z} z}{1 - e^{-2z} - 2e^{-z}}\\
&= 1 + 2e^{-z} - \frac{2(z + e^z - 4)}{e^z(e^z - 2) - 1}.
\end{aligned}$$

3.

$$\Pr(C_i \le z \mid C_{i-1} \le z) = \frac{\Pr((X_i + X_{i+1})/2 \le z, (X_{i-1} + X_i)/2 \le z)}{\Pr((X_{i-1} + X_i)/2 \le z)}.$$

First, consider the denominator:

$$\begin{aligned}\Pr((X_{i-1} + X_i)/2 \le z) &= \int_0^z 2e^{-2t} \left(\int_0^{z-t} 2e^{-2s}\, ds \right) dt \\ &= 1 - e^{-2z}(1 + 2z).\end{aligned}$$

Now, the numerator. First, we'll require $X_{i-1}/2 \le z$, then $X_i/2 \le z - X_{i-1}/2$, then $X_{i+1}/2 \le z - X_i/2$:

$$\begin{aligned}&\Pr((X_i + X_{i+1})/2 \le z, (X_{i-1} + X_i)/2 \le z) \\ &= \int_0^z 2e^{-2t} \left(\int_0^{z-t} 2e^{-2s} \left[\int_0^{z-s} 2e^{-2r}\, dr \right] ds \right) dt \\ &= 1 - e^{-4z} - 4e^{-2z} z.\end{aligned}$$

Thus,

$$\Pr(C_i \le z \mid C_{i-1} \le z) = \frac{1 - e^{-4z} - 4e^{-2z} z}{1 - e^{-2z}(1 + 2z)}.$$

4.

$$\Pr(R_n \le z \mid C_{n-2} \le z) = \frac{\Pr(X_{n-1}/2 + X_n \le z, (X_{n-2} + X_{n-1})/2 \le z)}{\Pr((X_{n-2} + X_{n-1})/2 \le z)}$$

We already know the denominator from the previous step. We also know the numerator, by symmetry with the numerator in step 2. Thus:

$$\Pr(R_n \le z \mid C_{n-2} \le z) = \frac{1 - 2e^{-3z} + 3e^{-2z} - 2e^{-z} - 2e^{-2z} z}{1 - e^{-2z}(1 + 2z)}.$$

Putting the four pieces together and simplifying:

$$\Pr(M_k \leq z) = \left(1 - e^{-2z} - 2e^{-z}\right) \left(\frac{1 - 2e^{-3z} + 3e^{-2z} - 2e^{-z} - 2e^{-2z}z}{1 - e^{-2z}(1+2z)}\right)$$
$$\cdot \left(1 + 2e^{-z} - \frac{2(z + e^z - 4)}{e^z(e^z - 2) - 1}\right) \left(\frac{1 - e^{-4z} - 4e^{-2z}z}{1 - e^{-2z}(1+2z)}\right)^{n-4}$$
$$= \frac{e^{-4z}(-2 + e^z(3 + e^z(-2 + e^z) - 2z))^2}{-1 + e^{2z} - 2z} \left(\frac{1 - e^{-4z} - 4e^{-2z}z}{1 - e^{-2z}(1+2z)}\right)^{n-4}. \tag{$*$}$$

We want to take the limit of $(*)$ as $n \to \infty$. We'll focus on the second part first, since the limit of a product is the product of the limits (as we will see, both limits are well-defined). Define

$$\ell(z) = \lim_{n\to\infty} \left(\frac{1 - e^{-4z} - 4e^{-2z}z}{1 - e^{-2z}(1+2z)}\right)^{n-4}.$$

Take the log to handle the exponent:

$$\log \ell(z) = \lim_{n\to\infty} (n-4) \log \left(\frac{1 - e^{-4z} - 4e^{-2z}z}{1 - e^{-2z}(1+2z)}\right)$$
$$= \lim_{n\to\infty} [(n-4)\log\left(1 - e^{-4z} - 4e^{-2z}z\right) - (n-4)\log\left(1 - e^{-2z}(1+2z)\right)].$$

Now we'll split the limit into its two terms and plug in $(\log n + \log\log n + x)/2$

for z in $\ell(z)$. The first term:

$$\lim_{n\to\infty}(n-4)\log\left(1-e^{-4(\log n+\log\log n+x)/2}-4e^{-2(\log n+\log\log n+x)/2}(\log n+\log\log n+x)/2\right)$$

$$= \lim_{n\to\infty}(n-4)\log\left(1-e^{-2x}n^{-2}\log^{-2}n-2e^{-x}n^{-1}\log^{-1}n(\log n+\log\log n+x)\right) \qquad \text{(simplify)}$$

$$= \lim_{n\to\infty}\frac{\log\left(1-e^{-2x}n^{-2}\log^{-2}n-2e^{-x}n^{-1}\log^{-1}n(\log n+\log\log n+x)\right)}{(n-4)^{-1}} \qquad \text{(rearrange)}$$

$$= \lim_{n\to\infty}-(n-4)^2\frac{d}{dn}\log\left(1-e^{-2x}n^{-2}\log^{-2}n-2e^{-x}n^{-1}\log^{-1}n(\log n+\log\log n+x)\right) \qquad \text{(l'Hôpital's rule)}$$

$$= \lim_{n\to\infty}-(n-4)^2\frac{2e^{x}n\log n(\log n+1)(\log n+\log\log n+x-1)+2\log n+2}{e^{x}n^2\log^2 n(e^{x}n\log n-2\log n-2\log\log n-2x)-n\log n} \qquad \text{(take derivative)}$$

$$= \lim_{n\to\infty}\frac{-2e^{x}n^3\log^3 n-O(n^3\log^2 n)}{e^{2x}n^3\log^3 n-O(n^2\log^3 n)} \qquad \text{(isolate highest order terms)}$$

$$= -2e^{-x}.$$

Plugging $(\log n+\log\log n+x)/2$ into the second term and following the same strategy as above:

$$\lim_{n\to\infty}(n-4)\log\left(1-e^{-2((\log n+\log\log n+x)/2)}(1+2((\log n+\log\log n+x)/2))\right)$$
$$= \lim_{n\to\infty}-(n-4)^2\frac{d}{dn}\log\left(1-e^{-x}n^{-1}\log^{-1}n(1+\log n+\log\log n+x)\right)$$
$$= \lim_{n\to\infty}(n-4)^2\frac{(\log n+1)(\log n+\log\log n+x)}{n\log n\left(-e^{x}n\log n+\log n+\log\log n+x+1\right)}$$
$$= \lim_{n\to\infty}\frac{n^2\log^2 n+O(n^2\log n\log\log n)}{-e^{x}n^2\log^2 n+O(n\log^2 n)}$$
$$= -e^{-x},$$

where splitting the limit is allowed because both limits are finite. Thus,

$$\begin{aligned}\log\ell((\log n+\log\log n+x)/2) &= -2e^{-x}-(-e^{-x})\\ &= -e^{-x}.\end{aligned}$$

We then have $\ell((\log n + \log\log n + x)/2) = e^{-e^{-x}}$. Going back to the first part of $(*)$ (recall that we are plugging in $z = (\log n + \log\log n + x)/2$),

$$\begin{aligned}
\lim_{n\to\infty} \frac{e^{-4z}(-2+e^z(3+e^z(-2+e^z)-2z))^2}{-1+e^{2z}-2z} &= \lim_{z\to\infty} \frac{e^{-4z}(-2+e^z(3+e^z(-2+e^z)-2z))^2}{-1+e^{2z}-2z} \\
&= \lim_{z\to\infty} \frac{e^{2z}-O(e^z)}{e^{2z}-O(z)} \\
&= 1.
\end{aligned}$$

Combining these findings gives us the limits of $(*)$ as $n \to \infty$, again plugging in $z = (\log n + \log\log n + x)/2$ and using the results above,

$$\lim_{n\to\infty} \Pr\left(M_k \leq \frac{\log n + \log\log n + x}{2}\right) = e^{-e^{-x}}.$$

As we saw at the beginning of the proof, to convert from the max sum of exponential RVs to the max plurality vote share (in the limit), we simply multiply by n. We can additionally convert back to $k+1=n$ to prove the claim:

$$\begin{aligned}
e^{-e^{-x}} &= \lim_{n\to\infty} \Pr\left(M_k \leq \frac{\log n + \log\log n + x}{2}\right) \\
&= \lim_{n\to\infty} \Pr\left(nV_k \leq \frac{\log n + \log\log n + x}{2}\right) \\
&= \lim_{k\to\infty} \Pr\left(V_k \leq \frac{\log(k+1) + \log\log(k+1) + x}{2(k+1)}\right).
\end{aligned}$$

□

Since $e^{-e^{-x}} \to 1$ and $e^{-e^{x}} \to 0$ as $x \to \infty$, we immediately have the following corollary of Lemma 3.

Corollary 13. *Let $n = k+1$. For any function $g(n)$ with $\lim_{k\to\infty} g(n) = \infty$,*

$$\lim_{k\to\infty} \Pr\left(\frac{\log n + \log\log n - g(n)}{2n} \leq V_k \leq \frac{\log n + \log\log n + g(k)}{2n}\right) = 1.$$

Intuitively, Corollary 13 states that the asymptotic winning plurality vote share is almost exactly $\frac{\log(k+1)+\log\log(k+1)}{2(k+1)}$ with probability 1. We now provide a useful lemma before proving Theorem 25.

Lemma 11. *Let A_t and B_t be events with $\lim_{t\to\infty} \Pr(A_t) = 1$ and $\lim_{t\to\infty} \Pr(B_t) = \beta > 0$. Then $\lim_{t\to\infty} \Pr(A_t \mid B_t) = 1$.*

Proof. Using the Law of Total Probability and some basic probability facts,

$$\begin{aligned}
\Pr(A_t \mid B_t) &= \frac{\Pr(A_t \cap B_t)}{\Pr(B_t)} \\
&= \frac{\Pr(B_t) - \Pr(\overline{A_t} \cap B_t)}{\Pr(B_t)} \\
&\geq 1 - \frac{\Pr(\overline{A_t})}{\Pr(B_t)}
\end{aligned}$$

Thus,

$$\begin{aligned}
\lim_{t\to\infty} \Pr(A_t \mid B_t) &\geq \lim_{t\to\infty} 1 - \frac{\Pr(\overline{A_t})}{\Pr(B_t)} \\
&= 1 - \frac{0}{\beta} \\
&= 1.
\end{aligned}$$

We then have $\lim_{t\to\infty} \Pr(A_t \mid B_t) = 1$.

□

Finally, we can prove Theorem 25.

Proof of Theorem 25. Consider a plurality election with k candidates on the circle with circumference 1, with points on the circle mapped to the interval $[0, 1)$ (we'll say the point on the circle corresponding to the endpoints of the interval maps to 0). Let C_k be the position of the plurality winner on the circle with candidates positioned uniformly at random. By rotational symmetry, C_k is uniform over the interval $[0,1)$. Consider a particular configuration of k candidates on the circle. When we break the circle to make it the unit interval, we only change the vote shares of the candidates closest to 0 and 1 (call them x_ℓ and x_r, respectively). Let $x_{\ell'}$ be the second-closest candidate to 0 and let $x_{r'}$ be the second-closest candidate to 1.

Consider $\Pr(x_{\ell'} < y) = 1 - \Pr(x_{\ell'} \geq y)$. The event $x_{\ell'} \geq y$ can be partitioned into two cases: either $x_\ell < y$ or $x_\ell \geq y$. Thus,

$$\begin{aligned}\Pr(x_{\ell'} \geq y) &= \Pr(x_{\ell'} \geq y, x_\ell < y) + \Pr(x_{\ell'} \geq y, x_\ell \geq y)\\ &= \Pr(x_\ell < y)\Pr(x_{\ell'} \geq y \mid x_\ell < y) + (1-y)^k\\ &= (1-(1-y)^k)(1-y)^{k-1} + (1-y)^k.\end{aligned}$$

Now, pick $y = \frac{\log k}{4k}$. We then have:

$$\begin{aligned}\lim_{k\to\infty} \Pr\left(x_{\ell'} < \frac{\log k}{4k}\right) &= \lim_{k\to\infty}\left[1 - \Pr\left(x_{\ell'} \geq \frac{\log k}{4k}\right)\right]\\ &= \lim_{k\to\infty}\left[1 - \left(1-\left(1-\frac{\log k}{4k}\right)^k\right)\left(1-\frac{\log k}{4k}\right)^{k-1} - \left(1-\frac{\log k}{4k}\right)^k\right]\end{aligned}$$

Note that $\lim_{k\to\infty}(1-\log k/(4k))^k = 0$, since $\lim_{k\to\infty}(1-x/k)^k = e^{-x}$. We also have $\lim_{k\to\infty}(1-\log k/(4k))^{k-1} = 0$, since $(1-\log k/(4k))^{k-1} = (1-\log k/(4k))^k/(1-\log k/(4k))$ and $\lim_{k\to\infty}(1-\log k/(4k)) = 1$. This means

$$\begin{aligned}\lim_{k\to\infty} \Pr\left(x_{\ell'} < \frac{\log k}{4k}\right) &= \lim_{k\to\infty}\left[1 - \left(1-\left(1-\frac{\log k}{4k}\right)^k\right)\left(1-\frac{\log k}{4k}\right)^{k-1} - \left(1-\frac{\log k}{4k}\right)^k\right]\\ &= 1-(1-0)\cdot 0 - 0\\ &= 1. \end{aligned} \tag{F.1}$$

Symmetrically, we also have $\lim_{k\to\infty}\Pr\left(x_{r'} > 1 - \frac{\log k}{4k}\right) = 1$. We can therefore bound the asymptotic vote shares of x_ℓ and x_r on both the circle and the unit interval. Neither can get more votes (in either setting) than the distance between $x_{\ell'}$ and $x_{r'}$ on the circle; i.e., $b = x_{\ell'} + 1 - x_{r'}$ is an upper bound on the vote shares of x_ℓ and x_r on both the circle and the unit interval. We can use the above facts to find an asymptotic bound on b. First, consider the probability that *both* $x_{r'}$ and $x_{\ell'}$ are close to the boundaries:

$$\begin{aligned}
&\lim_{k\to\infty} \Pr\left(x_{r'} > 1 - \frac{\log k}{4k}, x_{\ell'} < \frac{\log k}{4k}\right) \\
&= \lim_{k\to\infty} \Pr\left(x_{\ell'} < \frac{\log k}{4k}\right) \cdot \lim_{k\to\infty} \Pr\left(x_{r'} > 1 - \frac{\log k}{4k} \mid x_{\ell'} < \frac{\log k}{4k}\right) \\
&= \lim_{k\to\infty} \Pr\left(x_{r'} > 1 - \frac{\log k}{4k} \mid x_{\ell'} < \frac{\log k}{4k}\right) && \text{(by Equation (F.1))} \\
&= 1. && \text{(by Lemma 11)}
\end{aligned}$$

If $x_{r'} > 1 - \frac{\log k}{4k}$, then $1 - x_{r'} < \frac{\log k}{4k}$. Thus, if both $x_{r'} > 1 - \frac{\log k}{4k}$ and $x_{\ell'} < \frac{\log k}{4k}$, we then have $b = x_{\ell'} + 1 - x_{r'} < \frac{\log k}{2k}$. Therefore $\lim_{k\to\infty} \Pr(b < \frac{\log k}{2k}) = 1$. That is, the asymptotic vote share of the leftmost and rightmost candidates are both less than $\frac{\log k}{2k}$ with probability 1. Meanwhile, we know from Lemma 3 that the asymptotic winning vote share on the unit interval is larger than $\frac{\log k}{2k}$ with probability 1; i.e., with probability 1, neither x_ℓ nor x_r is the winner (on either the circle or unit interval). Since no other vote shares change when we go between the unit interval and the circle, the winner on the unit interval is the same as the winner on the circle with probability 1 as $k \to \infty$. Thus, $\lim_{k\to\infty} \Pr(P_k \le x) = \lim_{k\to\infty} \Pr(C_k \le x) = x$. □

Proof of Theorem 27. Begin the same way as in the proof of Theorem 26, minimizing x's vote shares with candidates at $c - \epsilon$ and $1 - c + \epsilon$. We thus have

$$v(x) = F\left(\frac{x + 1 - c + \epsilon}{2}\right) - F\left(\frac{c - \epsilon + x}{2}\right).$$

If f is monotonic and non-decreasing over $[0, 1/2]$, then x has the smallest vote share when $x = c$. At this edge of the interval, x's vote share is at least

$$\begin{aligned}
v(x) &\geq F\left(\frac{c+1-c+\epsilon}{2}\right) - F\left(\frac{c-\epsilon+c}{2}\right) \\
&= F\left(\frac{1+\epsilon}{2}\right) - F\left(c - \frac{\epsilon}{2}\right) \\
&> F\left(1/2\right) - F\left(c\right) && (F \text{ increasing}) \\
&= 1/2 - F\left(c\right). && (\text{symmetry of } f)
\end{aligned}$$

Suppose $c \leq F^{-1}(1/6)$. Then:

$$\begin{aligned}
1/2 - F\left(c\right) &\geq 1/2 - F(F^{-1}(1/6)) \\
&= 1/2 - 1/6 \\
&= 1/3.
\end{aligned}$$

Thus x cannot be eliminated next. The IRV winner must therefore be in $[c, 1-c]$ by the same argument as in Theorem 23. □

Proof of Theorem 28. As in Theorem 26, minimize x's vote shares with candidates at $c - \epsilon$ and $1 - c + \epsilon$. If f is monotonic and non-increasing over $[0, 1/2]$, then x has the smallest vote share when $x = 1/2$:

$$\begin{aligned}
v(x) &\geq F\left(\frac{x+1-c+\epsilon}{2}\right) - F\left(\frac{c-\epsilon+x}{2}\right) \\
&= F\left(\frac{3}{4} - \frac{c-\epsilon}{2}\right) - F\left(\frac{1}{4} + \frac{c-\epsilon}{2}\right) \\
&= 2\left[F(1/2) - F\left(1/4 + \frac{c-\epsilon}{2}\right)\right] && (\text{symmetry of } f) \\
&= 1 - 2F\left(1/4 + \frac{c-\epsilon}{2}\right). && (\text{symmetry of } f)
\end{aligned}$$

Suppose $c \leq 2(F^{-1}(1/3) - 1/4)$. Then we have:

$$\begin{aligned}
1 - 2F\left(1/4 + \frac{c-\epsilon}{2}\right) &\geq 1 - 2F\left(1/4 + \frac{2(F^{-1}(1/3) - 1/4) - \epsilon}{2}\right) \\
&= 1 - 2F\left(F^{-1}(1/3) - \frac{\epsilon}{2}\right) \\
&> 1 - 2F\left(F^{-1}(1/3)\right) && (F \text{ increasing}) \\
&= 1 - 2/3 \\
&= 1/3.
\end{aligned}$$

As before, x cannot be eliminated next and some candidate in $[c, 1-c]$ must win under IRV. □

Proof of Theorem 29. Suppose there is exactly one candidate $\ell \in [0, c]$ and at least one candidate each in $(c, 1-c)$ and $[c, 1]$ (if there are no candidates in $(c, 1-c)$, the claim is vacuously true). The smallest vote share ℓ could have occurs when $\ell = 0$ and there is a candidate at $c + \epsilon$. In this case, ℓ's vote share is

$$F\left(\frac{c+\epsilon}{2}\right) > F(c/2).$$

If $c \geq 2F^{-1}(1/3)$, then

$$\begin{aligned}
F(c/2) &\geq F\left(2F^{-1}(1/3)/2\right) \\
&= 1/3.
\end{aligned}$$

Thus, ℓ cannot be eliminated next. By a symmetric argument, the last candidate r in $[1-c, 1]$ is guaranteed more than a third of the vote. As long we we begin with at least one candidate in $[0, c]$ and at least one candidate in $[1-c, 1]$, once there is only one candidate remaining in each of these intervals, they will survive elimination until all candidates in $(c, 1-c)$ are eliminated. At this point, the IRV winner is guaranteed to be in $[0, c]$ or $[1-c, 1]$. □

Proof of Theorem 30. First, if $x_1 = 0$, consider a voter distribution with density function f that increases monotonically over $[0, 1]$. Let r be the position of the candidate immediately to the right of x_1. The candidate at r must have a higher vote share than x_1, since they split the interval $[0, r]$ and the right half of this interval has more voter mass (as f is increasing). Thus, x_1 cannot win under plurality, regardless of how many additional candidates we add. A symmetric argument shows that there are cases where a candidate at $x_1 = 1$ cannot win under plurality.

Now we show how to add candidates to the initial set $x_1, \ldots, x_\kappa$ so that x_1 becomes the plurality winner if $x_1 \notin \{0, 1\}$. Since f is continuous and $f(x_1) > 0$, there must exist some $\delta > 0$ such that $|f(x_1) - f(x_1 + t)| < f(x_1)/4$ for $t \in [-\delta, \delta]$. This argument still holds if we make δ smaller, so we ensure than $\delta < \min\{x_1, 1 - x_1\}$ (both are positive since $x_1 \notin \{0, 1\}$). Now add candidates $\ell = x_1 - \delta$ and $r = x_1 + \delta$. Then the vote share ℓ gets on its right is less than $\frac{5}{8}\delta f(x_1)$:

$$\begin{aligned}
\int_{\ell}^{(x_1+\ell)/2} f(t)dt &< \int_{x_1-\delta}^{(x_1+x_1-\delta)/2} (f(x_1) + f(x_1)/4)dt \\
&= \frac{5}{4} f(x_1)(x_1 - \delta/2 - x_1 + \delta) \\
&= \frac{5}{8}\delta f(x_1).
\end{aligned}$$

The same argument shows the vote share r gets on its left is less than $\frac{5}{8}\delta f(x_1)$. Meanwhile, the vote share of x_1 is more than $\frac{3}{4}\delta f(x_1)$:

$$\begin{aligned}
\int_{(x_1+\ell)/2}^{(x_1+r)/2} f(t)dt &> \int_{(x_1+x_1-\delta)/2}^{(x_1+x_1+\delta)/2} (f(x_1) - f(x_1)/4)dt \\
&= \frac{3}{4} f(x_1)(x_1 + \delta/2 - x_1 + \delta/2) \\
&= \frac{3}{4}\delta f(x_1).
\end{aligned}$$

Thus x_1 has a higher vote share than ℓ gets on its right and than r gets on its left. Now, we repeatedly add candidates to the left of ℓ adjacent to the candidates with the maximum vote shares in $[0, \ell]$. We can make the maximum vote share in $[0, \ell]$ arbitrarily small (except ℓ's) by adding enough candidates in this way—in particular, we can make it smaller than $v(x_1)$. We can also make ℓ's total vote share smaller than $v(x_1)$, since the vote share ℓ gets on its right is strictly smaller than $v(x_1)$. Doing the same in $[r, 1]$ then ensures x_1 is the plurality winner. $\square$

F.3 Derivations of f_{P_3} and f_{R_3}

In this section, we derive the probability density functions of the position of the plurality and IRV winners for 1-Euclidean profiles with $k = 3$ uniformly placed candidates and continuous uniform voters. We then compute the variances of these distributions. The calculations for plurality are in Appendix F.3.1 and the calculations for IRV are in Appendix F.3.2. First, we provide an overview of our approach.

Let $X_1, \ldots, X_k \sim \text{Unif}(0, 1)$ be the random positions of the k candidates and let $W \in \{X_1, \ldots, X_k\}$ be the position of the winner. Let $X_{(i)}$ denote the ith order statistic of $X_1, \ldots, X_k$.

The density of the winner's position at a point w, denoted $f(w)$, is k times the probability that a particular candidate at w is the winner (times the density of that candidate's position at w, which is 1). We sum over the possible order statistics of the winner to compute $f(w)$:

$$
\begin{aligned}
f(w) &= k \Pr(w \text{ wins}) \\
&= k \sum_{i=1}^{k} \Pr(w \text{ wins}, w = X_{(i)})
\end{aligned}
$$

What is the probability that a candidate at position w with order statistic i wins? Say the winner is candidate 1. We can choose which $i-1$ candidates are to their left. The remaining $k-i$ candidates are to their right. Then, we integrate over the positions of the other candidates where the candidate at w wins.

$$
\begin{aligned}
&\Pr(w \text{ wins}, w = X_{(i)}) \\
&= \binom{k}{i-1} \underbrace{\int_0^{x_i} \cdots \int_0^{w}}_{i-1} \underbrace{\int_{x_i}^{1} \cdots \int_{w}^{1}}_{k-1} \mathbf{1}[w \text{ wins given positions } x_2, \ldots, x_k] dx_k \ldots dx_2
\end{aligned}
$$

We can also note that the win probability (and therefore the winner density) is symmetric about 0.5: $f(w) = 1 - f(w)$. We therefore only need to consider $w \in [0, 0.5]$.

F.3.1 1d plurality winner distribution, $k = 3$

We'll compute $\Pr(w \text{ wins}, w = X_{(i)})$ for $i = 1, 2, 3$ and $w \in [0, 0.5]$. That is, we'll compute the win probability of a candidate at a point w in the cases where they are the leftmost, the middle, and the rightmost of the three candidates. In all cases, we'll call the winner candidate 1 and the losers candidates 2 and 3, at positions x_2 and x_3.

1. $w = X_{(1)}$. Consider the order $w < x_2 < x_3$ (we'll multiply by 2 later to account for the ordering $w < x_3 < x_2$). For w to beat x_2, we need:

$$\begin{aligned} & w + (x_2 - w)/2 > (x_2 - w)/2 + (x_3 - x_2)/2 \\ \Leftrightarrow \quad & w > (x_3 - x_2)/2 \\ \Leftrightarrow \quad & 2w > x_3 - x_2 \\ \Leftrightarrow \quad & x_3 < 2w + x_2 \end{aligned} \tag{F.2}$$

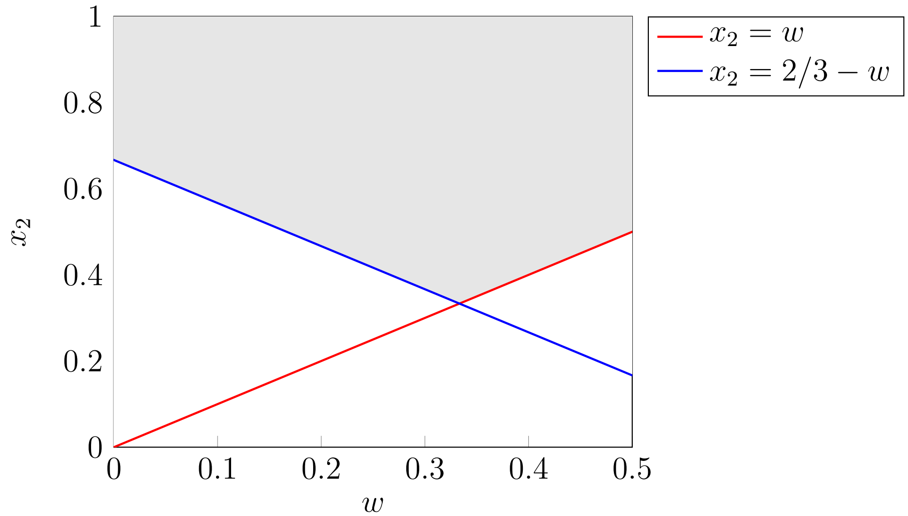
For w to beat x_3, we need

$$\begin{aligned} & w + (x_2 - w)/2 > 1 - x_3 + (x_3 - x_2)/2 \\ \Leftrightarrow \quad & 2w + x_2 - w > 2 - 2x_3 + x_3 - x_2 \\ \Leftrightarrow \quad & w + x_2 > 2 - x_3 - x_2 \\ \Leftrightarrow \quad & x_3 > 2 - 2x_2 - w \end{aligned} \tag{F.3}$$

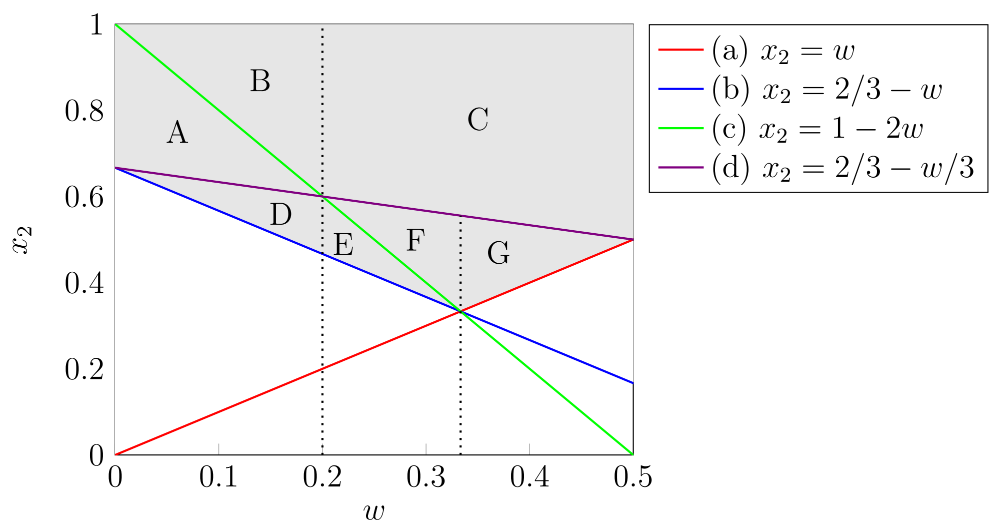
For both (F.2) and (F.3) to be feasible, w and x_2 cannot both be too small. The inequalities match at

$$\begin{aligned} & 2w + x_2 = 2 - 2x_2 - w \\ \Leftrightarrow \quad & 3w + 3x_2 = 2 \\ \Leftrightarrow \quad & w + x_2 = 2/3. \end{aligned}$$

We therefore need $w + x_2 > 2/3$ to satisfy both (F.2) and (F.3). We summarize the constraints on w and x_2 in the following plot, where the gray region contains points where w can win (given $w < x_2 < x_3$).



The lower bound on x_3 for w to win is the minimum of x_2 and $2 - 2x_2 - w$, while the upper bound is the maximum of 1 and $2w + x_2$. We'll plot the lines where these bounds change, namely $x_2 = 2 - 2x_2 - w \Leftrightarrow x_2 = 2/3 - w/3$ and $1 = 2w + x_2 \Leftrightarrow x_2 = 1 - 2w$, and label the regions over which we can easily integrate:



Above line (c), the upper bound for x_3 is 1; below (c), it's $2w + x_2$. Above line (d), the lower bound for x_3 is x_2; below, it's $2 - 2x_2 - w$. Lines (c) and (d) intersect at $w = 1/5$, while lines (a), (b), and (c) intersect at $w = 1/3$.

With this information in hand, we can compute the integral describing the win probability of w by summing the integrals for regions A–G and multiplying by 2 to account for the ordering $w < x_3 < x_2$:

$$
\begin{aligned}
A: &\quad \int_{2/3-w/3}^{1-2w} \int_{x_2}^{2w+x_2} dx_3\, dx_2 &&= 2w/3 - 10w^2/3 \\
B: &\quad \int_{1-2w}^{1} \int_{x_2}^{1} dx_3\, dx_2 &&= 2w^2 \\
C: &\quad \int_{2/3-w/3}^{1} \int_{x_2}^{1} dx_3\, dx_2 &&= 1/18 + w/9 + w^2/18 \\
D: &\quad \int_{2/3-w}^{2/3-w/3} \int_{2-2x_2-w}^{2w+x_2} dx_3\, dx_2 &&= 2w^2/3 \\
E: &\quad \int_{2/3-w}^{1-2w} \int_{2-2x_2-w}^{2w+x_2} dx_3\, dx_2 &&= 1/6 - w + 3w^2/2 \\
F: &\quad \int_{1-2w}^{2/3-w/3} \int_{2-2x_2-w}^{1} dx_3\, dx_2 &&= -2/9 + 14w/9 - 20w^2/9 \\
G: &\quad \int_{w}^{2/3-w/3} \int_{2-2x_2-w}^{1} dx_3\, dx_2 &&= -2/9 + 14w/9 - 20w^2/9
\end{aligned}
$$

For $w \in [0, 1/5]$, the win probability is $2(2w/3 - 10w^2/3 + 2w^2 + 2w^2/3) = 4w/3 - 4w^2/3$.

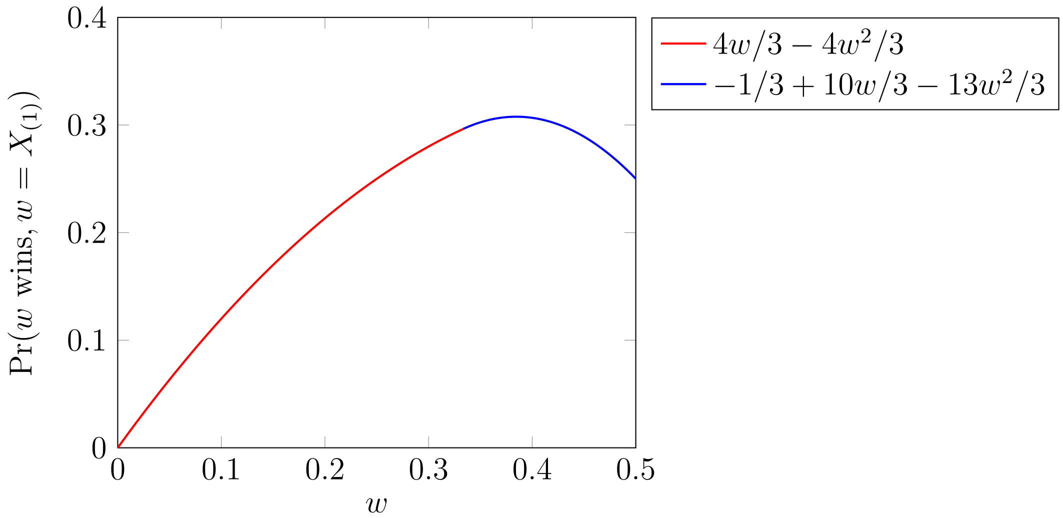
For $w \in [1/5, 1/3]$, the win probability is $2(1/6 - w + 3w^2/2 + -2/9 + 14w/9 - 20w^2/9 + 1/18 + w/9 + w^2/18) = 4w/3 - 4w^2/3$.

Finally, for $w \in [1/3, 1/2]$, the win probability is $2(-2/9 + 14w/9 - 20w^2/9 + 1/18 + w/9 + w^2/18) = -1/3 + 10w/3 - 13w^2/3$.

To summarize, the win probability is:

$$
\Pr(w \text{ wins}, w = X_{(1)}) = \begin{cases} 4w/3 - 4w^2/3, & w \in [0, 1/3] \\ -1/3 + 10w/3 - 13w^2/3, & w \in [1/3, 1/2] \end{cases} \tag{F.4}
$$

Visualizing this:



2. $w = X_{(2)}$. Consider the ordering $x_2 < w < x_3$ (we'll multiply by 2 later to account for $x_3 < w < x_2$). For w to beat x_2, we need:

$$\begin{aligned} & (w - x_2)/2 + (x_3 - w)/2 > x_2 + (w - x_2)/2 \\ \Leftrightarrow \quad & (x_3 - w)/2 > x_2 \\ \Leftrightarrow \quad & x_3 > 2x_2 + w \end{aligned} \tag{F.5}$$

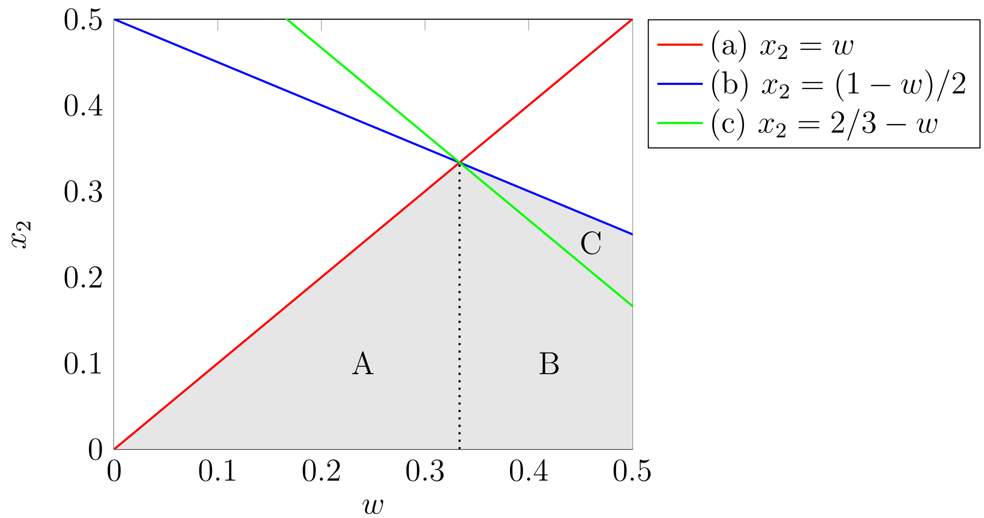
In order for this to be feasible, we need $2x_2 + w < 1 \Leftrightarrow x_2 < (1 - w)/2$. For w to beat x_3, we need:

$$\begin{aligned} & (w - x_2)/2 + (x_3 - w)/2 > 1 - x_3 + (x_3 - w)/2 \\ \Leftrightarrow \quad & (w - x_2)/2 > 1 - x_3 \\ \Leftrightarrow \quad & x_3 > 1 - (w - x_2)/2 \end{aligned} \tag{F.6}$$

These bounds are equal if

$$\begin{aligned}
& 2x_2 + w = 1 - (w - x_2)/2 \\
\Leftrightarrow\quad & 4x_2 + 2w = 2 - w + x_2 \\
\Leftrightarrow\quad & 3x_2 + 3w = 2 \\
\Leftrightarrow\quad & x_2 = 2/3 - w
\end{aligned}$$

Again, we can split the w–x_2 plane using this line to make integration easy. Note that the lines $x_2 = 2/3 - w$ and the line $x_2 = w$ intersect at 1/3.



Above line (c), the constraint $x_3 > 2x_2 + w$ dominates. Below line (c), $x_3 > 1 - (w - x_2)/2$ dominates. Integrating in the regions A–C:

$$\begin{aligned}
A:&\quad \int_0^w \int_{1-(w-x_2)/2}^1 dx_3\, dx_2 && = w^2/4 \\
B:&\quad \int_0^{2/3-w} \int_{1-(w-x_2)/2}^1 dx_3\, dx_2 && = -1/9 + 2w/3 - 3w^2/4 \\
C:&\quad \int_{2/3-w}^{(1-w)/2} \int_{2x_2+w}^1 dx_3\, dx_2 && = 1/36 - w/6 + w^2/4
\end{aligned}$$

Recall that we need to multiply by 2 to account for the ordering $x_3 < w < x_2$.

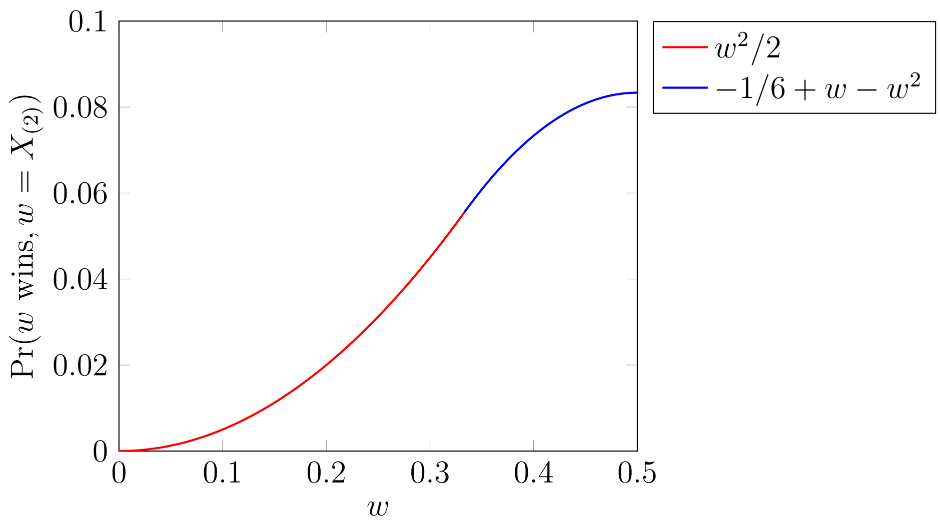
For $w \in [0, 1/3]$, the win probability is $2(w^2/4) = w^2/2$.

For $w \in [1/3, 1/2]$, the win probability is $2(-1/9 + 2w/3 - 3w^2/4 + 1/36 - w/6 + w^2/4) = -1/6 + w - w^2$.

To summarize:

$$\Pr(w \text{ wins}, w = X_{(2)}) = \begin{cases} w^2/2, & w \in [0, 1/3] \\ -1/6 + w - w^2, & w \in [1/3, 1/2] \end{cases} \tag{F.7}$$

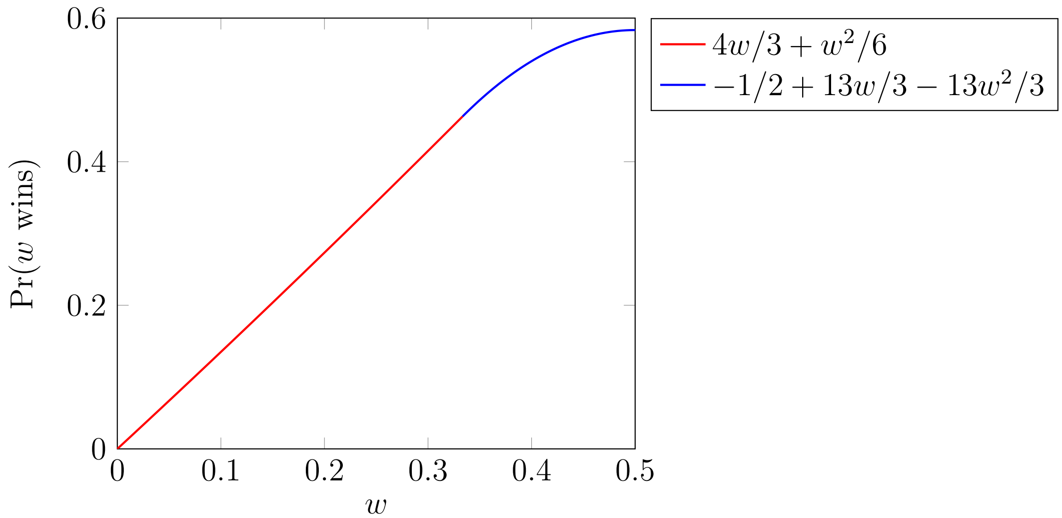
Visualizing this:



3. $w = X_{(3)}$. This means $x_2 < w$ and $x_3 < w$. Since $w \leq 0.5$, w always wins. We also have $\Pr(x_2 < w) = w$ and $\Pr(x_3 < w) = w$. Thus, $\Pr(w \text{ wins}, w = X_{(3)} \mid w \in [0, 0.5]) = w^2$.

Adding the three cases, we arrive at $\Pr(w \text{ wins})$. For $w \in [0, 1/3]$, the sum is $4w/3 - 4w^2/3 + w^2/2 + w^2 = 4w/3 + w^2/6$. For $w \in [1/3, 1/2]$, the sum is $-1/3 + 10w/3 - 13w^2/3 - 1/6 + w - w^2 + w^2 = -1/2 + 13w/3 - 13w^2/3$. Summarizing and plotting:

$$\Pr(w \text{ wins}) = \begin{cases} 4w/3 + w^2/6, & w \in [0, 1/3] \\ -1/2 + 13w/3 - 13w^2/3, & w \in [1/3, 1/2] \end{cases} \tag{F.8}$$



Scaling by 3, the variance of P_3 is then:

$$6\left[\int_0^{1/3} (w - 1/2)^2 (4w/3 + w^2/6)\, dw + \int_{1/3}^{1/2} (w - 1/2)^2 (-1/2 + 13w/3 - 13w^2/3)\, dw\right]$$

$$= 23/540 \approx 0.043$$

F.3.2 1d IRV winner distribution, $k = 3$

We'll perform the same type of analysis, but for IRV instead of plurality. In addition to breaking down cases by the order statistic of the winner, we'll also consider the IRV elimination order.

1. $w = X_{(1)}$. Consider the order $w < x_2 < x_3$ (we'll multiply by 2 later to account for the ordering $w < x_3 < x_2$).

(a) Candidate 2 is eliminated first. Since 2 has a smaller vote share than the winner,

$$\begin{aligned} & (x_2 - w)/2 + (x_3 - x_2)/2 < w + (x_2 - w)/2 \\ \Leftrightarrow \quad & (x_3 - x_2)/2 < w \\ \Leftrightarrow \quad & x_3 < 2w + x_2 \end{aligned} \tag{F.9}$$

Since 2 has a smaller vote share than 3,

$$\begin{aligned} & (x_2 - w)/2 + (x_3 - x_2)/2 < 1 - x_3 + (x_3 - x_2)/2 \\ \Leftrightarrow \quad & (x_2 - w)/2 < 1 - x_3 \\ \Leftrightarrow \quad & x_3 < 1 - (x_2 - w)/2. \end{aligned} \tag{F.10}$$

For this constraint to be feasible, we need

$$\begin{aligned} & x_2 < 1 - (x_2 - w)/2 \\ \Leftrightarrow \quad & 2x_2 < 2 - x_2 + w \\ \Leftrightarrow \quad & x_2 < 2/3 + w/3 \end{aligned}$$

The constraints are equal if

$$\begin{aligned} & 2w + x_2 = 1 - (x_2 - w)/2 \\ \Leftrightarrow \quad & 4w + 2x_2 = 2 - x_2 + w \\ \Leftrightarrow \quad & 3w + 3x_2 = 2 \\ \Leftrightarrow \quad & x_2 = 2/3 - w \end{aligned}$$

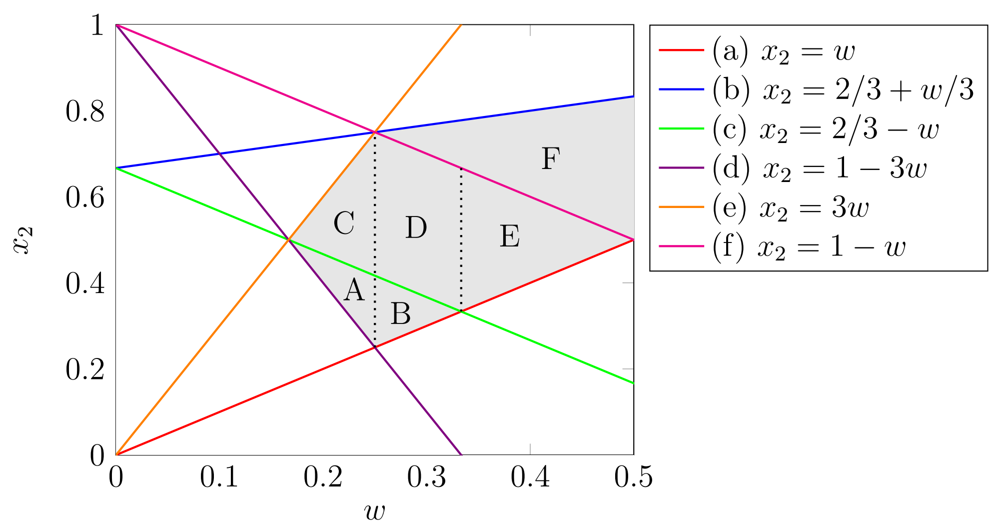
Once 2 is eliminated, whoever is closer to 1/2 is the winner. Thus, for w to win, we need $x_3 > 1 - w$. This constraint equals constraint F.9 if

$$\begin{aligned} & 2w + x_2 = 1 - w \\ \Leftrightarrow \quad & x_2 = 1 - 3w \end{aligned}$$

If w is too small (i.e., left of the line $x_2 = 1 - 3w$), then we cannot satisfy both $x_3 > 1 - w$ and $x_3 < 2w + x_2$. The constraint $x_3 > 1 - w$ equals constraint F.10 if

$$\begin{aligned} & 1 - (x_2 - w)/2 = 1 - w \\ \Leftrightarrow \quad & 2 - x_2 + w = 2 - 2w \\ \Leftrightarrow \quad & x_2 = 3w \end{aligned}$$

Again, if we are to the left of this line, we cannot satisfy both $x_3 > 1-w$ and $x_3 < 1-(x_2-w)/2$. Finally, the lower bound on x_3 is the maximum of $1 - w$ and x_2. Above the line $x_2 = 1 - w$, the lower bound x_2 dominates; below, $1 - w$ dominates.



Below line (c), constraint (F.9) dominates; above line (c), constraint (F.10) dominates. Below line (f), the lower bound on x_3 is $1-w$; above line (f), the lower bound is x_2. Lines (c), (d), (e) intersect at $w = 1/6$; lines (b), (e), (f) intersect at $w = 1/4$; lines (a) and (d) intersect at $w = 1/4$; lines (a) and (c) intersect at $w = 1/3$.

Integrating for each region and multiplying by 2 to account for the order $w < x_3 < x_2$:

$$\begin{aligned}
A: &\quad \int_{1-3w}^{2/3-w} \int_{1-w}^{2w+x_2} dx_3\, dx_2 &&= 1/18 - 2w/3 + 2w^2 \\
B: &\quad \int_{w}^{2/3-w} \int_{1-w}^{2w+x_2} dx_3\, dx_2 &&= -4/9 + 10w/3 - 6w^2 \\
C: &\quad \int_{2/3-w}^{3w} \int_{1-w}^{1-(x_2-w)/2} dx_3\, dx_2 &&= 1/9 - 4w/3 + 4w^2 \\
D: &\quad \int_{2/3-w}^{1-w} \int_{1-w}^{1-(x_2-w)/2} dx_3\, dx_2 &&= -5/36 + 2w/3 \\
E: &\quad \int_{w}^{1-w} \int_{1-w}^{1-(x_2-w)/2} dx_3\, dx_2 &&= -1/4 + 2w - 3w^2 \\
F: &\quad \int_{1-w}^{2/3+w/3} \int_{x_2}^{1-(x_2-w)/2} dx_3\, dx_2 &&= 1/12 - 2w/3 + 4w^2/3
\end{aligned}$$

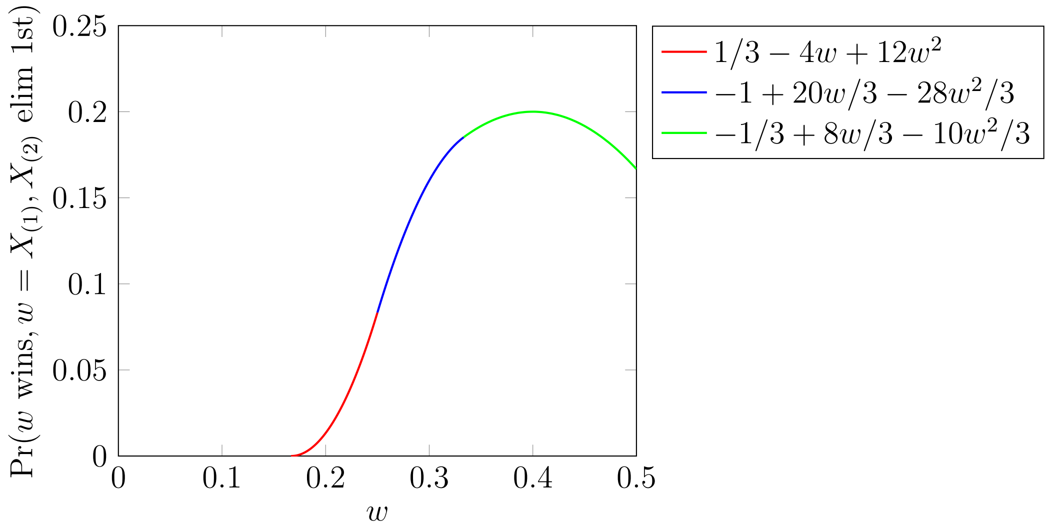
For $w \in [1/6, 1/4]$, the win probability is $2(1/18 - 2w/3 + 2w^2 + 1/9 - 4w/3 + 4w^2) = 1/3 - 4w + 12w^2$.

For $w \in [1/4, 1/3]$, the win probability is $2(-4/9 + 10w/3 - 6w^2 - 5/36 + 2w/3 + 1/12 - 2w/3 + 4w^2/3) = -1 + 20w/3 - 28w^2/3$.

For $w \in [1/3, 1/2]$, the win probability is $2(-1/4 + 2w - 3w^2 + 1/12 - 2w/3 + 4w^2/3) = -1/3 + 8w/3 - 10w^2/3$.

Summarizing and visualizing:

$$\Pr(w \text{ wins}, w = X_{(1)}, X_{(2)} \text{ elim 1st}) \\
= \begin{cases} 1/3 - 4w + 12w^2, & w \in [1/6, 1/4] \\ -1 + 20w/3 - 28w^2/3, & w \in [1/4, 1/3] \\ -1/3 + 8w/3 - 10w^2/3, & w \in [1/3, 1/2] \end{cases}$$



(b) Candidate 3 is eliminated first. Since candidate 3 has a smaller vote share than candidate 2:

$$
\begin{aligned}
& (x_3 - x_2)/2 + 1 - x_3 < (x_3 - x_2)/2 + (x_2 - w)/2 \\
\Leftrightarrow\quad & 1 - x_3 < (x_2 - w)/2 \\
\Leftrightarrow\quad & x_3 > 1 - (x_2 - w)/2
\end{aligned}
$$

Since candidate 3 has a smaller vote share than the winner:

$$
\begin{aligned}
& (x_3 - x_2)/2 + 1 - x_3 < w + (x_2 - w)/2 \\
\Leftrightarrow\quad & x_3 - x_2 + 2 - 2x_3 < 2w + x_2 - w \\
\Leftrightarrow\quad & -x_3 + 2 < w + 2x_2 \\
\Leftrightarrow\quad & x_3 > 2 - w - 2x_2
\end{aligned}
$$

For this to be feasible, we need:

$$
\begin{aligned}
& 2 - w - 2x_2 < 1 \\
\Leftrightarrow\quad & 1 - w < 2x_2 \\
\Leftrightarrow\quad & x_2 > (1 - w)/2
\end{aligned}
$$

The constraints $x_3 > 1 - (x_2 - w)/2$ and $x_3 > 2 - w - 2x_2$ are equal when:

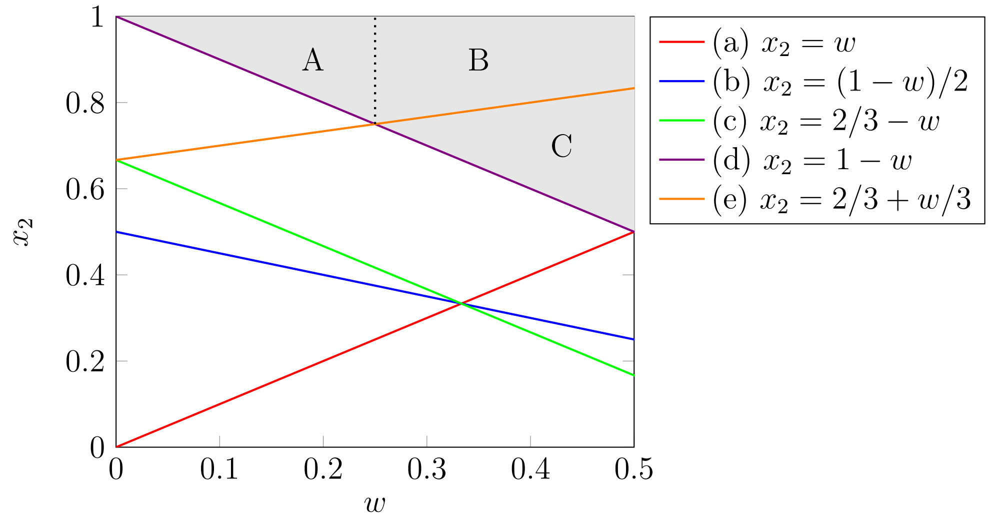
$$
\begin{aligned}
& 1 - (x_2 - w)/2 = 2 - w - 2x_2 \\
\Leftrightarrow \quad & 2 - x_2 + w = 4 - 2w - 4x_2 \\
\Leftrightarrow \quad & 3x_2 = 2 - 3w \\
\Leftrightarrow \quad & x_2 = 2/3 - w
\end{aligned}
$$

Above the line $x_2 = 2/3 - w$, the constraint $x_3 > 1 - (x_2 - w)/2$ dominates; below, $x_3 > 2 - w - 2x_2$ dominates.

In order for w to win, it must be closer to the center than x_2. This requires that $x_2 > 1 - w$ (since $w < 0.5$ and $w < x_2$). Thus, we never need to worry about the constraint $x_3 < 2 - w - 2x_2$ (since the line $x_2 = 1 - w$ is above the line $2/3 - w$). The lower bound on x_3 is thus the maximum of x_2 and $1 - (x_2 - w)/2$. These are equal if

$$
\begin{aligned}
& x_2 = 1 - (x_2 - w)/2 \\
\Leftrightarrow \quad & 2x_2 = 2 - x_2 + w \\
\Leftrightarrow \quad & 3x_2 = 2 + w \\
\Leftrightarrow \quad & x_2 = 2/3 + w/3
\end{aligned}
$$

Above the line $x_2 = 2/3 + w/3$, the lower bound on x_3 is x_2; below, it's $1 - (x_2 - w)/2$.



Lines (d) and (e) intersect at $w = 1/4$. Integrating over the regions:

$$A : \int_{1-w}^{1} \int_{x_2}^{1} dx_3\, dx_2 = w^2/2 \tag{F.11}$$

$$B : \int_{2/3+w/3}^{1} \int_{x_2}^{1} dx_3\, dx_2 = 1/18 - w/9 + w^2/18 \tag{F.12}$$

$$C : \int_{1-w}^{2/3+w/3} \int_{1-(x_2-w)/2}^{1} dx_3\, dx_2 = -5/36 + 7w/9 - 8w^2/9 \tag{F.13}$$

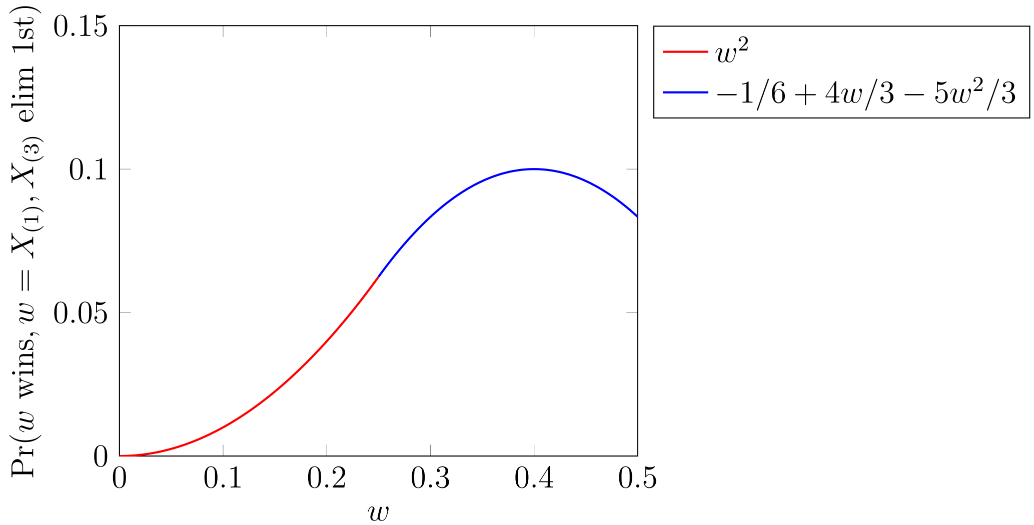
Multiplying by 2 to account for the ordering $w < x_3 < x_2$:

For $w \in [0, 1/4]$, the win probability is $2(w^2/2) = w^2$.

For $w \in [1/4, 1/2]$, the win probability is $2(1/18 - w/9 + w^2/18 - 5/36 + 7w/9 - 8w^2/9) = -1/6 + 4w/3 - 5w^2/3$.

$$\Pr(w \text{ wins}, w = X_{(1)}, X_{(3)} \text{ elim 1st}) = \begin{cases} w^2, & w \in [0, 1/4] \\ -1/6 + 4w/3 - 5w^2/3, & w \in [1/4, 1/2] \end{cases} \tag{F.14}$$

Plotting:



2. $w = X_{(2)}$. Consider the order $x_2 < w < x_3$ (we'll multiply by 2 later to account for $x_3 < w < x_2$).

   (a) Candidate 2 is eliminated first. Since candidate 2 has a smaller vote share than the winner,

$$\begin{aligned} & x_2 + (w - x_2)/2 < (x_3 - w)/2 + (w - x_2)/2 \\ \Leftrightarrow \quad & x_2 < (x_3 - w)/2 \\ \Leftrightarrow \quad & 2x_2 < x_3 - w \\ \Leftrightarrow \quad & x_3 > 2x_2 + w. \end{aligned}$$

   For this to be feasible, we need

$$\begin{aligned} & 2x_2 + w < 1 \\ \Leftrightarrow \quad & x_2 < (1 - w)/2. \end{aligned}$$

Since candidate 2 has a smaller vote share than candidate 3,

$$\begin{aligned} & x_2 + (w - x_2)/2 < 1 - x_3 + (x_3 - w)/2 \\ \Leftrightarrow\quad & 2x_2 + w - x_2 < 2 - 2x_3 + x_3 - w \\ \Leftrightarrow\quad & x_2 + 2w < 2 - x_3 \\ \Leftrightarrow\quad & x_3 < 2 - x_2 - 2w. \end{aligned}$$

For this to be feasible, we need

$$\begin{aligned} & 2 - x_2 - 2w > w \\ \Leftrightarrow\quad & x_2 < 2 - 3w. \end{aligned}$$

Since $w \leq 0.5$, this is always satisfied. The upper bound on x_3 is the minimum of 1 and $2 - x_2 - 2w$. There are equal if

$$\begin{aligned} & 1 = 2 - x_2 - 2w \\ \Leftrightarrow\quad & x_2 = 1 - 2w. \end{aligned}$$

To the left of the line $x_2 = 1 - 2w$, the upper bound on x_3 is 1; to the right, it's $2 - x_2 - 2w$.

The upper and lower bounds on x_3 are equal when

$$\begin{aligned} & 2x_2 + w = 2 - x_2 - 2w \\ \Leftrightarrow\quad & 3x_2 = 2 - 3w \\ \Leftrightarrow\quad & x_2 = 2/3 - w. \end{aligned}$$

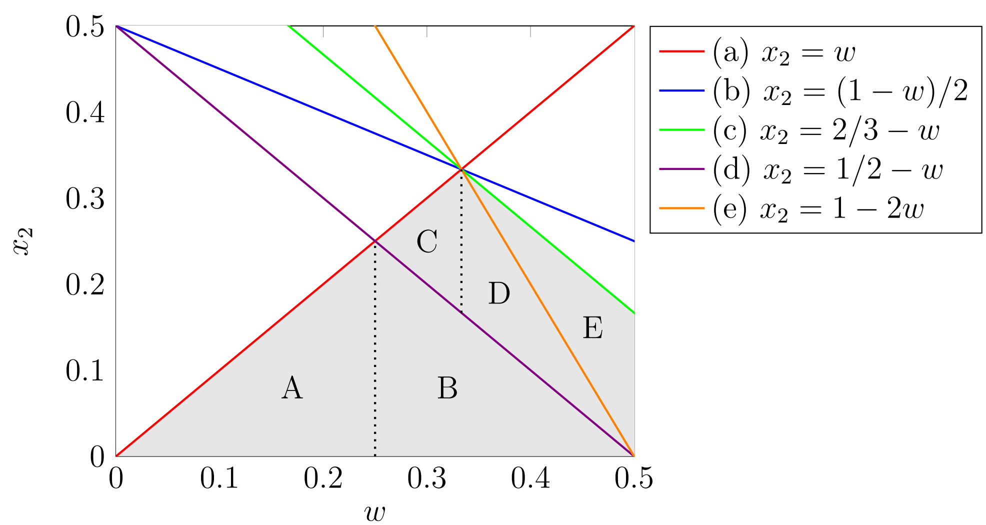
For both constrains to be feasible, we need to be to the left of the line $x_2 = 2/3 - w$. In order for w to win, it needs to be closer to the center than candidate 3. That is, we need $x_3 > 1 - w$. This equals the upper bound constraint on x_3 if

$$1 - w = 2 - x_2 - 2w$$
$$\Leftrightarrow \quad x_2 = 1 - w$$

Since we already need to be left of the line $2/3 - w$, we don't need to worry about being to the left of $1 - w$. Finally, the two lower bounds on x_3 are equal if

$$1 - w = 2x_2 + w$$
$$\Leftrightarrow \quad x_2 = 1/2 - w.$$

To the left of the line $x_2 = 1/2 - w$, the lower bound on x_3 is $1 - w$; to the right, the lower bound is $2x_2 + w$.



Lines (a) and (d) intersect at $w = 1/4$; lines (a) and (c) intersect at $w = 1/3$. Integrating over the regions:

$$
\begin{aligned}
A: &\quad \int_0^{w}\int_{1-w}^{1} dx_3\,dx_2 && = w^2\\
B: &\quad \int_0^{1/2-w}\int_{1-w}^{1} dx_3\,dx_2 && = w/2-w^2\\
C: &\quad \int_{1/2-w}^{w}\int_{2x_2+w}^{1} dx_3\,dx_2 && = -1/4+3w/2-2w^2\\
D: &\quad \int_{1/2-w}^{1-2w}\int_{2x_2+w}^{1} dx_3\,dx_2 && = -1/4+3w/2-2w^2\\
E: &\quad \int_{1-2w}^{2/3-w}\int_{2x_2+w}^{2-x_2-2w} dx_3\,dx_2 && = 1/6-w+3w^2/2
\end{aligned}
$$

We now sum and multiply by 2 to account for the ordering $x_3 < w < x_2$.

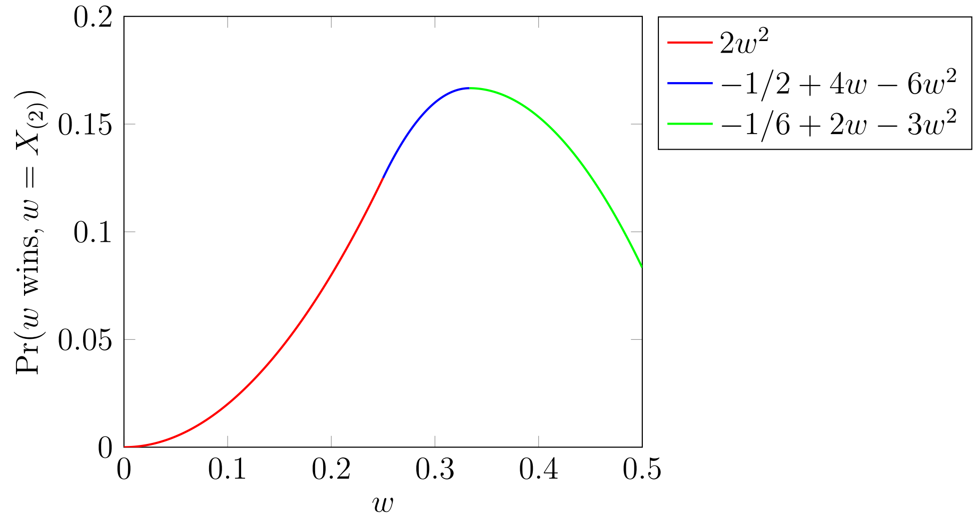
For $x \in [0, 1/4]$, the win probability is $2w^2$.

For $x \in [1/4, 1/3]$, the win probability is $2(w/2 - w^2 - 1/4 + 3w/2 - 2w^2) = -1/2 + 4w - 6w^2$.

For $x \in [1/3, 1/2]$, the win probability is $2(w/2 - w^2 - 1/4 + 3w/2 - 2w^2 + 1/6 - w + 3w^2/2) = -1/6 + 2w - 3w^2$.

Summarizing and visualizing:

$$
\Pr(w \text{ wins}, w = X_{(2)}, X_{(1)} \text{ elim 1st}) = \begin{cases} 2w^2, & w \in [0, 1/4] \\ -1/2 + 4w - 6w^2, & w \in [1/4, 1/3] \\ -1/6 + 2w - 3w^2, & w \in [1/3, 1/2] \end{cases} \tag{F.15}
$$



(b) Candidate 3 is eliminated first. Since candidate 3 has a smaller vote share than the winner,

$$
\begin{aligned}
& 1 - x_3 + (x_3 - w)/2 < (x_3 - w)/2 + (w - x_2)/2 \\
\Leftrightarrow\quad & 1 - x_3 < (w - x_2)/2 \\
\Leftrightarrow\quad & x_3 > 1 - (w - x_2)/2
\end{aligned}
$$

This is feasible if

$$
\begin{aligned}
& 1 - (w - x_2)/2 < 1 \\
\Leftrightarrow\quad & -w + x_2 < 0 \\
\Leftrightarrow\quad & x_2 < w,
\end{aligned}
$$

which is always true. Since candidate 3 has a smaller vote share than candidate 2,

$$
\begin{aligned}
& 1 - x_3 + (x_3 - w)/2 < x_2 + (w - x_2)/2 \\
\Leftrightarrow\quad & 2 - 2x_3 + x_3 - w < 2x_2 + w - x_2 \\
\Leftrightarrow\quad & 2 - x_3 < x_2 + 2w \\
\Leftrightarrow\quad & x_3 > 2 - x_2 - 2w
\end{aligned}
$$

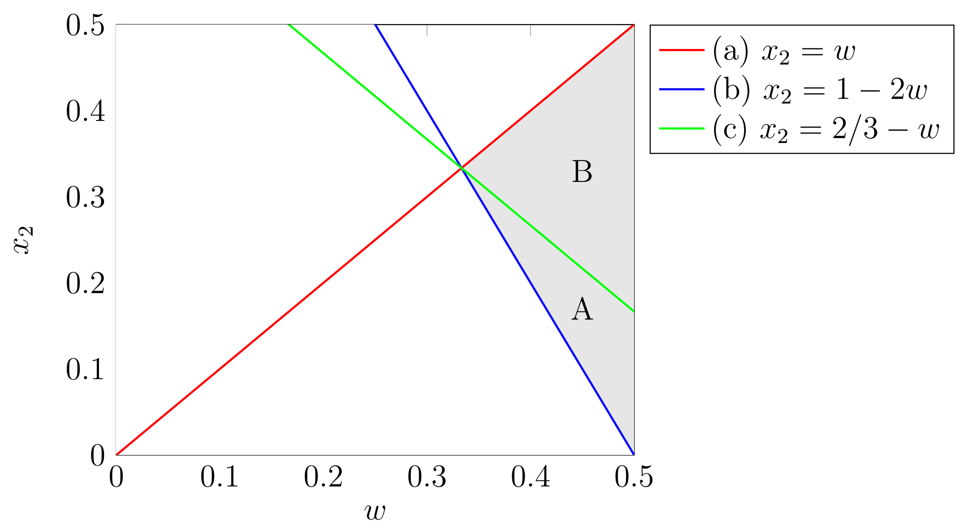
This is feasible if

$$
\begin{aligned}
& 2 - x_2 - 2w < 1 \\
\Leftrightarrow \quad & x_2 > 1 - 2w.
\end{aligned}
$$

The two lower bounds on x_3 are equal if

$$
\begin{aligned}
& 1 - (w - x_2)/2 = 2 - x_2 - 2w \\
\Leftrightarrow \quad & 2 - w + x_2 = 4 - 2x_2 - 4w \\
\Leftrightarrow \quad & 3x_2 = 2 - 3w \\
\Leftrightarrow \quad & x_2 = 2/3 - w
\end{aligned}
$$

Above the line $2/3 - w$, the lower bound on x_3 is $1 - (w - x_2)/2$; below the line, it's $2 - x_2 - 2w$. As long as candidate 3 is eliminated first, w wins since it's closer to the center than x_2.



Lines (a), (b), and (c) all intersect at $w = 1/3$. Integrating over the two regions and multiplying by 2 to account for the ordering $x_3 < w < x_2$:

$$A: \quad \int_{1-2w}^{2/3-w} \int_{2-x_2-2w}^{1} dx_3\, dx_2 \qquad = 1/18 - w/3 + w^2/2$$

$$B: \quad \int_{2/3-w}^{w} \int_{1-(w-x_2)/2}^{1} dx_3\, dx_2 \qquad = 1/9 - 2w/3 + w^2$$

For $w \in [1/3, 1/2]$, the win probability is $2(1/18 - w/3 + w^2/2 + 1/9 - 2w/3 + w^2) = 1/3 - 2w + 3w^2$. Thus:

$$\Pr(w \text{ wins}, w = X_{(2)}, X_{(3)} \text{ elim 1st}) = 1/3 - 2w + 3w^2, \qquad w \in [1/3, 1/2] \tag{F.16}$$

3. $w = X_{(3)}$. If both $x_2 < w$ and $x_3 < w$, then w wins by IRV. Thus, $\Pr(w \text{ wins}, w = X_{(3)}) = w^2$ for $w \in [0, 0.5]$.

We can finally sum over the three cases to arrive at $\Pr(w \text{ wins})$.

For $w \in [0, 1/6]$, the sum is $w^2 + 2w^2 + w^2 = 4w^2$.

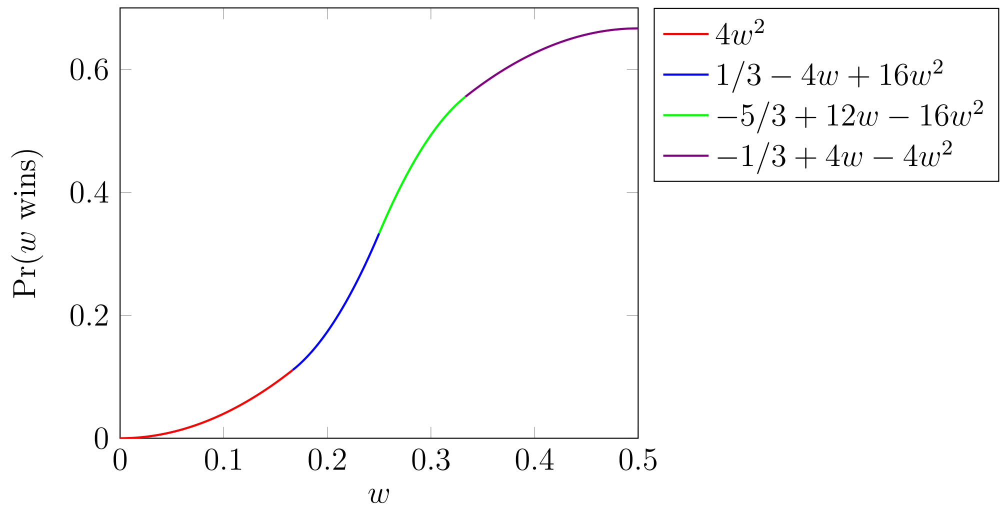
For $w \in [1/6, 1/4]$, the sum is $w^2 + 2w^2 + w^2 + 1/3 - 4w + 12w^2 = 1/3 - 4w + 16w^2$.

For $w \in [1/4, 1/3]$, the sum is $w^2 - 1/2 + 4w - 6w^2 - 1/6 + 4w/3 - 5w^2/3 - 1 + 20w/3 - 28w^2/3 = -5/3 + 12w - 16w^2$.

For $w \in [1/3, 1/2]$, the sum is $w^2 + 1/3 - 2w + 3w^2 - 1/6 + 2w - 3w^2 - 1/6 + 4w/3 - 5w^2/3 - 1/3 + 8w/3 - 10w^2/3 = -1/3 + 4w - 4w^2$.

Summarizing and plotting:

$$\Pr(w \text{ wins}) = \begin{cases} 4w^2, & w \in [0, 1/6] \\ 1/3 - 4w + 16w^2, & w \in [1/6, 1/4] \\ -5/3 + 12w - 16w^2, & w \in [1/4, 1/3] \\ -1/3 + 4w - 4w^2, & w \in [1/3, 1/2] \end{cases} \tag{F.17}$$



To get the winner position distribution f_{R_3}, we scale by three. The variance of f_{R_3} is thus:

$$\begin{aligned} &6\Bigg[\int_0^{1/6} (w-1/2)^2(4w^2)\,dw + \int_{1/6}^{1/4} (w-1/2)^2(1/3-4w+16w^2)\,dw \\ &\quad + \int_{1/4}^{1/3} (w-1/2)^2(-5/3+12w-16w^2)\,dw + \int_{1/3}^{1/2} (w-1/2)^2(-1/3+4w-4w^2)\,dw\Bigg] \\ &\quad = 25/864 \approx 0.029 \end{aligned} \tag{F.18}$$

Table F.1: Quotes for and against a moderating effect of IRV

[Under ranked-choice voting,] voters can support their favorites while still voting effectively against their least favorite. Having more competition encourages better dialogue on issues. Civility is substantially improved. Needing to reach out to more voters leads candidates to reduce personal attacks and govern more inclusively. —Howard Dean, former Governor of Vermont (Dean, 2016)
We need an electoral system that breaks the current stranglehold of the two-party monopoly, one that would allow voters to choose between a much more nuanced range of positions than "extreme" versus "moderate," would allow third-party candidates to run without being spoilers and would encourage more civil campaigning and political discourse. The solution is to adopt ranked-choice voting for all state and federal elections [....] We need [...] reforms that will allow the American people to reassert our power over a party system that is badly broken and compel candidates to appeal to a far broader swathe of us than a narrow "base." —Anne-Marie Slaughter, CEO of New America (Slaughter, 2019)
Quite to the contrary, [ranked-choice voting] may give life to more strident candidates, hoping to siphon first-place ballots from extreme voters who will give second preference to whichever major party is closest to them. This could result in more comity between the major-party candidates, as fringier competitors blot the airwaves with attacks. Or it might produce strategic coalitions sniping at each other, leaving us effectively back where we started. —Simon Waxman, former managing editor at Boston Review (Waxman, 2016)
However, ranked-choice voting makes it more difficult to elect moderate candidates when the electorate is polarized. For example, in a three-person race, the moderate candidate may be preferred to each of the more extreme candidates by a majority of voters. However, voters with far-left and far-right views will rank the candidate in second place rather than in first place. Since ranked-choice voting counts only the number of first-choice votes (among the remaining candidates), the moderate candidate would be eliminated in the first round, leaving one of the extreme candidates to be declared the winner. [...] The ranked-choice system that is being used around the country to conduct elections with more than two candidates is biased towards extreme candidates and away from moderate ones. —Nathan Atkinson, Assistant Professor at University of Wisconsin Law School, and Scott C. Ganz, Associate Teaching Professor at Georgetown University) (Atkinson and Ganz, 2022)

F.4 Additional figures

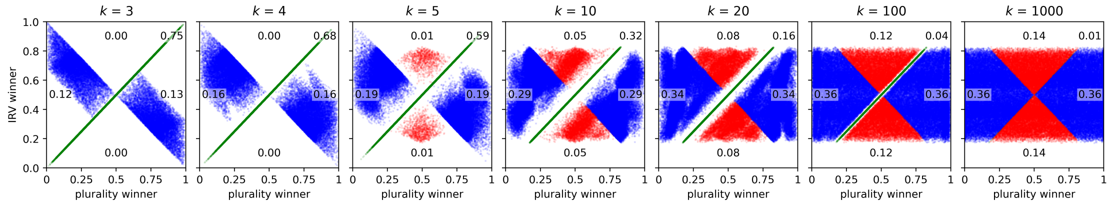


Figure F.1: Plurality vs. IRV winner positions in 100,000 simulation trials for increasing candidate count k (with uniform voters and candidates). Blue points are trials where the IRV winner was more moderate than the plurality winner, while red points are trials where the plurality winner was more moderate. Green points are trials where the winners were identical. Numbers in each quadrant show the proportion of trials falling in that region (the top right number is the proportion of same-winner trials). Notice that cases where the IRV winner is more extreme only appear beginning at $k = 5$, in accordance with Theorem 31. Note the probabilistic moderating effect of IRV compared to plurality: IRV does not elect extreme candidates as k grows large, but plurality does.

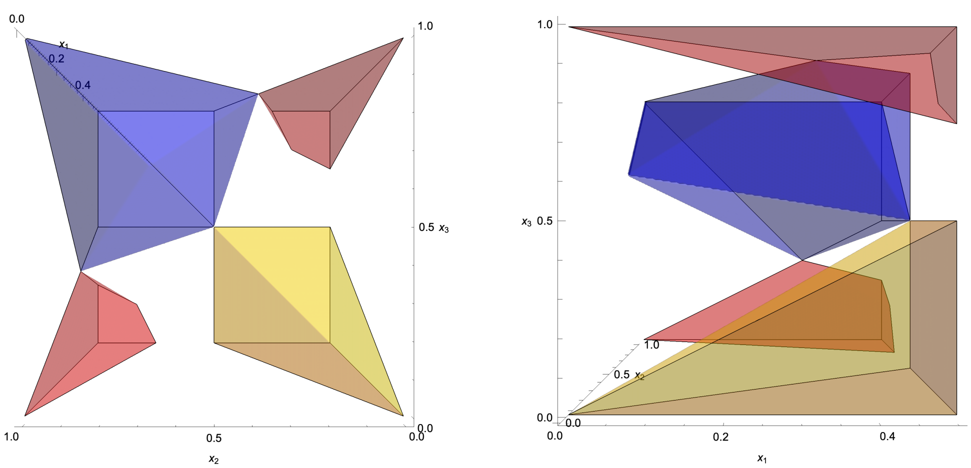


Figure F.2: Plurality winner regions for $k = 3$ with uniform voters and candidates. Colored polyhedra show the regions where a candidate at position x_1 is the plurality winner against candidates at x_2 and x_3. Regions are only shown for $x_1 \leq 0.5$, since the other half of is symmetric. The color of a region corresponds to the order statistic of the winner. Blue: winner is the leftmost, red: winner is in the middle, yellow: winner is the rightmost. The left view has the plane of the page at $x_1 = 0$, looking towards increasing x_1. The right view has the plane of the page at $x_2 = 0$, with x_1 increasing from left to right.

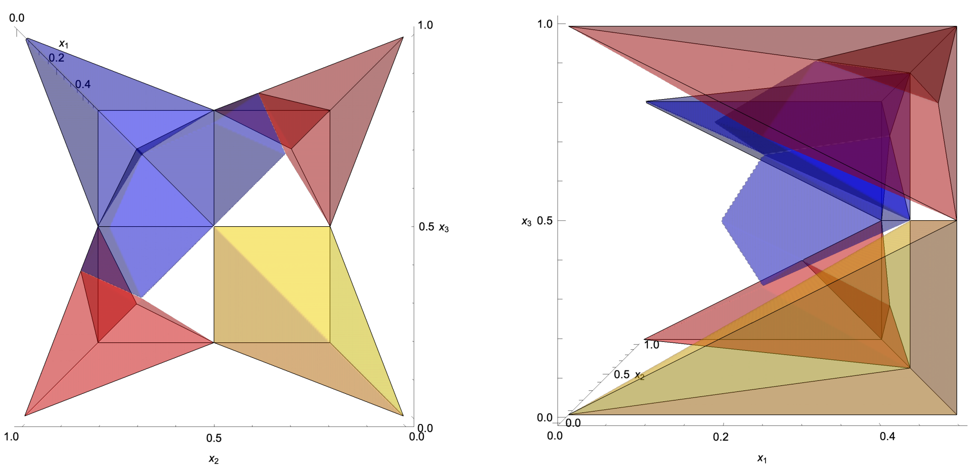


Figure F.3: IRV winner regions for $k = 3$ with uniform voters and candidates. See Figure F.2 for details about the visualization.

APPENDIX G

TECHNICAL DETAILS FOR CHAPTER 8: REPLICATING ELECTORAL SUCCESS

"What ho!" I said.

"What ho!" said Motty.

"What ho! What ho!"

"What ho! What ho! What ho!"

After that it seemed rather difficult to go on with the conversation.

P. G. Wodehouse, *My Man Jeeves*, 1919

(Bertie Wooster, an expert in applied bounded rationality.)

G.1 Additional plots

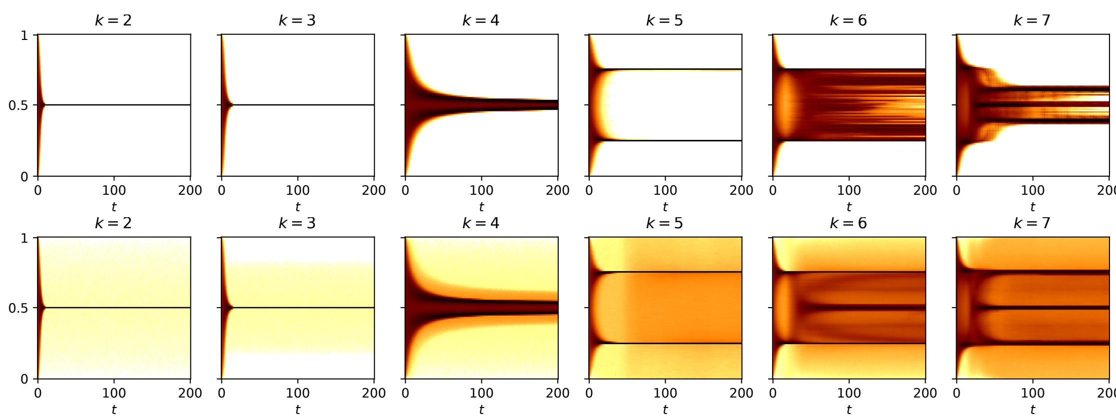


Figure G.1: Replicator dynamics runs just as in Figure 8.3, but without enhanced symmetry. For $k > 6$, the behavior of the Monte Carlo trials becomes inconsistent without enhanced symmetry, particularly without ϵ-uniform noise. See Figure 8.3 for more details.

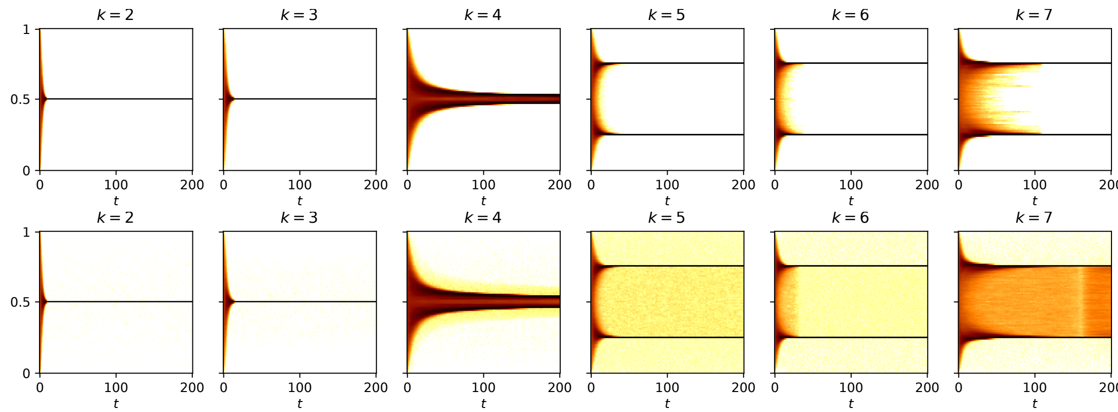


Figure G.2: Replicator dynamics runs with enhanced symmetry just as in Figure 8.3, but showing only a single trial instead of aggregating 50 runs. With enhanced symmetry, the behavior is very consistent across runs.

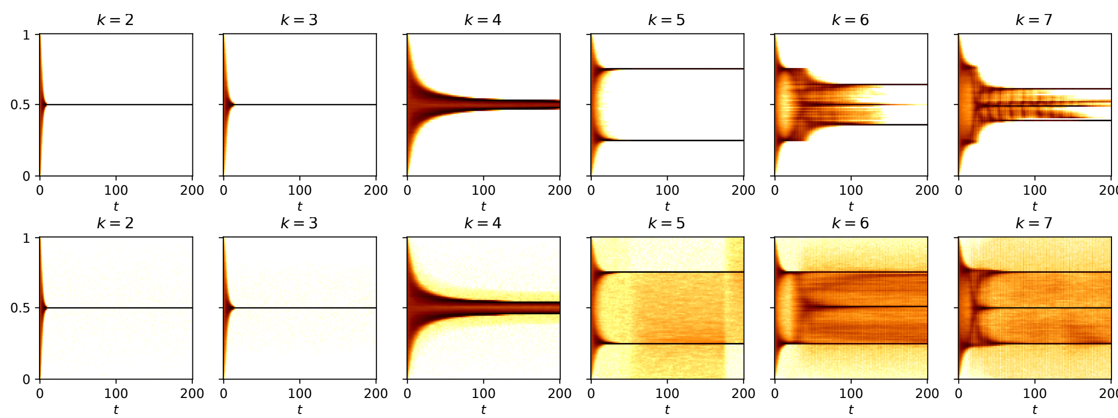


Figure G.3: Replicator dynamics runs just as in Figure G.1 (no enhanced symmetry), but showing only a single trial instead of aggregating 50 runs to highlight the inconsistent behavior for $k = 6$ and 7 without enhanced symmetry.

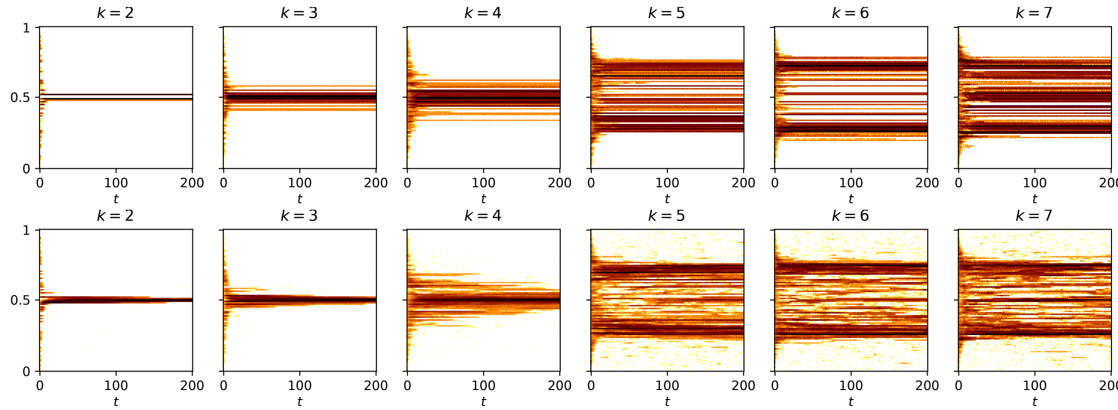


Figure G.4: Replicator dynamics runs with only 50 elections per generation, without enhanced symmetry. Each plot shows 50 trials. The top row has no noise, while the bottom row uses 0.01-uniform noise. Even with a small sample size, our main finding holds.

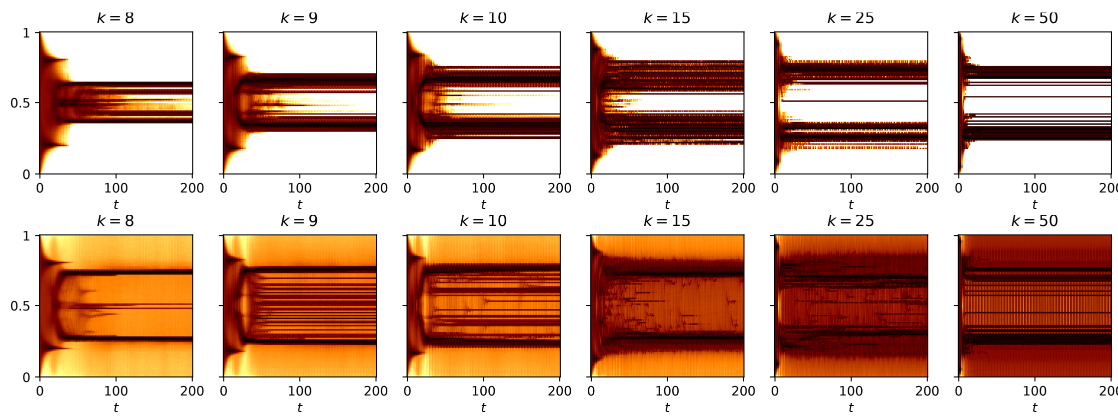


Figure G.5: Replicator dynamics runs just as in Figure 8.4, but without enhanced symmetry. As with smaller values of k, the behavior becomes more chaotic without enhanced symmetry.

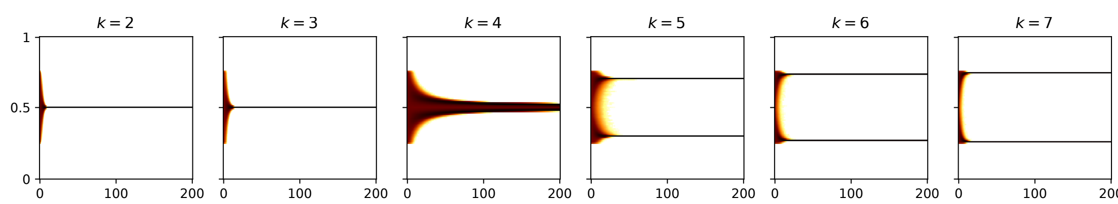


Figure G.6: Replicator dynamics with initial candidate distribution Uniform$(1/4, 3/4)$. These plots show 50 trials with 100,000 elections per generation, no noise, and without enhanced symmetry. The dynamics are very well-behaved with $(1/4, 3/4)$ support, removing the need for enhanced symmetry; compare to Figure G.1.

G.1.1 Additional variant plots

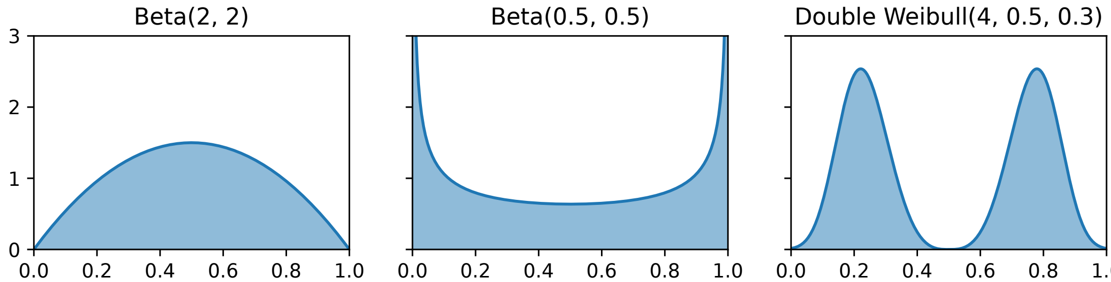


Figure G.7: PDFs of different voter distributions used in Figure 8.7.

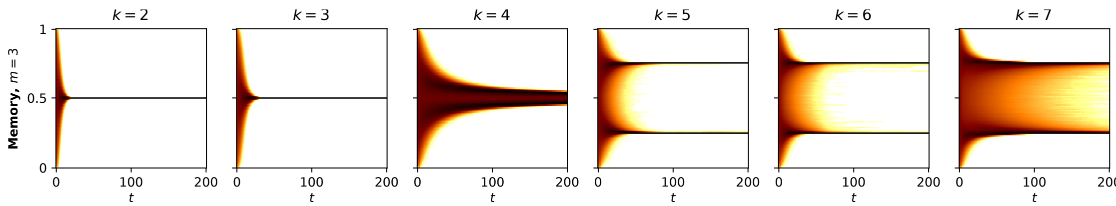


Figure G.8: Replicator dynamics with $m = 3$ generations of memory, no enhanced symmetry, and 50 trials per plot. There is no qualitative difference between $m = 3$ and $m = 2$ (compare to Figure 8.7).

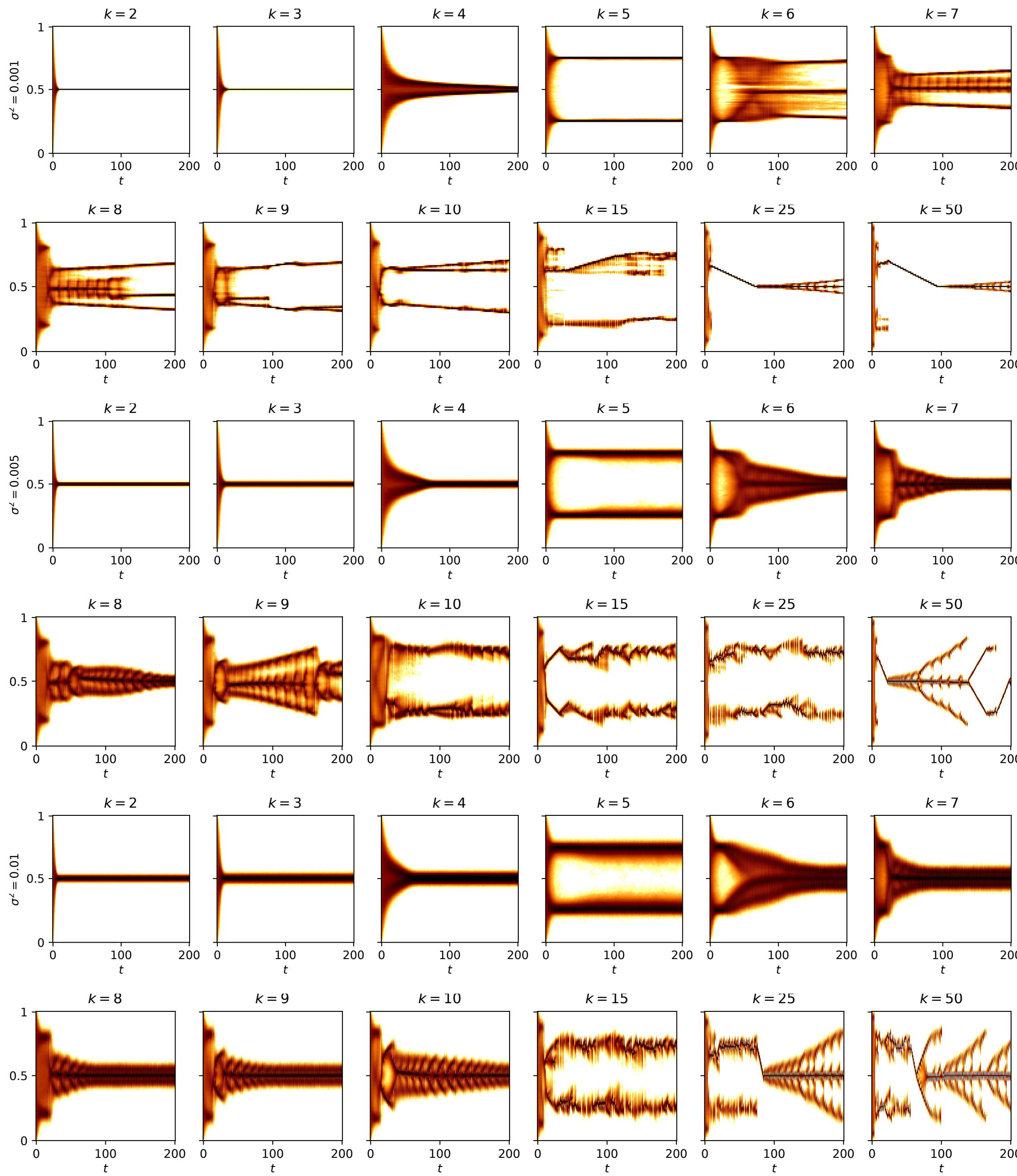


Figure G.9: Single trials of the replicator dynamics with perturbation noise and 100,000 elections per generation. The first two rows use $\sigma^2 = 0.001$, the middle two use $\sigma^2 = 0.005$, and the bottom two use $\sigma^2 = 0.01$. Perturbation noise combined with Monte-Carlo asymmetries can result in complex and unpredictable branching with higher k.

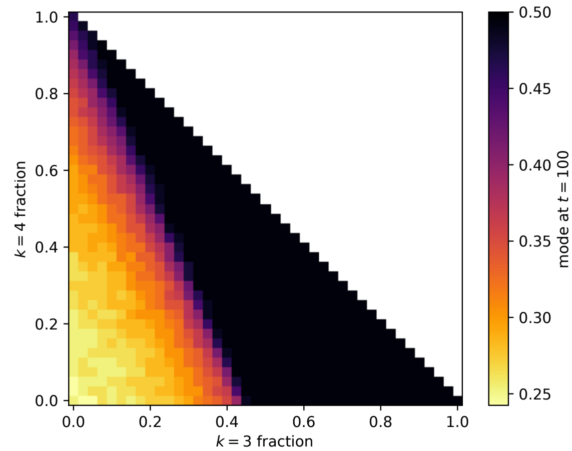


Figure G.10: Heatmap showing the position of the candidate distribution mode at $t = 100$ when elections have a mixture of $k = 3, 4$, and 5 candidates each (only modes $\leq 1/2$ are shown). These simulations use 100,000 elections per generation, with k split between $3, 4$, and 5 in different proportions at each point. The fraction of elections with 3 candidates varies along the x axis, while the fraction with 4 candidates varies along the y axis. Any remaining elections have $k = 5$. For instance, the lower left corner has all 100,000 elections use $k = 5$, while the point $(1/3, 1/3)$ has an even mix of candidate counts. When either the $k = 3$ or $k = 4$ fraction is high enough (but especially $k = 3$), the distribution converges to the center, with the mode at $1/2$. However, with enough $k = 5$ elections, two clusters emerge, and more $k = 5$ elections pushes them farther apart.

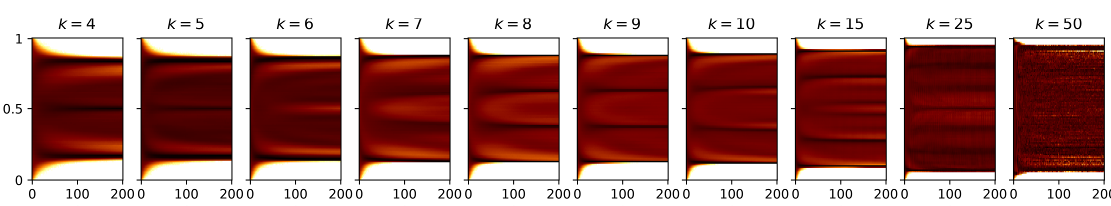


Figure G.11: Replicator dynamics with top-h copying where $h = 3$, no enhanced symmetry, 50 trials per plot, and 100,000 elections per generation.

G.2 Additional proofs

G.2.1 Proofs from Section 8.2

Theorem 34. *Let $F_0 \in \mathcal{F}$. For all $x < 1/2$ and $t > 0$,*

$$F_{3,t}(x) \leq 3/4 \cdot F_{3,t-1}(x) + F_{3,t-1}(x)^3. \tag{8.3}$$

This can be written as a looser closed form

$$F_{3,t}(x) \leq F_0(x) \cdot \left[3/4 + F_0(x)^2\right]^t. \tag{8.4}$$

Proof. Let $x < 1/2$ and define $p = F_{3,t-1}(x)$. Consider the following cases for the positions of the three candidates $X_{1,t}, X_{2,t}$, and $X_{3,t}$. Call candidates in $(x, 1-x)$ *inner.*

1. All three candidates in $[0, 1/2)$ (and the symmetric case). First suppose all three are in $[0, 1/2)$ (the other side is symmetric). If there is at least one inner candidate (w.p. $1/2^3 - p^3$), then the winner is inner. Accounting for symmetry, an inner candidate wins in this case w.p. $2(1/2^3 - p^3) = 1/4 - 2p^3$.
2. Two candidates in $(x, 1/2)$ and one in $(1/2, 1-x)$ (and the symmetric case). Since all candidates are inner, an inner candidate wins. Accounting for symmetry, an inner candidate wins in this case w.p. $2\left[3(1/2-p)^3\right] = 6(1/2-p)^3$.
3. Two candidates in $[0, x)$ and one in $(1/2, 1-x)$ (and the symmetric case). The candidate in $(1/2, 1-x)$ wins with vote share at least $1/2$. Accounting for symmetry, an inner candidate wins in this case w.p. $2\left[3p^2(1/2-p)\right] = 6p^2(1/2-p)$.

4. One candidate in $[0, x)$, one in $(x, 1/2)$, and one in $(1/2, 1-x)$. Label them 1, 2, and 3, respectively. Candidate 3 gets vote share $1 - (X_3 + X_2)/2 = [(1 - X_3) + (1 - X_2)]/2$, while candidate 1 gets vote share $(X_1 + X_2)/2$. Since $X_3 < 1 - x$, $1 - X_3 > X_1$; and since $X_2 < 1/2$, $1 - X_2 > X_2$. Thus candidate 3 has higher vote share than candidate 1 and an inner candidate wins. Accounting for symmetry, an inner candidate wins in this case w.p. $2\left[3 \cdot 2p(1/2 - p)^2\right] = 12p(1/2 - p)^2$.

Adding up these cases yields a lower bound on the probability that an inner candidate wins:

$$\begin{aligned}
&\Pr(x < \text{Plurality}(X_{1,t}, X_{2,t}, X_{3,t}) < 1 - x) \\
&\geq 1/4 - 2p^3 + 6(1/2 - p)^3 + 6p^2(1/2 - p) + 12p(1/2 - p)^2 \\
&= 1 - 3/2 \cdot p - 2p^3.
\end{aligned}$$

By symmetry, this yields the claimed upper bound on the probability a candidate in $[0, x]$ wins:

$$\begin{aligned}
F_{3,t}(x) &= \Pr(\text{Plurality}(X_{1,t}, X_{2,t}, X_{3,t}) \leq x) \\
&= (1 - \Pr(x < \text{Plurality}(X_{1,t}, X_{2,t}, X_{3,t}) < 1 - x))/2 \\
&\leq \left[1 - (1 - 3/2 \cdot p - 2p^3)\right]/2 \\
&= 3/4 \cdot p + p^3 \\
&= 3/4 \cdot F_{3,t-1}(x) + F_{3,t-1}(x)^3.
\end{aligned}$$

We now show the closed form bound by induction on t. We'll simultaneously show that $F_{3,t}(x) \leq F_0(x)$. For the base case $t = 0$, we have $F_{3,t}(x) \leq F_0(x)$. Now for $t > 0$, suppose the claims hold for $t-1$. Using the bound above, we know that

$$\begin{aligned}
F_{3,t}(x) &\leq 3/4 \cdot F_{3,t-1}(x) + F_{3,t-1}(x)^3 \\
&= F_{3,t-1}(x) \cdot \left[3/4 + F_{3,t-1}(x)^2\right] \\
&\leq F_{3,t-1}(x) \cdot \left[3/4 + F_0(x)^2\right] && \text{(by IH)} \\
&\leq F_0(x) \cdot \left[3/4 + F_0(x)^2\right]^{t-1} \cdot \left[3/4 + F_0(x)^2\right] && \text{(by IH)} \\
&= F_0(x) \cdot \left[3/4 + F_0(x)^2\right]^t
\end{aligned}$$

This is the main claim we wanted to show. We can now also show the supporting fact that $F_{3,t}(x) \leq F_0(x)$. For $x < 1/2$, $F_0(x) \leq 1/2$ by symmetry. Thus $3/4 + F_0(x)^2 \leq 3/4 + 1/2^2 = 1$, so by the inequality above, $F_{3,t}(x) \leq F_0(x) \cdot [3/4 + F_0(x)^2]^t \leq F_0(x) \cdot 1^t$. □

Lemma 4. *Let $F_0 \in \mathcal{F}$. For all $x \in (1/3, 1/2)$ and $t \geq 0$, $F_{4,t}(x) \leq F_{4,0}(x)$.*

Proof. Let $x \in (1/3, 1/2)$ and $p = F_{4,t-1}(x)$. We'll find a lower bound on the probability an inner candidate in $(x, 1-x)$ wins. Consider the following cases for candidate positions in a $k = 4$ plurality election:

1. All four candidates in $[0, 1/2)$ (and the symmetric case). An inner candidate wins if at least one candidate is inner. Accounting for symmetry, an inner candidate wins in this case w.p. $2(1/2^4 - p^4) = 1/8 - 2p^4$.
2. Three candidates in $[0, 1/2)$ and one in $(1/2, 1-x)$ (and the symmetric case). The candidate on the right has a higher vote share than any outer candidate on the left (as in Theorem 34 Case 4), so an inner candidate wins. Accounting for symmetry, an inner candidate wins in this case w.p. $2(4 \cdot 1/2^3 \cdot (1/2 - p)) = 1/2 - p$.
3. Two candidates in $(x, 1/2)$ and two in $(1/2, 1-x)$. All candidates are inner, so an inner candidate wins. This occurs w.p. $\binom{4}{2} \cdot (1/2 - p)^4 = 6(1/2 - p)^4$.

4. Two candidates in $[0, x)$ and two in $(1/2, 1-x)$ (and the symmetric case). Since $x > 1/3$, the rightmost candidate gets vote share greater than $1/3$. Meanwhile, the leftmost candidate gets vote share less than $1/3$. The second-leftmost candidate gets vote share less than $(2/3)/2 = 1/3$ (the candidates flanking it are closer together than 0 and $1 - x < 2/3$). Thus an inner candidate wins. Accounting for symmetry, an inner candidate wins in this case w.p. $2\binom{4}{2}p^2(1/2-p)^2 = 12p^2(1/2-p)^2$

5. Two candidates in $(x, 1/2)$, one in $(1/2, 1-x)$, and one in $(1-x, 1]$ (and the symmetric case). Label the candidates 1–4 in left–right order. By symmetry, candidate 3 is farther from $1/2$ than candidate 2 with probability $2/3$: all $3! = 6$ orderings of distance from $1/2$ between candidates 1–3 are equiprobable and only the 2 where candidate 3 is closest to $1/2$ fail this property. In this scenario, candidate 4 has vote share $1-(X_3+X_4)/2 = ((1-X_3)+(1-X_4))/2$ and candidate 1 has vote share $(X_1+X_2)/2$. Since $1 - X_3 < X_2$ (candidate 2 is closer to the center than 3) and $1 - X_4 < X_2$ (since $X_2 > x$ and $X_4 > 1-x$), candidate 1 has a larger vote share than candidate 4, the only outer candidate. Thus an inner candidate wins. Accounting for symmetry, an inner candidate wins in this case w.p. $2 \cdot 4 \cdot 3 \cdot 2/3 \cdot p(1/2-p)^3 = 16p(1/2-p)^3$

Combining all five cases gives a lower bound on the probability that an inner candidate wins:

$$
\begin{aligned}
&\Pr(x < \text{Plurality}(X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}) < 1 - x) \\
&\geq 1/8 - 2p^4 + 1/2 - p + 6(1/2-p)^4 + 12p^2(1/2-p)^2 + 16p(1/2-p)^3 \\
&= 1 - 2p \\
&= 1 - 2 \cdot F_{4,t-1}(x).
\end{aligned}
$$

By symmetry, this means

$$\begin{aligned}
F_{4,t}(x) &= \Pr(\text{Plurality}(X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}) \leq x) \\
&= \left[1 - \Pr(x < \text{Plurality}(X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}) < 1 - x)\right]/2 \\
&\leq \left[1 - (1 - 2 \cdot F_{4,t-1}(x))\right]/2 \\
&= F_{4,t-1}(x).
\end{aligned}$$

The claim then follows by induction on t. □

Theorem 35. *Let $F_0 \in \mathcal{F}$. For all $x \in (1/3, 1/2)$ and $t \geq 0$,*

$$F_{4,t}(x) \leq F_0(x) \cdot \left[1 - 4(1/2 - F_0(x/3 + 1/3))^3\right]^t. \tag{8.5}$$

Proof. Let $x \in (1/3, 1/2)$ and $p = F_{4,t-1}(x)$. By the argument in the proof of Lemma 4, an inner candidate wins with probability at least $1 - 2p$. We can strengthen this bound using Lemma 4 and one more case omitted from that analysis (which can't easily be used there): three candidates in $(x/3+1/3, 1/2)$ and one in $(1-x, 1]$ (and the symmetric case). Note that $x/3 + 1/3 = x + 2/3 \cdot (1/2 - x)$ is the point two-thirds of the way from x to $1/2$. The leftmost candidate gets vote share more than $x/3 + 1/3$. Meanwhile, the lone outer candidate gets vote share less than $x + (1 - x - (x/3 + 1/3))/2 = x/3 + 1/3$, so an inner candidate wins. By Lemma 4, we know $F_{4,t-1}(x/3 + 1/3) \leq F_{4,0}(x/3 + 1/3)$. Thus, a candidate is in $(x/3 + 1/3, 1/2)$ with probability $1/2 - F_{4,t-1}(x/3 + 1/3) \geq 1/2 - F_{4,0}(x/3 + 1/3)$. Therefore, accounting for symmetry, an inner candidate wins in this case w.p. at least $2 \cdot 4 \cdot (1/2 - F_{4,0}(x/3 + 1/3))^3 \cdot p = 8p(1/2 - F_{4,0}(x/3 + 1/3))^3$.

Combining this new case with the cases from the proof of Lemma 4, an inner candidate wins w.p. at least $1 - 2p + 8p(1/2 - F_{4,0}(x/3 + 1/3))^3$. By symmetry, this means

$$
\begin{aligned}
F_{4,t}(x) &\le \left[1-(1-2p+8p(1/2-F_{4,0}(x/3+1/3))^3)\right]/2 \\
&= \left[2p-8p(1/2-F_{4,0}(x/3+1/3))^3\right]/2 \\
&= p\left[1-4(1/2-F_{4,0}(x/3+1/3))^3\right] \\
&= F_{4,t-1}(x)\cdot\left[1-4(1/2-F_{4,0}(x/3+1/3))^3\right].
\end{aligned}
$$

The claim then follows by induction on t. □

Theorem 36. *Let $F_0 \in \mathcal{F}$. For any $k \ge 5$, there exists some $x < 1/2$ such that $\lim_{t\to\infty} F_{k,t}(x) \ne 0$. That is, the candidate distribution does not converge to a point mass at $1/2$.*

Proof. Suppose $F_{k,t-1}(1/4) \le \alpha$ for some small α. Let $x \in (1/4, 1/2)$ and $F_{k,t-1}(x) = p$, so $F_{k,t-1}(x) - F_{k,t-1}(1/4) \ge p - \alpha$. We'll lower bound the probability that the winner is is an *outer* candidate outside of $(x, 1-x)$, focusing mainly on cases where all candidates are in $(1/4, 3/4)$ so we can apply Lemma 5.

If all candidates are in $[0, 1/2)$, then an outer candidate only wins if all candidates are left of x, which occurs w.p. p^k. Accounting for the symmetric case gives an outer candidate win probability of $2p^k$ when all candidates are on the same side. Now suppose there is at least one candidate on each side. If the left- and rightmost candidates are in $(1/4, x)$ and $(1-x, 3/4)$, respectively, then an outer candidate wins by Lemma 5. We can find the probability this occurs as the probability that all candidates are in $(1/4, 3/4)$ minus the probability that all candidates are in $(1/4, 1-x]$ or in $[x, 3/4)$—since this means there is at least one candidate each in $(1-x, 3/4)$ and $(1/4, x)$. Since $F_{k,t-1}(1/4) \le \alpha$, the following is a lower bound on the probability the leftmost candidate is at $X_{1,t} \in (1/4, x)$ and the rightmost is at

$X_{k,t} \in (x-1, 3/4)$:

$$\begin{aligned}
&\Pr(X_{1,t} \in (1/4, x), X_{k,t} \in (x-1, 3/4)) \\
&\geq \underbrace{(1-2\alpha)^k}_{\text{all in } (1/4,3/4)} - \Big[\underbrace{(1-\alpha-p)^k}_{\text{all in } (1/4,1-x]} + \underbrace{(1-\alpha-p)^k}_{\text{all in } [x,3/4)} - \underbrace{(1-2p)^k}_{\text{all in } [x,1-x]}\Big] \\
&\hfill \text{(by inclusion–exclusion)} \\
&= (1-2\alpha)^k - 2(1-\alpha-p)^k + (1-2p)^k.
\end{aligned}$$

Combining this with the case where all candidates are on the same side (and then dividing by 2 to account for symmetry) yields a lower bound on $F_{k,t}(x)$:

$$\begin{aligned}
F_{k,t}(x) &\geq \left[2p^k + (1-2\alpha)^k - 2(1-\alpha-p)^k + (1-2p)^k\right]/2 \\
&= p^k + (1-2\alpha)^k/2 - (1-\alpha-p)^k + (1-2p)^k/2. \qquad \text{(G.1)}
\end{aligned}$$

We can now use this bound to prove non-convergence. Suppose for a contradiction that $\lim_{t\to\infty} F_{k,t}(x) = 0$ for all $x < 1/2$. Then there exists some t^* such that $F_{k,t}(1/4) \leq \alpha = \left[1 - (499/512)^{1/k}\right]/2$ for all $t > t^*$. But now consider $x^* = F^{-1}_{k,t^*}(1/4) < 1/2$. Since $F_{k,t}(1/4) \leq \alpha$ for all $t > t^*$, we can use the fact above to show inductively that $F_{k,t}(x^*) \geq 1/4$ for all $t \geq t^*$. For the base case $t = t^*$, the claim is vacuously true: $F_{k,t^*}(x^*) = 1/4 \geq 1/4$. Now suppose for $t > t^*$ that $F_{k,t-1}(x^*) \geq 1/4$. Then $z = F^{-1}_{k,t-1}(1/4) \leq x^*$. From (G.1), we then have:

$$\begin{aligned}
F_{k,t}(z) &\geq 1/4^k + (1-2\alpha)^k/2 - (1-\alpha-1/4)^k + (1-2/4)^k/2 \\
&> (1-2\alpha)^k/2 - (3/4-\alpha)^k && \text{(throw away terms)} \\
&\geq (1-2\alpha)^k/2 - (3/4)^5 && (\text{since } k \geq 5,\ \alpha > 0) \\
&= \left(1 - 2\left[1-(499/512)^{1/k}\right]/2\right)^k/2 - (3/4)^5 && (\text{plug in } \alpha) \\
&= 499/1024 - 243/1024 \\
&= 1/4.
\end{aligned}$$

By the monotonicity of the CDF, $F_{k,t}(x^*) > 1/4$, since $z \leq x^*$. By induction, $F_{k,t}(x^*) \geq 1/4$ for all $t \geq t^*$ This contradicts that $\lim_{t\to\infty} F_{k,t}(x) = 0$ for all $x < 1/2$. $\square$

G.2.2 Proofs from Section 8.3

Our proofs with ϵ-uniform noise make extensive use of the following lemma, which allows us to translate the convergence of an iterated map bounding a sequence into an eventual bound on the sequence.

Lemma 12. *Consider an iterated map $x_t = f(x_{t-1})$ where $f : [0, 1/2) \to [0, 1/2)$ is non-decreasing. Suppose $\lim_{t\to\infty} x_t = c$ for all $x_0 \in I \subseteq [0, 1/2)$.*

1. *If $y_t \leq f(y_{t-1})$ for all $t > 0$, then $\limsup_{t\to\infty} y_t \leq c$ for all $y_0 \in I$.*
2. *If $y_t \geq f(y_{t-1})$ for all $t > 0$, then $\liminf_{t\to\infty} y_t \geq c$ for all $y_0 \in I$.*

Proof. Let $y_0 \in I$ and define $x_0 = y_0$. Suppose $y_t \leq f(y_{t-1})$ for all $t > 0$. We'll show $y_t \leq x_t$ by induction. The base case $t = 0$ holds by the definition of x_0. Suppose for $t > 0$ that $y_{t-1} \leq x_{t-1}$. Then $y_t \leq f(y_{t-1}) \leq f(x_{t-1}) = x_t$ (since f is non-decreasing), so $y_t \leq x_t$ for all t by induction. Thus, $\limsup_{t\to\infty} y_t \leq \limsup_{t\to\infty} x_t = \lim_{t\to\infty} x_t = c$. The second claim with $y_t \geq f(y_{t-1})$ follows from the exact same argument with each $\leq$ replaced by $\geq$ and $\limsup$ replaced by $\liminf$. $\square$

Lemma 6. *For all initial $p \in [0, 1/2]$, $\epsilon \in (0, 1)$, and $x \in [0, 1/2)$, the quadratic iterated map $p' = 2p^2(1-\epsilon)^2 + 4px\epsilon(1-\epsilon) + 2x^2\epsilon^2$ converges to the fixed point $p^* = \frac{1-4x\epsilon(1-\epsilon)-\sqrt{1-8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2} \leq \epsilon$.*

Proof. We begin by looking for the fixed points of the map:

$$
\begin{aligned}
&2p^2(1-\epsilon)^2 + 4px\epsilon(1-\epsilon) + 2x^2\epsilon^2 = p \\
&\Leftrightarrow 2(1-\epsilon)^2p^2 + (4x\epsilon(1-\epsilon) - 1)p + 2x^2\epsilon^2 = 0.
\end{aligned}
$$

Applying the quadratic formula and simplifying yields the two fixed points:

$$
\begin{aligned}
p_1^* &= \frac{1 - 4x\epsilon(1-\epsilon) - \sqrt{1 - 8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2} \\
p_2^* &= \frac{1 - 4x\epsilon(1-\epsilon) + \sqrt{1 - 8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2}.
\end{aligned}
$$

We'll show that p_1^* is stable and that $p_1^* \leq \epsilon$ while p_2^* is unstable and $p_2^* > 1/2$. To see that $p_1^* \leq \epsilon$, consider $\epsilon - p_1^*$:

$$
\epsilon - \frac{1 - 4x\epsilon(1-\epsilon) - \sqrt{1 - 8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2} = \frac{4(1-\epsilon)^2\epsilon - 1 + 4x\epsilon(1-\epsilon) + \sqrt{1 - 8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2}. \tag{G.2}
$$

It suffices to show the numerator is non-negative. Taking its derivative with respect to x shows the numerator is decreasing in x:

$$
\begin{aligned}
\frac{\partial}{\partial x}\left[4(1-\epsilon)^2\epsilon - 1 + 4x\epsilon(1-\epsilon) + (1 - 8\epsilon x(1-\epsilon))^{1/2}\right] &= 4\epsilon(1-\epsilon) - \frac{4\epsilon(1-\epsilon)}{(1 - 8\epsilon x(1-\epsilon))^{1/2}} \\
&< 4\epsilon(1-\epsilon) - 4\epsilon(1-\epsilon) \\
&= 0.
\end{aligned}
$$

Thus, it suffices to show the function is non-negative when $x = 1/2$. For $x = 1/2$,

$$
\begin{aligned}
&4(1-\epsilon)^2\epsilon - 1 + 4x\epsilon(1-\epsilon) + \sqrt{1 - 8\epsilon x(1-\epsilon)} \\
&= 4(1-\epsilon)^2\epsilon - 1 + 2\epsilon(1-\epsilon) + \sqrt{1 - 4\epsilon(1-\epsilon)} \\
&= 4\epsilon^3 - 10\epsilon^2 + 6\epsilon - 1 + \sqrt{(1-2\epsilon)^2} \\
&= 4\epsilon^3 - 10\epsilon^2 + 6\epsilon - 1 + |1 - 2\epsilon|.
\end{aligned}
$$

Consider the cases $\epsilon \leq 1/2$ and $\epsilon > 1/2$. If $\epsilon \leq 1/2$,

$$\begin{aligned}
4\epsilon^3 - 10\epsilon^2 + 6\epsilon - 1 + |1 - 2\epsilon| &= 4\epsilon^3 - 10\epsilon^2 + 4\epsilon \\
&= 2\epsilon(2 - \epsilon)(1 - 2\epsilon) \\
&\geq 0. && \text{(since } \epsilon \leq 1/2\text{)}
\end{aligned}$$

If $\epsilon > 1/2$,

$$\begin{aligned}
4\epsilon^3 - 10\epsilon^2 + 6\epsilon - 1 + |1 - 2\epsilon| &= 4\epsilon^3 - 10\epsilon^2 + 8\epsilon - 2 \\
&= 2(1 - \epsilon)^2(2\epsilon - 1) \\
&\geq 0 && \text{(since } \epsilon > 1/2\text{)}
\end{aligned}$$

Therefore the numerator in (G.2) is non-negative, so $p_1^* \leq \epsilon$. To show p_1^* is stable, consider the derivative of the iterated map in (8.8):

$$\frac{\partial}{\partial p}\left[2p^2(1-\epsilon)^2 + 4px\epsilon(1-\epsilon) + 2x^2\epsilon^2\right] = 4(1-\epsilon)^2 p + 4x\epsilon(1-\epsilon). \tag{G.3}$$

Plugging in p_1^*:

$$\begin{aligned}
4(1-\epsilon)^2 p_1^* + 4x\epsilon(1-\epsilon) &= 4(1-\epsilon)^2 \frac{1 - 4x\epsilon(1-\epsilon) - \sqrt{1 - 8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2} + 4x\epsilon(1-\epsilon) \\
&= 1 - \sqrt{1 - 8\epsilon x(1-\epsilon)} \\
&< 1 - \sqrt{1 - 4\epsilon(1-\epsilon)} && (x < 1/2) \\
&\leq 1. && (\epsilon(1-\epsilon) \leq 1/4)
\end{aligned}$$

Thus the derivative of the iterated map at p_1^* has magnitude strictly less than 1, so p_1^* is a stable fixed point.

Now, consider the other fixed point p_2^*:

$$
\begin{aligned}
p_2^* &= \frac{1 - 4x\epsilon(1-\epsilon) + \sqrt{1 - 8\epsilon x(1-\epsilon)}}{4(1-\epsilon)^2} \\
&> \frac{1 - 2\epsilon(1-\epsilon) + \sqrt{1 - 4\epsilon(1-\epsilon)}}{4(1-\epsilon)^2} && (x < 1/2) \\
&= \frac{1 - 2\epsilon(1-\epsilon) + \sqrt{(1-2\epsilon)^2}}{4(1-\epsilon)^2} \\
&= \frac{1 - 2\epsilon(1-\epsilon) + |1-2\epsilon|}{4(1-\epsilon)^2}.
\end{aligned}
$$

If $\epsilon \leq 1/2$,

$$
\begin{aligned}
p_2^* &> \frac{1 - 2\epsilon(1-\epsilon) + 1 - 2\epsilon}{4(1-\epsilon)^2} \\
&= \frac{2(1-\epsilon)^2}{4(1-\epsilon)^2} \\
&= 1/2.
\end{aligned}
$$

If $\epsilon > 1/2$,

$$
\begin{aligned}
p_2^* &> \frac{1 - 2\epsilon(1-\epsilon) - 1 + 2\epsilon}{4(1-\epsilon)^2} \\
&= \frac{2\epsilon^2}{4(1-\epsilon)^2} \\
&> \frac{2(1/2)^2}{4(1-1/2)^2} \\
&= 1/2.
\end{aligned}
$$

In either case, $p_2^* > 1/2$. Additionally, plugging p_2^* into the derivative (G.3) yields $1 + \sqrt{1 - 8x\epsilon(1-\epsilon)} > 1$ (for $x < 1/2$), showing p_2^* is unstable. Thus, for $p \in [0, 1/2]$, the quadratic map converges to the stable fixed point $p_1^* \leq \epsilon$. $\square$

Lemma 13. *For any $\epsilon \in (0, 1/3)$, the cubic iterated map given by*

$$p' = 3/4 \cdot [\epsilon/2 + (1-\epsilon)p] + [\epsilon/2 + (1-\epsilon)p]^3$$

converges to $p^* \leq 1.5\epsilon$ for all initial $p \in [0, 1/2)$. Moreover, the map is non-decreasing in p on $[0, 1/2)$.

Proof. The fixed points of this map can be found using the cubic formula (equivalently, we used Mathematica):

$$\begin{aligned}
p_1^* &= \frac{1}{4}\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}} - \frac{1+2\epsilon}{4(1-\epsilon)} \\
p_2^* &= -\frac{1}{4}\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}} - \frac{1+2\epsilon}{4(1-\epsilon)} \\
p_3^* &= \frac{1}{2}.
\end{aligned}$$

We can ignore the negative fixed point p_2^*, since p can never be negative. We'll show that for $\epsilon < 1/3$, $p_1^* \in [0, 1.5\epsilon]$, p_1^* is stable, and the cubic map converges to p_1^* for $p \in [0, 1/2)$. To begin with, we'll show $p_1^* \geq 0$:

$$\begin{aligned}
p_1^* &= \frac{1}{4}\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}} - \frac{1+2\epsilon}{4(1-\epsilon)} \\
&= \frac{(1+15\epsilon)^{1/2}}{4(1-\epsilon)^{3/2}} - \frac{(1+2\epsilon)(1-\epsilon)^{1/2}}{4(1-\epsilon)^{3/2}} \\
&= \frac{(1+15\epsilon)^{1/2} - ((1+2\epsilon)^2)^{1/2}(1-\epsilon)^{1/2}}{4(1-\epsilon)^{3/2}} \\
&= \frac{(1+15\epsilon)^{1/2} - ((1+2\epsilon)^2(1-\epsilon))^{1/2}}{4(1-\epsilon)^{3/2}} \\
&= \frac{(1+15\epsilon)^{1/2} - (1+3\epsilon-4\epsilon^3)^{1/2}}{4(1-\epsilon)^{3/2}} \\
&\geq 0. \qquad \text{(since } 1+15\epsilon > 1+3\epsilon-4\epsilon^3\text{)}
\end{aligned}$$

Now we'll show that $p_1^* \leq 1.5\epsilon$. To do this, we'll show $1.5\epsilon - p_1^* \geq 0$:

$$\begin{aligned}
1.5\epsilon - p_1^* &= 1.5\epsilon - \frac{(1+15\epsilon)^{1/2} - (1+3\epsilon-4\epsilon^3)^{1/2}}{4(1-\epsilon)^{3/2}} \\
&= \frac{6\epsilon(1-\epsilon)^{3/2}}{4(1-\epsilon)^{3/2}} - \frac{(1+15\epsilon)^{1/2} - (1+3\epsilon-4\epsilon^3)^{1/2}}{4(1-\epsilon)^{3/2}} \\
&= \frac{6\epsilon(1-\epsilon)^{3/2} - (1+15\epsilon)^{1/2} + (1+3\epsilon-4\epsilon^3)^{1/2}}{4(1-\epsilon)^{3/2}}.
\end{aligned}$$

It suffices to show the numerator is non-negative on $[0, 1/3)$:

$$
\begin{aligned}
&6\epsilon(1-\epsilon)^{3/2} - (1+15\epsilon)^{1/2} + (1+3\epsilon-4\epsilon^3)^{1/2} \\
&= 6\epsilon(1-\epsilon)(1-\epsilon)^{1/2} - (1+15\epsilon)^{1/2} + (1+2\epsilon)(1-\epsilon)^{1/2} \\
&= (1-\epsilon)^{1/2}\left[6\epsilon(1-\epsilon) + (1+2\epsilon)\right] - (1+15\epsilon)^{1/2} \\
&= (1-\epsilon)^{1/2}(1+8\epsilon-6\epsilon^2) - (1+15\epsilon)^{1/2} \\
&= \left[(1-\epsilon)(1+8\epsilon-6\epsilon^2)^2\right]^{1/2} - (1+15\epsilon)^{1/2} \\
&= (1+15\epsilon+36\epsilon^2-148\epsilon^3+132\epsilon^4-36\epsilon^5)^{1/2} - (1+15\epsilon)^{1/2}.
\end{aligned}
$$

To show this is non-negative, it suffices to show $36\epsilon^2 - 148\epsilon^3 + 132\epsilon^4 - 36\epsilon^5$ is non-negative. Factoring yields

$$36\epsilon^2 - 148\epsilon^3 + 132\epsilon^4 - 36\epsilon^5 = 4\epsilon^2(1-3\epsilon)(9-10\epsilon+3\epsilon^2).$$

Finally, we can see this is non-negative for $\epsilon \in (0, 1/3)$, so $p_1^* \leq 1.5\epsilon$ for $\epsilon \in (0, 1/3)$.

Now, to show p_1^* is a stable fixed point, we can take the derivative of the cubic map at p_2^*:

$$
\begin{aligned}
&\frac{\partial}{\partial p}\left(3/4 \cdot [\epsilon/2 + (1-\epsilon)p] + [\epsilon/2 + (1-\epsilon)p]^3\right) \\
&= \frac{\partial}{\partial p}\left[p^3(1-\epsilon)^3 + \frac{3}{2}p^2\epsilon(1-\epsilon)^2 + \frac{3}{4}p(1-\epsilon)\left(\epsilon^2+1\right) + \frac{1}{8}\epsilon^3 + \frac{3}{8}\epsilon\right] \\
&= 3(1-\epsilon)^3p^2 + 3\epsilon(1-\epsilon)^2p + \frac{3}{4}(1-\epsilon)(1+\epsilon^2) \qquad \text{(G.4)}
\end{aligned}
$$

Plugging in p_1^* and simplifying yields

$$
\begin{aligned}
&3(1-\epsilon)^3\left(\frac{1}{4}\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}-\frac{1+2\epsilon}{4(1-\epsilon)}\right)^2+3\epsilon(1-\epsilon)^2\left(\frac{1}{4}\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}-\frac{1+2\epsilon}{4(1-\epsilon)}\right)\\
&\qquad+\frac{3}{4}(1-\epsilon)(1+\epsilon^2)\\
&=3(1-\epsilon)^3\left(\frac{1+15\epsilon}{16(1-\epsilon)^3}+\frac{(1+2\epsilon)^2}{16(1-\epsilon)^2}-\frac{1+2\epsilon}{8(1-\epsilon)}\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}\right)+\frac{3}{4}\epsilon(1-\epsilon)^2\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}\\
&\qquad-\frac{3}{4}\epsilon(1-\epsilon)(1+2\epsilon)+3/4(1-\epsilon)(1+\epsilon^2)\\
&=\frac{3(1+15\epsilon)}{16}+\frac{3(1-\epsilon)(1+2\epsilon)^2}{16}-\frac{3(1-\epsilon)^2(1+2\epsilon)}{8}\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}+\frac{3}{4}\epsilon(1-\epsilon)^2\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}\\
&\qquad-\frac{3}{4}\epsilon(1-\epsilon)(1+2\epsilon)+3/4(1-\epsilon)(1+\epsilon^2)\\
&=\frac{3}{16}(1+15\epsilon)+\frac{3}{16}(1-\epsilon)(1+2\epsilon)^2+(1-\epsilon)^2\left(\frac{3}{4}\epsilon-\frac{3}{8}(1+2\epsilon)\right)\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}\\
&\qquad-\frac{3}{4}\epsilon(1-\epsilon)(1+2\epsilon)+3/4(1-\epsilon)(1+\epsilon^2)\\
&=-\frac{3}{8}(1-\epsilon)^2\sqrt{\frac{1+15\epsilon}{(1-\epsilon)^3}}+\frac{9}{8}+\frac{15}{8}\epsilon\\
&=-\frac{3}{8}\sqrt{(1+15\epsilon)(1-\epsilon)}+\frac{9}{8}+\frac{15}{8}\epsilon.
\end{aligned}
$$

To see this is positive for $\epsilon < 1/3$, note that $\frac{3}{8}\sqrt{(1+15\epsilon)(1-\epsilon)} < \frac{3}{8}\sqrt{(1+15/3)} \approx 0.92 < 9/8$. We can also show the derivative of the cubic map at p_2^* is less than 1. To do this, we'll show that 1 minus the derivative at p_1^* is positive:

$$
\begin{aligned}
1-\left(-\frac{3}{8}\sqrt{(1+15\epsilon)(1-\epsilon)}+\frac{9}{8}+\frac{15}{8}\epsilon\right)&=\frac{3}{8}\sqrt{(1+15\epsilon)(1-\epsilon)}-\frac{1}{8}-\frac{15}{8}\epsilon\\
&=\sqrt{\frac{9}{64}(1+15\epsilon)(1-\epsilon)}-\sqrt{\left(\frac{1}{8}+\frac{15}{8}\epsilon\right)^2}.
\end{aligned}
$$

By the monotonicity of square roots, it suffices to show that the following quadratic is positive:

$$\frac{9}{64}(1+15\epsilon)(1-\epsilon) - \left(\frac{1}{8}+\frac{15}{8}\epsilon\right)^2 = -\frac{45\epsilon^2}{8} + \frac{3\epsilon}{2} + \frac{1}{8}$$
$$= \frac{1}{8}(1-3\epsilon)(15\epsilon+1).$$

which we can see is positive for $\epsilon \in (0, 1/3)$. Thus, the derivative of the cubic map at p_1^* is positive but less than 1, so p_1^* is a stable fixed point. The fixed point at $1/2$ is unstable, in contrast: plugging $p_3^* = 1/2$ into the derivative (G.4) and simplifying yields $3/2(1-\epsilon)$, which is larger than 1 for $\epsilon < 1/3$. Thus the cubic map converges to p_1^* for initial values in $[0, 1/2)$.

Finally, to show the map is non-decreasing in p, notice that derivative Equation (G.4) is non-negative for $p \geq 0$ and $\epsilon \in (0, 1]$. □

Theorem 39. *Let $F_0 \in \mathcal{F}$. For any $\epsilon \in (0, 1/3)$ and $x \in [0, 1/2)$, $\limsup_{t\to\infty} F^{\epsilon}_{3,t}(x) \leq 1.5\epsilon$.*

Proof. Let $x < 1/2$ and define $p = F^{\epsilon}_{3,t-1}(x)$. With ϵ-uniform noise, $\Pr(X_{i,t} \leq x) = \epsilon x + (1-\epsilon)p$. The case analysis from Theorem 34 then proceeds exactly the same way, so we can replace p by $\epsilon x + (1-\epsilon)p$ in the bound from Theorem 34 to get the equivalent bound with ϵ-uniform noise:

$$F^{\epsilon}_{3,t}(x) \leq 3/4 \cdot [\epsilon x + (1-\epsilon)p] + [\epsilon x + (1-\epsilon)p]^3. \tag{G.5}$$

While it would be possible to work directly with the cubic map (G.5), its fixed points are extremely messy. As such, we instead analyze the upper bound given by $x < 1/2$ and then use Lemma 12:

$$F^{\epsilon}_{3,t}(x) < 3/4 \cdot [\epsilon/2 + (1-\epsilon)p] + [\epsilon/2 + (1-\epsilon)p]^3. \tag{G.6}$$

If $F_0(x) < 1/2$, then Lemma 13 states that the map (G.6) upper bounding $F^{\epsilon}_{3,t}(x)$ converges to $p^* \leq 1.5\epsilon$. Thus, applying Lemma 12 gives $\limsup_{t\to\infty} F^{\epsilon}_{3,t}(x) \leq p^* \leq$

1.5ϵ as claimed. If $F_0(x) = 1/2$ (which is possible since we don't require that F_0 is positive near $1/2$), then applying (G.6),

$$
\begin{aligned}
F_{3,1}^{\epsilon}(x) &< 3/4 \cdot [\epsilon/2 + (1-\epsilon)/2] + [\epsilon/2 + (1-\epsilon)/2]^3 \\
&= 3/4 \cdot 1/2 + [1/2]^3 \\
&= 1/2.
\end{aligned}
$$

Thus $F_{3,1}^{\epsilon}(x) < 1/2$, so we can apply Lemmas 12 and 13 with initial $p = F_{3,1}^{\epsilon}(x)$ rather than $F_0(x)$. $\square$

Lemma 7. *Let $F_0 \in \mathcal{F}$. With ϵ-uniform noise, for any $\epsilon \in (0,1]$, $x \in (1/3, 1/2)$, and $t > 0$,*

$$F_{4,t}^{\epsilon}(x) \leq \epsilon x + (1-\epsilon) F_{4,t-1}^{\epsilon}(x).$$

Thus, $F_{4,t}^{\epsilon}(x) \leq \max\{x, F_{4,0}^{\epsilon}(x)\}$.

Proof. Let $x \in (1/3, 1/2)$. Just as in Theorem 39, we can take the bound from Lemma 4 and replace $F_{4,t-1}^{\epsilon}(x)$ with $\epsilon x + (1-\epsilon)F_{4,t-1}^{\epsilon}(x)$ to get the claimed upper bound with ϵ-uniform noise. The second part of the claim follows by induction after noting $F_{4,1}^{\epsilon}(x) \leq \epsilon x + (1-\epsilon)F_{4,0}^{\epsilon}(x) \leq \epsilon \max\{x, F_{4,0}^{\epsilon}(x)\} + (1-\epsilon)\max\{x, F_{4,0}^{\epsilon}(x)\} = \max\{x, F_{4,0}^{\epsilon}(x)\}$. $\square$

Theorem 40. *Let $F_0 \in \mathcal{F}$. For any $\epsilon \in (0,1]$ and $x \in (1/3, 1/2)$, let $\beta = 1/2 - \epsilon(x/3 + 1/3) - (1-\epsilon)\max\{x/3 + 1/3, F_0(x/3 + 1/3)\}$. Then $\beta \in (0, 1/2]$ and $\limsup_{t\to\infty} F_{4,t}^{\epsilon}(x) \leq \frac{1}{8\beta^3}\epsilon$.*

Proof. Let $p = F_{4,t-1}^{\epsilon}(x)$. By Lemma 7 and symmetry, an inner candidate in $(x, 1-x)$ wins with probability at least $1 - 2p(1-\epsilon) - 2x\epsilon$. We'll strengthen this bound in the same way as in Theorem 35, using the case with three candidates in

$(x/3+1/3, 1/2)$ and one in $(1-x, 1]$ (and the symmetric case). With ϵ-uniform noise and accounting for symmetry, the probability this case occurs is

$$\begin{aligned}
&8[\epsilon x + (1-\epsilon)p][1/2 - \epsilon(x/3+1/3) - (1-\epsilon)F^{\epsilon}_{4,t-1}(x/3+1/3)]^3 \\
&\geq 8[\epsilon x + (1-\epsilon)p][1/2 - \epsilon(x/3+1/3) - (1-\epsilon)\max\{x/3+1/3, F_0(x/3+1/3)\}]^3 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Lemma 7)} \\
&= 8[\epsilon x + (1-\epsilon)p]\beta^3.
\end{aligned}$$

Then, adding this case to the cases implicitly used in Lemma 7 (see Lemma 4 for the list of cases), we find

$$\begin{aligned}
&\Pr(x < \mathrm{Plurality}(X^{\epsilon}_{1,t}, X^{\epsilon}_{2,t}, X^{\epsilon}_{3,t}, X^{\epsilon}_{4,t}) < 1-x) \\
&\geq 1 - 2p(1-\epsilon) - 2x\epsilon + 8[\epsilon x + (1-\epsilon)p]\beta^3.
\end{aligned}$$

By symmetry,

$$\begin{aligned}
F^{\epsilon}_{4,t}(x) &= \left[1 - \Pr(x < \mathrm{Plurality}(X^{\epsilon}_{1,t}, X^{\epsilon}_{2,t}, X^{\epsilon}_{3,t}, X^{\epsilon}_{4,t}) < 1-x)\right]/2 \\
&\leq p(1-\epsilon) + x\epsilon - 4[\epsilon x + (1-\epsilon)p]\beta^3 \\
&= p(1-\epsilon)(1-4\beta^3) + \epsilon x(1-4\beta^3). \qquad \text{(G.7)}
\end{aligned}$$

We'll show that the iterated map (G.7) upper bounding $F^{\epsilon}_{4,t}(x)$ converges to a fixed point upper bounded by $\frac{1}{8\beta}\epsilon$ and then apply Lemma 12. First, we'll find the fixed point (unique, since this is a linear map):

$$\begin{aligned}
&p^*(1-\epsilon)(1-4\beta^3) + \epsilon x(1-4\beta^3) = p^* \\
\Leftrightarrow\quad &p^*[(1-\epsilon)(1-4\beta^3) - 1] + \epsilon x(1-4\beta^3) = 0 \\
\Leftrightarrow\quad &p^* = \frac{\epsilon x(1-4\beta^3)}{1-(1-\epsilon)(1-4\beta^3)}. \qquad \text{(G.8)}
\end{aligned}$$

To show convergence to p^*, it suffices to show that the slope of the map is in $(-1, 1)$ (any such linear map converges to its unique fixed point, e.g., by the Banach fixed-point theorem). First, we can show $\beta \in (0, 1/2]$:

$$\begin{aligned}
\beta &= 1/2 - \epsilon(x/3 + 1/3) - (1-\epsilon)\max\{x/3 + 1/3, F_0(x/3 + 1/3)\} \\
&> 1/2 - \epsilon(1/2) - (1-\epsilon)\max\{1/2, F_0(x/3 + 1/3)\} && \text{(since } x < 1/2) \\
&= 1/2 - \epsilon(1/2) - (1-\epsilon)(1/2) && \text{(since } F_0(x/3 + 1/3) \leq 1/2) \\
&= 0.
\end{aligned}$$

Thus, the slope $(1-\epsilon)(1-4\beta^3) \in [0, 1)$, so the map (G.7) converges to p^* for all initial values p and is non-decreasing in p. Now we can upper bound p^*:

$$\begin{aligned}
p^* &= \frac{\epsilon x(1-4\beta^3)}{1-(1-\epsilon)(1-4\beta^3)} \\
&< \frac{\epsilon/2}{1-(1-4\beta^3)} && (\epsilon \geq 0,\ x < 1/2) \\
&= \frac{\epsilon}{8\beta^3}.
\end{aligned}$$

Thus, by Lemma 12, $\limsup_{t\to\infty} F^{\epsilon}_{4,t}(x) \leq p^* < \frac{1}{8\beta^3}\epsilon$. ☐

Theorem 41. *Let $F_0 \in \mathcal{F}$. For $k \geq 5$, the candidate distribution does not approximately converge to the center under replicator dynamics with ϵ-uniform noise.*

Proof. Suppose for a contradiction that the candidate distribution does approximately converge to the center. That is, suppose that for all $c > 0$ and $x < 1/2$, there exists some $\epsilon_{\max} > 0$ such that with ϵ-uniform noise, for any $\epsilon \in (0, \epsilon_{\max}]$, $\limsup_{t\to\infty} F^{\epsilon}_{k,t}(x) < c$. If $\limsup_{t\to\infty} F^{\epsilon}_{k,t}(x) < c$, then there is some t^* such that for all $t \geq t^*$, $F^{\epsilon}_{k,t}(x) \leq c$. In particular, let $\epsilon^*_{\max}$ and t^* be the corresponding values for $x = 1/4$. Then for any ϵ-uniform noise with $\epsilon \leq \epsilon^*_{\max}$, $F^{\epsilon}_{k,t^*}(1/4) \leq c$. Additionally, consider the point $z = (F^{\epsilon}_{k,t^*})^{-1}(1/4)$. By our assumption, there is some $\epsilon'_{\max}$ and t' such that if $\epsilon < \epsilon'_{\max}$, then $F^{\epsilon}_{k,t}(z) \leq c$ for all $t \geq t'$. We can make

ϵ and c as small as needed, so we'll pick:

$$c < \left[1 - (125/128)^{1/k}\right]/3 \tag{G.9}$$

$$\epsilon < \min\left\{\epsilon^*_{\text{max}}, \epsilon'_{\text{max}}, \left[1 - (125/128)^{1/k}\right]/3\right\}. \tag{G.10}$$

Note that $\left[1-(125/128)^{1/k}\right]/3$ is largest at $k = 5$, when its value is approximately 0.0016. Also note that we must have $t' > t^*$, since $F^{\epsilon}_{k,t^*}(z) = 1/4 > c$.

Now, we can apply the same argument as in Theorem 36, finding a lower bound on $F^{\epsilon}_{4,t^*+1}(x)$ (for $x \in (1/4, 1/2)$) given that only a c-fraction of the winners in generation t^* are left of $1/4$ (note that this parameter was called α in the proof of Theorem 36). Let $p = F^{\epsilon}_{k,t^*}(x)$ with ϵ-uniform noise. For brevity, we will avoid repeating the argument from Theorem 36 and instead substitute directly into the resulting bound (G.1). Replacing p with $\Pr(X^{\epsilon}_{i,t^*} \leq x) = \epsilon x + (1-\epsilon)p$ and α with $\Pr(X^{\epsilon}_{i,t^*} \leq 1/4) \leq \epsilon/4 + (1-\epsilon)c$ in (G.1) then yields

$$\begin{aligned} F^{\epsilon}_{4,t^*+1}(x) \geq{}& [\epsilon x + (1-\epsilon)p]^k + (1 - 2[\epsilon/4 + (1-\epsilon)c])^k/2 \\ &- (1 - [\epsilon/4 + (1-\epsilon)c] - [\epsilon x + (1-\epsilon)p])^k + (1 - 2[\epsilon x + (1-\epsilon)p])^k/2. \end{aligned} \tag{G.11}$$

We will now derive a contradiction: that $F^{\epsilon}_{k,t}(z)$ never goes below $1/4$ as t increases from t^* to t', when it should go below c. We know $z > 1/4$ by the monotonicity of the CDF, since $F^{\epsilon}_{k,t^*}(1/4) \leq c$. So, we can apply the lower bound (G.11) to z (where $p = 1/4$):

$$\begin{aligned}
F^{\epsilon}_{k,t^*+1}(z) &\geq [\epsilon z + (1-\epsilon)/4]^k + (1 - 2[\epsilon/4 + (1-\epsilon)c])^k/2 \\
&\qquad - (1 - [\epsilon/4 + (1-\epsilon)c] - [\epsilon z + (1-\epsilon)/4])^k + (1 - 2[\epsilon z + (1-\epsilon)/4])^k/2 \\
&> (1 - 2[\epsilon/4 + (1-\epsilon)c])^k/2 - (1 - [\epsilon/4 + (1-\epsilon)c] - [\epsilon z + (1-\epsilon)/4])^k \\
&= (1 - \epsilon/2 - 2(1-\epsilon)c)^k/2 - (1 - [\epsilon/4 + (1-\epsilon)c] - [\epsilon z + (1-\epsilon)/4])^k \\
&> (1 - \epsilon - 2c)^k/2 - (1 - (1-\epsilon)/4)^k \\
&\geq (1 - \epsilon - 2c)^k/2 - (3/4 + \epsilon/4)^5. \qquad \text{(G.12)}
\end{aligned}$$

Now consider each term in (G.12). By our upper bounds on c and ϵ,

$$\begin{aligned}
(1-\epsilon-2c)^k/2 &> \left(1 - \left[1 - (125/128)^{1/k}\right]/3 - 2\left[1 - (125/128)^{1/k}\right]/3\right)^k/2 \\
&= \left(1 - \left[1 - (125/128)^{1/k}\right]\right)^k/2 \\
&= 125/256.
\end{aligned}$$

Meanwhile, ϵ is also small enough that $(3/4+\epsilon/4)^5 < \frac{244}{1024}$ (solving for ϵ reveals we need $\epsilon < 0.0024$, which we have ensured). This then means that $(1-\epsilon-2c)^k/2 - (3/4+\epsilon/4)^5 > \frac{125}{256} - \frac{244}{1024} = 1/4$. Following the above chain of inequalities, this shows that $F^{\epsilon}_{k,t^*+1}(z) > 1/4$.

We will now show by induction that for all $t > t^*$, $F^{\epsilon}_{k,t}(z) > 1/4$, a contradiction (since by our assumption, $F^{\epsilon}_{k,t'}(z) \leq c$ with $t' > t^*$). We have just shown the base case $t = t^* + 1$ above. Then, suppose as an inductive hypothesis that $F^{\epsilon}_{k,t}(z) \geq 1/4$ for $t > t^*$. Define $w = (F^{\epsilon}_{k,t})^{-1}(1/4)$; by the inductive hypothesis, we know $w \leq z$. By the argument above, $F^{\epsilon}_{k,t+1}(w) > 1/4$ (the argument only requires that $F^{\epsilon}_{k,t}(1/4) \leq c$ and $w \in (1/4, 1/2)$, which are satisfied here). Thus, by the monotonicity of the CDF, $F^{\epsilon}_{k,t+1}(z) > 1/4$. By induction, we then have $F^{\epsilon}_{k,t'}(z) > 1/4$, a contradiction. $\square$

G.2.3 Proofs from Section 8.4

Before proving Theorem 42, we need a few supporting lemmas. We begin with a result from the proof of Theorem 36, specialized to the case when all candidates are in $(1/4, 3/4)$.

Lemma 14. *Suppose $F_0 \in \mathcal{F}$ is supported on $(1/4, 3/4)$. For $k \geq 5$ and $x \in (1/4, 1/2)$,*

$$F_{k,t}(x) \geq 1/2 + F_{k,t-1}(x)^k - (1 - F_{k,t-1}(x))^k + (1 - 2F_{k,t-1}(x))^k/2.$$

Proof. We can use the same argument as in Theorem 36 to find a lower bound on $F_{k,t}(x)$, but now $F_{k,t-1}(x) \leq \alpha = 0$ since F_0 (and therefore all subsequent $F_{k,t}$) is supported only on $(1/4, 3/4)$. Plugging $\alpha = 0$ into (G.1) yields the claim, noting $p = F_{k,t-1}(x)$. □

This gives us an iterated map which bounds $F_{k,t}(x)$ from below. We can show that this map converges to 1/2 in a large interval around 1/2, meaning that the candidate distribution converges to one with no mass in this interval. We cannot give an explicit form for the basin of attraction of this map since it depends on a root of a polynomial of order k, but we can show the interval grows in k and characterize it for $k = 5$.

Lemma 15. *For all $k \geq 5$, the iterated map given by $p' = 1/2 + p^k - (1-p)^k + (1-2p)^k/2$ converges to 1/2 for all initial $p \in ([1 - \sqrt{3/7}]/2 = 0.172\ldots, 1/2]$. Moreover, this map in non-decreasing in p on $[0, 1/2)$.*

Proof. First, we'll show 1/2 is a stable fixed point of the map. Indeed, $1/2 + (1/2)^k - (1-1/2)^k + (1-2(1/2))^k/2 = 1/2 + 1/2^k - 1/2^k = 1/2$. The stability of this fixed point is determined by the derivative

$$\frac{\partial}{\partial p}\left(1/2 + p^k - (1-p)^k + (1-2p)^k/2\right) = kp^{k-1} + k(1-p)^{k-1} - k(1-2p)^{k-1}. \tag{G.13}$$

At 1/2, the derivative is $k(1/2)^{k-1} + k(1-1/2)^{k-1} - k(1-1)^{k-1} = k(1/2)^{k-2}$. For $k \geq 5$, $k(1/2)^{k-2} < 1$, showing the fixed point is stable.

For $k = 5$, we can find the other fixed points of the map by factoring:

$$\begin{aligned}
&1/2 + p^5 - (1-p)^5 + (1-2p)^5/2 = p \\
&\Leftrightarrow 1/2 + p^5 - (1-p)^5 + (1-2p)^5/2 - p = 0 \\
&\Leftrightarrow p(1-p)(1-2p)\left(-7p^2 + 7p - 1\right) = 0. \\
&\Leftrightarrow p \in \left\{0, (1-\sqrt{3/7})/2, 1/2, (1+\sqrt{3/7})/2, 1\right\} = \{0, 0.172\ldots, 0.5, 0.827\ldots, 1\}.
\end{aligned}$$

Plugging in the $k = 5$ fixed point $(1 - \sqrt{3/7})/2 = 0.172\ldots$ to the derivative (G.13) yields ≈ 1.43, so this fixed point in unstable. Next, note that the map monotonically increases in p for $p \in (0, 1/2)$, since the derivative (G.13) is positive (as $1 - p > 1 - 2p$; similarly, the map is non-increasing on $[0, 1/2)$, as claimed). Thus, for $k = 5$, the map is larger than p but smaller than 1/2 for p in $([1 - \sqrt{3/7}]/2, 1/2)$ and initial values in this range converge to the stable fixed point 1/2.

The final step is to show the map is increasing in k for $p \in (0, 1/2)$, which means that the basin of attraction only grows in k. To do this, consider the derivative of the map with respect to k:

$$\begin{aligned}
&\frac{\partial}{\partial k}\left(1/2 + p^k - (1-p)^k + (1-2p)^k/2\right) \\
&= \frac{1}{2}(1-2p)^k \log(1-2p) - (1-p)^k \log(1-p) + p^k \log p.
\end{aligned}$$

We establish this is positive in the following lemma.

Lemma 16. *For all $k \geq 3$ and $0 < p < 1/2$,*

$$1/2(1-2p)^k \log(1-2p) - (1-p)^k \log(1-p) + p^k \log p > 0.$$

Proof. We thank River Li[1] for a key idea behind this analysis, based on the following integral trick:

$$\begin{aligned}\int_0^1 \frac{x-1}{1+t(x-1)}\,dt &= \log(1+t(x-1))\big|_{t=0}^{1} \\ &= \log(1+1(x-1)) - \log(1+0(x-1)) \\ &= \log x.\end{aligned}$$

Now, apply this identity to the function in question for $k = 3$:

$$\begin{aligned}&1/2(1-2p)^3 \log(1-2p) - (1-p)^3 \log(1-p) + p^3 \log p \\ &= 1/2(1-2p)^3 \int_0^1 \frac{-2p}{1+t(-2p)}\,dt \; - (1-p)^3 \int_0^1 \frac{-p}{1+t(-p)}\,dt \; + p^3 \int_0^1 \frac{p-1}{1+t(p-1)}\,dt \\ &= \int_0^1 \left(-\frac{p(1-2p)^3}{1-2pt} + \frac{p(1-p)^3}{1-pt} - \frac{p^3(1-p)}{1+pt-t} \right) dt.\end{aligned}$$

We'll show that the integrand is positive for all $t \in [0,1]$, which implies the integral is also positive. Converting to a common denominator,

$$\begin{aligned}&-\frac{p(1-2p)^3}{1-2pt} + \frac{p(1-p)^3}{1-pt} - \frac{p^3(1-p)}{1+pt-t} \\ &= \frac{-p(1-2p)^3(1-pt)(1+pt-t) + p(1-p)^3(1-2pt)(1+pt-t) - p^3(1-p)(1-2pt)(1-pt)}{(1-2pt)(1-pt)(1+pt-t)}.\end{aligned}$$

Since $0 < p < 1/2$ and $0 \leq t \leq 1$, the denominator is positive, so we just need to show the numerator is positive. We can factor:

$$\begin{aligned}&-p(1-2p)^3(1-pt)(1+pt-t) + p(1-p)^3(1-2pt)(1+pt-t) - p^3(1-p)(1-2pt)(1-pt) \\ &= p^2(1-2p)(3-4p-4t+4pt+p^2t+t^2+pt^2-4p^2t^2+2p^3t^2) \\ &= p^2(1-2p)\left[(1+p-4p^2+2p^3)t^2 - (4-4p-p^2)t + 3 - 4p\right].\end{aligned}$$

[1] https://math.stackexchange.com/q/4789978

Again, since p^2 and $(1-2p)$ are positive, we just need to show the right factor is positive. Now, notice that $4 - 4p - p^2 > 0$ (since $p < 1/2$), and $t \leq \frac{t^2+1}{2}$, so

$$
\begin{aligned}
&(1 + p - 4p^2 + 2p^3)t^2 - (4 - 4p - p^2)t + 3 - 4p \\
&\geq (1 + p - 4p^2 + 2p^3)t^2 - (4 - 4p - p^2)\frac{t^2+1}{2} + 3 - 4p \\
&= \frac{1}{2}\left[2 - 4p + p^2 - (2 - 6p + 7p^2 - 4p^3)t^2\right]
\end{aligned}
$$

Now, $2 - 6p + 7p^2 - 4p^3$ is positive for $p \in (0, 1/2)$. We can see this since its derivative, $-6 + 14p - 12p^2$ is negative (achieving a maximum of $-23/12$ at $p = 7/12$) and the polynomial has a zero at $p = 1/2$. Thus, we can shrink the function by replacing t^2 by 1:

$$
\begin{aligned}
\frac{1}{2}\left[2 - 4p + p^2 - (2 - 6p + 7p^2 - 4p^3)t^2\right] &\geq \frac{1}{2}\left[2 - 4p + p^2 - (2 - 6p + 7p^2 - 4p^3)\right] \\
&= p(1-p)(1-2p).
\end{aligned}
$$

Finally, we see that this is positive for all $p \in (0, 1)$, which implies that

$$1/2(1-2p)^3 \log(1-2p) - (1-p)^3 \log(1-p) + p^3 \log p > 0.$$

We can now use this as a base case $k = 3$ in an inductive argument. For the inductive case ($k \geq 3$), suppose

$$1/2(1-2p)^k \log(1-2p) - (1-p)^k \log(1-p) + p^k \log p > 0.$$

Note that the first and third terms are negative, while the middle term is positive (because of the logs). So, let $x = \min\{1/2(1-2p)^k \log(1-2p), p^k \log p\}$. We then have

$$
\begin{aligned}
&1/2(1-2p)^{k+1}\log(1-2p) - (1-p)^{k+1}\log(1-p) + p^{k+1}\log p \\
&\geq (1-2p)x - (1-p)^{k+1}\log(1-p) + px \\
&\hspace{20em}\text{(replace both terms by their minimum)} \\
&= (1-p)x - (1-p)(1-p)^k\log(1-p) \\
&\geq (1-p)1/2(1-2p)^k\log(1-2p) - (1-p)(1-p)^k\log(1-p) + (1-p)p^k\log p \\
&= (1-p)\left[1/2(1-2p)^k\log(1-2p) - (1-p)^k\log(1-p) + p^k\log p\right] \\
&> 0. \qquad\qquad \text{(by IH)}
\end{aligned}
$$

The claim then holds for all $k \geq 3$ by induction. □

Therefore, since the map only increases in k, the basin of attraction for the stable fixed point at $1/2$ can only grow as k increases from 5. □

Theorem 42. *Suppose $F_0 \in \mathcal{F}$ is supported on $(1/4, 3/4)$. Let $\ell = [1-\sqrt{3/7}]/2 = 0.172\ldots$. For $k \geq 5$ and $x \in (F_0^{-1}(\ell), 1/2)$, $\lim_{t\to\infty} F_{k,t}(x) = 1/2$.*

Proof. Applying Lemma 12 to the bound from Lemma 14 and the convergence and monotonicity from Lemma 15 gives $\liminf_{t\to\infty} F_{k,t}(x) \geq 1/2$. Meanwhile, $F_{k,t}(x) \leq 1/2$ for all t by symmetry, so $\limsup_{t\to\infty} F_{k,t}(x) \leq 1/2$. Therefore $\lim_{t\to\infty} F_{k,t}(x) = 1/2$. □

Theorem 43. *Suppose $F_0 \in \mathcal{F}$ is supported on $(1/4, 3/4)$. For any $k \geq 2$ and $t \geq 0$,*

$$f_{k,t}(1/2) = f_0(1/2) \cdot \left[k(1/2)^{k-2}\right]^t. \tag{8.9}$$

Proof. By Lemma 5, only the left- or rightmost candidate can win. Thus, if a candidate at $1/2$ is the winner, all other candidates must either be left of $1/2$ or right of 1/2. Moreover, if all other candidates are on one side, then a candidate at 1/2 wins. Thus, a candidate at 1/2 wins if and only if all other candidates fall on the left or the right. Note that multiple candidates are at 1/2 with probability 0, since the candidate distribution is atomless. By symmetry, this occurs with probability $2 \cdot (1/2)^{k-1} = (1/2)^{k-2}$. Therefore $\Pr(\text{Plurality}(1/2, X_{2,t}, \ldots, X_{k,t})) = (1/2)^{k-2}$. By Equation (8.2), we then have

$$\begin{aligned} f_{k,t}(1/2) &= k \cdot f_{k,t-1}(1/2) \cdot \Pr(\text{Plurality}(1/2, X_{2,t}, \ldots, X_{k,t})) \\ &= k \cdot f_{k,t-1}(1/2) \cdot (1/2)^{k-2}. \end{aligned}$$

We can now prove the claim by induction on t. For $t = 0$, indeed $f_{k,0}(1/2) = f_{k,0}(1/2) \cdot \left[k(1/2)^{k-2}\right]^0$. For $t \geq 1$, applying the inductive hypothesis to the above inequality yields

$$\begin{aligned} f_{k,t}(1/2) &= k \cdot f_{k,t-1}(1/2) \cdot (1/2)^{k-2} \\ &= k \cdot f_{k,0}(1/2) \cdot \left[k(1/2)^{k-2}\right]^{t-1} \cdot (1/2)^{k-2} \\ &= f_{k,0}(1/2) \cdot \left[k(1/2)^{k-2}\right]^t. \end{aligned}$$

□

G.2.4 Proofs from Section 8.7

Theorem 44. *Suppose F_0 places probability mass p at 1/2. For any $k \geq 2$, there is some $p_k^* < 1$ such that if $p > p_k^*$, the candidate distribution converges to a point mass at 1/2 under the replicator dynamics with left–right tie-breaking. One of the fixed points of $p^k + kp^{k-1}(1-p)$ is such a p_k^*.*

Proof. Let p' denote the mass at 1/2 in generation $t+1$. If all k candidates are at 1/2, then so is the winner. Similarly, if all but one candidate are at 1/2, then the lone deviant loses with vote share less than 1/2 (with left–right tie-breaking). Thus, $p' \geq p^k + kp^{k-1}(1-p)$. For any k, this lower bound is larger than p for p sufficiently close to 1. To see this, take the derivative at $p = 1$: $\frac{d}{dp}\left[p^k + kp^{k-1}(1-p)\right] = kp^{k-1} + k(k-1)p^{k-2} - k^2p^{k-1}$, which is 0 at $p = 1$. Thus, for any small enough ϵ, $p^k + kp^{k-1}(1-p)$ is larger than $1 - \epsilon$ when evaluated at $1 - \epsilon$. Thus, p will converge to 1 by the monotone convergence theorem. □

Theorem 45. *With $k \geq 4$ and left–right tie-breaking, for any $x \in (1/4, 1/2)$, the strategy where each candidate picks uniformly at random between x and $1 - x$ is a SMSNE.*

Proof. By symmetry, every candidate has a $1/k$ win probability if they all follow this strategy. Suppose a deviant chooses a distribution that is supported on a point besides x and $1 - x$. If they choose a point between x and $1 - x$, they lose unless all other candidates pick the same side, which occurs w.p. $2(1/2)^{k-1} = 1/2^{k-2}$. For $k \geq 4$, this is at most $1/k$ (and strictly less for $k > 4$), so sampling points in $(x, 1-x)$ does not increase with probability. Alternatively, if the deviant samples a point in $[0, x)$ (or symmetrically, $(1-x, 1)$), they certainly lose unless no candidates pick x, which occurs with probability $1/2^{k-1}$—smaller than $1/k$ for $k \geq 4$. Thus deviating to a point left of x only hurts. Combining the above findings, a deviant does not benefit by sampling any point other than x or $1 - x$. Finally, a deviant does not benefit by changing the probability with which they sample either point by symmetry of the other candidates' choices. Since no deviation is beneficial, the strategy is a Nash equilibrium. □

Theorem 46. *Suppose F_0 places probability mass p at x and at $1-x$, for $1/4 < x < 1/2$. For any $k \geq 5$, there exists some $p^*_k < 1/2$ such that if $p > p^*_k$, the candidate distribution converges to point masses at x and $1-x$ under the replicator dynamics.*

In particular, one of the fixed points of $(2p)^k/2 + k(1-2p)((2p)^{k-1} - 2p^{k-1})/2$ *is such a* p^*_k.

Proof. Let p' denote the mass at x in generation $t+1$. If all candidates are at x or $1-x$ (w.p. $(2p)^k$), then a candidate at x wins with probability $1/2$ by symmetry. Alternatively, suppose all but one candidate are at x or $1-x$. The probability that there is at least one candidate at both x and $1-x$ and a wildcard is $k(1-2p)((2p)^{k-1}-2p^{k-1})$. In such a case, the wildcard loses if they are in the middle (since they get vote share less than $1/4$) and they lose if they are on the outside (to the opposite outside candidate). Thus, $p' \geq (2p)^k/2+k(1-2p)((2p)^{k-1}-2p^{k-1})/2$. For $k \geq 5$, this is larger than p for p sufficiently close to $1/2$. To see this, take the derivative at $p = 1/2$: $f'(p) = \frac{d}{dp}\left[(2p)^k/2 + k(1-2p)((2p)^{k-1} - 2p^{k-1})/2\right]$. We then find $f'(1/2) = 2^{2-k}k$, which is smaller than 1 for $k \geq 5$. Thus, p will converge to $1/2$ by the monotone convergence theorem. ☐

Theorem 48. *The following are (some[2] of the) PSNEs with uniform voters, complete plurality maximizing candidates, and left–right tie-breaking:*

1. *Any* $k \geq 2$*: all* k *candidates at* $1/2$.
2. *Any* $k \geq 4$*: for any* $x \in (1/4, 1/2)$, $\lfloor k/2 \rfloor$ *candidates at* x, $\lfloor k/2 \rfloor$ *candidates at* $1-x$*, and the last candidate (if* k *is odd) at either* x *or* $1-x$.
3. *Any* $k \geq 5$*:* $\lfloor (k-1)/2 \rfloor$ *candidates at* $1/4$, $\lfloor (k-1)/2 \rfloor$ *candidates at* $3/4$*, one candidate at* $1/2$*, and the last candidate (if* k *is even) at either* $1/4$ *or* $3/4$.
4. *Even* k*: Cox's equilibrium; two candidates at each of the points* $1/k, 3/k, \ldots, (k-1)/k$.

[2] In Appendix G.2.4, we show that for $k \leq 5$, this list of PSNEs is exhaustive (Theorem 50); for $k > 6$, there may be others.

Proof. We show in each case than no deviation is beneficial.

1. If all k candidates are at $1/2$, then the winner is chosen uniformly from the leftmost and rightmost candidate at $1/2$, who each get vote share $1/2$. If any one candidate moves to some point away from $1/2$, they get vote share strictly less than $1/2$, while the middle candidate opposite them gets vote share $1/2$ and wins. Thus, no candidate can benefit by deviating.
2. Since $k \geq 4$, both points x and $1 - x$ have at least two candidates. The candidates who end up being the outermost at x and $1-x$ each get vote share x, while the innermost candidates get vote share $1/2 - x$, which is strictly smaller since $x > 1/4$. Any candidate who moves towards the edge gets vote share strictly less than x and loses to the other side. Any candidate who moves into $(x, 1-x)$ gets vote share $1/2 - x$ and loses. Finally, no candidate benefits by moving from x to $1 - x$ (or vice-versa), since they will always be in a lottery to be the outermost which has at least as many candidates as their original point (even if k is odd and a deviant moves from the more populated point).
3. Since $k \geq 5$, both points $1/4$ and $3/4$ have at least two candidates. There is a three-way tie with vote share $1/4$ between the leftmost candidate, the rightmost candidate, and the one at $1/2$—the inner candidates at $1/4$ and $3/4$ get vote share strictly less than $1/4$. As in the previous case, every candidate certainly loses if they move left of $1/4$ or right of $3/4$. The side candidates also lose if they move into $(1/4, 3/4)$. As before, there is no benefit to switching from $1/4$ to $3/4$ or vice-versa. Finally, the candidate at $1/2$ only shrinks their win probability by moving to $1/4$ or $3/4$ (and worsens their margin against a competitor by moving to any other point in $(1/4, 3/4)$). Thus, no candidate benefits by deviating.

4. Every candidate gets vote share $1/k$ and has a chance to win. If any candidate moves, their partner will get vote share more than $1/k$ and the deviant will still have vote share at most $1/k$, so no one can deviate beneficially.

□

Lemma 17. *Any PSNE with uniform voters, complete plurality mixmizing candidates, and left–right tie-breaking must satisfy the following properties:*

(a) Any point occupied by one candidate cannot be between another occupied point and a boundary.

(b) Any point with at least three candidates must be adjacent to a boundary.

(c) Any point with two candidates not adjacent to a boundary must have the same vote share on both sides.

(d) In any two-point equilibrium, the points must be equidistant from $1/2$*.*

Proof. (a) Otherwise, the candidate can move away from the boundary to increase their vote share and decrease an opponent's vote share.

(b) Otherwise, one of the candidates could move distance ϵ either to the right or left of the point to guarantee the maximum possible vote share (instead of having probability $< 1/3$ of being on that side). This only decreases other vote shares—except the new left- or rightmost candidate created, which only has vote share $\epsilon/2$ (note this requires at least three candidates; with only two, moving increases the vote share of the partner). When adjacent to a boundary, this doesn't work—moving ϵ towards a boundary would decrease the vote share achieved (even if it's guaranteed), which could create a plurality loss, as in Equilibria 2 and 3 from Theorem 48.

(c) If not, then one of the candidates can move ϵ towards the side with higher vote share to guarantee it. For small enough ϵ, the deviant will have higher vote share than their former partner. This also decreases the vote share of the bordering candidates the deviant moved towards. Thus, this either increases the plurality win probability of the deviant or at least decreases the expected margin against the winner.

(d) Suppose not, and call the points x and y. Assume without loss of generality that $x < y$ and $x < 1-y$. Let $z = (y-x)/2$ be the vote share inner candidates get. If $z \geq y$, then a candidate at x can move right by ϵ to improve their winning chances, getting certain vote share z rather than a chance at it. If $z < 1 - y$, then a candidate at y can move right by ϵ to guarantee a win. Thus the points must be equidistant from $1/2$.

□

Theorem 50. *The following is a complete list of the PSNEs with uniform voters,complete plurality maximizing candidates, and left–right tie-breaking for small k:*

$k = 2$: $(1/2, 1/2)$

$k = 3$: $(1/2, 1/2, 1/2)$

$k = 4$:*(a)* $(1/2, 1/2, 1/2, 1/2)$

(b) $(1/4, 1/4, 3/4, 3/4)$

(c) $(x, x, 1 - x, 1 - x)$*, for any* $x \in (1/4, 1/2)$

$k = 5$:*(a)* $(1/2, 1/2, 1/2, 1/2, 1/2)$

(b) $(1/4, 1/4, 1/2, 3/4, 3/4)$

(c) $(x, x, 1 - x, 1 - x, 1 - x)$*, for any* $x \in (1/4, 1/2)$

(d) $(x, x, x, 1-x, 1-x)$, *for any* $x \in (1/4, 1/2)$.

Proof. We know by Theorem 48 that these are all Nash equilibria, so we only need to show no other equilibria exist.

$k = 2$: If a candidate is at a point other than $1/2$, then they can move to $1/2$ and do strictly better (regardless of their opponent's position), so no other equilibrium is possible.

$k = 3$: We know no point with one candidate can be adjacent to a boundary in equilibrium by Lemma 17. So all candidates must be at the same point. If that point is anything other than $1/2$, it would not be an equilibrium, so $(1/2, 1/2, 1/2)$ must be the unique equilibrium.

$k = 4$: There is no way to have a single-candidate point not adjacent to a boundary, since no partition of 4 that includes a 1 has two numbers larger than 1 to flank the single-candidate point. Thus, any equilibrium either has two points with two candidates each or one point with all four candidates. The latter type of equilibrium must be at $1/2$, so we only need to characterize the two-point equilibria.

We know by Lemma 17 that in two-point equilibria, the points must be equidistant from $1/2$ and so can be written as x and $1-x$. Now, we can show that we must have $x \in [1/4, 1/2)$. If $x < 1/4$, then a candidate at x can move right by ϵ to guarantee the winning inner vote share rather than a $1/2$ chance at it. Thus, the only two point equilibria are those claimed.

$k = 5$: A single-point equilibrium must be at $1/2$.

A two-point equilibrium cannot be a 1–4 split since the lone candidate would be adjacent to a boundary, so any two-point equilibrium must be a 2–3 split. By Lemma 17, the points must be equidistant from $1/2$, so call them x and $1-x$. We cannot have $x < 1/4$, or else a candidate at x would move right by ϵ to guarantee a winning vote share. Unlike for $k = 4$, we also cannot have $x = 1/4$. If we did, consider the point with 3 candidates. One of them could move to $1/2$ to guarantee vote share $1/4$, which would be tied for the winning share, whereas they only had a $2/3$ chance of getting that vote share before. Thus the only two-point equilibria are those claimed.

We cannot have four- or five-point equilibria, since we would then be forced to place single-candidate points adjacent to the boundary. However, we can have a three-point equilibrium with a 2-1-2 split (a 3-1-1 is impossible for the same boundary reason). So we only need to show that the claimed 2-1-2 equilibrium is the only one. First, the lone candidate must be at the midpoint of the two outer points to optimize its most competitive margin. Next, we'll show the outer points must be equidistant from the boundaries. Suppose not: say the outer points are x and y with $x < 1-y$. If the inner vote share at x ($(y-x)/4$) is smaller than x, then a candidate at x has no chance of winning. But by moving to $x-\epsilon$ for some small ϵ, they can guarantee the larger vote share and reduce their expected losing margin. If the inner vote share at x is larger than the outer vote share, then a candidate at x loses to the lone inner candidate; but again, they can move to $x+\epsilon$ improve their expected losing margin. The only remaining option is that the inner and outer shares at x are equal (so the inner share is x and the middle candidate gets vote share $2x$). In that case, consider subcases based on $1-y$. If $1-y < 2x$, then the middle candidate always wins. Since $1-y > x$, a candidate at y can move to $y+\epsilon$ to reduce their expected losing margin against the middle candidate. If $1-y > 2x$, then a candidate at y can move to $y+\epsilon$ to guarantee a win rather than a 1/2 chance. If $1-y = 2x$, then a candidate at x can move to y, giving it a chance to enter the winning lottery for vote share $2x$ (note that the candidate they leave behind at x now also gets vote share $2x$).

Now that we know the outer points are equidistant from the boundaries, the middle candidate must then be at 1/2. We can now show that the only possible outer points x and $1-x$ are given by $x = 1/4$. If $x > 1/4$, then the middle candidate cannot win; but they could move to x to join a lottery for the winning vote share. If $x < 1/4$, then a candidate at x cannot win. If the inner vote share at x is larger than x, a candidate at x can move to $x+\epsilon$ to reduce their expected losing margin against the middle candidate. Symmetrically, if x is larger than the inner vote share, then a candidate at x can move to $x-\epsilon$ to reduce their expected losing margin. Finally, consider the case where the inner and outer vote shares are equal ($x = 1/6$). A candidate at x can move into $(1/6, 1/2)$, keeping the same vote share 1/6 while reducing the vote share of the winning candidate at 1/2, thus improving their losing margin. Therefore, the only three-point equilibrium is the one claimed with $x = 1/4$.

$\square$

G.3 Formal definitions of variants

To handle non-uniform voter distributions, we define $\text{Plurality}_V(x_1, \ldots, x_k)$ to be the position of the plurality winner among $x_1, \ldots, x_k$ if the voter distribution is V.

Definition 8. Given an initial candidate distribution F_0 and a candidate count k, and a distribution of voters V, the *replicator dynamics for candidate positioning with voter distribution* V are, for all $t > 0$,

$$F_{k,t}(x) = \Pr(\text{Plurality}_V(X_{1,t}, \ldots, X_{k,t}) \leq x),$$
$$X_{i,t} \sim F_{k,t-1}, \ \forall i = 1, \ldots, k.$$

Definition 9. Given an initial candidate distribution F_0, a candidate count k, and m generations of memory, the *replicator dynamics for candidate positioning with* m *generations of memory* are, for all $t > 0$,

$$F_{k,t}(x) = \Pr(\text{Plurality}(X_{1,t}, \ldots, X_{k,t}) \leq x),$$
$$X_{i,t} \sim \begin{cases} F_{k,t-1} & \text{w.p. } \frac{1}{m} \\ \ldots & \\ F_{k,t-m} & \text{w.p. } \frac{1}{m}. \end{cases}$$

Definition 10. Given an initial candidate distribution F_0, a candidate count k, and a variance $\sigma^2 \in [0, 1]$, the *replicator dynamics for candidate positioning with* σ^2*-perturbation noise* are, for all $t > 0$,

$$F_{k,t}(x) = \Pr(\text{Plurality}(X_{1,t}, \ldots, X_{k,t}) \leq x),$$
$$X_{i,t} \sim \min(1, \max(0, F_{k,t-1} + \mathcal{N}(0, \sigma^2))), \ \forall i = 1, \ldots, k.$$

Definition 11. Given an initial candidate distribution F_0, and candidate count proportions $p_2, p_3, \ldots, p_{k_{\max}}$, the *replicator dynamics for candidate positioning with* *variable candidate counts* are, for all $t > 0$,

$$F_t(x) = \sum_{k=2}^{k_{\max}} p_k \cdot \Pr(\text{Plurality}(X_{1,t}, \ldots, X_{k,t}) \leq x),$$
$$X_{i,t} \sim F_{t-1}.$$

Let $F_{k,t}^{(i)}$ denote the distribution of the i-th place candidate generation t with k candidates per election, where $i \leq k$. We define $F_{k,0}^{(i)} = F_0$ for all k and all i, although we typically write F_0 since the initial distribution does not depend on k. Under this notation $F_{k,t}^{(1)} = F_{k,t}$ where $F_{k,t}$ is the CDF of the winner distribution. Then,

Definition 12. Given an initial candidate distribution F_0, a candidate count k, and $h \leq k$, the *replicator dynamics for candidate positioning with top-h copying* are, for all $t > 0$,

$$F_{k,t}(x) = \Pr(\text{Plurality}(X_{1,t}, \ldots, X_{k,t}) \leq x),$$
$$X_{i,t} \sim \begin{cases} F_{k,t}^{(1)} & \text{w.p. } \frac{1}{h} \\ \ldots & \\ F_{k,t}^{(h)} & \text{w.p. } \frac{1}{h}. \end{cases}$$

BIBLIOGRAPHY

Ehsan Abbasnejad, Scott Sanner, Edwin V Bonilla, and Pascal Poupart. Learning community-based preferences via Dirichlet process mixtures of Gaussian processes. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 1213–1219, 2013.

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Atila Abdulkadiroglu, Parag Pathak, Alvin E Roth, and Tayfun Sonmez. Changing the Boston school choice mechanism. Working Paper 11965, National Bureau of Economic Research, 2006. URL http://www.nber.org/papers/w11965.

Elvio Accinelli, Filipe Martins, Jorge Oviedo, Alberto Pinto, and Luis Quintas. Who controls the controller? A dynamical model of corruption. *The Journal of Mathematical Sociology*, 41(4):220–247, 2017.

Arpit Agarwal, Prathamesh Patil, and Shivani Agarwal. Accelerated spectral ranking. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 70–79. PMLR, 2018.

Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.

Greg M Allenby and Peter E Rossi. Hierarchical Bayes models. In Rajiv Grover and Marco Vriens, editors, *The handbook of marketing research: Uses, misuses, and future advances*, pages 418–440. Sage, 2006.

R Michael Alvarez and Jonathan Nagler. When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science*, 42(1):55–96, 1998.

Ofra Amir, Barbara J Grosz, Krzysztof Z Gajos, Sonja M Swenson, and Lee M Sanders. From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1419–1428. ACM, 2015.

Ammar Ammar, Sewoong Oh, Devavrat Shah, and Luis Filipe Voloch. What's your choice?: learning the mixed multi-nomial. In *ACM SIGMETRICS Performance Evaluation Review*, pages 565–566. ACM, 2014.

Simon P Anderson, Andre De Palma, and Jacques-Francois Thisse. *Discrete choice theory of product differentiation.* MIT Press, 1992.

Rie K Ando and Tong Zhang. Learning on graph with laplacian regularization. In *Advances in Neural Information Processing Systems*, 2006.

Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950.

Kenneth J Arrow. *Social Choice and Individual Values.* John Wiley & Sons, 1951.

Nanthan Atkinson and Scott C. Ganz. The flaw in ranked-choice voting: rewarding extremists. *The Hill*, 2022. URL https://thehill.com/opinion/campaign/3711206-the-flaw-in-ranked-choice-voting-rewarding-extremists/. Accessed 2022-08-30.

Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavorial Research*, 46(3):399–424, 2011.

Jonn Axsen and Kenneth S Kurani. Social influence, consumer behavior, and low-carbon energy transitions. *Annual Review of Environment and Resources*, 37: 311–340, 2012.

Manel Ayadi, Nahla Amor, Jérôme Lang, and Dominik Peters. Single transferable vote: Incomplete knowledge and communication issues. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1288–1296, 2019.

Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485, 2017.

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54, 2006.

Ricardo Baeza-Yates, Di Jiang, Fabrizio Silvestri, and Beverly Harrison. Predicting the next app that you are going to use. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 285–294, 2015.

Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80, 2018.

N Balakrishnan and Subrahmaniam Kocherlakota. On the double Weibull distribution: order statistics and estimation. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 161–178, 1985.

Abhijit V Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

John Bartholdi, Craig A Tovey, and Michael A Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6: 157–165, 1989.

Richard R Batsell and John C Polking. A new class of market share models. *Marketing Science*, 4(3):177–198, 1985.

Johann Bauer, Mark Broom, and Eduardo Alonso. The stabilization of equilibria in evolutionary game dynamics through mutation: mutation limits in evolutionary games. *Proceedings of the Royal Society A*, 475(2231):20190355, 2019.

Dorothea Baumeister, Piotr Faliszewski, Jérôme Lang, and Jörg Rothe. Campaigns for lazy voters: truncated ballots. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 577–584, 2012.

Moshe Ben-Akiva and Bruno Boccara. Discrete choice models with latent choice sets. *Int. Journal of Research in Marketing*, 12(1), 1995.

David Ben-Shimon, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle. RecSys challenge 2015 and the YOOCHOOSE dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 357–358, 2015.

Jonathan Bendor, Daniel Diermeier, David A Siegel, and Michael Ting. *A behavioral theory of elections*. Princeton University Press, 2011.

Austin R Benson and Jon Kleinberg. Found graph data and planted vertex covers. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Austin R Benson, Ravi Kumar, and Andrew Tomkins. On the relevance of irrelevant alternatives. In *Proceedings of the 25th International Conference on World Wide Web*, pages 963–973, 2016.

Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):E11221–E11230, 2018a.

Austin R Benson, Ravi Kumar, and Andrew Tomkins. A discrete choice model for subset selection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 37–45, 2018b.

B Douglas Bernheim. A theory of conformity. *Journal of Political Economy*, 102 (5):841–877, 1994.

Chandra R Bhat and Rachel Gossen. A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B: Methodological*, 38:767–787, 2004.

Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. Analysis of implicit choice set generation using a constrained multinomial logit model. *Transportation Research Record*, 2175(1), 2010.

Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34, 1948.

Duncan Black. *The theory of committees and elections.* Springer, 1958.

H. D. Block and Jacob Marschak. Random orderings and stochastic theories of responses. In Ingram Olkin, Sudhist G. Ghurye, Wassily Hoeffding, William G. Madow, and Henry B. Mann, editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford University Press, 1960.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

Michelle Blom, Vanessa Teague, Peter J Stuckey, and Ron Tidhar. Efficient computation of exact IRV margins. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 480–488. 2016.

Lawrence Blume and David Easley. Evolution and market behavior. *Journal of Economic Theory*, 58(1):9–40, 1992.

Anna Bogomolnaia and Jean-François Laslier. Euclidean preferences. *Journal of Mathematical Economics*, 43(2):87–98, 2007.

Tobias Böhmelt, Lawrence Ezrow, Roni Lehrer, and Hugh Ward. Party policy diffusion. *American Political Science Review*, 110(2):397–410, 2016.

Damien Bol, Arnaud Dellis, and Mandar Oak. Endogenous candidacy in plurality rule elections: Some explanations of the number of candidates and their polarization. *Available at SSRN 2704859*, 2016.

Massimo Bordignon, Tommaso Nannicini, and Guido Tabellini. Moderating political extremism: single round versus runoff elections under plurality rule. *American Economic Review*, 106(8):2349–70, 2016.

Tilman Börgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.

Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 197–214, 2012.

Amanda Bower and Laura Balzano. Preference modeling with context-dependent salient features. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1067–1077, 2020.

Shaun Bowler, David M Farrell, and Robin T Pettitt. Expert opinion on electoral systems: So which electoral system is "best"? *Journal of Elections, Public Opinion & Parties*, 15(1):3–19, 2005.

Markus Brill, Jonas Israel, Evi Micha, and Jannik Peters. Individual representation in approval-based committee voting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4892–4899, 2022.

William A Brock and Steven N Durlauf. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260, 2001.

David Brownstone, David S Bunch, Thomas F Golob, and Weiping Ren. A transactions choice model for forecasting demand for alternative-fuel vehicles. *Research in Transportation Economics*, (4):87–129, 1996.

Elizabeth Bruch, Fred Feinberg, and Kee Yeun Lee. Extracting multistage screening rules from online dating activity data. *Proceedings of the National Academy of Sciences*, 113(38):10530–10535, 2016.

Michael Bruter, Robert S Erikson, and Aaron B Strauss. Uncertain candidates, valence, and the dynamics of candidate position-taking. *Public Choice*, 144: 153–168, 2010.

Martin Burda, Matthew Harding, and Jerry Hausman. A Bayesian mixed logit–probit model for multinomial choice. *Journal of Econometrics*, 147(2):232–246, 2008.

Craig M Burnett and Vladimir Kogan. Ballot (and voter) "exhaustion" under instant runoff voting: An examination of four ranked-choice elections. *Electoral Studies*, 37:41–49, 2015.

Steven Callander. Bandwagons and momentum in sequential voting. *The Review of Economic Studies*, 74(3):653–684, 2007.

Randall L Calvert. Robustness of the multidimensional voting model: Candidate motivations, uncertainty, and convergence. *American Journal of Political Science*, 29(1):69–95, 1985.

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007.

Mona Chalabi. Did third-party candidates Jill Stein and Gary Johnson lose Clinton the election? *The Guardian*, 2016. URL https://www.theguardian.com/us-news/2016/nov/10/third-party-candidate-gary-johnson-jill-stein-clinton-loss. Accessed 2024-06-25.

John R Chamberlin and Michael D Cohen. Toward applicable social choice theory: A comparison of social choice functions under spatial model assumptions. *American Political Science Review*, 72(4):1341–1356, 1978.

Henry W Chappell and William R Keech. Policy motivation and party differences in a dynamic spatial model of party competition. *American Political Science Review*, 80(3):881–899, 1986.

Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 227–236, 2016a.

Shuo Chen and Thorsten Joachims. Predicting matchups and preferences in context. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Minin*, pages 775–784, 2016b.

Xi Chen, Jefrey Lijffijt, and Tijl De Bie. Quantifying and minimizing risk of conflict in social networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1197–1205. ACM, 2018.

Man-Wah Cheung. Imitative dynamics for games with continuous strategy space. *Games and Economic Behavior*, 99:206–223, 2016.

Yann Chevaleyre, Jérôme Lang, Nicolas Maudet, and Jérôme Monnot. Possible winners when new candidates are added: The case of scoring rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 762–767, 2010.

Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Discrete choice, permutations, and reconstruction. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 576–586. SIAM, 2018a.

Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Learning a mixture of two multinomial logits. In *Proceedings of the 35th International Conference on Machine Learning*, pages 961–969, 2018b.

Alex Chin, Yatong Chen, Kristen M. Altenburger, and Johan Ugander. Decoupled smoothing on graphs. In *Proceedings of the 2019 World Wide Web Conference*, pages 263–272, 2019.

Joseph A Coll. Demographic disparities using ranked-choice voting? Ranking difficulty, under-voting, and the 2020 Democratic primary. *Politics and Governance*, 9(2):293–305, 2021.

Andrew M Colman. Depth of strategic reasoning in games. *Trends in Cognitive Sciences*, 7(1):2–4, 2003.

Vincent Conitzer, Matthew Rognlie, and Lirong Xia. Preference functions that score rankings and maximum likelihood estimation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 109–115, 2009.

Clyde H Coombs. Psychological scaling without a unit of measurement. *Psychological Review*, 57(3):145, 1950.

Clyde H Coombs. *A theory of data.* Wiley, 1964.

Thomas H Cormen, Charles Eric Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms.* MIT Press, 2001.

Gary W Cox. Electoral equilibrium under alternative voting institutions. *American Journal of Political Science*, pages 82–108, 1987.

Nick Craswell, Arjen P de Vries, and Ian Soboroff. Overview of the TREC 2005 Enterprise Track. In *The Fourteenth Text REtrieval Conference Proceedings*, 2005.

Donald A Darling. On a class of problems related to the random division of an interval. *The Annals of Mathematical Statistics*, 24(2):239–253, 1953.

James M Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2): 250–273, 2014.

Nicolas de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, 1785.

André De Palma, Victor Ginsburgh, Yorgo Y Papageorgiou, and J-F Thisse. The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica*, 53(4):767–781, 1985.

Howard Dean. How to move beyond the two-party system. *The New York Times*, 2016. URL https://www.nytimes.com/2016/10/08/opinion/howard-dean-how-to-move-beyond-the-two-party-system.html. Accessed 2024-06-25.

Eddie Dekel and Michele Piccione. Sequential voting procedures in symmetric binary elections. *Journal of Political Economy*, 108(1):34–55, 2000.

Daniel DellaPosta, Yongren Shi, and Michael Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

Amaud Dellis, Alexandre Gauthier-Belzile, and Mandar Oak. Policy polarization and strategic candidacy in elections under the alternative-vote rule. *Journal of Institutional and Theoretical Economics*, 173(4):565–590, 2017.

Arnaud Dellis. Would letting people vote for multiple candidates yield policy moderation? *Journal of Economic Theory*, 144(2):772–801, 2009.

Peter M DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3):909–968, 2003.

Yvo Desmedt and Edith Elkind. Equilibria of plurality voting with abstentions. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, pages 347–356, 2010.

Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274, 2001.

Larry Diamond. The second-most important vote on Nov. 8. *Foreign Policy*, 2016. URL https://foreignpolicy.com/2016/10/13/the-second-most-important-vote-on-nov-8-trump-lepage-maine-democracy/. Accessed 2024-06-25.

Todd Donovan, Caroline Tolbert, and Kellen Gracey. Campaign civility under preferential and plurality voting. *Electoral Studies*, 42:157–163, 2016.

Jay K Dow and James W Endersby. Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral Studies*, 23(1): 107=122, 2004.

Anthony Downs. *An economic theory of democracy*. Harper & Row, 1957.

Douglas Drinen, Kathleen Grace Kennedy, and William M Priestley. An optimization problem with a surprisingly simple solution. *The American Mathematical Monthly*, 116(4):328–341, 2009.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification*. John Wiley & Sons, 2001.

John Duggan and César Martinelli. The political economy of dynamic elections: Accountability, commitment, and responsiveness. *Journal of Economic Literature*, 55(3):916–984, 2017.

Maurice Duverger. *Political parties: Their organization and activity in the modern state*. Methuen and Wiley, London and New York, 2nd edition, 1959. Translated by Barbara North and Robert North.

David Easley and Jon Kleinberg. *Networks, crowds, and markets*. Cambridge University Press, 2010.

B Curtis Eaton and Richard G Lipsey. The principle of minimum differentiation reconsidered: Some new developments in the theory of spatial competition. *The Review of Economic Studies*, 42(1):27–49, 1975.

Soroush Ebadian, Anson Kahng, Dominik Peters, and Nisarg Shah. Optimized distortion and proportional fairness in voting. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 563–600, 2022.

Edith Elkind, Martin Lackner, and Dominik Peters. Preference restrictions in computational social choice: recent progress. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 4062–4065, 2016.

Edith Elkind, Piotr Faliszewski, and Piotr Skowron. A characterization of the single-peaked single-crossing domain. volume 54, pages 167–181. Springer, 2020.

Edith Elkind, Martin Lackner, and Dominik Peters. Preference restrictions in computational social choice: A survey. *arXiv:2205.09092*, 2022.

Peter Emerson. The original Borda count and partial voting. *Social Choice and Welfare*, 40(2):353–358, 2013.

Lawrence Ezrow, Tobias Böhmelt, Roni Lehrer, and Hugh Ward. Follow the foreign leader? Why following foreign incumbents is an effective electoral strategy. *Party Politics*, 27(4):716–729, 2021.

Peter S Fader and Leigh McAlister. An elimination by aspects model of consumer response to promotion calibrated on UPC scanner data. *Journal of Marketing Research*, 27(3):322–332, 1990.

FairVote. Details about ranked choice voting, 2022. URL https://www.fairvote.org/rcv. Accessed 2023-03-15.

Timothy J Feddersen, Itai Sened, and Stephen G Wright. Rational voting and candidate entry under plurality rule. *American Journal of Political Science*, 34 (4):1005–1016, 1990.

Fred Feinberg, Elizabeth Bruch, Michael Braun, Brett Hemenway Falk, Nina Fefferman, Elea McDonnell Feit, John Helveston, Daniel Larremore, Blakeley B McShane, Alice Patania, and Mario L Small. Choices in networks: a research framework. *Marketing Letters*, 31:349–359, 2020.

John A Ferejohn and Randall L Calvert. Presidential coattails in historical perspective. *American Journal of Political Science*, 28(1):127–146, 1984.

Peter C Fishburn and Steven J Brams. Manipulability of voting by sincere truncation of preferences. *Public Choice*, 44(3):397–410, 1984.

Florida Legislature. State of Florida Chapter No. 2022-73, Senate Bill No. 524, 2022. URL http://laws.flrules.org/2022/73. Accessed 2023-03-15.

Jean Guillaume Forand. Two-party competition with persistent policies. *Journal of Economic Theory*, 152:64–91, 2014.

Jon Fraenkel and Bernard Grofman. A neo-Downsian model of the alternative vote as a mechanism for mitigating ethnic conflict in plural societies. *Public Choice*, 121:487–506, 2004.

Jon Fraenkel and Bernard Grofman. Does the alternative vote foster moderation in ethnically divided societies? The case of Fiji. *Comparative Political Studies*, 39(5):623–651, 2006a.

Jon Fraenkel and Bernard Grofman. The failure of the alternative vote as a tool for ethnic moderation in Fiji: A rejoinder to Horowitz. *Comparative Political Studies*, 39(5):663–666, 2006b.

David A Freedman and Richard A Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, 2008.

Daniel Friedman. Evolutionary games in economics. *Econometrica: Journal of the Econometric Society*, 59(3):637–666, 1991.

Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *ACM SIGIR Forum*, 51(3):32–41, 2018.

Michele Jonsson Funk et al. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

Guillermo Gallego and Huseyin Topaloglu. Constrained assortment optimization for the nested logit model. *Management Science*, 60(10):2583–2601, 2014.

Joshua S Gans and Michael Smart. Majority voting with single-crossing preferences. *Journal of Public Economics*, 59(2):219–237, 1996.

Nikhil Garg, Lodewijk L Gelauff, Sukolsak Sakshuwong, and Ashish Goel. Who is in your top three? optimizing learning in elections with many candidates. In

Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 7, pages 22–31, 2019.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 81–90. ACM, 2017.

Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, 41(4):587–601, 1973.

Vasilis Gkatzelis, Daniel Halpern, and Nisarg Shah. Resolving the optimal metric distortion conjecture. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1427–1438. IEEE, 2020.

Frank Goetzke and Tilmann Rave. Bicycle use in Germany: Explaining differences between municipalities with social network effects. *Urban Studies*, 48(2):427–437, 2011.

Chaitanya S Gokhale and Arne Traulsen. Evolutionary games in the multiverse. *Proceedings of the National Academy of Sciences*, 107(12):5500–5504, 2010.

Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. People of opposing views can share common interests. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 281–282. ACM, 2014.

Mark S Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6), 1973.

James Green-Armytage, T Nicolaus Tideman, and Rafael Cosman. Statistical evaluation of voting rules. *Social Choice and Welfare*, 46:183–212, 2016.

Bernard Grofman. Downs and two-party convergence. *Annual Review of Political Science*, 7:25–46, 2004.

Bernard Grofman and Scott L Feld. If you like the alternative vote (aka the instant runoff), then you ought to know about the Coombs rule. *Electoral Studies*, 23 (4):641–659, 2004.

Tim Groseclose. A model of candidate location when one candidate has a valence advantage. *American Journal of Political Science*, 45(4):862–886, 2001.

Bettina Grün and Friedrich Leisch. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, 25:225–247, 2008.

Harsh Gupta and Mason A. Porter. Mixed logit models and network formation, 2022.

Lars Hagen and Andrew Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design*, 11(9):1074–1085, 1992.

Daniel Halpern, Gregory Kehne, Ariel D Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich. Representation with incomplete votes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5657–5664, 2023.

Nick Hanley, Robert E Wright, and Vic Adamowicz. Using choice experiments to value the environment. *Environmental and Resource Economics*, 11(3-4): 413–428, 1998.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics, 2nd edition, 2009.

Jerry Hausman and Daniel McFadden. Specification tests for the multinomial logit model. *Econometrica*, 52(5):1219–1240, 1984.

Jerry A Hausman and David A Wise. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica*, 46(2):403–426, 1978.

Jan Hązła, Ali Jadbabaie, Elchanan Mossel, and M. Amin Rahimian. Reasoning in Bayesian opinion exchange networks is PSPACE-hard. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1614–1648, 2019.

Jan Hązła, Ali Jadbabaie, Elchanan Mossel, and M Amin Rahimian. Bayesian decision making in groups is hard. *Operations Research*, 69(2):632–654, 2021.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 639–648, 2020.

Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71 (4):1161–1189, 2003.

Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, 2003.

Christina Hoffman, Jakini Kauba, Julie Reidy, and Thomas Weighill. Proportionality in multi-winner RCV elections: A simulation study with ballot truncation. *Available at SSRN 3942892*, 2021.

Saul D Hoffman and Greg J Duncan. Multinomial and conditional logit discrete-choice models in demography. *Demography*, 25(3):415–427, 1988.

Robert E Hogan. Gubernatorial coattail effects in state legislative elections. *Political Research Quarterly*, 58(4):587–597, 2005.

Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2):026107, 2002.

Lars Holst. On the lengths of the pieces of a stick broken at random. *Journal of Applied Probability*, 17(3):623–634, 1980.

Donald L Horowitz. Strategy takes a holiday: Fraenkel and grofman on the alternative vote. *Comparative Political Studies*, 39(5):652–662, 2006.

Donald L Horowitz. Where have all the parties gone? Fraenkel and Grofman on the alternative vote–yet again. *Public Choice*, 133(1):13–23, 2007.

Harold Hotelling. Stability in competition. *The Economic Journal*, 39(153):41–57, 1929.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin Benson. Combining label propagation and simple models out-performs graph neural networks. In *9th International Conference on Learning Representations*, 2021.

Joel Huber, John W Payne, and Christopher Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9(1):90–98, 1982.

Samuel Ieong, Nina Mishra, and Or Sheffet. Predicting preference flips in commerce search. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1795–1802, 2012.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Andreas Irmen and Jacques-François Thisse. Competition in multi-characteristics spaces: Hotelling was almost right. *Journal of Economic Theory*, 78(1):76–102, 1998.

Matthew O Jackson and Brian W Rogers. Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915, 2007.

Junteng Jia and Austin R Benson. A unifying generative model for graph learning algorithms: Label propagation, graph convolutions, and combinations. *SIAM Journal on Mathematics of Data Science*, 4(1):100–125, 2022.

Junteng Jia and Austion R Benson. Residual correlation in graph neural network regression. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 588–598, 2020.

Emily M Jin, Michelle Girvan, and Mark EJ Newman. Structure of growing social networks. *Physical Review E*, 64(4):046132, 2001.

Sarah John and Andrew Douglas. Candidate civility and voter engagement in seven cities with ranked choice voting. *National Civic Review*, 106(1):25–29, 2017.

Kaggle. Personalize Expedia hotel searches - ICDM 2013, 2013a. URL https://www.kaggle.com/c/expedia-personalized-sort.

Kaggle. Allstate purchase prediction challenge, 2013b. URL https://www.kaggle.com/c/allstate-purchase-prediction-challenge. Accessed 2024-06-25.

Anson Kahng, Mohamad Latifian, and Nisarg Shah. Voting with preference intensities. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 5697–5704, 2023.

Nathan Kallus and Madeleine Udell. Revealed preference at scale: Learning personalized preferences from assortment choices. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 821–837. ACM, 2016.

Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 583–588, 2003.

Eric Kamwa. Scoring rules, ballot truncation, and the truncation paradox. *Public Choice*, 192:79–97, 2022.

Richard M Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, editors, *Complexity of Computer Computations*, pages 85–103. Springer, 1972.

Aaron Kaufman, Gary King, and Mayya Komisarchik. How to measure legislative district compactness if you only know it when you see it. *American Journal of Political Science*, 64(3):533–550, 2017.

Michael W. Kearney. Presidential election county results 2016, 2018. URL https://github.com/mkearney/presidential_election_county_results_2016. Accessed 2021-04-13.

D Marc Kilgour, Jean-Charles Grégoire, and Angèle M Foley. The prevalence and consequences of ballot truncation in ranked-choice elections. *Public Choice*, 184 (1):197–218, 2020.

Jinhee Kim, Soora Rasouli, and Harry Timmermans. Expanding scope of hybrid choice models allowing for mixture of social influences and latent attitudes: Application to intended purchase of electric cars. *Transportation Research Part A: Policy and Practice*, 69:71–85, 2014.

Jinhee Kim, Soora Rasouli, and Harry JP Timmermans. Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence. *Transport Reviews*, 38(4):499–523, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Jon Kleinberg, Sendhil Mullainathan, and Johan Ugander. Comparison-based choices. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 127–144. ACM, 2017.

Ken Kollman, John H Miller, and Scott E Page. Adaptive parties in spatial elections. *American Political Science Review*, 86(4):929–937, 1992.

Ken Kollman, John H Miller, and Scott E Page. Political parties and electoral landscapes. *British Journal of Political Science*, 28(1):139–158, 1998.

Kathrin Konczak and Jérôme Lang. Voting procedures with incomplete preferences. In *IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, volume 20, 2005.

Frank S Koppelman and Chandra Bhat. A self instructing course in mode choice modeling: multinomial and nested logit models. Technical report, 2006. URL http://www.ce.utexas.edu/prof/bhat/COURSES/LM_Draft_060131Final-060630.pdf.

Gerald H Kramer. A dynamical model of political equilibrium. *Journal of Economic Theory*, 16(2):310–334, 1977.

Martha Kropf. Using campaign communications to analyze civility in ranked choice voting elections. *Politics and Governance*, 9(2):280–292, 2021.

Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *IEEE 16th International Conference on Data Mining*, pages 221–230, 2016.

Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, pages 933–943, 2018a.

Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and VS Subrahmanian. REV2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 333–341, 2018b.

Anna-Sophie Kurella. *Issue Voting and Party Competition*, chapter The Evolution of Models of Party Competition, pages 11–25. Springer, 2017.

James P Langan. Instant runoff voting: A cure that is likely worse than the disease. *William & Mary Law Review*, 46:1569–1595, 2004.

Daniel B Larremore, Aaron Clauset, and Abigail Z Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

Jean-François Laslier and Bilge Ozturk Goktuna. Opportunist politicians and the evolution of electoral competition. *Journal of Evolutionary Economics*, 26:381–406, 2016.

Michael Laver and Kenneth Benoit. The evolution of party systems between elections. *American Journal of Political Science*, 47(2):215–233, 2003.

Geoffrey C Layman, Thomas M Carsey, and Juliana Menasce Horowitz. Party polarization in american politics. *Annual Review of Political Science*, 9:83–110, 2006.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1, 2007.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World Wide Web*, pages 641–650, 2010a.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370, 2010b.

Michael P Leung. Inference in models of discrete choice with social interactions using network data. *Available at SSRN 3446926*, 2019.

Omer Lev and Jeffrey S Rosenschein. Convergence of iterative voting. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 611–618, 2012.

Michael E Lewyn. Two cheers for instant runoff voting. *Phoenix Law Review*, 6: 117, 2012.

Faming Liang, Qifan Song, and Peihua Qiu. An equivalent measure of partial correlation coefficients for high-dimensional gaussian graphical models. *Journal of the American Statistical Association*, 110(511):1248–1265, 2015.

Q Vera Liao and Wai-Tat Fu. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 184–196. ACM, 2014.

Marji Lines. Approval voting and strategy analysis: A Venetian example. *Theory and Decision*, 20(2):155–172, 1986.

V Losert and Ethen Akin. Dynamics of games and genes: Discrete versus continuous time. *Journal of Mathematical Biology*, 17:241–251, 1983.

Li Lu, Y Connie Yuan, and Poppy Lauretta McLeod. Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review*, 16(1):54–75, 2012.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 1959.

R. Duncan Luce and Patrick Suppes. Preference, utility, and subjective probability. In R. Duncan Luce, Robert Bush, and Eugene Galanter, editors, *Handbook of Mathematical Psychology, Volume 3*, pages 249–410. Wiley, 1965.

Jiaqi Ma, Xingjian Zhang, and Qiaozhu Mei. Fast learning of mnl model from general partial rankings with application to network formation modeling. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 715–725, 2022.

Thomas R Magrino, Ronald L Rivest, Emily Shen, and David Wagner. Computing the margin of victory in IRV elections. In *Proceedings of the 2011 Conference on Electronic Voting Technology/Workshop on Trustworthy Elections*, 2011.

Rahul Makhijani and Johan Ugander. Parametric models for intransitivity in pairwise rankings. In *The World Wide Web Conference*, pages 3056–3062. ACM, 2019.

Michael Maness, Cinzia Cirillo, and Elenna R Dugundji. Generalized behavioral framework for choice models of social influence: Behavioral and data concerns in travel behavior. *Journal of Transport Geography*, 46:137–150, 2015.

Charles F Manski. The structure of random utility models. *Theory and Decision*, 8(3):229–254, 1977.

Charles F Manski and Steven R Lerman. The estimation of choice probabilities from choice based samples. *Econometrica*, 45(8):1977–1988, 1977.

Benjamin M Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. Collaborative filtering and the missing at random assumption. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 267–275, 2007.

Jacob Marschak. Binary choice constraints on random utility indicators. Cowles Foundation Discussion Papers 74, Cowles Foundation for Research in Economics, Yale University, 1959. URL https://EconPapers.repec.org/RePEc:cwl:cwldpp:74.

Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y Ozbay. Revealed attention. *American Economic Review*, 102(5):2183–2205, 2012.

Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505, 2017.

Nicholas Mattei and Toby Walsh. PrefLib: A library for preferences. In *Algorithmic Decision Theory*, pages 259–270. Springer, 2013.

Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. In *Advances in Neural Information Processing Systems*, pages 172–180, 2015.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

Daniel McFadden. Modeling the choice of residential location. In *Transportation Forecasting and Travel Behavior*, Transportation Research Record, pages 72–77. Transportation Research Board, 1978.

Daniel McFadden. Sociality, rationality, and the ecology of choice. In Stephane Hess and Andrew Daly, editors, *Choice Modelling: The State-of-the-art and The State-of-practice*, pages 1–17. Emerald Group Publishing Limited, 2010.

Daniel McFadden and Kenneth Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

Daniel McFadden, Kenneth Train, and William B Tye. An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model. In *Forecasting Passenger and Freight Travel*, Transportation Research Record, pages 39–46. Transportation Research Board, 1977.

Anthony J McGann, Bernard Grofman, and William Koetzle. Why party leaders are more extreme than their members: Modeling sequential elimination elections in the US House of Representatives. *Public Choice*, 113(3-4):337–356, 2002.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537, 2001.

Walter R Mebane Jr. Partisan messages, unconditional strategies and coordination in american elections. Revised version of a paper originally presented at the Annual Meeting of the Political Methodology Society (https://public.websites.umich.edu/~wmebane/egamesim.pdf), 2005.

Reshef Meir, Maria Polukarov, Jeffrey Rosenschein, and Nicholas Jennings. Convergence to equilibria in plurality voting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 823–828, 2010.

Reshef Meir, Omer Lev, and Jeffrey S Rosenschein. A local-dominance theory of voting equilibria. In *Proceedings of the 15th ACM Conference on Economics and Computation*, pages 313–330, 2014.

Vijay Menon and Kate Larson. Computational aspects of strategic behaviour in elections with top-truncated ballots. *Autonomous Agents and Multi-Agent Systems*, 31(6):1506–1547, 2017.

Samuel Merrill. A comparison of efficiency of multicandidate electoral systems. *American Journal of Political Science*, pages 23–48, 1984.

Russell Merris. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 197:143–176, 1994.

Paul Mitchell. The single transferable vote and ethnic conflict: the evidence from northern ireland. *Electoral Studies*, 33:246–257, 2014.

James Moody, Wendy D Brynildsen, D Wayne Osgood, Mark E Feinberg, and Scott Gest. Popularity trajectories and substance use in early adolescence. *Social Networks*, 33(2):101–112, 2011.

Sean A Munson, Stephanie Y Lee, and Paul Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 419–428, 2013.

Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 369–378. International World Wide Web Conferences Steering Committee, 2018.

Roger B Myerson and Robert J Weber. A theory of voting equilibria. *American Political Science Review*, 87(1):102–114, 1993.

Nina Narodytska and Toby Walsh. The computational impact of partial votes on strategic voting. In *European Conference on Artificial Intelligence*, pages 657–662. 2014.

Paulo Natenzon. Random choice and learning. *Journal of Political Economy*, 127 (1):419–457, 2019.

Francis Neely and Corey Cook. Whose votes count? undervotes, overvotes, and ranking in san francisco's instant-runoff elections. *American Politics Research*, 36(4):530–554, 2008.

Francis Neely and Jason McDaniel. Overvoting and the equality of voice under instant-runoff voting in San Francisco. *California Journal of Politics and Policy*, 7(4):1–27, 2015.

Richard R Nelson, Giovanni Dosi, Constance E Helfat, Andreas Pyka, Pier Paolo Saviotti, Keun Lee, Sidney G Winter, Kurt Dopfer, and Franco Malerba. *Modern evolutionary economics: An overview*. Cambridge University Press, 2018.

Richard G Niemi and William H Riker. The choice of voting systems. *Scientific American*, 234(6):21–27, 1976.

Joïlle Noailly, Jeroen CJM van den Bergh, and Cees A Withagen. Evolution of harvesting strategies: replicator and resource dynamics. *Journal of Evolutionary Economics*, 13:183–200, 2003.

NobelPrize.org. Press release: The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2000, 2000. URL https://www.nobelprize.org/prizes/economic-sciences/2000/press-release/. Accessed 2024-06-17.

Salvatore Nunnari and Jan Zápal. Dynamic elections and ideological polarization. *Political Analysis*, 25(4):505–534, 2017.

Hannu Nurmi. *Comparing voting systems*. D. Reidel, 1987.

Svetlana Obraztsova, Evangelos Markakis, Maria Polukarov, Zinovi Rabinovich, and Nicholas Jennings. On the convergence of iterative voting: how restrictive should restricted dynamics be? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pages 993–999, 2015.

Svetlana Obraztsova, Zinovi Rabinovich, Edith Elkind, Maria Polukarov, and Nicholas R Jennings. Trembling hand equilibria of plurality voting. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 440–446, 2016.

Jörg Oechssler and Frank Riedel. Evolutionary dynamics on infinite strategy spaces. *Economic Theory*, 17:141–162, 2001.

Sewoong Oh and Devavrat Shah. Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems*, pages 595–603, 2014.

Sewoong Oh, Kiran K Thekumparampil, and Jiaming Xu. Collaboratively learning preferences from ordinal data. In *Advances in Neural Information Processing Systems*, pages 1909–1917, 2015.

Yohsuke Ohtsubo and Amnon Rapoport. Depth of reasoning in strategic form games. *The Journal of Socio-Economics*, 35(1):31–47, 2006.

Martin J Osborne. Spatial models of political competition under plurality rule: A survey of some explanations of the number of candidates and the positions they take. *Canadian Journal of Economics*, pages 261–301, 1995.

Jan Overgoor, Austin Benson, and Johan Ugander. Choosing to grow a graph: Modeling network formation as discrete choice. In *Proceedings of the 2019 World Wide Web Conference*, pages 1409–1420, 2019.

Jan Overgoor, George Pakapol Supaniratisai, and Johan Ugander. Scaling choice models of relational social data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1990–1998, 2020.

Antonio Páez, Darren M Scott, and Erik Volz. A discrete-choice approach to modeling social influence on individual decision making. *Environment and Planning B: Planning and Design*, 35(6):1055–1069, 2008.

Thomas R Palfrey. Spatial equilibrium with entry. *The Review of Economic Studies*, 51(1):139–156, 1984.

Wei Pan, Nadav Aharony, and Alex Pentland. Composite social network for predicting mobile apps installation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, volume 25, pages 821–827, 2011.

Pietro Panzarasa, Tore Opsahl, and Kathleen M Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, 60(5): 911–932, 2009.

Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035, 2019.

Karlson Pfannschmidt, Pritha Gupta, Björn Haddenhorst, and Eyke Hüllermeier.

Learning context-dependent choice functions. *International Journal of Approximate Reasoning*, 140:116–155, 2022.

Charles R Plott. A notion of equilibrium and its possibility under majority rule. *The American Economic Review*, 57(4):787–806, 1967.

Keith T Poole and Howard Rosenthal. The polarization of american politics. *The Journal of Politics*, 46(4):1061–1079, 1984.

Keith T Poole and Howard Rosenthal. Patterns of congressional voting. *American Journal of Political Science*, 35(1):228–278, 1991.

Drazen Prelec, Birger Wernerfelt, and Florian Zettelmeyer. The role of inference in context effects: Inferring what you want from what is available. *Journal of Consumer Research*, 24(1):118–125, 1997.

Stephen Ragain and Johan Ugander. Pairwise choice Markov chains. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Thomas C Ratliff. A comparison of Dodgson's method and Kemeny's rule. *Social Choice and Welfare*, 18(1):79–89, 2001.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Michel Regenwetter, Aeri Kim, Arthur Kantor, and Moon-Ho R Ho. The unexpected empirical consensus among consensus methods. *Psychological Science*, 18(7):629–635, 2007.

Benjamin Reilly. Centripetalism and electoral moderation in established democracies. *Nationalism and Ethnic Politics*, 24(2):201–221, 2018.

Jeffrey Richelson. A comparative analysis of social choice functions. *Behavioral Science*, 20(5):331–337, 1975.

Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.

William H Riker. The two-party system and duverger's law: An essay on the history of political science. *American Political Science Review*, 76(4):753–766, 1982.

Robert P Rooderkerk, Harald J Van Heerde, and Tammo HA Bijmolt. Incorporating context effects into a choice model. *Journal of Marketing Research*, 48 (4):767–780, 2011.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Matthew Rosenberg, Nicole Perlroth, and David E Sanger. 'Chaos is the point': Russian hackers and trolls grow stealthier in 2020, 2020. URL https://www.nytimes.com/2020/01/10/us/politics/russia-hacking-disinformation-election.html. Accessed 2024-06-25.

Nir Rosenfeld, Kojin Oshiba, and Yaron Singer. Predicting choice with set-dependent aggregation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8220–8229, 2020.

Robert W Rosenthal. A model of far-sighted electoral competition. *Mathematical Social Sciences*, 2(3):289–297, 1982.

William E Roth. On direct product matrices. *Bulletin of the American Mathematical Society*, 40(6):461–468, 1934.

Stuart Rothenberg. How third-party votes sunk clinton, what they mean for trump, 2019. URL https://www.rollcall.com/2019/07/29/how-third-party-votes-sunk-clinton-what-they-mean-for-trump/. Accessed 2024-06-25.

Donald B Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26, 1977.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Francisco JR Ruiz, Susan Athey, and David M Blei. SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27, 2020.

Paat Rusmevichientong, Zuo-Jun Max Shen, and David B Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58(6):1666–1680, 2010.

Paat Rusmevichientong, David Shmoys, Chaoxu Tong, and Huseyin Topaloglu. Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management*, 23(11):2023–2039, 2014.

Donald G Saari and Jill Van Newenhizen. The problem of indeterminacy in approval, multiple, and truncated voting systems. *Public Choice*, 59(2):101–120, 1988.

Karolina Safarzyńska and Jeroen CJM van den Bergh. Evolutionary models in economics: a survey of methods and building blocks. *Journal of Evolutionary Economics*, 20:329–373, 2010.

Karolina Safarzynska and Jeroen CJM van den Bergh. Beyond replicator dynamics: Innovation–selection dynamics and optimal diversity. *Journal of Economic Behavior & Organization*, 78(3):229–245, 2011.

Michael Saltsman and Rebekah Paxton. Start spreading the news—ranked-choice voting is a mess. *The Wall Street Journal*, 2021. URL https://www.wsj.com/articles/start-spreading-the-newsranked-choice-voting-is-a-mess-11625178082. Accessed 2024-06-25.

Anand D Sarwate, Stephen Checkoway, and Hovav Shacham. Risk-limiting audits and the margin of victory in nonplurality elections. *Statistics, Politics, and Policy*, 4(1):29–64, 2013.

Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, 1975.

Pier Paolo Saviotti and GS Mani. Competition, variety and technological evolution: a replicator dynamics model. *Journal of Evolutionary Economics*, 5:369–392, 1995.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1670–1679, 2016.

Peter Schuster and Karl Sigmund. Replicator dynamics. *Journal of Theoretical Biology*, 100(3):533–538, 1983.

Arjun Seshadri and Johan Ugander. Fundamental limits of testing the independence of irrelevant alternatives in discrete choice. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 65–66, 2019.

Arjun Seshadri, Alex Peysakhovich, and Johan Ugander. Discovering context effects from raw choice data. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Arjun Seshadri, Stephen Ragain, and Johan Ugander. Learning rich rankings. In *Advances in Neural Information Processing Systems*, volume 33, pages 9435–9446, 2020.

Eldar Shafir, Itamar Simonson, and Amos Tversky. Reason-based choice. *Cognition*, 49(1-2):11–36, 1993.

Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

Charles R Shipan and Craig Volden. The mechanisms of policy diffusion. *American Journal of Political Science*, 52(4):840–857, 2008.

Gernot Sieg and Christof Schulz. Evolutionary dynamics in the voting game. *Public Choice*, 85(1-2):157–172, 1995.

Herbert A Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

Herbert A Simon. Rational decision making in business organizations. *The American Economic Review*, 69(4):493–513, 1979.

Itamar Simonson. Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16(2):158–174, 1989.

Itamar Simonson and Amos Tversky. Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29(3):281–295, 1992.

Anne-Marie Slaughter. How to fix polarization: Ranked choice voting, 2019. URL https://www.politico.com/interactives/2019/how-to-fix-politics-in-america/polarization/ranked-choice-voting/. Accessed 2023-04-08.

Kenneth A Small and Cheng Hsiao. Multinomial logit specification tests. *International Economic Review*, 52(5):619–627, 1985.

Dale O Stahl and Paul W Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6):1467, 1985.

Leslie S Stratton, Dennis M O'Toole, and James N Wetzel. A multinomial logit model of college stopout and dropout behavior. *Economics of Education Review*, 27(3):319–331, 2008.

Sherrie Swensen. Minutes of the political subdivisions interim committee. https://le.utah.gov/MtgMinutes/publicMeetingMinutes.jsp?Com=INTPOL&meetingId=17750, 2:06:27, 2021. Accessed 2023-03-15.

Kalyan Talluri and Garrett Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

Elie Tamer. The ET interview: Professor Charles Manski. *Econometric Theory*, 35(2):233–294, 2019.

Peter D Taylor and Leo B Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.

Tennessee Legislature. State of Tennessee Public Chapter No. 621, Senate Bill No. 1820, 2022. URL https://publications.tnsosfiles.com/acts/112/pub/pc0621.pdf. Accessed 2023-03-15.

David RM Thompson, Omer Lev, Kevin Leyton-Brown, and Jeffrey Rosenschein. Empirical analysis of plurality election equilibria. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, pages 391–398, 2013.

LL Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

Kiran Tomlinson and Austin R Benson. Choice set optimization under discrete choice models of group decisions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9514–9525. 2020.

Kiran Tomlinson and Austin R Benson. Learning interpretable feature context effects in discrete choice. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1582–1592, 2021.

Kiran Tomlinson and Austin R. Benson. Graph-based methods for discrete choice. *Network Science*, 12(1):21–40, 2024.

Kiran Tomlinson, Johan Ugander, and Austin R. Benson. Choice set confounding in discrete choice. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1571–1581, 2021.

Kiran Tomlinson, Johan Ugander, and Jon Kleinberg. Ballot length in instant runoff voting. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 5841–5849, 2023.

Kiran Tomlinson, Tanvi Namjoshi, Johan Ugander, and Jon Kleinberg. Replicating electoral success. *arXiv:2402.17109*, 2024a.

Kiran Tomlinson, Johan Ugander, and Jon Kleinberg. The moderating effect of instant runoff voting. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 9909–9917, 2024b.

Kenneth E Train. *Discrete choice methods with simulation.* Cambridge University Press, 2009.

George Tridimas. Democracy without political parties: the case of ancient athens. *Journal of Institutional Economics*, 15(6):983–998, 2019.

Jennifer S Trueblood, Scott D Brown, Andrew Heathcote, and Jerome R Busemeyer. Not just for consumers: Context effects are fundamental to decision making. *Psychological Science*, 24(6):901–908, 2013.

Amos Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281–299, 1972.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management Science*, 39(10):1179–1189, 1993.

US Congress. H.R.4464 - Ranked Choice Voting Act, 2019. URL https://www.congress.gov/bill/116th-congress/house-bill/4464. Accessed 2023-03-15.

Alexei Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67:056104, 2003.

Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pages 37–42, 2009.

Joan L Walker, Emily Ehlers, Ipsita Banerjee, and Elenna R Dugundji. Correcting for endogeneity in behavioral choice models with social influence variables. *Transportation Research Part A: Policy and Practice*, 45(4):362–374, 2011.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6638–6647, 2019.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 426–431, 2020.

Henk Jan Wassenaar, Wei Chen, Jie Cheng, and Agus Sudjianto. Enhancing discrete choice demand modeling for decision-based design. *Journal of Mechanical Design*, 127(4):514–523, 2005.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Simon Waxman. Ranked-choice voting is not the solution. *Democracy*, 2016. URL https://democracyjournal.org/arguments/ranked-choice-voting-is-not-the-solution/. Accessed 2023-08-03.

Shlomo Weber. On hierarchical spatial competition. *The Review of Economic Studies*, 59(2):407–425, 1992.

Donald Wittman. Candidates with policy preferences: A dynamic model. *Journal of Economic Theory*, 14(1):180–189, 1977.

Donald Wittman. Candidate motivation: A synthesis of alternative theories. *American Political Science Review*, 77(1):142–157, 1983.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6861–6871. PMLR, 2019.

Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. Evidence for a bimodal distribution in human communication. *Proceedings of the National Academy of Sciences*, 107(44):18803–18808, 2010.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020.

Lirong Xia. Computing the margin of victory for various voting rules. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 982–999, 2012.

Lirong Xia and Vincent Conitzer. Determining possible and necessary winners given partial orders. *Journal of Artificial Intelligence Research*, 41:25–67, 2011.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Ya Xu, Justin S Dyer, and Art B Owen. Empirical stationary correlations for semi-supervised learning on graphs. *The Annals of Applied Statistics*, 4(2):589–614, 2010.

Ye Xu, Mu Lin, Hong Lu, Giuseppe Cardone, Nicholas Lane, Zhenyu Chen, Andrew Campbell, and Tanzeem Choudhury. Preference, context and communities: a multi-faceted approach to predicting smartphone app usage patterns. In *Proceedings of the 2013 International Symposium on Wearable Computers*, pages 69–76, 2013.

Shuang-Hong Yang, Bo Long, Alexander J Smola, Hongyuan Zha, and Zhaohui Zheng. Collaborative competitive filtering: learning recommender using context of user choice. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 295–304, 2011.

John I Yellott Jr. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977.

Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 555–564, 2017.

Hao Yin, Austin R Benson, and Johan Ugander. Measuring directed triadic closure with closure coefficients. *Network Science*, 8(4):551–573, 2020.

Danqing Zhang, Kimon Fountoulakis, Junyu Cao, Michael Mahoney, and Alexei Pozdnoukhov. Social discrete choice models. *arXiv preprint arXiv:1703.07520*, 2017.

Zhibing Zhao and Lirong Xia. Learning mixtures of Plackett-Luce models from structured partial orders. In *Advances in Neural Information Processing Systems*, volume 32, pages 10143–10153, 2019.

Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of Plackett–Luce models. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2906–2914.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pages 321–328, 2003.

Kuangqi Zhou, Yanfei Dong, Kaixin Wang, Wee Sun Lee, Bryan Hooi, Huan Xu, and Jiashi Feng. Understanding and resolving performance degradation in graph convolutional networks. *arXiv preprint arXiv:2006.07107*, 2020.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02–107, 2002.