# New Paradigms in Social Choice

a dissertation presented
by
Daniel Halpern
to
The Department of Computer Science

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Computer Science

Harvard University
Cambridge, Massachusetts
June 2025

Thesis advisor: Professor Ariel D. Procaccia Daniel Halpern

# New Paradigms in Social Choice

## Abstract

Several current preference aggregation challenges — ranging from aligning large language models to summarizing civic discussions — closely resemble classical questions in social choice theory. Yet they often fall outside the scope of traditional frameworks: the outputs may be unconventional, individuals may express preferences over only a tiny fraction of alternatives, or real-world instances may exhibit much richer structure. This thesis bridges that gap by developing new theoretical frameworks, formal analyses, and algorithms tailored to these settings, offering principled approaches to collective decision-making in modern, complex environments.

**Part I** addresses AI alignment. We introduce an axiomatic framework for evaluating alignment methods and show that widely-used loss-minimization techniques violate some of the most fundamental principles of preference aggregation. We further demonstrate that aligning a single model faces inherent limitations, regardless of the algorithm used. To overcome this challenge, we propose using an ensemble of models that collectively capture a broader range of human preferences and provide both theoretical and empirical evidence for the effectiveness of this approach.

**Part II** focuses on preference elicitation in settings where individuals can feasibly provide only limited feedback across a large space of alternatives, an increasingly common restriction in contexts such as civic participation platforms. We design algorithms that elicit the *right* information for meaningful aggregation, with provable guarantees, and develop theoretical tools to characterize the fundamental trade-offs imposed by such informational constraints.

**Part III** studies liquid democracy, a flexible voting system in which individuals may delegate their votes to others. While much of the literature emphasizes worst-case failures, we propose a semi-random model that better captures realistic delegation behavior. Under mild assumptions, we show that liquid democracy avoids many of the previously-identified pitfalls. These results are further supported by real-world

experiments in classroom and corporate settings, offering empirical evidence of the paradigm's practical viability.

# Table of Contents

For Grandma and Savta.

# Acknowledgments

To Ariel — working with you has been one of the true highlights of my PhD. As an advisor, you strike a rare balance: you're always up for a quick chat and reply to emails at record speed, yet never overbearing; you're overflowing with amazing research ideas, yet always encourage us to explore and collaborate freely; you're a ping pong rival and friend, yet also a thoughtful mentor. It's clear to everyone that you're more than just a brilliant researcher, you always put your students first. When they thrive, you thrive. My PhD was, simply put, so much fun, and that's in no small part thanks to you.

To Nisarg — you're living proof that Ariel isn't just a fantastic advisor; he teaches others how to be fantastic advisors too. Your superhuman speed at grasping and generating ideas, paired with your infectious excitement, was the best introduction to research I could have hoped for. It's been a joy to grow from student, to advisee, to coauthor, to academic sibling, and soon, to colleague. Thank you for being there every step of the way.

To my committee, David and Yiling — thank you for your kindness and support throughout my PhD. You both have cultivated such a vibrant, welcoming community through the EconCS group, and I feel incredibly lucky to be a part of it.

To all of my collaborators — I genuinely look forward to the first 15 minutes of every meeting, chatting about anything and everything. Thank you for making the "work" part of the PhD feel like anything but.

To all my friends — thank you for the ski trips and game nights, pickleball matches and crossword sessions, the dinners, drinks, and, of course, a good old SEC lunch. You made this time unforgettable.

To my family — David and Sara, we did it! We're all doctors. Mom, thank you for the years of support. As the only creative in the family, our apartment would look drab without your art on the walls. More than that, you've taught me to appreciate the beauty and creativity in all aspects of life. Such as propagating plants and helping them thrive, wandering through art museums on my birthday. Thank you for being there when I need you most. And Dad, the PhD was just one chapter in a much

longer story. It started with math puzzles on hikes, sprinting home after contests to solve the problems I missed together, and grew into long conversations about research ideas and the field as a whole. You walked this road before me, but never pushed me toward it. It was just a really lucky coincidence I happened to love the same things you did.

To Chickpea and Squash, perhaps the two most different cats the world has ever seen — thank you for your tireless efforts to insert !sdf#$we into every draft. The story wouldn't be complete without it.

And finally, to my partner, Evelyn — thank you for listening patiently, watching closely, proofreading tirelessly, advising wisely, caring deeply, laughing freely, and loving unconditionally. This journey wouldn't have been the same without you and neither would I. I'm so grateful to share it all with you.

# 0 Introduction

This thesis lies within the realm of social choice theory — the study of how to aggregate individual preferences into a collective decision. For centuries, this field has offered a rich conceptual and mathematical toolbox for tackling such problems, with applications ranging from political elections to resource allocation. These foundations have given rise to a deep understanding of desirable properties for aggregation, along with fundamental trade-offs they entail.

Typical social choice frameworks, however, often rely on strong assumptions: for example, we are given a well-defined set of candidates, full access to each participant's preferences, and the goal is to simply pick one winner. These settings, while stylized, are far from trivial. They have inspired a wealth of elegant results, remain the subject of active research, and apply to a staggering number of domains. Yet, many modern preference aggregation problems, particularly those driven by rapidly evolving technologies, pose challenges that do not fit neatly into this classical mold.

Take, for example, the problem of *aligning large language models*. AI companies routinely collect large-scale datasets containing entries of the form: "Between these two answers to the prompt, which do you prefer?" This data is used to adjust the language model to better reflect these human preferences. But how should this adjustment be done? This is fundamentally a social choice problem: individuals may have differing views on which responses are better, yet their collective preferences are

distilled into a single model. But what even are principled ways to evaluate such aggregation? And how do current methods measure up?

Similarly, consider online platforms for civic participation and opinion aggregation such as *Polis*[1] [193] or *Remesh*.[2] These platforms host large-scale discussions on pressing issues, for example, "How should our city regulate Uber?" Participants submit free-text comments on the topic, and then evaluate statements submitted by others, indicating agreement or disagreement. The primary goal is to "summarize" the landscape of opinions. Again, this is fundamentally a social choice problem — we are aggregating the feedback of the participants into a single summary. In fact, it might initially appear that this problem falls well within the scope of existing methods. But there is an extra challenge: a single discussion may receive thousands of user-submitted opinions, while each participant can only feasibly evaluate a small subset. How can we meaningfully aggregate with such sparse information? Furthermore, how should these platforms select which statements to show to users in order to yield the highest-quality outcome?

These modern dilemmas underscore the need to extend and adapt the tools of social choice theory. This thesis aims to tackle these challenges head on. Building on classical notions of fairness, representation, and efficiency, it introduces new theoretical frameworks tailored to modern contexts and proposes algorithms with provable guarantees. Where possible, these approaches are evaluated on real-world datasets to validate their practicality. The overarching goal is to bridge the gap between centuries of social choice insights and the growing need for new collective decision-making systems in today's complex and information-rich world.

## 0.1 Thesis Overview

### Part I: AI Alignment

The first part of this thesis develops principled approaches to understanding AI alignment. We begin by framing alignment as a social choice problem. Specifically, we assume access to a dataset of pairwise preferences over prompt/response pairs—formally, elements of a set $\mathcal{X}$. The dataset consists of comparisons of the form $x \succ x'$ for

[1] pol.is
[2] remesh.com

$x, x' \in \mathcal{X}$.

A prominent approach to incorporating these preferences into AI systems is *reinforcement learning with human feedback (RLHF)* [157, 196, 216]. RLHF first requires aggregating the preferences into a reward function $r : \mathcal{X} \to \mathbb{R}$ that assigns a numerical score to each prompt/response pair, representing its quality. The choice of aggregation method from preferences to reward is agnostic to the rest of the pipeline — this is precisely where insights from social choice theory become valuable. RLHF then uses techniques from reinforcement learning to train the model to favor higher-reward outputs. Crucially, the reward function must generalize beyond the observed data to assign scores to all possible prompt/response pairs. This necessitates selecting a function $r_\theta$ from a parameterized family $\{r_\theta \mid \theta \in \Theta\}$. Typically, language models can produce a feature vector in $\mathbb{R}^d$ for each element $x \in \mathcal{X}$, and the class of reward functions can then be linear (or slightly more expressive) functions of these vectors. From a social choice perspective, we can evaluate the chosen $r_\theta$ by the ranking it induces over $\mathcal{X}$, allowing us to import classic techniques and desiderata from social choice theory.

Chapter 1: Axioms for AI Alignment The first chapter introduces an axiomatic framework for aggregating preferences into a reward function. Assume the pairwise comparisons arise from $n$ agents, each with a complete ranking over a finite subset $C \subseteq \mathcal{X}$. The reward function $r_\theta$ itself induces a ranking over $C$, framing the task as a quintessential social choice problem: aggregate $n$ rankings over $C$ into a single ranking.

Classic social choice desiderata now apply. For example, *Pareto Optimality*, one of the most fundamental, requires that if all $n$ agents prefer $a \succ b$, then the chosen reward $r$ should reflect this, with $r(a) > r(b)$. However, the challenge is that since the reward must come from a parameterized class (e.g., linear in the feature vectors for this chapter), we cannot simply apply a traditional voting rule.

The first results pertain to a broad class of existing approaches where the reward function is chosen to minimize a loss of the form

$$\mathcal{L}(r) = \sum_{x \succ x'} \ell(r(x') - r(x)),$$

where $\ell$ penalizes assigning higher reward to less-preferred responses. A canonical

choice is $\ell(s) = \log(1 + e^s)$, corresponding to choosing $r$ as the maximum likelihood estimate (MLE) of the *Bradley-Terry model* [49, 157]. Some approaches directly apply this loss to all elicited pairwise comparisons; others first replace repeated preferences over the same pair by the majority [44, 80, 122, 123, 134, 135, 165], i.e., if 60% prefer $a \succ b$, and 40% prefer $b \succ a$, then these are all aggregated into a single $a \succ b$ entry in the dataset.

We first show that for both of these methods, under mild conditions on $\ell$, fail to satisfy Pareto optimality, i.e., there are instances where all voters prefer $a \succ b$, but the chosen function assigns $b$ a higher reward than $a$.

We complement these results with a method that satisfies Pareto Optimality along with several other desirable social choice properties, representing a proof of concept that these requirements are not overly stringent.

Chapter 2: Pairwise Calibrated Rewards for Pluralistic Alignment A key limitation of producing a single ranking is that it cannot reflect the diversity of opinions in a population. For instance, if 60% prefer $a \succ b$ and 40% prefer $b \succ a$, a single reward must misrepresent a substantial minority.

Motivated by new ideas in *pluralistic alignment* [194], we propose a more general framework: rather than a single reward function, we select a small ensemble of reward functions that collectively capture diverse preferences. Each of these could be used to train a corresponding language model. Unlike prior work that assumes participants are grouped by predefined demographic information, we define a desirable output property, *pairwise calibration*: If a $p$ fraction of annotators prefers $a \succ b$, then approximately a $p$ fraction of the reward functions in the ensemble should have $r(a) > r(b)$.

We show a variety of theoretical results. First, finding a perfectly pairwise-calibrated ensemble is computationally hard, although, under mild conditions, a small approximately calibrated ensemble exists. Second, such ensembles can be constructed without relying on outlier reward functions that deviate too far from the average voter, i.e., even treated as single rewards they would still perform reasonably well on the population. And third, ensembles that perform well on sufficiently large datasets generalize to unseen examples.

We complement these theoretical results with experiments, demonstrating that approximate pairwise calibration is achievable on several real-world preference datasets via heuristic training methods.

## Part II: Efficient Preference Elicitation

This part addresses the increasingly pertinent challenge of preference aggregation in scenarios where it is infeasible for participants to express preferences over the entire set of alternatives $C$. In such contexts, a system must intelligently select which limited subset of comparisons or evaluations to elicit from each voter.

These challenges arise across a range of domains. In AI alignment, for instance, it is entirely impractical to ask each annotator for preferences over the full set $\mathcal{X}$. Similarly, civic engagement platforms like Polis must aggregate opinions across huge numbers of user-submitted statements, even though individual participants can evaluate only a small subset. Even in traditional domains, such as selecting awardees at academic conferences, similar challenges arise: each reviewer can asses only a small batch of papers, a tiny fraction of all submissions, yet collective decisions must be made from this sparse data.

The three chapters in this part explore these issues using diverse problem formulations, objectives, and elicitation formats.

Chapter 3: Representation with Incomplete Votes This chapter focuses on achieving fair representation in online civic platforms where users provide approval votes (agree/disagree) on a limited subset of comments. The set of alternatives $C$ represents the submitted comments.

Drawing inspiration from multi-winner approval voting, we aim to select a set of comments that is *proportional*, reflecting the diverse opinions on the population. One central criterion we consider from the literature is *Justified Representation (JR)* [14]: a selected set of $k$ comments satisfies JR if no group, comprising at least $1/k$ of the population, unanimously disagrees with all $k$ chosen comments while also unanimously agreeing with another comment not in the set. If such a group and comment exist, it suggests that this segment "deserves" to have their viewpoint included for the selection to be truly representative. JR is a relatively minimal fairness requirement; the literature also offers a spectrum of stronger variants that impose stricter guarantees.

The key challenge lies in eliciting sufficient preference information, as existing methods typically assume full knowledge of each participant's opinions over all comments—a condition clearly unmet in platforms like Polis. We model preference elicitation as follows: at each time step, the platform selects a small subset $Q \subseteq C$ of at most $t$

comments, where $t$ reflects a limit of how many comments an individual will respond to. This set is shown to the next arriving participant, who is assumed to be a random draw from the population. This participant then indicates whether they agree or disagree with each comment in $Q$.

We begin by analyzing the current approach, which, to a first approximation, selects $Q$ completely at random. We show that this approach fails to gather enough data to identify a JR set — regardless of the downstream selection algorithm — unless the number of participants exceeds the number of submitted comments by orders of magnitude above what is typically seen in real Polis discussion.

To address this, we propose an *adaptive* querying method that dynamically selects $Q$, guided by responses gathered from earlier users. This adaptive approach enables the platform to focus feedback on comments that are more likely to contribute to proportional outcomes. This method provably achieves JR and stronger fairness guarantees, significantly outperforming the non-adaptive baseline.

We support these theoretical results with experiments on real-world Polis data, demonstrating that our algorithms meet and often exceed their formal guarantees, achieving practical fairness in realistic contexts.

Chapter 4: Computing Voting Rules with Incomplete Votes In this chapter, we consider a similar preference elicitation model as in the Polis setting, but with a key difference: voters are assumed to possess *ranked* preferences over $C$. That is, each voter has an implicit complete ranking of all candidates, and, when queried on a subset, they reveal the induced ranking over that subset.

The objective here is also distinct. Rather than directly optimizing for certain representative fairness properties, we assume there is a fixed voting rule that would be applied if full rankings were available, and investigate whether it is possible to determine the winner using only limited-size queries. For example, consider the classic *plurality* rule, which selects the candidate ranked first most often. Can we recover the plurality winner by asking arriving voters to rank only a small subset of candidates?

This problem is practically motivated by settings such as political polling in large primaries, where it may be infeasible to survey voters over the entire field. Can we reliably infer the winner using only head-to-head comparisons or polls involving a few candidates?

We provide a complete characterization of the class of *positional scoring rules* that

are computable under such limited-query constraints for a candidate set $C$ of size $m$, using queries of size $t$. Notably, *the Borda count* is computable even when queries are restricted to pairwise comparisons, i.e., $t = 2$. In stark contrast, Plurality is not computable using any subset smaller than the full candidate set, i.e., $t < m$. This impossibility result is information-theoretic and holds even for randomized algorithms. This intractability also extends to *Single Transferable Vote (STV)*, another commonly used rule in political elections.

For computable rules, we further analyze the query complexity — the number of participants required to identify the correct outcome. We provide tight bounds for both deterministic and randomized settings. These results offer a comprehensive picture of what is and is not achievable under limited elicitation.

Chapter 5: Metric Distortion with Elicited Pairwise Comparisons This chapter adopts a classical model that assumes it is possible to embed voters and candidates in a metric space, where voters prefer candidates closer to them [7, 69]. This space may represent, for example, a space of opinions, such as voters and candidates being placed on a political compass for political contexts, or how much a statement emphasizes various dimensions of an issue if candidates are statements. A reasonable goal here is to choose a candidate $c \in C$ that minimizes the average distance to all voters.

However, we cannot assume access to explicit distances, especially since the metric space is often only implicitly defined. Instead, we assume each voter answers $t$ pairwise comparisons: "Which of these two candidates do you prefer?" Prior work has shown that if full rankings (all pairwise preferences) are known, it is possible to choose a candidate whose total distance is within a constant factor of the optimal, regardless of the metric [7, 84, 127].

Here, we ask what approximation guarantees are possible when each voter provides only $t$ comparisons. We derive nearly tight bounds on approximation quality as a function of $m$ and $t$, and study the impact of various forms of adaptivity. For instance, can the next query to a voter depend on their previous answers? Or, can it depend on answers from earlier voters? Each kind of (non)adaptivity has normative and practical implications, discussed more in the chapter. We establish tight bounds across all such settings, offering a fine-grained understanding of the trade-offs between elicitation cost and decision quality.

## Part III: Liquid Democracy

The final part concerns a relatively new voting paradigm known as *liquid democracy*—a system made feasible only by modern technology. Liquid democracy occupies a middle ground between two familiar systems: *representative democracy*, where officials vote on behalf of citizens, and *direct democracy*, where all individuals vote directly on every issue themselves. In liquid democracy, voters have a choice: they can either vote directly on an issue or *delegate* their vote to any other individual. The term "liquid" reflects the transitive nature of delegations: if $i$ delegates to $j$ and $j$ delegates to $k$, $k$ votes on behalf of all three. In this way, votes can "flow" through the network of delegations.

Liquid democracy has been gaining prominence in political arenas [129] and, more recently, in blockchain-based systems due to its decentralized nature [10]. Yet despite its popularity, is liquid democracy an effective method of preference aggregation from a formal, social-choice theoretic standpoint?

We are not the first to formally analyze liquid democracy (e.g., [37, 40, 50, 86, 115]). However, prior work has generally cast liquid democracy in a negative light, highlighting various shortcomings. In this part, we offer a more nuanced perspective, presenting both theoretical and empirical evidence that supports the system's effectiveness. On the theoretical side, we present a *semi-random* model of delegation. While, worst-case analysis reveals limitations, we argue that such extremes are unrealistic in large-scale implementations. Instead, when delegations include some degree of randomness—reflecting the imperfect, noisy nature of real-world decision-making—the performance of liquid democracy significantly improves.

Complementing our theoretical analysis, we report results from some of the first real-world experiments conducted on liquid democracy. These include experiments in both classroom and corporate environments, which validate key assumptions of our theoretical model and demonstrate the system's potential in practice.

**Chapter 6: Tracking Truth with Liquid Democracy** We discuss both the theoretical and empirical contributions of this chapter.

*Theoretical contributions.* Building on prior work [115], we analyze the performance of liquid democracy using the *epistemic approach* from social choice theory, which evaluates decision-making systems based by their ability to identify a correct outcome.

We assume that voters are facing a choice between two options one of which is objectively better. Each voter $i$ is characterized by an expertise level $p_i \in [0, 1]$, representing the probability that they choose the correct option. Under direct democracy, the correct outcome is selected if a majority of voters do so.

In liquid democracy, where voters may delegate their vote to others, the system's outcome is determined by a delegation graph—a directed graph where an edge from $i$ to $j$ indicates that $i$ delegated to $j$. Each voter accumulates a weight equal to the number of individuals who (transitively) delegated to them, and the final decision is made by a weighted majority. This setup enables a principled comparison between liquid and direct democracy: Given a delegation graph, we can compare the probability, over the votes cast by the individuals, that each system leads to the correct outcome.

Of course, in some instances, liquid democracy outperforms direct democracy, and in others, it underperforms. The key question, then, is: how likely is it that liquid democracy performs well?

To answer this, we introduce a semi-random model defined by three components: (1) a distribution $\mathcal{D}$ over voter expertise levels; (2) a function $q : [0, 1] \to [0, 1]$ specifying the probability that a voter with expertise $p$ chooses to delegate (ideally, lower-expertise voters are more likely to delegate); and (3) a function $\varphi : [0, 1]^2 \to \mathbb{R}_+$ encoding the propensity for a voter of expertise $x$ to delegate to a voter of expertise $y$.

With the parameters $\mathcal{D}$, $q$, and $\varphi$ fixed, we sample delegation instances and compare the performance of liquid and direct democracy. We show that under natural assumptions — such as those with more expertise being less likely to delegate, and voters tending to delegate to those with more expertise — liquid democracy not only matches direct democracy with high probability but often outperforms it significantly.

*Empirical contributions.* To validate our theoretical framework, we conducted six experiments with 168 participants in classroom and corporate settings, yielding 62 delegation graphs and more than 11,000 recorded votes. Participants answered sets of binary true/false questions and could either vote directly or delegate their vote to a peer. All experiments took place in environments with pre-existing social ties, allowing participants to make informed delegation choices. From this data, we estimated individual expertise and analyzed delegation behavior.

Empirically, we found that liquid democracy often matched or outperformed direct democracy in identifying correct outcomes. Delegation patterns aligned with our model assumptions: higher-expertise individuals were less likely to delegate, and vot-

ers tended to delegate to more competent peers. Notably, we observed little evidence of concentration of power — situations where a small number of voters accumulate a disproportionate share of the vote weight. These scenarios have been shown in prior theoretical work to severely degrade performance, yet in our experiments, such cases were rare. This suggests that the extreme cases driving negative results in worst-case analyses are not representative of how liquid democracy operates in practice.

## 0.2 Bibliographic Notes and Excluded Research

The research in this thesis is based on joint work with a variety of coauthors. Chapter 1 is based on joint work with Luise Ge, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu [82]. Chapter 2 is based on joint work with Evi Micha, Ariel D. Procaccia, and Itai Shapira [97]. Chapter 3 is based on joint work with Gregory Kehne, Ariel D. Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich [92]. Chapter 4 is based on joint work with Safwan Hossain and Jamie Tucker-Foltz [96]. Chapter 5 is based on joint work with Soroush Ebadian and Evi Micha [65]. Chapter 6 is based on joint work with Adam Berinsky, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, and Manon Revel [29], a journal paper combining two conference papers: a theoretical one with Halpern, Jadbabie, Mossel, Procaccia and Revel [94], and an experimental one with Berinsky, Revel, and Jadbabie [172].

To maintain conciseness and coherence in this thesis, a significant portion of my PhD research has been omitted. This includes work in the following areas:

- Efforts to relax stringent assumptions in social choice theory that typically lead to pessimistic outcomes, such as axiom satisfaction of voting rules under smoothed analysis [77], incentive compatibility of rules under distributional assumptions [95], and the existence of envy-free allocations in fair division under distributional assumptions [24].
- Work on tradeoffs of certain social choice objectives using partial rankings [34].
- Research in online fair division, where items arriving over time must be fairly allocated, focusing on tradeoffs of various assumptions on agent values [100], and on guarantees using partial information [25] (conference version [23]).

- Contributions toward developing new models and desiderata in multi-agent systems, including novel representation guarantees in approval voting [99]; desiderata for determining the appropriate size of representative bodies [173]; frameworks and new representation criteria for selecting interconnected citizens' assemblies [98]; and models analyzing tradeoffs between engagement and diversity in social media [21].
- Research on strategic information revelation of buyers in a market [93].

# Part I

# AI Alignment

# 1

# Axioms for AI Alignment

## 1.1 Introduction

The alignment of AI models with human values is widely recognized as a crucial task. A prominent method for this task, *reinforcement learning with human feedback* (RLHF), has been used in different applications, such as robotics [30, 131] and recommendations [1, 200]. Recently, RLHF has attracted significant attention as a tool for fine-tuning large language models (LLMs) [157, 196, 216]. A typical implementation of RLHF involves learning a reward model using a pre-trained LLM, which is then utilized to fine-tune an existing LLM. During the learning step, human feedback is provided in the form of ordinal comparisons, and a reward function is learned from these. The most common learning method assumes an underlying *random utility model* such as the *Bradley-Terry-Luce (BTL)* model [49, 157] and computes a reward function that corresponds to a maximum likelihood estimator for the observed comparisons.

Is this the "right" way of aggregating individual preferences towards a socially desirable reward function? To answer this question, we draw on *social choice theory*, a field that studies collective decison making through a mathematical lens [36]. The maximum likelihood estimation approach is in line with a storied body of work that assumes that different human participants have preferences stemming from noisy estimation of a common ground truth, and the goal is to learn this ground truth as

accurately as possible [210]. But this is not the case when it comes to questions of AI alignment, where individuals can have legitimate differences of opinion rooted in different values or priorities.

We argue that when preferences are truly heterogeneous, the *axiomatic approach*—which rose to prominence in social choice with the work of Arrow [12]—may be more suitable. This approach analyzes the desirability of aggregation methods through their satisfaction of certain axioms that capture notions of consensus, fairness, and economic efficiency. Specifically, we are interested in the axiomatic properties of aggregation methods that take ordinal preferences as input and output a reward function. We address the following two research questions:

*What axioms are satisfied by aggregation methods used by existing RLHF algorithms? And are there alternative aggregation methods that offer stronger axiomatic guarantees?*

### 1.1.1 Our Approach

In social choice theory, axioms are typically defined for rules that map rankings over candidates to a single winner (social choice functions) or a ranking of the candidates (social welfare functions). By contrast, we are interested in rules that assign a reward to each candidate. This gap is easy to bridge, though: we simply consider a ranking of the candidates by decreasing reward.

A much more significant gap is that in classical social choice, all relevant candidates appear in the input preferences, whereas in our setting (where candidates correspond, e.g., to prompts and their responses), we are only given preferences over a relatively small set of candidates *identified by their (known) features*, and we need to generalize from this information. In practice, this entails using a restricted—commonly, parametric—class of reward models which map candidate features to real-valued rewards, and which we fit to existing data.

Specifically, we assume that a *linear* reward function defined by a parameter vector determines the reward of each candidate by computing the inner product of the parameter vector and the feature vector of the candidate; these modeling choices are consistent with prior and concurrent work [83, 213, 215] and aim to capture the practice of RLHF.[1] Each human participant (henceforth referred to as a *voter*) is associ-

[1]We can represent the reward model as an embedding layer $\phi(x)$ which is then linearly ag-

ated with a parameter vector, which is unknown to us and is used to specify ordinal preferences over the candidates. Our task is to design *linear rank aggregation rules*, which aggregate rankings induced by these individual linear functions[2] into a collective ranking that is also induced by a linear function; this is a new paradigm in social choice, for which we coin the term *linear social choice*.

To evaluate linear rank aggregation rules, we adapt fundamental axioms from social choice theory [36]. The first is *Pareto optimality (PO)*, which requires that if a candidate $a$ is ranked above candidate $b$ in *every* input ranking, then the resulting ranking should rank $a$ above $b$. This is seen as a basic requirement and is satisfied by *every* standard voting method in the classical setting.

The second axiom is *pairwise majority consistency (PMC)*: If there exists a reward function that generates a ranking where, for each pair of candidates, a majority of voters agree with the ranking, then the resulting ranking should match that ranking.

This axiom is an extension of *Condorcet consistency* to rankings, and is satisfied by some, but not all, standard voting methods in the classical setting.

### 1.1.2 Our Results

We start by examining, in Section 1.3.1, a family of loss-based rules that finds a ranking induced by a parameter vector that optimizes a measure of loss; this measure increases for every disagreement with a voter on a pair of alternatives, where the larger the difference in rewards, the larger the penalty. Crucially, by plugging in binary cross-entropy loss we can recover the BTL model. Our first main result is that whenever the loss function is weakly convex and nondecreasing, or strictly convex — conditions satisfied by binary cross-entropy loss, as well as, e.g., exponential and hinge loss — the corresponding rule fails both PMC and PO. This result suggests that the prevailing practice of RLHF is flawed from an axiomatic viewpoint.

In Section 1.3.2, we take a first step towards addressing this shortcoming. We modify the loss-based formulation to focus on majority preferences instead of those of in-

gregated to compute the final reward. If we fix the embedding function $\phi$ and treat its output as the feature representation of the outcomes, such as prompt-response pairs, the resulting reward model is linear in $\phi(x)$; see, e.g. [215].

[2]In practice, typical RLHF datasets consist of pairwise comparisons, not complete rankings. Assuming rankings as input makes our exposition cleaner, and it is not a fundamental limitation, as we discuss in Section 1.5.

dividuals. This modification defines a family of rules that are PMC, but we show that all of them fail PO by establishing an even stronger impossibility result: In stark contrast to the classical setting, any *linear* rank aggregation rule that depends only on majority preferences must fail PO.

In order to achieve both PO and PMC, we design (in Section 1.4) a linear rank aggregation rule that we call *leximax Copeland subject to PO*. Not only does it satisfy our two main axioms, it also satisfies two additional ones, *majority consistency* and *winner monotonicity*.

To summarize, while widely applied rules fail to meet basic axioms, there are alternative methods that are desirable from this viewpoint. Our approach, therefore, provides a discriminative lens through which to evaluate RLHF methods and AI alignment methods more broadly.

### 1.1.3 Related Work

A number of recent papers have sought to build connections between social choice theory and RLHF [43, 55, 60, 83, 147, 158, 188, 197, 213]; this surge of interest points, in our view, to the importance of the agenda.

Three of those papers are position papers that conceptually support our work in that they discuss the possibility of applying an axiomatic approach to RLHF [55, 60, 147], although they do not provide any technical results. By contrast, existing technical papers on RLHF do not take an axiomatic approach. Of the technical papers, the one that is most closely related to ours is that of Siththaranjan et al. [188]. They show, among other results, that the ranking induced by the reward function that the MLE estimator of the Bradley-Terry-Luce Model returns follows the famous Borda count rule when unrestricted reward functions are allowed. In the classical setting, Borda count has strong axiomatic guarantees, including PO (but not PMC). However, it cannot be realized as a linear rank aggregation rule, and it is arguably impractical for RLHF.

More recently, there has been a growing emphasis on *pluralistic alignment* [55, 82, 148, 170, 194] — the development of alignment methods that account for the diversity of human values and perspectives. We discuss this emerging literature in more detail in Chapter 2, where it is more directly relevant to the contributions.

Our work builds on an earlier study by Noothigattu et al. [155], which explores the

axiomatic properties of reward functions defined as MLE estimators of underlying random utility models. The key difference is that their approach allows for general reward functions, not just linear ones, and they do not consider features at all. Unlike our findings, they show that the BTL model satisfies Pareto Optimality under these conditions. Additionally, they find that pairwise majority consistency is violated even without assuming linearity. However, their results strongly depend on varying the number of comparisons across different pairs of candidates. By contrast, our findings demonstrate that pairwise majority consistency is violated even when the number of comparisons is equal across all pairs of candidates.

## 1.2 The Linear Social Choice Model

Let $C$ be a set of $m$ distinct prompt/responses, referred to as *candidates*, and let $V = \{1, \ldots, n\}$ be a set of $n$ human participants, known as *voters*. We denote by $\mathbb{R}^d$ the $d$-dimensional real space in which both candidate feature vectors and chosen parameter vectors lie.

Each candidate $c \in C$ is associated with a distinct feature vector $\mathbf{x}_c \in \mathbb{R}^d$. A parameter vector $\theta \in \mathbb{R}^d$ induces a linear reward function $r_\theta : C \mapsto R$ defined by taking the dot product with feature vectors $r_\theta(c) = \langle \theta, \mathbf{x}_c \rangle$. We will primarily be interested in how these parameterized functions rank the candidates by reward. Let $R^{a \succ b} = \{\theta \mid r_\theta(a) \geq r_\theta(b)\}$ be the region where the reward of $a$ is at least as large as that of $b$. Note that $R^{a \succ b}$ and $R^{b \succ a}$ split $\mathbb{R}^d$ into two half spaces, separated by the hyperplane orthogonal to $\mathbf{x}_a - \mathbf{x}_b$. Parameter vectors $\theta$ on the hyperplane have $r_\theta(a) = r_\theta(b)$, while rankings in the interior of either half-space strictly rank one over the other.

For a ranking $\sigma$ over the candidates, we say that $\theta$ induces $\sigma$, denoted $\theta \triangleright \sigma$, if $a \succ_\sigma b$ implies $r_\theta(a) \geq r_\theta(b)$. Let $R^\sigma = \{\theta \mid \theta \triangleright \sigma\}$ be the set of vectors $\theta$ that induce it. Note that this can be written as the intersection of corresponding half spaces $R^\sigma = \bigcap_{a,b:a \succ_\sigma b} R^{a \succ b}$. Further, the collection of $\{R^\sigma\}$ essentially form a partition of $\mathbb{R}^d$, covering the space and intersecting only at their boundaries.

We call a $\theta$ *non-degenerate* if it is fully on one side of each of the separating hyperplanes, i.e., $r_\theta(a) \neq r_\theta(b)$ for all $a, b \in C$. Non-degenerate parameter vectors lie in the interior of some $R^\sigma$, and thus induce exactly one ranking. We call $\sigma$ *feasible* if $R^\sigma$ has

a nonempty interior, i.e., is induced by some nondegenerate $\theta$.[3]

Each voter $i \in V$ submits a ranking over the candidate $\sigma_i$. We assume that the feature space is rich enough that voter preferences can be captured via non-degenerate parameter vectors. In other words, we assume that each $\sigma_i$ is feasible. We refer to the vector of voter rankings $\pi = (\sigma_i)_{i \in V}$ as a *profile*. Further, for two candidates $a, b$, we write $n_{a \succ b}(\pi) := |\{i \in V \mid a \succ_{\sigma_i} b\}|$ for the number of voters that prefer $a$ to $b$, and $w_{a \succ b}(\pi) = n_{a \succ b}(\pi)/n$ for the proportion of such voters. When the profile $\pi$ is clear from context, we may shorten these to $n_{a \succ b}$ and $w_{a \succ b}$, respectively.

We define a *parameter aggregation rule* as a function that takes as input a profile $\pi$ and outputs a parameter vector $\theta^*$. Our goal is to design parameter aggregation rules such that $r_{\theta^*}$ satisfies desirable properties with respect to the voter preferences. However, as the properties we care about will only be with respect to how $r_{\theta^*}$ *ranks* the candidates, it will be more convenient to work with what we call *linear rank aggregation rules* that take as input a profile $\pi$ and output a *feasible* ranking $\sigma$. There is a natural way to interpret a parameter aggregation rule as a linear rank aggregation rule, namely, output any feasible ranking induced by $\theta^*$. The exact properties of the parameter aggregation rule could in principle be sensitive to the tie-breaking of non-degenerate outputs; however, all of our results will be robust to such tie-breaking.[4]

We pay special attention to a prominent family of rules from social choice theory referred to as $C1$ *rules* [75], whose outputs depend only on majority relationships, i.e., they only need to know for each pair of candidates $(a, b)$ whether the majority prefers $a$ or $b$.

[3]The number of feasible rankings is in general upper bounded by $m^{O(d)}$ due to how many regions $\binom{m}{2}$ hyperplanes can partition the space [66]. Further, under mild conditions on the feature vector locations, the exact number of feasible rankings is known [57, 87].

[4]One may wonder if there are any computational barriers to converting between parameter and linear rank aggregation rules. However, this is not the case. In particular, for *every* set of pairwise comparisons $R = \{a_1 \succ b_1, a_2 \succ b_2, a_3 \succ a_3, \ldots\}$, we can efficiently (i) check if there is a feasible ranking $\sigma$ consistent with $R$, and (ii) if such a $\sigma$ exists, find a nondegenerate $\theta$ inducing such a $\sigma$. This can be done by finding a $\theta$ satisfying the following system of linear inequalities, or determining that no such $\theta$ can satisfy them (which can be done using a linear program): $r_\theta(a) \geq r_\theta(b) + 1,\ \forall a \succ b \in R$. Any $\theta$ satisfying the system would be nondegenerate and induce a $\sigma$ consistent with $R$. Furthermore, if a feasible ranking $\sigma$ is consistent with $R$, then taking any nondegenerate $\theta$ inducing $\sigma$ and scaling up its values would satisfy all inequalities. This means that given a ranking $\sigma = c_1 \succ c_2 \succ \cdots \succ c_m$, we can check whether or not it is feasible, and so, find a $\theta$ inducing it by running this with $R = \{c_1 \succ c_2, \ldots, c_{m-1} \succ c_m\}$. Additionally, given a possibly degenerate $\theta$, we can find a feasible ranking $\sigma$ which $\theta$ induces by running this with $R = \{a \succ b \mid r_\theta(a) > r_\theta(b)\}$.

In our study, we examine several axioms borrowed from social choice theory to evaluate the reasonableness (fairness) of our aggregation mechanisms. These axioms include:

**Definition 1.2.1** (Pareto Optimality)**.** *A linear rank aggregation rule $f$ satisfies* Pareto optimality *if, whenever every voter prefers candidate $a$ over candidate $b$ on $\pi$, i.e., $w_{a\succ b}(\pi) = 1$, then candidate $a$ is ranked higher than candidate $b$ in the output ranking, i.e., $a \succ_{f(\pi)} b$.*

**Definition 1.2.2** (Pairwise Majority Consistency (PMC))**.** *A ranking $\sigma$ is called a* PMC ranking for profile $\pi$ *if for all $a, b \in C$, $a \succ_\sigma b$ if and only if a majority of voters rank $a \succ_{\sigma_i} b$, i.e., $w_{a\succ b} > 1/2$. A linear rank aggregation rule satisfies PMC if, when a PMC ranking $\sigma$ exists for the input profile $\pi$ and $\sigma$ is feasible, then $f(\pi) = \sigma$.*

Note that a PMC ranking for each $\pi$ need not exist, but when one does, it is unique. The words "$\sigma$ is feasible" allude to the possibility that no non-degenerate parameter vector $\theta$ induces the unique PMC ranking. Indeed, we have such an example; see Appendix B of the full version for details.

Our research question, then, is whether these axioms can be simultaneously satisfied by *linear* rank aggregation rules. Our approach seeks to provide a concrete illustration of how theoretical insights from social choice can inform practical algorithm design in RLHF.

## 1.3 Loss-Based Rules

### 1.3.1 Standard Loss Formulation

We begin our study of linear social choice by considering a quite broad yet natural class of rules that capture how RLHF is currently being done. Their core idea is the following: when considering parameter vector $\theta$, for each voter $i$ that ranks a pair of candidates $a \succ_i b$, we should incur some loss for giving $b$ a higher reward than $a$. To formalize this, let $\ell : \mathbb{R} \to \mathbb{R}$ be a *loss function*, which we assume is nonnegative. We can then choose a parameter vector minimizing

$$\mathcal{L}(\theta; \pi, \ell) = \sum_{a \neq b \in C} n_{a\succ b}(\pi) \cdot \ell(r_\theta(b) - r_\theta(a)).$$

Note that the BTL model fits within this framework using $\ell(x) = \ln(1 + e^x)$, i.e., *binary cross-entropy loss*.[5] One caveat to this approach, however, is that an optimal $\theta$ need not be well-defined: it is possible that no minimum is attained. Fortunately, since we only care about rankings induced by optimal parameter vectors, we can conveniently remedy this by saying the output is any ranking that is induced by parameter vectors that are arbitrarily close to optimal. More formally, we say that a linear rank aggregation rule $f$ *minimizes* $\ell$ if for all $\sigma = f(\pi)$,

$$\inf_{\theta:\theta\in R^\sigma} \mathcal{L}(\theta; \pi, \ell) = \inf_{\theta} \mathcal{L}(\theta; \pi, \ell). \tag{1.1}$$

Even if no minimum is attained, there is always a choice of feasible ranking $\sigma$ such that Equation (1.1) is satisfied.

With this definition in hand, we proceed to our first main result, which spells rather bad news for this class of rules: *Any* loss-based aggregation rule using a nondecreasing and convex loss function (of which BTL is one, and hinge loss is another) will fail our two core axioms, PMC and PO. This paints a negative picture for current RLHF methods with respect to their social choice guarantees. Note that we will exclude the discussion of loss functions that have its global minimum at zero like ReLU because the loss minimizer will be zero, making all rankings vacuous consequently. And we have focused on convex loss functions due to its practical ease in optimization.

**Theorem 1.3.1.** *If a linear rank aggregation rule $f$ optimizes a loss function $\ell$ that satisfies $\inf_x \ell(x) < \ell(0)$ and is either nondecreasing and weakly convex, or strictly convex (and possibly nonmonotone), then $f$ fails PMC and PO.*

*Proof.* Fix a loss function $\ell$ satisfying the theorem conditions. Note that since $\ell$ is convex, we may also assume it is continuous [174, Corollary 10.1.1]. Furthermore, since $\inf_x \ell(x) < \ell(0)$, we know that there exists $x \neq 0$ such that $\ell(x) < \ell(0)$. The case where $x > 0$ is relatively simple (as such loss functions lead to unnatural behavior), and we handle it at the end of the proof. For now, we assume that there exists $x < 0$ such that $\ell(x) < \ell(0)$. Note that this also implies that for all $y \geq 0$, $\ell$

[5] Typically, BTL is presented as choosing $\theta$ maximizing the likelihood of seeing the pairwise comparisons we observed, assuming that $\Pr[a \succ b] = \frac{e^{r_\theta(a)}}{e^{r_\theta(a)}+e^{r_\theta(b)}}$. That is, we choose $\theta$ maximizing $\prod_{a\neq b} \left(\frac{e^{r_\theta(a)}}{e^{r_\theta(a)}+e^{r_\theta(b)}}\right)^{n_{a\succ b}}$. By taking the log and swapping the sign, we see that this is equivalent to minimizing $\sum_{a\neq b} \log\left(1 + e^{r_\theta(b)-r_\theta(a)}\right)$.

is lower bounded by the affine linear function connecting $(x, \ell(x))$ and $(0, \ell(0))$, and thus, $\lim_{x\to\infty} \ell(x) = \infty$.

We begin with a small instance of just three candidates $C^{core} = \{a, b, c\}$ to gain some traction on how $\ell$ behaves. We will later extend this instance with additional candidates to demonstrate a profile where PO and PMC fail. The candidates will have feature vectors $\mathbf{x}_a := (2, 1)$, $\mathbf{x}_b := (1, 1)$, and $\mathbf{x}_c := (0, 0)$, respectively. Furthermore, a $p$-fraction of voters (for $p$ to be chosen later) will rank $a \succ b \succ c$, while the remaining $(1-p)$-fraction will have inverted preferences, ranking $c \succ b \succ a$.[6]

Let

$$\mathcal{L}^{core}(\theta) := \sum_{x\neq y\in C^{core}} w_{x\succ y}\ell(r_\theta(y) - r_\theta(x))$$

with $w_{x\succ y} \in \{1-p, p\}$ be the loss function on this instance (scaling $n_{x\succ y}$ down to $w_{x\succ y}$ leads to an equivalent formulation). Let $g(x) = p \cdot \ell(-x) + (1-p)\ell(x)$. Note that we can rewrite $\mathcal{L}^{com}$ as

$$\mathcal{L}^{core}(\theta) = g(r_\theta(a) - r_\theta(b)) + g(r_\theta(a) - r_\theta(c)) + g(r_\theta(b) - r_\theta(c)).$$

Note that $r_\theta(c) = 0$ for all $\theta$, so we can simplify this to

$$\mathcal{L}^{core}(\theta) = g(r_\theta(a) - r_\theta(b)) + g(r_\theta(a)) + g(r_\theta(b)).$$

We will consider an unconstrained version of this problem where we are free to choose rewards $r_a, r_b \in \mathbb{R}$ arbitrarily, and later show by which vectors $\theta$ these optimal values can be induced. That is, we will first find $r_a, r_b \in \mathbb{R}$ minimizing

$$\mathcal{L}^{unconstr}(r_a, r_b) := g(r_a - r_b) + g(r_a) + g(r_b).$$

Let $OPT^{core} = \{\theta \mid \mathcal{L}^{core}(\theta) = \inf_{\theta'} \mathcal{L}^{core}(\theta')\}$ and $OPT^{unconstr}\{(r_a, r_b) \mid \mathcal{L}^{unconstr}(r_a, r_b) = \inf_{r'_a, r'_b} \mathcal{L}^{unconstr}(r'_a, r'_b)\}$ be the set of minimizers for these two loss functions. We now establish the following results about these optimal sets.

**Lemma 1.3.2.** *There exists a rational $p \in (1/2, 1]$ and values $A_1 < A_2$ with $A_2 > 0$ such that $OPT^{unconstr}$ is nonempty and for all $(r_a, r_b) \in OPT^{unconstr}$, $r_a > A_2$ and*

[6] It so happens that these rankings are feasible, but for now we will not worry about this as loss function-based rules still make sense regardless of whether the inputs are feasible. For the final example, we will ensure that the rankings are feasible.

$r_b \leq A_1$.

**Lemma 1.3.3.** *Suppose Lemma 1.3.2 holds for values $p, A_1$ and $A_2$, then, for this same choice of $p$, $OPT^{core}$ is nonempty and there exist $A_3$ and $A_4$ with $A_3 > 0$ such that for all $(\theta_1, \theta_2) \in OPT^{core}$, $\theta_1 > A_3$ and $\theta_2 < A_4$.*

*Proof of Lemma 1.3.2.* We begin with some observations about $g$. First, we have that since $\ell$ is nonnegative, $g$ is nonnegative. This along with the fact that $\lim_{x\to\infty} \ell(x) = \infty$, we have that both $\lim_{x\to\infty} g(x) = \infty$ and $\lim_{x\to-\infty} g(x) = \infty$. Together, these imply that $\mathcal{L}^{unconstr}$ attains a minimum. Indeed, $\mathcal{L}^{unconstr}(0,0) = 3g(0)$, and there is some bound $B$ such that for all $x > B$ and $x < -B$, $g(x) > 3g(0)$. We can therefore restrict the optimization problem to $r_a, r_b \in [-B, B]$ without changing the solutions. Since $\mathcal{L}^{unconstr}$ is continuous and $[-B, B]^2$ is compact, a minimum is attained.

Next, note that $g$ is convex because compositions of convex functions with monotonic functions and convex combinations of convex functions are convex [174]. From this, we claim that if there is an optimal solution $(r_a, r_b)$, then $(r_a, r_a/2)$ is also an optimal solution. Indeed, fix such an $(r_a, r_b)$,

$$\begin{aligned}
\mathcal{L}^{unconstr}(r_a, r_a/2) &= g(r_a - r_a/2) + g(r_a) + g(r_a/2) \\
&= g(r_a) + 2g(r_a/2) \\
&= g(r_a) + 2g(1/2(r_a - r_b) + 1/2r_b) \\
&\leq g(r_a) + 2(1/2g(r_a - r_b) + 1/2g(r_b)) \\
&= \mathcal{L}^{unconstr}(r_a, r_b),
\end{aligned}$$

where the inequality comes from convexity. This implies that $(r_a, r_a/2)$ is also optimal.

By above, we have that if $(r_b, r_a)$ is optimal, it must be the case that $r_a$ minimizes

$$h(r_a) := 2g(r_a/2) + g(r_a).$$

Observe that $h$ is again convex by monotonic composition and convex combinations.

Next, we will make use of the following facts about convex functions. Although they need not be differentiable, right- and left-hand derivatives always exist. For a

function $k$ these are defined as

$$k'_+(x) = \lim_{h \to 0^+} \frac{k(x+h) - k(x)}{h}$$
$$k'_-(x) = \lim_{h \to 0^-} \frac{k(x+h) - k(x)}{h}.$$

Further, if $k$ is convex, we have that (i) $k'_+$ and $k'_-$ are nondecreasing, (ii) for every $x$, $k'_-(x) \leq k'_+(x)$, (iii) $x$ minimizes $k$ if and only if $k'_-(x) \leq 0 \leq k'_+(x)$, and (iv) $k'_+$ and $k'_-$ are right and left continuous, respectively [174]. They also follow standard linearity and chain rule properties, which allow for simpler computation, e.g., if $k(x) = a\alpha(x) + b\beta(x)$, then $k'_+(x) = a\alpha'_+(x) + b\beta'_+(x)$ and if $k(x) = \alpha(\gamma x)$, then $k'_+(x) = \gamma\alpha'_+(\gamma x)$ if $\gamma \geq 0$, and $k'_+(x) = \gamma\alpha'_0(\gamma x)$ (all of these for left-hand derivatives swapping the places of $+$ and $-$) [52].

Next, we claim that we can find a valid $p$ (rational with $1/2 < p < 1$) and $w > 0$ such that $g'_+(w) > 0$, while $h'_+(w) < 0$. To that end, expanding the first derivative, we have

$$g'_+(x) = -p\ell'_-(-x) + (1-p)\ell'_+(x).$$

As long as $\ell'_-(-x) + \ell'_+(x) > 0$, this is strictly more than 0 for $p$ satisfying

$$p < \frac{\ell'_+(x)}{\ell'_-(-x) + \ell'_+(x)}. \tag{1.2}$$

For $h$,

$$\begin{aligned} h'_+(x) &= 2 \cdot 1/2 g'_+(x/2) + g'_+(x) \\ &= -p\ell'_-(-x/2) + (1-p)\ell'_+(x/2) - p\ell'_-(-x) + (1-p)\ell'_+(x). \\ &= -p\big[\ell'_-(-x/2) + \ell'_-(-x)\big] + (1-p)\big[\ell'_+(x/2) + \ell'_+(x)\big]. \end{aligned}$$

As long as $\ell'_-(-x/2) + \ell'_-(-x) + \ell'_+(x/2) + \ell'_+(x) > 0$, then this is strictly less than 0 for

$$p > \frac{\ell'_+(x/2) + \ell'_+(x)}{\ell'_-(-x/2) + \ell'_-(-x) + \ell'_+(x/2) + \ell'_+(x)}. \tag{1.3}$$

Now, choose $w > 0$ such that $\ell'_-(-w/2) > \ell'_-(-w) \geq 0$. This is possible by using the following procedure. We know $\ell'_-(0) > 0$ (as otherwise $\ell(0) < \ell(x)$ for all $x < 0$, contradicting our assumption on $\ell$). We will split into cases depending on whether $\ell$ is nondecreasing or strictly-convex (at least one must be true by the theorem assumptions).

First, suppose $\ell$ is non-decreasing. This implies that $\ell'_-(x) \geq 0$ for all $x$. We know that there is a point $x < 0$ such that $\ell'_-(x) < \ell'_-(0)$ (as otherwise $\ell$ would eventually become negative). Let $d = \ell'_-(x)$. Take $w$ such that $-w = \sup\{x \mid \ell'_-(x) = d\}$. Note that $\ell'_-(-w) = d$ because $\ell'_-$ is left continuous (from (iv) above). Since $-w < 0$, so $-w/2 > -w$. This implies that $\ell'_-(-w/2) > d$, as otherwise $-w$ would not be the supremum of such points. In addition, we have that $d \geq 0$ because $\ell$ is nondecreasing.

Next, suppose $\ell$ is strictly convex. then $\ell'_-$ is strictly increasing. Further, since it is left continuous and $\ell'_0(0) > 0$, there is a $\gamma > 0$ such that for all $x \in [-\gamma, 0]$, $\ell'_-(x) \geq 0$. Therefore, choosing $w = \gamma$ will do: $\ell'_-(-\gamma) \geq 0$ by choice of $\gamma$, and $-\gamma/2 > -\gamma$, so $\ell'_-(-\gamma/2) > \ell'_-(-\gamma)$ since $\ell'_-$ is strictly increasing.

For this choice of $w$, we first claim that the preconditions of denominators being positive hold for (1.2) and (1.3). Indeed, let us write $z_1, z_2, z_3, z_4$ for $\ell'_-(-2w)$,$\ell'_-(-w)$, $\ell'_+(w)$, $\ell'_+(2w)$. We know that $z_1 \leq z_2 \leq z_3 \leq z_4$ by properties of convexity, and since $\ell'_-(-w/2) > \ell'_-(-w) \geq 0$, we have that $0 \leq z_1 < z_2$. The denominators are of the form $z_1 + z_4$ and $z_1 + z_2 + z_3 + z_4$, which are now both necessarily positive. In addition, we also claim that a rational $p$ with $1/2 < p < 1$ satisfying both inequalities (1.2) and (1.3) will exist. Note that inequality (1.2) can now be represented as $\frac{z_4}{z_1+z_4}$, and we have that $\frac{z_4}{z_1+z_4} \leq 1$ because $0 \leq z_1 < z_4$ Additionally, inequality (1.3) can be represented as $\frac{z_3+z_4}{z_1+z_2+z_3+z_4}$ which is at least $1/2$ because $z_3 + z_4 > z_1 + z_2$. Finally, we have

$$\frac{z_3}{z_2 + z_3} \leq \frac{z_4}{z_2 + z_4} < \frac{z_4}{z_1 + z_4}$$

which implies that

$$\frac{z_3 + z_4}{z_1 + z_2 + z_3 + z_4} < \frac{z_4}{z_1 + z_4}.$$

Hence, there exists some rational $p$ in this interval, which is necessarily between $1/2$ and 1, as needed.

To summarize, we have found a valid $p$ and value $w > 0$ such that $h'_+(w) < 0$ while $g'_+(w) > 0$. Because $h'_+$ is right continuous (from (iv) above), there is some $\gamma > 0$ such that $h'_+(w+\gamma) < 0$ as well. We will now show for all $(r_a, r_b) \in OPT^{unconstr}$, $r_a > w+\gamma$

and $r_b \leq w$. Indeed, note that $r_a$ must minimize $h$, so $h'_+(r_a) \geq 0$ (from (iii) above), which implies $r_a > w + c$. For $r_b$, suppose for a contradiction $r_b > w$ as well. Note that $g'_+(w) > 0$ means $g$ is increasing to the right of $w$. Let $d = \min(r_a, r_b) - w > 0$. Consider $(r'_a, r'_b) = (r_a - d, r_b - d)$. We then have

$$\begin{aligned}
\mathcal{L}^{unconstr}(r'_a, r'_b) &= g(r'_a - r'_b) + g(r'_a) + g(r'_b) \\
&= g(r_a - r_b) + g(r'_a) + g(r'_b) \\
&< g(r_a - r_b) + g(r_a) + g(r_b) \\
&= \mathcal{L}^{unconstr}(r_a, r_b).
\end{aligned}$$

where the equality holds because $r'_a - r'_b = r_a - r_b$ and the inequality because $g$ is increasing to the right of $w$. Therefore, we reach a contradiction, because then $(r_a, r_b)$ would not be optimal. Thus, we have found values satisfying the lemma statement with $A_1 = w$ and $A_2 = w + \gamma$. □

*Proof of Lemma 1.3.3.* Fix $p$ inducing a nonempty $OPT^{unconstr}$ satisfying Lemma 1.3.2 with values $A_1$ and $A_2$. We first claim that for any $(r_a, r_b)$ (regardless of optimality), it is possible to find a $\theta$ such that $r_\theta(a) = r_a$ and $r_\theta(b) = r_b$. Indeed, note that

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} r_\theta(a) \\ r_\theta(b) \end{pmatrix}.$$

Since $M = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ is invertible with inverse $M^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$, for any $r_a, r_b$, we can simply set $\theta = M^{-1} \begin{pmatrix} r_a \\ r_b \end{pmatrix}$. Thus, $OPT^{com}$ is nonempty, and is simply the image of $OPT^{unconstr}$ under $M^{-1}$. Now, fix $(r_a, r_b) \in OPT^{unconstr}$, and let $\theta = M^{-1} \begin{pmatrix} r_a \\ r_b \end{pmatrix}$. By assumption, we have $r_a > A_2$ and $r_b \leq A_1$. This implies that $\theta_1 = r_a - r_b > A_2 - A_1$, while $\theta_2 = 2r_b - r_a < 2A_1 - A_2$. Thus, setting $A_3 = A_2 - A_1$ and $A_4 = 2A_1 - A_2$ satisfy the desired properties. □

We will now explicitly construct a family of instances with candidate feature vectors parameterized by a value $\varepsilon \in \mathbb{R}$ such that for sufficiently small $\varepsilon > 0$, the output of $f$ fails the two axioms. Fix $p, A_3$ and $A_4$ from Lemma 1.3.3, and choose $\delta$ with

$0 < \delta < 1$ such that $\delta A_4 - A_3 < 0$ ($\delta < A_3/A_4$ works if $A_4 > 0$, and otherwise, any $0 < \delta < 1$ will do).

Each instance will have six candidates, which we will think of as two groups of three, $C = C^{core} \cup C^{copies}$. The first group $C^{core} = \{a, b, c\}$ will be the same as the three-candidate instance from above, while the second group $C^{copies} = \{a', b', c'\}$ will be new. The candidates $a, b, c$ will still be located at $\mathbf{x}_a := (2, 1)$, $\mathbf{x}_b := (1, 1)$, and $\mathbf{x}_c := (0, 0)$, respectively. The candidates $a'$, $b'$, $c'$ will be located near their undecorated counterparts at $\mathbf{x}_{a'} := \mathbf{x}_a + (-\varepsilon, 0)$, $\mathbf{x}_{b'} := \mathbf{x}_b + (-\varepsilon, 0)$ and $\mathbf{x}_{c'} := \mathbf{x}_c + (-\varepsilon, \delta \cdot \varepsilon)$.

Next, we describe the voter preferences. A $p$-fraction of voters will have the ranking $a \succ a' \succ b \succ b' \succ c' \succ c$, and the remaining $(1 - p)$-fraction of voters will have ranking $c' \succ c \succ b' \succ b \succ a' \succ a$. As long as $0 < \varepsilon < 1$ (which will be the case for our final chosen $\varepsilon$), these are both feasible rankings. The former is induced by the nondegenerate feature vector $(\delta, 2)$[7] and the latter by $(-1, 0)$.[8]

For each $\varepsilon \in \mathbb{R}$, let

$$\mathcal{L}^{\varepsilon}(\theta) = \sum_{x \neq y \in C} w_{x \succ y} \ell(r_\theta(y) - r_\theta(x))$$

with $w_{x \succ y} \in \{0, 1 - p, p, 1\}$ be the loss function we are optimizing using candidate locations parameterized by $\varepsilon$.

We will show that for sufficiently small $\varepsilon > 0$, $\inf_{\theta \in R^{c' \succ c}} \mathcal{L}^{\varepsilon}(\theta) > \inf_\theta \mathcal{L}^{\varepsilon}(\theta)$. This means that $f$ must output a ranking with $c \succ c'$. Observe that this is a PO violation because all voters agree that $c' \succ c$. Furthermore, this is a PMC violation because a majority of voters have the ranking $a \succ a' \succ b \succ b' \succ c' \succ c$, yet this is not the output.

Let $OPT(\varepsilon)$ be the set of vectors optimizing $\mathcal{L}^{\varepsilon}$. The rest of the proof will follow from the following two lemmas.

**Lemma 1.3.4.** *$OPT(0) \subseteq \overline{R^{c' \succ c}}$.*

**Lemma 1.3.5.** *Suppose $OPT(0) \subseteq \overline{R^{c' \succ c}}$, then, for sufficiently small $\varepsilon > 0$,*

$$\inf_{\theta : \theta \in R^{c' \succ c}} \mathcal{L}^{\varepsilon}(\theta) > \inf_\theta \mathcal{L}^{\varepsilon}(\theta).$$

[7] This induces rewards $2 + 2\delta, 2 + (2 - \varepsilon)\delta, 2 + \delta, 2 + (1 - \varepsilon)\delta, \delta\varepsilon$ and 0 for $a, a', b, b', c'$, and $c$, respectively.

[8] This induces rewards $\varepsilon, 0, -1 + \varepsilon, -1, -2 + \varepsilon$ and $-2$. for $c', c, b', b, a'$, and $a$, respectively.

*Proof of Lemma 1.3.4.* We first claim that $OPT(0) = OPT^{core}$. This will follow from showing

$$\mathcal{L}^0(\theta) = 4 \cdot \mathcal{L}^{core}(\theta) + 3 \cdot \ell(0).$$

Indeed, in $\mathcal{L}^0$, the copied candidates are in exactly the same location as their counterparts. Hence, each term in $\mathcal{L}^{core}$ appears 4 times, one for each combination of original and copy. In addition to these, there are the 6 terms for each ordered pair of $(a, a')$, $(b, b')$, and $(c, c')$. Note that, each $r_\theta(x) = r_\theta(x')$ for each $x \in \{a, b, c\}$ regardless of $\theta$ since they are in the same location. Therefore, the $\ell$ portion is always $\ell(0)$, and the corresponding $w_{x \succ x'}$ and $w_{x' \succ x}$ terms add up to one for each pair. Hence, the total sum of these terms is $3 \cdot \ell(0)$. Since $\mathcal{L}^0$ is equivalent to $\mathcal{L}^{core}$ up to positive scaling and translation, they have the same optima.

Finally, fix a $\theta \in OPT(0)$. Since $\theta \in OPT^{core}$, by Lemma 1.3.3, $\theta_1 > A_3$ and $\theta_2 < A_4$. Thus,

$$r_\theta(c') = \varepsilon \cdot (-\theta_1 + \delta\theta_2) < \varepsilon(-A_3 + \delta A_4) < 0 = r_\theta(c)$$

by choice of $\delta$. Hence, $\theta \in \overline{R^{c' \succ c}}$ □

*Proof of Lemma 1.3.5.* We first show that when optimizing each $\mathcal{L}^\varepsilon$, it is sufficient to consider only $\theta$ coming from a bounded region. Indeed, observe that $\mathcal{L}^\varepsilon(\mathbf{0}) = \binom{6}{2}\ell(0)$ for all $\varepsilon$. Since $\lim_{x \to \infty} \ell(x) = \infty$, we can find some $B > 0$ such that for all $x > B$, $\ell(x) > \frac{\binom{6}{2}\ell(0)}{1-p} = \frac{\mathcal{L}^\varepsilon(\mathbf{0})}{1-p}$. For a pair of candidates $x \neq y \in C^{com}$, in the two terms concerning these candidates, we have

$$\begin{aligned}
&w_{x \succ y}\ell(r_\theta(y) - r_\theta(x)) + w_{y \succ x}\ell(r_\theta(x) - r_\theta(y)) \\
&\geq (1-p)\left(\ell(r_\theta(y) - r_\theta(x)) + \ell(r_\theta(x) - r_\theta(y))\right) \\
&\geq (1-p)\ell(|r_\theta(y) - r_\theta(x)|).
\end{aligned}$$

Applying this to $\{a, b\}$ and $\{b, c\}$,

$$\begin{aligned}
\mathcal{L}^\varepsilon(\theta) &\geq (1-p)\left(\ell(|r_\theta(a) - r_\theta(b)| + \ell(|r_\theta(b)) - r_\theta(c)|)\right) \\
&= (1-p)(\ell(|\theta_1|) + \ell(|\theta_1 + \theta_2|)
\end{aligned}$$

This implies that we may restrict our attention to $\theta$ in the region $R^{bounded} = \{\theta \mid$

$|\theta_1| \leq B, |\theta_2| \leq 2B\}$. Indeed, for $\theta \notin R^{bounded}$, either $|\theta_1| > B$ or $|\theta_1 + \theta_2| > B$. In either case, we have $\mathcal{L}^{\varepsilon}(\theta) \geq (1-p)\ell(B) > \mathcal{L}^{\varepsilon}(\mathbf{0})$.

Note that $\mathcal{L}^{\varepsilon}(\theta)$ is continuous not only in $\theta$, but also in $\varepsilon$. Additionally, $R^{bounded}$ is closed and bounded, and hence, compact. Therefore, by Berge's Maximum Theorem, $OPT(\varepsilon)$ is nonempty and *upper semi-continuous* in $\varepsilon$ [28]. As per the definition of upper semi-continuous, since $OPT(0) \subseteq \overline{R^{c' \succ c}}$, an open set, for sufficiently small $\varepsilon > 0$, $OPT(\varepsilon) \subseteq \overline{R^{c' \succ c}}$. Finally, note that $R^{c' \succ c} \cap R^{bounded}$ is compact, so a minimum is attained, and this minimum must therefore be strictly larger than the values attained by members of $OPT(\varepsilon)$. □

Finally, we handle the case that exists $x > 0$ such that $\ell(x) < \ell(0)$. Note that by convexity, this implies that for all $y < 0$, $\ell(y) > \ell(0) > \ell(x)$, so $\inf_{y \leq 0} \ell(y) < \inf_y \ell(y)$. Now, consider an instance with two candidates $\{a, b\}$ located at $\mathbf{x}_a = (1, 0)$ and $\mathbf{x}_b = (0, 1)$, and a single voter ranking $a \succ b$ (feasible via the parameter vector $(1, 0)$). It is possible to achieve a loss of $\ell(x)$, e.g., by outputting the parameter vector $(0, x)$. On the other hand, any $\theta$ inducing the ranking $a \succ b$ will be lower bounded by $\ell(0) > \ell(x)$ from above. Hence, $f$ must output $b \succ a$, which is both a PO and PMC violation. □

### 1.3.2 Majority-Based Loss Formulation

Despite the negative results for loss-function-based rules, we may hope for a remedy using slightly different information. Specifically, we consider a similar loss-based function that rather than getting penalized for disagreeing with each voter only gets penalized if it disagrees with a majority of voters. That is, we choose $\theta$ minimizing

$$\mathcal{L}^{maj}(\theta; \pi, \ell) = \sum_{a \neq b \in C} \mathbb{I}[w_{a \succ b}(\pi) > 1/2] \cdot \ell(r_{\theta(b)} - r_{\theta(a)}).$$

Defining a parameter aggregation function based on this loss suffers from the same caveat as before, that in some cases no optimal $\theta$ exists. Nevertheless, we can apply an analogous fix for a ranking variant. We say that a linear rank aggregation rule $f$ *minimizes $\ell$ in the majority formulation* if for all $\sigma = f(\pi)$,

$$\inf_{\theta : \theta \in R^{\sigma}} \mathcal{L}^{maj}(\theta; \pi, \ell) = \inf_{\theta} \mathcal{L}^{maj}(\theta; \pi, \ell).$$

We first show that this does indeed help achieve PMC with essentially all loss functions.

**Theorem 1.3.6.** *Fix a nondecreasing loss function $\ell$ with $\ell(0) > \inf_x \ell(x)$. If a linear rank aggregation rule $f$ minimizes $\ell$ in the majority formulation, then $f$ satisfies PMC.*

*Proof.* Without loss of generality, we may assume that $\inf_x \ell(x) = 0$, as otherwise we could translate $\ell$ without affecting the optimization problem. Fix a profile $\pi$ with feasible PMC ranking $\sigma$, and let $\theta^{PMC}$ be a non-degenerate parameter vector that induces $\sigma$.

First, we show that $\inf_\theta \mathcal{L}^{maj}(\theta; \pi, \ell) = 0$. Indeed, note that $c \cdot \theta^{PMC} \in R^\pi$ for all $c > 0$. Further, note that for any $a, b$ with $w_{a \succ b}(\pi) > 1/2$, $r_{\theta^{PMC}}(b) - r_{\theta^{PMC}}(a) < 0$. Therefore, by making $c$ large, the nonzero terms in $\mathcal{L}^{maj}$ will have an input to $\ell$ negative and becoming arbitrarily large in magnitude. Since $\ell$ is nondecreasing, these approach the infemum of 0.

Next, for any $\sigma' \neq \sigma$, $\inf_\theta \mathcal{L}^{maj}(\theta; \pi, \ell) \geq \ell(0)$. Indeed, there must be some pair of candidates $a, b$ with $a \succ_\sigma b$ and $b \succ_{\sigma'} a$. For any $\theta \in R^{\sigma'}$, $r_\theta(b) \geq r_\theta(a)$, so $\ell(r_\theta(b) - r_\theta(a)) \geq \ell(0)$, and this lower bounds the loss function. □

Note that if the $\ell(0) > \inf_x \ell(x)$ condition is not satisfied, i.e., $\ell(0) = \inf_x \ell(x)$, then *all* linear rank aggregation rules $f$ minimize $\ell$ in the majority formulation, so satisfying this is a vacuous condition. Indeed, the parameter vector $\mathbf{0}$ of all 0s achieves optimal loss of $\ell(0)$ for each pair and is consistent with every ranking $\sigma$. Therefore, the condition $\ell(0) > \inf_x \ell(x)$ is as innocuous as possible to rule out these edge cases.

However, despite this good news for PMC, we show that this does not help in achieving PO. In fact, our negative result extends to every $C1$ linear rank aggregation rule. Note that if $f$ minimizing $\ell$ in the majority formulation breaks ties consistently (i.e., if multiple feasible rankings are optimal, then it consistently chooses the same one), then it is C1. We then have the following result.

**Theorem 1.3.7.** *All $C1$ linear rank aggregation rules fail PO.*

*Proof.* We construct an explicit instance and pairwise majority relationships such that no matter what feasible ranking a rule picks, there is an underlying profile where that output was a PO violation.

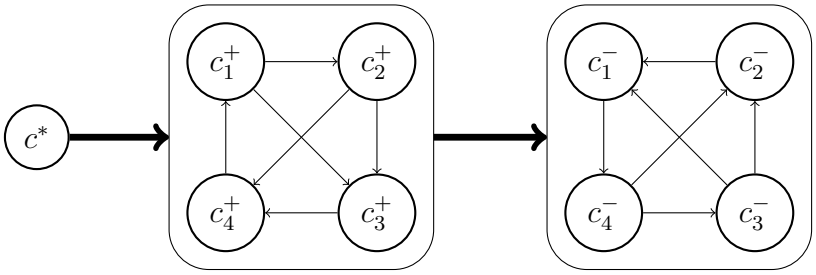



**Figure 1.1:** Graph showing pairwise majority relationship between candidates. Regular edges show relationships among $c_i^+$ candidates and among $c_i^-$ candidates. Thick edges indicate that $c^*$ pairwise beats all candidates, and each $c_i^+$ pairwise beats each $c_j^-$ candidate.

We will have 9 candidates; 8 will be labeled $c_i^+$ and $c_i^-$ for $i = 1, 2, 3, 4$, and one labeled $c^*$. They will have feature vectors in $\mathbb{R}^4$. Each $c_i^\pm$ will be located at $\mathbf{x}_{c_i^\pm} = \pm e_i$ where $e_i$ is the $i$'th standard basis vector, i.e., $c_2^+$ is at $(0, 1, 0, 0)$ and $c_4^-$ is at $(0, 0, 0, -1)$. Finally, $c^*$ will be located at $(1/5, 1/5, 1/5, 1/5)$.

There will be 5 voters. Their pairwise majority graph will be as follows. Candidate $c^*$ will pairwise beat all others. In addition, each $c_i^+$ will pairwise beat each $c_j^-$. Among the $c_i^+$ candidates, there will be a cycle $c_1^+ \succ c_2^+ \succ c_3^+ \succ c_4^+ \succ c_1^+$ and between the remaining two pairs $c_1^+ \succ c_3^+$ and $c_2^+ \succ c_4^+$. The $c_i^-$ candidates will be the exact reverse of this, i.e., a cycle $c_4^- \succ c_3^- \succ c_2^- \succ c_1^- \succ c_4^-$, along with $c_3^- \succ c_1^-$ and $c_4^- \succ c_2^-$. A pictorial representation can be found in Figure 1.1.

A C1 rule must pick a $\theta$ solely based on the pairwise majority graph. We will show that regardless of what $\theta$ it outputs, this will lead to a PO violation.

To that end, the first fact we will show is that no $\theta$ can rank $c^*$ first. Indeed, for any $\theta$, $r_\theta(c^*) = \frac{1}{5}\sum_i \theta_i \leq \frac{1}{5}\sum_i |\theta_i| < \max_i |\theta_i|$. On the other hand, for $i$ maximizing $|\theta_i|$, at least one of $c_i^\pm$ will achieve this reward, strictly larger than $r_\theta(c^*)$. Hence, regardless of the output $\theta$, some candidate must be ranked above $c^*$. We will show that this leads to a PO violation.

To construct profiles consistent with the pairwise majority graph, voters will always have rankings of the following form:

$$c_i^+ \succ c^* \succ c_j^+ \succ c_k^+ \succ c_\ell^+ \succ c_\ell^- \succ c_k^- \succ c_j^- \succ c_i - \tag{1.4}$$

for some $\{i, j, k, \ell\} = \{1, 2, 3, 4\}$. In other words, they will rank a single + candidate above $c^*$ and the rest in some order, followed by all $-$ candidates in the reverse order.

This is always achievable with the voter vector that puts values $1, 3\varepsilon, 2\varepsilon, 1\varepsilon$ in entires $i, j, k$, and $\ell$, respectively, for some small $\varepsilon > 0$.

Fix an output $\theta$ with induced ranking $\sigma$. There must be at least one candidate $c \neq c^*$ ranked above $c^*$. We now split into cases depending on which candidate this is. For each choice, we will construct a profile consistent with the pairwise majority graph where the candidate above $c^*$ is Pareto dominated by $c^*$.

The profiles with five voters each are presented in the table below. Each voter's ranking only be specified by an ordering over the + candidates, assuming they otherwise take the form described in (1.4). For example, the notation $1 : (2, 1, 3, 4)$ implies one voter has the ranking in the form of (1.4) with $(i, j, k, \ell) = (2, 1, 3, 4)$. Note that $c^*$ is always ranked second, so if a candidate is never ranked first, they are Pareto dominated by $c^*$.

| $c_1^+$ | $c_2^+$ | $c_3^+$ | $c_4^+$ |
|---|---|---|---|
| 1: $(2, 1, 3, 4)$ | 2: $(1, 2, 3, 4)$ | 2: $(1, 2, 3, 4)$ | 2: $(1, 2, 3, 4)$ |
| 1: $(3, 1, 2, 4)$ | 2: $(3, 2, 4, 1)$ | 2: $(2, 3, 4, 1)$ | 2: $(2, 4, 1, 3)$ |
| 1: $(3, 2, 4, 1)$ | 1: $(4, 1, 2, 3)$ | 1: $(4, 1, 2, 3)$ | 1: $(3, 4, 1, 2)$ |
| 2: $(4, 1, 2, 3)$ | | | |

Finally, if a $-$ candidate is ranked above $c^*$, then any of the following profiles work, as all $-$ candidates are Pareto dominated by $c^*$ with rankings shown in (1.4). □

This result is quite unfortunate, because if there were a rule that is both $C1$ and PO, we would automatically achieve PMC: Whenever there is a feasible PMC ranking, a $C1$ rule cannot distinguish between this profile and a profile where all voters submit this ranking, hence, under the PO criterion, it must output it. Furthermore, whenever there is a PMC ranking, outputting it is necessarily PO, as for every pair, a majority of voters agree with the PMC ranking. Interestingly, in the proof, we construct a profile which has a PMC ranking, yet it is not feasible, and no matter how a $C1$ linear rank aggregation rule breaks ties, there is an underlying profile in which this output violates PO.

## 1.4 Social Choice-Based Rule

In light of the above negative results, in this section, we ask whether there are linear rank aggregation rules that concurrently satisfy our two core axioms, PO and PMC. We answer this question affirmatively by presenting a new method based on a prominent rule from voting theory.

The *Copeland rule* assigns a *Copeland score* to each alternative equal to the number of other alternatives it beats in a pairwise competition, i.e., the score for $a$ is $|\{b \mid w_{a \succ b} > 1/2\}|$. It then ranks the candidates in descending order according to their Copeland scores (breaking ties arbitrarily). It is known that Copeland satisfies PO, PMC, and additional axiomatic properties. However, in linear social choice, since not every ranking is feasible, we cannot always output the Copeland ranking.

We, therefore, define a new linear rank aggregation rule, which we call *leximax Copeland*. This rule chooses a feasible ranking as follows. It ranks first the candidate with the highest Copeland score that can be feasibly ranked first under some parameter vector $\theta$. Subject to this first position, it ranks second the candidate with the highest Copeland score which can be feasibly ranked second, and continues this process for subsequent positions.

Copeland's rule is a $C1$ rule because it only requires the majority relationships between the candidates. Analogously, leximax Copeland is also a $C1$ linear rank aggregation rule. Therefore, by Theorem 1.3.7, it does not satisfy the PO criterion. To address this issue, we define a variant called *leximax Copeland subject to PO (LCPO)*, which incorporates the PO criterion. Under LCPO, for every pair of alternatives where one dominates the other, the rule restricts rankings to place the dominating alternative above the dominated one. The rule remains well-defined since the set of feasible rankings when enforcing the PO criterion is non-empty, as whenever $a$ dominates $b$, all the rankings in the input profile rank $a$ above $b$. Note that if the Copeland ranking is feasible, then this rule outputs that ranking, since unrestricted Copeland satisfies PO.

In addition to PO and PMC, we wish to show that LCPO satisfies two additional properties, which we define presently.

**Definition 1.4.1** (majority consistency)**.** *A linear rank aggregation rule satisfies* majority consistency *if when a candidate $a$ is ranked first by a majority of voters in the input profile, $a$ is ranked first in the output ranking.*

Majority consistency ensures that the collective decision reflects the preference of the majority when there is a clear favorite. This principle aligns with PMC, but specifically focuses on the majority's favorite alternative. However, as we discussed above, a PMC ranking does not necessarily exist, and even when it exists, it is not necessarily feasible. By contrast, when a majority winner exists, this candidate is necessarily ranked first by a majority of voters in the input profile, who themselves (by assumption) submit feasible rankings. Therefore, we need not handle the case where it is impossible to rank the majority winner first.

**Definition 1.4.2** (winner monotonicity)**.** *A linear rank aggregation rule satisfies* winner monotonicity *if, when a candidate $a$ is ranked first in the output ranking, elevating $a$ in any voter's preference does not cause $a$ to lose their top position in the updated aggregate ranking.*

Winner monotonicity ensures that improving a leading candidate's position among individual voters will not result in that candidate's demotion.

We now state and prove the main result of this section.

**Theorem 1.4.3.** *LCPO satisfies PO, PMC, majority consistency and winner monotonicity.*

*Proof.* LCPO trivially satisfies PO since it always outputs a ranking that respects the PO criterion. Moreover, since Copeland satisfies PMC, and whenever Copeland's ranking is in the domain, leximax Copeland subject to PO returns this ranking, it clearly satisfies PMC.

Note that if an alternative $a$ is ranked first by at least half of the voters, then $a$ has the highest Copeland score, meaning that leximax Copeland subject to PO will rank this candidate first if this is possible. We see that this is indeed possible, by noticing that there is at least one feasible ranking in the input profile where $a$ is ranked first, and any such input ranking is feasible (by assumption) and satisfies the PO requirement. Therefore, majority consistency is satisfied.

It remains to show that LCPO satisfies winner monotonicity. Suppose that on input profile $\pi$, the rule outputs a ranking $\sigma$ where candidate $a$ is ranked first. Now, consider a profile $\pi'$ which is similar to $\pi$ with the only exception being a ranking in which $a$ is placed in a higher position. Let $S$ be the set of agents that ranked above $a$ in Copeland's ranking under $\pi$ and let $S'$ be the set of agents that ranked above $a$ in

Copeland's ranking under $\pi'$. Note that $S' \subseteq S$, since when moving from $\pi$ to $\pi'$, only the Copeland score of $a$ can increase, and therefore it is not possible for a candidate $b$ to beat $a$ under $\pi'$ but not under $\pi$.

Now, suppose that $R$ and $R'$ are the set of rankings that satisfy the PO criterion with respect to $\pi$ and $\pi'$, respectively. We show that $R' \subseteq R$. First, note that since $a$ is ranked first under $\pi$, no alternative dominates $a$ in $\pi$, as otherwise the PO criterion would be violated. Therefore, we get that no other alternative dominates $a$ in $\pi'$ as well. Moreover, note that if $b$ dominates $c$ in $\pi$, then this remains true in $\pi'$ as well. On the other hand, it is possible that $a$ dominates an alternative $b$ in $\pi'$ but not in $\pi$. From all of the above, we conclude that $R' \subseteq R$.

Since $a$ is ranked first in $\sigma$, we get that for every candidate $b \in S$, there is no ranking in $R$ in which $b$ is ranked first, since otherwise, LCPO would output such a ranking. This also means that for every candidate $b$ in $S'$, there is no ranking in $R'$ in which $b$ is ranked first, since $R' \subseteq R$ and $S' \subseteq S$. Moreover, note that every ranking in $R$ in which $a$ is ranked first is also in $R'$ since it satisfies all the PO restrictions of $\pi'$. Therefore, under $\pi'$, LCPO outputs a ranking in which $a$ is ranked first. □

Leximax Copeland subject to PO can be implemented in polynomial time by solving $O(|C|^2)$ relatively small linear programs. Specifically, given an input profile, we sequentially choose the candidate that is ranked in position $r+1$ as follows. We denote by $\sigma_r$ the partial ranking, where the first $r$ positions have been fixed. For each candidate $c$ that has not been ranked yet, we want to check if there is a parameter vector that adheres to the partial ranking $\sigma_r$, respects the Pareto optimality criterion and ranks $c$ at position $r+1$. Since all these constraints can be expressed as pairwise comparisons, we can use a linear program such as the one described in Footnote 4 to check if such a feasible ranking exists. Among the candidates meeting this criterion, we select the one with the highest Copeland score for position $r$.

## 1.5 Discussion

We conclude with a discussion of several extensions and limitations of our approach and results.

First of all, we wish to emphasize that our results are theoretical. While they highlight some shortcomings of the current practice of RLHF, our goal was not to "outperform" existing RLHF methods. Rather, we see our model as giving a framework

for understanding and comparing rules and methods — it is a (useful, we believe) lens through which researchers and engineers can examine their AI alignment methods.

Second, as written, our model has voters give their complete rankings, while in practice, this would be infeasible. In the real world, we are likely to elicit only relatively few pairwise comparisons per person. For our negative results, this assumption only makes them stronger: the BTL model fails PO and PMC *even* when it has access to complete voter rankings. By contrast, for the positive results, specifically implementing leximax Copeland subject to PO, this ostensibly seems like a serious limitation. However, the complete rankings are not necessary for computing this rule, rather, all we need to know are PO dominance relationships and majority directions. We can therefore apply the rule whenever we can determine this information at least approximately through, e.g., sampling. An alternative approach is to infer a complete ranking of each voter by fitting a parameter vector based on their pairwise responses; this process of learning a complete ranking and then running voting rules has been used before in a variety of settings [136, 154]

Third, our work initiates the study of the axiomatic method in our linear social choice model. However, we leave open many questions about which axioms are compatible and finding rules that achieve them. It should be clear by now that the primary challenge in linear social choice is that not every ranking over the candidates can be output. This means that essentially all known aggregation rules cannot be used directly without at least some modification. A natural direction to tackle is to try to find methods of converting known voting rules into linear aggregation ones while maintaining some of their axiomatic properties. To this end, we conclude with some preliminary results, and somewhat surprising findings within this space.

Some rules which optimize over rankings can be naturally transformed. For example, consider the Kemeny rule, which returns the ranking with the smallest pairwise disagreement over all votes. This can easily be transformed to the linear setting by simply outputting the optimal feasible ranking. In fact, in Appendix C.2 of the full version, we show that this rule carries over the property of *separability*,[9] a social choice axiom that is violated by Copeland (in the classical setting) and leximax Copeland subject to PO (in our setting). We show this in Appendix C.1. However, quite strikingly, although separability remains, this transformation makes Kemeny no longer PO (Appendix C.2).

[9]Formally, *ranking separability* to distinguish it from the single-winner version.

Finally, note that the "leximax" portion of leximax Copeland can be seen as a general purpose tool for mapping traditional rules to a linear aggregation ones. In Appendix C.3, we explore leximax plurality (run leximax on the ranking of candidates by plurality scores), and show that it satisfies majority consistency, winner monotonicity, and separability. Additionally, the "subject to PO" can be seen as another "tool" for enforcing the Pareto optimality criterion when a rule does not independently satisfy it. However, enforcing PO can again cause somewhat surprising results. For example, in Appendix C.2, we show that linear Kemeny subject to PO, while now trivially satisfying PO, again violates separability. These observations indicate the challenges inherent in linear social choice, and we hope the many open questions will lead to fruitful followup research.

# 2

# Pairwise Calibrated Rewards for Pluralistic Alignment

## 2.1 Introduction

Recall the standard high-level implementation of RLHF for AI alignment, as outlined in Chapter 1: first, train a reward model — typically using the Bradley-Terry (BT) framework [35] — on pairwise preference data [18, 204]; then, fine-tune a pre-trained model to align with the resulting reward signal.

Implicit in current RLHF implementations is the assumption of a shared human intuition — a common ground among evaluators about what constitutes desirable behavior. While this assumption may hold for alignment objectives such as ensuring model safety, it generally does not apply to tasks where interpretations of "right" behavior inherently diverge across backgrounds, cultures, and beliefs [9, 125, 166, 211]. BT-based reward models compress these diverse inputs into one scalar, blurring conflicting viewpoints into a one-size-fits-all model. This aggregation leads to a misspecified objective [41, 43] that struggles when faced with plural or contradictory feedback, marginalizing minority perspectives and failing to capture the full spectrum of human values [124, 161, 176, 192]. Once such a reward is learned, algorithms like PPO [180] then push the policy to maximize this signal, triggering *preference col-*

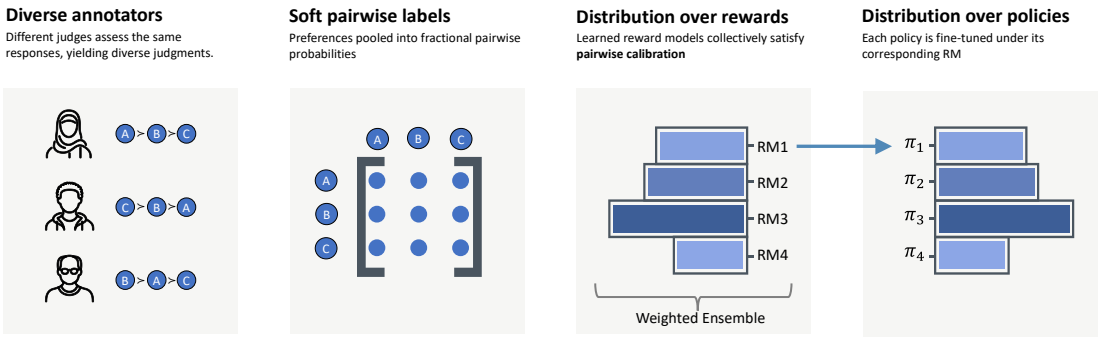



**Figure 2.1:** We explicitly model human preferences as a distribution over reward functions without relying on predefined annotator groups or identities. By enforcing pairwise calibration, this distribution faithfully captures the inherent diversity in aggregate human judgments. Each calibrated reward function then induces a distinct policy.

*lapse* [207], where majority views are further amplified, and response diversity is reduced [120, 126, 132, 160, 185].

To address these shortcomings, we replace the single-reward assumption with a *distribution* over reward functions; each independently assigns desirability scores to responses and they collectively span plausible interpretations of human judgment. By fine-tuning a distinct LLM for each reward in the support, the approach yields a distribution over policies. At inference time, these policies can be used in several ways [71, 194]: present a range of viewpoints or fold them into one inclusive answer, let users pick the policy that matches their taste, or sample responses directly from the distribution.

A straightforward approach to derive a distribution over reward functions is to partition annotators into predefined groups — such as demographic segments — or to cluster them by the similarity of their recorded preferences [43, 45, 158, 195]. Such methods implicitly assume that each group holds a stable, coherent preference vector that carries over from one context to another. Human preferences, however, are fluid and highly context-dependent [107]; demographic or ideological attributes explain only part of the observed variation, and annotators who match on all recorded traits can still disagree sharply [42]. Moreover, overly fine-grained groups may lack enough data for reliable training.

Without explicit annotator identification or predefined groups, learning a distri-

bution over rewards requires disentangling the hidden preference dimensions inside aggregated feedback. Concretely, rather than associating preferences with stable annotator identities, we look for a *small* set of internally consistent reward functions — an *ensemble* — that together explain conflicting human judgments across contexts. Our aim is to achieve a novel property we call *pairwise calibration*: for every pair of candidate responses, the share of reward functions in the support of the ensemble favoring one response matches the fraction of annotators who express that preference. A pairwise-calibrated ensemble gives each reward function a coherent viewpoint and its own policy. Taken together, these policies mirror a broader range of human preferences and avoid preference collapse (see Figure 2.1). Thus, pairwise calibration provides a principled mechanism to preserve pluralism, without imposing rigid clusters or relying on annotator tags.

Our results. In Section 2.2 we present our model and formally define *pairwise calibration*. Section 2.3 shows that while ensembles can trivially satisfy pairwise calibration when they are sufficiently large (e.g., at the scale of the number of annotators), constructing such ensembles is NP-hard (Theorem 2.3.1). More importantly, our main goal is to find practical ensembles with small support. In light of these observations, we show in Theorem 2.3.3 that an ensemble of size $O(1/\varepsilon)$ is sufficient to achieve an average calibration error of $\varepsilon$, and Theorem 2.3.4 further shows that including extreme outlier preferences is unnecessary, as an outlier-free ensemble remains nearly pairwise-calibrated. Moreover, Theorem 2.3.6 shows that achieving pairwise calibration can be learned with a limited number of pairwise comparisons, providing a generalization guarantee. Section 2.4 asks whether such ensembles can be efficiently constructed in practice and answers with a forward stagewise additive modeling (FSAM) procedure: each iteration trains a reward model on the current residual calibration error and re-weights it into the mixture, giving a tractable method for constructing a compact ensemble. Finally, Section 2.5 demonstrates that these ensembles attain lower calibration mean squared error (MSE) than the theoretically optimal single deterministic reward and that the reward models within each ensemble are diverse.

Related work. Our work contributes to the growing field of *pluralistic alignment*, training AI systems to accommodate the natural diversity of human values and perspectives [55, 82, 148, 170, 194]. We briefly cover the broad approaches taken in this

emerging area.

*Scalar Reward Aggregation Approaches.* To mitigate the majority-bias of a single reward, recent methods introduce *additional* loss terms or algorithmic modifications that slow or penalize the policy from over-optimizing for the majority. For instance, Xiao et al. [207] add a custom regularizer to reduce preference collapse in PPO updates, while Chen et al. [46] temper the update magnitude whenever annotator feedback shows high dispersion. Beyond such direct regularization, some works increase the reward model's capacity to represent preference diversity — for example, by weighting pairwise data based on annotator or context uncertainty [105] or learning a Bayesian distribution over reward function parameters [202]. Despite these refinements, methods reliant on a single scalar reward still risk oversimplifying genuine diversity in user opinions, motivating richer multi-reward solutions.

*Clustering and Group-Based Reward Modeling.* A more direct way to represent divergent preferences is to train multiple reward models, each matched to a distinct segment of the human population [43, 45, 158, 195]. This can be accomplished by clustering annotators based on demographics or inferred preference patterns and then fitting a group-specific reward function. For example, Chakraborty et al. [43] show that optimizing a mixture of cluster-specific rewards via a max-min objective can protect minority viewpoints from being overshadowed by the majority. Although powerful, these methods typically rely on stable, coherent groups, which may not reflect real-world fluidity in user preferences [107], and group membership can sometimes be ambiguous or proprietary. Additionally, overly granular clusters risk sparse data, whereas coarse groupings might obscure meaningful sub-group heterogeneity [42]. By contrast, we avoid the need for explicit clusters or demographic tags. Our approach learns a small ensemble of reward functions — each representing an internally consistent viewpoint — and ensures pairwise calibration so that overall, the ensemble reflects the full distribution of observed human judgments.

*Personalized and Identity-Conditional Alignment.* A further alternative is to directly personalize the alignment objective at the level of individual annotators, assuming each user has a unique reward function [163]. While highly expressive, such methods become cumbersome at scale, and excessive personalization can reinforce biases or amplify social divides [124]. Consequently, many personalization approaches use latent embeddings or partial fine-tuning, striking a balance between capturing user-specific signals and avoiding an explosion of fully individualized models. In prac-

tice, personalization-based methods still risk "echo chambers," prompting calls for a more bounded approach to preference diversity [124]. We share this concern, and thus focus on small ensembles as a middle path: each ensemble component aligns with a coherent subset of preferences, preserving pluralism while limiting the societal risks of extreme personalization.

## 2.2 Our Approach

Notation. Let $\mathcal{X}$ be the context space and let $\mathcal{Y}$ be the response space. A reward function $r_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ assigns a scalar value indicating the desirability of a response given a context, where $\theta \in \Theta$ is a set of parameters. A *$k$-ensemble* (or simply an ensemble) is a tuple $(r_{\theta_1}, \ldots, r_{\theta_k}; \alpha_1, \ldots, \alpha_k)$, where $\alpha_1, \ldots, \alpha_k$ are mixture weights forming a probability distribution over the reward functions (each $\alpha_j \geq 0$ and $\sum_{j=1}^{k} \alpha_j = 1$).

We denote by $\mathcal{N}$ the population of annotators, and we write $i \sim \mathcal{N}$ to mean drawing an annotator uniformly from the population. We assume each annotator $i \in \mathcal{N}$ has a preference relation $\succ_i$ such that for any context $x \in \mathcal{X}$ and distinct responses $y_1, y_2 \in \mathcal{Y}$, $y_1 \succ_i y_2 \mid x$ indicates that annotator $i$ prefers $y_1$ to $y_2$ given context $x$. We call a tuple $(x, y_1, y_2)$ a *response triple* or simply a *triple*. We assume the reward space is rich enough such that the annotators are *reward-inducible*: for each $i \in \mathcal{N}$, there is a reward $r_\theta$ such that $y_1 \succ_i y_2 \mid x$ exactly when $r_\theta(x, y_1) > r_\theta(x, y_2)$.

A policy $\pi(y \mid x)$ is a probability distribution over outputs $y$ conditioned on a context $x$. We assume there is an underlying distribution over contexts $D_{\mathcal{X}}$ and base policy $\pi_0$. Fine-tuned policies are typically learned by optimizing responses to maximize expected rewards according to the corresponding reward function $r_\theta$.

Pairwise calibration. Our key conceptual innovation is the requirement that the ensemble faithfully reflects human preference frequencies. Intuitively, an ensemble is *pairwise calibrated* if the pairwise comparisons it induces align with the proportion of human annotators who prefer one response over another. Formally, consider a context $x \in \mathcal{X}$ and two distinct candidate responses $y_1, y_2 \in \mathcal{Y}$. We define $p^\star(x, y_1, y_2)$ to be the true fraction of annotators that prefer $y_1$ to $y_2$ given context $x$:

$$p^\star(x, y_1, y_2) := \Pr_{i \sim \mathcal{N}}[y_1 \succ_i y_2 \mid x] \in [0, 1]. \tag{2.1}$$

Given an ensemble $\mathbf{r} = (r_{\theta_j}; \alpha_j)_{j \in [k]}$, denote the induced preference probability by

$$\hat{p}^{\mathbf{r}}(x, y_1, y_2) := \sum_{j=1}^{k} \alpha_j \cdot \mathbf{1}[r_{\theta_j}(x, y_1) > r_{\theta_j}(x, y_2)].^{1}$$

If $\mathbf{r}$ is clear from context, we may drop it from the $\hat{p}$ notation.

**Definition 2.2.1** (Pairwise calibration)**.** *We say an ensemble* $\mathbf{r}$ *is* $\varepsilon$*-pairwise calibrated if:*

$$\mathbb{E}_{x \sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot|x)} \left[ (\hat{p}^{\mathbf{r}}(x, y_1, y_2) - p^{\star}(x, y_1, y_2))^2 \right] \leq \varepsilon.$$

If an ensemble is 0-pairwise calibrated, we call it *perfectly* pairwise calibrated. We use squared error for analytical convenience, though other divergences that penalize discrepancies between $\hat{p}$ and $p^{\star}$ (e.g., absolute error or cross-entropy) could also serve this purpose. Calibration is defined here at the reward level based only on how they order the responses. This is not the only natural definition: pairwise calibration could instead depend directly on the probabilities induced by the learned policies. Our choice is primarily motivated by tractability and the observation that downstream policies are largely influenced by the ordering of rewards. See Appendix A of the full version for further discussion and justification of this design choice.

Non-identifiability. Once annotator identities are dropped, the only observable data are the pairwise preference probabilities $p^{\star}$ in Equation (2.1). Prior work [169, 184, 188] shows that the same $p^{\star}$ can be generated by infinitely many distinct mixtures of annotator-specific preferences. Because of this, recovering the "true" annotator preference distribution is information-theoretically impossible; our goal is therefore to learn *any* ensemble that is pairwise calibrated to $p^{\star}$, rather than to reconstruct hidden annotator groups.

Policy ensemble inference. A pairwise-calibrated ensemble induces a mixture of policies, $(\pi_1, \ldots, \pi_k)$, where $\pi_j$ is aligned to $r_{\theta_j}$; yet this alone does not tell us *how* to deploy that distribution at inference time. Below we outline three complementary

[1] We assume that chosen reward functions never produce ties. This is a minor assumption, as in continuous parameter spaces, the parameters that induce ties $r_\theta(x, y_1) = r_\theta(x, y_2)$ with $y_1 \neq y_2$ typically form a measure-zero set, so any perturbation almost surely removes them.

approaches — inspired by Sorensen et al. [194] — that can be selectively employed to serve different pluralism purposes depending on the context. Together they offer a practical toolkit for pluralistic alignment.

In *balanced pluralism* mode, the system queries every LLM in the support. After sampling the resulting set of candidate outputs, it can either present them in an Overton-style slate, useful for open-ended advice, policy deliberation, and brainstorming; or distill them into a single consensus statement, aligning with prior work on consensus building [19, 74].

In *steerable pluralism* mode, the system, on a per-prompt basis, selects one policy $\pi_j$ whose persona or metadata best matches a declared user preference and produces that policy's output. The result is a single voice that fits personal assistants, brand-specific copy, or group-specific safety policies, without the overhead of user-specific fine-tuning or elaborate prompt engineering.

In *distributional pluralism* mode, the system samples a policy from the mixture in proportion to its weight and returns that policy's output. Applied over many requests, this procedure maintains population-level diversity and counteracts the tendency of aligned models to exhibit low output diversity by recycling a small set of high-reward responses [120, 126, 160]. This mode is especially valuable in creative workflows — text-to-image generation, story creation, and recommendation engines — where safety constraints must coexist with a rich variety of outputs.

Small support. We deliberately restrict the ensemble to a small support. A compact ensemble is easier to train, tune, and deploy, and it generalizes better (see Theorem 2.3.6). Furthermore, the inference methods discussed require storing and efficiently querying the full ensemble, constraining $k$ to stay relatively small. Finally, Section 2.3.1 shows that only $O(1/\varepsilon)$ reward functions suffice for $\varepsilon$-pairwise calibration, so increasing $k$ yields diminishing gains.

Beyond these practical reasons, we intentionally target a small support for normative considerations. Limiting the ensemble size mitigates societal risks inherent in extreme personalization — such as reinforcing existing biases, promoting echo chambers, and exacerbating polarization (see *personalization within bounds* [124]). Taken together, Theorems 2.3.3 and 2.3.4 show that a limited-support distribution over rewards preserves meaningful diversity without over-tailored personalization.

Outlier reward functions. While pairwise calibration ensures the ensemble *as a whole* faithfully reflects aggregate human preference frequencies, we are also interested in the behavior of the individual reward functions $r_{\theta_j}$ that constitute the ensemble. A well-calibrated ensemble could, in principle, still contain individual reward functions that represent extreme or undesirable viewpoints.

To quantify how much a single reward function $r_\theta$ deviates from the overall population preferences, we define its *disagreement score* as:

$$\Phi(\theta) = \Pr_{\substack{x\sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot|x), \\ i\sim\mathcal{N}}} \left[\mathbf{1}[r_\theta(x, y_1) > r_\theta(x, y_2)] \neq \mathbf{1}[y_1 \succ_i y_2 \mid x]\right].$$

The disagreement score measures how frequently the reward function disagrees with the population. More specifically, it is the probability that for a randomly selected annotator $i \sim \mathcal{N}$ and context response triple $(x, y_1, y_2)$, $r_\theta$ prefers one response while $i$ the other.

Of course, on populations with high amounts of diversity, it may be impossible for any reward to have particularly low disagreement score. Given this, we say a reward function $r_\theta$ is a $(\beta, \gamma)$-*outlier* if $\Phi(\theta) > \beta \cdot \inf_{\theta' \in \Theta} \Phi(\theta') + \gamma$, which normalizes disagreement relative to the best-achievable score.

## 2.3 Theoretical Guarantees

Having defined pairwise calibration, we now address three fundamental questions: (i) *Can approximately pairwise-calibrated ensembles be supported on a relatively small set of reward functions?* (ii) *Can such calibration be achieved without using extreme outlier rewards?* (iii) *Do ensembles calibrated on a finite dataset generalize to unseen populations?* We answer all three in the affirmative.

### 2.3.1 Pairwise-Calibrated Ensembles with Small Support

A uniform mixture over the annotators' true reward functions would, in principle, achieve perfect pairwise calibration. However, this approach is impractical for two reasons. First, the true reward functions are never observed — only partial pairwise information provided. Second, this ensemble's support size would need to scale with the number of annotators, making the approach computationally infeasible.

For now, we set aside the second issue and focus on the first. We show that *finding* a perfectly calibrated ensemble is computationally hard even in an extremely simplified setting: there is only a single context, (i.e., $|\mathcal{X}| = 1$), $m$ possible responses (i.e., $|\mathcal{Y}| = m$), and we have access to the true fractions of annotators who prefer $y_1$ to $y_2$ for all pairs, (i.e., all of the values $p^\star(x, y_1, y_2)$).

Under these conditions, the set of pairwise comparisons we wish to calibrate to is finite. A straightforward application of Carathéodory's theorem guarantees that there exists a perfectly calibrated ensemble of size only $\Theta(m^2)$ (see Appendix C.1 off the full version). Yet, even under these strong simplifying assumptions — and a guarantee that only a relatively small ensemble is required — it is computationally hard to find one.

**Theorem 2.3.1.** *If $P \neq NP$, then there does not exist a polynomial-time algorithm for finding a perfectly pairwise calibrated ensemble when $|\mathcal{X}| = 1$, $|\mathcal{Y}| = m$ for some finite $m$ and given access to $p^\star(x, y_1, y_2)$ for all $(x, y_1, y_2)$.*[2]

This setup is analogous to one in *social choice theory* [36]; see Appendix C.2 of the full version for details. At a high level, the proof shows that if we could efficiently find perfectly calibrated ensembles, we would be able to determine whether certain values of $p^\star$ are realizable by *any* underlying distribution over reward functions. This provides a *membership oracle* to the set of realizable $p^\star$, which turn out to form a useful convex set known as the *linear ordering polytope*. Membership oracles allow us to optimize linear functions over the polytope, thereby enabling us, if we are careful with various approximations, to solve classic NP-hard problems such as *Minimum Feedback Arc Set*.

*Proof of Theorem 2.3.1.* For convenience, we adopt standard terminology of *social choice theory* [36] in this proof. Specifically, we call elements of $\mathcal{Y} = \{y_1, \ldots, y_m\}$ *candidates*. Let $\mathrm{LO}(\mathcal{Y})$ denote the set of linear orders (or rankings) over $\mathcal{Y}$. The preferences of the voters (i.e., annotators) are represented by a distribution over rankings. A distribution assigns a probability $p_\sigma \geq 0$ to each $\sigma \in \mathrm{LO}(\mathcal{Y})$ such that $\sum_{\sigma \in \mathrm{LO}(\mathcal{Y})} p_\sigma = 1$.

[2]Formally, the algorithm takes as input the rational values $p^\star(x, y_1, y_2)$, encoded as pairs of binary integers, and must run in time polynomial in the size of this instance.

Each ranking $\sigma$ can be represented by its *incidence vector* $\mathbf{x}^\sigma \in \{0,1\}^{\binom{m}{2}}$. For each pair $\{y_i, y_j\}$ with $i < j$,

$$x_{ij}^\sigma = \begin{cases} 1 & \text{if } y_i \text{ is ranked before } y_j \text{ in } \sigma, \\ 0 & \text{if } y_j \text{ is ranked before } y_i \text{ in } \sigma. \end{cases}$$

A *(weighted) tournament graph* corresponding to a distribution over rankings is the weighted average of these incidence vectors $\mathbf{t} = \sum_{\sigma \in \mathrm{LO}(\mathcal{Y})} p_\sigma \mathbf{x}^\sigma$. Each coordinate $t_{ij}$ represents the fraction of voters that rank $y_i$ before $y_j$. The set of all realizable tournament graphs is the convex hull of the incidence vectors:

$$\mathrm{TG} = \mathrm{conv}\left\{\mathbf{x}^\sigma \mid \sigma \text{ is a linear ordering}\right\}.^{3}$$

We show that deciding if a given vector $\mathbf{y} \in \mathbb{R}^{\binom{m}{2}}$ is a valid tournament graph (i.e., $\mathbf{y} \in \mathrm{TG}$) is NP-hard. This implies that the potentially harder problem of finding a distribution over rankings $p_\sigma$ consistent with a given $\mathbf{t} \in TG$ is also NP-hard.

To establish this, we reduce from the NP-hard problem of *Minimum Feedback Arc Set (MFAS)* [116]. Given a directed graph $G = (V, E)$ on $m$ vertices (which we identify with $\mathcal{Y}$) and an integer $k$, MFAS asks if there is an ordering of the vertices $\sigma$ such that at most $k$ edges in $E$ are *feedback arcs* (i.e., an edge $(y, y')$ such that $y'$ appears before $y$ in $\sigma$). The number of feedback arcs for an ordering $\sigma$ is given by the linear function

$$N_{\mathrm{FAS}}(x^\sigma) = \sum_{i<j \ (y_j, y_i) \in E} x_{ij}^\sigma + \sum_{i<j \ (y_i, y_j) \in E} (1 - x_{ij}^\sigma).$$

Furthermore, since $N_{\mathrm{FAS}}$ is a linear function, its minimum on the convex hull is achieved at a vertex. Hence, $\min_{\sigma \in \mathrm{LO}(\mathcal{Y})} N_{\mathrm{FAS}}(\mathbf{x}^\sigma) = \min_{\mathbf{x} \in TG} N_{\mathrm{FAS}}(\mathbf{x})$.

Suppose we have a polynomial-time algorithm (an oracle) that decides whether a given (rational) vector $\mathbf{y}$ is in TG (in particular, this algorithm should run in time polynomial in the binary encoding of $\mathbf{y}$). We call this the *Tournament-Oracle Problem*. We show this allows for a polynomial-time algorithm for MFAS using the following theorem:

**Theorem 2.3.2** (Grötschel et al. [89], Corollary 4.3.12 (rephrased))**.** *Let $K \subset \mathbb{R}^d$ be a convex set, such that there exists a point $\mathbf{x}_0 \in \mathbb{R}^d$, and numbers $0 < r < R$ such*

[3] This is often called the *linear ordering polytope*.

*that $B(\mathbf{x}_0, r) \subseteq K \subseteq B(\mathbf{x}_0, R)$, where $B(\mathbf{x}_0, r)$ is the ball of radius $r$ centered at $\mathbf{x}_0$. Suppose we have a membership oracle $\mathcal{O}$ for $K$. Then, for every $\mathbf{c} \in \mathbb{R}^d$, there exists an algorithm that outputs a point $\mathbf{z} \in \mathbb{R}^d$ such that*

$$\mathbf{c}^\top \mathbf{z} \leq \min_{\mathbf{x} \in K} \mathbf{c}^\top \mathbf{x} + \varepsilon\big(\max_{\mathbf{x} \in K} \mathbf{c}^\top \mathbf{x} - \min_{\mathbf{x} \in K} \mathbf{c}^\top \mathbf{x}\big).$$

*The algorithm runs in time polynomial in $\log(1/\varepsilon)$, $\log(R/r)$, the bit-size of $\mathbf{c}$, and $d$. The membership oracle is queried with rational points of polynomially bounded encoding sizes.*

We apply Theorem 2.3.2 with $K = \mathrm{TG}$, hence $d = \binom{m}{2} = O(m^2)$. Let $\mathbf{x}_0 \in \mathbb{R}^d$ be the point with all coordinates $1/2$. Since $\mathbf{x}_0$ can be induced by the average of all $m!$ possible rankings $\mathbf{x}^\sigma$, it is in TG. The squared $L_2$ distance from $\mathbf{x}_0$ to any vertex $\mathbf{x}^\sigma \in \{0,1\}^d$ is

$$\|\mathbf{x}^\sigma - \mathbf{x}_0\|_2^2 = \sum_{k=1}^{d} \left(\mathbf{x}_k^\sigma - \frac{1}{2}\right)^2 = \frac{d}{4},$$

implying that $R \leq \sqrt{d/2} = O(m)$. McGarvey's theorem [144] shows that there exist tournaments with 0 or 1 in a single coordinate and $1/2$ everywhere else. Convex combinations of McGarvey's tournament graphs and $\mathbf{x}_0$ can generate any tournament such that every coordinate is within $1/\binom{m}{2}$ of $1/2$. Hence, $r = 1/\sqrt{\binom{m}{2}} \geq 1/m$ is sufficient. Thus, $\log(R/r) = O(\log m)$.

For MFAS, we want to minimize $N_{\mathrm{FAS}}$. The objective function's linear part is defined by $\mathbf{c}_{\mathrm{FAS}}$, whose components are in $\{-1, 0, 1\}$, and therefore $O(d) = O(m^2)$ bit complexity. The minimum value of $N_{\mathrm{FAS}}$, as it must occur at a vertex in $\{0,1\}^d$, is an integer. Furthermore, $0 \leq N_{\mathrm{FAS}}(\mathbf{x}) \leq \binom{m}{2}$ for all $\mathbf{x} \in \mathrm{TG}$.

Applying the theorem with $\varepsilon < 1/(2 \cdot \binom{m}{2})$ and rounding the solution gives a polynomial time algorithm for finding the optimal value of MFAS, which, assuming $P \neq NP$, is a contradiction. □

Next, we address the second consideration. We show that small-support ensembles can indeed achieve (approximate) pairwise calibration.

**Theorem 2.3.3.** *For any $\varepsilon > 0$, there exists a $O(\varepsilon^{-1})$-ensemble that is $\varepsilon$-pairwise-calibrated.*

This is shown using the probabilistic method: by sampling $O(1/\varepsilon)$ reward functions according to a particular strategy, the expected pairwise calibration error is at most $\varepsilon$, guaranteeing the existence of at least one ensemble meeting the criterion.

*Proof of Theorem 2.3.3.* Fix $\varepsilon > 0$. For each $i \in \mathcal{N}$, let $r_{\theta_i}$ be the reward function that induces $i$'s preferences (i.e., $y_1 \succ y_2 \mid x \iff r_{\theta_i}(x, y_1) > r_{\theta_i}(x, y_2)$). Let $k = \lceil \frac{1}{4\varepsilon} \rceil$. This choice implies $k \geq \frac{1}{4\varepsilon}$, so $\varepsilon \leq \frac{1}{4k}$.

Consider randomly constructing a $k$-ensemble $\mathbf{r}$ as follows: Randomly select $k$ annotators $i_1, \ldots, i_k \sim \mathcal{N}$ uniformly with replacement. The ensemble is then:

$$\mathbf{r} = (r_{\theta_{i_1}}, \ldots, r_{\theta_{i_k}}; 1/k, \ldots, 1/k).$$

Let $\mathcal{R}$ denote this distribution over ensembles.

For each triple $(x, y_1, y_2)$ consider

$$\mathbb{E}_{\mathbf{r}\sim\mathcal{R}}\Big[(\hat{p}^{\mathbf{r}}(x, y_1, y_2) - p^*(x, y_1, y_2))^2\Big].$$

The estimator $\hat{p}^{\mathbf{r}}(x, y_1, y_2)$ is obtained by drawing

$$T \sim \mathrm{Bin}\big(k, p^*(x, y_1, y_2)\big),$$

and setting $\hat{p}^{\mathbf{r}}(x, y_1, y_2) = T/k$. Therefore

$$\mathbb{E}_{\mathbf{r}\sim\mathcal{R}}\Big[(\hat{p}^{\mathbf{r}}(x, y_1, y_2) - p^*(x, y_1, y_2))^2\Big] = \frac{\mathrm{Var}(T)}{k^2} \leq \frac{p^*(x, y_1, y_2)\big(1 - p^*(x, y_1, y_2)\big)}{k} \leq \frac{1}{4k}.$$

Next, the expected calibration error of a randomly chosen ensemble $\mathbf{r}$ satisfies

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{R}}\Big[\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, y_1, y_2\sim\pi_0(\cdot|x)}\big(\hat{p}^{\mathbf{r}}(x, y_1, y_2) - p^*(x, y_1, y_2)\big)^2\Big] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, y_1, y_2\sim\pi_0(\cdot|x)}\Big[\mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{R}}\big(\hat{p}^{\mathbf{r}}(x, y_1, y_2) - p^*(x, y_1, y_2)\big)^2\Big] \\
&\leq \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}, y_1, y_2\sim\pi_0(\cdot|x)}\Big[\frac{1}{4k}\Big] = \frac{1}{4k}.
\end{aligned}$$

Since the expected calibration error of a randomly chosen $\mathbf{r} \sim \mathcal{R}$ is at most $\frac{1}{4k}$, this implies that the same bound holds for at least one deterministic $\mathbf{r}$, as needed. □

### 2.3.2 Pruning and Outlier Control

Matching aggregate preferences is not enough if some reward functions deviate so sharply from the population that they could endorse behaviors most annotators would strongly reject. The next question is whether achieving calibration *forces* us to include such extreme outliers. Using the previously defined *disagreement score* $\Phi(\theta)$ as our proxy for how far a reward strays from mainstream judgments, our next result demonstrates it is possible to remove these extreme outliers without sacrificing too much pairwise calibration.

**Theorem 2.3.4.** *Suppose there exists a $k$-ensemble $\mathbf{r}$ that is $\varepsilon$-pairwise calibrated. Then, for all $\beta \geq 2$, the total weight of reward functions in $\mathbf{r}$ that are $(\beta, (\beta+1)\cdot\sqrt{\varepsilon})$-outliers is at most $\frac{1}{\beta-1}$. Furthermore, there exists another $\ell$-ensemble (for $\ell \leq k$) $\mathbf{r}'$ that is $(\sqrt{\varepsilon} + \frac{1}{\beta-1})^2$-pairwise calibrated and does not contain any $(\beta, (\beta+1)\cdot\sqrt{\varepsilon})$-outliers, and this $\mathbf{r}'$ can be computed with access only to $\mathbf{r}$.*

*Proof of Theorem 2.3.4.* Let $\mathbf{r} = (r_{\theta_j}; \alpha_j)_{j=1}^k$ be a $k$-ensemble that is $\varepsilon$-pairwise calibrated. Fix $\beta \geq 2$, and let $\gamma = (\beta+1)\sqrt{\varepsilon}$. Recall the disagreement score for a reward function $r_\theta$:

$$\Phi(\theta) = \Pr_{\substack{x\sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot|x), \\ i\sim\mathcal{N}}} \left[\mathbf{1}[r_\theta(x,y_1) > r_\theta(x,y_2)] \neq \mathbf{1}[y_1 \succ_i y_2 \mid x]\right].$$

Let $\Phi_{\min} = \inf_\theta \Phi(\theta)$. Then, $r_\theta$ is a $(\beta,\gamma)$-outlier if $\Phi(\theta) > \beta \cdot \Phi_{\min} + \gamma$.

Define the random variables

$$V^\theta = \mathbf{1}\{r_\theta(x,y_1) > r_\theta(x,y_2)\}, \qquad P^\star = p^\star(x,y_1,y_2), \qquad \hat{P} = \hat{p}(x,y_1,y_2),$$

The randomness for these variables is over the draw of a triple $(x,y_1,y_2)$ according to the underlying distribution (e.g., $x \sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot \mid x)$). For any two random variables $X$ and $Y$ (over triples), let

$$d(X,Y) = \mathbb{E}_{x\sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot|x)}[|X - Y|].$$

First, we show that $\Phi(\theta) = d(V^\theta, P^\star)$. To this end, we can rewrite the disagreement

score as

$$\begin{aligned}
&\Pr_{\substack{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x),\\ i\sim\mathcal{N}}}[\mathbf{1}[r_\theta(x,y_1)>r_\theta(x,y_2)]\neq\mathbf{1}[y_1\succ_i y_2\mid x]]\\
&=\mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x)}\left[\Pr_{i\sim\mathcal{N}}[\mathbf{1}[r_\theta(x,y_1)>r_\theta(x,y_2)]\neq\mathbf{1}[y_1\succ_i y_2\mid x]]\right].
\end{aligned}$$

Furthermore, for a fixed triple $(x, y_1, y_2)$, the inner term is:

$$\begin{aligned}
&\Pr_{i\sim\mathcal{N}}[\mathbf{1}[r_\theta(x,y_1)>r_\theta(x,y_2)]\neq\mathbf{1}[y_1\succ_i y_2\mid x]]\\
&=V^\theta\cdot\left(1-\Pr_{i\sim\mathcal{N}}[y_1\succ y_2\mid x]\right)+(1-V^\theta)\cdot\Pr_{i\sim\mathcal{N}}[y_1\succ y_2\mid x]\\
&=V^\theta\cdot(1-P^\star)+(1-V^\theta)\cdot P^\star.
\end{aligned}$$

Since $V^\theta$ is binary and $P^\star\in[0,1]$, this simplifies to $|V^\theta - P^\star|$. Thus, the disagreement score is indeed $\Phi(\theta) = d(V^\theta, P^\star)$.

Define $\hat{\Phi}(\theta) = d(V^\theta, \hat{P})$ and $\hat{\Phi}_{\min} = \inf_\theta \hat{\Phi}(\theta)$ as an analogous "disagreement score" with respect to $\hat{P}$. Next, we claim that if $\theta$ is a $(\beta,\gamma)$-outlier, then $d(V^\theta, \hat{P}) > \beta\cdot\hat{\Phi}_{\min}$. By the triangle inequality,

$$|d(V^\theta, P^\star) - d(V^\theta, \hat{P})| \le d(P^\star, \hat{P}).$$

The ensemble $\mathbf{r}$ is $\varepsilon$-pairwise calibrated, meaning $\mathbb{E}[(\hat{P} - P^\star)^2] \le \varepsilon$. By Jensen's inequality,

$$\mathbb{E}[|\hat{P} - P^\star|]^2 \le \mathbb{E}[(\hat{P} - P^\star)^2] \le \varepsilon.$$

Hence, $d(P^\star, \hat{P}) \le \sqrt{\varepsilon}$. So, for all $\theta$, we have

$$|d(V^\theta, \hat{P}) - d(V^\theta, P^\star)| \le \sqrt{\varepsilon}.$$

This also implies that $|\Phi_{\min} - \hat{\Phi}_{\min}| \le \sqrt{\varepsilon}$.

Combining these observations, we see that if $\theta$ is a $(\beta,\gamma)$-outlier, then $d(V^\theta, P^\star) > \beta\cdot\Phi_{\min} + (\beta+1)\cdot\sqrt{\varepsilon}$, which implies that $d(V^\theta, \hat{P}) > \beta\hat{\Phi}_{\min}$.

Next, we claim that for a fixed $\theta\in\Theta$,

$$d(V^\theta, \hat{P}) = \mathbb{E}_{\theta'\sim\mathbf{r}}[d(V^\theta, V^{\theta'})],$$

where $\theta' \sim \mathbf{r}$ denotes drawing from the ensemble, i.e., $\theta_j$ with probability $\alpha_j$. Expanding the definitions:

$$\mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x)}[|V^\theta - \hat{P}|] = \mathbb{E}_{\theta'\sim\mathbf{r}}[\mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x)}[|V^\theta - V^{\theta'}|]].$$

Swapping the order of the expectations:

$$\mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x)}[|V^\theta - \hat{P}|] = \mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x)}[\mathbb{E}_{\theta'\sim\mathbf{r}}[|V^\theta - V^{\theta'}|]].$$

For a fixed triple $(x, y_1, y_2)$, the equality $|V^\theta(x, y_1, y_2) - \hat{P}(x, y_1, y_2)| = \mathbb{E}_{\theta'\sim\mathbf{r}}[|V^\theta(x, y_1, y_2) - V^{\theta'}(x, y_1, y_2)]$ holds, because $V^\theta(x, y_1, y_2)$ is deterministic (in $\{0, 1\}$, and $V^{\theta'}$ is a Bernoulli random variable whose expectation is $\hat{p}(x, y_1, y_2)$.

By Markov's inequality, for any $\theta$, we have that

$$\begin{aligned}&\Pr_{\theta'\sim\mathbf{r}}[d(V^{\theta'}, V^\theta) \geq (\beta - 1) \cdot d(V^\theta, \hat{P})] \\ =\ &\Pr_{\theta'\sim\mathbf{r}}[d(V^{\theta'}, V^\theta) \geq (\beta - 1) \cdot \mathbb{E}_{\theta'\sim\mathbf{r}}[d(V^{\theta'}, V^\theta)]] \leq \tfrac{1}{\beta-1}.\end{aligned}$$

By the triangle inequality, $d(V^{\theta'}, \hat{P}) \leq d(V^{\theta'}, V^\theta) + d(V^\theta, \hat{P})$. Therefore, if $d(V^{\theta'}, \hat{P}) \geq \beta \cdot d(V^\theta, \hat{P})$, that implies that $d(V^{\theta'}, \hat{P}) \geq (\beta - 1) \cdot d(V^\theta, V^{\theta'})$. This gives us that

$$\Pr_{\theta'\sim\mathbf{r}}[d(V^{\theta'}, \hat{P}) \geq \beta \cdot d(V^\theta, \hat{P})] \leq \frac{1}{\beta - 1}.$$

Taking the infimum over choices of $\theta$ yields (e.g., using $\theta$ such that $d(V^\theta, \hat{P}) \to \hat{\Phi}_{\min}$)

$$\Pr_{\theta'\sim\mathbf{r}}[d(V^{\theta'}, \hat{P}) > \beta \cdot \hat{\Phi}_{\min}] \leq \tfrac{1}{\beta-1}.$$

Finally, as $(\beta, \gamma)$-outliers must satisfy $d(V^{\theta'}, \hat{P}) > \beta \cdot \hat{\Phi}_{\min}$, this immediately implies

$$\Pr_{\theta'\sim\mathbf{r}}[\theta' \text{ is a } (\beta, \gamma)\text{-outlier}] \leq \frac{1}{\beta - 1},$$

yielding the first part of the theorem.

Now, construct $\mathbf{r}'$ as follows: Order the $\theta_1, \ldots, \theta_k$ from $\mathbf{r}$ in non-decreasing order by $d(\theta, \hat{P})$ as $\theta_{i_1}, \ldots, \theta_{i_k}$. Remove $\theta_j$ if they are in the top $\frac{1}{\beta-1}$ of total weight according to this ordering (i.e., if $j$ satisfies that $\sum_{j'=j}^{k} \alpha_{j'} \leq \frac{1}{\beta-1}$). Then, renormalize the remaining $\alpha_j$s to sum to 1. The previously derived bound shows that at most $\frac{1}{\beta-1}$ of

the probability mass is on $\theta$ satisfying $d(\theta, \hat{P}) > \beta \cdot \Phi_{\min}$. Removing those with largest $d(\cdot, \hat{P})$ values ensures all such $\theta$ are removed. By the above arguments, this also implies that $\mathbf{r}'$ contains no $(\beta, \gamma)$-outliers. Furthermore, this computation depended only on $\mathbf{r}$ (i.e., we did not need access to true proportions $p^\star$).

It remains to show that $\mathbf{r}'$ is $(\sqrt{\varepsilon} + \frac{1}{\beta-1})^2$-pairwise calibrated. To this end, let $\hat{P}'$ be the analogous random variable to $\hat{P}$ for $\mathbf{r}'$, i.e., it takes on values $\hat{p}^{\mathbf{r}'}(x, y_1, y_2)$. We claim that $|\hat{P}' - \hat{P}| \leq \frac{1}{\beta-1}$ surely.

Let $S_{\text{rem}}$ be the set of indices of removed $\theta_j$. Let $\alpha_{\text{rem}} = \sum_{j \in S_{\text{rem}}} \alpha_j$. We know $\alpha_{\text{rem}} \leq \frac{1}{\beta-1}$. Let $\alpha_{\text{tot}} = \sum_{j \notin S_{\text{rem}}} \alpha_j = 1 - \alpha_{\text{rem}}$. For a fixed triple $(x, y_1, y_2)$, let $V_j = V^{\theta_j}(x, y_1, y_2) \in \{0, 1\}$. Then $\hat{P}(x, y_1, y_2) = \sum_{j \notin S_{\text{rem}}} \alpha_j V_j + \sum_{j \in S_{\text{rem}}} \alpha_j V_j$. And $\hat{P}'(x, y_1, y_2) = \frac{\sum_{j \notin S_{\text{rem}}} \alpha_j V_j}{\alpha_{\text{tot}}}$.

Consider the difference:

$$\begin{aligned}
\hat{P}' - \hat{P} &= \frac{\sum_{j \notin S_{\text{rem}}} \alpha_j V_j}{\alpha_{\text{tot}}} - \left( \sum_{j \notin S_{\text{rem}}} \alpha_j V_j + \sum_{j \in S_{\text{rem}}} \alpha_j V_j \right) \\
&= \left( \frac{1}{\alpha_{\text{tot}}} - 1 \right) \sum_{j \notin S_{\text{rem}}} \alpha_j V_j - \sum_{j \in S_{\text{rem}}} \alpha_j V_j \\
&= \frac{1 - \alpha_{\text{tot}}}{\alpha_{\text{tot}}} \sum_{j \notin S_{\text{rem}}} \alpha_j V_j - \sum_{j \in S_{\text{rem}}} \alpha_j V_j \\
&= \frac{\alpha_{\text{rem}}}{\alpha_{\text{tot}}} \sum_{j \notin S_{\text{rem}}} \alpha_j V_j - \sum_{j \in S_{\text{rem}}} \alpha_j V_j.
\end{aligned}$$

Let $A = \sum_{j \notin S_{\text{rem}}} \alpha_j V_j$ and $B = \sum_{j \in S_{\text{rem}}} \alpha_j V_j$. We know $0 \leq A \leq \sum_{j \notin S_{\text{rem}}} \alpha_j = \alpha_{\text{tot}}$, and $0 \leq B \leq \sum_{j \in S_{\text{rem}}} \alpha_j = \alpha_{\text{rem}}$. So, $\hat{P}' - \hat{P} = \frac{\alpha_{\text{rem}}}{\alpha_{\text{tot}}} A - B$.

Now, $0 \leq A \leq \alpha_{\text{tot}}$ and $0 \leq B \leq \alpha_{\text{rem}}$, so this difference is bounded in $[-\alpha_{\text{rem}}, \alpha_{\text{rem}}]$ and thus bounded in $[-\frac{1}{\beta-1}, \frac{1}{\beta-1}]$.

Finally, we show the pairwise calibration bound. We have just shown that $|\hat{P} - \hat{P}'| \leq \frac{1}{\beta-1}$, which implies that

$$\sqrt{\mathbb{E}_{x \sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot|x)}[(\hat{P} - \hat{P}')^2]} \leq \frac{1}{\beta - 1}.$$

Furthermore, $\varepsilon$-pairwise calibration implies that

$$\sqrt{\mathbb{E}_{x \sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot|x)}[(P^\star - \hat{P})^2]} \leq \sqrt{\varepsilon}.$$

By the triangle inequality for $L_2$ norms (Minkowski inequality),

$$\sqrt{\mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x)}[(P^\star-\hat{P}')^2]} \le \sqrt{\varepsilon}+\frac{1}{\beta-1}.$$

This implies that $\mathbf{r}'$ is $\left(\sqrt{\varepsilon}+\frac{1}{\beta-1}\right)^2$-pairwise calibrated, as needed. □

This proof begins with a Markov-style argument: most reward functions in $\mathbf{r}$ cannot deviate substantially from the ensembles aggregate predictions ($\hat{p}^{\mathbf{r}}$). Furthermore, we can bound the disagreement score of individual reward functions based on how much they deviate from $\mathbf{r}$ and the initial pairwise calibration of $\mathbf{r}$. Functions exceeding this bound are identified as outliers. We then demonstrate that removing these outliers—-a process requiring comparison only against $\hat{p}^{\mathbf{r}}$ (not the true $p^\star$ values)—-has minimal impact on the overall pairwise calibration.

### 2.3.3 Generalization

In practice, we do not have direct access to the true preference proportions $p^\star$. Rather, we typically work with a finite dataset $\mathcal{D}$ of pairwise comparisons. We assume the dataset has the following form: $\mathcal{D} = \{(x^{(i)}, y_1^{(i)}, y_2^{(i)}, p^{(i)})\}_{i=1}^N$, where each entry is generated by sampling a context $x \sim D_{\mathcal{X}}$, two candidate responses $y_1, y_2 \sim \pi_0(\cdot \mid x)$, and $n$ annotators $i_1, \ldots, i_n \sim \mathcal{N}$. The empirical preference is then set to $p = \frac{1}{n}\sum_{j=1}^n \mathbf{1}[y_1 \succ_{i_j} y_2 \mid x]$, the fraction of annotators who prefer $y_1$ over $y_2$.

Let $\mathcal{R}_k$ be the set of all $k$-ensembles. Our goal is to, using the dataset, find an ensemble $\mathbf{r} \in \mathcal{R}_k$ that achieves small pairwise calibration error on the true population:

$$\mathcal{L}(\mathbf{r}) = \mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x)}\left[(\hat{p}^{\mathbf{r}}(x,y_1,y_2) - p^\star(x,y_1,y_2))^2\right].$$

However, we only have access to the dataset, and the corresponding empirical loss:

$$\widehat{\mathcal{L}}(\mathbf{r}) = \mathbb{E}_{(x,y_1,y_2,p)\sim\mathcal{D}}\left[(\hat{p}^{\mathbf{r}}(x,y_1,y_2) - p)^2\right].$$

Using $\widehat{\mathcal{L}}$ to estimate $\mathcal{L}$ presents an inherent challenge: $\widehat{\mathcal{L}}$ is a biased estimator of $\mathcal{L}$.

**Lemma 2.3.5.** *For any ensemble $\mathbf{r} \in \mathcal{R}_k$, $\mathbb{E}_{\mathcal{D}}[\widehat{\mathcal{L}}(\mathbf{r})] = \mathcal{L}(\mathbf{r}) + C$, where*

$$C := \mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x),p\sim Bin(n,p^\star(x,y_1,y_2)/n}\left[(p - p^\star(x,y_1,y_2))^2\right].$$

*Proof.* First, note that by linearity over expectation,

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}(\mathbf{r})] = \mathbb{E}_{x\sim D_{\mathcal{X}},y_1,y_2\sim\pi_0(\cdot|x),p\sim\mathrm{Bin}(n,p^\star(x,y_1,y_2))/n}[(\hat{p}^{\mathbf{r}}(x,y_1,y_2)-p)^2].$$

Fix a triple $(x, y_1, y_2)$. For brevity, we will write $p^\star$ and $\hat{p}^{\mathbf{r}}$ instead of $p^\star(x, y_1, y_2)$ and $\hat{p}^{\mathbf{r}}(x, y_1, y_2)$, respectively. The inner term then becomes

$$\mathbb{E}_{p\sim\mathrm{Bin}(n,p^\star)/n}[(\hat{p}^{\mathbf{r}}-p)^2].$$

Furthermore, observe that $\mathbb{E}_{p\sim\mathrm{Bin}(n,p^\star)/n}[p] = p^\star$. Hence,

$$\begin{aligned}
\mathbb{E}_{p\sim\mathrm{Bin}(n,p^\star)/n}[(\hat{p}^{\mathbf{r}}-p)^2] &= \mathbb{E}_{p\sim\mathrm{Bin}(n,p^\star)/n}[(\hat{p}^{\mathbf{r}}-p^\star+p^\star-p)^2] \\
&= \mathbb{E}_{p\sim\mathrm{Bin}(n,p^\star)/n}[(\hat{p}^{\mathbf{r}}-p^\star)^2 - 2(\hat{p}^{\mathbf{r}}-p^\star)(p^\star-p) + (p^\star-p)^2] \\
&= (\hat{p}^{\mathbf{r}}-p^\star)^2 - 2(\hat{p}^{\mathbf{r}}-p^\star)(p^\star-p^\star) + \mathbb{E}_{p\sim\mathrm{Bin}(n,p^\star)/n}[(p^\star-p)^2] \\
&= (\hat{p}^{\mathbf{r}}-p^\star)^2 + \mathbb{E}_{p\sim\mathrm{Bin}(n,p^\star)/n}[(p^\star-p)^2].
\end{aligned}$$

Taking expectation over $x \sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot \mid x)$ yields the lemma statement. □

The term $C$ represents an irreducible error introduced due to only $n$ annotators responding per data point. Even a perfectly pairwise calibrated ensemble $\mathbf{r}$ will incur error $C$ in expectation over $\mathcal{D}$.

Despite this limitation, learning-theoretic tools allow us to provably estimate $\mathcal{L}$ up to the constant shift $C$. This implies that minimizing the empirical loss $\widehat{\mathcal{L}}$ remains a sound strategy for obtaining a model with optimal pairwise calibration.

To formalize this, let $\mathcal{F} = \{(x, y_1, y_2) \mapsto \mathbf{1}[r_\theta(x, y_1) \geq r_\theta(x, y_2)] \mid \theta \in \Theta\}$ be the class of binary comparison functions induced by the reward models, i.e., these are functions parameterized by $\Theta$, mapping triples $(x, y_1, y_2)$ to $\{0, 1\}$, outputting 1 exactly when $r_\theta(x, y_1) \geq r_\theta(x, y_2)$. Defining $\mathcal{L}'(\mathbf{r}) = \mathcal{L}(\mathbf{r}) + C$, we have:

**Theorem 2.3.6.** *Suppose $\mathcal{F}$ has finite VC-dimension $d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a dataset $\mathcal{D}$ of $N$ i.i.d. samples,*

$$\sup_{\mathbf{r}\in\mathcal{R}_k}\left|\mathcal{L}'(\mathbf{r}) - \widehat{\mathcal{L}}(\mathbf{r})\right| \leq \sqrt{\frac{2d'\log(\frac{eN}{d'})}{N}} + \sqrt{\frac{\log\frac{1}{\delta}}{2N}},$$

*where $d' = 20(d+1)k \cdot \log(2(d+1)k) \in O(kd\log(kd))$.*

The proof applies uniform convergence bounds for agnostic PAC learning. However, to do so requires bounding the pseudo dimension of the loss class $\{((x, y_1, y_2), p) \mapsto (\hat{p}^{\mathbf{r}}(x, y_1, y_2) - p)^2 \mid \mathbf{r} \in \mathcal{R}_k\}$. To bound this, we apply a sequence of transformations to $\mathcal{F}$, bringing it closer to the loss class, and show each step individually does not substantially increase the pseudo-dimension.

As a concrete example, suppose we have an embedding function $\varphi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^\ell$ which maps context-response pairs into $\mathbb{R}^\ell$, and that the reward functions $\{r_\theta \mid \theta \in \Theta\}$ are linear functions over these embeddings: $r_\theta(x, y) = \theta \cdot \varphi(x, y)$ where $\Theta = \mathbb{R}^\ell$. This setup corresponds to the common practice of learning a reward model by removing the final layer of a pretrained language model (yielding an embedding function), attaching a new linear head that maps the embedding to a scalar reward, and training only this final layer on preference data. Here, the comparison functions become $\mathbf{1}[\theta \cdot \varphi(x) \geq \theta \cdot \varphi(y))] = \mathbf{1}[\theta \cdot (\varphi(x) - \varphi(y)) \geq 0]$ which correspond to linear classifiers over the embedding space. This is known to have VC-dimension at most $\ell$ [149].

*Proof of Theorem 2.3.6.* We invoke the following uniform-convergence result:

**Theorem 2.3.7** (Mohri et al. [149], Theorem 11.8 (rephrased))**.** *Let $\mathcal{H}$ be a family of real-valued functions and*

$$\mathcal{G} = \{(x, y) \mapsto \mathcal{L}(h(x), y) \mid h \in \mathcal{H}\}$$

*the family of loss functions associated to $\mathcal{H}$. Assume that $\mathrm{Pdim}(\mathcal{G}) = d^\star$ and that the loss function $\mathcal{L}$ is non-negative and bounded by $M$. Let $\mathcal{D}$ be a distribution over $(x, y)$ and let $\bar{\mathcal{D}}$ be a sample of size $N$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of samples,*

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(h(x), y)] - \mathbb{E}_{(x,y)\sim\bar{\mathcal{D}}}[\mathcal{L}(h(x), y)] \right| \leq M\sqrt{\frac{2d^\star \log(\frac{eN}{d})}{N}} + M\sqrt{\frac{\log \frac{1}{\delta}}{2N}}$$ [4]

For our purposes, $\mathcal{H}$ is the set of functions $\{\hat{p}^{\mathbf{r}} \mid \mathbf{r} \in \mathcal{R}_k\}$ where each $\hat{p}^{\mathbf{r}}$ maps $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. The loss function $\mathcal{L}$ is the squared error. The domain points are triples $(x, y_1, y_2)$, drawn from the usual $x \sim D_{\mathcal{X}}, y_1, y_2 \sim \pi_0(\cdot \mid x)$. The target the

[4] The theorem statement in Mohri et al. [149] states one-sided error, but the proof implies two-sided one.

label is the empirical proportion

$$p \sim \mathrm{Bin}\big(n, p^\star(x, y_1, y_2)\big)/n,$$

Since $p \in [0, 1]$, $\mathcal{L}$ is bounded by $M = 1$.

Furthermore,

$$\hat{\mathcal{L}}(\mathbf{r}) = \mathbb{E}_{(x,y)\sim\bar{\mathcal{D}}}\big[\mathcal{L}(h(x), y)\big],$$

and by Lemma 2.3.5,

$$\mathcal{L}'(\mathbf{r}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}(h(x), y)\big].$$

It therefore suffices to show that $\mathrm{Pdim}(\mathcal{G}) \le 20(d+1)k \cdot \log(2(d+1)k)$, where

$$\mathcal{G} = \{((x, y_1, y_2), p) \mapsto (\hat{p}^{\mathbf{r}}(x, y_1, y_2) - p)^2 \mid \mathbf{r} \in \mathcal{R}\}.$$

For notational convenience, inputs to function classes (e.g., $(x, y_1, y_2)$ or $((x, y_1, y_2), p)$) may be generically denoted by $z$.

Recall the following standard definitions and results from learning theory. Fix a function class $\mathcal{F}'$ and a set of $m$ points, $\{z_1, \ldots, z_m\}$. For a binary vector $b \in \{0, 1\}^m$, called a *sign pattern*, $\mathcal{F}'$ realizes $b$ if there exists $f \in \mathcal{F}'$ such that $\mathbf{1}[f(z_i) \ge 0] = b_i$ for all $i$. We say $\mathcal{F}'$ *shatters* the set if all $2^m$ sign patterns are realized. The VC-dimension of $\mathcal{F}'$ is the size of the largest set it shatters. If $\mathcal{F}'$ is real-valued, its pseudo-dimension is the VC dimension of $\{(z, t) \mapsto f(z) - t \mid f \in \mathcal{F}'\}$. If the latter function class shatters a set, we say that $\mathcal{F}'$ pseudo-shatters that set. We write $\mathrm{Pdim}(\mathcal{F}')$ for the pseudo-dimension of $F'$. Finally, Sauer's Lemma [179] states that a class of VC-dimension $d'$ induces at most $(em/d')^{d'}$ sign patterns on $m$ points.

We upper-bound $\mathrm{Pdim}(\mathcal{G})$ by analyzing a sequence of function classes $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ and bounding their pseudo-dimensions $d_1, d_2, d_3$:

1. Linear combinations: Let $\mathcal{F}_1 = \{\sum_{j=1}^k \alpha_j \cdot f_j(z) \mid f_j \in \mathcal{F}, \alpha_j \in \mathbb{R}^k\}$, the set of $k$-sized linear combinations of functions in $\mathcal{F}$.

Fix $m$ points $(z_1, t_1), \ldots, (z_m, t_m)$. First, consider the maximum number of sign patterns of $z_1, \ldots, z_m$ which can be realized by (the binary-valued) $\mathcal{F}$. Since $\mathcal{F}$ has VC-dimension $d$, by Sauer's Lemma, this is at most $(e \cdot m/d)^d$. Thus, if we are picking $k$ such functions from $\mathcal{F}$, this can induce $(em/d)^{kd} \le (em)^{kd}$ combinations of sign patterns.

Fixing a choice of patterns $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \{0,1\}^m$. We now consider how many sign patterns of $\{(z_i, t_i)\}$ can be induced by various choices of $\boldsymbol{\alpha}$. Note that a certain $\boldsymbol{\alpha}$ will induce sign pattern with $b_i = \mathbf{1}[\boldsymbol{\alpha} \cdot (v_{1i}, \ldots, v_{ki}) - t_i \geq 0]$. Consider the set of $k$-dimensional linear functions mapping $\mathbb{R}^k \to \mathbb{R}$, $\{\mathbf{z} \mapsto \boldsymbol{\alpha} \cdot \mathbf{z} \mid \boldsymbol{\alpha} \in \mathbb{R}^k\}$. Note that this class has pseudo-dimension $k$ [149]. This immediately implies, by Sauer's Lemma and the definition of pseudo-dimension, that the number of sign patterns inducible on $\{(z_i, t_i)\}$ is at most $(em/k)^k \leq (em)^k$.

Together, these imply that the total number of sign patterns realizable by $\mathcal{F}_1$ is at most $(em)^{(d+1)k}$. Thus, if $\mathcal{F}_1$ can pseudo-shatter this set, all sign patterns must be realizable, so $2^m \leq (em)^{(d+1)k}$. Equivalently, $\log(2)m \leq (d+1)k \cdot \log(m)$, implying

$$m \leq \frac{(d+1)k}{\log(2)} \cdot \log(m) \leq 2 \cdot (d+1) \cdot k \cdot \log m.$$

By Lemma A.1 of Shalev-Shwartz and Ben-David [182], $m \leq 4(d+1)k \cdot \log(2(d+1)k)$. Hence, $d_1 \leq 4(d+1)k \cdot \log(2(d+1)k)$.

2. Affine shifts: Let $\mathcal{F}_2 = \{(z, y) \mapsto f(z) - y \mid f \in \mathcal{F}_1\}$.

If $\mathcal{F}_2$ pseudo-shatters

$$((z_1, y_1), t_1), \ldots, ((z_m, y_m), t_m),$$

then $\mathcal{F}_1$ pseudo-shatters

$$(z_1, t_1 - y_1), \ldots, (z_m, t_m - y_m),$$

implying that the pseudo-dimension of $\mathcal{F}_2$ is at most that of $\mathcal{F}_1$. Hence, $d_2 \leq d_1$.

3. Squaring: Let $\mathcal{F}_3 = \{z \mapsto f(z)^2 \mid f \in \mathcal{F}_2\}$.

Fix $m$ points $S = \{(z_1, t_1), \ldots (z_m, t_m)\}$ pseudo-shattered by $\mathcal{F}_3$. Assume each $t_i > 0$, as otherwise that point alone cannot be shattered. Furthermore assume that $f(z_i) \neq t_i$ for any $f \in \mathcal{F}_3$, otherwise we can adjust $t_i$ such that the set is still pseudo-shattered and equality does not hold.

Now, consider the sign patterns of function $f \in \mathcal{F}_2$ on the $2m$ points

$$S' = \{(z_1, \sqrt{t_1}), \ldots, (z_m, \sqrt{t_m}), (z_1, -\sqrt{t_1}), \ldots, (z_m, \sqrt{t_m})\}.$$

Observe that if two functions $f, f' \in \mathcal{F}_2$ have that $f(z)^2$ and $f'(z)^2$ differ in sign pattern on $S$, then $f(z)$ and $f'(z)$ differ in sign pattern on $S'$, as if $f(x_i)^2 < t_i$ and $f'(x_i)^2 \geq t_i$, then that implies $-\sqrt{t_i} < f(x_i) < \sqrt{t_i}$ and either $f'(x_i) > \sqrt{t_i}$ or $f'(x_i) < \sqrt{t_i}$, so $f$ and $f'$ must differ on at least one point. Therefore, the number of sign patterns $\mathcal{F}_2$ can realize on $S'$ must be at least as large as the number of sign patterns $\mathcal{F}_3$ can realize on $S$. By Sauer's lemma, the number of sign patterns $\mathcal{F}_2$ can realize on $S'$ is at most $\left(\frac{e \cdot 2m}{d_2}\right)^{d_2}$. Since $\mathcal{F}_3$ pseudo-shatters these points, $2^m \leq \left(\frac{e \cdot 2m}{d_2}\right)^{d_2}$, and, equivalently, $2^{m/d_2}/(m/d_2) \leq 2e$. Now if $m \geq 5d_2$, then, since $2^z/z$ is increasing for $z \geq 2$, the LHS is at least $2^5/5 > 2e$. Therefore, $d_3 \leq 5 \cdot d_2$.

Finally, observe that after all of these transformations, $\mathcal{G}$ is in fact a subset of $\mathcal{F}_3$. Therefore,

$$\operatorname{Pdim}(\mathcal{G}) \leq d_3 \leq 5d_2 \leq 5d_1 \leq 20(d+1)k \log\big(2(d+1)k\big),$$

as needed. □

## 2.4 Implementation via Residual Reward Calibration

Given that only a small mixture is theoretically sufficient, the practical question is: *can we construct it in practice?* In principle, we could learn an ensemble by solving the full optimization problem:

$$\min_{\substack{\alpha_1,\ldots,\alpha_k \\ \theta_1,\ldots,\theta_k}} \mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}} \left[ \left(p(x,y_1,y_2) - \hat{p}(x,y_1,y_2)\right)^2 \right]. \tag{2.2}$$

[5]

However, directly minimizing this objective requires a computationally intensive search over both the mixture weights and the reward model parameters, rendering it impractical.

Instead, we propose a heuristic approach based on *forward stagewise additive modeling (FSAM)* (see Hastie et al. [103] for an overview) that decomposes the problem into a sequence of more tractable subproblems. At each step, we fit a new reward model to the current residual error of the ensemble, keeping previously learned models fixed; pick a mixing weight that best reduces that error, and append the new model to the mixture. We observe that, despite its heuristic nature, this approach tractably

[5] Throughout this section, we use $p(x, y_1, y_2)$ to denote the observed fraction of annotators that responded $y_1 \succ y_2 \mid x$.

recovers ensembles that are approximately pairwise-calibrated on real-world datasets.

We start by detailing how we train the first reward model (i.e., $k = 1$). Even in this case, Equation (2.2) already departs from vanilla RLHF in two ways. First, the target is the observed preference fractions, which we refer to as the *soft labels*. Even in cases where multiple annotators disagreed on which response is better, typical alignment datasets flatten these to binary labels $p(x, y_1, y_2) \in \{0, 1\}$ by collapsing multiple annotators' votes into a single "majority-decision" label [44, 80, 122, 123, 134, 135, 165]. In fact, several high-profile alignment deployments train exclusively on these binary signals and treat them as a gold standard, discarding the underlying soft-label information as noise [17, 196, 199].[6] By contrast, we *need* access to this soft label data, as we aim to match the observed preference fractions. This training regime is conceptually distinct from methods that merely smooth hard labels — such as label smoothing [150, 201] — or from approaches that try to infer a latent consensus from soft labels [48, 79]. Second, we present our method in terms of the MSE loss, while typical RLHF uses cross-entropy. MSE aligns cleanly with later ensemble iterations, though the same construction could be carried out with a cross-entropy loss as well.

Having defined the $k = 1$ special case, we now turn to the rest of the ensemble. At iteration $j > 1$ we expand the current ensemble $\mathbf{r}_{j-1} = (r_{\theta_1}, \ldots, r_{\theta_{j-1}}; \alpha_1, \ldots, \alpha_{j-1})$ by adding a new reward model $r_{\theta_j}$. Let

$$\varepsilon_j(x, y_1, y_2) = p(x, y_1, y_2) - \hat{p}^{\mathbf{r}_{j-1}}(x, y_1, y_2)$$

be the residual calibration error. We fit $r_{\theta_j}$ by minimizing

$$\mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}}\left[\left(\varepsilon_j(x, y_1, y_2) - \sigma\left(r_{\theta_j}(x, y_1) - r_{\theta_j}(x, y_2)\right)\right)^2\right],$$

using the sigmoid $\sigma(z) = 1/(1 + \exp(-z))$ as a smooth proxy for the binary vote indicator.

Each reward model starts from the same supervised fine-tuned (SFT) checkpoint: we drop the token-prediction head, attach a freshly initialized single-node reward head, keep all other parameters frozen from SFT, and then fine-tune on preference

[6] This choice is surprising because in the BT model, soft-label BCE matches the true log-likelihood, and the vote fraction reveals how small the underlying score gap is between the two responses (see Appendix E of the full version).

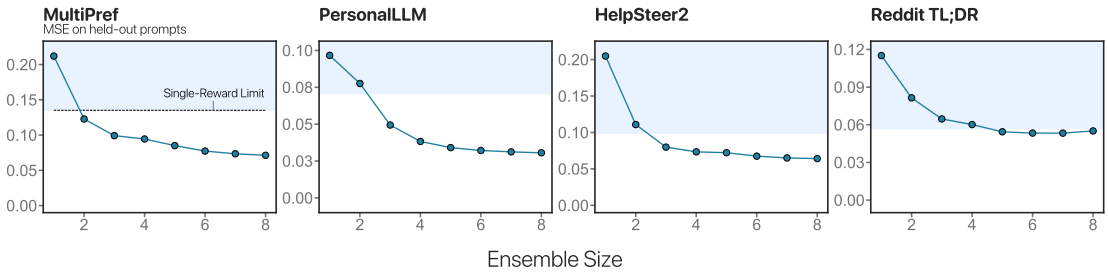



**Figure 2.2:** MSE calibration error on held-out prompts as a function of ensemble size $k$. Each curve is obtained by training a mixture of eight weak reward models with our FSAM procedure, which greedily adds one base model at a time to minimize the residual calibration error. The shaded band marks the floor for any single deterministic reward model, $\sum_i \min\{p_i^2, (1 - p_i)^2\}$. Across all datasets, ensembles of only two to four rewards already beat this single-model bound.

pairs.[7] Once $r_{\theta_j}$ is trained, we add it to the ensemble, and reoptimize coefficients $(\alpha_1, \ldots, \alpha_j)$ to minimize the training MSE, given $r_{\theta_1}, \ldots, r_{\theta_j}$.

The method incrementally concentrates capacity on the "difficult" comparisons — those with the largest residual magnitude. Residuals may be negative or exceed 1, applying even more pressure to the subsequent reward model to converge toward the soft labels.

The procedure can terminate after a fixed $k$ iterations or, alternatively, via early stopping when the validation loss stops improving. Each additional iteration adds one fine-tuning pass, so runtime grows linearly with the number of reward models.

## 2.5 Empirical Results

We now evaluate the FSAM procedure to test whether this heuristic can learn pairwise-calibrated ensembles on real alignment datasets. We focus on two questions: (i) *Can a small ensemble of weak reward models match the observed vote fractions more accurately than any single reward model?* (ii) *Do the individual reward models in the ensemble capture distinct preference patterns rather than duplicating one another?* For both, we find positive results.

**Datasets.** To be suitable for our experiments, datasets must satisfy two conditions: (i) they must include — or allow us to reconstruct — pairwise comparisons between distinct LLM outputs; and (ii) every comparison must be rated by more than one

[7] Training details appear in Appendix F.1 of the full version.

annotator. We use four public datasets that satisfy these requirements and exhibit annotator disagreement: `MultiPref` [146], `PersonalLLM` [217], `HelpSteer2` [203], and `Reddit TL;DR` [196]. Summary statistics appear in Table 2.1; further details about the datasets, including pre-processing, can be found in Appendix F of the full version. These datasets vary substantially in both domain and size.

| Name | Pref. Pairs | Unique Prompts | Annotation | Avg # Annots. | Avg $p$ |
|---|---|---|---|---|---|
| `MultiPref` [146] | 9,413 | 4,791 / 532 | Human annotators | 4.0 | 63% |
| `PersonalLLM` [217] | 263,256 | 9,402 / 1,000 | Model-based scores | 10 | 76% |
| `HelpSteer2` [203] | 21,000 | 10,000 / 1,000 | Human annotators | 3.5 | 74% |
| `Reddit TL;DR` [196] | 3,217 | 729 / 845 | Human annotators | 7.56 | 84% |

**Table 2.1:** Summary statistics for the four preference datasets (see Appendix F.2 of the full version for details). $p$ is defined here as the mean majority-agreement score, i.e., the average share of annotators who chose the majority option; by definition, this share is at least 0.5.

Ensemble of weak reward models vs. an optimal single reward model. We fit an ensemble of k=8 reward models on each dataset, starting from supervised fine-tuned checkpoints of Meta-Llama-3-8B [2] (a juggernaut in the small-model bracket). We compare the ensemble to a *theoretically optimal single deterministic reward model* that always selects the majority-preferred answer for every comparison. Such a model minimizes binary error and its mean-squared calibration loss cannot fall below $\sum_i \min\{p_i^2, (1-p_i)^2\}$. A mixture of weak deterministic reward models offers a finer-grained set of outputs but, a priori, need not perform better.

Our results (Figure 2.2) show that, in many cases, an ensemble of only 2-4 such rewards already achieves noticeably better calibration on held-out prompts. Gains tend to taper off around six models. Thus, combining a handful of weaker reward models delivers a calibration accuracy that no single deterministic model can match.

Note that performance is reported as mean-squared calibration error. Other empirical pluralistic alignment experiments frequently report *accuracy*, but this is not meaningful here because the task is not majority classification and annotator identities are unavailable (see Appendix E of the full version).

Do the reward models capture distinct preferences? To test diversity at the policy level, we use Best-of-$N$ (BoN) sampling as a proxy for fine-tuning [90, 152]. Concretely, we select 50 prompts that the ensemble was not trained on, generate

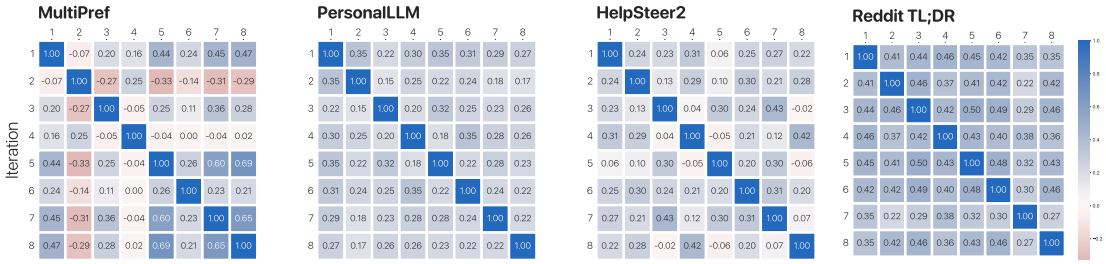



**Figure 2.3:** Pairwise Kendall — $\tau$ correlation scores between reward models in each ensemble, obtained by ranking 100 high-temperature continuations for 50 prompts. A score of 0 corresponds to random rankings, while a score of 1 corresponds to identical rankings; lower scores thus indicate greater diversity. The Reddit dataset — the smallest and focused on summarization — records the highest scores.

100 diverse continuations from the frozen SFT model using high-temperature sampling, and score each response with every reward model in the ensemble. We quantify distinctness by the normalized Kendall — $\tau$ rank correlation coefficient [119], which counts the number of pairwise disagreements between two ranking lists, averaged over all prompts. Prompts for the first three datasets were taken from the PRISM dataset [125] (covering value-centric and controversial topics); prompts for the Reddit dataset were taken from its validation set.

The reward models within each ensemble are distinct and qualitatively different (Figure 2.3). This diversity is further illustrated by comparing the top-ranked response chosen by each model (Appendix G of the full version).

## 2.6 Discussion

Our approach raises several important considerations regarding its scope, alternatives, and limitations.

First, for highly contested or sensitive issues, pairwise calibration may not always be appropriate. We are not aiming for the model to express its *own* opinion or mirror the population; rather, a neutral answer — or, when necessary, outright refusal — s often the safer choice. However, while refusal policy is a simple fallback for disallowed or highly sensitive content, over-reliance can hamper the system's usefulness when thoughtful, context-aware answers are needed. Even ostensibly neutral responses may embed hidden framing biases, so it remains essential to understand and address those

biases and to let the model meaningfully adapt to different user contexts, cultures, and values whenever such adaptation is appropriate.

Second, one might ask whether a single, unified model could achieve similar pluralistic alignment via alternate techniques. Indeed, one possible approach is to directly instruct the model within its context to adopt a particular perspective—known as *in-context steering* or *persona modulation* [42, 107, 113, 159]. However, recent empirical evidence suggests that language models instructed to emulate specific viewpoints frequently produce inconsistent or overly stereotyped responses, and often fail to capture nuanced differences between distinct user groups [121]. By contrast, our approach explicitly incorporates architectural support for pluralism by training multiple distinct reward functions.

Finally, while pairwise calibration ensures that the ensemble accurately reflects the preferences of annotators for *pairs* of responses given a context, it does not guarantee accurate representation of higher-order judgments over larger sets of responses. Ideally, for any reasonably sized set of $t$ responses, the proportion of reward functions ranking a given response highest would (approximately) match this preference in the population. However, recovering such higher-order preference structures is information-theoretically infeasible using pairwise data alone. Capturing these relationships would require richer elicitation, such as rankings or best-of-$\ell$ judgments over larger response sets.

# Part II

# Efficient Preference Elicitation

# 3

# Representation with Incomplete Votes

## 3.1 Introduction

A recent surge of interest in empowering citizens through online civic participation has spurred the development of a number of platforms [11, 76, 108, 110, 175, 183]. A particularly successful example is *Polis* [193],[1] an open-source "system for gathering, analyzing, and understanding what large groups of people think in their own words." It has been widely used by local and national government agencies around the world. Most notably, it is the basis of vTaiwan, a system commissioned by the government of Taiwan, whose participatory process — involving thousands of ordinary citizens — has led to new regulation of ride-sharing services and financial technology. A similar (albeit commercial) system called *Remesh*[2] allows users to "save resources by democratizing insights in live, flexible conversations with up to 1,000 people at the same time."[3]

The key idea underlying both systems is simple and broadly applicable: Participants can submit free-text comments about the discussion topic at hand and choose to agree or disagree with others' comments presented to them by the platform. An

[1] `https://pol.is`
[2] `https://www.remesh.ai`
[3] See also `consider.it`, `citizens.is`, `make.org`, `kialo.com`.

essential part of the process is the aggregation of these opinions toward an "understanding of what large groups of people think." Polis, for instance, displays a list of comments that received the most support among participants to whom they were shown. But this aggregation method may fail to represent minority groups, even those that are very large: if 51% of participants agree with one set of comments, while 49% of participants agree with another set of comments, only comments from the first set will appear on this list. Polis has recognized this problem and sought to mitigate it by employing a second, more elaborate procedure [193].[4] While this procedure has produced interesting results in practice, it does not guarantee summarizations that are representative of the discussion in a rigorous sense.

In this paper, we reexamine opinion aggregation in systems such as Polis and Remesh through the lens of computational social choice [36]. We observe that *selecting a subset of comments based on agreements and disagreements is equivalent to electing a committee based on approval votes*. From this viewpoint, the primary aggregation method used by Polis corresponds to classical approval voting (AV). There is substantial work — starting with the paper of Aziz et al. [14] — on *approval-based committee elections* that seeks to avoid the shortcomings of AV by guaranteeing that the selected committee satisfies fairness notions. To define one such notion (which is not satisfied by AV), note that if the size of the committee is $k$ and the number of voters is $n$, a subset of $n/k$ voters is large enough to demand a seat on the committee if they agree on at least one candidate. This intuition is captured by a property called *justified representation (JR)*, which guarantees that every such subset of voters has an approved candidate on the committee.

There is a major gap, however, between the literature on approval-based committee elections and the reality of systems like Polis and Remesh: these systems only have access to partial votes. For example, in the discussion facilitated by Polis around ridesharing regulation in Taiwan, 197 comments were submitted, but each participant only cast a vote on 10.57 comments on average — roughly 5% of all comments. Our main conceptual insight is that we can overcome the partial-information gap via statistical estimation and adaptive querying (i.e., by deciding which comments to show to incoming users based on previous votes).[5]

[4]The idea is to find clusters of participants with similar opinions and then ensure that each cluster is represented by comments that distinguish it from the others.

[5]There is a body of work in computational social choice related to incomplete votes. For

Our approach and results. In our model, each voter (user) can be asked to express their opinion (approval/disapproval) about at most $t$ candidates (comments). More formally, a *query* asks a randomly-chosen voter for their approval votes on a subset of candidates $S$ of size $|S| \leq t$. Note that this query model is consistent with how Polis works, where participants express their agreement or disagreement with the comments shown to them by the system. We can view the response to such a query, i.e., the approval votes of a single voter, as noisy information about the profile of the entire population of voters (restricted to these $t$ candidates). Therefore, we refer to these real-world queries as *noisy queries*.

Before we discuss this setting, we consider a simplified setting in Section 3.3, where queries yield the profile of the entire population of voters on the $t$ candidates in the query. While such *exact queries* are not realistic, they provide an abstraction that is easier to study and allows us to derive lower bounds on the number of queries required to achieve JR (which apply also to the noisy-query setting, since it is strictly harder). We start by studying the required number of queries of *non-adaptive* algorithms, which decide on their queries before any votes are cast. While non-adaptive algorithms may be preferable in some cases (e.g., because no voter can influence what alternatives are shown to other voters or because computation can be performed offline), we show that they are impractical because they must ask at least $\Omega(m^{11})$ queries (and hence voters) to achieve JR, where $m$ is the number of candidates.

Therefore, we focus on *adaptive* algorithms in the rest of the paper. In Section 3.3.2 we adapt a local-search algorithm of Aziz et al. [15] to the case of exact queries and show that it can achieve JR (and even stronger properties) with $\mathcal{O}(mk^2 \log k)$ queries.

In Section 3.4, we move on to the realistic, noisy-query model, where a query corresponds to a single voter. Since we need to estimate the answer to each exact query using multiple noisy queries to control uncertainty, the query complexity of the adaptive algorithm for the same guarantees increases to $\mathcal{O}\left(mk^6 \log k \log m\right)$. By applying

instance, some papers aim to find winning committees, given incomplete approval votes, or to fill in the missing votes, given knowledge about the domain of approval profiles [109, 198, 214]. However, these papers are primarily concerned with the computational complexity of these problems, while we focus on information-theoretic questions. There is also related work that studies the problem of determining the winner given only partial rankings [73, 206], but this setting is mathematically different from ours. Furthermore, prior work does not consider the adaptive setting, where we query voters sequentially and decide on the next question based on previous votes.

martingale theory, we develop an extension of this algorithm that allows the reuse of votes in a statistically sound way.

In Section 3.5 we show empirically (on real datasets from Polis and Reddit) that this extension allows us to find committees satisfying (approximate) JR (and stronger properties) despite access to little information (i.e., few voters, each voting on only a small fraction of the comments).

## 3.2 Preliminaries

We begin by introducing the standard approval-based committee selection setting [14]. For $s \in \mathbb{N}$, we use the notation $[s] = \{1, \dots, s\}$. We have a set $N = [n]$ of $n$ voters and a set $C$ of $m$ candidates. Each voter $i \in N$ approves a set of candidates $A_i \subseteq C$. We refer to the vector $\mathbf{A} = (A_1, \dots, A_n)$ as an *approval profile*. The goal is to choose a *committee* $W \subseteq C$ of size $k \leq m$. The value $k$ is called the *target committee size*. We refer to an algorithm that takes as input the profile and candidates and outputs a committee of size $k$ as a *$k$-committee-selection algorithm*.

**Notions of representation.** We say that a group of voters $V \subseteq N$ is *$\ell$-large* if $|V| \geq \ell \cdot \frac{n}{k}$; $V$ is *$\ell$-cohesive* if $|\bigcap_{i \in V} A_i| \geq \ell$. Aziz et al. [14] introduced the following two fairness notions:

**Definition 3.2.1** (Justified Representation (JR))**.** *A committee $W$ provides JR if for every 1-large, 1-cohesive group of voters $V$, there exists a voter $i \in V$ who approves a member of $W$, i.e., $|A_i \cap W| \geq 1$.*

**Definition 3.2.2** (Extended Justified Representation (EJR))**.** *A committee $W$ provides EJR if for every $\ell \in [k]$ and every $\ell$-large, $\ell$-cohesive group of voters $V$, there exists a voter $i \in V$ who approves at least $\ell$ members of $W$, i.e., $|A_i \cap W| \geq \ell$.*

We also study the following approximate version of EJR:

**Definition 3.2.3** ($\alpha$-Extended Justified Representation ($\alpha$-EJR))**.** *A committee $W$ provides $\alpha$-EJR if for every $\ell \in [k]$ and every $\frac{\ell}{\alpha}$-large, $\ell$-cohesive group of voters $V$, there exists a voter $i \in V$ who approves at least $\ell$ members of $W$, i.e., $|A_i \cap W| \geq \ell$.*

Fernández et al. [72] proposed another notion of representation called the *average satisfaction* of a group of voters $V$ for a committee $W$, defined as $\text{avs}_W(V) = \frac{1}{|V|} \sum_{i \in V} |A_i \cap W|$. Related to this quantity, we define the following property:

**Definition 3.2.4** ($\alpha$-Optimal Average Satisfaction ($\alpha$-OAS))**.** *A committee $W$ provides $\alpha$-OAS if for every $\lambda \in [0, k]$ and every $\frac{\lambda+1}{\alpha}$-large, $\lambda$-cohesive group of voters $V$, we have $avs_W(V) \geq \lambda$.*

This property measures how close a committee is to the maximum average satisfaction that can be guaranteed to hold for all elections. To see this, note that the condition above is equivalent to the following condition: for every $\ell \in [\frac{1}{\alpha}, \frac{k+1}{\alpha}]$ and every $\ell$-large, $(\alpha\ell - 1)$-cohesive group of voters $V$, we have $\text{avs}_W(V) \geq \alpha\ell - 1$. This implies a *proportionality guarantee* [191] of $g(\ell, k) = \alpha\ell - 1$. Since there is no selection rule that satisfies a proportionality guarantee with $g(\ell, k) > \ell - 1$ for all elections [15, 191], $\alpha = 1$ is the best we can hope for, so we refer to 1-OAS simply as OAS.

**Proportional approval voting.** *Proportional Approval Voting (PAV)* is a widely-studied committee selection algorithm: given an approval profile $\mathbf{A}$ and a committee size $k$, it returns a committee $W$ of size $k$ maximizing the *PAV score*, defined as

$$\text{PAV-SC}(W) := \frac{1}{n} \sum_{i \in N} \sum_{j=1}^{|A_i \cap W|} \frac{1}{j}.$$

PAV satisfies EJR and OAS [15, 72], but is NP-hard to compute [16]. Consequently, Aziz et al. [15] propose a local search approximation of PAV (LS-PAV), which continues to satisfy EJR and OAS, but, unlike PAV, runs in polynomial time. As we shall see, LS-PAV is a useful basis for algorithms in our query model.

## 3.3 Exact Queries

In the exact-query setting, the response $R$ to a query $Q$ consists of a proportion $p_S$ for every subset $S \subseteq Q$, where $p_S$ is the proportion of voters who only approve the candidates in $S$ among the queried candidates $Q$, i.e.,

$$p_S := \frac{1}{n} \sum_{i \in N} \mathbb{I}[A_i \cap Q = S],$$

where $\mathbb{I}$ is the indicator function. We refer to an algorithm that makes queries of size $t$, receives this type of response, and outputs a committee of size $k$ as a $(k, t)$-*committee selection algorithm with exact queries*. We say an algorithm is *adaptive* if the queries it chooses depend on responses from previous queries. Note that we allow

all of our algorithms to be randomized. In the following, we ask how many queries are needed to guarantee the notions of representation introduced in Section 3.2.

### 3.3.1 Nonadaptive Algorithms

In this section, we think of $m$ as large (many comments will be submitted to the system), while we think of $k$ and $t$ as small constants (since we wish to select only a few comments and voters have limited time). Since we are primarily interested in *lower* bounds on the query complexity of non-adaptive algorithms, we consider only JR, the weakest fairness criterion.

An initial observation is that, if $t \geq k$, JR can always be guaranteed with $O(m^k)$ queries, as simply querying every set of $k$ candidates provides all the information necessary to run PAV. For $k = 1$, this bound is tight, as voters could all approve only a single candidate, which will take a linear number of queries to find. Our first result is a tight quadratic lower bound for $k = 2$.

**Theorem 3.3.1.** *For any constants $k$ and $t$ such that $k \geq 2$, and any $\varepsilon > 0$, any non-adaptive $(k,t)$-committee selection algorithm that makes fewer than $\Omega(m^2)$ queries satisfies JR with probability at most $\varepsilon$.*

This result provides a separation between the non-adaptive and the adaptive settings: In Section 3.3.2, we discuss an adaptive $(k,t)$-committee selection algorithm guaranteeing JR with only $O(m)$ queries for any $k$ and $t$ such that $k < t$.

Theorem 3.3.1 follows from a more general result that we present formally in Appendix B of the full version. Here, we illustrate the argument by considering the special case where $t = k = 2$ and $\varepsilon = \frac{5}{6}$: Consider an adversary that picks a random set of 3 candidates, call them 1, 2, and 3, and answers queries according to the approval matrix visualized in Figure 3.1(a): half of the voters approve only candidate 1, and the other half of the voters approve only candidates 2 and 3. To satisfy JR, the algorithm needs to include candidate 1 in the committee. However, if the algorithm never queries $\{1,2\}$, $\{1,3\}$, or $\{2,3\}$, it receives no information that can distinguish candidates 1, 2, and 3 from each other, so it can do no better than selecting a random pair from these three candidates, which will succeed with probability $\frac{2}{3}$. This problematic case will occur frequently if the number of queries is not very large, say $\frac{1}{18} \cdot \binom{m}{2}$: Since there are $\binom{m}{2}$ pairs of candidates, the probability that the algorithm queries any

randomly selected pair of candidates is at most $\frac{1}{18}$. By the union bound, the probability that the algorithm queries any of $\{1,2\}$, $\{1,3\}$, or $\{2,3\}$ is at most $3 \cdot \frac{1}{18} = \frac{1}{6}$. To summarize, for the algorithm to succeed, it either needs to get lucky during the querying phase, which happens with probability at most $\frac{1}{6}$, or during the selection phase, which happens with probability at most $\frac{2}{3}$. By the union bound, the algorithm succeeds with probability at most $\frac{1}{6} + \frac{2}{3} = \frac{5}{6}$.

A natural follow-up question is whether the $O(m^k)$ upper bound is tight for larger $k$. Interestingly, this is *not* the case for $k \geq 3$:

**Theorem 3.3.2.** *For any $t \geq \frac{2}{3}k$, there exists a $(k,t)$-committee selection algorithm guaranteeing JR with $O(m^t)$ exact queries.*

*Proof.* Consider Algorithm 1, described below.

**Algorithm 1** $(k,t)$-non-adaptive (for $t \geq \frac{2}{3}k$)

Query every set of candidates of size $t$
**for** $i \leftarrow 1, 2, \ldots, t$ **do**
    $c_i \leftarrow$ approval winner among voters not approving $c_1, c_2, \ldots, c_{i-1}$
**for** $i \leftarrow t+1, \ldots, \lfloor \frac{3}{2}t \rfloor$ **do**
    $c_i \leftarrow$ arbitrary default candidate
    **for** $c \in C$ **do**
        $A \leftarrow$ set of voters approving $c$ but not any of $\{c_1, c_2, \ldots, c_{t-1}\}$
        $B \leftarrow$ set of voters approving $c$ but not any of $\{c_1, c_2, \ldots, c_{i-1}\} \setminus \{c_{\lfloor t/2 \rfloor}, \ldots, c_{t-1}\}$
        **if** $|A| \geq \frac{n}{k}$ and $|B| \geq \frac{n}{k}$ **then**
            $c_i \leftarrow c$
**return** $\{c_1, c_2, \ldots, c_k\}$

Provided that $t \geq \frac{2}{3}k$, it is straightforward to verify that the **if** condition can be checked using only information about sets of voters of size $t$. Thus, Algorithm 1 is indeed a non-adaptive $(k,t)$-committee selection algorithm with exact queries.

For each $i \in \{1, 2, \ldots, \lfloor \frac{3}{2}t \rfloor\}$, we say that a voter is *satisfied on round $i$* if it approves of $c_i$, but none of the previously selected candidates $c_1, c_2, \ldots, c_{i-1}$, and we say that a voter is *satisfied by round $i$* if it was satisfied on some round $j \leq i$. We

prove that the final committee satisfies JR by counting the fraction of voters that are satisfied on each round. Indeed, JR is equivalent to the property that there is no 1-cohesive $\frac{1}{k}$-fraction of voters that is left unsatisfied by the $k^{\text{th}}$ round.

The case where $k \leq t$ is easy: on each of the first $k$ rounds, either we satisfy a $\frac{1}{k}$ fraction of voters, or there is no 1-cohesive set of $\frac{n}{k}$ unsatisfied voters. Thus, by round $k$, either all voters are satisfied, or the remaining set of unsatisfied voters has no 1-cohesive set of size $\frac{n}{k}$.

Now suppose that $k > t$. For each $i$, let $x_i$ denote the fraction of voters that are satisfied on round $i$. Note that the sequence of $x_i$ are weakly decreasing for $i \leq t$. Again, if any $x_i < \frac{1}{k}$, it means that there is no 1-cohesive $\frac{1}{k}$-fraction of unsatisfied voters after round $i$, so JR is already satisfied. So assume each $x_i \geq \frac{1}{k}$ for all $i \leq t$. Further, if on any round $i > t$ we fail to find a candidate $c$ making the **if** condition true, we claim that JR is already satisfied. For if JR were not satisfied, then there would be some candidate $c$ approved by a $\frac{1}{k}$-fraction of voters $S$ who approve of no previous candidates. Clearly, we would then have $S \subseteq A$ and $S \subseteq B$, so $A$ and $B$ both contain at least $\frac{1}{k}$ fractions of voters.

Thus, we may assume that, for each $i > t$, candidate $c_i$ satisfies the **if** statement on round $i$. Consider an arbitrary round $i$. Let $A$ and $B$ denote the respective values of the variables on the iteration of the inner loop where $c_i$ was set to its ultimate value. Observe that the candidates enumerated in the definitions of $A$ and $B$ cover all previously selected candidates. This means that $A \cap B$ is precisely the set of voters approving $c_i$ and not any of the previous candidates; in other words, $A \cap B$ is the set of voters satisfied on round $i$. On the other hand, since candidates $c_1$,$c_2$,$\ldots$,$c_{\lfloor t/2 \rfloor}$ are enumerated in the definitions of both sets, it follows that $A \cup B$ is a set of voters approving $c_i$ but not any of $c_1$,$c_2$,$\ldots$,$c_{\lfloor t/2 \rfloor}$. This means that $A \cup B$ can contain at most an $x_{\lfloor t/2 \rfloor}$ fraction of voters, for otherwise candidate $c_i$ should have been selected earlier, on round $\lfloor t/2 \rfloor$. Thus, we may lower bound the fraction of voters satisfied on round $i$ as

$$\frac{1}{n}\left(|A \cap B|\right) = \frac{1}{n}\left(|A| + |B| - |A \cup B|\right) \geq \frac{1}{n}\left(\frac{n}{k} + \frac{n}{k} - n x_{\lfloor t/2 \rfloor}\right) = \frac{2}{k} - x_{\lfloor t/2 \rfloor}.$$

Summing over each of the first $k$ rounds, the number of satisfied voters is

$$\begin{aligned}\sum_{i=1}^{k}(\#\text{ satisfied voters on round } i) &\geq \sum_{i=1}^{t} x_i + \sum_{i=t+1}^{k}\left(\frac{2}{k} - x_{\lfloor t/2\rfloor}\right)\\ &= \sum_{i=1}^{t} x_i - \sum_{i=t+1}^{k} x_{\lfloor t/2\rfloor} + (k-t)\frac{2}{k}\\ &= \sum_{i=1}^{k-t}\left(x_i - x_{\lfloor t/2\rfloor}\right) + \sum_{i=k-t+1}^{k} x_i + (k-t)\frac{2}{k}.\end{aligned}$$

Since the $x_i$ are decreasing and $k - t \leq \lfloor t/2\rfloor$ when $t \geq 2k/3$, this first term is nonnegative, and $x_i \geq 1/k$ for $i \leq t$ in this case. Therefore we have

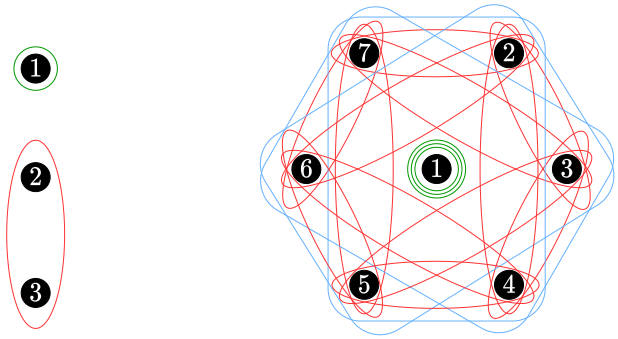
$$\begin{aligned}\sum_{i=1}^{k}(\#\text{ satisfied voters on round } i) &\geq \sum_{i=k-t+1}^{k} x_i + (k-t)\frac{2}{k}\\ &\geq (t-(k-t))\frac{1}{k} + (k-t)\frac{2}{k}\\ &= \frac{2(k-t)+(2t-k)}{k}\\ &= 1.\end{aligned}$$

Since all voters are satisfied by round $k$, the final committee satisfies JR. □

However, the exponent does have a dependence on $k$. In particular, we find that guaranteeing JR requires $\Omega(m^3)$ queries starting at $k = 6$. The adversary employs an analogous strategy, now picking 7 random candidates and imposing the approval matrix depicted in Figure 3.1(b). Satisfying JR requires that the algorithm include candidate 1, which is indistinguishable from the other six candidates unless the algorithm makes $\Omega(m^3)$ queries, since every candidate is approved by $\frac{6}{18}$ of the voters and every pair of candidates is approved by $\frac{2}{18}$ of the voters.

In Appendix B of the full version, we describe a computational search we conducted to find similar instances for larger values of $k$. The best lower bound obtained is as follows.

**Theorem 3.3.3.** *For any $\varepsilon > 0$, there exists a target committee size $k$ with $k = \Theta(\log 1/\varepsilon)$ such that for all $t$, any non-adaptive $(k,t)$-committee selection algorithm*




(a) $k = 2$ (b) $k = 6$

**Figure 3.1:** Adversarial approval matrices. Each region represents a disjoint, equally-sized set of voters who approve only the candidates within the region. In (a), queries of size $t \geq 2$ are needed to distinguish the candidates; in (b), we need $t \geq 3$.

*with exact queries that makes fewer than $\Omega(m^{11})$ queries satisfies JR with probability at most $\varepsilon$.*

This theorem closes the book on the (im)practicality of non-adaptive committee selection algorithms. We therefore turn our attention to adaptive algorithms.

### 3.3.2 An Efficient Adaptive Algorithm

In this section, we propose an adaptive algorithm based on LS-PAV [15], and we show that it achieves EJR and OAS with a practically-feasible number of queries.

For convenience, we introduce the following notation: For a committee $W$ and candidates $c \in W$ and $c' \notin W$, let

$$\Delta(W, c', c) := \text{PAV-SC}(W \cup \{c'\} \setminus \{c\}) - \text{PAV-SC}(W)$$

denote the difference in PAV score obtained by replacing $c$ with $c'$ in $W$. Additionally, let

$$\Delta(W, c) := \text{PAV-SC}(W \cup \{c\}) - \text{PAV-SC}(W)$$

denote the marginal gain in PAV score by adding $c$ to $W$.

LS-PAV starts with an arbitrary committee $W$ and repeatedly replaces a committee member $c \in W$ with a candidate $c' \notin W$, provided the improvement to the PAV

**Algorithm 2** $(k,t)$-$\alpha$-PAV

Choose $W \in \binom{C}{k}$, $c \in W$, and $c' \notin W$ arbitrarily
$\gamma \leftarrow \infty$
**while** $\gamma \geq \frac{1}{\alpha k}$ **do**
    $W \leftarrow W \cup \{c'\} \setminus \{c\}$
    Choose $\mathcal{Q} = \{Q_i\}_i$, with $|Q_i| = t$, s.t. $W \subseteq \bigcap \mathcal{Q}$ and $C \subseteq \bigcup \mathcal{Q}$
    $c' \leftarrow \arg\max_{x \notin W} \Delta(W, x)$ ▷ (using $\mathcal{Q}$)
    $c \leftarrow \arg\max_{x \in W} \Delta(W, c', x)$ ▷ (using $\mathcal{Q}$)
    $\gamma \leftarrow \Delta(W, c')$
**return** $W$

score satisfies $\Delta(W, c', c) \geq \frac{1}{k^2}$. Aziz et al. [15] show that after at most $\mathcal{O}(k^2 \log k)$ swaps, no such swap pairs $c, c'$ remain, at which point $W$ satisfies OAS and EJR.

We first observe that LS-PAV can be implemented using exact queries: For any set of candidates $S$, PAV-SC$(S)$ can be computed using any query $Q \supseteq S$, as it is sufficient to know the proportion of voters that approve each subset of $S$. Hence, for any $W$, $c \in W$, and $c' \notin W$, $\Delta(W, c', c)$ can be computed using a query $Q$ that contains both $W$ and $c'$. Using $\left\lceil \frac{m-k}{t-k} \right\rceil$ queries of size $t$, we can cover all $m - k$ candidates that are not in $W$, which leads to an overall query complexity of $\mathcal{O}\left(mk^2 \log k\right)$.

We next present a version of LS-PAV, which we call $\alpha$-PAV (Algorithm 2), that has the same query complexity as LS-PAV for finding a committee that satisfies EJR and OAS, but lower query complexity for approximate ($\alpha < 1$) $\alpha$-EJR and $\alpha$-OAS. Besides introducing the approximation parameter $\alpha$, we make two other modifications to LS-PAV: First, Algorithm 2 terminates when there is no candidate $c'$ such that $\Delta(W, c') \geq \frac{1}{k}$ (for $\alpha = 1$), while LS-PAV terminates when there is no pair $c, c'$ such that $\Delta(W, c', c) \geq \frac{1}{k^2}$. As we shall see in Lemma 3.3.6, the termination condition of Algorithm 2 is weaker than that of LS-PAV, implying that it may terminate earlier. Second, instead of considering all possible swaps $c, c'$, we only consider adding the candidate $c'$ with the largest $\Delta(W, c')$. This modification makes the algorithm slightly simpler and more computationally efficient (by a factor of $k$).

**Theorem 3.3.4.** *For any $m \geq t > k$, Algorithm 2 yields a committee satisfying*

*$\alpha$-OAS and $\alpha$-EJR while making at most*

$$\left\lceil \frac{m-k}{t-k} \right\rceil \frac{\alpha k^2}{(1-\alpha)k+1} H_k$$

*queries, where $H_k$ is the $k^{th}$ harmonic number. For $\alpha = 1$, this leads to a query complexity of $\mathcal{O}\left(mk^2 \log k\right)$ while for any fixed $\alpha < 1$, this leads to a query complexity of $\mathcal{O}\left(mk \log k\right)$.*

The proof of Theorem 3.3.4 essentially follows from the following two lemmas, the first of which uses the notation

$$\Delta^*(W) := \max_{c \in C} \Delta(W, c).$$

Proofs of the lemmas and the formal implication can be found in Appendix D of the full version.

**Lemma 3.3.5.** *If a committee $W$ satisfies $\Delta^*(W) < \frac{1}{\alpha k}$, then $W$ satisfies $\alpha$-EJR and $\alpha$-OAS.*

**Lemma 3.3.6.** *For any committee $W$ and $c \notin W$, we have that $\max_{x \in W} \Delta(W, c, x) \geq \frac{(k+1)\Delta(W,c)-1}{k}$. In particular, if $\Delta(W, c) \geq \frac{1}{\alpha k}$, then $\max_{x \in W} \Delta(W, c, x) \geq \frac{(1-\alpha)k+1}{\alpha k^2}$.*

Lemma 3.3.5 guarantees that when Algorithm 2 terminates the desired fairness properties are satisfied. Lemma 3.3.6 establishes that the PAV score increases over the algorithm's run. This bounds the number of swaps it performs since PAV-SC$(W)$ is at most $H_k$.

Lemma 3.3.5 is a generalization of the lower bound from Lemma 1 of Skowron [191]. This generalization is useful because it states that to establish EJR and OAS of any given committee $W$ (no matter how it is derived), it is sufficient to prove that $\Delta^*(W)$ is small; hence it can be used as a certificate of satisfaction. In Appendix E of the full version, we show that standard PAV and LS-PAV satisfy $\Delta^*(W) < \frac{n}{k}$, which is noteworthy in that it provides a simple proof of the known result that they satisfy EJR and OAS.

We observe that, for exact queries, an $\alpha$-approximation with $\alpha < 1$ improves the query complexity by a factor of $k$. In the next section, we will see that such an approximation yields an even larger improvement in query complexity for noisy queries, as it also reduces the accuracy with which we need to estimate $\Delta(W, c', c)$.

## 3.4 Noisy Queries

We now turn to a query model that includes the noise we abstracted away in Section 3.3. In order to represent voters arriving to the platform one-by-one, we assume that each time the algorithm performs a query $Q \subseteq C$ a voter $i \in N$ is selected independently and uniformly at random[6]and then the algorithm observes their votes on the queried candidates $Q \cap A_i$. We refer to an algorithm that performs queries of size $t$, receives as a response the votes of a single voter, and outputs a committee of size $k$ as a $(k, t)$-*committee selection algorithm with noisy queries*.

To see the connection between this query model and the previous one, note that an algorithm with noisy queries can approximate an exact query $Q$ by estimating the values of $p_S$ by taking the empirical proportion of repeated samples. By standard sample complexity bounds, using $\Theta\left(\log(2^t/\delta)/\varepsilon^2\right)$ queries, a noisy-query algorithm could guarantee $\pm\varepsilon$ estimates of $p_S$ for all $S \subseteq Q$ with probability $1 - \delta$. Hence if an exact-query algorithm requires no more than $\text{poly}(m)$ queries with additive $\varepsilon$ error, then it can be implemented using a factor of $\Theta(\log m)$ more noisy queries and yield a correct result with probability $1 - \delta$. What's more, this log factor is in some cases necessary when moving from the exact-query to the noisy-query setting. In Appendix C of the full version, we demonstrate instances for which a non-adaptive exact-query algorithm needs only $\Theta(m)$ queries, while in order to be correct with any fixed probability $\delta$, a non-adaptive noisy-query algorithm requires $\Omega(m \log m)$ queries.

Conversely, notice that one can use exact queries to simulate noisy queries. Indeed, $p_S$ is exactly the probability that an incoming voter will vote yes on candidates $S$ and no on $Q \setminus S$ in response to a query $Q$. An algorithm with access to exact query values can simply sample a voter response and feed it to a noisy-query algorithm. Therefore, the lower bounds on the query complexity of exact-query, non-adaptive algorithms, in particular Theorem 3.3.3, apply to noisy-query, non-adaptive algorithms as well. As the number of candidates becomes large, adaptivity is therefore necessary to attain

[6]Note that a voter-profile $A_i$ may be queried more than once during the run of the algorithm because we sample *with replacement*. This model simplifies the statistical analysis and has a natural interpretation: Rather than thinking of a finite population of voters, we draw samples from an underlying population distribution where each profile $A_1, .., A_n$ has the same frequency (probability). Furthermore, our model approaches sampling without replacement if the size of the underlying population $n$ is large compared to the number of queried voters, hence both models are qualitatively interchangeable.

**Algorithm 3** $(k,t)$-noisy-$\alpha$-PAV

$\ell \leftarrow \left\lceil 288 \left(\frac{\alpha k^2}{(1-\alpha)k+1}\right)^2 \log\left(\frac{8mk^4}{\delta}\right)\right\rceil$
Choose $W \in \binom{C}{k}$, $c \in W$, and $c' \notin W$ arbitrarily
$\gamma \leftarrow \infty$
**while** $\gamma \geq 1/(\alpha k) - ((1-\alpha)k+1)/(12\alpha k^2)$ **do**
    $W \leftarrow W \cup \{c'\} \setminus \{c\}$
    Choose $\mathcal{Q} = \{Q_i\}_i$, with $|Q_i| = t$, such that $W \subseteq \bigcap \mathcal{Q}$ and $C \subseteq \bigcup \mathcal{Q}$
    Ask each query $Q \in \mathcal{Q}$ to $\ell$ new voters
    $\hat{\Delta}(W,x) \leftarrow$ estimate of $\Delta(W,x)$ using $\ell$ voters from query $Q$ containing $W \cup \{x\}$ ▷ $\forall x \notin W$
    $\hat{\Delta}(W,x,y) \leftarrow$ estimate of $\Delta(W,x,y)$ using $\ell$ voters from $Q$ containing $W \cup \{x\}$ ▷ $\forall x \notin W, \forall y \in W$
    $c' \leftarrow \arg\max_{x \notin W} \hat{\Delta}(W,x)$
    $c \leftarrow \arg\max_{x \in W} \hat{\Delta}(W,c',x)$
    $\gamma \leftarrow \hat{\Delta}(W,c')$
**return** $W$

theoretical guarantees — mirroring the approach of online platforms in practice.

A natural starting point is the exact-query adaptive algorithm, namely Algorithm 2. Indeed, it can be adapted to the noisy setting by replacing exact queries with a sufficient number of noisy queries, $\ell$, to obtain high-probability bounds on $\Delta$, yielding Algorithm 3.

The key is to choose $\ell$ large enough that if the termination condition is not met, i.e., we have $\hat{\Delta}(W,c') < \frac{1}{\alpha k} - \frac{(1-\alpha)k+1}{12\alpha k^2}$, the resulting swap is guaranteed to yield a positive improvement in the PAV-score, such that the number of steps of the algorithm is bounded. With the choice of $\ell$ in Algorithm 3, we obtain the following theorem, whose proof can be found in Appendix F of the full version.

**Theorem 3.4.1.** *For any $m \geq t > k$, with probability at least $1-\delta$, Algorithm 3 returns a committee that satisfies $\alpha$-EJR and $\alpha$-OAS after querying no more than*

$$578 H_k \left\lceil \frac{m-k}{t-k} \right\rceil \left(\frac{\alpha k^2}{(1-\alpha)k+1}\right)^3 \log\left(\frac{4mk^4}{\delta}\right)$$

*voters. For any fixed $\delta > 0$, if $\alpha = 1$, this leads to a query complexity of $\mathcal{O}\left(mk^6 \log k \log m\right)$, and if $\alpha < 1$, this leads to a query complextiy of $\mathcal{O}\left(mk^3 \log k \log m\right)$.*

While Algorithm 3 achieves good worst-case query complexity, it may be suboptimal on certain instances because of two reasons: (i) after each swap, Algorithm 3 discards all previous information so each candidate is reassessed from scratch, and (ii) it presents each candidate $c \notin W$ to the same number of voters, even though it may quickly become apparent that some candidates are more promising than others.

To address issue (i), we can use all past votes to compute bounds on $\Delta$. A difficulty with this approach is that past voters may not have voted on all candidates in $W$ (which is necessary to directly estimate $\Delta(W, c)$), since they may have been queried on a different committee $W'$. But we can nonetheless use these past votes to obtain upper and lower bounds on estimated values. To address issue (ii), we can present promising candidates to voters more often. Further, it is possible to perform swaps as soon as we are confident they yield an increase of the PAV-score of at least some value $\varepsilon$, rather than first querying a predetermined number of voters as in Algorithm 3.

These ideas are incorporated into an algorithm called ucb-$\alpha$-PAV; see Appendix G of the full versionfor a formal description of the algorithm and an analysis of its query complexity.

### 3.5 Experiments

Since the analysis in the theoretical sections considers worst-case approval profiles, it is possible that, in practice, we may be able to find good committees with fewer queries than required by Theorem 3.4.1. We investigate this question empirically on real data from online discussions with only a few hundred voters, each voting on only a fraction of all comments.

Datasets. Polis provides open-use data from real deliberations hosted on their platform.[7] These include, for instance, a discussion organized by the government of Taiwan, which led to the successful regulation of Uber. Since participants typically only vote on a fraction of comments, most votes are missing. To be able to simulate

[7] `https://github.com/compdemocracy/openData`

the proposed adaptive algorithms, we first infer these missing votes using a matrix factorization library, LensKit.[8] Importantly, we infer votes only for the purpose of the experiments; if our algorithms were executed during the discussion, they would adaptively query users about the relevant comments without relying on any inference method.

In most datasets, we observe several comments that are nearly universally approved. Since these comments make achieving EJR and OAS trivial, we remove comments approved by more than 60% of participants. This step may also be appropriate in practice to gain insights into participants' opinions beyond uncontroversial issues.

The number of queried voters $L$ ranges from 87 to 1000 across the 13 datasets (see Appendix H of the full version for details). For all datasets, we assume that each voter votes on $t = 20$ comments. Since the total number of comments $m$ ranges from 31 to 1719 across datasets, the percentage of comments each voter votes on, $t/m$, ranges from 1% to 65%. For each dataset, we run the algorithms with target committee sizes $k = 5, 7, 10$. Hence, there are a total of $13 \cdot 3 = 39$ experiments (times 10 random seeds).

The second dataset we consider consists of Reddit discussions.[9] To obtain an interesting dataset, we combined voting data from two subreddits, r/politics and r/Conservative, which are arguably situated at opposite ends of the American political spectrum. More details about this dataset can also be found in Appendix H of the full version.

ALGORITHMS. We evaluate noisy-$\alpha$-PAV and ucb-$\alpha$-PAV. Both query $L$ voters in random order, each of whom votes on $t = 20$ comments. To enable these algorithms to swap candidates after querying only a small number of voters, we make the following modifications: For both algorithms we treat $\ell$, the number of times we ask voters about each candidate, as a parameter. In addition, for ucb-$\alpha$-PAV, we replace the numerator in the confidence intervals $err_s$ with a parameter $\theta$. Both $\ell$ and $\theta$ were chosen based on validation on a separate dataset, see Appendix H of the full version for details. We run both algorithms on all the $L$ voters, rather than terminating as soon as we can guarantee $\Delta^*(W) < \frac{1}{\alpha k}$ (and hence EJR and OAS).

To obtain an upper bound on the attainable performance, we execute $\alpha$-PAV with

[8] `https://lenskit.org`

[9] `https://www.kaggle.com/datasets/josephleake/huge-collection-of-reddit-votes`

access to exact queries. To obtain the best possible $\alpha$, we let $\alpha$-PAV run as long as the swap increases the PAV score, i.e., $\Delta(W, c', c) > 0$, instead of terminating as soon as $\Delta(W, c') < 1/k$ (which would be sufficient to guarantee EJR and OAS).

To verify that the proposed algorithms do indeed take the complementarity of different candidates into account, we also compare against standard approval voting (AV) with access to all votes, which simply selects the $k$ candidates with the most approval votes.

Performance Metric. As a performance metric, we use $\hat{\alpha} := \frac{1}{k\Delta^*(W)}$, where $W$ is the committee selected by the respective algorithm. According to Lemma 3.3.5, $\alpha > \hat{\alpha}$, so this implies $\hat{\alpha}$-EJR and $\hat{\alpha}$-OAS. As discussed in Section 3.2, $\alpha = 1$ is the best that can be guaranteed across all possible approval profiles. Note that $\alpha$ may be larger than $\hat{\alpha}$, hence obtaining $\hat{\alpha} = 1$ is a sufficient, but not a necessary condition for OAS and EJR. Nevertheless, we will use $\hat{\alpha}$ as a metric for two reasons: first, verifying whether $\alpha \geq 1$ (i.e., whether a committee satisfies EJR and OAS) is computationally hard [14], which makes it impractical for evaluation; and the stronger condition $\hat{\alpha} \geq 1$ provides the additional benefit that EJR and OAS can easily be verified through Lemma 3.3.5. Second, one could argue that $\hat{\alpha}$ is a meaningful quantity in its own right since it (or rather its inverse $1/\hat{\alpha}$) measures how much voter satisfaction could be improved by adding another candidate (giving lower weight to voters who already have many approved candidates).

Polis Results. In Figure 3.2, we show the $\hat{\alpha}$ achieved on all the Polis datasets for each of the four algorithms. Recall that higher $\hat{\alpha}$ is better and that $\hat{\alpha} \geq 1$ implies OAS and EJR. As expected, $\alpha$-PAV performs best since it has access to exact queries. Note that it often achieves an $\alpha$ substantially larger than 1, which means that the corresponding instance allows for better representation than can be guaranteed in the worst case. AV performs surprisingly well in most experiments, but in 38% of the cases, it yields $\hat{\alpha}$ smaller than 1 (and sometimes much smaller). We conclude that for some datasets, it is important to take the complementarity of candidates into account rather than selecting them individually. The challenge for the proposed algorithms is to do so while being sample-efficient. We see that noisy-$\alpha$-PAV often fails to achieve an $\hat{\alpha} \geq 1$. We know from Theorem 3.4.1 that given enough queries, noisy-$\alpha$-PAV achieves $\hat{\alpha} \geq 1$, so this failure is due to the low number of queried voters. By con-

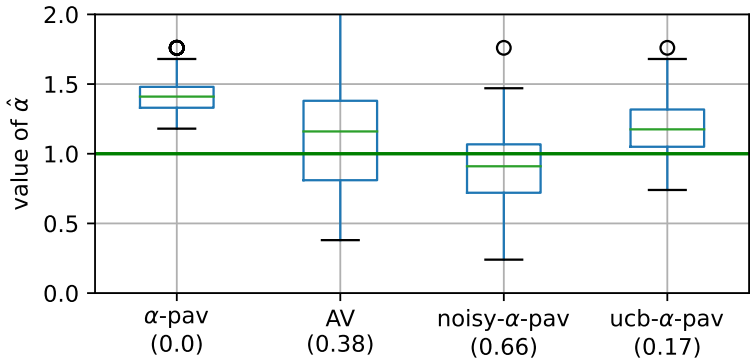



**Figure 3.2:** Boxplots where datapoints correspond to the 39 Polis problems ($\times 10$ random seeds). The top / bottom whiskers indicate the maximal / minimal points (except outliers, which are marked by circles), the line in the middle is the median, and the bottom and top of the boxes are the 1st and 3rd quartiles, respectively. The numbers in parenthesis are the fractions of problems where the respective algorithm yields a $\hat{\alpha} \leq 1$.

trast, ucb-$\alpha$-PAV yields $\hat{\alpha} \geq 1$ in 83% of the cases, and $\hat{\alpha} \geq 0.75$ in all cases, which indicates that the proposed extensions (i.e., querying promising candidates more often, swapping as soon as possible, and reusing voters) indeed lead to more efficient use of data.

Reddit Results. To illustrate why approval voting can perform poorly despite having access to the full votes, we execute the algorithms on the Reddit dataset described above. In this experiment, AV achieves only $\hat{\alpha} = 0.68$. To understand why this happens, we show in Figure 3.3 the fraction of voters who have at least $1, \ldots, 10$ approved comments in the committee. We see that AV yields a committee where a high fraction of voters approve many candidates, e.g., about 60% of voters approve 7 or more candidates, whereas for $\alpha$-PAV, this is the case for only about 40%. This comes at the cost of a high fraction of voters who are poorly represented by AV, e.g., about 25% of voters get at most one approved candidate, whereas for $\alpha$-PAV, this percentage is less than 10%. This is to be expected as approval voting does not take the complementarity of candidates into account and can therefore lead to less equitable results. Finally, we observe that ucb-$\alpha$-PAV achieves an $\hat{\alpha}$ close to 1, and its approval fractions look similar to $\alpha$-PAV, i.e., more equitable than AV. It is interesting that ucb-$\alpha$-PAV performs well on this example, since it only has access to $t = 20$ votes for

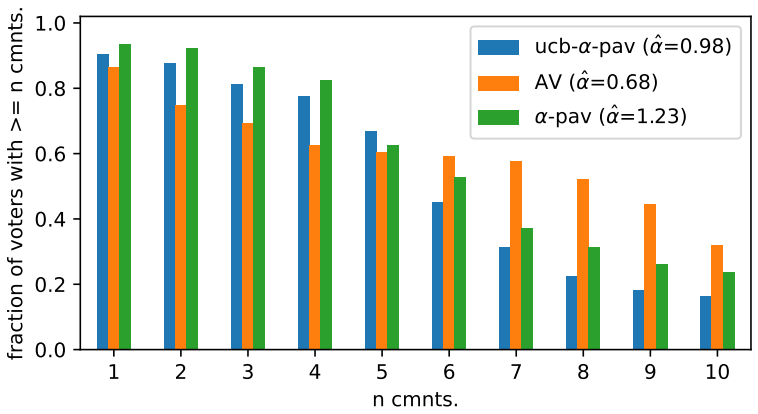



**Figure 3.3:** Results on Reddit dataset (with $L = 608, m = 2135, k = 10$): the fraction of voters ($y$-axis) that approve of at least $1, 2, .., 10$ candidates ($x$-axis) among the selected committee of size $k = 10$.

each of the $L = 608$ queried candidates, while it has to select from a large number of comments, $m = 2135$.

## 3.6 Discussion

This work bridges the gap between online civic-participation systems, such as Polis, and committee-election methods by enabling them to handle incomplete votes. To deploy the proposed algorithms on such platforms, two practical issues must be considered.

First, our adaptive approach requires control over what the Polis creators call *comment routing* [193]: the algorithm that decides which comments are shown to which participants. If on a given platform a comment-routing algorithm is already in place, shared control is possible: each algorithm could determine part of the slate of comments shown to a participant, or the participants themselves can be divided between the algorithms.

Second, in our analysis, we assumed that all comments have been submitted — or all candidates are known — at the time we run our algorithms. Nevertheless, our algorithms can be extended straightforwardly to a growing set of comments, but we would inevitably lose the representation guarantees for comments that were submitted late if not enough participants could vote on them. In practice, this could be resolved by

setting a comment submission deadline, which has been done previously by Polis.

An alternative to our approach would be to complete partial approval votes using collaborative filtering [171]. The completed approval votes can then be aggregated through any approval-based committee election rule, such as PAV. The disadvantage of this approach is that it is unlikely to lead to worst-case guarantees of the type we establish in this paper.

Finally, we emphasize that our approach may be applicable to social media more generally. For instance, as mentioned in Section 3.5, Reddit users also approve or disapprove comments through upvotes and downvotes. However, Reddit uses these inputs to produce a *ranking* of the comments, in contrast to our goal of selecting a subset. There is work on obtaining justified-representation-type guarantees for rankings [190], which could possibly be extended to the setting of incomplete votes using the techniques developed in this paper. More broadly, this article provides insights into how to fairly represent opinions of groups given incomplete information, which may be relevant for the design of more constructive online ecosystems.

# 4

# Computing Voting Rules with Incomplete Votes

## 4.1 Introduction

While Chapter 3 focused on opinion aggregation, the difficulty of expressing preferences over a large set of alternatives extends well beyond that context. Traditional social choice frameworks typically assume that voting rules have access to each voter's complete ordinal preferences over all candidates. Indeed, this is seen in practice as well with the widening adoption of ranked-choice voting systems [70], requiring voters to submit such information. Whether such information can actually be reliably elicited depends significantly on the context. If the number of candidates is small and voters have strong opinions, it may indeed be reasonable for them to provide a complete ranking. However, these assumptions do not hold in many scenarios. Primary elections in the United States, for example, routinely field large numbers of candidates, with many being unfamiliar to voters [104].

A classic line of work in behavioral economics and psychology supports the premise that individuals struggle in such scenarios. In his seminal work, Schwartz [181] puts forth the *paradox of choice*: individuals incur increased anxiety when faced with too many alternatives, which often leads them to take a default action, defer, or not par-

ticipate altogether [111]. Recent literature has shown this phenomenon to hold specifically in the voting and social choice setting. Cunow et al. [58, 59] experimentally show that even increasing the number of candidates from 3 to 6 leads voters to spend less effort learning candidates' policy positions and instead rely on arbitrary heuristics. This voter frustration is also evidenced in practice: incomplete ballots are quite common, which are often completely exhausted under elimination voting long before the final candidate is elected [38].

In Chapter 3, specifically for opinion aggregation, we asked the questions: what meaningful conclusions can be drawn from querying users over such limited data? And how should the platform select such queries? This was done in the context of approval votes, where the goal is to select a representative "committee" of size $k$. Voters from the population arrive randomly and can be presented with at most $t < m$ candidates (opinions) at a time, over which they can express their approval or disapproval.

Our paper extends this framework to consider ordinal preferences, with the more classic goal of selecting a single winner (rather than a committee) under a given rule. There is a distribution over rankings of the $m$ candidates representing the underlying voter preferences that is unknown to the algorithm. It may, however, *query* randomly arriving individuals about any subset of $t$ candidates, where $t < m$; with suitable samples, the algorithm can determine the corresponding ranking distribution over this subset.[1] This leads to the fundamental question: what is the set of voting rules that are implementable with such small-sized queries chosen by the platform? Conceptually, this question asks about the axiomatic implications of cognitive barriers to preference elicitation in social choice.

### 4.1.1 Contributions

We begin our investigation with the ubiquitous plurality rule. Here, we find a surprisingly negative result: determining a plurality winner cannot be done using queries of any size $t < m$ (Theorem 4.4.3). That is, even if one has access to the distribution of the population's rankings over every $m - 1$ subset of the $m$ candidates, it is still impossible to correctly identify who received the most first-place votes. In fact, even

[1] Clearly, having access to the exact distribution is strictly more informative than having to approximate it through repeated samples. Our negative results hold for this idealized setting and thus immediately apply to the weaker and more realistic query model.

a randomized algorithm can only correctly choose a plurality winner with probability $\frac{1}{m}$ in the worst case, i.e., there are instances where, no matter which queries an algorithm makes, it can do no better than picking a candidate uniformly at random. The proof follows from a novel construction of pairs of profiles that have different plurality winners but induce the same distribution on any subset of $m-1$ candidates. Section 4.3 is dedicated to explaining this construction, which forms the basis for all of the impossibility results in this paper.

Plurality is just one example of a *positional scoring rule*, whereby candidates receive points corresponding to their rank position in each ballot, with the winner being the candidate with the most aggregated points [209]. While plurality requires queries of size $m$, it is known that another positional scoring rule, the Borda count, only requires pairwise margins to determine a winner and, hence, can be computed with queries of size 2. For general $t$-sized queries, one straightforward algorithm is to query every subset of size $t$ and give each candidate a certain number of points depending on which of the $t$ positions they appear. Under a different query model, Bentert and Skowron [27] give a family of scoring rules for each size $t$ for which this algorithm works.[2] A similar argument for the same families of rules holds for our model as well, which we prove in Lemma 4.4.4 for completeness, along with a more detailed comparison between our work and theirs. We then proceed to our second main result (Theorem 4.4.5) that shows that these rules are, in fact, the *only* ones that can be implemented with queries of size $t$. We thus obtain a complete characterization of all computable positional scoring rules under a limited query model and also give visualizations of this space. Our analysis is then extended to the *single transferrable vote (STV)* rule (also known as *instant-runoff voting*) wherein we find a negative result akin to Plurality: algorithms with limited query size ($t < m$) cannot correctly implement this rule. All of these negative results again hold not only for deterministic algorithms but for randomized ones that are correct with probability strictly better than $\frac{1}{m}$ (i.e. better than randomly guessing).

We next use our characterization to study the rules that *are* computable with small-sized queries. Specifically, for any scoring rule requiring $t^*$-sized queries, we lower bound the number of $t$-sized queries ($t^* \leq t < m$) needed to compute the winner under this rule (Theorem 4.5.1). If $t$ and $t^*$ are treated as constants, then $\Theta(m^{t^*})$ queries are needed in the worst case. This asymptotic bound holds even for random-

[2] In their model, each voter submits a ranking over a *randomly* selected set of $t$ candidates.

ized algorithms that are correct with probability $\frac{1}{m} + \varepsilon$ for any constant $\varepsilon > 0$. If on the other hand an algorithm covers a $\delta$ fraction of all $t^*$-subsets, we give an upper bound on the success probability. Although this is not tight, we exactly determine the optimal success probability for Borda count with $m = 3$ (Theorem 4.5.2) with a surprisingly intricate construction, and leave open the general case.

### 4.1.2 Related Work

Our work contributes to a line of literature on the information-theoretic aspects of voting and elections, specifically on what can be accomplished with incomplete information. There are a variety of lenses through which to study this problem. One line of work considers the communication complexity of computing various voting rules [54]. Another studies when using incomplete votes can guarantee a candidate must or cannot be the winner, regardless of the missing information [130, 206]. Others consider different "approximation" objectives such as minimax regret [139] and distortion [168]. For a more complete survey, see chapter 10 by Brandt et al. [36].

More specific to our work are models where the partial information given is $t$-wise comparisons. The special case of $t = 2$ corresponds to only being given pairwise comparisons. All information about pairwise comparisons can be summarized in a *weighted tournament graph*, a widely studied object in social choice theory [36]. For example, there is a classification of common voting rules into those that can be computed using just the tournament graph (such as *Borda Count*, *Minimax*, *Kemeny*, and *Copeland*) and those that cannot [75]. Any of these weighted tournament solutions can trivially be computed with queries of size $t = 2$. Beyond information-theoretic results, there are even bounds on query complexity, such as how many queries to the tournament graph are needed to compute Condorcet winners [167]. Our paper is a natural extension of this literature to the realm of more powerful $t$-wise comparison queries for $t > 3$.

Related, but technically incomparable, is when voters reveal a ranking of their top $t$ candidates for a fixed value of $t$ [73, 156]. This was one of the two models studied by Bentert and Skowron [27]. The other has voters revealing a ranking over a *random* set of $t$ candidates. Positive results here translate to our model, although negative ones do not. We describe this connection and their results more in-depth in Section 4.4.1.

Finally, note that we largely study information-theoretic impossibilities and thus

do not focus on the randomized arrivals aspect; other works do consider the sample complexity of computing various rules [61]; however, they assume randomly sampled voters reveal their complete preferences over all candidates.

## 4.2 Preliminaries

### 4.2.1 Voter preferences

For a positive integer $s$, let $[s] := \{1, \ldots, s\}$. A *ranking* or *preference* over a set $C$ of $m$ *candidates* is a bijection $\sigma : [m] \to C$, where $\sigma(j)$ represents the $j$'th most-preferred candidate according to $\sigma$. We use the standard notation $a \succ_\sigma b$ to denote that $a$ is preferred to $b$ under $\sigma$, i.e., $\sigma^{-1}(a) < \sigma^{-1}(b)$, where $\sigma^{-1}$ is the inverse mapping from candidates to rankings. We write $\mathcal{L}(C)$ to denote the set of all $m!$ rankings over the candidates in $C$. For a subset of candidates $S \subseteq C$, we write $\sigma|_S$ to denote the ranking $\sigma$ restricted to the candidates in $S$, i.e., $\sigma|_S \in \mathcal{L}(S)$, with $\sigma|_S(j)$ being the $j$'th most preferred among those in $S$ according to $\sigma$.

For a permutation over the candidates $\pi : C \to C$, we will write $\pi \circ \sigma$ for the ranking $\sigma$ permuted by $\pi$, i.e., $(\pi \circ \sigma)(j) = \pi(\sigma(j))$. Of particular interest will be permutations that swap a single pair of candidates. For this reason, for two candidates $a$ and $b$, we define $\pi^{ab}$ to be the $(a, b)$-*transposition*, the permutation that swaps $a$ and $b$, i.e., $\pi^{ab}(a) = b$, $\pi^{ab}(b) = a$, and $\pi^{ab}(c) = c$ for all $c \neq a, b$. Further, we will write $\sigma^{a \leftrightarrow b}$ for $\pi^{ab} \circ \sigma$.

A *preference profile* (or simply a *profile*) is a distribution $\boldsymbol{\sigma}$ over preferences $\mathcal{L}(C)$, representing the proportion of voters in the population that have each ranking. For example,

$$\Pr_{\sigma \sim \boldsymbol{\sigma}}[\sigma = a \succ b \succ c] = \frac{1}{5}$$

means that $\frac{1}{5}$ of the voters have the ranking $a \succ b \succ c$.[3] We denote by $\Pi(C)$ the space of all possible preference profiles over a candidate set $C$. For a profile $\boldsymbol{\sigma} \in \Pi(C)$, $\sigma \sim \boldsymbol{\sigma}$ denotes sampling a preference $\sigma$ from distribution $\boldsymbol{\sigma}$. For a set of rankings $R \subseteq \mathcal{L}(C)$, we will also use the notation $\sigma \sim \text{Unif}(R)$ to denote sampling a ranking

[3] More often in social choice, a profile is a ranking assignment for a finite number of $n$ agents. The distributional definition is essentially equivalent insofar as voter identity is not important (as is the case for all rules we study) while having the additional benefit of making our query model and proof techniques easier to understand. However, we could have equivalently used the more traditional definitions, and all of our results would still hold.

$\sigma$ uniformly from $R$. We extend the restriction $\sigma|_S$, permutation $\pi \circ \sigma$, and transposition $\sigma^{a \leftrightarrow b}$ operations from rankings to profiles in the natural way. More formally, for restrictions, we write $\boldsymbol{\sigma}|_S$ to denote the distribution $\boldsymbol{\sigma}$ when restricted to candidates in $S$, i.e., $\boldsymbol{\sigma}|_S$ is an element of $\Pi(S)$ induced by sampling $\sigma \sim \boldsymbol{\sigma}$ and outputting $\sigma|_S$. For permutations, $\pi \circ \boldsymbol{\sigma}$ is the distribution induced by sampling $\sigma \sim \boldsymbol{\sigma}$ and outputting $\pi \circ \boldsymbol{\sigma}$. For transpositions, $\boldsymbol{\sigma}^{a \leftrightarrow b} = \pi^{ab} \circ \boldsymbol{\sigma}$. For a set $S$, we also use the notation $\mathrm{Unif}(S)$ to denote the uniform distribution over elements of $S$.

### 4.2.2 Voting rules

A *voting rule* $f$ maps preference profiles to a set of winning candidates. If a candidate $c$ is amongst the winners for a voting rule $f$, we refer to this candidate as an $f$-winner.

We will primarily focus on *positional scoring rules* (or simply *scoring rules*), which are a very practical and well-studied class. These are parameterized by a *scoring vector* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$.[4] Intuitively, a voter with ranking $\sigma$ gives $\alpha_1$ points to their first place candidate $\sigma(1)$, $\alpha_2$ points to their second place candidate $\sigma(2)$, and so on, with the winner of the profile being the candidate with the most aggregated points. Written in our distributional notation, the score of a candidate $c$ on profile $\boldsymbol{\sigma}$ is $\mathrm{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(c) := \mathbb{E}_{\sigma \sim \boldsymbol{\sigma}}[\alpha_{\sigma^{-1}(c)}]$, and the winning candidates on a profile $\boldsymbol{\sigma}$ are those with maximal score. Some common scoring rules include *plurality*, parameterized by $(1, 0, \ldots, 0)$, *veto*, parameterized by $(0, \ldots, 0, -1)$, and *Borda count*, parameterized by $(m-1, m-2, \ldots, 0)$. Since the plurality score will come up quite frequently, we will write $\mathrm{plu}_{\boldsymbol{\sigma}}(c)$ instead of using the $\mathrm{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(c)$ notation, and note that this simplifies to $\mathrm{plu}_{\boldsymbol{\sigma}}(c) = \Pr_{\sigma \sim \boldsymbol{\sigma}}[\sigma(1) = c]$. For conciseness, we will use $\boldsymbol{\alpha}$-winner, plurality winner, veto winner, and Borda winner to refer to $f$-winners of the scoring rule induced by $\boldsymbol{\alpha}$, plurality, veto, and Borda count, respectively. Note that $\boldsymbol{\alpha}$-winners are invariant under modifying $\boldsymbol{\alpha}$ via translation or multiplication by a positive constant, e.g., veto-winners (with vector $(0, \ldots, 0, -1)$) coincide with both $(1, \ldots, 1, 0)$-winners and $(\frac{1}{m-1}, \ldots, \frac{1}{m-1}, 0)$-winners.

In addition to scoring rules, we will also consider the rule *Single Transferable Vote (STV)*, which is defined as follows. For a profile $\boldsymbol{\sigma}$, if there is a single candidate, it

[4] Often, scoring vectors are restricted to be nonnegative and nonincreasing, but, for our purposes, it will be more convenient to allow for arbitrary vectors.

returns that candidate. Otherwise, it chooses a candidate $c$ with *minimal* plurality score, deletes them from the profile, and recurses on the rest. More formally, it chooses $c \in \arg\min_{c'} \text{plu}_{\boldsymbol{\sigma}}(c')$, and runs STV on $\boldsymbol{\sigma}|_{C\setminus\{c\}}$. This must eventually terminate, as a candidate is removed at each iteration. Further, for $m = 2$, this coincides with plurality. Note that in some cases, there are ties for the minimal score candidate. Hence, we will say that $c$ is an STV winner if some sequence of valid eliminations results in $c$ being the winner.

### 4.2.3 Query model

We consider (possibly randomized) algorithms that are allowed to adaptively submit queries to the underlying distribution $\boldsymbol{\sigma}$. For a fixed parameter $t$, each query consists of a subset of candidates $Q \subseteq C$ with $|Q| \leq t$ (referred to as a *query of size* $t$, or a *$t$-query* for short) and returns the distribution $\boldsymbol{\sigma}|_Q$. As mentioned previously, having access to the exact distribution is strictly more informative than one approximated by a random voter arriving (a voter $\sigma \sim \boldsymbol{\sigma}$ arrives, and the algorithm learns $\sigma|_Q$). All our impossibility results hold in this idealized setting and thus immediately apply to the more realistic one.

Two profiles $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$ are said to be *$t$-indistinguishable* if for all subsets $Q \subseteq C$ with $|Q| \leq t$, $\boldsymbol{\sigma}^1|_Q = \boldsymbol{\sigma}^2|_Q$. That is, regardless of whether the profile is $\boldsymbol{\sigma}^1$ or $\boldsymbol{\sigma}^2$, any query $Q$ with $|Q| \leq t$ will have the same response. Importantly, if $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$ are $t$-indistinguishable, but $f(\boldsymbol{\sigma}^1) \cap f(\boldsymbol{\sigma}^2) = \emptyset$, then no $t$-query algorithm can always output an $f$-winner. Note that to check whether two profiles are $t$-indistinguishable, it suffices to check only queries $Q$ of size exactly $t$, as if $Q' \subseteq Q$, then $\boldsymbol{\sigma}|_{Q'} = (\boldsymbol{\sigma}|_Q)|_{Q'}$, so $\boldsymbol{\sigma}^1|_Q = \boldsymbol{\sigma}^2|_Q$ implies $\boldsymbol{\sigma}^1|_{Q'} = \boldsymbol{\sigma}^2|_{Q'}$. Finally, notice that the special case of 2-indistinguishable is equivalent to $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$ having the same *weighted tournament graph*, the complete-directed graph, where the nodes are candidates, and the weight on edge $(a, b)$ is the proportion of voters that prefer $a$ to $b$ [36]. In this sense, the collection of distributions $\{\boldsymbol{\sigma}|_Q\}_{Q\subseteq C:|Q|=t}$ are a generalization of the weighted tournament graph to arbitrary $t \geq 2$ (e.g., for $t = 3$, rather than consisting of proportions of people that prefer $a$ to $b$, it consists of the proportions of people that prefer $a$ to $b$ to $c$ for all such combinations).

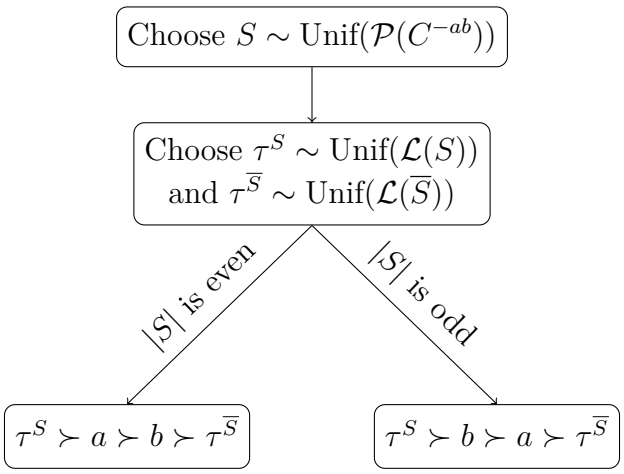



**Figure 4.1:** A process inducing the distribution over rankings for $\boldsymbol{\sigma}$.

## 4.3 An Indistinguishable Construction

We begin with a construction of two indistinguishable profiles which will be used throughout our technical results.

**Lemma 4.3.1.** *For any $m$ and pair of candidates $a, b \in C$, there is a profile $\boldsymbol{\sigma}$ such that (i) $\text{plu}_{\boldsymbol{\sigma}}(a) \neq \text{plu}_{\boldsymbol{\sigma}}(b)$ and (ii) $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^{a \leftrightarrow b}$ are $(m-1)$-indistinguishable.*

*Proof.* Fix $m$, $a$, and $b$. Let $C^{-ab} = C \setminus \{a, b\}$ be the set of remaining candidates. We have that $|C^{-ab}| = m - 2$. We define $\boldsymbol{\sigma}$ to be the distribution induced by the following random process. First, pick a set $S \subseteq C^{-ab}$ uniformly at random (i.e., each of the $2^{m-2}$ sets with equal probability). Formally, let $\mathcal{P}(C^{-ab})$ denote the power set of $C^{-ab}$, and we will sample $S \sim \text{Unif}(\mathcal{P}(C^{-ab}))$. Then choose a uniformly random ranking $\tau^S$ of the candidates in $S$ and a uniformly random ranking $\tau^{\overline{S}}$ of the candidates in $\overline{S} := C^{-ab} \setminus S$. Finally, if $|S|$ is even, output $\tau^S \succ a \succ b \succ \tau^{\overline{S}}$ and if $|S|$ is odd, output $\tau^S \succ b \succ a \succ \tau^{\overline{S}}$. This process is visually represented in Figure 4.1.

First, observe that it is indeed the case that $\text{plu}_{\boldsymbol{\sigma}}(a) \neq \text{plu}_{\boldsymbol{\sigma}}(b)$. The only way that either $a$ or $b$ could be ranked first is if the set $S$ is empty. In that case, $|S|$ is even, so $a$ will be ranked first. This does happen with positive probability ($1/2^{m-2}$), so the plurality score of $a$ is positive. However, this can never happen for candidate $b$.

Next, we show that $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^{a \leftrightarrow b}$ are $(m-1)$-indistinguishable. Note that a ranking from $\boldsymbol{\sigma}^{a \leftrightarrow b}$ can be sampled by running the process for $\boldsymbol{\sigma}$ but swapping the outcomes

of the "$|S|$ is even" and "$|S|$ is odd" branches.

Fix any $Q \subseteq C$ with $|Q| = m - 1$ and fix some ranking $\tau \in \mathcal{L}(Q)$. We want to show that observing $\tau$ is equiprobable under both $\boldsymbol{\sigma}|_Q$ and $\boldsymbol{\sigma}^{a \leftrightarrow b}|_Q$.

First, suppose $Q$ contains only one of $a$ or $b$. Without loss of generality, suppose it contains $a$. Since $Q$ does not contain $b$, if we run the process of Figure 4.1 and pick $\tau^S$ and $\tau^{\overline{S}}$, regardless of whether we follow the "$|S|$ is even" or "$|S|$ is odd" branch the output when restricted to $Q$ is the same. This is due to $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^{a \leftrightarrow b}$ differing only in which branch to follow, and both branches are identical apart from the ordering of $a$ and $b$, which occur consecutively in both. Thus, outputting $\tau$ is equiprobable in both restricted profiles.

Next, suppose $Q$ contains both $a$ and $b$. Again, since in any ranking of $\boldsymbol{\sigma}$ or $\boldsymbol{\sigma}^{a \leftrightarrow b}$, $a$ and $b$ always appear adjacent to each other, if $a$ and $b$ are not adjacent in $\tau$, it occurs with probability 0 in both profiles. As such, suppose they are adjacent in $\tau$, and without loss of generality, let $a \succ_\tau b$ (the other case is symmetric). Let $T$ be the set of candidates ranked above $a$ in $\tau$, and $L$ be the set of candidates ranked below $b$. Therefore, $Q = T \cup L \cup \{a, b\}$. Note to sample $\sigma$ with $\sigma|_Q = \tau$ under either $\boldsymbol{\sigma}$ or $\boldsymbol{\sigma}^{a \leftrightarrow b}$, it must be the case that when selecting $S$, $S \cap Q = T$. Further, conditioned on this, the probability of getting both $T$ and $L$ in the order matching $\tau$ is simply $\frac{1}{|T|!|L|!}$. Finally, the order of $a$ and $b$ will match $\tau$ exactly when $|S|$ is even. Putting this together, we have that the probability of sampling $\sigma$ with $\sigma|_Q = \tau$ under $\boldsymbol{\sigma}$ is exactly

$$\frac{1}{|T|!|L|!} \Pr[(S \cap Q = T) \wedge (|S| \text{ is even})].$$

For $\boldsymbol{\sigma}^{a \leftrightarrow b}$, it is identical but with "even" switched with "odd." Hence, to show equality, it suffices to show that:

$$\Pr[(S \cap Q = T) \wedge (|S| \text{ is even})] = \Pr[(S \cap Q = T) \wedge (|S| \text{ is odd})].$$

This is equivalent to showing that

$$\Pr[|S| \text{ is even} \,|\, S \cap Q = T] = \Pr[|S| \text{ is odd} \,|\, S \cap Q = T].$$

Let us now consider how to sample $S$ from the conditional distribution given $S \cap Q = T$. Note that $S$ satisfies this exactly when $T \subseteq S$ and $L \cap S \neq \emptyset$. Hence, to sample such an $S$, we can sample $S'$ uniformly from $C^{-ab} \setminus (T \cup L)$ and output $S' \cup T$. By

the assumptions that $|Q| = m - 1 < |C|$ and $\{a, b\} \subseteq Q$, we have that $C^{-ab} \setminus (T \cup L)$ is nonempty. It is known that when sampling a subset uniformly at random from a non-empty set, it is equiprobable whether it is of even or odd carnality.[5] Hence, $|S'|$ is equally likely to be even or odd, which implies that $|S|$ conditioned on $S \cap Q = T$ is equally likely to be even or odd, as needed. □

## 4.4 Uncomputable Voting Rules

In this section, we prove that there is a family of voting rules that *cannot* be computed using limited query sizes. In particular, we give a complete characterization of which positional scoring rules can be computed in our query model for each choice of $t$. In addition, we analyze the widely adopted STV rule. We begin, however, with two lemmas, which together give sufficient conditions for a scoring rule to *not* be computable using limited queries. The second will also be helpful for the STV impossibility.

**Lemma 4.4.1.** *Fix a vector $\boldsymbol{\alpha}$, and suppose there exists a profile $\boldsymbol{\sigma}$ and two candidates $a$ and $b$ such that* $\text{score}_{\boldsymbol{\sigma}}^{\boldsymbol{\alpha}}(a) \neq \text{score}_{\boldsymbol{\sigma}}^{\boldsymbol{\alpha}}(b)$ *yet $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^{a \leftrightarrow b}$ are $t$-indistinguishable. Then, there exists a family of profiles, $\{\boldsymbol{\sigma}^c\}_{c \in C}$ such that all are $t$-indistinguishable from one another, but each candidate $c \in C$ uniquely maximizes* $\text{score}_{\boldsymbol{\sigma}^c}^{\boldsymbol{\alpha}}$.

*Proof.* Let $\boldsymbol{\sigma}$, $a$, and $b$ be the profile and candidates satisfying the lemma conditions. Fix a candidate $c$. We describe the distribution $\boldsymbol{\sigma}^c$ as follows. Assume without loss of generality that $\text{score}_{\boldsymbol{\sigma}}^{\boldsymbol{\alpha}}(a) > \text{score}_{\boldsymbol{\sigma}}^{\boldsymbol{\alpha}}(b)$. We first sample a permutation over the candidates $\pi$ uniformly at random. If $\pi(b) = c$, we return a ranking sampled from $\pi \circ \boldsymbol{\sigma}^{a \leftrightarrow b}$; otherwise, we return a ranking sampled from $\pi \circ \boldsymbol{\sigma}$. A visual representation can be found in Figure 4.2. We will compare this constructed profile with respect to another, which we call $\boldsymbol{\sigma}^{\text{unif}}$. The profile $\boldsymbol{\sigma}^{\text{unif}}$ is constructed by picking a permutation $\pi$ uniformly at random and returning a sample $\sigma \sim \pi \circ \boldsymbol{\sigma}$ regardless of $\pi$, shown in Figure 4.3.

We first claim that $\boldsymbol{\sigma}^{\text{unif}}$ is in fact the uniform distribution over $\mathcal{L}(C)$. Indeed, an equivalent way of sampling from $\boldsymbol{\sigma}^{\text{unif}}$ is to sample in the reverse order, first sampling

[5] Fix an element $x$. For any subset $S$ of the remaining elements, it is equally likely to pick $S$ and $S \cup \{x\}$. One of these has even parity and the other has odd.

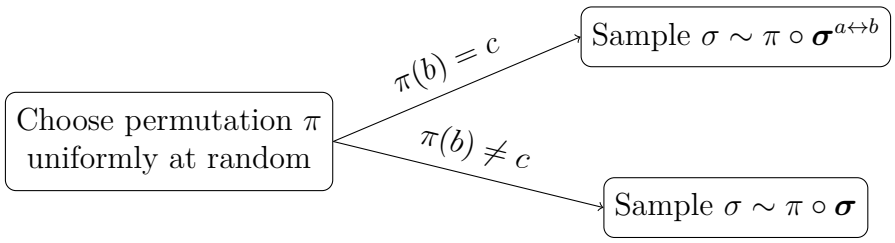



**Figure 4.2:** A process inducing the distribution over rankings for $\boldsymbol{\sigma}^c$.

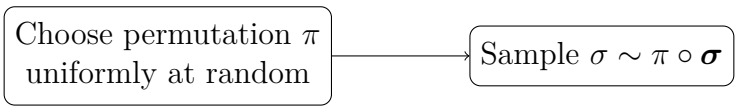



**Figure 4.3:** A process inducing the distribution over rankings for $\boldsymbol{\sigma}^{\text{unif}}$.

$\sigma \sim \boldsymbol{\sigma}$ and then applying a randomly selected permutation $\pi$ to $\sigma$. This makes it clear that $\boldsymbol{\sigma}^{\text{unif}}$ is a mixture over uniform distributions and is hence uniform.

We can think of both $\boldsymbol{\sigma}^c$ and $\boldsymbol{\sigma}^{\text{unif}}$ as a mixture of $m!$ different profiles, the one associated with each choice of $\pi$. Abusing notation slightly, we will write $\boldsymbol{\sigma}^c_\pi$ and $\boldsymbol{\sigma}^{\text{unif}}_\pi$ for the profile sampled when we had the permutation $\pi$. More concretely,

$$\boldsymbol{\sigma}^c_\pi = \begin{cases} \pi \circ \boldsymbol{\sigma} & \text{if } \pi(b) \neq c \\ \pi \circ \boldsymbol{\sigma}^{a \leftrightarrow b} & \text{if } \pi(b) = c \end{cases},$$

while $\boldsymbol{\sigma}^{\text{unif}}_\pi = \pi \circ \boldsymbol{\sigma}$ for all $\pi$.

Note that the scores $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c')$ are equal for all $c'$ by symmetry. We will show both (i) $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c}(c) > \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c)$ while $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c}(c') \leq \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c')$ for all $c' \neq c$ and (ii) $\boldsymbol{\sigma}^c$ is $t$-indistinguishable from $\boldsymbol{\sigma}^{\text{unif}}$. The first shows that the $c$ is the unique score maximizer, and the second shows that all the constructed profiles are $t$-indistinguishable from each other since each is $t$-indistinguishable from $\boldsymbol{\sigma}^{\text{unif}}$.

For the first, consider the difference $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c}(c') - \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c')$ for an arbitrary candidate $c'$. Because scores are linear, we can split them across our mixture definitions to get

$$\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c}(c') = \mathbb{E}_\pi[\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c_\pi}(c')] \text{ and } \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c') = \mathbb{E}_\pi[\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}_\pi}(c')].$$

Plugging this into the difference, by linearity of expectation, we have that

$$\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c}(c') - \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c') = \mathbb{E}_\pi[\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c_\pi}(c') - \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}_\pi}(c')].$$

Now, for any $\pi$ with $\pi(b) \neq c$, the difference inside the expectation is 0 because $\boldsymbol{\sigma}^c_\pi = \boldsymbol{\sigma}^{\text{unif}}_\pi$. Fix some $\pi$ such that $\pi(b) = c$. In this case, $\boldsymbol{\sigma}^c_\pi = \pi \circ \boldsymbol{\sigma}^{a\leftrightarrow b}$ while $\boldsymbol{\sigma}^{\text{unif}}_\pi = \pi \circ \boldsymbol{\sigma}$. Note that

$$\text{score}^{\boldsymbol{\alpha}}_{\pi\circ\boldsymbol{\sigma}^{a\leftrightarrow b}}(c') = \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{a\leftrightarrow b}}(\pi^{-1}(c')) = \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(\pi^{ab}(\pi^{-1}(c')))$$

where the first equality moves the application of $\pi$ and the second does the same, using the fact that $\boldsymbol{\sigma}^{a\leftrightarrow b}$ is simply achieved by the $\pi^{ab}$ permutation (which is its own inverse). We also have:

$$\text{score}^{\boldsymbol{\alpha}}_{\pi\circ\boldsymbol{\sigma}}(c') = \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(\pi^{-1}(c)).$$

Hence, this difference is only nonzero if $\pi^{ab}(\pi^{-1}(c')) \neq \pi^{-1}(c')$, i.e., $\pi^{-1}(c') \in \{a, b\}$. If $\pi^{-1}(c') = a$, then the difference is $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(b) - \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(a)$, while if $\pi^{-1}(c') = b$, then the difference is $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(a) - \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(b)$. By the lemma assumptions, $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(a) > \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(b)$; hence, the difference is positive in the first case and negative in the second. To summarize, $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c_\pi}(c') - \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}_\pi}(c')$ is nonzero only when $\pi(b) = c$ and either $\pi(b) = c'$ (in which case it is positive) or $\pi(a) = c'$ (in which case it is negative). From this description, we see that when $c' = c$, this can *only* take on positive values (and does whenever $\pi(b) = c$), and when $c' \neq c$, this can *only* take on negative values (and does whenever $\pi(b) = c$ and $\pi(a) = c'$). Hence, $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^c}(c') - \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c')$ is positive when $c' = c$ and negative when $c' \neq c$, as needed.

To complete the proof, we show that $\boldsymbol{\sigma}^c$ and $\boldsymbol{\sigma}^{\text{unif}}$ are $t$-indistinguishable. To that end, fix $Q \subseteq C$ of size $t$. The key fact we will use is that if $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$ are $t$-indistinguishable, then $\pi\circ\boldsymbol{\sigma}^1$ and $\pi\circ\boldsymbol{\sigma}^2$ are $t$-indistinguishable for all $\pi$. This implies that $(\pi\circ\boldsymbol{\sigma}^{a\leftrightarrow b})|_Q = (\pi \circ \boldsymbol{\sigma})|_Q$ for all $\pi$. Therefore, both $\boldsymbol{\sigma}^c|_Q$ and $\boldsymbol{\sigma}^{\text{unif}}|_Q$ are mixtures over the exact same $m!$ distributions, and are hence equal, as needed. □

**Lemma 4.4.2.** *Fix a voting rule $f$ and suppose there are $m$ profiles that are all $t$-indistinguishable, but each has a distinct singleton $f$-winner. Then, for all (possibly randomized) algorithms $A$ which on input profile $\boldsymbol{\sigma}$ can make queries of size at most $t$ to $\boldsymbol{\sigma}$ and output a candidate, there is a profile $\boldsymbol{\sigma}^*$ with a unique $f$-winning candidate $c$, such that the probability $A$ outputs $c$ on $\boldsymbol{\sigma}^*$ is at most $1/m$.*

*Proof.* Let $\{\boldsymbol{\sigma}^c\}_{c \in C}$ denote the set of $m$ profiles that are all $t$-indistinguishable, and let $c$ be the $f$-winner on profile $\boldsymbol{\sigma}^c$. Note that an algorithm run on any of these profiles will receive the exact same responses to queries. Hence, its output must be identical for all of them. There must be some candidate $a^*$ which it outputs with probability at most $1/m$, and hence, $\boldsymbol{\sigma}^{a^*}$ satisfies the desired properties. □

Taken together, these two lemmas outline sufficient conditions wherein limited query algorithms cannot compute the winner. Combined with the construction presented in Lemma 4.3.1, it allows us to immediately conclude the following result about the impossibility of computing a plurality winner with any restricted query size.

**Theorem 4.4.3.** *For any number of candidates $m \geq 2$, for all $t < m$, no randomized $t$-query algorithm can always output a plurality winner with probability more than $1/m$.*

*Proof.* In Lemma 4.3.1, we proved the existence of a profile $\boldsymbol{\sigma}$ that has distinct plurality scores for two candidates $a, b$, and is $m - 1$ indistinguishable from its transposed profile $\boldsymbol{\sigma}^{a \leftrightarrow b}$. As such, we can apply Lemma 4.4.1 and show the existence of $m$ profiles with distinct plurality winners that are all $m - 1$ indistinguishable. Applying Lemma 4.4.2 on these $m$ profiles directly gives us the desired result for plurality. □

While the result for plurality follows immediately, the question of computing arbitrary scoring rules with a limited query size requires a more involved approach. This is tackled next.

### 4.4.1 Characterization of scoring rules

We now consider an arbitrary scoring vector $\boldsymbol{\alpha}$ and determine the exact query size needed to compute an $\boldsymbol{\alpha}$-winner. Fix $1 \leq t \leq m$, and fix a candidate set $C$ of size $m$. The following notation will be convenient for discussing arbitrary positional scoring rules. For any preference profile $\boldsymbol{\sigma} \in \Pi(C)$ and candidate $c \in C$, we define $\mathrm{pos}(\boldsymbol{\sigma}, c) \in \mathbb{R}^m$ to be the vector of positional occurrences of candidate $c$ across all rankings in profile $\boldsymbol{\sigma}$. More formally, for each $j \in [m]$,

$$\mathrm{pos}(\boldsymbol{\sigma}, c)_j := \Pr_{\sigma \sim \boldsymbol{\sigma}}[\sigma(j) = c].$$

Thus, a positional scoring rule is a voting rule parameterized by scoring vector $\boldsymbol{\alpha}$ which selects a candidate $c$ maximizing $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(c) = \boldsymbol{\alpha} \cdot \text{pos}(\boldsymbol{\sigma}, c)$. Now for each $k \in [t]$, consider the scoring vector $\boldsymbol{\alpha}^k \in \mathbb{R}^m$ given by:

$$\alpha^k_j = \binom{j-1}{k-1}\binom{m-j}{t-k}.$$

We define the subspace spanned by these $k$ vectors as $R_{m,t}$. Formally,

$$R_{m,t} := \text{span}(\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \ldots, \boldsymbol{\alpha}^t) \subseteq \mathbb{R}^m.$$

To build some intuition for this space, it can be shown that for $t \geq 2$, $R_{m,t}$ contains the commonly used Borda score, corresponding to $\boldsymbol{\alpha} = (m-1, m-2, \ldots, 0)$ (see Figure 4.5 for a visualization). We next show that any scoring rule in this space can be computed with $t$-sized queries and give a constructive algorithm. This is essentially shown in Theorem 1 of Bentert and Skowron [27] but under a different model. For completeness, we present the proof for our setting below:

**Lemma 4.4.4** (Theorem 1 of Bentert and Skowron [27])**.** *For any $t \leq m$ and $\boldsymbol{\alpha} \in R_{m,t}$, given $\boldsymbol{\sigma} \in \Pi(C)$ and a candidate $c \in C$, it is possible to compute* $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(c)$ *with queries of size $t$.*

*Proof.* We will show that for each $k \leq t$, it is possible to compute $\text{score}^{\boldsymbol{\alpha}^k}_{\boldsymbol{\sigma}}(c)$. For any $\boldsymbol{\alpha} \in R_{m,t}$, since $\boldsymbol{\alpha} = \lambda_1\boldsymbol{\alpha}^1 + \cdots + \lambda_t\boldsymbol{\alpha}^t$ for some scalars $\lambda_1, \ldots, \lambda_t$, by linearity, $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(c) = \lambda_1 \text{score}^{\boldsymbol{\alpha}^1}_{\boldsymbol{\sigma}}(c) + \cdots + \lambda_t \text{score}^{\boldsymbol{\alpha}^t}_{\boldsymbol{\sigma}}(c)$. Hence, as long as we can compute the score for each of the basis vectors, we can do so for any vector in the span.

Fix a candidate $c$ and index $j \in [m]$, and let $\sigma$ be any ranking such that $\sigma(j) = c$. Since there are $m$ candidates in $C$, there are $\binom{m}{t}$ possible sets $S \subseteq C$ of size $t$. The number of such subsets $S$ for which the restricted ranking $\sigma|_S$ puts $c$ in a position $k$ is given by $\binom{j-1}{k-1}\binom{m-j}{t-k}$, since $S$ must contain the special candidate $c$, along with $k-1$ of the $j-1$ candidates ranked above $c$ in $\sigma$, and $t-k$ of the $m-j$ candidates ranked below $c$ in $\sigma$. Thus, the probability that $\sigma|_S(k) = c$ for a uniformly random subset $S$ of size $t$ is

$$\frac{\binom{j-1}{k-1}\binom{m-j}{t-k}}{\binom{m}{t}}.$$

Consider the following algorithm. For any input preference profile $\boldsymbol{\sigma} \in \Pi(C)$ and candidate $c \in C$, we first compute the probability that, if we draw a set $S$ of $t$ dis-

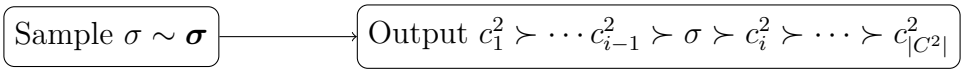



**Figure 4.4:** A process inducing $\boldsymbol{\sigma}^i$. This is parameterized by two disjoint sets, $C^1$ and $C^2$ and two candidates $a, b \in C^1$. The profile $\boldsymbol{\sigma} \in \Pi(C^1)$ satisfies the conditions of Lemma 4.3.1 with $C_1, a$, and $b$. The indexing $c_1^2 \succ \cdots \succ c_{|C^2|}^2$ is an arbitrary order of the candidates in $C^2$.

tinct candidates uniformly at random from $C$ and draw a preference $\sigma \sim \boldsymbol{\sigma}$, candidate $c$ will be at position $k$ in the ranking $\sigma$ restricted to $S$ - i.e the event $\sigma|_s(k) = c$. Clearly, this probability can be determined with queries of size $t$ by brute-forcing over all subsets $S$ of size $t$. We then output this probability multiplied by $\binom{m}{t}$. Observe that we may write the output of this algorithm as

$$\begin{aligned}
\binom{m}{t} \Pr_{\substack{S \subseteq C,\ |S|=t \\ \sigma \sim \boldsymbol{\sigma}}}[\sigma|_S(k) = c] &= \binom{m}{t} \sum_{j=1}^{t} \Pr_{\sigma \sim \boldsymbol{\sigma}}[\sigma(j) = c] \Pr_{\substack{S \subseteq C,\ |S|=t \\ \sigma \sim \boldsymbol{\sigma}}}[\sigma|_S(k) = c \mid \sigma(j) = c] \\
&= \binom{m}{t} \sum_{j=1}^{t} \mathrm{pos}(\boldsymbol{\sigma}, c)_j \frac{\binom{j-1}{k-1}\binom{m-j}{t-k}}{\binom{m}{t}} \qquad \text{by the calculation above} \\
&= \sum_{j=1}^{t} \boldsymbol{\alpha}_j^k \mathrm{pos}(\boldsymbol{\sigma}, c)_j = \boldsymbol{\alpha}^k \cdot \mathrm{pos}(\boldsymbol{\sigma}, c) = \mathrm{score}_{\boldsymbol{\sigma}}^{\boldsymbol{\alpha}^k}(c). \qquad \square
\end{aligned}$$

We now move to our main result which generalizes Theorem 4.4.3 by proving that $R_{m,t}$is exactly the space of all scoring rules computable with limited queries of size $t$, thus giving a complete characterization.

**Theorem 4.4.5.** *For any number of candidates $m \geq 2$, any $1 \leq t \leq m$, and any vector $\boldsymbol{\alpha} \in \mathbb{R}^m$*

1. *If $\boldsymbol{\alpha} \in R_{m,t}$ then there a $t$-query algorithm that always outputs a candidate $c$ maximizing $\mathrm{score}_{\boldsymbol{\sigma}}^{\boldsymbol{\alpha}}(c)$ on any input profile $\boldsymbol{\sigma}$.*

2. *If $\boldsymbol{\alpha} \notin R_{m,t}$, then no randomized $t$-query algorithm can always output an $\boldsymbol{\alpha}$-winner with probability more than $\frac{1}{m}$.*

Statement (1) is the easier part and follows immediately from Lemma 4.4.4. Our main focus here is statement (2), which leverages the construction from Section 4.3. We define a sequence of $m - t$ profiles as follows. We partition the candidate set

$C$ into two disjoint pieces, a set $C_1$ of size $t+1$ with two distinguished candidates $a, b \in C_1$, and a set $C_2$ of size $m-t-1$. Let $\boldsymbol{\sigma} \in \Pi(C_1)$ be a profile satisfying the conditions of Lemma 4.3.1 with $C_1$, $a$ and $b$, i.e., $\text{plu}_{\boldsymbol{\sigma}}(a) \neq \text{plu}_{\boldsymbol{\sigma}}(b)$ and $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^{a \leftrightarrow b}$ are $t$-indistinguishable. Let $c_1^2 \succ \cdots \succ c_{|C_2|}^2$ be an arbitrary order of the candidates in $C^2$. For each $i \in [m-t]$, we extend $\boldsymbol{\sigma}$ to a profile on all $m$ candidates $\boldsymbol{\sigma}^i \in \Pi(C)$ by inserting it in the $C^2$ order after the first $i-1$ candidates. A visual representation of this can be found in Figure 4.4. Note that, for each $i \in [m-t]$, we clearly have that $\boldsymbol{\sigma}^i$ is $t$-indistinguishable from $(\boldsymbol{\sigma}^i)^{a \leftrightarrow b}$ since $\boldsymbol{\sigma}$ is $t$-indistinguishable from $\boldsymbol{\sigma}^{a \leftrightarrow b}$, as for any $t$-query query $Q$, the candidates from $C_2$ are all in the same order, and $\boldsymbol{\sigma}|_{Q \cap C_1} = \boldsymbol{\sigma}^{a \leftrightarrow b}|_{Q \cap C_1}$ because $|Q \cap C_1| \leq t$. We now show for any vector not in the span, the score for $a$ and $b$ on one of these $t$-indistinguishable profiles is not the same.

**Lemma 4.4.6.** *For any $\boldsymbol{\alpha} \notin R_{m,t}$, there exists some $i \in [m-t]$ such that* $\text{score}_{\boldsymbol{\sigma}^i}^{\boldsymbol{\alpha}}(a) \neq \text{score}_{\boldsymbol{\sigma}^i}^{\boldsymbol{\alpha}}(b)$.

*Proof.* For each $i \in [m-t]$, we define the vector

$$s^i := \text{pos}(\boldsymbol{\sigma}^i, a) - \text{pos}(\boldsymbol{\sigma}^i, b) \in \mathbb{R}^m.$$

Suppose toward a contradiction that candidates $a$ and $b$ have the same scores in each $\boldsymbol{\sigma}^i$ according to $\boldsymbol{\alpha}$. This implies that

$$\boldsymbol{\alpha} \cdot \text{pos}(\boldsymbol{\sigma}^i, a) = \boldsymbol{\alpha} \cdot \text{pos}(\boldsymbol{\sigma}^i, b).$$

Since the score under each basis vector $\text{score}_{\boldsymbol{\sigma}^i}^{\boldsymbol{\alpha}^k}$ for $k \in [t]$ can be computed with queries of size $t$ by Lemma 4.4.4 and each $\boldsymbol{\sigma}^i$ is $t$-indistinguishable from $(\boldsymbol{\sigma}^i)^{a \leftrightarrow b}$, we know that $a$ and $b$ must have the same scores in $\text{score}_{\boldsymbol{\sigma}^i}^{\boldsymbol{\alpha}^k}$ for each $\boldsymbol{\sigma}^i$ and $k \in [t]$ as well. In other words, for each $0 \leq k \leq t$, and for each $i \in [m-t]$, we have

$$\boldsymbol{\alpha}^k \cdot \text{pos}(\boldsymbol{\sigma}^i, a) = \boldsymbol{\alpha}^k \cdot \text{pos}(\boldsymbol{\sigma}^i, b).$$

We can therefore equivalently write for all $\boldsymbol{\alpha}' \in \{\boldsymbol{\alpha}, \boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \ldots, \boldsymbol{\alpha}^t\}$ and all $i \in [m-t]$,

$$\boldsymbol{\alpha}' \cdot s^i = 0.$$

Consider the following two subspaces of $\mathbb{R}^m$:

$$R := \text{span}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \ldots, \boldsymbol{\alpha}^t)$$
$$S := \text{span}(s^1, s^2, \ldots, s^{m-t})$$

What we have just shown is that the vectors generating $R$ are each orthogonal to the vectors generating $S$, so the spaces are orthogonal to each other. Therefore, $\dim(R) + \dim(S) \leq m$. To obtain a contradiction, we will show that $\dim(R) = t + 1$ and $\dim(S) = m - t$.

First consider $R$. For each $k \in [t]$, observe that the $j^{\text{th}}$ entry of $\boldsymbol{\alpha}^k$ is zero for all $j < k$, since $\binom{j-1}{k-1} = 0$. On the other hand, the $k^{\text{th}}$ entry of $\boldsymbol{\alpha}^k$ is

$$\binom{k-1}{k-1}\binom{m-k}{t-k} = \binom{m-k}{t-k} > 0$$

because $k \leq t \leq m$. Thus, arranging the vectors $\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \ldots, \boldsymbol{\alpha}^t$ as the rows of a matrix, we have a triangle of zeros below nonzero diagonal entries of the form

$$M = \begin{pmatrix} \boldsymbol{\alpha}^1 \\ \boldsymbol{\alpha}^2 \\ \vdots \\ \boldsymbol{\alpha}^t \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}^1_1 > 0 & \boldsymbol{\alpha}^1_2 & \cdots & \boldsymbol{\alpha}^1_t \\ 0 & \boldsymbol{\alpha}^2_2 > 0 & \cdots & \boldsymbol{\alpha}^2_t \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\alpha}^t_t > 0 \end{pmatrix}$$

Clearly, $M$ has full rank, so the $\boldsymbol{\alpha}^k$ vectors are all linearly independent. Furthermore, $\boldsymbol{\alpha}$ is independent of all of these $t$ basis vectors, since we are assuming $\boldsymbol{\alpha} \notin R_{m,t}$. Hence, the $t + 1$ vectors generating $R$ are independent, so $R$ has dimension $t + 1$.

Now consider $S$. For each $i \in [m - t]$, there is no way that either of the two special candidates $a$ and $b$ could be ranked above position $i$ in any preference $\sigma$ in the support of $\boldsymbol{\sigma}^i$, since $a$ and $b$ belong to $C_1$, not $C_2$. Thus, for each $j < i$, we have

$$s^i_j = \text{pos}(\boldsymbol{\sigma}^i, a)_j - \text{pos}(\boldsymbol{\sigma}^i, b)_j = 0 - 0 = 0.$$

On the other hand, $a$ and $b$ occur exactly at position $i$ in $\boldsymbol{\sigma}^i$ with the same probability that they occur first in $\boldsymbol{\sigma}$ satisfying Lemma 4.3.1. By Lemma 4.3.1, the probability of $a$ occurring at the first position is different than the probability of $b$ occurring

at the first position. In other words,

$$s^i_i = \text{pos}(\boldsymbol{\sigma}^i, a)_i - \text{pos}(\boldsymbol{\sigma}^i, b)_i = \text{pos}(\boldsymbol{\sigma}, a)_1 - \text{pos}(\boldsymbol{\sigma}, b)_1 = \text{plu}_{\boldsymbol{\sigma}}(a) - \text{plu}_{\boldsymbol{\sigma}}(b) \neq 0.$$

Thus, by the same triangle-of-zeros argument as before, we conclude that the $s^i$ vectors are independent, so $\dim(S) = m - t$. □

*Proof of Theorem 4.4.5.* First suppose $\boldsymbol{\alpha} \in R_{m,t}$. By Lemma 4.4.4, on input $\boldsymbol{\sigma}$, we can compute $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}(c)$ for each candidate $c$ using queries of size $t$. Therefore, we can output a candidate maximizing this score.

On the other hand, suppose $\boldsymbol{\alpha} \notin R_{m,t}$. Then let $\boldsymbol{\sigma}^i$ be as in Lemma 4.4.6. Since $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(a) \neq \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(b)$, yet $\boldsymbol{\sigma}^i$ is indistinguishable from $(\boldsymbol{\sigma}^i)^{a \leftrightarrow b}$, the conclusion follows immediately from Lemmas 4.4.1 and 4.4.2. □

To make this space of computable scoring rules interpretable, we visualize the subspaces $R_{m,t}$ for $m = 4$. Since translating a scoring vector by a constant and scaling by a positive value do not affect the induced rule, all rules in $\mathbb{R}^4$ can be normalized such that they are contained within the 3-dimensional simplex (a tetrahedron). For instance, the scoring vector $(3, 2, 1, 0)$ for Borda is equivalent to $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}, 0)$ while the one for veto $(0, 0, 0, -1)$ corresponds to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$. Figure 4.5 depicts this 3-simplex of scoring rules for $m = 4$ candidates and highlights its intersection with the subspaces $R_{4,3}$, $R_{4,2}$, and $R_{4,1}$ along with other rules of interest.[6] Note that $R_{4,4}$ corresponds to the whole simplex.

### 4.4.2 Single Transferrable Vote

Next, we consider the Single Transferrable Vote (STV) which cannot be parameterized by a scoring vector. Nonetheless, similar to plurality, we find a strong negative result about its computability with any limited-sized queries.

**Theorem 4.4.7.** *For any number of candidates $m \geq 2$, for all $t < m$, no randomized $t$-query algorithm can always output an STV winner with probability more than $\frac{1}{m}$.*

*Proof.* Observe that when $m = 2$, the STV winner is equivalent to the plurality winner, so this is directly implied by Theorem 4.4.3. Fix $m \geq 3$. We would like to apply

[6] Although as a subspace of $\mathbb{R}^m$, $R_{m,t}$ is $t$-dimensional, when we restrict to the simplex, we lose a dimension. Hence $R_{4,4}$ becomes 3-dimensional, $R_{4,3}$ becomes 2-dimensional and so on.

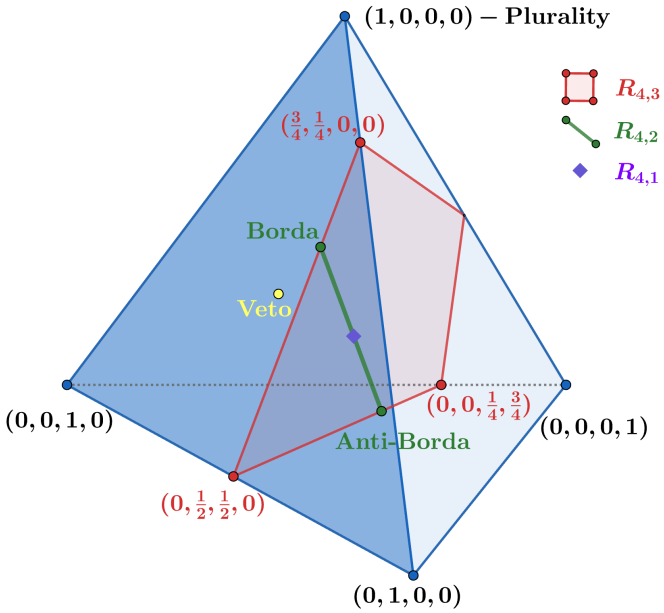



**Figure 4.5:** The space of positional scoring rules for $m = 4$ candidates. The corners represent rules that give all weight to a single position, with the top being Plurality. The red kite-shaped region is the 2-dimensional subspace $R_{4,3}$ spanned by the vectors $\boldsymbol{\alpha}^1 = (3, 1, 0, 0)$, $\boldsymbol{\alpha}^2 = (0, 1, 1, 0)$, and $\boldsymbol{\alpha}^3 = (0, 0, 1, 3)$, which are normalized in the simplex as the three red points at the corners of the kite. As we have shown, this subspace does not contain Plurality. Nested within this subspace is $R_{4,2}$, the green 1-dimensional subspace spanned by the Borda and Anti-Borda scoring vectors. The purple point at the middle of this line is the trivial voting rule that gives every candidate the same score, which is the only element of the 0-dimensional subspace $R_{4,1}$.

Lemma 4.4.2, and to do so, we will construct $m$ profiles $\{\boldsymbol{\sigma}_{STV}^{c_1}, \ldots, \boldsymbol{\sigma}_{STV}^{c_m}\}$ that are all $(m-1)$-indistinguishable, but the STV winner on profile $\boldsymbol{\sigma}_{STV}^{c_i}$ is candidate $c_i$. The construction of such profiles is as follows. Let $\alpha = (-1, 0, \ldots, 0)$, the negation of the plurality score vector. Since on any profile $\boldsymbol{\sigma}$ and any candidate pair $a, b$, we have that $\text{plu}_{\boldsymbol{\sigma}}(a) \neq \text{plu}_{\boldsymbol{\sigma}}(b) \iff sc_a^{\alpha}(\boldsymbol{\sigma}) \neq sc_b^{\alpha}(\boldsymbol{\sigma})$, we can use Lemmas 4.3.1 and 4.4.1 to obtain a set of $(m-1)$-indistinguishable profiles $\{\boldsymbol{\sigma}^c\}_{c \in M}$, where candidate $c$ is the unique $\alpha$ score maximizer on profile $\boldsymbol{\sigma}^c$; correspondingly, it is the unique plurality minimizer on profile $\boldsymbol{\sigma}^c$ due to the choice of $\alpha$.

Next, fix a directed cycle of the candidates $D := c_1 \to c_2 \to \cdots \to c_m \to c_1$. Let $R$ be the set of all rankings such that the first and last candidates are consecutive in the cycle, i.e., $R = \{\sigma \in \mathcal{L}(C) \mid (\sigma(1), \sigma(m)) \in D\}$. Choose $\varepsilon > 0$ such that $\varepsilon < \frac{1-\varepsilon}{m(m-1)}$ ($\varepsilon = \frac{1}{m^2}$ will do). Fix a candidate $c$, and let next$(c)$ be the subsequent candidate in

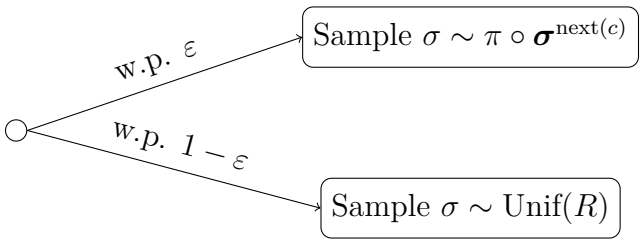



**Figure 4.6:** A process inducing $\boldsymbol{\sigma}^c_{STV}$.

the cycle, i.e., the unique candidate $c'$ such that $(c, c') \in D$. Define $\boldsymbol{\sigma}^c_{STV}$ as follows: with probability $\varepsilon$, output a sample $\sigma$ from $\boldsymbol{\sigma}^{\text{next}(c)}$, with remaining probability $1-\varepsilon$, select $\sigma$ from $\text{Unif}(R)$. A visual representation of this can be found in Figure 4.6.

Observe that for any pair $c, c'$, the profiles $\boldsymbol{\sigma}^c_{STV}$ and $\boldsymbol{\sigma}^{c'}_{STV}$ are $(m-1)$-indistinguishable. Indeed, their generating processes are the same with probability $1-\varepsilon$ (when we sample uniformly from $R$), and with probability $\varepsilon$ they differ due to $\boldsymbol{\sigma}^c$ and $\boldsymbol{\sigma}^{c'}$ which are themselves $(m-1)$-indistinguishable.

We next show that on each profile $\boldsymbol{\sigma}^c_{STV}$, the unique STV winner is $c$. Without loss of generality, consider the candidate $c = c_m$ (so $\text{next}(c) = c_1$, $\text{next}(c_1) = c_2$, and so on) as the argument holds symmetrically for any other $c$. We will show by strong induction that the $k$'th candidate to be eliminated is $c_k$. This implies that $c_m$ will be the final candidate remaining, and thus the STV winner.

We begin with the base case, that $c_1$ is the first to be eliminated. Note that when sampling uniformly from $R$, by symmetry, each candidate has the same plurality score. On the other hand, in $\boldsymbol{\sigma}^{\text{next}(c)}$, candidate $\text{next}(c) = c_1$ is the unique plurality minimizer. Thus, in the mixed profile $\boldsymbol{\sigma}^c_{STV}$, $c_1$ has the lowest plurality score and is eliminated first.

Next, suppose candidates $c_1, \ldots, c_{k-1}$ for $k \geq 2$ (and $k \leq m-1$) have been eliminated. We will show that the next to be eliminated is $c_k$. Let $C^k = C \setminus \{c_1, \ldots, c_{k-1}\}$ be the set of uneliminated candidates. We first consider the proportion of first-place votes each candidate $c \in C^k$ gets in $\text{Unif}(R)|_{C^k}$. Let $R = R^1 \sqcup \cdots \sqcup R^m$ be a partition of $R$ such that $R^i$ contains the rankings with $c_i$ ranked first. By symmetry, these are each the same size. Note that for $i \geq k$, $c_i \in C_k$, so rankings in $R^i$ will continue to rank $c_i$ first, each accounting for a $1/m$ proportion of the rankings. For $i \leq k-2$, since $\text{next}(c_i) \notin C^k$, by symmetry, the first place votes of $R^i$ will be distributed equally

among all candidates in $C^k$. For $R^{k-1}$ however, every $\sigma \in R^{k-1}$ ranks $c_k$ last, and, since $|C^k| \geq 2$, no votes will go to $c_k$; instead, they will be spread equally among $C^k \setminus \{c_k\}$. Hence, the plurality score of $c_k$ on $\text{Unif}(R)|_{C^k}$ will be $\frac{1}{m\cdot(|C^k|-1)} \geq \frac{1}{m(m-1)}$ smaller than all other candidates. By the choice of $\varepsilon$, $(1-\varepsilon)\frac{1}{m(m-1)} > \varepsilon$, so no matter how many first place votes $c_k$ gets on $\boldsymbol{\sigma}^{\text{next}(c)}$, $c_k$ has the smallest plurality score on $\boldsymbol{\sigma}^c_{STV}|_{C^k}$. Therefore, it is the next to be eliminated.

Finally, we apply Lemma 4.4.2 to this set of $m$ profiles to conclude that there exists a profile where no $(m-1)$-query algorithm can determine the STV winner with probability greater than $\frac{1}{m}$. □

## 4.5 Query Complexity

In the previous section, we characterized when it was information-theoretically possible to find winning candidates using a certain query size. We turn now to focusing on those cases when it *is* possible and prove bounds on the *query complexity*. In other words, when it is possible to determine the winner with limited-sized query size, how many such queries are needed?

For an integer $k \leq m$, we use the notation $\binom{S}{k}$ to denote the set of all subsets of $S$ of size $k$. Fix a scoring vector $\boldsymbol{\alpha}$ and let $t^*$ be the minimal value such that $\boldsymbol{\alpha} \in R_{m,t^*}$. Suppose we can make queries of size $t \geq t^*$ and wish to find a candidate maximizing $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}}$. As a benchmark, note that if we make enough queries to be able to deduce $\boldsymbol{\sigma}|_S$ for all possible $S \in \binom{C}{t^*}$, then it is information-theoretically possible to find this winning candidate. Indeed, one could simulate any $t^*$-query algorithm using this (say the one from Lemma 4.4.4) as they would have the responses for all $t$-sized query. Let $\text{cov}(m,t,t^*)$ be the minimum number of subsets of size $t$ needed to cover all subsets of size $t^*$ out of a set of size $m$. This value is referred to as a *covering number*; computing such covering numbers and optimal subset structures that induce them is a canonical problem in combinatorics with a rich history (see, e.g., Mills and Mullin [145]). For our purposes, a reasonable (and nearly tight) lower bound on covering numbers is

$$\text{cov}(m,t,t^*) \geq \frac{\binom{m}{t^*}}{\binom{t}{t^*}}.$$

This follows from a simple argument: There are $\binom{m}{t^*}$ subsets of size $t^*$, and each subset of size $t$ can cover at most $\binom{t}{t^*}$ subsets them. When $t$ and $t^*$ are treated as con-

stant, then this value is $\Omega(m^{t^*})$.

Our primary question is whether we can cleverly choose queries to use fewer than $\text{cov}(m, t, t^*)$ queries. As a motivating example, consider instead finding a *Condorcet winner* under our model. A Condorcet winner on profile $\boldsymbol{\sigma}$ is a candidate $a$ that beats all others in a pairwise competition. More formally, for all $b \neq a$, $\Pr_{\sigma \sim \boldsymbol{\sigma}}[a \succ_\sigma b] > 1/2$. Note that Condorcet winners need not exist,[7] however when they do, they are unique. From the definition, we can see that using queries of size $t^* = 2$ is sufficient to determine whether a Condorcet winner exists and, if so, determine this candidate as this only depends on pairwise margins. One option is to make all possible queries of size 2; this requires $\text{cov}(m, 2, 2) = \binom{m}{2} = \Theta(m^2)$ queries. However, as shown by Procaccia [167], there is a more clever way, requiring only $O(m)$ queries to compute this winner.[8]

We now ask whether such improvements can be found for scoring rules. Unfortunately, we show that for both deterministic and randomized algorithms, they cannot.

**Theorem 4.5.1.** *Fix $\boldsymbol{\alpha} \in \mathbb{R}^m$ and let $t^*$ be the minimal value such that $\boldsymbol{\alpha} \in R_{m,t^*}$. For $t^* \geq 2$, any deterministic $t$-query algorithm that always outputs a candidate maximizing $\text{score}_{\boldsymbol{\sigma}}^{\boldsymbol{\alpha}}$ must make at least $\text{cov}(m, t, t^*)$ queries in the worst case[9]. Further, any randomized algorithm making at most $\delta \frac{\binom{m}{t^*}}{\binom{t}{t^*}}$ queries outputs an $\boldsymbol{\alpha}$-winner with probability at most $\min(\delta + \frac{1}{m}, \delta + (1-\delta)\frac{1}{t^*})$ in the worst case.*

*Proof.* Fix $\boldsymbol{\alpha}$ and $t^*$. Let $C_1$ be a set of candidates of size $t^*$ with two distinguished candidates $a, b \in C_1$, and let $C_2 = \overline{C_1}$ be the remaining candidates. We construct $\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^{m-t^*}$ just as in Figure 4.4 from Section 4.4.1. Recall that each profile $\boldsymbol{\sigma}^i$ has a profile $\boldsymbol{\sigma} \in \Pi(C_1)$ satisfying the conditions of Lemma 4.3.1 with $a$ and $b$ "contained" in it. In addition, it has $i-1$ of the $C_2$ candidates ranked in a fixed order before $\boldsymbol{\sigma}$, and the rest are in a fixed order after. After describing these profiles in Section 4.4.1,

[7] The classic example is when a third of the voters have each of the rankings $a \succ b \succ c$, $c \succ a \succ b$, and $b \succ c \succ a$.

[8] The algorithm runs in two phases. First, run a knockout tournament among the candidates, where the candidate receiving more pairwise votes makes it on to the next round. If there is a Condorcet winner, then that candidate must be the winner of the knockout tournament. In the second phase, compare this winner to all other candidates they did not play. If they win all of these comparisons, they are the Condorcet winner, if not, there is no winner. This requires $2m - \lfloor \log m \rfloor - 2$ queries.

[9] It is only when $\boldsymbol{\alpha}$ is a constant vector does $t^* = 1$, a degenerate case that we ignore since any candidate can be considered the winner.

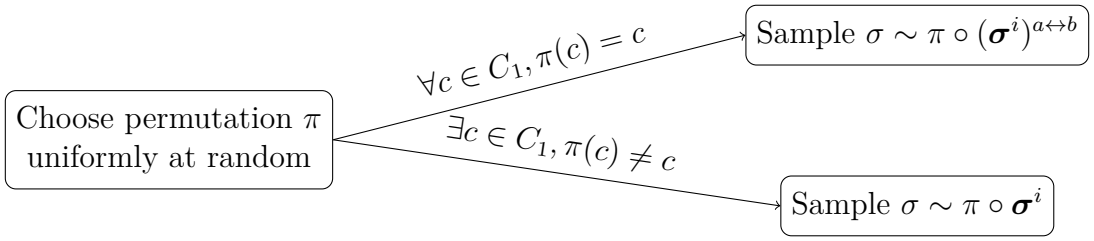



**Figure 4.7:** A process inducing $\boldsymbol{\sigma}^*$.

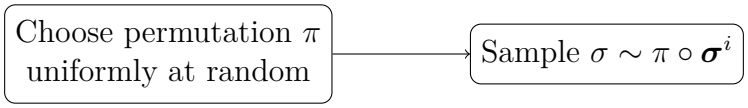



**Figure 4.8:** A process inducing $\boldsymbol{\sigma}^{\text{unif}}$.

we observed that each $\boldsymbol{\sigma}^i$ and $(\boldsymbol{\sigma}^i)^{a \leftrightarrow b}$ are $(t^* - 1)$-indistinguishable. However, we claim that an even stronger property is true: For any $t$ sized query $Q$ that does not contain $C_1$, $\boldsymbol{\sigma}^i$ and $(\boldsymbol{\sigma}^i)^{a \leftrightarrow b}$ are indistinguishable (note $t$ may be much larger than $t^*$). More formally, for all $Q$ such that $C_1 \not\subseteq Q$, $\boldsymbol{\sigma}^i|_Q = (\boldsymbol{\sigma}^i)^{a \leftrightarrow b}|_Q$. Indeed, the candidates of $C_2 \cap Q$ are always in the same order, either before or after the candidates of $C_1$. The candidates in $C_1$ will follow the distribution according to $\boldsymbol{\sigma}|_{Q \cap C_1}$ and $\boldsymbol{\sigma}^{a \leftrightarrow b}|_{Q \cap C_1}$. By Lemma 4.3.1, since $Q \cap C_1 \subsetneq C_1$, these are identical.

By definition of $t^*$, $\boldsymbol{\alpha} \notin R_{m,t^*-1}$. Hence, Lemma 4.4.6 implies that for one of these profiles, $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(a) \neq \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(b)$. Fix such an $i$, and without loss of generality, assume $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(a) > \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(b)$. Consider the profile $\boldsymbol{\sigma}^*$ induced by sampling a permutation $\pi$ uniformly at random, and, if $\pi(c) = c$ for all $c \in C_1$, sample $\sigma \sim \pi \circ (\boldsymbol{\sigma}^i)^{a \leftrightarrow b}$, otherwise, sample $\sigma \sim \pi \circ \boldsymbol{\sigma}^i$. This is shown in Figure 4.7.

Next, we will compare $\boldsymbol{\sigma}^*$ to another profile, $\boldsymbol{\sigma}^{\text{unif}}$, defined in Figure 4.8, where we sample from $\pi \circ \boldsymbol{\sigma}^i$ regardless of $i$. Note that $\boldsymbol{\sigma}^{\text{unif}}$ is the uniform distribution over all rankings which can be equivalently achieved by first sampling $\sigma \sim \boldsymbol{\sigma}^i$, and then outputting $\pi \circ \sigma$ for a $\pi$ that is uniformly selected.[10] We will show two things (i) $\boldsymbol{\sigma}^*|_Q = \boldsymbol{\sigma}^{\text{unif}}$ unless $C_1 \subseteq Q$, and (ii) $b$ is the unique $\boldsymbol{\alpha}$-winner on $\boldsymbol{\sigma}^*$. Note that (i) implies that on any query $Q$ with $C_1 \not\subseteq Q$, $\boldsymbol{\sigma}^*|_Q$ is the uniform distribution over rankings in $\mathcal{L}(Q)$.

[10]Note that although the generating process for $\boldsymbol{\sigma}^{\text{unif}}$ here is different than the one used in Lemma 4.4.1 and Figure 4.3, the resulting distribution is still the same.

For (i), fix a query $Q$ with $C_1 \not\subseteq Q$. Since the $\pi(c) \neq c$ for some $c \in C_1$ branch of $\boldsymbol{\sigma}^*$ is identical to $\boldsymbol{\sigma}^{\text{unif}}$, it suffices to focus on $\pi$ such that $\pi(c) = c$ for all $c \in C_1$. When this holds, $\pi$ can only reorder the candidates of $C_2$, leaving candidates in $C_1$ unchanged. Therefore, $(\pi \circ (\boldsymbol{\sigma}^i)^{a \leftrightarrow b})|_Q = (\pi \circ \boldsymbol{\sigma}^i)|_Q$. Hence, sampling from $\boldsymbol{\sigma}^*|_Q$ is equivalent to sampling $\pi$ uniformly at random and sampling from $(\pi \circ \boldsymbol{\sigma}^i)|_Q$, equivalent to sampling from $\boldsymbol{\sigma}^{\text{unif}}|_Q$.

For (ii), we will show that $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^*}(b) > \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(b)$ while $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^*}(c) \leq \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c)$ for all $c \neq b$. By symmetry, the scores of all candidates in $\boldsymbol{\sigma}^{\text{unif}}$ are the same, hence, this shows that $b$ is the unique winner. To that end, note that the only time in sampling $\boldsymbol{\sigma}^*$ and $\boldsymbol{\sigma}^{\text{unif}}$ that the scores will differ, is if we take the top branch, in which case $\pi(c) = c$ for all $c \in C_1$. By assumption, $\pi(a) = a$ and $\pi(b) = b$ as $a, b \in C_1$, so we are simply swapping $a$ and $b$. Since $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(a) > \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^i}(b)$, this strictly increases the score of $b$ on average, and strictly decreases the score of $a$. Hence, $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^*}(b) > \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(b)$, $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^*}(a) < \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(a)$, and $\text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^*}(c) = \text{score}^{\boldsymbol{\alpha}}_{\boldsymbol{\sigma}^{\text{unif}}}(c)$ for all $c \neq a, b$, as needed.

Fix a deterministic algorithm $t$-query algorithm $\mathcal{A}$ that outputs a candidate after making strictly fewer than $\text{cov}(m, t, t^*)$ queries. We will show that it cannot always output an $\boldsymbol{\alpha}$-winner. Consider a run of the algorithm where on every query $Q$, it receives in response the uniform distribution over $\mathcal{L}(Q)$. Suppose on this run, it outputs candidate $c$. Now, by the definition of the covering number, there must be a set $C'$ with $|C'| = t^*$ such that $C'$ was not contained in any query made by the algorithm. Since $t^* \geq 2$, $C' \setminus \{c\}$ is not empty. Let $c^* \in C' \setminus \{c\}$. Let $\pi$ be a permutation such that $\pi(b) = c^*$ and $\pi$ maps $C_1 \setminus \{b\}$ to $C' \setminus \{c^*\}$. Consider the running $\mathcal{A}$ on $\pi \circ \boldsymbol{\sigma}^*$. Note that on every query $Q$ not containing $C_1$, the response will be indistinguishable from $\boldsymbol{\sigma}^{\text{unif}}$, and hence, the uniform distribution over $Q$. Therefore, by above, $\mathcal{A}$ will return candidate $c$ on this instance. However, by construction, $c^*$ is the unique $\boldsymbol{\alpha}$-winner, and $c^* \neq c$, a contradiction.

Next, we will show that a randomized algorithm making at most $\delta \frac{\binom{m}{t^*}}{\binom{t}{t^*}}$ queries will output an $\boldsymbol{\alpha}$-winner with probability at most $\min(\delta + \frac{1}{m}, \delta + (1-\delta)\frac{1}{t^*})$. We will make use of Yao's Minimax Principle [208]. More specifically, will show that there is a distribution over profiles such that no deterministic algorithm can be correct with larger probability. This implies that no randomized algorithm can achieve a larger probability on a worst-case profile.

The distribution over instances we will choose is simply uniform over $\pi \circ \boldsymbol{\sigma}^*$ for all

permutations $\pi$. Fix an arbitrary deterministic algorithm $\mathcal{A}$ that always outputs a candidate after at most $\delta\binom{m}{t^*}/\binom{t}{t^*}$ queries. Consider a run of this algorithm where the response to every query $Q$ is the uniform distribution over $\mathcal{L}(Q)$, and suppose on this run, the output candidate is $c$. Let $\mathcal{Q}$ be the set of queries asked on this run. We have that $|\mathcal{Q}| \leq \delta\binom{m}{t^*}/\binom{t}{t^*}$ by assumption. Observe that since each query of size $t$ can cover $\binom{t}{t^*}$ sets of size $t^*$, at most a $\delta$-fraction of the $\binom{m}{t^*}$ $t^*$-sets are covered by $\mathcal{Q}$.

We claim that the algorithm must be incorrect for all $\pi \circ \boldsymbol{\sigma}^*$ such that both (i) $\pi(C_1) \not\subseteq Q$ for all $Q \in \mathcal{Q}$ and (ii) $\pi(b) \neq c$. Indeed, (i) ensures that the run of the algorithm will always lead to uniform responses, meaning $\mathcal{A}$ *must* output $c$, and (ii) ensures this is the incorrect choice. More formally, let

$$\mathcal{E}_1 = \{\pi \mid \pi(C_1) \not\subseteq Q \text{ for all } Q \in \mathcal{Q}\}$$

be the event that the first property holds, and

$$\mathcal{E}_2 = \{\pi \mid \pi(b) \neq c\}$$

be the event that the second does. The probability of success is at most $1 - \Pr[\mathcal{E}_1 \cap \mathcal{E}_2]$. We will upper bound this in two ways. First,

$$1 - \Pr[\mathcal{E}_1 \cap \mathcal{E}_2] = \Pr[\overline{\mathcal{E}_1} \cup \overline{\mathcal{E}_2}] \leq \Pr[\overline{\mathcal{E}_1}] + \Pr[\overline{\mathcal{E}_2}] \leq \delta + \frac{1}{m},$$

where the first inequality holds by the union bound, $\Pr[\overline{\mathcal{E}_1}] \leq \delta$ because at most a $\delta$-fraction of all $t^*$-subsets are covered, and $\Pr[\overline{\mathcal{E}_2}] \leq \frac{1}{m}$ by symmetry. Second,

$$\begin{aligned}
1 - \Pr[\mathcal{E}_1 \cap \mathcal{E}_2] &= 1 - \Pr[\mathcal{E}_2 \mid \mathcal{E}_1] \cdot \Pr[\mathcal{E}_1] \\
&\leq 1 - \left(1 - \frac{1}{t^*}\right) \cdot (1 - \delta) \\
&= 1 - (1 - \delta) + (1 - \delta) \cdot \frac{1}{t^*} \\
&= \delta + (1 - \delta) \cdot \frac{1}{t^*}.
\end{aligned}$$

Again, $\Pr[\mathcal{E}_1] \geq 1 - \delta$ holds because at most a $\delta$-fraction are covered. The other term $\Pr[\mathcal{E}_2 \mid \mathcal{E}_1] \cdot \Pr[\mathcal{E}_1]$ holds because conditioned on $\pi(C_1) = C'$ for *any* uncovered $C'$, the probability that $\pi(b) = c$ is at most $1 - \frac{1}{t^*}$. Indeed, this holds exactly if $c \in C'$, and

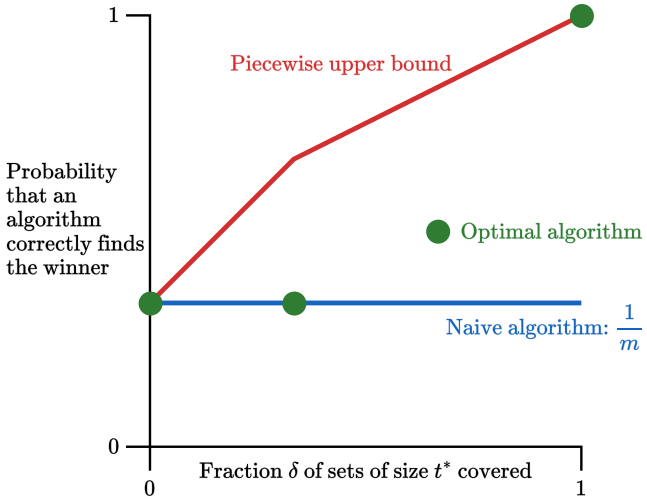



**Figure 4.9:** Success probabilities for randomized algorithms computing the Borda winner on $m = 3$ candidates making queries of size $t = t^* = 2$. The red piecewise-linear curve, which has similar shapes for larger parameters and other computable scoring rules, represents the upper bound on the success probability of any algorithm from Theorem 4.5.1. The blue curve is the best-known general lower bound obtained by the algorithm that makes no queries and simply guesses a candidate at random. The green points are the true success probabilities of the optimal algorithm given by Theorem 4.5.2.

is 0 otherwise. Together, these two bounds imply that the probaility of success is at most $\min(\delta + \frac{1}{m}, \delta + (1-\delta)\frac{1}{t^*})$. □

Theorem 4.5.1 completely settles the query complexity of deterministic algorithms for computing all positional scoring rules. However, for randomized algorithms, the story is not quite complete. On the one hand, if $t^*$ and $t$ are constants, and an algorithm would like to be correct with probability $\frac{1}{m} + c$ for a constant $c$, then $\Omega(m^{t^*})$ queries are needed (hiding constants depending on $t^*, t$, and $c$), and $O(m^{t^*})$ clearly suffice, as regardless of $t$, $\binom{m}{t^*}$ are certainly enough to cover all sets. On the other hand, if an algorithm can make $\delta\frac{\binom{m}{t^*}}{\binom{t}{t^*}}$ queries for a fixed $\delta$, we do not know the exact probability with which it can be correct. Figure 4.9 depicts the gap between the upper bound and the best general lower bound as functions of the parameter $\delta$. This lower bound is essentially the naive algorithm achieving $\frac{1}{m}$ by simply picking a candidate at random. However, as the following result shows, even for the simplest non-trivial case of $m = 3$ and $t = t^* = 2$, the true query complexity of computing the

(essentially unique) scoring rule in $R_{3,2}$ is strictly between our general bounds when $\delta = \frac{2}{3}$.

**Theorem 4.5.2.** *With $m = 3$ candidates, the optimal randomized algorithm making queries of size $t = 2$ to compute the Borda count winner succeeds with*

1. *worst-case success probability $\frac{1}{3}$ when allowed to make exactly one query, and*
2. *worst-case success probability $\frac{1}{2}$ when allowed to make exactly two queries.*

Before giving the proof, we note that the measure of worst-case optimality here is a bit finicky. On the negative side, we show that for all algorithms and any $\varepsilon > 0$, there are instances where they do not succeed with probability more than $\frac{1}{3} + \varepsilon$ for (1), and $\frac{1}{2} + \varepsilon$ for (2). At least for (1), this relaxation is necessary. Consider the algorithm that makes a single query to a uniformly random pair of candidates and selects between them with the probabilities given by the query response (i.e., if the algorithm learns that $\Pr_{\sigma \sim \boldsymbol{\sigma}}[a \succ_\sigma b] = p$, it picks $a$ with probability $p$ and $b$ with probability $1 - p$). It can be shown that this process selects each candidate with probability proportional to its Borda score.[11] Given any fixed profile, unless all three candidates have the same score, a maximal one will be selected with probability strictly greater than $\frac{1}{3}$ (and if they are all the same, then all are Borda winners, and hence the algorithm succeeds with probability 1). However, there are instances where the best Borda score is arbitrarily close to the others, resulting in a success probability no constant greater than $\frac{1}{3}$. Thus, by saying that "the optimal" randomized algorithm that makes a single query achieves a worst-case success probability of $\frac{1}{3}$, we really mean that it is not possible to surpass $\frac{1}{3}$ by any constant. In the proof of Theorem 4.5.2, we must construct a family of increasingly more difficult instances that bring the success probabilities closer to $\frac{1}{3}$ and $\frac{1}{2}$. This becomes quite complicated, involving a construction based on Fibonacci numbers to ensure query responses do not leak cardinal information about the relative strengths of candidates.

Proof of Theorem 4.5.2: Fix a set of candidates $C = \{a, b, c\}$. Throughout the proof, we use the scoring vector $(1, 0, -1)$ to compute Borda scores (which is equivalent to the more traditional choice of $(2, 1, 0)$ by translation). When writing scores,

[11] The probability it picks a candidate $c$ is equal to $\frac{1}{\binom{m}{2}} \sum_{c' \neq c} \Pr_{\sigma \sim \boldsymbol{\sigma}}[c \succ c']$. It is well known that the Borda score of $c$ is equal to $\sum_{c' \neq c} \Pr_{\sigma \sim \boldsymbol{\sigma}}[c \succ c']$ [36].

we drop the $(1, 0, -1)$ superscript in the score notation, using $\text{score}_{\boldsymbol{\sigma}}(c')$ to refer to $\text{score}_{\boldsymbol{\sigma}}^{(1,0,-1)}(c')$. We also use the convention that when $\boldsymbol{\sigma}$ is queried on $\{a, b\}$, the algorithm learns $\Pr_{\sigma\sim\boldsymbol{\sigma}}[a \succ_\sigma b]$, on $\{b, c\}$, the algorithm learns $\Pr_{\sigma\sim\boldsymbol{\sigma}}[b \succ_\sigma c]$, and on $\{a, c\}$, the algorithm learns $\Pr_{\sigma\sim\boldsymbol{\sigma}}[c \succ_\sigma a]$. These single numbers completely parameterize the distribution $\boldsymbol{\sigma}|_Q$ for each $Q$ of size 2. One can check that the following equalities hold for scores

$$\begin{aligned}
\text{score}_{\boldsymbol{\sigma}}(a) &= \Pr_{\sigma\sim\boldsymbol{\sigma}}[a \succ_\sigma b] - \Pr_{\sigma\sim\boldsymbol{\sigma}}[b \succ_\sigma c], \\
\text{score}_{\boldsymbol{\sigma}}(b) &= \Pr_{\sigma\sim\boldsymbol{\sigma}}[b \succ_\sigma c] - \Pr_{\sigma\sim\boldsymbol{\sigma}}[c \succ_\sigma a], \\
\text{score}_{\boldsymbol{\sigma}}(c) &= \Pr_{\sigma\sim\boldsymbol{\sigma}}[c \succ_\sigma a] - \Pr_{\sigma\sim\boldsymbol{\sigma}}[a \succ_\sigma b].
\end{aligned}$$

We begin with the lower bounds on the probabilities. First, note that it is always possible to succeed with probability $\frac{1}{3}$ by just picking a random one of the three candidates. This establishes the lower bound on (1).

For (2), consider the following algorithm. We pick a random candidate $c'$ and query both sets of size 2 containing that candidate. From this information, we are able to learn the Borda score of candidate $c'$. If the score is positive, we return $c'$. Otherwise, we randomly return one of the other two candidates. Observe that, with our choice of scoring vector $(1, 0, -1)$, the sum of all three Borda scores must be zero, so at most two are strictly positive. If none of them are positive, then they must all be zero, in which case every candidate is a Borda winner, and the algorithm succeeds with probability 1. If one Borda score is positive, then if that candidate is chosen as $c$, the algorithm succeeds with probability 1, and otherwise the algorithm succeeds with probability $\frac{1}{2}$. Since the former case happens with probability $\frac{1}{3}$, the total expected success probability is $\frac{2}{3}$. Finally, if two Borda scores are positive, then if the true Borda winner is chosen as $c'$, the algorithm succeeds with probability 1; if the other candidate with positive Borda score is chosen as $c'$, the algorithm incorrectly returns it, succeeding with probability 0; and if the candidate with negative Borda score is chosen as $c$, the algorithm succeeds with probability $\frac{1}{2}$. In total, the success probability is $\frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{2}$. Thus, the worst-case success probability is $\frac{1}{2}$.

For the upper bounds, we again use Yao's Minimax principle. That is, we will show that there are distributions over instances where no deterministic algorithm can output a Borda winner with probability more than $\frac{1}{3} + \varepsilon$ for any $\varepsilon > 0$. This implies that

no randomized algorithm can do so on every instance.

Define a family of distributions over profile $D_1, D_2, D_3, \ldots$ as follows. To generate $D_n$, we first sample three values, $p_1, p_2, p_3 \in [\frac{1}{3}, \frac{2}{3}]$ and return an arbitrary profile $\boldsymbol{\sigma}$ where

$$p_1 := \Pr_{\sigma \sim \boldsymbol{\sigma}}[a \succ_\sigma b],$$
$$p_2 := \Pr_{\sigma \sim \boldsymbol{\sigma}}[b \succ_\sigma c],$$
$$p_3 := \Pr_{\sigma \sim \boldsymbol{\sigma}}[c \succ_\sigma a].$$

Before showing how to sample $p_1, p_2, p_3$, we first show that such a profile $\boldsymbol{\sigma}$ satisfying the pairwise margins always exists. We claim that the following preference profile suffices.

| Ranking | Probability |
|---|---|
| $a \succ b \succ c$ | $p_2 - \frac{1}{3}$ |
| $a \succ c \succ b$ | $\frac{2}{3} - p_2$ |
| $b \succ c \succ a$ | $p_3 - \frac{1}{3}$ |
| $b \succ a \succ c$ | $\frac{2}{3} - p_3$ |
| $c \succ a \succ b$ | $p_1 - \frac{1}{3}$ |
| $c \succ b \succ a$ | $\frac{2}{3} - p_1$ |

The reader may verify that:

- All probabilities lie in $[0, 1]$ for $p_1, p_2, p_3 \in [\frac{1}{3}, \frac{2}{3}]$.
- The sum of all six probabilities is 1.
- The pairwise ranking probabilities are indeed given by $p_1$, $p_2$, and $p_3$.

As shown above, the $p_i$ values contain all relevant information for computing the Borda scores of each candidate: $\text{score}_{\boldsymbol{\sigma}}(a) = p_1 - p_3$, $\text{score}_{\boldsymbol{\sigma}}(b) = p_2 - p_1$, and $\text{score}_{\boldsymbol{\sigma}}(c) = p_3 - p_2$, along with the responses for all the queries. Hence, the exact construction of $\boldsymbol{\sigma}$ will not be important for the remainder of the proof.

Let $F_1, F_2, F_3, \ldots$ be the Fibonacci sequence shifted to the left by one, beginning with $F_1 := 1$, $F_2 := 2$, $F_3 := 3$, $F_4 := 5$, and so on. For any positive integer $n$,

we generate $p_1$, $p_2$, and $p_3$ for $D_n$ using the following process. First, sample $i$ uniformly from $\{1, 2, \ldots, n\}$, sample $s$ uniformly from $\{0, 1, 2, \ldots, nF_{n+2}\}$ (all integers from 0 to $nF_{n+2}$), and sample $r$ uniformly from $\{1, 2, 3, 4, 5, 6\}$. Then output the profile $\boldsymbol{\sigma}(i, s, r)$ defined by the table below, where $p_1$, $p_2$, and $p_3$ are defined from the auxiliary values $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ by the correspondence

$$p_j := \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot \hat{p}_j.$$

| Profile | True Borda winner | Scaled probabilities $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$ |
|---|---|---|
| $\boldsymbol{\sigma}(i, s, 1)$ | $a$ | $(s + F_{i+2}, s, s + F_i)$ |
| $\boldsymbol{\sigma}(i, s, 2)$ | $c$ | $(s + F_{i+2}, s, s + F_{i+1})$ |
| $\boldsymbol{\sigma}(i, s, 3)$ | $b$ | $(s + F_i, s + F_{i+2}, s)$ |
| $\boldsymbol{\sigma}(i, s, 4)$ | $a$ | $(s + F_{i+1}, s + F_{i+2}, s)$ |
| $\boldsymbol{\sigma}(i, s, 5)$ | $c$ | $(s, s + F_i, s + F_{i+2})$ |
| $\boldsymbol{\sigma}(i, s, 6)$ | $b$ | $(s, s + F_{i+1}, s + F_{i+2})$ |

Note that each $p_j \in [\frac{1}{3}, \frac{2}{3}]$ if and only if $\hat{p}_j \in [0, (n+1)F_{n+2}]$, so any pair $(i, s)$ is valid. We leave the reader to verify that the Borda winners are as stated in the table. We explicitly compute the winner in the final profile as an example:

$$\begin{aligned}
\text{score}_{\boldsymbol{\sigma}(s,i,6)}(a) &= p_1 - p_3 = \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot (\hat{p}_1 - \hat{p}_3) \\
&= \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot ((s) - (s + F_{i+2})) \\
&= \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot (-F_{i+2}) \\
\text{score}_{\boldsymbol{\sigma}(s,i,6)}(b) &= p_2 - p_1 = \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot (\hat{p}_2 - \hat{p}_1) \\
&= \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot ((s + F_{i+1}) - (s)) \\
&= \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot (F_{i+1}) \\
\text{score}_{\boldsymbol{\sigma}(s,i,6)}(c) &= p_3 - p_2 = \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot (\hat{p}_3 - \hat{p}_2)
\end{aligned}$$

$$= \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot ((s + F_{i+2}) - (s + F_{i+1}))$$
$$= \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{((n+1)F_{n+2})} \cdot (F_i)$$

Thus, the winner is $b$, since $F_{i+1}$ is the largest out of $\{-F_{i+2}, F_{i+1}, F_i\}$.

Fix a deterministic algorithm $\mathcal{A}$ making at most one query of size 2. Observe that $D_n$ is completely symmetric with respect to $p_1$, $p_2$, and $p_3$ (in the sense that permuting $p_1 \mapsto p_2$, $p_2 \mapsto p_3$, and $p_3 \mapsto p_1$ gives the same distribution on preference profiles). Thus, we may assume without loss of generality that an algorithm queries $\{a, b\}$ and learns $\hat{p}_1 \in \{0, 1, 2, \ldots, (n+1)F_{n+2}\}$. Let $\mathcal{E}$ be the event that $i \in \{3, 4, 5, \ldots, n-2\}$ and $s \in [F_{n+2}, (n-2)F_{n+2}]$. By the union bound, the probability that $\mathcal{E}$ does not occur is at most

$$\begin{aligned}\Pr[\overline{\mathcal{E}}] &\leq \Pr[i \in \{1, 2, n-1, n\}] + \Pr[s \in [0, F_{n+2}] \cup [(n-2)F_{n+2}, nF_{n+2}]] \\ &\leq \frac{4}{n} + \frac{F_{n+2} + 2F_{n+2}}{nF_{n+2}} = \frac{7}{n}.\end{aligned}$$

When $\mathcal{E}$ occurs, we will have $\hat{p}_j \in [F_{n+2}, (n-1)F_{n+2}]$ for each $j$, so in particular this holds for $\hat{p}_1$. Suppose that the algorithm additionally learns $i$, which only makes it stronger. Then there are exactly six possible choices of the parameters $s$ and $c$ that could have led to the specific realization $\hat{p}_1$:

- $r = 1$ and $s = \hat{p}_1 - F_{i+2} \implies a$ is the winner.
- $r = 2$ and $s = \hat{p}_1 - F_{i+2} \implies c$ is the winner.
- $r = 3$ and $s = \hat{p}_1 - F_i \implies b$ is the winner.
- $r = 4$ and $s = \hat{p}_1 - F_{i+1} \implies a$ is the winner.
- $r = 5$ and $s = \hat{p}_1 \implies c$ is the winner.
- $r = 6$ and $s = \hat{p}_1 \implies b$ is the winner.

Since each possibility is equally likely, no matter which candidate the algorithm picks it succeeds with probability $\frac{1}{3}$. Thus, overall,

$$\Pr[\text{success}] = \Pr[\mathcal{E}] \Pr[\text{success} \mid \mathcal{E}] + \Pr[\overline{\mathcal{E}}] \Pr[\text{success} \mid \overline{\mathcal{E}}]$$

$$\leq \Pr[\text{success} \mid \mathcal{E}] + \Pr[\overline{\mathcal{E}}]$$
$$\leq \frac{1}{3} + \frac{7}{n}.$$

By picking $n$ sufficiently large, we see that no algorithm that makes a single query can achieve a worst-case success probability of $\frac{1}{3} + \varepsilon$ for any $\varepsilon > 0$.

For an algorithm that makes two queries, we similarly assume without loss of generality that the queries yielded the values of $\hat{p}_1$ and $\hat{p}_2$ in some order. Regardless of whether the second query was made adaptively or nonadaptively, we will argue that, from these responses the algorithm cannot determine the winner with probability greater than $\frac{1}{2} + \frac{7}{n}$. As before, it suffices to show that the algorithm succeeds with probability at most $\frac{1}{2}$ when $\hat{p}_1, \hat{p}_2 \in [F_{n+2}, (n-1)F_{n+2}]$. Here there are two cases to consider, depending on which observed value is larger.

First suppose $\hat{p}_1 > \hat{p}_2$, and let $j \in \{3, 4, 5, \ldots, n\}$ be such that $\hat{p}_1 - \hat{p}_2 = F_j$. Then there are exactly two possible choices of the parameters $s$, $i$, and $c$ that could have led to the specific realizations $\hat{p}_1$ and $\hat{p}_2$:

- $r = 1$, $i = j - 2$, and $s = \hat{p}_1 - F_{i+2} \implies a$ is the winner.
- $r = 2$, $i = j - 2$, and $s = \hat{p}_1 - F_{i+2} \implies c$ is the winner.

Thus, $a$ and $c$ are equally likely to be the winner, so no matter which candidate the algorithm returns, it will be correct with probability at most $\frac{1}{2}$.

Now suppose $\hat{p}_1 < \hat{p}_2$, and let $j \in \{3, 4, 5, \ldots, n\}$ be such that $\hat{p}_2 - \hat{p}_1 = F_j$. Then there are exactly four possibilities for $s$, $i$, and $c$:

- $r = 3$, $i = j-1$, and $s = \hat{p}_1 - F_{i+i} \implies b$ is the winner. Note that this is the first place we make use of the Fibonacci recurrence: $\hat{p}_2 - \hat{p}_1 = (s + F_{i+2}) - (s + F_i) = F_{i+1} = F_j$.
- $r = 4$, $i = j$, and $s = \hat{p}_1 - F_{i+1} \implies a$ is the winner (again using the Fibonacci recurrence).
- $r = 5$, $i = j$, and $s = \hat{p}_1 \implies c$ is the winner.
- $r = 6$, $i = j - 1$, and $s = \hat{p}_1 \implies b$ is the winner.

Clearly, $b$ is the best guess here, but it is still only correct with probability $\frac{1}{2}$.

Thus, picking $n$ sufficiently large as before, we conclude that no algorithm making only two queries can achieve a worst-case success probability of $\frac{1}{2} + \varepsilon$ for any $\varepsilon > 0$. □

## 4.6 Discussion

Voting rules are increasingly applied to aggregate preferences across a large range of candidates, from primary elections to online opinions. This can, however, be at odds with the cognitive and implementation challenges that exist when requiring individuals to specify preferences across a large selection. Naturally, in such scenarios, voters end up specifying preferences over a limited set, whether by explicit design or implicitly by submitting incomplete votes. Our work studies the implications of this phenomenon on the computability of voting rules.

For the large class positional scoring rules, we provide an exact information-theoretic characterization of what can and cannot be correctly computed under this model. Specifically, a decrease in query size equivalently diminishes the dimension of the computable scoring vector space. We explicitly characterize these spaces, finding that, while the Borda count is included for $t \geq 2$, Plurality is not for *any* limited-sized query. We also extend this strong impossibility to STV. From a practical perspective, these results demonstrate the pitfalls of common rules like Plurality and STV within the setting incomplete votes and point to the space of alternative rules. Future work could go beyond voting and further investigate this question of computability with limited-sized queries for other social choice rules or other more general functions, such as committee selection with rankings.

For rules computable with at least $t^*$-sized queries, we also give bounds on the query complexity of any deterministic or randomized algorithm making $t \geq t^*$ sized queries. While we show that deterministic algorithms must cover the space of all $t^*$ sized queries in the worst-case, thus giving a tight bound, the picture for randomized algorithms is far less clear. We give an upper bound on the success probability when using a given number of queries, yet no known general-purpose algorithm achieves anything close to it. In Theorem 4.5.2, we close this gap for a special case of the Borda rule by constructing surprisingly intricate hard instances. Closing the general query complexity gap in our randomized setting is an intriguing open problem whose technical depth is illustrated by this result.

# 5

# Metric Distortion with Elicited Pairwise Comparisons

## 5.1 Introduction

As we have discussed in the preceding chapters, when aggregating a population's preferences to make a decision, we must contend with the fact that our access to preferences may be limited. One canonical example are platforms such as *Polis*, discussed in Chapter 3. Another includes RLHF, discussed in Part I. Recall, the goal is to choose model parameters that yield desirable outcomes according to the participants, and we typically learn what is "desirable'' by asking comparison queries, i.e., on prompt x, do you prefer output y or output z? However, notice that these responses only give us partial information about the individuals' true underlying preferences. It would be completely infeasible to learn, say, a ranking for each person over all possible prompt-output combinations. In both these examples, human input is a scarce resource that must be carefully allocated.

When selecting the queries to show participants, it is crucial to consider various constraints that a designer might need to satisfy. For example, on platforms like Polis, it is not guaranteed that we can interact with a user, then acquire more information from other users, and later return to the original user for a second round of queries.

For this, we may want algorithms that are *single-pass*: the queries asked to earlier users do not depend on queries to ones arriving later. Another desirable property a practitioner might want to satisfy is a notion of fairness, say, the queries asked for each agent to be independent of the responses of other agents; this can help ensure that interaction with the system remains unaffected by the arrival time. The same need may also arise from temporal concerns: if agents all answer simultaneously, we simply do not have data from the other agents. Conversely, we may want to have queries be independent of an agent's *own* answers, ensuring an unbiased interaction with the system. Therefore, our main research questions are:

> *How can we design a limited number of pairwise comparison queries to achieve efficient outcomes? Furthermore, how well can we do so under various restrictions such as limited rounds of interactions and which answers a query can depend on?*

To understand the effectiveness of various query choices, we need a way to measure the quality of the outcome. For this, we turn to the celebrated *distortion framework* [168]. Here, it is assumed that the comparison feedback submitted by the agents is derived from more expressive underlying preferences, typically scalar values. The more expressive values give rise to concrete objective values for each of the outcomes. The goal is to design a procedure that minimizes *distortion*, the worst-case ratio between the best possible objective value and the objective value of the chosen outcome. Without additional assumptions, achieving bounded distortion becomes impossible, even with complete knowledge of each agent's ranking. In this work, we make the frequently-made natural assumption that agents have costs induced by distances in an underlying metric space [7]. Beyond just a theoretical necessity, this restriction seems especially aligned with the motivating examples discussed.[1] To summarize, our goal is to optimize metric distortion using a limited number of pairwise comparison queries.

[1] In Polis, as part of the aggregation process, users and comments are mapped to a common low-dimensional feature space representing "opinion" space summarizing the various views. Within opinion space, voters are more likely to approve comments near them. Hence, Polis is implicitly assuming it is possible to map users/comments to a metric space. In RLHF, the possible models generating input/output prompts live in a feature space. An agent is assumed to give feedback according to a reward function of these features. If the reward function is concave, it is natural to assume that an agent has an ideal point maximizing their reward function and prefers inputs closer to this point.

### 5.1.1 Our approach and results

We assume there are $m$ alternatives, and we can ask each agent up to $t$ pairwise comparison queries. A pairwise comparison query (or simply a query) involves two alternatives that an agent is asked to rank. Note that if $t$ is quite large relative to $m$ (e.g., at least $m \log_2(m)$), we can learn the full ranking of an agent.[2] However, in applications like RLHF and Polis, $t$ is much smaller.

We first show that a query algorithm making at most $t$ queries to each agent, even if it can ask these in an arbitrary order with arbitrary dependencies, will incur distortion $\Omega(m/t)$ (Section 5.3.1).

Next, we consider algorithms that match this lower bound. We are especially interested in algorithms that require few rounds of interactions, as it may not be reasonable to assume that agents return later in the process after answering queries. Of particular importance is the extreme case: *single-pass* algorithms needing only one round of interaction. Strikingly, we show that there exists a single-pass algorithm that achieves optimal distortion of $O(m/t)$ (Section 5.3.2).

Finally, we consider other kinds of restrictions on the adaptivity of the algorithm, i.e., how the queries depend on previous answers. We focus on two types of adaptivity. The first, called *inter-agent* adaptivity, indicates whether the queries made to an agent can depend on responses from other agents. The second, called *intra-agent* adaptivity, indicates whether the queries made to an agent can depend on previous responses from that agent. Here, we devise nearly tight (up to logarithmic factors) upper and lower bounds for the optimal distortion achievable under all combinations of these restrictions (Section 5.4). The results are summarized in Table 5.1. Note that all upper bounds are induced by single-pass algorithms, while the lower bounds hold for all algorithms, regardless of the number of passes.

### 5.1.2 Related Work

The most closely related work is that of Anagnostides et al. [5], who also consider the metric distortion problem while eliciting preferences. However, the preference elicitation method differs from ours; they assume that an algorithm can query two candidates $\{a, b\}$ and, in response, learns which of the two candidates a majority of voters

[2]This is equivalent to using comparison-based queries to sort a list. Hence, this can be done via any sorting algorithm, e.g., MergeSort which takes at most $m \log_2(m)$ queries [56].

| Inter-agent | Intra-agent | |
|---|---|---|
| | Yes | No |
| Yes | $\Theta\left(\frac{m}{t}\right)$ | $O\left(\frac{m}{t} + \log m\right)$ , $\Omega\left(\frac{m}{t}\right)$ |
| No | $O\left(\frac{m^2 \log^2 t}{t^2}\right)$, $\Omega\left(\frac{m^2}{t^2}\right)$ | $\Theta\left(\frac{m^2}{t}\right)$ |

**Table 5.1:** Summary of results for different inter- and intra-agent adaptivity combinations. All upper bounds are achieved with single-pass rules. All lower bounds hold regardless of the number of passes.

prefer. While this could be implemented in our model, note that this would require multiple rounds of interaction and also makes additional restrictions, e.g., all voters will be asked the same queries, and only the majority winner is given, rather than the magnitude of this majority. Hence, we are interested in similar questions under a different query model, which we view as more suited for our motivating applications.

Slightly further afield is considering metric distortion, where only partial information about each agent's ranking is given. For example, the same work [5] builds on work by Kempe [117] where voters reveal their top-$k$ candidates for a fixed value $k \leq m$. Kempe [117] also considers the more general problem of $b$ bits of information being communicated about a ranking using a communication scheme fixed upfront.

Metric distortion itself (using complete rankings) was introduced by Anshelevich et al. [6]. They show that no social choice rule has distortion better than 3, while the well-known Copeland rule achieves distortion 5, among bounds for other well-known rules. Followup work proved bounds for even more rules [85, 189], but none beat this barrier of 5. Significant progress was made by Munagala and Wang [151], who introduced a rule achieving distortion 4.236. The gap was finally closed by Gkatzelis et al. [84], who gave a rule achieving distortion 3. Anshelevich et al. [8] provide a survey of these results, as well as progress in non-metric distortion more broadly.

In addition to improvements in lowering distortion, other progress has been made. Kempe [118] gives an LP-duality framework for analyzing the distortion of existing rules. Kizilkaya and Kempe [127] and Kizilkaya and Kempe [128] both provide new rules achieving optimal metric distortion requiring much simpler analyses. We build on techniques from Kempe [118] as well as the rule of Kizilkaya and Kempe [127]; a more in-depth description can be found when we use them.

Preference elicitation, more broadly, has a rich history in computational social

choice with several variations. A more general summary can be found in Brandt et al. [36]; however, we focus on the most closely related works here. Lu and Boutilier [139] has the most similar style pairwise queries to our model; however, their objective is quite different from metric distortion. Amanatidis et al. [4] study the tradeoff of distortion and query complexity; however, their queries are quite dissimilar, directly eliciting cardinal values and relative magnitudes rather than information about the ranking. Preference elicitation questions are also often studied through the lens of communication complexity, i.e., the number of bits that need to be revealed rather than the number of queries of a specific type. For example, Conitzer and Sandholm [54] give the communication complexity of determining the winner of several common voting rules (where each agent privately knows their ranking). This has also been applied to non-metric distortion, where voters can reveal a certain number of bits about their utilities rather than just the rankings [141, 142].

## 5.2 Model

For $s \in \mathbb{N}$, define $[s] = \{1, 2, \dots, s\}$.

Metric Voting. We consider the setting of single-winner elections where we have a set $C$ of $m$ candidates and a set $V = [n]$ of $n$ voters. We assume there is a (pseudo)metric[3] over $C \sqcup V$ characterized by a distance function $d : (C \sqcup V) \times (C \sqcup V) \to \mathbb{R}_{\geq 0}$. That is, for all $i, j, k \in C \sqcup V$, we have that $d(i, i) = 0$, $d(i, j) = d(j, i)$, and $d(i, j) \leq d(i, k) + d(k, j)$ (the triangle inequality). Furthermore, each voter $i \in V$ has a *preference ranking* $\sigma_i = \sigma_i(1) \succ_i \cdots \succ_i \sigma_i(m)$ such that $a \succ_i b$ implies that $d(i, a) \leq d(i, b)$ — we allow ties to be broken arbitrarily. We write $a \succeq_i b$ for $a \succ_i b$ or $a = b$. The collection of all voters' preference rankings form a *preference profile* $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$. An *instance* is a tuple $(d, \boldsymbol{\sigma})$.

Utilitarianism. Given an instance $(d, \boldsymbol{\sigma})$, the *social cost* (or simply cost) of a candidate $c \in C$ is defined as $\text{cost}_d(c) := \sum_{i \in V} d(i, c)$. When $d$ is clear from context, we may drop it from the notation by simply writing $\text{cost}(c)$. Under this measure,

[3] A pseudometric is a generalization of a metric that allows two distinct points to be distance 0 from each other. In other words, we allow voters and candidates to be at the same location.

the optimal candidate is the one minimizing social cost. We will use the notation $\mathrm{opt}(d, \boldsymbol{\sigma}) := \min_{c \in C} \mathrm{cost}_d(c)$ for the minimum cost.

Algorithms with Pairwise Comparison Queries. We consider algorithms that do not have direct access to the underlying metric nor the preference profile; rather, they are only aware of $C$ and $V$ and can learn of voters' preferences via a sequence of *pairwise comparison queries*. That is, the algorithm can choose a voter $i$ and two candidates $\{a, b\}$ and learn which candidate voter $i$ prefers, $a \succ_i b$ or $b \succ_i a$. After receiving the responses, the algorithm (i.e., a voting rule) returns a candidate $c$ as the winner. More formally, by a slight abuse of notation, we let $\mathcal{A}(d, \boldsymbol{\sigma})$ denote the candidate returned by an algorithm $\mathcal{A}$ on instance $(d, \boldsymbol{\sigma})$.

We are interested in algorithms that do not make too many queries to each voter. If an algorithm makes at most $t$ queries to each voter $i$, we call it a *$t$-query* algorithm.

Distortion. The quantitative measure of efficiency we investigate in this work is *distortion*. For a given instance $(d, \boldsymbol{\sigma})$, the distortion of candidate $c$ is defined as

$$\mathrm{dist}(c \mid (d, \boldsymbol{\sigma})) \;=\; \frac{\mathrm{cost}_d(c)}{\mathrm{opt}(d, \boldsymbol{\sigma})},$$

i.e., the ratio of the social cost of $c$ to the optimal social cost. This value is always at least one, and the lower the ratio, the more efficient the candidate. The distortion of an algorithm $\mathcal{A}$ is defined as

$$\mathrm{dist}_m(\mathcal{A}) \;=\; \sup_{(d,\boldsymbol{\sigma}),|C|=m} \mathrm{dist}(\mathcal{A}(d, \boldsymbol{\sigma}) \mid (d, \boldsymbol{\sigma})),$$

where the supremum is over all instances with $m$ candidates and any number of voters. In other words, the distortion of an algorithm $\mathcal{A}$ is its worst-case guarantee in terms of the multiplicative approximation to the optimal social cost.

### 5.2.1 Algorithm Restrictions

In the setup described so far, an algorithm can make queries to voters in an arbitrary order, with the option to revisit voters it has already queried or choose each subsequent query based on arbitrary prior responses. This flexibility may be costly or even impractical in certain scenarios. Hence, we are interested in the optimal distortion

of algorithms subject to various constraints. We describe the restrictions of interest below.

Number of passes. We will say an algorithm is *p-pass* if the sequence of queries can be partitioned into at most $p$ contiguous sequences where agent indices are non-decreasing. We will call it *single-pass* if it is 1-pass.

Inter-agent adaptivity. An algorithm is *inter-agent nonadaptive* if the queries it makes to voter $i$ do not depend on responses from other voters $i' \neq i$. In other words, the choice of queries can be decomposed into $n$ subprocedures $A_1, \ldots, A_n$ such that $A_i$ only makes queries and learns the responses of voter $i$ (the aggregation given these responses can be arbitrary). Note that an inter-agent nonadaptive algorithm can always be implemented as a single-pass algorithm because we can run these procedures sequentially (first $A_1$, then $A_2$, and so on).

Intra-agent adaptivity. An algorithm is *intra-agent nonadaptive* if the queries asked to voter $i$ do not depend on responses to previous queries to that agent.

Full nonadaptivity. If an algorithm is inter- and intra-agent nonadaptive, we will call it *fully nonadaptive*. Such an algorithm can be implemented such that all queries are fixed upfront without depending on any of the responses.

## 5.3 $t$-Query Algorithms

We begin by considering optimal distortion bounds for $t$-query algorithms. First, we establish a lower bound for any algorithm. Then, we present an algorithm that not only matches this lower bound but is also single-pass.

### 5.3.1 Lower Bound

We start by establishing a lower bound on distortion when each agent can answer up to $t$ queries without any other restrictions on the algorithm.

**Theorem 5.3.1.** *All $t$-query algorithms incur a distortion of at least $\frac{m}{t} - 1$.*

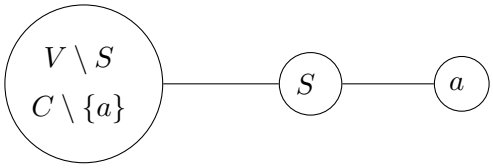



**Figure 5.1:** Metric for the case where the algorithm returns $a$.

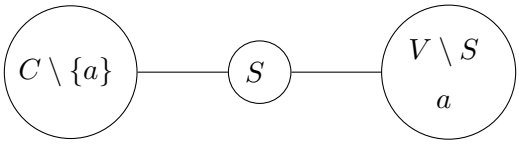



**Figure 5.2:** Metric for the case where the algorithm does not return $a$.

In fact, we show an even stronger result. This holds for any algorithm that makes at most $tn$ queries in total *regardless* of how they are distributed among the agents. Any algorithm making $t$ queries per voter trivially makes at most $tn$ queries.

*Proof.* Let $R = \{a_1 \succ_{i_1} b_1, a_2 \succ_{i_2} b_2, \ldots\}$ with $|R| \leq tn$ be an arbitrary sequence of responses the algorithm has queried and learned, and let $c$ be the candidate returned given these responses. Since there are at most $2|R| \leq 2tn$ candidate appearances in $|R|$, there exists a candidate $a \in C$ that appears in less than $2tn/m$ many queries. Let $S \subseteq V$ be the set of voters that were ever queried on $a$. We have that $|S| \leq 2tn/m$.

We do a case analysis based on $c = a$ or $c \neq a$. The metrics are shown in Figures 5.1 and 5.2, where the distance function is defined by the path length. First, we show that there are profiles for both metrics that are consistent with the responses in $R$. In both metrics, voters in $S$ are equidistant from all candidates, and voters in $V \setminus S$ are equidistant from all candidates except $a$. We construct the profile as follows. We choose rankings for voters in $S$ in an arbitrary way consistent with $R$. For the voters in $V \setminus S$, depending on the case, we either place $a$ at the top or bottom of their ranking and complete the rest of the ranking arbitrarily in a way consistent with $R$. Since no voter in $V \setminus S$ was queried on $a$, this is always possible, and by the equal distances, these profiles are consistent with the metrics.

*Case $c = a$.* Consider the metric in Figure 5.1. Note that $\text{cost}(a) = 2(|V| - |S|) + |S| = 2n - |S|$. For any candidate $c' \in C \setminus \{a\}$, $\text{cost}(c') = |S|$. Therefore,

$$\text{dist}(c) \geq \frac{\text{cost}(c)}{\text{cost}(c')} = \frac{2n - |S|}{|S|} \geq \frac{m}{t} - 1.$$

*Case $c \neq a$.* Consider the metric in Figure 5.2. Note that $\text{cost}(a) = |S|$ and, since $c \neq a$, $\text{cost}(c) = |S| + 2(|V| - |S|)$. As in the previous case, we get a lower bound of $m/t - 1$. □

### 5.3.2 A Matching Upper Bound

Next, we consider designing algorithms to match this lower bound. Recall that using $m \log_2 m$ queries, it is possible to learn an agent's entire ranking. Hence, if $t$ is sufficiently large ($\geq m \log_2 m$), we can simply learn all of the agents' complete rankings and then run any constant-distortion rule such as *Copeland* [6] or *Plurality Veto* [127].

In fact, Plurality Veto can be implemented using $O(m)$ queries. As this will be useful to our later analysis, we describe the rule and query implementation in more detail. Plurality Veto runs in two phases. In the first, it determines the *plurality score* of each candidate, i.e., the number of voters that ranks that candidate first. The second begins with each candidate having a number of points equal to their plurality score. It sequentially goes through the voters (in an arbitrary order), and for each, asks the voter their *least favorite* candidate who still has a nonzero number of points; it then decrements that candidate by one point. The winning candidate is the last one remaining just before the final voter is asked to decrement.[4]

Note that both determining a favorite or least favorite candidate among a set $S$ requires $|S| - 1$ pairwise queries. Hence, this can be implemented by asking each voter at most $2m - 2$ pairwise queries. Therefore, when $t \geq 2m - 2$, Plurality Veto has optimal distortion.

However, there are still two drawbacks to this result. First, even though Plurality Veto requires only $O(m)$ queries, it needs to do this in two phases. Hence, its naïve implementation requires two passes. We may wonder if the same can be done with a single-pass algorithm. Second, and more important, this gives us no insight for how to handle $t < 2m - 2$, a crucial case for many applications. Our next result handles both of these matters, devising a single-pass algorithm that yields asymptotically optimal distortion for all $t$.

[4]Note that the total number of points given out is $n$, and each voter decreases by one. Hence, just before the final voter is asked, exactly one candidate will remain with a single point.

**Algorithm 4** Single-Pass Plurality Veto

Initialize $\text{plu}(c) = 0$
**for** each voter $i$ among the first $\lceil |V|/2 \rceil$ voters **do**
    $c_i \leftarrow \text{top}(\sigma_i, C)$
    $\text{plu}(c_i) \leftarrow \text{plu}(c_i) + 1$
**for** each voter $i$ among the next $\lceil |V|/2 \rceil - 1$ voters **do**
    $C \leftarrow \{c \in C \mid \text{plu}(c) \geq 1\}$
    $c'_i \leftarrow \text{bottom}(\sigma_i, C)$
    $\text{plu}(c'_i) \leftarrow \text{plu}(c'_i) - 1$
**return** the unique candidate $a$ with $\text{plu}(a) \geq 1$.

**Algorithm 5** $t$-Query Single-Pass Adaptive

$R \leftarrow \lceil (m-1)/t \rceil$;  $C_{\text{active}} \leftarrow C$
**for** each round $r = 1$ to $R$ **do**
    $C_r \leftarrow \min\{t+1, |C_{\text{active}}|\}$ candidates from $C_{\text{active}}$
    $V_r \leftarrow$ the next $\lfloor n/R \rfloor$ voters
    $c_r^* \leftarrow$ Algorithm 4$(C \leftarrow C_r, V \leftarrow V_r)$
    $C_{\text{active}} \leftarrow (C_{\text{active}} \setminus C_r) \cup \{c_r^*\}$
**return** the unique candidate $a \in C_{\text{active}}$

**Theorem 5.3.2.** *For $n \geq \lceil \frac{m-1}{t} \rceil$ and $t \leq m-1$, Algorithm 5 is a single-pass, $t$-query algorithm achieving distortion $6\lceil \frac{m-1}{t} \rceil + 1$.*

The formal proof of the theorem appears later in this section. Algorithm 5 uses Algorithm 4 as a subroutine on subsets of candidates over a sequence of $\lceil \frac{m-1}{t} \rceil$ rounds.[5] This subroutine is essentially a variant of Plurality Veto which runs in a single pass. In the special case of $t = m-1$, there is only one round, and the subroutine is used "as is" on the entire candidate set. To better illustrate the ideas, we first prove lemmas that are sufficient to show Algorithm 4 achieves the desired distortion bound in this special case. We then build on these ideas to handle the general case of $t \leq m-1$.

[5]Note that the $n \geq \lceil \frac{m-1}{t} \rceil$ requirement simply ensures that we can use at least one distinct voter per round.

### Warm-up Case: $t = m - 1$

In this case, Algorithm 5 simply runs Algorithm 4 and returns the chosen candidate. Hence, we directly analyze Algorithm 4.

*Algorithm Description.* Algorithm 4 uses two sequences of adaptive queries:

- $\text{top}(\sigma_i, C')$: A sequence of adaptive queries that finds the most preferred candidate of voter $i$ among a subset of candidates $C' \subseteq C$. This can be done using $|C'| - 1$ queries (compare the first two candidates, then compare the preferred one with the third candidate, and so on).
- $\text{bottom}(\sigma_i, C')$ : Similar to above but for the least preferred alternative of voter $i$ among $C' \subseteq C$.

Algorithm 4 accumulates plurality scores of the first $\lceil |V|/2 \rceil$ voters in $\text{plu}(c)$ (lines 2–4). Then, for each of the second $\lceil |V|/2 \rceil - 1$ voters, among the remaining candidates with positive plurality score $C_{\text{active}}$, the voter indicates her least preferred candidate $c'_i$ by decreasing the candidate's plurality score by one (lines 5-8). Since the number of increments is exactly one more than the number of decrements, only one candidate remains with a positive score, which subsequently is returned.

*Distortion Analysis.* To prove the distortion bound, we need the following technical definition and lemmas.

**Definition 5.3.3** ($(a,b)$-path)**.** *Fix two candidates $a, b \in C$. A sequence of distinct voters $v_1, \ldots, v_k$ is called a* voter path from $a$ to $b$ *(or an $(a,b)$-*path *for short) if there exists a sequence of $k+1$ not necessarily distinct candidates beginning with $a$ and ending with $b$, $a = c_0, c_1, \ldots, c_{k-1}, c_k = b$ such that, for each $j \in [k]$, $c_{j-1} \succeq_{i_j} c_j$. Slightly abusing notation, we will refer to a set of voters $V' \subseteq V$ as an $(a,b)$-path if some order of $V'$ is an $(a,b)$-path. We will say two $(a,b)$-paths are disjoint if they do not share any voters.*

The following is essential to the distortion analysis of Algorithms 4 and 5.

**Lemma 5.3.4.** *Fix $a, b \in C$. Suppose there exist $\ell$ disjoint $(a,b)$-paths. Then,*

$$\frac{cost(a)}{cost(b)} \leq \frac{2n}{\ell} + 1.$$

*Proof.* Fix $a, b \in C$. We first show a useful property about $(a, b)$-paths. Namely, if $V'$ is an $(a, b)$-path, then,

$$d(a, b) \leq 2 \cdot \sum\nolimits_{i \in V'} d(i, b). \tag{5.1}$$

Indeed, let $k := |V'|$, $i_1, \ldots, i_k$ be the order of agents in $V'$, and $a = c_0, \ldots, c_k = b$ be the corresponding candidate sequence. The inequality follows from the following:

$$\begin{aligned}
d(a, b) &= d(c_0, c_k) - d(c_k, c_k) && (d(c_k, c_k) = 0) \\
&= \sum_{j=1}^{k} (d(c_{j-1}, c_k) - d(c_j, c_k)) \\
&\leq \sum_{j=1}^{k} (d(c_{j-1}, i_j) + d(i_j, c_k) - d(c_j, c_k)) && \text{(Triangle inequality)} \\
&\leq \sum_{j=1}^{k} (d(c_j, i_j) + d(i_j, c_k) - d(c_j, c_k)) && (c_{j-1} \succeq_{i_j} c_j) \\
&\leq \sum_{j=1}^{k} (d(i_j, c_k) + d(i_j, c_k)) && \text{(Triangle inequality)} \\
&= 2 \sum_{j=1}^{k} d(i_j, b).
\end{aligned}$$

With Inequality (5.1) in hand, we are ready to prove the lemma. Let $V_1, \ldots, V_\ell$ be the disjoint voter sets of the $\ell$ $(a, b)$-paths. By the disjointness of $V_j$'s we have

$$\text{cost}(b) = \sum_{i \in V} d(i, b) \geq \sum_{j=1}^{\ell} \sum_{i \in V_j} d(i, b) \geq \sum_{j=1}^{\ell} \frac{d(a, b)}{2} = \ell \cdot \frac{d(a, b)}{2},$$

where the first inequality follows because of disjointness, and the second from Inequality (5.1). Using the triangle inequality, we have $\text{cost}(a) \leq \sum_{i \in V} (d(i, b) + d(a, b)) = \text{cost}(b) + n \cdot d(a, b)$. Hence,

$$\frac{\text{cost}(a)}{\text{cost}(b)} \leq \frac{\text{cost}(b) + n \cdot d(a, b)}{\text{cost}(b)} = 1 + n \cdot \frac{d(a, b)}{\text{cost}(b)} \leq 1 + n \cdot \frac{d(a, b)}{\ell \cdot d(a, b)/2} = \frac{2n}{\ell} + 1. \quad \square$$

The definition of $(a, b)$-paths and the result of Lemma 5.3.4 are similar to the definition of chains of preferences and Corollary 5.3 of Kempe [118]. The key difference

is that we require the voters along each path to be distinct, which allows us to prove stronger properties.

Next, using Lemma 5.3.4, we show that there are $\lceil n/2 \rceil$ disjoint voter paths from the candidate returned by Algorithm 4 to each other candidate.

**Lemma 5.3.5.** *There exist $\lceil |V|/2 \rceil$ disjoint voter paths from the candidate selected by Algorithm 4 to any other candidate.*

*Proof.* Let $a$ be the candidate returned by Algorithm 4 and $b \in C \setminus \{a\}$ be an arbitrary candidate. Let $V_1$ be the first $\lceil |V|/2 \rceil$ voters initializing the plurality scores and $V_2$ be the second set of voters decreasing the scores. Note that $|V_1| - 1 = |V_2|$. Create a one-to-one function $g : V_2 \to V_1$ as follows. As per the algorithm, for $i \in V_1$, $c_i$ is the most favorite candidate for $i$; and, for $i \in V_2$, $c'_i$ is the candidate indicated by $i$ as her least favorite among the still active candidates (line 7). Map $i$ to a corresponding voter $g(i)$ increasing the plurality score of $c'_i$. This way, $g$ is injective and $c'_i = c_{g(i)}$.

Next, we show there exist $\lceil n/2 \rceil$ disjoint voter paths from $a$ to $b$. First, we claim that for each $i \in V_2$, $\{i, g(i)\}$ is an $(a, b)$-path via the candidate sequence $a, c'_i, b$. Indeed, for the first comparison, $a \succ_i c'_i$ because $i$ indicates $c'_i$ as her least preferred alternative among the active set which necessarily contains $a$ as it remained active until the end. For the second, $c'_i \succ_{g(i)} b$ because $c'_i = c_{g(i)}$, and $c_{g(i)}$ was $g(i)$'s most preferred alternative. In addition, consider the single voter $i^* \in V_1 \setminus \{g(i') \mid i' \in V_2\}$ who does not appear in the range of $f$. Her top vote must have been $a$ as $a$ is the only candidate with positive score in the end. Thus, $a \succ_{i^*} b$ and the set $\{i^*\}$ alone is a voter path from $a$ to $b$. Note that each of the above voter paths are disjoint, so we have constructed $|V_2| + |\{i^*\}| = \lceil |V|/2 \rceil$ disjoint voter paths from $a$ to $b$. □

As an aside, note that if we directly apply Lemma 5.3.4 to Lemma 5.3.5, this shows that Algorithm 4 achieves distortion 5.

### General Case: $t \leq m - 1$

Now, we analyze Algorithm 5 in full generality.

*Algorithm Description.* Algorithm 5 runs in $R = \lceil (m-1)/t \rceil$ rounds. In each round $r \in [R]$, it takes up to $t + 1$ candidates from the active candidate set $C_{\text{active}}$, denoted $C_r$. It then runs Algorithm 4 with a new set of $\lfloor n/R \rfloor$ voters, denoted $V_r$, along with the candidates in $C_r$, which return a candidate $c^*_r \in C_r$. Then, Algorithm 5 removes

all candidates in $C_r$ from the active candidate set $C_{\text{active}}$ except for $c_r^*$. Because of this removal procedure, in each of the first $R-1$ rounds, it eliminates $t$ candidates from the active set. In the final round, there are at most $m-(R-1)t \le 1+Rt-(R-1)t \le t+1$ candidates remaining, from which a single winning candidate is kept in the active set and subsequently returned.

*Distortion Analysis.* To utilize Lemma 5.3.4 similar to that in Lemma 5.3.5, we find a sufficient number of disjoint voter paths from the candidate returned by Algorithm 5 to each of the other candidates.

*Proof of Theorem 5.3.2.* Let $a$ be the candidate returned by Algorithm 5 and $b \in C \setminus \{a\}$ be an arbitrary candidate.

Consider a directed graph $G$ over candidates described as follows. At each round $r$, draw a directed edge from the winning candidate $c_r^*$ to all candidates $C_r \setminus \{c_r^*\}$. Note that each candidate $c \neq a$ has exactly one incoming edge, i.e., if $c$ is eliminated at round $r$, the directed edge $c_r^* \to c$ is in $G$. Thus, $G$ forms a directed rooted tree at $a$. Let $a = c_0 \to \ldots \to c_k = b$ be the unique directed path in $G$ from $a$ to $b$. For all $j \in [k]$, the edge $c_{j-1} \to c_j$ is created in a different round $r_j$. By Lemma 5.3.5, there are $\ell := \lceil \lfloor n/R \rfloor / 2 \rceil$ voter paths from $c_{j-1}$ to $c_j$ on a disjoint set of voters $V_{r_j}$. We have that $\ell \ge n/(3R)$ because $n \ge R$ and $\lceil \lfloor x \rfloor / 2 \rceil \ge x/3$ for $x \ge 1$. Note that $V_{r_j}$'s are disjoint as the algorithm selects a new set of voters at each round. Thus, we can extend the $\ell$ paths from $c_0$ to $c_1$ with the $\ell$ paths from $c_1$ to $c_2$ by arbitrarily matching those two sets of paths and get $\ell$ disjoint paths from $c_0$ to $c_2$. Then, we extend the paths from $c_0 = a$ to $c_3$ and so on to $c_k = b$. Therefore, by Lemma 5.3.4 and that this holds for all $b \in C \setminus \{c\}$ we have

$$\text{dist}(a) \le \frac{2n}{\ell} + 1 \le 6R + 1 = 6 \left\lceil \frac{m-1}{t} \right\rceil + 1. \qquad \square$$

where in the last inequality we used $n \ge R = \left\lceil \frac{m-1}{t} \right\rceil$ and that $\lfloor x \rfloor \ge \frac{x}{2}$ for $x \ge 1$.

## 5.4 Algorithms with Adaptivity Constraints

We now turn to varying restrictions on algorithm adaptivity. We begin with a tight analysis of the fully nonadaptive setting (Section 5.4.1). Then, we relax the constraints

by requiring only one of inter- and intra-agent nonadaptivity at a time (Sections 5.4.2 and 5.4.3).

Note that when algorithms are intra-agent nonadaptive (Sections 5.4.1 and 5.4.3), we can no longer learn each agent's full ranking with $O(m \log m)$ queries. To do so, we instead may have to ask about all $\binom{m}{2} = \Theta(m^2)$ possible pairwise comparisons. Hence, even for $t = \omega(m \log m)$, we may not be able to achieve constant distortion. Furthermore, recall that any inter-agent nonadaptive algorithms can trivially be implemented to be single-pass.

### 5.4.1 Fully Nonadaptive Algorithms

We show that the optimal distortion value for $t$-query fully nonadaptive algorithms is $\Theta(m^2/t)$. This is a factor of $m$ worse than the unrestricted setting.

**Theorem 5.4.1.** *All fully nonadaptive $t$-query algorithms incur a distortion of at least $\frac{m^2}{t} - 1$. Furthermore, for $t \leq \binom{m}{2}$ and $n \geq \binom{m}{2}/t$, there exists a fully nonadaptive $t$-query algorithm bound with distortion at most $3m^2/t + 1$.*

Below, we outline the algorithm and then analyze its distortion.

*Algorithm Description.* There are $\binom{m}{2}$ pairs of candidates. The algorithm distributes queries among voters as equally is possible in a way that each pair of candidates $(c, c') \in \binom{C}{2}$ is given to at least $\lfloor nt/\binom{m}{2} \rfloor$ voters. This can be done in various ways. For instance, concatenate $\lceil nt/\binom{m}{2} \rceil$ copies of the list of all $\binom{m}{2}$ pairs and assign the $i$th voter the pairs in range $[(i-1)t + 1, (i+1)t]$. Such a distribution can be determined beforehand and, thus, is fully nonadaptive.

The algorithm aggregates the results as follows. For each pair $c, c' \in C$, draw a directed edge from $c$ to $c'$ if a majority of voters queried on $\{c, c'\}$ preferred $c$ to $c'$ (in the case of a tie, pick an arbitrary direction). Note that this is now a *tournament* graph, where between every two nodes, there is one directed edge. Finally, from this graph, select a *king* vertex $a$. A vertex is a king if it reaches all other vertices by at most two edges. In other words, $a$ is a king if for all other candidates $b \in C \setminus \{a\}$, either $a$ has an edge to $b$ or there exists another candidate $c$ such that $a$ has an edge to $c$ and $c$ has an edge to $b$. It is well-known that king vertices exist in tournament graphs (e.g., West [205]).

*Distortion Analysis.* To prove Theorem 5.4.1, we utilize the following lemma derived in Kempe [118] (a special case of Corollary 5.3).

**Lemma 5.4.2** ([118]). *If at least $\ell$ voters prefer $a$ to $b$, then $\frac{cost(a)}{cost(b)} \le \frac{2n}{\ell} - 1$. Furthermore, if there is a candidate $c \in C \setminus \{a, b\}$ such that $f$ voters prefer $a$ to $c$, and also $\ell$ voters prefer $c$ to $b$, then $\frac{cost(a)}{cost(b)} \le \frac{2n}{\ell} + 1$.*

*Proof of Theorem 5.4.1.* Let $a$ be the king vertex in the pairwise-majority graph described above. Fix a candidate $b \in C \setminus \{a\}$. Since $a$ is a king vertex, there is a path of length at most 2 from $a$ to $b$. Each edge indicates the existence of $\lceil \frac{1}{2} \lfloor nt / \binom{m}{2} \rfloor \rceil$ voters preferring the candidate on the tail of the edge to the candidate on the head. Since $n \ge \binom{m}{2}/t$ and $\lceil \frac{1}{2} \lfloor x \rfloor \rceil \ge x$ for $x \ge 1$, $\lceil \frac{1}{2} \lfloor nt / \binom{m}{2} \rfloor \rceil \ge nt/(3\binom{m}{2}) \ge (2nt)/(3m^2)$. By Lemma 5.4.2, we have $\mathrm{dist}(a) \le 3m^2/t + 1$. □

### 5.4.2 Inter-Agent Nonadaptive Algorithms

Moving away from the most restricted case of full nonadaptivity, we find that by allowing intra-agent adaptivity, we can design algorithms that achieve better distortion by a factor of roughly $\widetilde{O}(t)$ (up to logarithmic factors).

**Theorem 5.4.3.** *For $t \le m \log_2 m$ and $n \ge 2m^2 \log_2^2(t+1)/t^2$, there exists an inter-agent nonadpative $t$-query algorithm achieving a distortion of $\frac{10m^2 \log_2^2(t+1)}{t} + 1$.*

*Proof.* Let $\ell = \max\{\ell' \mid (2\ell') \log_2(2\ell') \le t\}$, so that we can sort a group of $2\ell$ candidates using up to $t$ queries. We first claim that $\ell \ge \frac{t}{2\log_2(t+1)}$. Indeed, plugging in $\ell' = \frac{t}{2\log_2(t+1)}$, we have

$$(2\ell') \log_2(2\ell') = \frac{t}{\log_2(t+1)} \log\left(\frac{t}{\log_2(t+1)}\right) \le \frac{t}{\log_2(t+1)} \log(t) \le t.$$

Further, note that $2\ell \le m$ because $t$ queries are only sufficient to sort at most $m$ candidates.

Partition the candidate set into $\lceil m/\ell \rceil$ groups of size $\le \ell$ denoted by $\mathcal{C} = \{C_1, \ldots, C_{\lceil m/\ell \rceil}\}$. Divide the voters as equally as possible among pairs $(C_{j_1}, C_{j_2}) \in \binom{\mathcal{C}}{2}$, i.e., each pair of candidate groups is assigned at least $\lfloor n / \binom{|\mathcal{C}|}{2} \rfloor$ voters. Note that

$$\begin{aligned} \binom{|\mathcal{C}|}{2} &= \frac{\lceil m/\ell \rceil \cdot (\lceil m/\ell \rceil - 1)}{2} \\ &\le \frac{(\lceil m/\ell \rceil - 1/2)^2}{2} \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\left(\frac{5m}{4\ell}\right)^2}{2} && \lceil x \rceil - 1/2 \leq \tfrac{5x}{4} \text{ for } x \geq 2 \\
&= \frac{25m^2}{32\ell^2} \\
&\leq \frac{25m^2 \log_2^2 t}{16t^2}.
\end{aligned}
$$

The voters assigned to $(C_{j_1}, C_{j_2})$ will be queried to learn their full preference ranking over $C_{j_1} \cup C_{j_2}$. This way, all pairs of candidates $c, c' \in C$ are compared by at least $\lfloor n/\binom{|\mathcal{C}|}{2} \rfloor$ voters.

Similar to the algorithm for Theorem 5.4.1, we create the majority pairwise comparison graph and select a king node of this complete directed graph. Each edge of this graph indicates $\lceil \frac{1}{2} \cdot \lfloor n/\binom{|\mathcal{C}|}{2} \rfloor \rceil$ voters preferring one candidate to another. Note that since $n \geq \frac{2m^2 \log_2^2 t}{t^2}$, $n/\binom{|\mathcal{C}|}{2} \geq 1$, so $\lceil \frac{1}{2} \cdot \lfloor n/\binom{|\mathcal{C}|}{2} \rfloor \rceil \geq \frac{n}{3\binom{|\mathcal{C}|}{2}}$ since $\lceil \frac{1}{2} \lfloor x \rfloor \rceil \geq x/3$ for all $x \geq 1$.

Thus, by Lemma 5.4.2, the king vertex of this graph achieves a distortion of at most

$$
\frac{2n}{\frac{n}{3\binom{|\mathcal{C}|}{2}}} + 1 \leq 6\binom{|\mathcal{C}|}{2} + 1 \leq \frac{10m^2 \log_2^2(t+1)}{t} + 1. \qquad \square
$$

We complement the above theorem by showing its optimality up to logarithmic factors with a surprisingly-intricate lower bound.

**Theorem 5.4.4.** *All inter-agent nonadaptive algorithms have distortion at least* $m^2/(4t^2)$.

*Proof.* Fix an inter-agent nonadaptive, $t$-query algorithm $\mathcal{A}$. Without loss of generality, suppose that $\mathcal{A}$ always makes exactly $t$ queries to each agent (if it does not, we can add arbitrary queries at the end and ignore the responses). Since the queries made by the algorithm to agent $i$ can only depend on responses from $i$, we can decompose the query-making portion of $\mathcal{A}$ into $n$ subroutines, $A_1, \ldots A_n$. The subroutine $A_i$ decides which pair of candidates to query agent $i$ on next, given their previous responses. We can represent each $A_i$ as a binary decision tree with height $t+1$. Each internal node is labelled by a pair of candidates. Depending on the response of which candidate is preferred, either the left or right branch is followed.[6] Note that some-

[6]This is very similar to representations of comparison-based sorting algorithms used to

times only one of the two branches can be induced by a ranking, e.g., if a certain response would lead to a cycle in the preferences. In such cases, a node has only one child.

Each ranking $\sigma$ that agent $i$ could have corresponds to a path in this tree from the root to a leaf, where, at each internal node, the branch is taken based on the comparison query response according to $\sigma$. Note that if two rankings $\sigma$ and $\sigma'$ follow the same path down the tree, they are indistinguishable according to $A_i$, i.e., they lead to the exact same queries asked and responses received.

We will say a pair of candidates $\{a, b\} \subseteq C$ *deceive* $A_i$ if there are two rankings $\sigma$ and $\sigma'$ such that $\sigma(1) = \sigma'(2) = a$, $\sigma(2) = \sigma'(1) = b$, and $\sigma$ and $\sigma'$ follow the same path in $A_i$. In other words, there is a run of the algorithm $A_i$ such that it is plausible that $a$ is ranked first and $b$ is second or that $b$ is first and $a$ is second, and these two scenarios are indistinguishable. The property that $\{a, b\} \subseteq C$ *deceive* $A_i$ is equivalent to there exists a path in $A_i$ such that neither $a$ nor $b$ lose in a pairwise comparison. Indeed, taking any ranking corresponding to this path and then moving $a$ to the top and $b$ second or $a$ second and $b$ to the top will lead to the exact same responses and correspond to the same path.

Let $D_i = \{\{a, b\} \mid \{a, b\} \text{ deceive } A_i\}$ be the set of pairs that deceive $A_i$ and let $E_i = \overline{D_i}$ be its complement, i.e., the remaining pairs that *do not* deceive $A_i$. We will show that $|E_i| \leq t^2$, that is, there are at most $t^2$ pairs that do not deceive $A_i$.

Consider an arbitrary path down the decision tree in $A_i$. Let $T$ be the set of candidates that lost a pairwise comparison at least once on this path. Since the path is of length $t$ and each node results in one candidate losing, $|T| \leq t$. By the above definition, if $\{a, b\} \in E_i$, we have that $\{a, b\} \cap T \neq \emptyset$. In other words, every pair in $E_i$ *must* contain at least one member of $T$. Indeed, if $\{a, b\} \cap T = \emptyset$, then this is a path in which neither lost $a$ nor $b$ lost. Let $E_i^a = \{b \mid \{a, b\} \in E_i\}$ be the set of candidates who are paired with $a$ in $E_i$. The above implies that $|E_i| \leq \sum_{a \in T} |E_i^a|$.

Fix some $a^* \in T$. We next show that $|E_i^{a^*}| \leq t$. Combined with the above argument, this implies that $|E_i| \leq |T| \cdot t \leq t^2$. Let $\sigma$ be an arbitrary ranking with $\sigma(1) = a^*$, and consider the path induced by $\sigma$. Let $T^{a^*}$ be the set of candidates who lose at least once on this path. We have that $a^* \notin T^{a^*}$ because $a^*$ is ranked first in $\sigma$. Furthermore, note that if $b \in E_i^{a^*}$, then $b \in T^{a^*}$. Indeed, if not, the above path would have neither $a^*$ nor $b$ losing, contradicting that $\{a^*, b\} \in E_i$. This means that

prove lower bounds on query complexity.

$E_i^{a^*} \subseteq T^{a^*}$. Finally, notice that $|T^{a^*}| \leq t$ because the path is of length $t$. This implies the desired bound.

Since $|E_i| \leq t^2$, and $|D_i| + |E_i| = \binom{m}{2}$, this implies that $|D_i| \geq \binom{m}{2} - t^2$. By an averaging argument, there must be a pair $\{a, b\}$ that deceives at least $\frac{\binom{m}{2}-t^2}{\binom{m}{2}} \cdot n = (1 - \frac{2t^2}{m(m-1)}) \cdot n$ of the algorithms $A_i$.

Fix such a set $\{a, b\}$. Let $V' = \{i \mid \{a, b\} \in D_i\}$ be the set of agents whose algorithms are deceived. By the above, we have that $|V'| \geq (1 - \frac{2t^2}{m(m-1)}) \cdot n$. For each $i \in V'$, let $\sigma_i^{ab}$ and $\sigma_i^{ba}$ be a pair of rankings indistinguishable to $A_i$ such that $\sigma_i^{ab}(1) = \sigma_i^{ba}(2) = a$ and $\sigma_i^{ab}(2) = \sigma_i^{ba}(1) = b$.

We can now construct a pair of profiles $\boldsymbol{\sigma}^{ab}$ and $\boldsymbol{\sigma}^{ba}$ where, in the first, each $i \in V'$ has ranking $\sigma_i^{ab}$, and in the second, each $i \in V'$ has ranking $\sigma_i^{ba}$. The remaining voters $V \setminus V'$ will have arbitrary rankings (but the same in both profiles). Note that $\boldsymbol{\sigma}^{ab}$ and $\boldsymbol{\sigma}^{ba}$ lead to the exact same queries and responses. Hence, the chosen candidate must be the same on both. Therefore, for at least one of $a$ and $b$ was *not* chosen to be winner. Without loss of generality, suppose it was $a$.

We will consider the profile $\boldsymbol{\sigma}^{ab}$ along with the following metric. Each $i \in V'$ has $d(i, a) = 0$. For all remaining pairs of voters and candidates $(i', c)$, we will have $d(i', c) = 1$. Note that these distances are consistent with $\boldsymbol{\sigma}^{ab}$. Furthermore, the cost for $a$ is at most $|V \setminus V'| \leq \frac{2t^2}{m(m-1)} \cdot n$, while the cost for any other candidate is $n$. Therefore, the distortion on this instance is at least

$$\frac{n}{\frac{2t^2}{m(m-1)} \cdot n} \geq \frac{m^2}{4t^2},$$

as needed. □

### 5.4.3 Intra-Agent Nonadaptive Algorithms

We now turn to designing $t$-query algorithms that are *intra-agent* nonadaptive, i.e., each voter is asked up to $t$ pairwise comparison queries all at once and returns her answers to all the $t$ queries together. Its distortion guarantee is as follows.

**Theorem 5.4.5.** *For $n \geq \lfloor m/t \rfloor + \lceil \log_2 t \rceil$ and $t \leq m - 1$, Algorithm 6 is a single-pass, intra-agent nonadaptive, $t$-query algorithm that achieves a distortion of at most $6(\lfloor m/t \rfloor + \lceil \log_2 t \rceil) + 1$.*

**Algorithm 6** $t$-Query Single-Pass Intra-agent Nonadaptive

$R \leftarrow \lfloor m/t \rfloor + \lceil \log_2 t \rceil$; $C_{\text{active}} \leftarrow C$
**for** each round $r = 1$ to $R$ **do**
    $\ell_r \leftarrow \min\{t, \lfloor |C_{\text{active}}|/2 \rfloor\}$
    Let $C_r \leftarrow \{(a_{r,1}, b_{r,1}), (a_{r,2}, b_{r,2}), \ldots, (a_{r,\ell_r}, b_{r,\ell_r})\}$ be $\ell_r$ pairs of distinct candidates from $C_{\text{active}}$
    **for** each voter $i$ among the next $\lfloor n/R \rfloor$ voters **do**
        Present the $\ell_r$ pairs $(a_{r,1}, b_{r,1}), \ldots, (a_{r,\ell_r}, b_{r,\ell_r})$
    **for** each pair in $\{(a_j, b_j) \mid j \in [\ell_r]\}$ **do**
        Let $c_{r,j}$ be the candidate losing the majority of comparisons (ties broken arbitrarily)
        $C_{\text{active}} \leftarrow C_{\text{active}} \setminus \{c_{r,j}\}$
**return** the unique candidate $a \in C_{\text{active}}$

Algorithm 6 runs in $R$ rounds (for $R$ that will be determined later) querying an equal number of $\lfloor n/R \rfloor$ new voters in each. For round $r \in [R]$, the corresponding group of voters are queried on $\ell_r$ disjoint pairs of candidates, arbitrarily selected from the active candidate set $C_{\text{active}}$. For each of the $\ell_r$ pairs, the candidate losing the majority of the $\lfloor n/R \rfloor$ votes is eliminated from $C_{\text{active}}$. This process goes on until one candidate remains, which is subsequently returned.

Since we want at most $t$ queries made to each voter, in each round, we can ask about at most $t$ pairs (i.e., $\ell_r \leq t$). While $|C_{\text{active}}| \geq 2t$, there are enough candidates such that the algorithm can, in fact, make all $t$ queries (i.e., $\ell_r = t$). However, after $\lfloor m/t \rfloor - 1$ rounds, fewer than $2t$ candidates will remain. Then, the algorithm pairs up as many candidates as possible in $C_{\text{active}}$ (one candidate will be unpaired when $|C_{\text{active}}|$ is odd), and $\ell_r = \lfloor |C_{\text{active}}|/2 \rfloor$ candidates are eliminated. This can go on for at most $\lceil \log_2 2t \rceil \leq \lceil \log_2 t \rceil + 1$ many rounds until a single candidate remains. Hence, there will be at most $\lfloor m/t \rfloor + \lceil \log_2 t \rceil$ rounds.

To prove the distortion bound, we again utilize Lemma 5.3.4 by finding a sufficient number of disjoint voter paths from the chosen candidate to the other candidates.

*Proof of Theorem 5.4.5.* Let $a$ be the candidate returned by Algorithm 6 and $b \in C \setminus \{a\}$ be an arbitrary candidate. Create the elimination directed graph $G$ as follows. For each candidate $c \neq a$, there is a candidate $g(c)$ that won over $c$ in a pairwise majority comparison by $\lfloor n/R \rfloor$ voters; draw a directed edge from $g(c)$ to $c$. Since each

candidate has a single incoming edge, $G$ forms a directed tree rooted at $a$.

There is a unique path from $a = c_0 \to c_1 \ldots \to c_k = b$. For all $j \in [k]$, let $V_j$ be the set of voters $V_j$ who made the $(c_{j-1}, c_j)$ comparison resulting in elimination of $c_j$. The sets $V_j$'s are disjoint since $c_{j-1}$ must be eliminated after $c_j$ is eliminated (in an earlier round). Let $V_j' \subseteq V_j$ be the voters preferring $c_{j-1}$ to $c_j$. For each $i \in V_j'$, there is a voter path from $c_{j-1}$ to $c_j$ on $\{i\}$. Hence, there are $|V_j'| \geq \lceil \lfloor n/R \rfloor / 2 \rceil$ many disjoint voter paths from $c_{j-1}$ to $c_j$. We can connect the voter paths from $c_0$ to $c_1$ and $c_1$ to $c_2$ by arbitrarily matching and extending them to voter paths $c_0$ to $c_2$. We can continue to make voter paths from $c_0 = a$ to $c_k = b$. All of these paths consist of unique sets of voters. By Lemma 5.3.4 and that the above holds for all $b \in C$, we have

$$\text{dist}(a) = \max_{b \in C} \frac{\text{cost}(a)}{\text{cost}(b)} \leq \frac{2n}{\lceil \lfloor n/R \rfloor / 2 \rceil} + 1 \leq \frac{6n}{\frac{n}{R}} + 1 = 6R + 1 = 6(\lfloor m/t \rfloor + \lceil \log_2 t \rceil) + 1,$$

where the last inequality holds by $n \geq R = \lfloor m/t \rfloor + \lceil \log_2 t \rceil$ and $\lfloor x \rfloor \geq x/2$ for $x \geq 1$. □

While the lower bound of Theorem 5.3.1 shows the optimality of the above result for $t = O(m/\log m)$, it is unclear what the best achievable distortion is for $t = \omega(m/\log m)$. Theorem 5.4.5 continues to give a $O(\log m)$ upper bound, but the lower bound becomes $O(m/t)$, and only constant once $t = \Omega(m)$. For $t = \Omega(m^2)$, we know it is possible to achieve $O(1)$ distortion because, at this point, it is possible to learn all agents' complete rankings and run any of the well-known constant-distortion rules. However, we leave it as an interesting open question to close this gap for intermediate values of $t$, or find the minimum $t$ with constant distortion.

## 5.5 Discussion

To summarize, we have studied the design of a limited number of pairwise queries to optimize metric distortion. We considered a variety of constraints on our algorithms regarding both the rounds of interactions and the dependency of queries on previous answers. In all cases, we provide either tight or nearly tight upper and lower bounds. In the case of inter- and intra-agent nonadaptivity, there remain logarithmic gaps.

As for follow-up work, the most direct would be to close these gaps. However, more broadly, while requiring few rounds of interaction and both inter- and intra-agent nonadaptivity capture an interesting set of constraints on the algorithms, there are a plethora of reasonable notions one could consider. For example, we may wish for

the algorithm to be *anonymous*, demanding that each agent's queries depend only on their ranking, $\sigma_i$, rather than their identity. An even stricter property is to require that all agents receive the *exact* same queries, ensuring a strong notion of equality in how different agents interact with the system. We leave optimizing distortion in these settings and the development of new desirable properties as intriguing directions for future work.

Lastly, in certain applications, there can be such an enormous number of alternatives that our bounds become completely impractical. For instance, in variations of Polis, the domain of alternatives can be extended beyond just submitted comments to *any* statement from a natural language, a seemingly limitless domain. In such cases, when it is impossible to ensure that each candidate is queried by at least one voter, achieving bounded distortion is hopeless. We need to either (i) make additional assumptions about the domain of alternatives (say, they arise from a low-dimensional feature space), (ii) consider queries that are more powerful than pairwise comparisons, or (iii) use other measures of efficiency. We again leave these as exciting directions for further work.

# Part III

# Liquid Democracy

# 6

# Tracking Truth with Liquid Democracy

## 6.1 Introduction

*Liquid democracy* is a voting paradigm that is conceptually situated between *direct democracy*, in which voters have direct influence over decisions, and *representative democracy*, where voters choose delegates who represent them for a period of time. Under liquid democracy, voters have a choice: they can either vote directly on an issue like in direct democracy, or delegate their vote to another voter, entrusting them to vote on their behalf. The defining feature of liquid democracy is that these delegations are *transitive*: if voter 1 delegates to voter 2 and voter 2 delegates to voter 3, then voter 3 votes (or delegates) on behalf of all three voters.

In recent years, liquid democracy has gained prominence around the world. The most impressive example is that of the German Pirate Party, which adopted the *LiquidFeedback* platform in 2010 [129]. Other political parties, such as the Net Party in Argentina and Flux in Australia, have run on the wily promise that, once elected, their representatives would be essentially controlled by voters through a liquid democracy platform. Companies are also exploring the use of liquid democracy for corporate governance; Google, for example, has run a proof-of-concept experiment [101]. Blockchain systems have also been experimenting with related weighted decentralized voting systems [26, 137].

Practitioners, however, recognize that there is a potential flaw in liquid democracy, namely, the possibility of *concentration of power*, in the sense that certain voters amass a relatively large number of delegations, giving them pivotal influence over the final decision. This scenario seems inherently undemocratic — and it is not a mere thought experiment. Indeed, in the LiquidFeedback platform of the German Pirate Party, a linguistics professor at the University of Bamberg received so many delegations that, as noted by Der Spiegel,[1] his "vote was like a decree."

Kahng et al. [115] examine liquid democracy's concentration-of-power phenomenon from a theoretical viewpoint and establish a troubling impossibility result in what has been called the *epistemic* setting, that is, one where there is a ground truth.[2] Informally, they demonstrate that, even under the strong assumption that voters delegate only to more "competent" voters, any "local mechanism" satisfying minimal conditions will, in certain instances, be subject to concentration of power, leading to relatively low accuracy. More specifically, Kahng et al. model the problem as a decision problem where voters decide on an issue with two outcomes, $\{0, 1\}$, where 1 is correct (the ground truth) and 0 is incorrect. Each of the voters $i \in \{1, \ldots, n\}$ is characterized by a *competence* $p_i \in [0, 1]$. The binary vote $V_i$ of each voter $i$ is drawn independently from a Bernoulli distribution, that is, each voter votes correctly with probability $p_i$. Under direct democracy, the outcome of the election is determined by a majority vote: the correct outcome is selected if and only if more than half of the voters vote for the correct outcome; that is, it is correct if and only if $\sum_{i=1}^{n} V_i \geq n/2$. Under liquid democracy, there exists a set of weights, $\mathsf{weight}_i$ for each $i \in [n]$, which represent the number of votes that voter $i$ gathered transitively after delegation (if voter $i$ delegates, then $\mathsf{weight}_i = 0$). The outcome of the election is then determined by a weighted majority; it is correct if and only if $\sum_{i=1}^{n} \mathsf{weight}_i V_i \geq n/2$.

Kahng et al. also introduce the concept of a *delegation mechanism*, which determines whether voters delegate and, if so, to whom they delegate. They are especially interested in *local mechanisms*, where the delegation decision of a voter depends only on their local neighborhood according to an underlying social network. They assume that voters delegate only to those with strictly higher competence, which excludes the possibility of cyclic delegations. To evaluate liquid democracy, Kahng et al. test the

[1] See https://tinyurl.com/y52j6nfs.

[2] The use of the term "epistemic" in this context is well-established in the social choice literature [138, 162].

intuition that society makes more informed decisions under liquid democracy than under direct democracy (especially given the foregoing assumption about upward delegation). To that end, they define the *gain* of a delegation mechanism to be the difference between the probability the correct outcome is selected under liquid democracy and the probability the correct outcome is selected under direct democracy. A delegation mechanism satisfies *positive gain* if its gain is strictly positive in some cases, and it satisfies *do no harm* if, for all $\varepsilon > 0$, its gain is at least $-\varepsilon$ for sufficiently large instances. Assuming that competence after delegation remains strictly above 1/2, this will follow from the law of large number that applies to the weighted majority with weights relatively spread out [91]. The main result of Kahng et al. is that local mechanisms can never satisfy these two requirements. Caragiannis and Micha [40] further strengthen this negative result by showing that there are degenerate instances where local mechanisms perform much worse than either direct democracy or dictatorship (the most extreme concentration of power).[3]

These results undermine the case for liquid democracy: the benefits of delegation appear to be reversed by concentration of power. However, the negative conclusion relies heavily on worst-case modeling assumptions. Our research represents a significant advance as it offers a comprehensive framework that not only captures the worst-case scenarios of previous works, but also provides insights into more intriguing "high probability" cases. In particular, in this paper, we provide a new theoretical model and extensive experiments that show that liquid democracy will typically satisfy a probabilistic version of *positive gain* and *do no harm* under minimal assumptions.

### 6.1.1 Our Contributions and Techniques

Our contributions are the following. First, building on the work of Kahng et al. [115], we provide a general framework to analyze the stochastic network dynamics of transitive delegations that captures liquid democracy's intricate interactions between local-delegation choices and global properties. Second, we identify large classes of delega-

[3]The former constructs an instance where even with arbitrarily many voters, a constant number will receive a majority of the delegations. The group has an average competence above 1/2. The probability liquid democracy gives the right answer can be upper bounded by a constant strictly below 1, while direct democracy is correct with probability approaching one. In the latter case, the voters' numbers and relative competence are chosen so that liquid democracy almost always gives the incorrect answer (as does direct voting), while dictatorship is correct with a constant probability.

tion models where liquid democracy performs well, in that delegations induce a sufficiently small amount of concentration of power and liquid democracy almost surely results in correct outcomes. Along the way, we prove new high-probability bounds on the size of the largest component in an infinite Pólya urn process;[4] this result may be of independent interest. Finally, we conduct the first series of lab experiments on liquid democracy that can test epistemic performance. This involved over $11,000$ votes from 168 participants in six experimental groups, where each group had pre-existing social ties. Our novel experimental design allows us to compare the performance of direct and liquid democracy, as well as to analyze properties of real voter delegation behavior. Importantly, the behaviors we observe align with one of the models we introduce, thus lending support to this approach. Taken together, these results exhibit a regime in which liquid democracy displays promising performance. We next elaborate on some of our specific techniques.

### Stochastic Delegations

Our point of departure from the existing literature is the way we model delegation in liquid democracy. To emphasize these differences, instead of calling these delegation functions *mechanisms*, we instead call them delegation *models*, as they are intended to capture independent voter behavior rather than prescribing to each voter to whom they must delegate. Our delegation models are defined by $M = (q, \varphi)$, where $q : [0,1] \to [0,1]$ is a function that maps a voter's competence to the probability they delegate and $\varphi : [0,1]^2 \to \mathbb{R}_{\geq 0}$ maps a pair of competencies to a weight. In this model, each voter $i$ votes directly with probability $1 - q(p_i)$ and, conditioned on delegating with probability $q(p_i)$, delegates to voter $j \neq i$ with probability proportional to $\varphi(p_i, p_j)$. These delegation functions do not model explicit reasoning; rather, they model behaviors that may be influenced by tacit knowledge captured by $q$ and $\varphi$. A voter does not need to "know" the competence of another voter to decide whether to delegate; rather, the delegation *probabilities* are influenced by competence, as captured by $\varphi$note that delegation cycles are possible, and we take a worst-case approach to dealing with them: If the delegations form a cycle, then all voters in the cycle are

[4] An infinite Pólya urn process models an urn process where each new ball picks its urn with a probability proportional to the size of the urn or creates its own urn with constant probability.

assumed to be incorrect (vote 0).[5]

The most significant difference between our model of delegation and that of Kahng et al. [115] is that in our model, each voter has a chance of delegating to any other voter, whereas in their model, an underlying social network restricts delegation options. Our model captures a connected world where, in particular, voters may have heard of experts on various issues even if they do not know them personally. Although our model eschews an explicit social network, it can be seen as embedded into the delegation process, where the probability that $i$ delegates to $j$ takes into account the probability that $i$ is familiar with $j$ in the first place. Another difference between our model and that of Kahng et al. [115] is that we model the competencies $p_1, \ldots, p_n$ as being sampled independently from a distribution $\mathcal{D}$. While this assumption is made mainly for ease of exposition, it allows us to avoid edge cases and obtain robust results.

### Delegation Models

Our goal is to identify delegation models that satisfy (probabilistic versions of) positive gain and do no harm. Our first technical contribution, in Section 6.2.1, is the formulation of general conditions on the model and competence distribution that are sufficient for these properties to hold (Lemma 6.2.3). In particular, to achieve the more difficult do no harm property, we present conditions that guarantee the maximum weight $\mathsf{max\text{-}weight}(G_n)$ accumulated by any voter is sub-linear with high probability and the expected increase in competence after delegation is at least a positive constant times the population size. These conditions prevent extreme concentration of power and ensure that the representatives after delegation are sufficiently better than the entire population to compensate for any concentration of power that does happen.

Although the proof is straightforward, the benefit of this lemma is that it then suffices to identify models and distribution classes that verify these conditions. A delegation model $M$ and a competence distribution $\mathcal{D}$ induce a distribution over delegation instances that generates random graphs in ways that relate to well-known graph processes, which we leverage to analyze our models. Specifically, we introduce three models, all shown to satisfy do no harm and positive gain under *any* continuous distribution over competence levels. The first models, *upward delegation* and *confidence-based*

[5] In LiquidFeedback, delegation cycles are, in fact, ignored.

*delegation*, are interesting but restricted case studies that demonstrate the robustness of our approach. By contrast, the *general continuous delegation* model is, as the name suggests, quite general. Moreover, it is realistic: its predictions are consistent with our experiments.

Upward Delegation: In Section 6.3, we consider a model according to which the probability $p$ of delegation is exogenous and constant across competencies, $q(p_i) = p$, and delegation can occur only to voters with strictly higher competence (the weight that any voter $i$ puts on another voter $j$ is $\varphi(p_i, p_j) = \mathbb{I}_{\{p_j - p_i > 0\}}$. This model captures the fact that there might be some reluctance to delegate regardless of the voter's competence but does assume that voters act in the interest of society by only delegating to voters that are more competent than they are.

To generate a random graph induced by such a model, one can add a single voter at a time in order of decreasing competence and allow the voter to either not delegate (with probability $1 - p$) and create their own disconnected component, or delegate to the creator of any other component with probability proportional the size of the component. This works because delegating to any voter in the previous components is possible (since they have strictly higher competence) and would result in the votes being concentrated in the originator of that component by transitivity. Such a process exactly generates a preferential attachment graph with a positive probability of not attaching to the existing components, also called an infinite Pólya urn process [187]. We can then show that, with high probability, no component grows too large so long as $p < 1$ (see Section 6.1.1 for an overview of this step). Further, continuity of the competence distribution ensures that enough lower competence voters delegate to higher competence voters to sufficiently increase the average.

Confidence-Based Delegation. In Section 6.4, we consider a model in which voters delegate with probability decreasing in their competencies and choose someone at random when they delegate. That is, the probability $q(p_i)$ that any voter $i$ delegates is decreasing in $p_i$ and the weight that any voter $i$ gives to any voter $j$ is $\varphi(p_i, p_j) = 1$. In other words, in this model, competence does not affect the probability of receiving delegations, only the probability of delegating.

To generate a random graph induced by such a model, one can begin from a random vertex and study the delegation tree that starts at that vertex. A delegation tree

is defined as a branching process, where a node $i$'s "children" are the nodes that delegated to node $i$. In contrast to classical branching processes, the probability for a child to be born increases as the number of people who already received delegations decreases. Nevertheless, we prove that, with high probability, as long as a delegation tree is no larger than $O(\log n)$, our heterogeneous branching process is dominated by a sub-critical graph branching process [3]. We can then conclude that no component has size larger than $O(\log n)$ with high probability. Next, we show that the expected competence among the voters that do not delegate is strictly higher than the average competence.

General Continuous Delegation. Finally, we consider a general model in Section 6.5 where the likelihood of delegation is fixed and the weight assigned to each voter when delegating is increasing in their competence. That is, each voter $i$ delegates with probability $q(p_i) = p$ and the weight that voter $i$ places on voter $j$ is $\varphi(p_i, p_j)$, where $\varphi$ is continuous and increases in its second coordinate. Thus, in this model, the delegation distribution is slightly skewed towards more competent voters.

To generate a random graph induced by such a model, we again consider a branching process, but now voters $j$ and $k$ place different weights on $i$ per $\varphi$. Therefore, voters have a *type* that governs their delegation behavior; this allows us to define a multi-type branching process with types that are continuous in $[0, 1]$. The major part of the analysis is a proof that, with high probability, as long as the delegation tree is no larger than $O(\log n)$, our heterogeneous branching process is dominated by a sub-critical Poisson multi-type branching process. In a manner similar to Confidence-Based Delegation, we also show that there is an expected increase in competence after delegation.

### Component Sizes in Infinite Pólya Urn Processes

Recall that to prove that upward delegation satisfies do no harm, we show that the largest component in an infinite Pólya urn process is sub-linear with high probability (Lemma 6.3.2). We briefly expand on the proof as this result was, to the best of our knowledge, not previously known in the random graph literature, and may be of independent interest. We begin by focusing on the first $t^\gamma$ bins (for a suitably chosen $\gamma$ depending on the attachment probability $p$) and derive an upper bound on the ex-

pected size of these bins. This allows us to use Markov's inequality and union bound over all bins to show that simultaneously all of them are sublinear in size with high probability.

Second, we take care of the remaining bins by observing that each additional bins's growth is isomorphic to a classic Pólya urn process with two bins, whose limiting dynamic follows a Beta distribution. We analyze the rate of convergence, which allows us to give sufficiently strong bounds using Chebyshev's inequality after exactly $t - t^\gamma$ steps, and union bound over all of these bins, concluding that all are sublinear with high probability.

### Consistency With Experiments

Lastly, we conduct six experiments to statistically estimate the functions $q$ and $\varphi$, and test the overall effectiveness of liquid democracy. Participants were presented with several yes or no questions on various topics. We call the set of questions related to each topic a *task*. For each task, participants could either choose to vote directly or delegate their vote (for all questions) to another participant. They only saw the questions in a task if they chose to vote directly. In a later phase, they were asked to answers the questions they had delegated (and not seen) to see how they would have voted. This setup allows us to do a few things. First, it induces a matched-pair design where, for each task and experiment, we can compare the accuracy of voting under liquid and direct democracy. Second, we use the answers to all questions to estimate participants' competencies. Using this information, we study how delegation behavior depends on competence and investigate whether it is consistent with the theoretical findings.

Results suggest that (i) competence is inversely correlated with the chance of delegation, and (ii) the likelihood of delegating to another voter increases with their competence. The results, therefore, support the assumptions and predictions made by the confidence-based and general continuous-delegation models. Taken together, these results exhibit a regime in which liquid democracy is overall more likely to pinpoint the truth than direct democracy.

### 6.1.2 Related work

The most closely related paper is that of Kahng et al. [115], which was discussed in detail above. It is worth noting, though, that they complement their negative result with a positive one: when the mechanism can restrict the maximum number of delegations (transitively) received by any voter to $o(\sqrt{\log n})$, do no harm and positive gain are satisfied. Imposing such a restriction would require a central planner that monitors and controls delegations. Gölz et al. [86] build on this idea: they study liquid democracy systems where voters may nominate multiple delegates and a central planner chooses a single delegate in order to minimize the maximum weight of any voter. Similarly, Brill and Talmon [37] introduces a process that allows voters to specify ordinal preferences over delegation options and possibly restricting or modifying delegations in a centralized way. Caragiannis and Micha [40], and then Becker et al. [22] also consider central planners; they show that, for given competencies, the problem of choosing among delegation options to maximize the probability of a correct decision is hard to approximate. In any case, implementing these proposals would require a fundamental rethinking of the practice of liquid democracy. By contrast, our positive results show that *decentralized* delegation models can be inherently self-regulatory, which supports the effectiveness of the current practice of liquid democracy.

More generally, there has been a significant amount of theoretical research on liquid democracy in recent years. To give a few examples: Green-Armytage [88] studies whether it is rational for voters to delegate their vote from a utilitarian viewpoint; Christoff and Grossi [50] examine a similar question but in the context of voting on logically interdependent propositions; Bloembergen et al. [31], Zhang and Grossi [212] and Dhillon et al. [62] study liquid democracy from a game-theoretic viewpoint.

Next, our work builds on the random graph literature, as our delegation processes are related to well-known stochastic graph processes. Upward delegation can be viewed as a generalization of the preferential attachment model where agents do not attach to the existing component(s) with a fixed probability. Classical preferential attachment models assume that a new node attaches to an existing node $n_0$ with probability (parameterized by an *attachment function*) depending on the degree of $n_0$ [20, 64]. In our framework, a new component may be created with sfixed probability, a setup introduced by Simon [187] and usually referred to as an infinite Pólya urn process. Others have studied the distribution of degrees [63], the distribution of the number of

components with $k$ people at time $t$ [51], and the conditions for the emergence of infinite components [53]. However, to the best of our knowledge, the existing results do not allow us to derive bounds on the size of the largest component with high probability after a finite amount of time.

In terms of our experiments on liquid democracy, ours is the first paper to conduct experiments with human subjects. Previous papers have studied different aspects of liquid democracy through experiments in corporate [101] and political environments Independent of and essentially cooncurrent with our work, Campbell et al. [39] tested a game-theoretic formulation of liquid democracy. Unlike our experiments, they used online platforms to gather participants who did not know each other. Participants were assigned a probability of being correct and asked whether they would want to delegate to others, with experts (those with the highest probability of being correct) being publicly known. The delegations were randomly assigned to the pre-determined experts in one set-up, and through the random dot kinematogram task in another one. The group sizes considered are 5 people with one expert, 15 people with 3 experts and 125 people with 25 experts. While this study reveals interesting connections between individuals' perceived competence and delegation behavior, it cannot investigate how experts are (or are not) identified endogenously through interpersonal knowledge embedded in a social networks, since the participants do not know each other.

Last, our work relates to recent advances in managerial studies that consider novel forms of governance, such as corporate governance [e.g., 106], blockchain technologies [e.g., 26, 137], and prediction markets [e.g., 13, 47].

## 6.2 Model

There is a set of $n$ voters, denoted $[n] = \{1, \dots, n\}$. We assume voters are making a decision on a binary issue with possible answers 0 and 1; there is a correct alternative (1) and an incorrect alternative (0). Each voter $i$ has a *competence level* $p_i \in [0, 1]$ which is the probability that $i$ votes correctly. We denote the vector of competencies by $\vec{p}_n = (p_1, \dots, p_n)$. When $n$ is clear from the context, we sometimes drop it from the notation.

Delegation graphs A *delegation graph* $G_n = ([n], E)$ on $n$ voters is a directed graph with voters as vertices and a directed edge $(i, j) \in E$ denoting that $i$ delegates

their vote to $j$. Again, if $n$ is clear from context, we occasionally drop it from the notation. The outdegree of a vertex in the delegation graph is at most 1 since each voter can delegate to at most one person. Voters that do not delegate have no outgoing edges. In a delegation graph $G_n$, the *delegations received* by a voter $i$, $\mathsf{dels}_i(G_n)$, is defined as the total number of people that (transitively) delegated to $i$ in $G_n$, (i.e., the total number of ancestors of $i$ in $G_n$). The *weight* of a voter $i$, $\mathsf{weight}_i(G_n)$, is $\mathsf{dels}_i(G_n) + 1$ (the number of delegation they received plus their own weight) if $i$ votes directly, and 0 if $i$ delegates. We define $\mathsf{max\text{-}weight}(G_n) = \max_{i\in[n]} \mathsf{weight}_i(G_n)$ to be the largest weight of any voter and define $\mathsf{total\text{-}weight}(G_n) = \sum_{i=1}^{n} \mathsf{weight}_i(G_n)$. Since each vote is counted at most once, we have that $\mathsf{total\text{-}weight}(G_n) \leq n$. However, note that if delegation edges form a cycle, then the weight of the voters on the cycle and voters delegating into the cycle are all set to 0 and hence will not be counted. In particular, this means that $\mathsf{total\text{-}weight}(G_n)$ may be strictly less than $n$.[6]

Delegation instances We call the tuple $(\vec{p}_n, G_n)$ a *delegation instance*, or simply an instance, on $n$ voters. Let $V_i = 1$ if voter $i$ would vote correctly if $i$ did vote, and $V_i = 0$ otherwise. Fixed competencies $\vec{p}_n$ induce a probability measure $\mathbb{P}_{\vec{p}_n}$ over the $n$ possible binary votes $V_i$, where $V_i \sim \text{Bern}(p_i)$. Given votes $V_1, \ldots, V_n$, we let $X_n^D$ be the number of correct votes under direct democracy, that is, $X_n^D = \sum_{i=1}^{n} V_i$. We let $X_{G_n}^F$ be the number of correct votes under liquid democracy with delegation graph $G_n$, that is, $X_{G_n}^F = \sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot V_i$. The probability that direct democracy and liquid democracy are correct are $\mathbb{P}_{\vec{p}_n}[X_n^D > n/2]$ and $\mathbb{P}_{\vec{p}_n}[X_{G_n}^F > n/2]$, respectively.

Gain of a delegation instance We define the *gain* of an instance as

$$\text{gain}(\vec{p}_n, G_n) = \mathbb{P}_{\vec{p}_n}[X_{G_n}^F > n/2] - \mathbb{P}_{\vec{p}_n}[X_n^D > n/2].$$

In words, it is the difference between the probability that liquid democracy is correct and the probability that majority is correct.

Randomization over delegation instances In general, we assume that both competencies and delegations are chosen randomly. Each voter's competence $p_i$ is

[6] This is a worst-case approach where cycles can only hurt the performance of liquid democracy, since this assumption is equivalent to assuming that all voters on the cycles vote incorrectly.

sampled i.i.d. from a fixed distribution $\mathcal{D}$ with support contained in $[0,1]$. Delegations will be chosen according to a *model* $M$. A model $M = (q, \varphi)$ is composed of two parts. The first $q : [0,1] \to [0,1]$ is a function that maps competencies to the probability that the voter delegates. The second $\varphi : [0,1]^2 \to \mathbb{R}_{\geq 0}$ maps pairs of competencies to a weight. A voter $i$ with competence $p_i$ will choose how to delegate as follows:

- With probability $1 - q(p_i)$ they do not delegate.
- With probability $q(p_i)$, $i$ delegates; $i$ places weight $\varphi(p_i, p_j)$ on each voter $j \neq i$ and randomly sample another voter $j$ to delegate to proportional to these weights. In the degenerate case where $\varphi(p_i, p_j) = 0$ for all $j \neq i$, we assume that $i$ does not delegate.

A competence distribution $\mathcal{D}$, a model $M$, and a number $n$ of voters induce a probability measure $\mathbb{P}_{\mathcal{D},M,n}$ over all instances $(\vec{p}_n, G_n)$ of size $n$.

We can now redefine the *do no harm (DNH)* and *positive gain (PG)* properties from Kahng et al. [115] in a probabilistic way.

**Definition 6.2.1** (Probabilistic do no harm)**.** *A model $M$ satisfies* probabilistic do no harm *with respect to a class $\mathfrak{D}$ of distributions if, for all distributions $\mathcal{D} \in \mathfrak{D}$ and all $\varepsilon, \delta > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,*

$$\mathbb{P}_{\mathcal{D},M,n}[\text{gain}(\vec{p}_n, G_n) \geq -\varepsilon] > 1 - \delta.$$

**Definition 6.2.2** (Probabilistic positive gain)**.** *A model $M$ satisfies* probabilistic positive gain *with respect to a class $\mathfrak{D}$ of distributions if there exists a distribution $\mathcal{D} \in \mathfrak{D}$ such that for all $\varepsilon, \delta > 0$, there exists $n_0 \in N$ such that for all $n \geq n_0$,*

$$\mathbb{P}_{\mathcal{D},M,n}[\text{gain}(\vec{p}_n, G_n) \geq 1 - \varepsilon] > 1 - \delta.$$

7

[7]Note that positive gain and do no harm relate to the notion of concentration of the weighted sum $\sum_{i=1}^{n} \mathsf{weight}_i V_i$. Indeed, the probability of direct democracy being correct approaches 1 as $n$ increases when the average competence is strictly above 1/2. As a result, do no harm is satisfied by a delegation model exactly when the probability that liquid democracy is correct also approaches 1. This happens when the competence after delegation remains strictly above 1/2, and the weighted sum $\sum_{i=1}^{n} \mathsf{weight}_i V_i$ concentrates. Positive gain also holds if there exists a setup where the average group competence is strictly below 1/2, and

### 6.2.1 Core Lemma

Next, we give a key lemma, which provides sufficient conditions for a model $M$ to satisfy probabilistic do no harm and probabilistic positive gain with respect to a class $\mathfrak{D}$ of distributions. This lemma will form the basis of all of our later results.

**Lemma 6.2.3.** *If $M$ is a model, $\mathfrak{D}$ a class of distributions, $n$ a number of persons, and for all distributions $\mathcal{D} \in \mathfrak{D}$, there is an $\alpha \in (0,1)$ and $C : \mathbb{N} \to \mathbb{N}$ with $C(n) \in o(n)$ such that*

$$\mathbb{P}_{\mathcal{D},M,n}\left[\mathsf{max\text{-}weight}(G_n) \leq C(n)\right] = 1 - o(1) \tag{6.1}$$

$$\mathbb{P}_{\mathcal{D},M,n}\left[\sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \sum_{i=1}^{n} p_i \geq 2\alpha n\right] = 1 - o(1), \tag{6.2}$$

*then $M$ satisfies probabilistic do no harm. If in addition, there exists a distribution $\mathcal{D} \in \mathfrak{D}$ and an $\alpha \in (0,1)$ such that*

$$\mathbb{P}_{\mathcal{D},M,n}\left[\sum_{i=1}^{n} p_i + \alpha n \leq n/2 \leq \sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \alpha n\right] = 1 - o(1), \tag{6.3}$$

*then $M$ satisfies probabilistic positive gain.*

*Proof.* We establish the two properties separately.

*Probabilistic do-no-harm:* We first show that a model $M$ that satisfies conditions (6.1) and (6.2) satisfies probabilistic do no harm. Fix an arbitrary competence distribution $\mathcal{D} \in \mathfrak{D}$ and let $\alpha$ and $C$ be such that (6.1) and (6.2) are satisfied. Without loss of generality, suppose that $C(n) \leq n$ for all $n$, as replacing any larger values of $C(n)$ with $n$ will not affect (6.1) (since $\mathsf{max\text{-}weight}(G_n) \leq n$ for all graphs $G_n$ on $n$ vertices). Fix $\varepsilon, \delta > 0$. We must identify some $n_0$ such that for all $n \geq n_0$, $\mathbb{P}_{\mathcal{D},M,n}[\mathrm{gain}(\vec{p}_n, G_n) \geq -\varepsilon] > 1 - \delta$.

We will begin by showing there exists $n_1 \in \mathbb{N}$ such that for all instances $(\vec{p}_n, G_n)$ on $n \geq n_1$ voters, if both

$$\mathsf{max\text{-}weight}(G_n) \leq C(n) \text{ and} \tag{6.4}$$

---

the average competence after delegation remains strictly above $1/2$ and the weighted sum $\sum_{i=1}^{n} \mathsf{weight}_i V_i$ concentrates. In turn, these established benchmarks are directly mapped to existing results in social choice theory on the convergence of weighted majorities [91].

$$\sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \sum_{i=1}^{n} p_i \geq 2\alpha n, \tag{6.5}$$

then

$$\text{gain}(\vec{p}_n, G_n) \geq -\varepsilon. \tag{6.6}$$

Since (6.4) and (6.5) each hold with probability $1 - o(1)$ by (6.1) and (6.2), for sufficiently large $n$, say $n \geq n_2$, they will each occur with probability at least $1 - \delta/2$. Hence, by a union bound, for all $n \geq n_2$, they both occur with probability at least $1 - \delta$. By taking $n_0 = \max(n_1, n_2)$, this implies that probabilistic do no harm is satisfied.

We now prove that, for sufficiently large $n$, (6.4) and (6.5) imply (6.6). First, we will show that

$$\text{gain}(\vec{p}_n, G_n) \geq -\mathbb{P}_{\vec{p}_n}[X_n^D > X_{G_n}^F]. \tag{6.7}$$

Indeed, we have that

$$\begin{aligned}\mathbb{P}_{\vec{p}_n}[X_n^D > n/2] &= \mathbb{P}_{\vec{p}_n}[X_n^D > n/2, X_{G_n}^F > n/2] + \mathbb{P}_{\vec{p}_n}[X_n^D > n/2, X_{G_n}^F \leq n/2] \\ &\leq \mathbb{P}_{\vec{p}_n}[X_{G_n}^F > n/2] + \mathbb{P}_{\vec{p}_n}[X_n^D > X_{G_n}^F]\end{aligned}$$

where the first transition holds by the law of total probability, and the second because the corresponding events are contained in each other. That is,

$$\{X_n^D > n/2, X_{G_n}^F > n/2\} \subseteq \{X_{G_n}^F > n/2\}$$

and

$$\{X_n^D > n/2, X_{G_n}^F \leq n/2\} \subseteq \{X_n^D > X_{G_n}^F\}.$$

Re-arranging the terms above yields (6.7).

Hence, for our purpose, it suffices to show that (6.4) and (6.5) imply

$$\mathbb{P}_{\vec{p}_n}\left[X_n^D > X_{G_n}^F\right] \leq \varepsilon.$$

Intuitively, we will use (6.5) to show the expected value of $X_n^D$ is well below the expected value of $X_{G_n}^F$. Then we will show both $X_n^D$ and $X_{G_n}^F$ concentrate well around their means, where for the latter we will need (6.4). Together, these observations imply that $X_{G_n}^F > X_n^D$ with high probability.

Fix an instance $(\vec{p}_n, G_n)$ on $n$ voters satisfying (6.4) and (6.5). We will show that

for sufficiently large $n$,

$$\mathbb{P}_{\vec{p}_n}\left[X_n^D < \sum_{i=1}^{n} p_i + \alpha n\right] > 1 - \varepsilon/2 \tag{6.8}$$

and

$$\mathbb{P}_{\vec{p}_n}\left[X_{G_n}^F > \sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \alpha n\right] > 1 - \varepsilon/2. \tag{6.9}$$

Note that since (6.5) holds for this instance, $\sum_{i=1}^{n} p_i + \alpha n \leq \sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \alpha n$. Therefore, when both events whose probability is considered in (6.8) and (6.9) hold, $X_n^D \leq X_n^F$. Hence,

$$\mathbb{P}_{\vec{p}_n}[X_n^D \leq X_{G_n}^F] \geq \mathbb{P}_{\vec{p}_n}\left[X_n^D < \sum_{i=1}^{n} p_i + \alpha n, X_{G_n}^F > \sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \alpha n\right] > 1 - \varepsilon$$

where the last inequality holds by a union bound. This implies that $\mathbb{P}_{\vec{p}_n}[X_n^D \leq X_{G_n}^F] < \varepsilon$, as needed.

It remains to be shown that (6.8) and (6.9) hold for sufficiently large $n$. For (6.8), this follows directly from Hoeffding's inequality. To prove (6.9), first note that, as shown in Kahng et al. [115],

$$\begin{aligned}
\mathrm{Var}_{\vec{p}_n}\left[X_{G_n}^F\right] &= \sum_{i=1}^{n} \mathsf{weight}_i(G_n)^2 \cdot p_i(1-p_i) \\
&\leq \frac{1}{4} \cdot \sum_{i=1}^{n} \mathsf{weight}_i(G_n)^2 \\
&\leq \frac{1}{4} \cdot \sum_{i=1}^{\lceil n/C(n) \rceil} C(n)^2 \\
&< nC(n) \in o(n^2),
\end{aligned}$$

where the first inequality holds because $p(1-p)$ is upper bounded by $1/4$, the second because $\sum_{i=1}^{n} \mathsf{weight}_i(G_n) \leq n$ with each $\mathsf{weight}_i(G_n) \leq C(n)$ so the value is maximized by setting as many terms to $C(n)$ as possible, and the final inequality holds because $C(n) \leq n$.

Hence, by Chebyshev's inequality,

$$\mathbb{P}_{\vec{p}_n}\left[X^F_{G_n} \leq \mathbb{E}_{\vec{p}_n}\left[X^F_{G_n}\right] - \alpha n\right] \leq \frac{\mathrm{Var}_{\vec{p}_n}[X^F_{G_n}]}{(\alpha n)^2}.$$

This bound is $o(1)$ because the numerator is $o(n^2)$ and the denominators is $\Omega(n^2)$. This implies that for sufficiently large $n$, it will be strictly less than $\varepsilon/2$, so (6.9) holds.

*Probabilistic positive gain:* Fix a distribution $\mathcal{D} \in \mathfrak{D}$ and an $\alpha \in (0,1)$ such that (6.3) holds. We want to show that $M$ satisfies probabilistic positive gain. Since $\mathcal{D} \in \mathfrak{D}$, it also satisfies (6.1) for some $C$. We show below that there exists an $n_3$ such that all instances $(\vec{p}_n, G_n)$ with $n \geq n_3$ voters satisfying (6.4) for which $\sum_{i=1}^n p_i + \alpha n \leq n/2 \leq \sum_{i=1}^n \mathsf{weight}_i(G_n) \cdot p_i - \alpha n$, we have that $\mathrm{gain}(\vec{p}_n, G_n) \geq 1 - \varepsilon$. As with the DNH part of the proof, since the events of (6.1) and (6.3) each hold with probability $1 - o(1)$, for sufficiently large $n$, say $n \geq n_4$, they each occur with probability at least $1 - \delta/2$. Hence, by a union bound, for all $n \geq n_4$, they both occur with probability $1 - \delta$. For $n_0 = \max(n_3, n_4)$, probabilistic positive gain is satisfied.

It remains to show that that if (6.1) and (6.3) hold for a specific instance $(\vec{p}_n, G_n)$, then $\mathrm{gain}(\vec{p}_n, G_n) \geq 1 - \varepsilon$ for sufficiently large $n$. Since $\mathcal{D} \in \mathfrak{D}$, (6.8) and (6.9) are both satisfied for sufficiently large $n$. When

$$\sum_{i=1}^n p_i + \alpha n \leq n/2 \leq \sum_{i=1}^n p_i - \mathsf{weight}_i(G_n) \cdot \alpha n$$

is satisfied as well, we get that $\mathbb{P}_{\vec{p}_n}\left[X^D_n > n/2\right] < \varepsilon/2$ and $\mathbb{P}_{\vec{p}_n}\left[X^L_{G_n} > n/2\right] > 1 - \varepsilon/2$, so $\mathrm{gain}(\vec{p}_n, G_n) > 1 - \varepsilon$ is immediate. □

In words, condition (6.1) ensures that, as the number of voters grows large, the weighted number of correct votes under liquid democracy will concentrate around its expectation, $\sum_{i=1}^n \mathsf{weight}_i(G_n) \cdot p_i$. Standard concentration results already imply this holds for direct democracy. Condition (6.2) ensures that these expectations are sufficiently separated. So with high probability, liquid democracy will have more correct votes than direct democracy, which is sufficient to guarantee DNH. Finally, Condition (6.3) ensures that in some cases, the expectations for direct and liquid votes will be below and over half the voters, respectively, which after applying concentration means there will likely be a large gain.

In the following sections, we investigate natural delegation models and identify conditions such that the models satisfy probabilistic do no harm and probabilistic positive gain. In all instances, we will invoke Lemma 6.2.3 after showing that its sufficient conditions are satisfied.

## 6.3 Strictly Upward Delegation Model

We now turn to the analysis of a simple model that assumes that voters either do not delegate with fixed exogenous probability or delegate to voters that have a competence greater than their own.

Formally, for a fixed $p \in [0,1]$ we let $M_p^U = (q, \varphi)$ be the model consisting of $q(p_i) = p$ for all $p_i \in [0,1]$, and $\varphi(p_i, p_j) = \mathbb{I}_{\{p_j > p_i\}}$ for all $i, j \in [n]$. That is, voter $i$ delegates with fixed probability $p$ and puts equal weight on all the more competent voters. In other words, if voter $i$ delegates, then $i$ does so to a more competent voter chosen uniformly at random. Note that a voter with maximal competence will place 0 weight on all other voters, and hence is guaranteed not to delegate. We refer to $M_p^U$ as the *Upward Delegation Model* parameterized by $p$.

**Theorem 6.3.1** (Upward Delegation Model)**.** *For all $p \in (0,1)$, $M_p^U$ satisfies probabilistic do no harm and probabilistic positive gain with respect to the class $\mathfrak{D}^C$ of all continuous distributions.*

The proof of the theorem relies on novel bounds we drive on the largest bin size in an infinite Pólya urn process [51, 187]. We first formally define the process and present our bound in Lemma 6.3.2. A *Pólya urn process with attachment probability* $p$ begins at time $t = 1$ with one ball in one bin. At each timestep $t > 1$, a new ball arrives. With probability $1 - p$, a new bin is created and the new ball is placed in that bin; with probability $p$, the ball joins an existing bin, and it does so with probability proportional to the number of balls in the bins, i.e., if there are three bins containing 1, 2, and 3 balls respectively, it joins each with probability $1/6, 2/6$, and $3/6$ respectively. We then have the following.

**Lemma 6.3.2.** *For all $p \in (0,1)$ and $t \geq 1$, let $L_t^p$ be the random variable denoting the maximum number of balls in any bin after running the infinite Pólya urn process with new-bin probability $p$ for $t$ steps. Then, there exists $\delta < 1$ depending only on $p$ such that for all $T \geq 1$, $\Pr[L_T^p \leq T^\delta] = 1 - o(1)$.*

*Proof.* Fix the parameter $p \in (0,1)$. Choose $\gamma$ to be a constant such that $3/4 < \gamma < 1$; note that $p + (1-p)\gamma < p + (1-p) = 1$. Choose $\delta$ (for the lemma statement) such that $p + (1-p)\gamma < \delta < 1$. Notice that we can choose $\gamma$ and $\delta$ such that $\delta$ is arbitrarily close to $3/4 + p/4$.

Let $B^{(k)}$ denote the $k$-th bin. Let $U_t^{(k)}$ be the size of $B^{(k)}$ at time $t$. Since there are at most $t$ bins by time $t$, notice that $L_t^p = \max(U_t^{(1)}, \ldots, U_t^{(t)})$. In general, our approach will be to analyze bins separately and show that $U_T^{(k)}$ remains below $T^\delta$ with high enough probability so that we can union bound over all possible $k \leq T$. That is, we will show

$$\sum_{k=1}^{T} \Pr[U_T^{(k)} > T^\delta] = o(1),$$

which also implies $\Pr[L_T^p > T^\delta] = o(1)$. Hence, it will be useful to consider this process more formally from the perspective of the $k$th bin, $B^{(k)}$. The $k$th bin $B^{(k)}$ is "born" at some time $t \geq k$, the $k$th time in which a ball does not join a pre-existing bin, at which point $U_t^{(k)} = 1$ (prior to this, $U_t^{(k)} = 0$). More specifically, the first bin $B^{(k)}$ is guaranteed to be born at time $t = 1$ and for all other $k > 1$, $B^{(k)}$ will be born at time $t \geq k$ with probability $\binom{t-1}{k-1}(1-p)^k p^{t-k}$, although these exact probabilities will be unimportant for our analysis. Once born, we have the following recurrence on $U_t^{(k)}$ describing the probability $B^{(k)}$ will be chosen at time $t$:

$$U_t^{(k)} = \begin{cases} U_{t-1}^{(k)} + 1 & \text{with probability } \frac{p \cdot U_{t-1}^{(k)}}{t-1} \\ U_{t-1}^{(k)} & \text{with probability } 1 - \frac{p \cdot U_{t-1}^{(k)}}{t-1}. \end{cases}$$

Let $W_t^{(k)}$ be the process for the size of bin that is born at time $k$. That is, $W_k^{(k)} = 1$, and for $k > t$, $W_t^{(k)}$ follows the exact same recurrence as $U_t^{(k)}$. Note that since the $k$th bin $B^{(k)}$ can only be born at time $k$ or later, we have that $W_t^{(k)}$ stochastically dominates $U_t^{(k)}$ for all $k$ and $t$. Hence, it suffices to show that

$$\sum_{k=1}^{T} \Pr[W_T^{(k)} > T^\delta] = o(1). \tag{6.10}$$

We split our analysis into two parts: the first consider the first $T^\gamma$ bins, while the second considers the last $T - T^\gamma$ bins.

We first show that $\sum_{k=1}^{T^\gamma} \mathbb{P}[W_T^{(k)} > T^\delta] = o(1)$. Note that the expectation of $W_n^{(k)}$

$$\mathbb{E}[W_n^{(k)}] = \frac{\Gamma(n+p)\Gamma(k)}{\Gamma(p+k)\Gamma(n)} \tag{6.11}$$

for all $k \leq n$, where $\Gamma$ represents the Gamma function. Recall that $W_k^{(k)} = 1$ and we have the following recurrence for all $t > k$:

$$W_t^{(k)} = \begin{cases} W_{t-1}^{(k)} + 1 & \text{with probability } \frac{p \cdot W_{t-1}^{(k)}}{t-1} \\ W_{t-1}^{(k)} & \text{with probability } 1 - \frac{p \cdot W_{t-1}^{(k)}}{t-1}. \end{cases}$$

By the tower property of expectation, for all $t \geq k+1$,

$$\begin{aligned} \mathbb{E}[W_t^{(k)}] &= \mathbb{E}[\mathbb{E}[W_t^{(k)} \mid W_{t-1}^{(k)}]] \\ &= \mathbb{E}[W_{t-1}^{(k)}(1 + \frac{p}{t-1})] \\ &= \mathbb{E}[W_{t-1}^{(k)}](1 + \frac{p}{t-1}). \end{aligned}$$

Thus, by a straightforward induction argument and the fact that $\mathbb{E}[W_k^{(k)}] = 1$,

$$\mathbb{E}[W_T^{(k)}] = \mathbb{E}[W_k^{(k)}] \prod_{i=k}^{T-1} (1 + \frac{p}{i}) = \prod_{i=k}^{T-1} (1 + \frac{p}{i}).$$

Expanding this, we have

$$\begin{aligned} \prod_{i=k}^{T-1} (1 + \frac{p}{i}) &= \prod_{i=k}^{T-1} \frac{i+p}{i} \\ &= \frac{1}{\prod_{i=k}^{T-1} i} \cdot \prod_{i=k}^{T-1} (i+p) \\ &= \frac{(k-1)!}{(T-1)!} \cdot \frac{\prod_{i=0}^{T-1} (i+p)}{\prod_{i=0}^{k-1} (i+p)} \\ &= \frac{(k-1)!}{(T-1)!} \frac{\frac{\Gamma(p+T)}{\Gamma(p)}}{\frac{\Gamma(k+p)}{\Gamma(p)}} \\ &= \frac{\Gamma(T+p)\Gamma(k)}{\Gamma(p+k)\Gamma(T)}, \end{aligned}$$

where the fourth equality holds because $\Gamma(x+1) = x\Gamma(x)$ for all $x \in \mathbb{R}$, and the last uses the fact that $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}$. This proves (6.11).

Using this along with Gautchi's inequality [81], $(t+p-1)^p \leq \frac{\Gamma(p+t)}{\Gamma(t)} \leq (t+p)^p$, to approximate the $\Gamma$ terms, we can apply Markov's inequality and use algebra to get $\sum_{k=1}^{n^\gamma} \mathbb{P}[W_n^k > n^\delta] = o(1)$. can now use Markov's inequality to show that for all $k$,

$$\Pr\left[W_T^{(k)} > T^\delta\right] \leq \frac{\mathbb{E}[W_T^{(k)}]}{T^\delta} = \frac{1}{T^\delta} \cdot \frac{\Gamma(T+p)}{\Gamma(T)} \cdot \frac{\Gamma(k)}{\Gamma(k+p)}.$$

Hence,

$$\sum_{k=1}^{T^\gamma} \Pr[W_T^{(k)} > T^\delta] \leq \sum_{k=1}^{T^\gamma} \frac{1}{T^\delta} \cdot \frac{\Gamma(T+p)}{\Gamma(T)} \cdot \frac{\Gamma(k)}{\Gamma(k+p)} = \frac{1}{T^\delta} \cdot \frac{\Gamma(T+p)}{\Gamma(T)} \cdot \sum_{k=1}^{T^\gamma} \frac{\Gamma(k)}{\Gamma(k+p)}.$$

What remains to be shown is that

$$\frac{1}{T^\delta} \cdot \frac{\Gamma(T+p)}{\Gamma(T)} \cdot \sum_{k=1}^{T^\gamma} \frac{\Gamma(k)}{\Gamma(k+p)} = o(1).$$

To do this, we will use Gautschi's inequality [81] which states that for all $x > 0$, since $p \in (0,1)$,

$$(x+p-1)^p \leq \frac{\Gamma(p+x)}{\Gamma(x)} \leq (x+p)^p$$

We then have that

$$\begin{aligned}
\frac{1}{T^\delta} \cdot \frac{\Gamma(T+p)}{\Gamma(T)} \cdot \sum_{k=1}^{T^\gamma} \frac{\Gamma(k)}{\Gamma(k+p)} &\leq \frac{(T+p)^p}{T^\delta} \cdot \sum_{k=1}^{T^\gamma} \frac{1}{(k+p-1)^p} \\
&= \frac{(T+p)^p}{T^\delta} \cdot \left( \frac{1}{p^p} + \frac{1}{(1+p)^p} + \sum_{k=3}^{T^\gamma} \frac{1}{(k+p-1)^p} \right) \\
&\leq \frac{(T+p)^p}{T^\delta} \cdot \left( \frac{1}{p^p} + \frac{1}{(1+p)^p} + \sum_{k=3}^{T^\gamma} \frac{1}{(k-1)^p} \right) \\
&= \frac{(T+p)^p}{T^\delta} \cdot \left( \frac{1}{p^p} + \frac{1}{(1+p)^p} + \sum_{k=2}^{T^\gamma - 1} \frac{1}{k^p} \right) \\
&\leq \frac{(T+p)^p}{T^\delta} \cdot \left( \frac{1}{p^p} + \frac{1}{(1+p)^p} + \sum_{k=2}^{T^\gamma} \frac{1}{k^p} \right)
\end{aligned}$$

$$\leq \frac{(T+p)^p}{T^\delta} \cdot \left( \frac{1}{p^p} + \frac{1}{(1+p)^p} + \int_1^{T^\gamma} \frac{1}{x^p}\, dx \right)$$
$$= \frac{(T+p)^p}{T^\delta} \cdot \left( \frac{1}{p^p} + \frac{1}{(1+p)^p} + \frac{x^{1-p}}{1-p}\Big|_1^{T^\gamma} \right)$$
$$= \frac{(T+p)^p}{T^\delta} \cdot \left( \frac{T^{\gamma(1-p)}}{1-p} + \frac{1}{p^p} + \frac{1}{(1+p)^p} - \frac{1}{1-p} \right).$$

Notice that asymptotically, this upper bound is $O(T^{-\delta+p+\gamma\cdot(1-p)})$. By our choice of $\delta$, $\delta > p + \gamma \cdot (1-p)$, so this implies that it is is $o(1)$, as desired.

Now consider the final $T - T^\gamma$ components. We will prove that $\Pr[W_T^{(T^\gamma+1)} > T^\delta] = o(1/T)$. Since $W_T^{(k)}$ stochastically dominates $W_T^{(k')}$ for all $k' \geq k$, this implies that $\Pr[W_T^{(k)} > T^\delta] = o(1/T)$ for all $k \geq T^\gamma + 1$. Hence,

$$\sum_{k=T^\gamma+1}^{T} \Pr\left[W_T^{(k)} > T^\delta\right] = o(1).$$

To do this, we compare the $W_t^{(T^\gamma+1)}$ process to another process, $V_t$. We define $V_0 = 1$, and for $t > 0$, take $V_t$ to satisfy the following recurrence:

$$V_t = \begin{cases} V_{t-1} + 1 & \text{with probability } \frac{V_{t-1}}{t+n^\gamma} \\ V_{t-1} & \text{with probability } 1 - \frac{V_{t-1}}{t+n^\gamma}. \end{cases}$$

This is identical to the $W$ recurrence with $t$ shifted down by $n^\gamma + 1$ except without the $p$ factor. Hence, $V_{T-T^\gamma+1}$ clearly stochastically dominates $W_T^{(T^\gamma+1)}$. For convenience in calculation, we will instead focus on bounding $V_T$ which itself stochastically dominates $V_{T-T^\gamma+1}$.

Next, note that the $V_t$ process is isomorphic to the following classic Pólya urn process. We begin with two bins, one with a single ball and the other with $n^\gamma$ balls. At each time, a new ball is added to one of the two bins with probability proportional to the bin size. The process $V_t$ is isomorphic to the size of the one-ball urn after $t$ steps. Classic results tell us that for fixed starting bin sizes $a$ and $b$, as the number of steps grows large, the possible proportion of balls in the $a$-bin follows a Beta$(a, b)$ distribution [67, 114, 140, 143, 164].

The mean and variance of such a Beta distribution would be sufficient to prove our necessary concentration bounds; however, for us, we need results after exactly

$T - T^\gamma$ steps, not simply in the limit. Hence, we will be additionally concerned with the speed of convergence to this Beta distribution.

Let $X_T = \frac{V_T}{T}$ and $Z_T \sim \text{Beta}(1, T^\gamma)$. From Janson [112], we know that the rate of convergence is such that, for any $p \geq 1$

$$\ell_p(X_T, Z_T) = \Theta(1/T) \tag{6.12}$$

where $\ell_p$ is the *minimal* $L_p$ *metric*, defined as

$$\ell_p(X, Y) = \inf\{\mathbb{E}[|X' - Y'|^p]^{1/p} \mid X' \overset{d}{=} X, Y' \overset{d}{=} Y\},$$

which can be thought of as the minimal $L_p$ norm over all possible couplings between $X$ and $Y$. For our purposes, the only fact about the $\ell_p$ metric we will need is that $\ell_p(X, 0) = \mathbb{E}[|X|^p]^{1/p}$ where 0 is the identically 0 random variable. Since $\ell_p$ is in fact a metric, the triangle inequality tells us that $\ell_p(0, X_n) \leq \ell_p(0, Z_n) + \ell_p(Z_n, X_n)$, so, combining with (6.12), we have that

$$\mathbb{E}[|X_T|^p]^{1/p} \leq \mathbb{E}[|Z_T|^p]^{1/p} + \Theta(1/T) \tag{6.13}$$

for all $p \geq 1$.

Note that since $Z_T \sim \text{Beta}(1, T^\gamma)$,

$$\mathbb{E}[Z_T] = \frac{1}{1 + T^\gamma} = \Theta(T^{-\gamma})$$

and

$$\text{Var}[Z_T] = \frac{T^\gamma}{(2 + T^\gamma)(1 + T^\gamma)^2} = \Theta(T^{-2\gamma}).$$

Given these results, we are ready to prove that $V_T$ is smaller than $T^\delta$ with probability $1 - o(1/T)$. Precisely, we want to show that $\Pr[X_T \geq T^{\delta-1}] = o(1)$. By Chebyshev's inequality,

$$\Pr[X_T \geq T^{\delta-1}] \leq \frac{\text{Var}[X_T]}{(T^{\delta-1} - \mathbb{E}[X_T])^2}.$$

Inequality (6.13) with $p = 1$ along with the fact that $X_T$ and $Z_T$ are always nonnegative implies that $\mathbb{E}[X_T] \leq \mathbb{E}[Z_T] + \Theta(1/T) = O(T^{-\gamma})$. Hence, $T^{\delta-1} - \mathbb{E}[X_T] = \Omega(T^{\delta-1})$

since $\delta - 1 > -1/2 > -\gamma$. We can therefore write:

$$\left(T^{\delta-1} - \mathbb{E}[X_T]\right)^2 = \Omega(T^{-2(\delta-1)}). \tag{6.14}$$

Inequality (6.13) with $p = 2$ implies that $\sqrt{\mathbb{E}[X_T^2]} \leq \sqrt{\mathbb{E}[Z_T^2]} + \Theta(1/T)$. Hence,

$$\begin{aligned}
\mathbb{E}[X_T^2] &\leq (\Theta(1/T) + \sqrt{\mathbb{E}[Z_T^2]})^2 \\
&\leq (\Theta(1/T) + \sqrt{\mathbb{E}[Z_T]^2 + \mathrm{Var}[Z_T]})^2 \\
&\leq (\Theta(1/T) + \sqrt{\Theta(T^{-2\gamma})})^2 \\
&= (\Theta(1/T) + \Theta(T^{-\gamma}))^2 \\
&= \Theta(T^{-\gamma})^2 \\
&= \Theta(T^{-2\gamma}).
\end{aligned}$$

Next, note that $\mathrm{Var}[X_T] \leq \mathbb{E}[X_T^2]$, so

$$\mathrm{Var}[X_T] = O(T^{-2\gamma}) \tag{6.15}$$

as well. Combining (6.14) and (6.15), we have that

$$\Pr[X_T \geq T^{\delta-1}] \leq \frac{\mathrm{Var}[X_T]}{(T^{\delta-1} - \mathbb{E}[X_T])^2} = O\left(T^{-2\gamma+2(1-\delta)}\right).$$

Since $-2\gamma + 2(1-\delta) < 1$, given our assumption that $3/4 < \gamma < \delta$, it follows that $\Pr[X_T \geq T^{\delta-1}] = o(1/T)$, which allows us to conclude that

$$\sum_{k=T^\gamma+1}^{T} \Pr[W_T^{(k)} > T^\delta] = o(1).$$

Since we showed earlier that $\sum_{k=1}^{T^\gamma} \Pr[W_T^{(k)} > T^\delta] = o(1)$, we have that

$$\sum_{k=1}^{T} \Pr[W_T^{(k)} > T^\delta] = o(1),$$

as needed. □

We are now ready to prove the theorem about Upward Delegation.

*Proof of Theorem 6.3.1.* To prove the theorem, we will prove that the Upward Delegation Model with respect to $\mathfrak{D}^C$ satisfies (6.1), (6.2), (6.3), which implies that the model satisfies probabilistic do no harm and positive gain by Lemma 6.2.3.

*Upward Delegation satisfies* (6.1)

To do this, we will simply show that the component sizes in $G_n$ sampled according to $\mathbb{P}_{D,M,n}$ have the same distribution as the bin sizes in a Pólya urn process with attachment probability $p$, and hence $\mathsf{max\text{-}weight}(G_n)$ follows the same distribution as $L_n^p$. Once we have shown this, (6.1) follows immediately from Lemma 6.3.2 as $n^\delta \in o(n)$.

To that end, fix some sampled competencies $\vec{p}_n$. Recall that each entry $p_i$ in $\vec{p}_n$ is sampled i.i.d. from $\mathcal{D}$, a continuous distribution. Hence, almost surely, no two competencies are equal. From now on, we condition on this probability 1 event. Now consider sampling the delegation graph $G_n$. By the design of the model $M_p^U$, we can consider a random process for generating $G_n$ that is isomorphic to sampling according to $\mathbb{P}_{D,M,n}$ as follows: first, order the competencies $p_{(1)} > p_{(2)} > \cdots > p_{(n)}$ (note that such strict order is possible by our assumption that all competencies are different) and rename the voters such that voter $i$ has competence $p_{(i)}$; then construct $G_n$ iteratively by adding the voters one at a time in decreasing order of competencies, voter 1 at time 1, voter 2 at time 2, and so on.

We start with the voter with the highest competence, voter 1. By the choice of $\varphi$, voter 1 places weight 0 on every other voter and hence by definition does not delegate. This voter forms the first component in the graph $G_n$, which we call $C^{(1)}$. Then, we add voter 2 who either delegates to voter 1 joining component $C^{(1)}$ with probability $p$, or starts a new component $C^{(2)}$ with probability $1-p$. Next, we add voter 3. If $2 \in C^{(1)}$ (that is, if 2 delegated to 1), 3 either delegates to 1 (either directly or through 2 by transitivity) with probability $p$ or she starts a new component $C^{(2)}$. If $2 \in C^{(2)}$, then 3 either delegates to 1 with probability $p/2$ and is added to $C^{(1)}$, or delegates to 2 with probability $p/2$ and is added to $C^{(2)}$, or starts a new component $C^{(3)}$. In general, at time $t$, if there are $k$ existing components $C^{(1)}, \ldots, C^{(k)}$, voter $t$ either joins each component $C(j)$ with probability $\frac{p|C(j)|}{t-1}$ or starts a new component with probability $1-p$. To construct $G_n$, we run this process for $n$ steps. Notice that this is

identical to the Pólya urn process with bins $B^{(k)}$ and balls replaced with components $C^{(k)}$ and voters being run for $n$ steps, as needed.

*Upward Delegation satisfies* (6.2)

We will show there exists $\alpha \in (0,1)$ such that $\sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \sum_{i=1}^{n} p_i \geq 2\alpha n$ with high probability, so (6.2) is satisfied. Note that in the present scheme, cycles are impossible, so do need to worry about ignored voters.

Since $\mathcal{D}$ is a continuous distribution, there exists $a < b$ such that $\pi_a := \mathcal{D}[\{p : p < a\}] > 0$ and $\pi_b := \mathcal{D}[\{p : p > b\}] > 0$. Let $N_{a,n}(\vec{p}_n)$ be the number of voters in $\vec{p}_n$ with competence $p_i < a$ and $N_{b,n}(\vec{p}_n)$ be the number of voters with competence $p_i > b$. When we sample competencies, since each is chosen independently, $N_{a,n} \sim \mathrm{Bin}(n, \pi_a)$ and $N_{b,n} \sim \mathrm{Bin}(n, \pi_b)$. By Hoeffding's inequality and the union bound, with probability $1 - o(1)$, there will be at least $\pi_a/2 \cdot n$ voters with competence $p_i < a$ and $\pi_b/2 \cdot n$ voters with competence $p_i > b$. Indeed,

$$\begin{aligned} \mathcal{D}^n[N_{a,n} > \frac{n\pi_a}{2}, N_{b,n} > \frac{n\pi_b}{2}] &= 1 - \mathcal{D}^n[\{N_{a,n} \leq \frac{n\pi_a}{2}\} \cup \{N_{b,n} \leq \frac{n\pi_b}{2}\}] \\ &\geq 1 - (\mathcal{D}^n[N_{a,n} \leq \frac{n\pi_a}{2}] + \mathcal{D}^n[N_{b,n} \leq \frac{n\pi_b}{2}]) \\ &\geq 1 - \exp(-\frac{n\pi_a^2}{2}) - \exp(-\frac{n\pi_b^2}{2}), \end{aligned} \tag{6.16}$$

where the first line comes from De Morgan's law, the second from the union bound, and the last from Heoffding's inequality.

Conditioned on this occurring, each voter with competence $p_i < a$ has probability at least $p\pi_b/2$ of delegating to a voter with competence at least $b$. As they each decide to do this independently, the number $N_{ab,n}$ of $n$ voters deciding to do this stochastically dominates a random variable following the $\mathrm{Bin}(\pi_a/2 \cdot n, p \cdot \pi_b/2)$ distribution. We can again apply Hoeffding's inequality to conclude that with probability $1 - o(1)$, at least $\pi_a \cdot \pi_b \cdot p/8 \cdot n$ voters do so. Indeed,

$$\begin{aligned} \mathcal{D}[N_{ab,n} > \frac{np\pi_a\pi_b}{8} \mid N_{a,n} > \frac{n\pi_a}{2}, N_{b,n} > \frac{n\pi_b}{2}] &\geq \mathcal{D}[\mathrm{Bin}(\frac{n\pi_a}{2}, \frac{p\pi_b}{2}) > \frac{np\pi_a\pi_b}{8}] \\ &\geq 1 - \exp(-\frac{np^2\pi_a\pi_b^2}{4}), \end{aligned} \tag{6.17}$$

where the first inequality holds because $N_{ab,n}$ stochastically dominates the corresponding binomial random variable and the second holds by Hoeffding's inequality. Finally,

using (6.16) and (6.17), we have

$$\begin{aligned}\mathcal{D}[N_{ab,n} > \frac{np\pi_a\pi_b}{8}] &\geq \mathcal{D}[N_{ab,n} > \frac{np\pi_a\pi_b}{8} \mid N_{a,n} > \frac{n\pi_a}{2}, N_{b,n} > \frac{n\pi_b}{2}] \\ &\quad \cdot \mathcal{D}[N_{a,n} > \frac{n\pi_a}{2}, N_{b,n} > \frac{n\pi_b}{2}] \\ &\geq 1 - o(1).\end{aligned}$$

Under these upward delegation models, delegations can only increase the total competence of all voters. Hence,

$$\sum_{i=1}^{n} \mathsf{dels}_i(G_n) \cdot p_i - \sum_{i=1}^{n} p_i \geq (b-a)N_{ab.n}.$$

Each of these $\pi_a \cdot \pi_b \cdot p/8 \cdot n$ voters results in a competence increase of at least $b - a$. Hence, under these high probability events, the total competence increase is at least $(b-a) \cdot \pi_a \cdot \pi_b \cdot p/8 \cdot n$. Indeed, since $\mathcal{D}[N_{ab,n} > \frac{np\pi_a\pi_b}{8}] = 1 - o(1)$, this implies $\mathcal{D}[\sum_{i=1}^{n} \mathsf{dels}_i(G_n) \cdot p_i - \sum_{i=1}^{n} p_i > \frac{np\pi_a\pi_b}{8}] = 1 - o(1)$. By choosing $\alpha = \frac{p\pi_a\pi_b}{8}(b-a)$, we see that there is an $\alpha \cdot n$ increase in competence with high probability, as needed.

We have proved that for any continuous distribution $\mathcal{D}$, and for well-behaved realizations of $\vec{p}_n$ which occur with high probability, the graph generated from the random delegation process yields an increase in the expected sum of the votes of at least $\alpha \cdot n$. We can then conclude that $M_p^U$ satisfies Equation (6.2) with respect to the class of continuous distributions.

*Upward Delegation satisfies* (6.3)

We now show that there exists a distribution $\mathcal{D}$ such that $\sum_{i=1}^{n} p_i + \alpha n \leq n/2 \leq \sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i - \alpha n$ with probability $1 - o(1)$ for some $\alpha > 0$. This implies that the model satisfies probabilistic positive gain by Lemma 1.3.3, and will conclude the proof.

We take $\mathcal{D}$ to be $\mathcal{D}_\eta$, the uniform distribution $\mathcal{U}[0, 1-2\eta]$ for some small $0 < \eta < p/512$. Let $\alpha = \eta/2$. Clearly, $\mu_{\mathcal{D}_\eta}$, the mean of $D_\eta$, is $1/2 - \eta$. Since each $p_i \overset{i.i.d.}{\sim} \mathcal{D}_\eta$, the $p_i$s are bounded independent random variables with mean $1/2 - \eta$, so Hoeffding's inequality directly implies that $\sum_{i=1}^{n} p_i \leq n/2 - n\eta/2 = n/2 - n\alpha$ with high probability.

Now consider $\mathcal{E}_F$, the event consisting of instances $(\vec{p}_n, G_n)$ such that $\sum_{i=1}^{n} \mathsf{weight}_i(G_n) \cdot p_i \geq n/2 + n\alpha$. We denote by $\mathcal{E}_D$ the event that $\sum_{i=1}^{n} p_i \geq n/2 - 3n\eta/2$. The same reasoning as before implies that $\Pr_{\mathcal{D}_\eta, M_p^U, n}(\mathcal{E}_D) = 1 - o(1)$.

Let $a = 1/4 - \eta/2$ and $b = 1/2 - \eta$, so we have that $\pi_a := \mathcal{D}_\eta[p_i < a] = 1/4$ and $\pi_b := \mathcal{D}_\eta[p_i > b] = 1/2$. We proved in the preceding derivation that $\sum_{i=1}^n \mathsf{weight}_i(G_n) \cdot p_i - \sum_{i=1}^n p_i > \frac{np\pi_1\pi_b}{8} = \frac{np}{64}(1-\eta)$ with high probability. Hence, if both this and $\mathcal{E}_D$ occur, which is the case with high probability, by the union bound, it follows that $\sum_{i=1}^n \mathsf{weight}_i(G_n) \cdot p_i > n/2 + n(\frac{p}{128}(1-\eta) - 3\eta/2)$ with high probability.

Since $\eta < \frac{p}{512} < 1/2$, we have that

$$\frac{p}{128}(1-\eta) - 3\eta/2 > \frac{p}{256} - 3\eta/2 > 2\eta - 3\eta/2 = \eta/2 = \alpha,$$

and we can conclude that $\mathcal{E}_F$ occurs with high probability. Hence, $M_p^U$ satisfies Equation (6.3).

### 6.3.1 Missing Details from the Proof of Theorem 6.4.1

We show that the Confidence-based Model satisfies (6.3).

*Confidence-Based Delegation satisfies* (6.3)

We finally show there exists a distribution $\mathcal{D}$ such that $\sum_{i=1}^n p_i + \alpha n \leq n/2 \leq \sum_{i=1}^n \mathsf{weight}_i(G_n) \cdot p_i - \alpha n$ with probability $1 - o(1)$. This implies that the model $M_q^C$ satisfies probabilistic positive gain by Lemma 1.3.3.

Using the notation of the analogous proof in Section 6.3, let $\mathcal{D}_\eta = \mathcal{U}[0, 1-2\eta]$ for $\eta \in [0, 1/2)$. Note that as a function of $\eta$, $\frac{\mathbb{E}_{\mathcal{D}_\eta}[q^+]}{\mathbb{E}_{\mathcal{D}_\eta}[\bar{q}]}$, the expected competence conditioned on not delegating, is continuous. Moreover, if $\eta = 0$, then

$$\frac{\mathbb{E}_{\mathcal{D}_0}[q^+]}{\mathbb{E}_{\mathcal{D}_0}[\bar{q}]} > \mu_{\mathcal{D}_0} = 1/2.$$

Hence, for sufficiently small $\eta > 0$,

$$\frac{\mathbb{E}_{\mathcal{D}_\eta}[q^+]}{\mathbb{E}_{\mathcal{D}_\eta}[\bar{q}]} > 1/2 > \mu_{\mathcal{D}_\eta}.$$

We choose $\mathcal{D}_\eta$ to be our distribution for this choice of $\eta$. As in the previous section, let $\mu^*_{\mathcal{D}_\eta} = \frac{\mathbb{E}_{\mathcal{D}_\eta}[q^+]}{\mathbb{E}[\bar{q}]}$. Note that $\mu_{\mathcal{D}_\eta} = 1/2 - \eta$. Let $\gamma = \min(\frac{1/2 - \mu_{\mathcal{D}_\eta}}{2}, \frac{\mu^*_{\mathcal{D}_\eta} - 1/2}{2})$ and $\alpha = \gamma$.

By the earlier argument for (6.2), we have that with high probability

$$\sum_{i=1}^{n} p_i \leq n(\mu + \gamma) \leq n/2 - \alpha n$$

and

$$\sum_{i=1}^{n} \mathsf{weight}_i(G) p_i \geq n(\mu^* - \gamma) \geq n/2 + \alpha n.$$

By the union bound, we have that both occur simultaneously with high probability, so (6.3) holds. □

## 6.4 Confidence-Based Delegation Model

We now explore a model according to which voters delegate with probability that is strictly decreasing (or, monotonically decreasing, that is $x < y$ implies $f(x) > f(y)$) in their competence and when they do decide to delegate, they do so by picking a voter uniformly at random. This models the case where voters do not need to know anything about their peers' competencies, but do have some sense of their own competence, and delegate accordingly.

Formally, for any $q$, let $M_q^C = (q, \varphi^1)$ where $\varphi^1(p_i, p_j) = 1$ for all $i, j \in [n]$. Voter $i$ puts equal weight on all the voters and hence samples one uniformly at random when they delegate. We refer to $M_q^C$ as the Confidence-Based Delegation Model.

**Theorem 6.4.1** (Confidence-Based Delegation Model)**.** *All models $M_q^C$ with monotonically decreasing $q$ satisfy probabilistic do no harm and probabilistic positive gain with respect to the class $\mathfrak{D}^C$ of all continuous distributions.*

*Proof.* We show that the Confidence-Based Model satisfy (6.1) and (6.2) here, and relegate showing (6.3) to Appendix 6.3.1.

*Confidence-Based Delegation satisfies* (6.1)

Fix some distribution $\mathcal{D} \in \mathfrak{D}^C$. We show there exists $C(n) \in O(\log n)$ such that (6.1) holds.

Note that when sampling an instance $(\vec{p}_n, G_n)$, the probability an arbitrary voter $i$ chooses to delegate is precisely $p := \mathbb{E}_{\mathcal{D}}[q]$. To see this, consider how a voter $i$ chooses whether to delegate: they first sample a competence $p_i \sim \mathcal{D}$ and then sample whether

or not to delegate from $\text{Bern}(q(p_i))$. Treating this as a single process, it is clear that the overall probability of choosing to delegate is exactly $\mathbb{E}_{\mathcal{D}}[q]$ by integrating out the competence.

Further, since $\mathcal{D}$ is continuous and $q$ is monotonically decreasing, $p \in (0, 1)$. When a voter does decide to delegate, they do so by picking another voter uniformly at random. Hence, we can consider the marginal distribution of delegation graphs directly (ignoring the competencies). We will show that when sampling a delegation graph, for any specific voter $i$, with probability $1 - o(1/n)$, $\mathsf{dels}_i(G_n) \leq C(n)$, which implies $\mathsf{weight}_i(G_n) \leq C(n)$. A union bound over all $n$ voters implies $\mathsf{max\text{-}weight}(G_n) \leq C(n)$ with probability $1 - o(1)$.

To that end, we will describe a branching process similar to the well-known *graph branching process* [3], which has the property that the distribution of its size exactly matches the distribution of $\mathsf{dels}_i(G_n)$ for an arbitrary voter $i$. We will compare this process to a known graph branching process that has size at most $O(\log n)$ with high probability. We will show our process is sufficiently dominated such that it too has size at most $O(\log n)$ with high probability. The branching process works as follows. Fix our voter $i$. We sample which other voters end up in $i$'s "delegation tree" (i.e., its ancestors in $G_n$) dynamically over a sequence of time steps. As is standard for these processes, all voters $V$ will be one of three types, live, dead, or neutral. Dead voters are those whose "children" (i.e., voters who delegate to them) we have already sampled. Live voters are voters who have decided to delegate, but whose children have not yet been sampled. Neutral voters are still in the "pool" and have yet to commit to a delegation. At time zero, $i$ is a live voter, there are no dead voters, and all other voters $V \setminus \{i\}$ are neutral. At each time step, we take some live voter $j$, sample which of the neutral voters choose to delegate to $j$, add these voters as live vertices, and update $j$ as dead. The procedure ends when there are no more live vertices, at which point the number of delegations received by $i$ is simply the total number of dead vertices.

Let us now describe this more formally. Following the notation of Alon and Spencer [3], let $Z_t$ denote the number of voters we sample to delegate at time $t$. Let $Y_t$ be the number of live vertices at time $t$; we have that $Y_0 = 1$. At time $t$, we remove one live vertex and add $Z_t$ more, so we have the recursion $Y_t = Y_{t-1} - 1 + Z_t$. We let $N_t$ be the number of neutral vertices at time $t$. We have that $N_0 = n - 1$, and $N_t = N_{t-1} - Z_t$. Note that after $t$ time steps, there are $t$ dead vertices and $Y_t$ live ones, so this is equiv-

alent to $N_t = n - 1 - t - Y_t$. To sample $Z_t$, we fix some live voter $j$ and ask how many of the neutral voters chose to delegate to $j$, conditioned on them not delegating to any of the dead voters. Note that when sampling at this step, there are $t-1$ dead voters and conditioned on the neutral voters not delegating to the dead ones, the probability they delegate to any of the other $n-t$ individuals (not including themselves) is exactly $\frac{p}{n-t}$, equally split between them for a total delegation probability of $p$. Hence $Z_t \sim \text{Bin}(N_{t-1}, \frac{p}{n-t}) \sim \text{Bin}(n-t-Y_{t-1}, \frac{p}{n-t})$. We denote by $\mathfrak{X}^D_{n,p}$ the random variable that counts the size of this branching process, i.e., the number of time steps until there are no more live vertices. Note that the number of delegations received by any voter has the same distribution as $\mathfrak{X}^D_{n,p}$.

Choose some constant $p'$ such that $p < p' < 1$. We will be comparing the $\mathfrak{X}^D_{n,p}$ to a graph branching process $\mathfrak{X}^G_{n,p'}$. The graph branching process is nearly identical, except the probability each of the neutral vertex joins our component is independent of the number of dead vertices and is simply $\frac{p'}{n}$. In other words, $Z_t \sim \text{Bin}(N_{t-1}, \frac{p'}{n})$. A key result about this branching process is the probability of seeing a component of a certain size $\ell$ decreases exponentially with $\ell$. In other words, there is some constant $c$ such that

$$\mathbb{P}_{\mathcal{D},M^C_q,n}[\mathfrak{X}^G_{n,p'} \leq c\log(n)] = 1 - o(1/n).$$

Take $C(n) = c\log(n)$. Note that as long as $t$ is such that $\frac{p}{n-t} \leq \frac{p'}{n}$, the sampling in the delegation branching process is dominated by the sampling in this graph branching process. Hence, as long as $\frac{p}{n-C(n)} \leq \frac{p'}{n}$, $\mathbb{P}[\mathfrak{X}^D_{n,p} \leq c\log(n)] \geq \mathbb{P}[\mathfrak{X}^G_{n,p'} \leq c\log(n)]$. Since $C(n) \in O(\log n)$, this is true for sufficiently large $n$, so for such $n$, $\mathbb{P}[\mathfrak{X}^D_{n,p} \leq c\log(n)] = 1 - o(1/n)$. By a union bound over all $n$ voters, this implies the desired result.

*Confidence Based Delegation satisfies* (6.2)

Let $\bar{q}$ be such that $\bar{q}(x) = 1 - q(x)$, so $\bar{q}$ represents the probability someone with competence $x$ does not delegate. Notice that $\mathbb{E}_{\mathcal{D}}[\bar{q}]$ is exactly the probability an arbitrary voter will not delegate. Let $q^+(x) = \bar{q}(x)x$ and let

$$\mu^* = \frac{\mathbb{E}_{\mathcal{D}}[q^+]}{\mathbb{E}_{\mathcal{D}}[\bar{q}]}.$$

Expanding the definition, we see that $\mu^*$ is exactly the expected value of a voter's competence, conditioned on them not voting. Let $\mu_{\mathcal{D}}$ the mean of the competence

distribution $\mathcal{D}$. We first show that $\mu^* > \mu_{\mathcal{D}}$. Indeed, since both $x$ and $\bar{q}(x)$ are strictly increasing functions of $x$, the Fortuin–Kasteleyn–Ginibre (FKG) inequality [78] tells us that $\mathbb{E}_{\mathcal{D}}[q^+] > \mathbb{E}_{\mathcal{D}}[\bar{q}] \cdot \mathbb{E}_{\mathcal{D}}[x] = \mathbb{E}_{\mathcal{D}}[\bar{q}] \cdot \mu_{\mathcal{D}}$. This implies that the expected competence conditioned on not delegating is strictly higher than the overall expected competence.

Next, we will show that for any constant $\gamma > 0$, with high probability, both $\sum_{i=1}^n p_i \leq (\mu + \gamma)n$ and $\sum_{i=1}^n \mathsf{weight}_i(G) p_i \geq (\mu^* - \gamma)n$. If we choose $\gamma = (\mu^* - \mu)/3$ and $\alpha = \gamma/2$, it follows that, with high probability,

$$\sum_{i=1}^n \mathsf{weight}_i(G) p_i - \sum_{i=1}^n p_i \geq 2\alpha n,$$

implying that (6.2) is satisfied.

Since the $p_i$s are bounded independent variables, it follows directly from Heoffding's inequality that $\sum_{i=1}^n p_i \leq n(\mu + \gamma)$ with high probability, so we now focus on showing $\sum_{i=1}^n \mathsf{weight}_i(G) \cdot p_i \geq (\mu^* - \gamma)n$ with high probability. To do this, we will first show that, with high probability, the delegation graph $G$ satisfies $\mathsf{dels}_i(G) \leq C(n)$ for all $i$ and $\mathsf{total\text{-}weight}(G) \geq n - C(n)\log^2 n$.

We showed in the earlier part of this proof that $\mathsf{dels}_i(G) \leq C(n)$ with high probability. We will now prove that $\mathbb{P}_{\mathcal{D}, M_q^C, n}[\mathsf{total\text{-}weight}(G) \geq n - C(n)\log^2 n \mid \mathsf{dels}_i(G) \leq C(n)] = 1 - o(1)$. To do this, we will first bound the number of voters that, with high probability, end up in cycles. Fix a voter $i$ and sample $i$'s delegation tree. Voter $i$ will only end up in a cycle if $i$ chooses to delegate to someone in this delegation tree. Since we are conditioning on $\mathsf{dels}_i(G) \leq C(n)$, the maximum size of this tree is $C(n)$. Hence, the total $\varphi$ weight that voter $i$ places on someone in the tree is at most $C(n)$, while the total weight they place on all voters is $n - 1$. Hence, the probability that $i$ delegates to someone in their tree can be at most $p \cdot C(n)/(n-1)$. Since this is true for each voter $i$, the expected number of voters in cycles is at most $np\frac{C(n)}{(n-1)} \in O(\log n)$. By Markov's inequality, the probability that more than $\log^2 n$ voters are in cycles is at most $np\frac{C(n)}{(n-1)\log^2 n} = O(1/\log n) = o(1)$.

Next, since we have conditioned on $\mathsf{dels}_i(G) \leq C(n)$, no single voter, and in particular no single voter in a cycle, can receive more than $C(n)$ delegations. So conditioned on the high probability event that there are at most $\log^2 n$ voters in cycles, there are at most $C(n)\log^2 n$ voters that delegate to those in cycles. This implies that $\mathsf{total\text{-}weight}(G) \geq n - C(n)\log^2 n + \log^2 n$ with high probability.

We now show that, conditioned on the graph satisfying these properties, the instance $(\vec{p}, G)$ satisfies $\sum_{i=1}^{n} \mathsf{weight}_i(G) \cdot p_i \geq n(\mu^* - \gamma)$ with high probability. Note that the competencies satisfy that those that don't delegate are drawn i.i.d. from the distribution of competencies conditioned on not delegating, which has mean $\mu^*$. Fix an arbitrary graph $G$ satisfying the properties. Suppose $M$ is the set of voters that do not delegate. Note that for each $i \in M$, $\mathsf{weight}_i(G) \leq C(n)$, by assumption. Further $\sum_{i \in M} \mathsf{weight}_i(G) \geq n - C(n)\log^2(n)$. Hence, when we sample the non-delegator $p_i$s, $\mathbb{E}[\sum_{i \in M} \mathsf{weight}_i(G) \cdot p_i] \geq (n - C(n)\log^2(n)) \cdot \mu^*$. Moreover,

$$\mathrm{Var}[\sum_{i \in M} \mathsf{weight}_i(G) \cdot p_i] \leq \sum_{i \in M} \mathsf{weight}_i(G)^2 \leq C(n) \cdot n.$$

This follows from the fact that $\mathrm{Var}[p_i] \leq 1$ and that we have fixed the graph $G$ and hence $\mathsf{weight}_i(G)$ for each $i$, so these terms can all be viewed as constants. In addition, we know that, for each voter $i$, $\mathsf{weight}_i(G) \leq C(n)$, and $\sum_{i=1}^{n} \mathsf{weight}_i(G) \leq n$. Hence, we can directly apply Chebyshev's inequality:

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}, M_q^C, n}\left[\sum_{i \in M} \mathsf{weight}_i(G) p_i < n(\mu^* - \gamma)\right] &< \frac{\mathrm{Var}[\sum_{i \in M} \mathsf{weight}_i(G) p_i]}{(\mathbb{E}[\sum_{i \in M} \mathsf{weight}_i(G) p_i] - n(\mu^* - \gamma))^2} \\
&\leq \frac{nC(n)}{(\gamma n - C(n)\log^2(n)\mu^*)^2} \\
&\in o(1),
\end{aligned}$$

where the final step holds because the numerator is $o(n^2)$ and the denominator is $\Omega(n^2)$. Hence, $\sum_{i \in M} \mathsf{weight}_i(G) p_i \geq n(\mu^* - \gamma)$ with high probability, as needed.

To summarize, we have proved that, conditioned on $\mathsf{dels}_i(G) \leq C(n)$ for all $i$ and $\mathsf{total\text{-}weight}(G) \geq n - C(n)\log^2 n, \sum_{i=1}^{n} \mathsf{weight}_i(G) \cdot p_i \geq n(\mu^* - \gamma/3)$ occurs with high probability. Given that, conditioned on $\mathsf{dels}_i(G) \leq C(n), \mathsf{total\text{-}weight}(G) \geq n - C(n)\log^2 n$ occurs with high probability and that $\mathsf{dels}_i(G) \leq C(n)$ occurs with high probability, we can conclude by the chain rule that the intersection of these events hold with high probability. Given that the probability of any of this event is greater than the probability of the intersection, we can conclude that $\sum_{i=1}^{n} \mathsf{weight}_i(G) \cdot p_i \geq n(\mu^* - \gamma/3)$ occurs with probability $1 - o(1)$, as desired. □

## 6.5 Continuous General Delegation Model

Finally, we study a model in which voters delegate with fixed probability, and they do so by picking a voter according to a continuous increasing delegation function. This is a general model in which delegations can either go to more or less competent neighbors but where more competent voters are more likely to be chosen over less competent ones.

Formally, let $M^S_{p,\varphi} = (q^p, \varphi)$ where $q^p$ is a constant function equal to $p$, that is, $q^p(x) = p$ for all $x \in [0,1]$, and $\varphi(x,y)$ is non-zero, continuous, and increasing in $y$. We then have the following.

**Theorem 6.5.1** (Continuous General Delegation Model)**.** *All models $M^S_{p,\varphi}$ with $p \in (0,1)$ and $\varphi$ that is non-zero, continuous, and increasing in its second coordinate satisfy probabilistic do no harm and probabilistic positive gain with respect to the class $\mathfrak{D}^C$ of all continuous distributions.*

*Proof.* Fix $M^S_{p,\varphi}$ and and $\mathcal{D} \in \mathfrak{D}^C$. Note that since $\varphi$ is continuous and always positive on the compact set $[0,1]^2$, $\varphi$ is in fact uniformly continuous and there are bounds $L, U \in \mathbb{R}^+$ such that $\varphi$ is bounded in the interval $[L,U]$. Additionally, we can assume without loss of generality that for all $x \in [0,1]$, $\mathbb{E}_{\mathcal{D}}[\varphi(x,\cdot)] = 1$. Indeed, $\mathbb{E}_{\mathcal{D}}[\varphi(x,\cdot)]$ is a positive, continuous function of $x$, so replacing $\varphi$ by $\varphi'(x,y) = \varphi(x,y)/\mathbb{E}_{\mathcal{D}}[\varphi(x,\cdot)]$ induces the same model and satisfies the desired property.

*The Continuous General Delegation Model satisfies* (6.1).

Our goal is to show there is some $C(n) \in O(\log n)$ such that, with high probability, no voter receives more than $C(n)$ delegations. To do this, just as in the proof of Theorem 6.4.1, we consider a branching process of the delegations received beginning with some voter $i$. We will show that under minimal conditions on the sampled competencies (which all occur with high probability), this branching process will be dominated by a well-known *subcritical multi-type Poisson branching process* [33], which has size $O(\log n)$ with high probability.

For a fixed competence vector $\vec{p}_n$, the branching process for the number of delegations received by a voter $i$ works as follows. We keep track of three sets of voters: those that are live at time $t$ ($L_t$), those dead at time $t$ ($D_t$), and those neutral at time $t$ ($N_t$). Unlike in the proof of Theorem 6.4.1, where it was sufficient to keep track of the number of voters in each category, here we must keep track of the voter identities

as well, as they do not all delegate with the same probability. At time zero, the only live voter is voter $i$ and the rest are neutral, so $L_0 = \{i\}$, $D_0 = \emptyset$, and $N_0 = [n] \setminus \{i\}$. As long as there are still live voters, we sample the next set of delegating voters $Z_t$ in time $t$ by choosing some live voter $j \in R_{t-1}$ and sampling its children. Once $j$'s children are sampled, $j$ becomes dead, and $j$'s children become live. All voters that did not delegate and were not delegated to remain neutral. The children are sampled independently; the probability they are included is the probability they delegate to $j$ conditioned on them not delegating to the dead voters in $D_{t-1}$. For each voter $k \in N_{t-1}$, $k$ will be included with probability

$$p \cdot \frac{\varphi(p_k, p_j)}{\sum_{k' \in [n] \setminus (D_{t-1} \cup \{k\})} \varphi(p_k, p_{k'})}.$$

This is precisely the probability $k$ delegates to $j$ conditioned on them not delegating to any voter in $D_{t-1}$. We continue this process until there are no more live voters, at which point the number of delegations is simply the number of dead voters, or equivalently, the total number of time steps. We denote by $\mathfrak{X}^D_{\vec{p}_n, i}$ the size of the branching process parameterized by competencies $\vec{p}_n$ and a voter $i \in [n]$.

Our goal will be to compare $\mathfrak{X}^D_{\vec{p}_n, i}$ to the outcome of a well-known multi-type Poisson branching process. In this branching process, there are a fixed finite number $k$ of types of voters.[8] The process itself is parameterized by a $k \times k$ matrix $M$, where $M_{\tau\tau'}$ is the expected number of children of type $\tau'$ a voter of type $\tau$ will have. The process is additionally parameterized by the type $\tau \in [k]$ of the starting voter. The random variable $Y_t$ keeps track of the number of live voters of each type; it is a vector of length $k$, where the $\tau$th entry is the number of live voters of type $\tau$. Hence, $Y_0 = e_\tau$, the (basis) vector with a 1 in entry $\tau$ and an entry 0 for all other types. We sample children by taking an arbitrary live voter of type $\tau'$ (the $\tau'$ component in $Y_{t-1}$ must be positive, indicating that there is such a voter), and sampling its children $Z_t$, which is also a vector of length $k$, each entry indicating the number of children of that type. The vector $Z_t$ is sampled such that the $\tau''$ entry is from the $\text{Pois}(M_{\tau'\tau''})$ distribution. That is, children of different types are sampled independently from a Poisson distribution, with the given expected value. We have the recursion $Y_t = Y_{t-1} + Z_t - e_{\tau'}$.

Note that this means that there is no "pool" of voters to choose from; in fact, it is

[8] In the literature, these are often called *particles*, but to be consistent with our other branching processes, we call them voters here.

possible for this process to grow unboundedly large (see [3, Section 11.6] for the classical description of the single-type Poisson branching process). Nonetheless, this process will still converge often enough to remain useful. We denote by $\mathfrak{X}^P_{M,\tau}$ the random variable that gives the size of this branching process, parameterized by expected-children matrix $M$ and starting voter type $\tau \in [k]$. Such a branching process is considered *sub-critical* if the largest eigenvalue of $M$ is strictly less than 1 [33]. In such a case, if we begin with voter of any type $\tau \in [k]$, the probability of the branching process surviving $\ell$ steps decreases exponentially in $\ell$. Hence, there is some $c$ such that for all $\tau \in [k]$,

$$\mathbb{P}[\mathfrak{X}^P_{M,\tau} \leq c \log(n)] = 1 - o(1/n).$$

To compare these branching processes, we make a sequence of adjustments to the original branching process that at each step creates a dominating branching process slightly closer in flavor to the multi-type Poisson. In the end, we will be left with a sub-critical multi-type Poisson process that we can bound.

Fix some $\varepsilon > 0$, which is a parameter in all of our steps. Later, we will choose $\varepsilon$ to be sufficiently small (specifically, such that $p\frac{(1+\varepsilon)^3}{1-2\varepsilon} < 1$) to ensure that the Poisson branching process is sub-critical. To convert from our delegation branching process to the Poisson branching process, we take a voter's type to be their competence (which completely characterizes their delegation behavior). However, to compare to the Poisson process, there must be a finite number of types. Hence, we partition the interval $[0, 1]$ into $B$ buckets, each of size $1/B$, such that voters in the same bucket delegate and are delegated to "similarly". We choose $B$ large enough such that all points in $[0, 1]^2$ within a distance of $\sqrt{2}/B$ of each other differ in $\varphi$ by at most $L \cdot \varepsilon$. (Recall that the range of $\varphi$ is in the interval $[L, U]$.) This is possible since $\varphi$ is uniformly continuous. Further, this implies any points $(x, y), (x', y')$ within a square with side length $1/B$ have the property that $\varphi(x, y) \leq \varphi(x', y') + L \cdot \varepsilon \leq (1 + \varepsilon) \cdot \varphi(x', y')$. Note that $B$ depends only on $\varphi$ and $\varepsilon$, and hence is a constant with respect to the number of voters $n$.

We say a voter $i$ is of type $\tau$ if $\frac{\tau-1}{B} < p_i \leq \frac{\tau}{B}$ for $1 \leq \tau \leq B$ (with a non-strict inequality for $\tau = 1$, so 0 is of type 1). Let $S_\tau = (\frac{\tau-1}{B}, \frac{\tau}{B}]$ be the set of competencies of type $\tau$ (except that, in the case that $\tau = 1$, we take $S_1$ to be the closed interval $[0, \frac{1}{B}]$). Let $\pi_\tau = \mathcal{D}[S_\tau]$ be the probability that a voter has type $\tau$. Since the types form a partition of $[0, 1]$, we have that $\sum_{\tau=1}^{B} \pi_\tau = 1$.

For any two types $\tau, \tau'$, we define

$$\varphi'(\tau, \tau') = \sup_{(x,y) \in S_\tau \times S_{\tau'}} \varphi(x, y).^{9}$$

We abuse notation by extending $\varphi'$ to operate directly on competencies in $[0, 1]$ by first converting competencies to types and then applying $\varphi'$. Then, $\varphi'$ has the property that for any $p_i, p_j \in [0, 1]$,

$$\varphi(p_i, p_j) \leq \varphi'(p_i, p_j) \leq (1 + \varepsilon)\varphi(p_i, p_j).$$

We have that for all $\tau$, if $x \in S_\tau$, then

$$\sum_{\tau'=1}^{B} \varphi'(\tau, \tau')\pi_{\tau'} = \mathbb{E}_{\mathcal{D}}[\varphi'(x, \cdot)] \leq (1 + \varepsilon) \cdot \mathbb{E}_{\mathcal{D}}[\varphi(x, \cdot)] = (1 + \varepsilon).$$

Hence, we define

$$\tilde{\varphi}(\tau, \tau') = \varphi'(\tau, \tau') \cdot \frac{(1 + \varepsilon)}{\sum_{\tau''=1}^{B} \varphi'(\tau, \tau'')\pi_{\tau''}}.$$

We again abuse notation to allow $\tilde{\varphi}$ to operate directly on competencies. We have that $\tilde{\varphi}(x, y) \geq \varphi'(x, y) \geq \varphi(x, y)$ for all competencies $x, y \in [0, 1]$ and further, for all $\tau$, $\sum_{\tau'=1}^{B} \tilde{\varphi}(\tau, \tau')\pi_{\tau'} = 1 + \varepsilon$.

The Poisson branching process we will eventually compare to is one with $B$ types parameterized by the expected-children matrix $M$, where

$$M_{\tau\tau'} = p\frac{(1 + \varepsilon)^2}{1 - 2\varepsilon}\tilde{\varphi}(\tau, \tau').$$

First, we show that $M$ has largest eigenvalue strictly less than 1 (for our choice of $\varepsilon$), so that the branching process will be subcritical. Indeed, $M$ has only positive entries, so we need only exhibit an eigenvector with all nonnegative entries such that the associated eigenvalue is strictly less than 1. The Perron-Frobenius theorem tells us this eigenvalue must be maximal.

The eigenvector we consider is $\vec{\pi} = (\pi_1, \ldots, \pi_B)$ (which has nonnegative entries, as each $\pi_\tau$ is a probability). We show it has eigenvalue $p\frac{(1+\varepsilon)^3}{1-2\varepsilon}$, strictly less than 1 due

[9] Note that, because $\varphi$ is increasing in its second coordinate, one can actually write $\tilde{\varphi}(\tau, \tau') = \sup_{x \in S_\tau} \varphi(x, \frac{\tau'}{B})$.

to our choice of $\varepsilon$. We show it has eigenvalue $p\frac{(1+\varepsilon)^3}{1-2\varepsilon}$, strictly less than 1 due to our choice of $\varepsilon$. Indeed, we have that

$$(M\vec{\pi})_\tau = \sum_{\tau'=1}^{B} \pi_\tau \tilde{\varphi}(\tau, \tau')\pi_{\tau'} = \pi_\tau p \frac{(1+\varepsilon)^3}{1-2\varepsilon}$$

by the definition of $\tilde{\varphi}$. Hence, $\vec{\pi}$ is our desired eigenvector.

Since $\mathfrak{X}^P_{M,\tau}$ is sub-critical for all $\tau$, we have that there is some $c$ such that for all $\tau \in [B]$, $\mathbb{P}[\mathfrak{X}^P_{M,\tau} \leq c\log(n)] = 1 - o(1/n)$. We take $C(n) = c\log(n)$.

Now we consider our branching process, $\mathfrak{X}^D_{\vec{p},i}$. To make the comparison, we will need some minimal concentration properties. We first show that the sampled competencies $\vec{p}$ satisfy these properties with high probability, and then show that, conditioned on these properties, the branching process $\mathfrak{X}^D_{\vec{p},i}$ is easily comparable to a Poisson process. The properties are the following:

1. For each voter $i \in [n]$, $\sum_{j\neq i} \varphi(p_i, p_j) \geq (1-\varepsilon)\cdot n$.
2. For each type $\tau \in [B]$, the number of voters of type $\tau$, $|\{i \mid p_i \in S_\tau\}| \leq (1+\varepsilon)\pi_\tau n$.

For the first property, fix the competence $p_i$ of a single voter $i$. Then when sampling the $p_j$s, $\sum_{j\neq i} \varphi(p_i, p_j)$ is the sum $n-1$ independent variables, all in the interval $[L, U]$, with mean 1. Hence, by Hoeffding's inequality, for all competencies $c$, $\mathcal{D}^n[\sum_{j\neq i} \varphi(p_i, p_j) \geq (1-\varepsilon)n \mid p_i = c] = 1 - o(1/n)$, where the $o(1/n)$ term is independent of $c$. By the law of total probability, this implies that even when $p_i$ is sampled as well, the $1 - o(1/n)$ bound continues to hold. By a union bound over all $n$ voters, this holds for everybody with probability $1 - o(1)$.

For the second property, note that the number of voters of type $\tau$ follows a $\mathrm{Bin}(n, \pi_\tau)$ distribution. A simple application of Hoeffding's inequality implies that for this $\tau$, $|\{i \mid p_i \in S_\tau\}| \leq (1+\varepsilon)\pi_\tau n$ (note that this holds even in the extreme cases where $\pi_\tau = 0$ or $\pi_\tau = 1$). As the number $B$ of types is fixed and independent of $n$, a union bound over all $B$ types implies this holds for all $\tau$ with probability $1 - o(1)$.

Now fix some voter competencies $\vec{p}$ such that both properties hold. We will first upper bound the probability a voter of type $\tau$ delegates to a voter of type $\tau'$. Hence, we can compare our branching process to one with these larger probabilities, and this will only dominate our original process.

To that end, since $|D_{t-1}| = t - 1 \leq t$ (recall that $D_{t-1}$ consists of the dead voters at time $t-1$), using the first property, we have that for all $i \in [n]$,

$$\sum_{j \in [n] \setminus (D_{t-1} \cup \{i\})} \varphi(p_i, p_j) \geq (1-\varepsilon)n - U \cdot t.$$

Hence, as long as $t \leq \varepsilon n/U$, $\sum_{j \in [n] \setminus (D_{t-1} \cup \{i\})} \varphi(p_i, p_j) \geq (1-2\varepsilon)n$.

Including the fact that $\varphi(p_i, p_j) \leq \tilde{\varphi}(p_i, p_j)$ for all $p_i$ and $p_j$, we have that for all time steps $t \leq \varepsilon n/U$,

$$p \cdot \frac{\varphi(p_i, p_j)}{\sum_{k' \in [n] \setminus (D_{t-1} \cup \{k\})} \varphi(p_i, p_{k'})} \leq \frac{p}{n} \cdot \frac{\tilde{\varphi}(p_i, p_j)}{(1-2\varepsilon)}.$$

Note that for sufficiently large $n$, $C(n) \leq \varepsilon n/U$, so from now on we restrict ourselves to such $n$.

Further, note that by the second property, there will never be more than $(1+\varepsilon)\pi_\tau n$ neutral voters of type $\tau$. Hence, if we take a voter of type $\tau'$ at time step $t \leq C(n)$, the number of children it will have of type $\tau$ will be stochastically dominated by a $\mathrm{Bin}((1+\varepsilon)\pi_\tau n, \frac{p}{n} \cdot \frac{\tilde{\varphi}(p_i,p_j)}{(1-2\varepsilon)})$, and this is independent for each $\tau$. As $n$ grows large, this distribution approaches a $\mathrm{Pois}(p\frac{(1+\varepsilon)}{1-2\varepsilon}\tilde{\varphi}(\tau, \tau'))$. In particular, this means that for sufficiently large $n$, it will be stochastically dominated by a $\mathrm{Pois}(p\frac{(1+\varepsilon)^2}{1-2\varepsilon}\tilde{\varphi}(\tau, \tau'))$ distribution (note the extra $(1+\varepsilon)$ factor). Hence, if voter $i$ is of type $\tau$, up to time $t \leq C(n)$, $\mathfrak{X}^D_{\vec{p},i}$ is dominated by $\mathfrak{X}^P_{M,\tau}$, so

$$\mathbb{P}_{\mathcal{D}, M^S_{\vec{p},\varphi}.n}[\mathfrak{X}^D_{\vec{p},i} \geq C(n)] \geq \mathbb{P}_{\mathcal{D}, M^S_{\vec{p},\varphi}.n}[\mathfrak{X}^P_{M,\tau} \geq C(n)] = 1 - o(1/n).$$

A union bound over all $n$ voters tells us this is true for all voters simultaneously with probability $1 - o(1)$, as needed.

*The Continuous General Delegation Model satisfies* (6.2).

To show (6.2) holds, we first show the following.

Let $\mu_{\mathcal{D}}$ be the mean of the competence distribution $\mathcal{D}$. For a fixed $x$, let $\varphi^+_x(y)$ be the function $\varphi(x, y) \cdot y$. We show that there is some $c > 0$ such that for all $x \in [0, 1]$,

$$\mathbb{E}_{\mathcal{D}}[\varphi^+_x] \geq \mu_{\mathcal{D}} + c. \tag{6.18}$$

Indeed, if we view $\mathbb{E}_{\mathcal{D}}[\varphi^+_x]$ as a function of $x$ for $x \in [0, 1]$, first note that it is a con-

tinuous function on a compact set, and hence it attains its minimum. Further, for all $x \in [0,1]$, since $\varphi(x,y)$ and $y$ are both increasing functions of $y$, by the FKG inequality [78],

$$\mathbb{E}_{\mathcal{D}}[\varphi^+] > \mathbb{E}_{\mathcal{D}}[\varphi(x,\cdot)] \cdot \mu_{\mathcal{D}} = \mu_D,$$

since, by assumption, $\mathbb{E}_{\mathcal{D}}[\varphi(x,\cdot)] = 1$. Hence, this attained minimum must be strictly larger than $\mu$, implying (6.18).

Since $\varphi(x,y)$ is normalized so that $\mathbb{E}_{y\sim\mathcal{D}}[\varphi(x,y)] = 1$, $\mathbb{E}_{y\sim\mathcal{D}}[\varphi(x,y)\cdot y]$ is the expected competence of the voter to whom someone of competence $x$ delegates to (prior to other competencies being drawn). Hence, (6.18) tells us that "on average", all voters (regardless of competence) tend to delegate to those with competence strictly above the mean. Ideally, we would choose $\alpha \approx c/2$ and hope that some concentration result tells us that the weighted competencies after delegation will be strictly above $\mu + c/2$ (the mean of all competencies will be close to $\mu$ by standard concentration results). However, proving this concentration result is surprisingly subtle, as there are many dependencies between different voter delegations. Indeed, if one voter with high competence and many delegations chooses to delegate "downwards" (that is, to someone with very low competence), this can cancel out all of the "expected" progress we had made thus far. Hence, the rest of this proof involves proving concentration *does* in fact hold. We prove this by breaking up the process of sampling instances into much more manageable pieces, where, in each, as long as nothing goes "too" wrong, concentration will hold.

In particular, we will prove that for all $\gamma > 0$, with high probability,

$$\sum_{i=1}^{n} \mathsf{weight}_i(G) \cdot p_i - \sum_{i=1}^{n} p_i \geq (c(1-p) - \gamma)n. \tag{6.19}$$

Fix such a $\gamma$. As in the previous part, fix $\varepsilon > 0$ which will paramaterize our steps. We will later choose $\varepsilon$ sufficiently small to get our desired result (precisely $\varepsilon$ such that $6\varepsilon + \varepsilon^2 < \gamma$). By choosing $\gamma < c(1-p)$, this value is positive, so we can choose $\alpha = \frac{c(1-p)-\gamma}{2}$ which proves Equation (6.2)

To that end, we define a sequence of six sampling steps that together are equivalent to the standard sampling process with respect to $\mathcal{D}$ and $M^S_{p,\varphi}$. In each step, we will show that with high probability, nothing "goes wrong", and conditioned on nothing

going wrong in all these steps, we will get the $\alpha$ improvement that we desire. The six steps are as follows:

1. Sample a set $M \subseteq [n]$ of voters that choose not to delegate. Each voter is included independently with probability $p$.

2. Sample competencies $p_i$ for $i \in [n] \setminus M$. Each $p_i$ is sampled i.i.d. from $\mathcal{D}$.

3. Sample competencies $p_j$ for $j \in M$. Each $p_j$ is sampled i.i.d. from $\mathcal{D}$.

4. Sample a set $R \subseteq [n] \setminus M$ of delegators that delegate to those in $M$. Each voter $i \in [n] \setminus M$ is included independently with probability $\frac{\sum_{j \in M} \varphi(p_i, p_j)}{\sum_{j \in [n] \setminus \{i\}} \varphi(p_i, p_j)}$, that is, the total $\varphi$ weight they put on voters in $M$ divided by the total $\varphi$ weight they put on all voters.

5. Sample delegations of voters in $[n] \setminus (M \cup R)$. At this point, we are conditioning on such voters delegating, and when they do delegate, they do so to voters in $[n] \setminus M$. Hence, for each $i \in [n] \setminus (M \cup R)$, they delegate to $j \in [n] \setminus (M \cup \{i\})$ with probability $\frac{\varphi(p_i, p_j)}{\sum_{j' \in [n] \setminus (M \cup \{i\})} \varphi(p_i, p_{j'})}$.

6. Sample delegations of voters in $R$. At this point, we are conditioning on such voters delegating to those in $M$. Hence, for each $i \in R$, they choose to delegate to $j \in M$ with probability $\frac{\varphi(p_i, p_j)}{\sum_{j' \in M} \varphi(p_i, p_{j'})}$.

We now analyze each step, describing what could "go wrong". Let $\mathcal{E}_1, \ldots, \mathcal{E}_6$ be the events that nothing goes wrong in each of the corresponding steps. We define these events formally below. Our goal is to show that $\mathbb{P}_{\mathcal{D}, M^S_{p,\varphi}, n}[\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_6] = 1 - o(1)$.

• Let $\mathcal{E}_1$ be the event that $(p-\varepsilon) \cdot n \leq |M| \leq (p+\varepsilon) \cdot n$. Note that $M$ is the sum of $n$ independent Bernouilli random variables with success probability $p$. It follows directly from a union bound over both variants of Hoeffding's inequality that

$$\mathbb{P}_{\mathcal{D}, M^S_{p,\varphi}, n}[(p - \varepsilon) \cdot n \leq |M| \leq (p + \varepsilon) \cdot n] = 1 - o(1).$$

• Let $\mathcal{E}_2$ be the event that $\sum_{i \in [n] \setminus M} p_i \leq n(\mu + \varepsilon)(1 - p + \varepsilon)$. Note that $\sum_{i \in [n] \setminus M} p_i$ is the sum of $n - |M|$ i.i.d. random variables with mean $\mu$. Conditioning on event $\mathcal{E}_1$, $|M|$ is lower bounded by $n(p - \varepsilon)$, implying that $n - |M| \leq n(1 - p + \varepsilon)$ as well. It

follows from Hoeffding's inequality that

$$\mathbb{P}_{\mathcal{D},M^S_{p,\varphi},n}\left[\sum_{i\in[n]\setminus M} p_i \le n(\mu+\varepsilon)(1-p+\varepsilon) \;\middle|\; \mathcal{E}_1\right] = 1-o(1)$$

which, combined with $\mathbb{P}_{\mathcal{D},M^S_{p,\varphi},n}[\mathcal{E}_1] = 1-o(1)$, proves that $\mathcal{E}_1 \cap \mathcal{E}_2$ occurs with probability $1-o(1)$.

- Let $\mathcal{E}_3$ be the event consisting of all instances $(\vec{p}, G)$ such that

$$\frac{\sum_{j\in M}\varphi(p_i,p_j)\cdot p_j}{\sum_{j\in M}\varphi(p_i,p_j)} \ge \frac{(1-\varepsilon)}{(1+\varepsilon)}(\mu+c)$$

for all $i \in [n]\setminus M$.

We show $\mathcal{E}_3$ occurs with high probability conditional on $\mathcal{E}_1$ and $\mathcal{E}_1$ (conditioning on $\mathcal{E}_2$ is unnecessary. but makes the final statement easier). Fix a set of voters $M$ and $p_i$ for $i \in [n]\setminus M$ satisfying $\mathcal{E}_1$ and $\mathcal{E}_2$. For each $i \in [n]\setminus M$, we will show that with probability $1-o(1/n)$, when we sample the $p_j$s for $j \in M$, they satisfy

$$\sum_{j\in M}\varphi(p_i,p_j) \le |M|(1+\varepsilon) \tag{6.20}$$

and

$$\sum_{j\in M}\varphi(p_i,p_j)\cdot p_j \ge |M|(1-\varepsilon)(\mu+c). \tag{6.21}$$

(6.20) follows from the fact that $\sum_{j\in M}\varphi(p_i,p_j)$ is the sum of $|M|$ bounded independent random variables

with mean $\mathbb{E}_{y\sim\mathcal{D}}[\varphi(p_i,y)] = 1$. By Hoeffding's inequlaity, since $|M|$ is linear in $n$, $\sum_{j\in M}\varphi(p_i,p_j)$ is at most $|M|(1+\varepsilon)$ with probability $1-o(1/n)$.

(6.21) follows from the fact that $\sum_{j\in M}\varphi(p_i,p_j)\cdot p_j$ is also the sum of $|M|$ bounded independent random variables with mean $\mathbb{E}_{\mathcal{D}}[\varphi^+_{p_i}]$. Again, since we have conditioned on $\mathcal{E}_1$, $|M|$ is lower bounded by $(p-\varepsilon)n$, which by Hoeffding's inequality implies that $\sum_{j\in M}\varphi(p_i,p_j)p_j$ is at least $|M|(1-\varepsilon)\,\mathbb{E}_{\mathcal{D}}[\varphi_{p_i}]$ with probability $1-o(1/n)$.

Finally, we can conclude via a union bound that $\frac{\sum_{j\in M}\varphi(p_i,p_j)\cdot p_j}{\sum_{j\in M}\varphi(p_i,p_j)} \ge \frac{(1-\varepsilon)}{(1+\varepsilon)}(\mu+c)$ with probability $1-o(1/n)$ for any $i \in [n]\setminus M$. Hence, by another union bound over the at most $n$ voters $i \in [n]\setminus M$, $\frac{\sum_{j\in M}\varphi(p_i,p_j)\cdot p_j}{\sum_{j\in M}\varphi(p_i,p_j)} \ge \frac{(1-\varepsilon)}{(1+\varepsilon)}(\mu+c)$ for all $i \in [n]\setminus M$ with high probability.

By the law of total probability, $\mathcal{E}_3$ conditioned on $\mathcal{E}_1$ and $\mathcal{E}_2$ occurs with probability $1 - o(1)$, which proves that $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ occurs with probability $1 - o(1)$ by the chain rule.

• Let $\mathcal{E}_4$ be the entire sample space. Nothing can "go wrong" during this sampling step. So trivially, $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ occurs with probability $1 - o(1)$.

• Let $\mathcal{E}_5$ be the event that $\mathsf{dels}_i(G) \leq C(n)$ for all $i \in [n] \setminus M$ and $\mathsf{total\text{-}weight}(G) \geq n - C(n)^2 \log(n)$ in the subgraph $G$ sampled (i.e., with delegations only from voters not in $R$ or $M$). We will show $\mathcal{E}_5$ occurs with high probability even when we sample a full delegation graph (that is, samples delegations for all voters), which implies it continues to hold even when we sample only some delegations (recall that at this step we have only sampled delegations from voters in $[n] \setminus (M \cup R)$).

The proof of this is very similar to the one in Theorem 6.4.1, with one extra step to allow for different $\varphi$ weights.

It was proved in the previous part of this proof that, for all voters $i$, we have that $\mathsf{dels}_i(G) \leq C(n)$ with probability $1 - o(1)$ (not conditioned on anything) when we sample entire delegation graphs, so we can safely condition on this fact. We now prove that $\mathbb{P}_{\mathcal{D},M^S_{p,\varphi},n}[\mathsf{total\text{-}weight}(G_n) \geq n - O(\log^3 n) \mid \mathsf{dels}_i(G) \leq C(n)] = 1 - o(1)$.

We begin by bounding the number of voters that end up in cycles. Fix some voter $i$, and let us begin by sampling their delegation tree.

Since we are conditioning on the tree having size at most $C(n)$, the most weight that voter $i$ can place on all of the voters in $i$'s delegation tree is $U \cdot C(n)$. The minimum weight that $i$ can place on all voters is $L(n-1)$. Hence, the probability that $i$ delegates to someone in $i$'s tree conditional the delegation tree having size at most $C(n)$ is at most $p \cdot \frac{U \cdot C(n)}{L \cdot (n-1)}$. Since $i$ was arbitrary, this implies that the expected number of voters in cycles can be at most $n \cdot p \cdot \frac{U \cdot C(n)}{L \cdot (n-1)} \in O(\log n)$.

Applying Markov's inequality just as in the analogous proof in the previous section, the probability that more than $\log^2 n$ voters are in cycles is at most $np\frac{UC(n)}{L(n-1)\log^2 n} = O(1/\log n) = o(1)$. Further, the total number of people that could delegate to voters in cycles is at most $C(n)$ times the number of voters in cycles. Hence, with probability $1 - o(1)$, there are at most $C(n) \cdot \log^2 n$ voters delegating to those in cycles. This implies the desired bound. Hence, we have proved that $\mathbb{P}_{\mathcal{D},M^S_{p,\varphi},n}[\mathcal{E}_5] = 1 - o(1)$. Since we have already shown that $\mathbb{P}_{\mathcal{D},M^S_{p,\varphi},n}[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4] = 1 - o(1)$, a union bound implies that $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5$ occurs with probability $1 - o(1)$ as well.

• We now consider the sixth step. To define $\mathcal{E}_6$, we need some new notation. Fix

competencies $\vec{p}$ and a partial delegation graph $G$ such that $(\vec{p}, G)$ is in the first five events. We define $Q_i$ for $i \in R$ to be the random variable representing the competence of the voter to whom $i$ delegates. Since we know $i$ delegates to a voter in $M$, note that

$$Q_i(G) = p_j \text{ with probability } \frac{\varphi(p_i, p_j)}{\sum_{j' \in M} \varphi(p_i, p_{j'})} \text{ for all } j \in M.$$

Let $\mathcal{E}_6$ be the event consisting of all instance $(\vec{p}, G)$ such that that $\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i(G) \geq \frac{(1-\varepsilon)^2}{1+\varepsilon}(\mu + c)(1 - p - 2\varepsilon) \cdot n$. We show that $\mathbb{P}_{\mathcal{D}, M^S_{p,\varphi}, n}[\mathcal{E}_6 \mid \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_5] = 1 - o(1)$. This, combined with the the fact $\mathbb{P}_{\mathcal{D}, M^S_{p,\varphi}, n}[\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_5] = 1 - o(1)$ (shown earlier), implies that $\mathbb{P}_{\mathcal{D}, M^S_{p,\varphi}, n}[\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_6] = 1 - o(1)$. It follows from the definition of $Q_i$ that

$$\mathbb{E}[Q_i] = \sum_{j \in M} \frac{\varphi(p_i, p_j)}{\sum_{j' \in M} \varphi(p_i, p_{j'})} \cdot p_j = \frac{\sum_{j \in M} \varphi(p_i, p_j) \cdot p_j}{\sum_{j \in M} \varphi(p_i, p_j)}.$$

By conditioning on $\mathcal{E}_3$, we have that $\mathbb{E}[Q_i] \geq \frac{(1-\varepsilon)}{(1+\varepsilon)}(\mu + c)$ for each $i \in R$. Hence, $\mathbb{E}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i] \geq (n - |M| - C(n)^2 \log(n)) \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot (\mu + c)$, since we are conditioning on $\mathcal{E}_3$ and $\mathcal{E}_5$. Further, for sufficiently large $n$, $C(n)^2 \log(n) \leq \varepsilon n$; since we are conditioning on $\mathcal{E}_1$, $|M| \leq (p + \varepsilon)n$, so we have that for sufficiently large $n$,

$$\mathbb{E}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i] \geq (1 - p - 2\varepsilon) \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot (\mu + c) \cdot n \in \Omega(n).$$

Next, consider $\mathrm{Var}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i]$. Since each $Q_i$ takes on values in $[0, 1]$, $\mathrm{Var}[Q_i] \leq 1$. Further, each summand is independent, as each $Q_i$ is independent and we have fixed $G$, so we can can view $\mathsf{dels}_i(G)$ as a constant. Hence, $\mathrm{Var}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i] \leq \sum_{i \in R} \mathsf{dels}_i(G)^2 \in o(n^2)$ since, for all $i$, $\mathsf{dels}_i(G) \leq C(n) \in O(\log n)$ and $\sum_i \mathsf{dels}_i(G) \leq n$. Hence,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}, M^S_{p,\varphi}, n}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i < \frac{(1-\varepsilon)^2}{1+\varepsilon}(\mu + c)(1 - p - 2\varepsilon) \cdot n] \\
&\leq \mathbb{P}_{\mathcal{D}, M^S_{p,\varphi}, n}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i < (1-\varepsilon)\,\mathbb{E}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i]] \\
&\leq \frac{\mathrm{Var}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i]}{\varepsilon^2 \cdot \mathbb{E}[\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i]^2} \in o(1)
\end{aligned}$$

where the second inequality is due to Chebyshev's inequality, which is $o(1)$ because

the numerator is $o(n^2)$ and the denominator is $\Omega(n^2)$. This implies the desired result.

Finally, we show that for all instance $(\vec{p}, G) \in \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_6$, (6.2) holds, and hence so does (6.19). We have that $\sum_{i=1}^n w_i(G) \cdot p_i = \sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i(G) + \sum_{j \in M} p_j$, because in $G$ each voter $i \in R$ delegates all of their $\mathsf{dels}_i(G)$ votes to the voter in $M$ with competence $Q_i(G)$. Hence, $\sum_{i=1}^n w_i(G) \cdot p_i - \sum_{i=1}^n p_i = \sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i(G) - \sum_{i \in [n] \setminus (M \cup R)} p_i$. Since $(\vec{p}, G) \in \mathcal{E}_2$, we have that $\sum_{i \in [n] \setminus M} p_i \leq n(\mu + \varepsilon)(1 - p + \varepsilon)$. Since $(\vec{p}, G) \in \mathcal{E}_6$, we have that $\sum_{i \in R} \mathsf{dels}_i(G) \cdot Q_i(G) \geq \frac{(1-\varepsilon)^2}{1+\varepsilon}(\mu + c)(1 - p - 2\varepsilon) \cdot n$. Hence, this difference is at least

$$\begin{aligned}((\mu + c)(1 - p - 2\varepsilon) - (\mu + \varepsilon)(1 - p + \varepsilon))n &\geq (c(1 - p) - 3\varepsilon\mu - 2\varepsilon c - (1 - p)\varepsilon - \varepsilon^2)n \\ &\geq (c(1 - p) - 6\varepsilon - \varepsilon^2)n\end{aligned}$$

where the second inequality holds because, $c, (1 - p), \mu \leq 1$. By choosing $\varepsilon$ such that $6\varepsilon + \varepsilon^2 \leq \gamma$ ($\varepsilon = \min(\gamma/7, 1)$ will do), (6.19) follows.

*The Continuous General Delegation Model Satisfies* (6.3)

We now show that there exists a distribution $\mathcal{D}$ and $\alpha > 0$ such that $\sum_{i=1}^n p_i + \alpha n \leq n/2 \leq \sum_{i=1}^n \mathsf{weight}_i(G_n) \cdot p_i - \alpha n$ with probability $1 - o(1)$. This implies that the model $M^S_{p,\varphi}, n$ satisfies probabilistic positive gain by Lemma 1.3.3.

As in earlier arguments, let $\mathcal{D}_\eta = \mathcal{U}[0, 1 - 2\eta]$ for $\eta \in [0, 1/2)$. Note that

$$f(\eta) = \inf_{x \in [0,1]} \{\mathbb{E}_{D_\eta}[\varphi_x^+]\} \cdot (1 - p) - 3\eta/2$$

is a continuous function of $\eta$.

Moreover, $f(0) > 0$. Hence, for sufficiently small $\eta > 0$, $f(\eta) > 0$.

Consider $\mathcal{D}_\eta$ for some $\eta > 0$ such that $f(\eta) > 0$. Let $\alpha = \min(\eta/2, f(\eta)/2)$. Since $\mu_{\mathcal{D}_\eta} = 1/2 - \eta$, by Hoeffding's inequality, $\sum_{i=1}^n p_i \leq (1/2 - \eta/2)n \leq n/2 - \alpha n$ with high probability.

Next, note that we can choose $c = \inf_{x \in [0,1]} \{\mathbb{E}_{D_\eta}[\varphi_x^+]\}$ in order to satisfy (6.18). Hence, by choosing $\gamma = f(\eta)/2$, it follows from (6.19) that

$$\sum_{i=1}^n \mathsf{weight}_i(G) \cdot p_i - \sum_{i=1}^n p_i \geq (c(1 - p) - f(\eta)/2)n = (3\eta/2 + f(\eta)/2)n \geq (3\eta/2 + \alpha)n$$

with high probability. Further, by Hoeffding's inequality, $\sum_{i=1}^n p_i \geq (1/2 - 3\eta/2)n$

with high probability, so by the union bound applied to these inequalities,

$$\sum_{i=1}^{n} \mathsf{weight}_i(G) \cdot p_i \geq n/2 + \alpha n$$

with high probability, as needed. □

## 6.6 Liquid Democracy in Experiments

In six experiments, we statistically estimate the functions $q$ and $\varphi$ to assess the real-world implications of our theoretical findings. These experiments rely on a novel design measuring the vote n anon-strategic, non-incentivized liquid democracy stting, while simultaneously estimating voters' competence. Our empirical result are consistent exhibits a regime in which liquid democracy enhances collective intelligence, leveraging interpersonal knowledge embedded in social networks and identifying diverse sets of experts. Data and code are available at http://tinyurl.com/osf-liqdem.[10]

### 6.6.1 Experimental Design

#### Experiments and Material

We conducted $E = 6$ experiments after an initial pre-test[11] between March 21st and November 27th, 2022.[12] In each experiment $e$, a group of participants[13] performed $|\mathcal{T}_e|$ tasks[14] Each task consisted of 8 questions on the corresponding topic that

[10] Full link: https://osf.io/skxwg/?view_only=3671d431bcfd4a9cb94ded5aa86a0a95

[11] A description of the setup and results from the pre-study can be found in Appendix L of the full version and initial results can be found in [172].

[12] Our protocol E-3948 was approved and exempted by the university Committee on the Use of Humans as Experimental Subjects.

[13] Note that liquid democracy depends on the potential for beneficial delegation. It is therefore necessary to work with participants that have at least a passing familiarity with each other. Experiments were conducted in places such as classrooms and company workshops, where pre-existing group structures guaranteed such conditions. While significant preparation was needed to ensure correct experimental set-ups for these environments, this design did have the benefit of producing high-quality data with few missing entries and minimal drop-out.

[14] $|\mathcal{T}_e| = 4$, except for experiment $e = 6$ in which $|\mathcal{T}_e| = 12$: the final experiment was conducted over a longer period of time, allowing more tasks to be completed.

were primarily taken and adapted from the work of Simoiu et al. [186].[15] A total of $N = 168$ individuals participated. They hailed from over 30 countries; 33% were female, 1% were non-binary, 64% were male, and 2% preferred to self-describe. Each experiment $e$ had a number of participants $N_e$ ranging between 14 to 50. A description of the settings and group sizes are presented in Appendix C, and the survey material can be found in Appendix D of the full version.

### Survey Flow

Participants began by providing informed consent and inputting their name. Next, they completed the following steps.

*First experimental stage:* Participants were presented with a task and could either answer a series of questions related to that theme or delegate the task to a peer. For instance, a task read: "You will be shown images of architectural landmarks from around the world, and asked to select the country where the landmark is located," followed by "Do you want to vote yourself or delegate your vote to a trusted peer?" If they chose to vote themselves, they were taken to the 8 questions contained in the task. If they chose to delegate, they were asked to select the name of their delegate and then immediately directed to the next task. Importantly, when deciding whether or not to delegate, participants did not see the questions.

*Second experimental stage:* Participants were then asked to answer "additional questions." These were all the questions corresponding to tasks they had chosen to delegate in the first stage. We collected this data at the end of the experiment so as not to prime the participants on the exercise.[16]

Finally, optional background questions were asked on the last page. Note that the order in which tasks, questions within each task, and the "True/False" options ap-

[15] To be consistent with the theoretical setup under study, we converted all categorical questions into binary questions. For example, for a question from Simoiu et al. [186] of the form "Where is this famous landmark from?" with four options (Italy, Tibet, Greece, or Brazil), we selected a possible answer (e.g., Brazil) to reformulate the question as: "Is this famous landmark from Brazil?" In more detail, we first randomly selected which questions would be correct (to avoid the sense that most questions are incorrect) and then, for the incorrect ones, drew a wrong option at random. We found multiple inconsistencies in the Simoiu et al. [186] data that we corrected, and the prediction questions pertained to events that had passed, so these were replaced with new ones.

[16] We validated this approach with a robustness check on the time spent by participants as a function of how often they delegated (see Appendix K.1 of the full version).

peared were all randomized. The entire flow is summarized in Appendix D of the full version.

### Data Collected

Let $[N]$ be the set of $N$ participants and let $[E]$, the set of $E$ experiments. Each experiment $e \in [E]$ has $N_e$ participants so that $N = \sum_{e \in [E]} N_e$. $[N_e]$ denotes the subset of voters in experiment $e$ and $\mathcal{T}$ is the set of tasks surveyed ($|\mathcal{T}| = 15$). For each task $t \in \mathcal{T}$ there is a set $R_t$ of 8 corresponding questions. We let $R = \bigcup_t R_t$ be the set of all questions. For each participant $i$, $e(i) \in [E]$ is the experiment they participated in; for each question $r$, $t(r) \in \mathcal{T}$ is its corresponding task.

In the experiments, we collect (i) the direct vote to each question $i$ answered $v_{i,r} \in \{0, 1\}$ where 1 means correct and 0 means incorrect and (ii) the binary signal $\delta_{i,t}$ equal to 1 if $i$ delegated on task $t$ and 0 otherwise (note that $\delta_{i,t}$ is constant at the task level), along with which voter they delegated to. From this collected data, we can compute $w_{i,t}$, the weight of voter $i$ on task $t$. This is $i'$s total weight after adding up all transitive delegations; it is set to 0 when $i$ delegates. Figure 6.1 provides an example of a collected delegation graph.

In rare cases, a delegation could not be included for a couple of possible reasons. First, if a participant delegated to somebody who did not complete the survey. In this case, we would simply ignore the delegation (assuming they directly voted). Second, in an instance of a cycle (e.g., participant $i$ delegated to participant $j$ who delegated to participant $i$). These were also ignored (i.e., assumed that no voter on the cycle delegated). In many real-world implementations, such participants would be notified of the cycle and asked to choose a new delegate or vote directly.

## 6.6.2 Delegation and Competence Statistics

Over the 1096 (participant/task) pairs, we observed a total of 505 delegations meaning participants delegated 47% of the time ($std = 0.49$) The rate varied across experiments from 32% ($std = 0.49$) in experiment 2 to 54% ($std = 0.50$) in experiment 5 and across tasks from 22% ($std = 0.50$) in task $T_8$ to 80% ($std = 0.40$) in task $T_{15}$. Among those who voted directly, 15% received only one delegation besides their own (hence had weight 2 in the decision), 6% received two delegations, and just about 5% received five or more delegations. However, in one experiment, over half the votes

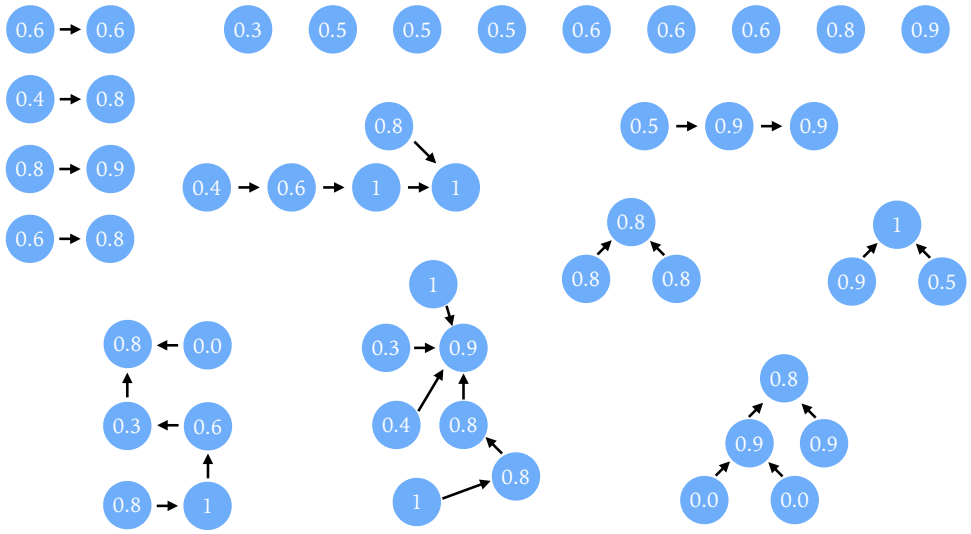



**Figure 6.1:** Delegation graphs for task $T_7$ ("You will be given upcoming European men soccer games and asked to predict the games' outcome.") from Experiment 6. Each node is a voter and the node's number represents the rounded expertise $\eta_{i,t}$ of a given voter $i$ for task $t$, computed using Item Response Theory, (see Appendix E of the full version).

were delegated to a single participant. Additionally, throughout all the experiments, we observed only four delegation cycles, and all were only of size two (where $a$ delegates to $b$, and $b$ delegates back to $a$). These occurred in Experiment 4 with $N_4 = 27$, and in Experiment 6 with $N_6 = 50$. Examples of additional delegation graphs can be found in Appendix F of the full version.

### Estimating Competence

In order to evaluate how delegation behavior relates to competence, we need to estimate participants' competence. We denote by $\eta_{i,t}$ the estimated competence of participant $i$ in task $t$. Naively, participants' competence per task could be estimated by averaging the number of correct answers given on all 8 questions of that task, $\eta_{i,t}^{\text{naive}} = \frac{\sum_{r \in R_t} v_{i,r}}{|R_t|}$. However, such a computation does not account for the questions' heterogeneity. We thus estimate $\eta_{i,t}$ using the Item Response Theory framework (IRT) [133], which provides a widely used parametric model to estimate competence $\eta_{i,t}$ and question difficulty from repeated measurements. We explain the parametric

estimation in Appendix E of the full version.[17]

#### Gender-based statistics

While we might worry that delegation patterns vary across gender due to significant differences in confidence [e.g., 68, 177], we actually find no significant differences in these experiments, neither in measured competence in tasks nor in propensity to delegate. ANOVA tests for the propensity to delegate (resp. competence) across gender shows no significant differences with $p = 0.464$ (resp. $p = 0.112$). Tukey tests for pairwise mean comparison further validate the absence of significant differences across the different genders (see Appendix G of the full version).

### 6.6.3 Estimating the Probability of Delegating as a Function of Competence

We now turn to estimating $q$ and $\varphi$ as a function of voters' competence. Recall that $q(\eta)$ represents the probability that somebody of competence $\eta$ chooses to delegate. We have observations $\delta_{i,t}$ encoding participant $i$'s delegation choice for task $t$, and an estimate $\eta_{i,t}$ of $i$'s competence on task $t$. We use these to estimate the relationship between competence $\eta_{i,t}$ and the probability of delegating $q(\eta_{i,t})$.

#### Methods

To estimate $q$, we fit a logistic model, regressing $\delta_{i,t}$ against $\eta_{i,t}$. The following equation shows the relationship we wish to fit, where $\alpha_0$ is the intercept and $\beta^q$ is the effect size we measure:

$$\log\left(\frac{\Pr[\delta_{i,t}=1]}{1-\Pr[\delta_{i,t}=1]}\right) = \alpha_0 + \beta^q \eta_{i,t} + \varepsilon_i. \tag{6.22}$$

To account for potential correlation in the error term within participants' answers, when estimating the parameters in Equation (6.22), we cluster standard errors at the

[17] While $\eta^{\text{naive}}_{i,t}$ takes on one of nine values (multiples of 1/8), $\eta_{i,t}$ (computed using IRT) is a continuous variable that can take on arbitrary values in $\mathbb{R}$. We normalize so that $\eta_{i,t} \in [0,1]$, and assume this to be the competence, the probability of being correct. Note that these different methods yield a correlation between $\eta^{\text{naive}}_{i,t}$ and $\eta_{i,t}$ of more than 94%.

participant level. We also test for the data normality; results for these test can be found in Appendix H of the full version.

We repeat the procedure above on data sets filtered by task, this time having a distinct $\beta_t$ for each task $t$ to measure the task-specific estimates. Additionally, we run these with fixed-effects for individuals and tasks to more directly measure the impact of competence (rather than just looking at population trends). Appendix I of the full version.

#### Results

We find $\beta^q = -2.24$, with standard error $s.e. = 0.42$, statistics $z = -7.12$ and p-value $p = 10^{-7}$. In turn, we estimate that $q(\eta_{i,t}) = \widehat{\Pr}[\delta_{i,t} = 1] = \frac{1}{1+\exp^{-(-1.39-2.24\times\eta_{i,t})}}$, suggesting that the probability of delegating decreases with competence. We can also test for monotonic dependence through a model-free method using a Pearson correlation test and its associated p-value. We find a correlation coefficient of $-0.17$ and $p < 5 \times 10^{-8}$.

### 6.6.4 Estimating Weight Function Used to Delegate

Recall that in the theoretical model, a voter with competence $\eta_1$ delegates to another with competence $\eta_2$ with probability proportional to $\varphi(\eta_1, \eta_2)$.

#### Methods

We first bucket the observed competence levels into $B$ clusters $c_1, \ldots, c_B$. We assume that $\varphi$ is constant across inputs in the same bucket, and fit it based on bucket "centers," $\eta_1, \ldots \eta_B$, which are simply taken to be the mean values of the competences in each bucket, i.e., $\eta_\ell = \frac{\sum_{i,t:\eta_{i,t}\in c_\ell} \eta_{i,t}}{|\{(i,t)|\eta_{i,t}\in c_\ell\}|}$. This means we can estimate $\varphi(x, y)$ using the number of delegations from any competence $x'$ to competence $y'$ where $x'$ and $y'$ fall in the same bucket as $x$ and $y$, respectively. Finally, we determine the Kendall tau rank correlation coefficient between $\varphi(x, y)$ and $y$ with its associated p-value to test for monotonic relation between $\varphi$ and its second coordinate.

Bucketing strategies. We discritize the segment $[0, 1]$ into $B$ buckets. We do so using several methods (to ensure the robustness of our approach); we describe here

the $k$-means clustering procedure and discuss the rest in Appendix J.1 of the full version.

To bucket using $k$-means, we optimize for $B$ clusters, $c_1, \ldots, c_B$, that minimize $\sum_{k=1}^{B} \sum_{\eta_{i,t} \in c_k} \left( \eta_{i,t} - \frac{\sum_{\eta_{i,t} \in c_k} \eta_{i,t}}{|c_k|} \right)^2$. In words, we compute a partition of the $[0, 1]$ segment such that the total squared distance from elements to their cluster centers is minimized. We use the standard $k$-means clustering algorithm to find the clusters [102].

Estimation of $\varphi$ for a given delegation graph. Next, we wish to fit a function $\varphi$. For given experiment $e$ and task $t$, we estimate $\varphi_{e,t}(\eta_\ell, \eta_k)$ for each $\ell, k \in [B]$, so for conciseness, we write $\varphi^{\ell}_{e,t}(\eta_k) := \varphi_{e,t}(\eta_\ell, \eta_k)$.[18]

Observe that the experiment can then be viewed as a multinomial trial: from the perspective of an $\ell-$participant, there are $B$ choices to pick from, where the probability of picking a $k-$participant is proportional to all the $\varphi^{\ell}_{e,t}(\eta_k)$ for $k \in [B]$ and the number of participants of each competence level. We observe instances of these choices, the $z^{\ell}_k$ that are the observed number of times that someone of type $\ell$ delegated to someone of type $k$. It then suffices to find the maximum likelihood estimators for $\varphi^{\ell}_{e,t}(\eta_k)$ as a function of $z^{\ell}_k$, $n_k$ and $n_\ell$. We provide more technical details in Appendix J.2 of the full version.

Testing for monotonic dependence of $\varphi$ in its second coordinate. Finally, we test for potential monotonic dependence of $\varphi^{\ell}_{e,t}(\eta_k)$ as a function of $\eta_k$ visualizing the $\varphi^{\ell}_{e,t}(\eta_k)$ in Figure 6.2 and computing the Kendall tau rank correlation coefficient between $\varphi^{\ell}_{e,t}(\eta_k)$ and $\eta_k$ for a fixed $\ell$. and its associated p-value. The Kendall tau rank correlation coefficient evaluates the similarity between two vectors of rank – its form and significance is detailed in Appendix J.3 of the full version.

[18] Note that because the number of participants in each bucket changes for different experiments/tasks, it is difficult to fit a single function. Instead, we first found the most likely $\varphi$ to have generated each experiment/task and then combined these to find an overall best fit.

### Results

We show here the results for using k-means bucketing with $B = 4$.[19] Additional results for other strategies and other numbers of buckets can be found in Appendix J.5 of the full version. Descriptions of these four buckets can be found in Table 6.1.

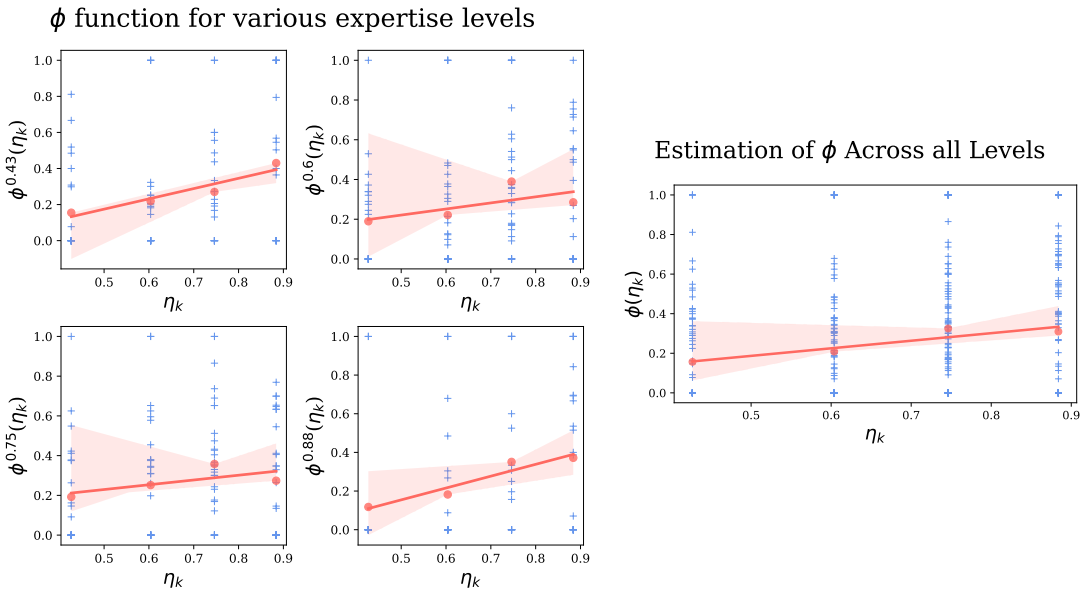
**Table 6.1:** Bucket Descriptions

| Bucket | Interval | Mean competence | Proportion of participants |
|---|---|---|---|
| $c_1$ | $[0.00, 0.514]$ | 0.43 | 16% |
| $c_2$ | $[0.515, 0.674]$ | 0.60 | 32% |
| $c_3$ | $[0.677, 0.814]$ | 0.75 | 35% |
| $c_4$ | $[0.818, 1.00]$ | 0.88 | 17% |

In Figure 6.2, the blue crosses in each column show the $\varphi^{\ell}_{e,t}(\eta_k)$ for all $(e, t)$ (for a particular $\eta_\ell$ in the four plots on the left and for all combined in the right plot) as a function $\eta_k$. The pink points represent the average across all experiments and tasks for a given $\eta_k$, and the regression line corresponds to an ordinary least square regression on the mean values. We show this both pooled only for those in the same bucket, as well as all grouped together.

To test the significance of the trends observed in Figure 6.2, we test whether the Kendall tau rank correlation coefficient between $\varphi^{\ell}_{e,t}(\eta_k)$ and $\eta_k$ signals significant associations, both at the overall level and when fixing $\ell$, or $\eta_\ell$, the first coordinate in $\varphi_{e,t}(\eta_\ell, \eta_k)$. Table 6.2 shows the resulting correlation coefficients and significance tests. Both the trends observed in Figure 6.2 and in Table 6.2 confirm that there is a statistically significant increase in $\varphi_{e,t}(\eta_\ell, \eta_k)$ as a function $\eta_\ell$ across all expertise levels,confirming that voters behave according to the general continuous delegation model. We further note that these significant trends are valid at the granularity of three of the four buckets (the third bucket $c_3$ exhibits non-statistically significant positive Kendall tau rank correlation.) We also run the same tests partitioned into tasks. The results can be found in Appendix J.4 of the full version. We last check that these results are not sensitive to the bucketing strategy in Appendix J.5 of the full version.

[19] This was the optimal number found using the *Kneedle algorithm* [178]




**(a)** Estimated values separated by input bucket. **(b)** Estimated values pooled together.

**Figure 6.2:** Pooled estimates of $\varphi^{\ell}_{e,t}$, both for each bucket individually, and grouped together. The blue crosses show the values computed for $\varphi^{\ell}_{e,t}(\eta_k)$. The pink dots show the average across all values for that $\eta_k$, and the pink lines correspond to a linear regression over the mean values. We observe increasing trends across the board, with slope (coefficient of determination) being $0.53(0.90), 0.28(0.46), 0.29(0.47)$ and $0.60(0.92)$, respectively, for individual buckets, and $0.38(0.85)$ for the pooled test. The shaded area represents the $95\%$ confidence interval.

### 6.6.5 Experimental Conclusions

We found that voters' likelihood to delegate decreases with their competence (as suggested by the confidence-based model). In addition, voters are more likely to delegate to someone of increasing competence (as suggested by the general continuous model). Note that, in fact, the general continuous model can be generalized to allow monotonically decreasing $q$ as well (indeed, it suffices to consider the expectation of $q(p_i)$ taken over the distribution of competence as the constant probability of voting). Our empirical results are hence consistent with a general continuous delegation model. Unsurprisingly, the upward-delegation model we described based on [40, 115] that leads to a catastrophic concentration of power is not consistent with experimental data: voters do not delegate only to higher-competence agents, and we do not observe constant delegation competence.

**Table 6.2:** Summary of correlation effects

| | Overall | For fixed $\ell$ | | | |
|---|---|---|---|---|---|
| | | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
| Correlation | $0.17^{****}$ | $0.29^{**}$ | $0.12^{*}$ | 0.11 | $0.28^{***}$ |
| P-value | $2 \times 10^{-5}$ | $2 \times 10^{-2}$ | $9 \times 10^{-2}$ | $1 \times 10^{-1}$ | $3 \times 10^{-3}$ |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01; $^{****}$p<0.0001 | | |

### 6.6.6 Additional Results

In Appendix K of the full version, we run additional tests and check other properties of the collected data. These include comparing the frequency of correctness between liquid and direct democracy, analyzing the increase in competence, and analyzing the concentration of power, using both the maximum weight, and the power of small coalitions. Note that the latter results are particularly relevant to the concentration of power in liquid democracy, a topic that was extensively discussed in previous work. Recall that, while previous work exhibited scenarios in which concentration of power occurs, our work is concerned with measuring whether such an extreme concentration of power is likely. In all but one of the 32 instances, we found no evidence of concentration of power. Anecdotally, we further observe that a number roughly the square root of the number of voters controls half of the votes.

## 6.7 Discussion

Our paper relies on a set of assumptions and modeling choices that are worth discussing.

First, a prominent feature of our model is that there is no underlying social network, that is, there is no restriction on whom a voter may delegate to. As we explained in Section 6.1, we believe this is realistic in a variety of scenarios. But we can, in fact, extend our results to a model where a directed social network is first sampled, and then a $(q, \varphi)$-model is followed. The social network must be sampled such that the neighbors of each voter are chosen uniformly at random, although the number of such neighbors could follow any small-tailed distribution. Intuitively, delegation proportional to weighting the neighbors of $i$ (rather than the entire population) is equiv-

alent to a possibly different weighting over the entire population. [20]An open research direction is to consider graph topologies not covered by these dynamics.

Second, building on Kahng et al. [115], we assume that there exists a true best alternative. Needless to say, this assumption is necessary if we wish to "defend" liquid democracy against their conclusions. It is also an extremely well-studied assumption that dates back to the 18th century [210]. The existence of a ground truth is easily justified in the contexts of prediction markets or corporate governance, where alternative policies can be measured in terms of concrete metrics like "estimated revenue in five years," and these metrics can be communicated to voters. That said, some decisions inherently rely on other subjective criteria that we do not capture.

Third, again like previous papers [22, 40, 115], we assume that voters vote independently. Admittedly, this is not a realistic assumption; relaxing it, as it was relaxed for the classic Condorcet Jury Theorem [91, 153], is a natural direction for future work.

Fourth, our models do not take strategic behaviors into account. In the same vein, our experiments do not involve explicit incentives and, while we do not have reasons to believe that significant strategic voting occurs, studying it these issues is beyond the scope of this work. It would be of interest to extend our results to a more game-theoretic setting, and relate them to work focusing on game-theoretic issues in liquid democracy [31, 62, 212]. Along those lines, experiments with monetary incentives would be interesting.

Fifth, our framework for stochastic delegations open interesting directions for more research. For instance, one could characterize all the delegation models satisfying positive gain and do no harm. One may also consider more general local-delegation models that would depend on the competences of all voters.

More generally, our work aims to provide a better understanding of a prominent shortcoming of liquid democracy: concentration of power. But there are others. For example, any voter can see the complete delegation graph under current liquid democracy systems — a feature that helps voters make informed delegation decisions (because one's vote can be transitively delegated). This may lead to voter coercion, however, and the tradeoff between transparency and security is poorly understood.

Finally, to summarize, we have introduced a general framework to investigate stochas-

[20]This extension does not carry over to undirected networks, since if voters have a small number of neighbors, we would expect many 2-cycles to form after delegation, which, under the worst-case cycle approach, may not be canceled out by the overall increase in competence.

tic dynamics in liquid democracy and proved new conditions for the convergence of weighted majorities; we then identified regimes in which liquid democracy leads to correct outcomes with high probability. In that sense, our work is to liquid democracy what the Condorcet Jury Theorem is to direct democracy. There are many reasons to be excited about the potential of liquid democracy [32]. We believe that our results provide another such reason and hope that our techniques will be useful in continuing to build the theoretical and empirical understanding of this compelling paradigm.

# 7

# Discussion

This thesis set out to bridge the rich heritage of social choice theory with the evolving challenges of modern collective decision-making. It extends and adapts classical principles of fairness, representation, and efficiency to contemporary contexts shaped by emerging technologies, such as AI alignment, large-scale online participation, and novel democratic processes.

While meaningful progress has been made, the journey from theoretical insight to practical deployment remains ongoing. As with all modeling efforts, our frameworks rest on assumptions — some arguably too weak, others perhaps too strong. Rather than viewing these as flaws, we see them as opportunities for deeper exploration. Although each chapter addresses immediate extensions, this concluding reflection takes a broader view to outline several promising directions for future research.

The principled study of AI alignment, as explored in Part I, is still in its early stages, even as AI systems grow increasingly pervasive and impactful in everyday life. Our axiomatic and definition-driven approach to specifying the desirable properties of aligned systems offers a promising foundation. However, significant challenges remain. Alignment, as currently pursued, conflates multiple objectives. On one hand, it is used to ensure safety and factual accuracy. On the other, RLHF is also a cornerstone in producing models that are more helpful and user-aligned. Balancing these dual aims, objective correctness and subjective alignment, remains an open and important prob-

lem. Moreover, while formal theory provides valuable guidance in simplified settings, the real world, with its biased, noisy datasets and intricate model architectures, poses a fundamentally harder problem.

In Part II, we considered the problems of aggregation under limited information. Perhaps our approach, specifically in Chapter 3, was too conservative in avoiding structural assumptions about individual preferences. It is plausible that preferences are much more structured, say, lying in relatively low-dimensional latent spaces. If such structure exists, it could enable more efficient algorithms with stronger guarantees. At the same time, some assumptions may have been overly idealized. For instance, treating participants as being randomly sampled fails to capture how arrival time and self-selection might correlate with underlying preferences. These dynamics pose critical challenges to fairness and representativeness, and designing systems that remain robust under such conditions is an important direction for future work.

Further, the goals of summarization and aggregation are highly context-dependent. When a single decision must be made, broad consensus may be the priority; however, when multiple outcomes can coexist, ensuring a diversity of viewpoints appear in the summary becomes more important. Minority priorities may be easily accommodated with minimal impact on others, yet such considerations may be overlooked by approaches solely focused on broad consensus. This highlights the risk of one-size-fits-all systems and the need to consider downstream applications.

Finally, in Part III, we turned to the comparison of democratic systems, with a focus on liquid democracy. A central challenge here lies in identifying realistic assumptions about voter behavior. Social choice theory is replete with impossibility results in the worst-case. This part showed that under plausible, semi-random delegation models, some of these barriers can be circumvented. Still, the question of what constitutes a reasonable model of preferences or behavior remains foundational, and largely unresolved. Many theoretical guarantees hinge on exactly these assumptions, making their careful justification all the more important.

Looking ahead, I hope the gaps identified — and partially bridged — in this thesis inspire future research into richer theoretical models and more practical, end-to-end systems. As automated decision-making increasingly shapes our world, social choice theory offers a critical lens for keeping human preferences at the core. With continued effort, we can build systems that are not only capable, but also provably inclusive, adaptable, and fair.

# References


[1] Nir Ailon and Mehryar Mohri. Preference-based learning to rank. *Machine Learning*, 80:189–211, 2010.

[2] AI@Meta. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, 2024. Accessed 11 May 2025.

[3] Noga Alon and Joel H. Spencer. *The Probabilistic Method*. John Wiley & Sons, 2016.

[4] G. Amanatidis, G. Birmpas, A. Filos-Ratsikas, and A. A. Voudouris. Peeking behind the ordinal curtain: Improving distortion via cardinal queries. *Artificial Intelligence*, 296:103488, 2021.

[5] I. Anagnostides, D. Fotakis, and P. Patsilinakos. Metric-distortion bounds under limited information. *Journal of Artificial Intelligence Research*, 74:1449–1483, 2022.

[6] E. Anshelevich, O. Bhardwaj, and J. Postl. Approximating optimal social choice under metric preferences. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 777–783, 2015.

[7] E. Anshelevich, O. Bhardwaj, E. Elkind, J. Postl, and P. Skowron. Approximating optimal social choice under metric preferences. *Artificial Intelligence*, 264: 27–51, 2018.

[8] E. Anshelevich, A. Filos-Ratsikas, N. Shah, and A. A. Voudouris. Distortion in social choice problems: The first 15 years and beyond. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4294–4301, 2021. Survey Track.

[9] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sour-

but, et al. Foundational challenges in assuring alignment and safety of large language models. arXiv:2404.09932, 2024.

[10] Ch Anwar ul Hassan, Muhammad Hammad, Jawaid Iqbal, Saddam Hussain, Syed Sajid Ullah, Hussain AlSalman, Mogeeb AA Mosleh, and Muhammad Arif. A liquid democracy enabled blockchain-based electronic voting system. *Scientific programming*, 2022(1):1383007, 2022.

[11] Pablo Aragón, Andreas Kaltenbrunner, Antonio Calleja-López, Andrés Pereira, Arnau Monterde, Xabier E Barandiaran, and Vicenç Gómez. Deliberative platform design: The case study of the online discussions in decidim Barcelona. In *Proceedings of the 9th International Conference on Social Informatics (SocInf)*, pages 277–287, 2017.

[12] K. Arrow. *Social Choice and Individual Values*. Wiley, 1951.

[13] Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management science*, 63 (3):691–706, 2017.

[14] H. Aziz, M. Brill, E. Elkind, R. Freeman, and T. Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 42(2):461–485, 2017.

[15] H. Aziz, E. Elkind, S. Huang, M. Lackner, L. Sánchez-Fernández, and P. Skowron. On the complexity of extended and proportional justified representation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 902–909, 2018.

[16] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Simon Mackenzie, Nicholas Mattei, and Toby Walsh. Computational Aspects of Multi-Winner Approval Voting. In *Proceedings of the 14th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 107–115, 2015.

[17] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al.

Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862, 2022.

[18] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073, 2022.

[19] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 38176–38189, 2022.

[20] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[21] Fabian Baumann, Daniel Halpern, Ariel D. Procaccia, Iyad Rahwan, Itai Shapira, and Manuel Wüthrich. Optimal engagement-diversity tradeoffs in social media. In *Proceedings of the 33rd International World Wide Web Conference (WWW)*, page 288–299, 2024.

[22] Ruben Becker, Gianlorenzo D'angelo, Esmaeil Delfaraz, and Hugo Gilbert. Unveiling the truth in liquid democracy with misinformed voters. In *Proceedings of the 7th International Conference on Algorithmic Decision Theory (ADT)*, pages 132–146, 2021.

[23] Gerdus Benade, Daniel Halpern, and Alexandros Psomas. Dynamic fair division with partial information. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3703–3715, 2022.

[24] Gerdus Benade, Daniel Halpern, Alexandros Psomas, and Paritosh Verma. On the existence of envy-free allocations beyond additive valuations. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC)*, page 1287, 2024.

[25] Gerdus Benadè, Daniel Halpern, and Alexandros Psomas. Dynamic fair division with partial information. *Operations Research*, 2025.

[26] Alon Benhaim, Brett H Falk, and Gerry Tsoukalas. Scaling blockchains: Can committee-based consensus help? *Management Science*, 69(11):6525–6539, 2023.

[27] M. Bentert and P. Skowron. Comparing election methods where each voter ranks only few candidates. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2218–2225, 2020.

[28] Claude Berge. *Topological Spaces*. Oliver and Boyd, 1963.

[29] Adam J Berinsky, Daniel Halpern, Joseph Y Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D Procaccia, and Manon Revel. Tracking truth with liquid democracy. *Management Science*, 2025.

[30] Erdem Bıyık, Nicolas Huynh, Mykel J Kochenderfer, and Dorsa Sadigh. Active preference-based Gaussian process regression for reward learning. arXiv:2005.02575, 2020.

[31] Daan Bloembergen, Davide Grossi, and Martin Lackner. On rational delegations in liquid democracy. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1796–1803, 2019.

[32] Christian Blum and Christina Isabel Zuber. Liquid democracy: Potentials, problems, and perspectives. *Journal of Political Philosophy*, 24(2):162–182, 2016.

[33] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[34] Allan Borodin, Daniel Halpern, Mohamad Latifian, and Nisarg Shah. Distortion in voting with top-t preferences. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 116–122, 2022.

[35] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[36] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[37] Markus Brill and Nimrod Talmon. Pairwise liquid democracy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 137–143, 2018.

[38] C. M. Burnett and V. Kogan. Ballot (and voter) "exhaustion" under instant runoff voting: An examination of four ranked-choice elections. *Electoral Studies*, 37:41–49, 2015.

[39] Joseph Campbell, Alessandra Casella, Lucas de Lara, Victoria R Mooers, and Dilip Ravindran. Liquid democracy. two experiments on delegation in voting. Technical report, National Bureau of Economic Research, 2022.

[40] Ioannis Caragiannis and Evi Micha. A contribution to the critique of liquid democracy. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 116–122, 2019.

[41] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv:2307.15217, 2023.

[42] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. Persona: A reproducible testbed for pluralistic alignment. arXiv:2407.17387, 2024.

[43] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. MaxMin-RLHF: Towards equitable alignment of large language models with diverse human preferences. arXiv:2402.08925, 2024.

[44] Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 101928–101968, 2024.

[45] Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[46] Haoxian Chen, Hanyang Zhao, Henry Lam, David Yao, and Wenpin Tang. Mallowspo: Fine-tune your LLM with preference dispersions. arXiv:2405.14953, 2024.

[47] M Keith Chen, Jonathan E Ingersoll Jr, and Edward H Kaplan. Modeling a presidential prediction market. *Management Science*, 54(8):1381–1394, 2008.

[48] Jae Hyeon Cho, Minkyung Park, and Byung-Jun Lee. Vpo: Leveraging the number of votes in preference optimization. arXiv:2410.22891, 2024.

[49] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 4302–4310, 2017.

[50] Zoé Christoff and Davide Grossi. Binary voting with delegable proxy: An analysis of liquid democracy. In *Proceedings of the 16th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 134–150, 2017.

[51] Fan Chung, Shirin Handjani, and Doug Jungreis. Generalizations of polya's urn problem. *Annals of combinatorics*, 7(2):141–153, 2003.

[52] F.H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley New York, 1983.

[53] Andrea Collevecchio, Codina Cotar, and Marco LiCalzi. On a preferential attachment and generalized pólya's urn model. *The Annals of Applied Probability*, 23(3):1219–1253, 2013.

[54] V. Conitzer and T. Sandholm. Communication complexity of common voting rules. In *Proceedings of the 6th ACM Conference on Economics and Computation (EC)*, pages 78–87, 2005.

[55] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey

Schoelkopf, et al. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 9346–9360, 2024.

[56] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2nd edition, 2001.

[57] Thomas M Cover. The number of linearly inducible orderings of points in $d$-space. *SIAM Journal on Applied Mathematics*, 15(2):434–439, 1967.

[58] S. Cunow, S. Desposato, A. Janusz, and C. Sells. Less is more: The paradox of choice in voting behavior. *Electoral Studies*, 69:102230, 2021.

[59] S. Cunow, S. Desposato, A. Janusz, and C. Sells. Too much of a good thing? longer ballots reduce voter participation. *Journal of Elections, Public Opinion and Parties*, pages 1–18, 2023.

[60] Jessica Dai and Eve Fleisig. Mapping social choice theory to RLHF. arXiv:2404.13038, 2024.

[61] P. Dey and A. Bhattacharyya. Sample complexity for winner prediction in elections. In *Proceedings of the 14th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1421–1430, 2015.

[62] Amrita Dhillon, Grammateia Kotsialou, Dilip Ravindran, and Dimitrios Xefteris. Information aggregation with delegation of votes. arXiv:2306.03960, 2023.

[63] Eleni Drinea, Mihaela Enachescu, and Michael D Mitzenmacher. Variations on random graph models for the web. Technical report, Harvard Computer Science Group, 2001.

[64] Richard Durrett. *Random graph dynamics*. Cambridge University Press, 2007.

[65] Soroush Ebadian, Daniel Halpern, and Evi Micha. Metric distortion with elicited pairwise comparisons. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2791–2798, 2024.

[66] H. Edelsbrunner. *Algorithms in Combinatorial Geometry*, volume 10 of *EATCS Monographs on Theoretical Computer Science*. Springer, 1987.

[67] Florian Eggenberger and George Pólya. Über die statistik verketteter vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik*, 3(4):279–289, 1923.

[68] Jessica Ellis, Bailey K Fosdick, and Chris Rasmussen. Women 1.5 times more likely to leave stem pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. *PloS one*, 11(7):e0157447, 2016.

[69] James M Enelow and Melvin J Hinich. *The spatial theory of voting: An introduction*. CUP Archive, 1984.

[70] FairVote. Research and data on RCV in practice, 2024. URL https://fairvote.org/resources/data-on-rcv/.

[71] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. arXiv:2406.15951, 2024.

[72] Luis Sánchez Fernández, Edith Elkind, Martin Lackner, Norberto Fernández García, Jesús Arias-Fisteus, Pablo Basanta-Val, and Piotr Skowron. Proportional justified representation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 670–676, 2017.

[73] Y. Filmus and J. Oren. Efficient voting via the top-$k$ elicitation scheme: A probabilistic approach. In *Proceedings of the 15th ACM Conference on Economics and Computation (EC)*, pages 295–312, 2014.

[74] Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC)*, page 985, 2024.

[75] P. C. Fishburn. Condorcet social choice functions. *SIAM Journal on Applied Mathematics*, 33(3):469–487, 1977.

[76] James Fishkin, Nikhil Garg, Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, Sukolsak Sakshuwong, Alice Siu, and Sravya Yandamuri. Deliberative democracy with the Online Deliberation platform. In *Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 1–2, 2019.

[77] Bailey Flanigan, Daniel Halpern, and Alexandros Psomas. Smoothed analysis of social choice revisited. In *Proceedings of the 19th Conference on Web and Internet Economics (WINE)*, pages 290–309, 2023.

[78] Cees M. Fortuin, Pieter W. Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[79] Hiroki Furuta, Kuang-Huei Lee, Shixiang Shane Gu, Yutaka Matsuo, Aleksandra Faust, Heiga Zen, and Izzeddin Gur. Geometric-averaged preference optimization for soft preference labels. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 57076–57114, 2024.

[80] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv:2209.07858, 2022.

[81] Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics*, 38(1):77–81, 1959.

[82] Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI alignment from human feedback. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 80439–80465, 2024.

[83] Luise Ge, Brendan Juba, and Yevgeniy Vorobeychik. Learning Linear Utility Functions From Pairwise Comparison Queries. arXiv:2405.02612, 2024.

[84] V. Gkatzelis, D. Halpern, and N. Shah. Resolving the optimal metric distortion conjecture. In *Proceedings of the 61st Symposium on Foundations of Computer Science (FOCS)*, pages 1427–1438, 2020.

[85] A. Goel, A. K. Krishnaswamy, and K. Munagala. Metric distortion of social choice rules: Lower bounds and fairness properties. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)*, pages 287–304, 2017.

[86] Paul Gölz, Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. The fluid mechanics of liquid democracy. In *Proceedings of the 14th Conference on Web and Internet Economics (WINE)*, pages 188–202, 2018.

[87] Henry W Gould. A note on the number of linearly inducible orderings of points in $d$-space. *SIAM Journal on Applied Mathematics*, 26(3):528–530, 1974.

[88] James Green-Armytage. Direct voting and proxy voting. *Constitutional Political Economy*, 26(2):190–220, 2015.

[89] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*. Springer, 1988.

[90] Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. arXiv:2406.00832, 2024.

[91] Olle Häggström, Gil Kalai, and Elchanan Mossel. A law of large numbers for weighted majority. *Advances in Applied Mathematics*, 37(1):112–123, 2006.

[92] D. Halpern, G. Kehne, A. D. Procaccia, J. Tucker-Foltz, and M. Wüthrich. Representation with incomplete votes. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5657–5664, 2023.

[93] Daniel Halpern, Gregory Kehne, and Jamie Tucker-Foltz. Can buyers reveal for a better deal? In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 314–320, 2022.

[94] Daniel Halpern, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, and Manon Revel. In defense of liquid democracy. In *Proceedings of the 24th ACM Conference on Economics and Computation (EC)*, page 852, 2023.

[95] Daniel Halpern, Rachel Li, and Ariel D Procaccia. Strategyproof voting under correlated beliefs. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 39744–39754, 2023.

[96] Daniel Halpern, Safwan Hossain, and Jamie Tucker-Foltz. Computing voting rules with elicited incomplete votes. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC)*, page 941–963, 2024.

[97] Daniel Halpern, Evi Micha, Ariel D. Procaccia, and Itai Shapira. Pairwise calibrated rewards for pluralistic alignment. arXiv:2506.06298, 2025.

[98] Daniel Halpern, Ariel D Procaccia, Ehud Shapiro, and Nimrod Talmon. Federated assemblies. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, pages 13897–13904, 2025.

[99] Daniel Halpern, Ariel D Procaccia, and Warut Suksompong. The proportional veto principle for approval ballots. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, 2025. Forthcoming.

[100] Daniel Halpern, Alexandros Psomas, Paritosh Verma, and Daniel Xie. Online envy minimization and multicolor discrepancy: Equivalences and separations. In *Proceedings of the 26th ACM Conference on Economics and Computation (EC)*, 2025. Forthcoming.

[101] Steve Hardt and Lia C. R. Lopes. Google votes: A liquid democracy experiment on a corporate social network. *Technical Disclosure Commons*, 2015.

[102] John A Hartigan, Manchek A Wong, et al. A k-means clustering algorithm. *Applied statistics*, 28(1):100–108, 1979.

[103] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[104] S. Hirano and J. M. Snyder Jr. *Primary elections in the United States*. Cambridge University Press, 2019.

[105] Ilgee Hong, Zichong Li, Alexander Bukharin, Yixiao Li, Haoming Jiang, Tianbao Yang, and Tuo Zhao. Adaptive preference scaling for reinforcement learning with human feedback. In *Proceedings of the 37rth Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 107249–107269, 2024.

[106] Jiekun Huang. Thy neighbor's vote: Peer effects in proxy voting. *Management Science*, 69(7):4169–4189, 2023.

[107] Eunjeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. Aligning language models to user opinions. In *Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[108] Luca Iandoli, Mark Klein, and Giuseppe Zollo. Enabling On-Line Deliberation and Collective Decision-Making through Large-Scale Argumentation: A New Approach to the Design of an Internet-Based Mass Collaboration Platform. *International Journal of Decision Support System Technology (IJDSST)*, 1(1):69–92, January 2009. URL https://www.igi-global.com/article/enabling-line-deliberation-collective-decision/1745.

[109] Aviram Imber, Jonas Israel, Markus Brill, and Benny Kimelfeld. Approval-Based Committee Voting under Incomplete Information. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5076–5083, 2022.

[110] Takayuki Ito, Shota Suzuki, Naoko Yamaguchi, Tomohiro Nishida, Kentaro Hiraishi, and Kai Yoshino. D-Agree: Crowd Discussion Support System Based on Automated Facilitation Agent. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 13614–13615, 2020.

[111] S. S. Iyengar and M. R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, 79(6): 995, 2000.

[112] Svante Janson. Rate of convergence for traditional pólya urns. *Journal of Applied Probability*, 57(4):1029–1044, 2020.

[113] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 10622–10643, 2023.

[114] Norman L Johnson and Samuel Kotz. *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. Wiley, 1978.

[115] Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. Liquid democracy: An algorithmic perspective. *Journal of Artificial Intelligence Research*, 70:1223–1252, 2021.

[116] Richard M Karp. On the computational complexity of combinatorial problems. *Networks*, 5(1):45–68, 1975.

[117] D. Kempe. Communication, distortion, and randomness in metric voting. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2087–2094, 2020.

[118] D. Kempe. An analysis framework for metric voting based on LP duality. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2079–2086, 2020.

[119] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2): 81–93, 1938.

[120] Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[121] Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. Randomness, not representation: The unreliability of evaluating cultural alignment in LLMs. arXiv:2503.08688, 2025.

[122] Kyuyoung Kim, Ah Jeong Seo, Hao Liu, Jinwoo Shin, and Kimin Lee. Margin matching preference optimization: Enhanced model alignment with granular feedback. arXiv:2410.03145, 2024.

[123] Hannah Rose Kirk, Andrew M Bean, Bertie Vidgen, Paul Röttger, and Scott A Hale. The past, present and better future of feedback learning in large language models for subjective human preferences and values. arXiv:2310.07629, 2023.

[124] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. arXiv:2303.05453, 2023.

[125] Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 105236–105344, 2024.

[126] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. arXiv:2310.06452, 2023.

[127] F. E. Kizilkaya and D. Kempe. Plurality veto: A simple voting rule achieving optimal metric distortion. In *Proceedings of the 31th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 349–355, 2022.

[128] F. E. Kizilkaya and D. Kempe. Generalized veto core and a practical voting rule with optimal metric distortion. In *Proceedings of the 24th ACM Conference on Economics and Computation (EC)*, page 913–936, 2023.

[129] Christoph Carl Kling, Jérôme Kunegis, Heinrich Hartmann, Markus Strohmaier, and Steffen Staab. Voting behaviour and power in online democracy: A study of LiquidFeedback in germany's pirate party. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, 2015.

[130] K. Konczak and J. Lang. Voting procedures with incomplete preferences. In *Proceedings of the 2nd Multidisciplinary Workshop on Advances in Preference Handling (M-PREF)*, 2005.

[131] A. Kupcsik, D. Hsu, and W. S. Lee. Learning dynamic robot-to-human object handover from human feedback. *Robotics Research: Volume 1*, pages 161–176, 2018.

[132] Thom Lake, Eunsol Choi, and Greg Durrett. From distributional to Overton pluralism: Investigating large language model alignment. arXiv:2406.17692, 2024.

[133] John Patrick Lalor and Pedro Rodriguez. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13, 2023.

[134] Nathan Lambert. *Reinforcement Learning from Human Feedback*. Online, 2024. URL https://rlhfbook.com.

[135] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The history and risks of reinforcement learning and human feedback. arXiv:2310.13595, 2023.

[136] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. In *Proceedings 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing,*, pages 1–35, 2019.

[137] Chao Li, Runhua Xu, and Li Duan. Liquid democracy in dpos blockchains. In *Proceedings of the 5th ACM International Symposium on Blockchain and Secure Critical Infrastructure*, pages 25–33, 2023.

[138] Christian List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001.

[139] T. Lu and C. Boutilier. Robust approximation and incremental elicitation in voting protocols. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 287–293, 2011.

[140] Hossam Mahmoud. *Pólya Urn Models*. CRC Press, 2009.

[141] D. Mandal, A. D. Procaccia, N. Shah, and D. P. Woodruff. Efficient and thrifty voting by any means necessary. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019. Forthcoming.

[142] D. Mandal, N. Shah, and D. P. Woodruff. Optimal communication-distortion tradeoff in voting. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC)*, pages 795–813, 2020.

[143] Andrei A. Markov. Sur quelques formules limites du calcul des probabilités. *Bulletin de l'Académie des Sciences*, 11(3):177–186, 1917.

[144] David C McGarvey. A theorem on the construction of voting paradoxes. *Econometrica: Journal of the Econometric Society*, pages 608–610, 1953.

[145] W. H. Mills and R. C. Mullin. Coverings and packings. In J. H. Dinitz and D. R. Stinson, editors, *Contemporary Design Theory: A Collection of Surveys*, chapter 9. Wiley, 1995.

[146] Lester James V Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and

Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. AI feedback. arXiv:2410.19133, 2024.

[147] Abhilash Mishra. AI alignment and social choice: Fundamental limitations and policy implications. arXiv:2310.16048, 2023.

[148] Abhilash Mishra. Ai alignment and social choice: Fundamental limitations and policy implications. arXiv:2310.16048, 2023.

[149] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[150] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Proceedings of the 32th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[151] K. Munagala and K. Wang. Improved metric distortion for deterministic social choice rules. In *Proceedings of the 20th ACM Conference on Economics and Computation (EC)*, pages 245–262, 2019.

[152] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332, 2021.

[153] Shmuel Nitzan and Jacob Paroush. Collective decision making and jury theorems. *The Oxford Handbook of Law and Economics*, 1, 2017.

[154] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1587–1594, 2018.

[155] Ritesh Noothigattu, Dominik Peters, and Ariel D Procaccia. Axioms for learning from pairwise comparisons. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 17745–17754, 2020.

[156] J. Oren, Y. Filmus, and C. Boutilier. Efficient vote elicitation under candidate uncertainty. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 309–316, 2013.

[157] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 27730–27744, 2022.

[158] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. RLHF from heterogeneous feedback via personalization and preference aggregation. arXiv:2405.00254, 2024.

[159] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. arXiv:2411.00860, 2024.

[160] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. arXiv:2202.03286, 2022.

[161] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. arXiv:2212.09251, 2022.

[162] Marcus Pivato. A statistical approach to epistemic democracy. *Episteme*, 9(2): 115–137, 2012.

[163] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. arXiv:2408.10075, 2024.

[164] George Pólya. Sur quelques points de la théorie des probabilités. *Annales de l'institut Henri Poincaré*, 1(2):117–161, 1930.

[165] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. arXiv:2110.05699, 2021.

[166] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence. arXiv:2211.13069, 2022.

[167] A. D. Procaccia. A note on the query complexity of the Condorcet winner problem. *Information Processing Letters*, pages 390–393, 2008.

[168] A. D. Procaccia and J. S. Rosenschein. The distortion of cardinal preferences in voting. In *Proceedings of the 10th International Workshop on Cooperative Information Agents (CIA)*, pages 317–331, 2006.

[169] Ariel D Procaccia, Benjamin Schiffer, and Shirley Zhang. Clone-robust AI alignment. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. Forthcoming.

[170] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[171] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[172] Manon Revel, Daniel Halpern, Adam Berinsky, and Ali Jadbabaie. Liquid democracy in practice: An empirical analysis of its epistemic performance. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2022.

[173] Manon Revel, Tao Lin, and Daniel Halpern. How many representatives do we need? the optimal size of a congress voting on binary issues. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9431–9438, 2022.

[174] R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[175] Matthew J Salganik and Karen E C Levy. Wiki surveys: open and quantifiable social data collection. *PloS one*, 10(5):e0123483, May 2015. URL http://dx.doi.org/10.1371/journal.pone.0123483.

[176] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 29971–30004, 2023.

[177] Heather Sarsons and Guo Xu. Confidence men? Evidence on confidence and gender among top economists. In *AEA Papers and Proceedings*, volume 111, pages 65–68. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2021.

[178] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *Proceedings of the 31st International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 166–171, 2011.

[179] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[180] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

[181] B. Schwartz. *The paradox of choice: Why more is less*. Harper Perennial, 2004.

[182] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

[183] Daichi Shibata, Ahmed Moustafa, Takayuki Ito, and Shota Suzuki. On Facilitating Large-Scale Online Discussions. In *PRICAI 2019: Trends in Artificial Intelligence*, pages 608–620. Springer International Publishing, 2019. URL http://dx.doi.org/10.1007/978-3-030-29911-8_47.

[184] Ali Shirali, Arash Nasr-Esfahany, Abdullah Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel Procaccia. Direct alignment with heterogeneous preferences. arXiv:2502.16320, 2025.

[185] Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of LLM generated content. arXiv:2504.12522, 2025.

[186] Camelia Simoiu, Chiraag Sumanth, Alok Mysore, and Sharad Goel. Studying the "wisdom of crowds" at scale. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 171–179, 2019.

[187] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42 (3/4):425–440, 1955.

[188] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[189] P. Skowron and E. Elkind. Social choice under metric preferences: scoring rules and STV. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 706–712, 2017.

[190] P. Skowron, M. Lackner, M. Brill, D. Peters, and E. Elkind. Proportional rankings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 409–415, 2017.

[191] Piotr Skowron. Proportionality Degree of Multiwinner Rules. In *Proceedings of the 22nd ACM Conference on Economics and Computation (EC)*, pages 820–840, 2021.

[192] Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse preference learning for capabilities and alignment. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025. Forthcoming.

[193] C. Small, M. Bjorkegren, T. Erkkilä, L. Shaw, and C. Megill. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Revista De Pensament I Anàlisi*, 26(2), 2021.

[194] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing

Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. arXiv:2402.05070, 2024.

[195] Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. arXiv:2503.15484, 2025.

[196] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Proceedings of the 33th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3008–3021, 2020.

[197] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. arXiv:2401.04056, 2024.

[198] Zoi Terzopoulou, Alexander Karpov, and Svetlana Obraztsova. Restricted domains of dichotomous preferences with possibly incomplete information. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5726–5733, 2021.

[199] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.

[200] Paolo Viappiani and Craig Boutilier. Optimal Bayesian recommendation sets and myopically optimal choice query sets. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2010.

[201] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of RLHF in large language models part ii: Reward modeling. arXiv:2401.06080, 2024.

[202] Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. Aligning language models with human preferences via a Bayesian approach. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 49113–49132, 2023.

[203] Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. arXiv:2406.08673, 2024.

[204] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. arXiv:2112.04359, 2021.

[205] D. B. West. *Introduction to Graph Theory*, volume 2. Prentice Hall Upper Saddle River, 2001.

[206] L. Xia and V. Conitzer. Determining possible and necessary winners given partial orders. *Journal of Artificial Intelligence Research*, 41:25–67, 2011.

[207] Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. On the algorithmic bias of aligning Large language models with RLHF: Preference collapse and matching regularization. arXiv:2405.16455, 2024.

[208] A. C. Yao. Probabilistic computations: Towards a unified measure of complexity. In *Proceedings of the 17th Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.

[209] H. P. Young. Social choice scoring functions. *SIAM Journal of Applied Mathematics*, 28(4):824–838, 1975.

[210] H. P. Young. Condorcet's theory of voting. *The American Political Science Review*, 82(4):1231–1244, 1988.

[211] Jiahao Yuan, Zixiang Di, Shangzixin Zhao, and Usman Naseem. Cultural palette: Pluralising culture alignment via multi-agent palette. arXiv:2412.11167, 2024.

[212] Yuzhe Zhang and Davide Grossi. Power in liquid democracy. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5822–5830, 2021.

[213] Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. Provable multi-party reinforcement learning with diverse human feedback. arXiv:2403.05006, 2024.

[214] Aizhong Zhou, Yongjie Yang, and Jiong Guo. Parameterized Complexity of Committee Elections with Dichotomous and Trichotomous Votes. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 503–510, 2019.

[215] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 43037–43067, 2023.

[216] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv:1909.08593, 2019.

[217] Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. arXiv:2409.20296, 2024.