Explainable Voting

Dominik Peters[1] **Ariel D. Procaccia**[2] **Alexandros Psomas**[3] **Zixin Zhou**[4]

Abstract

The design of voting rules is traditionally guided by desirable axioms. Recent work shows that, surprisingly, the axiomatic approach can also support the generation of explanations for voting outcomes. However, no bounds on the size of these explanations is given; for all we know, they may be unbearably tedious. We prove, however, that outcomes of the important Borda rule can be explained using $O(m^2)$ steps, where m is the number of alternatives. Our main technical result is a general lower bound that, in particular, implies that the foregoing bound is asymptotically tight. We discuss the significance of our results for AI and machine learning, including their potential to bolster an emerging paradigm of automated decision making called virtual democracy.

1. Introduction

As the reader is no doubt aware, voting has a storied history, both as a cornerstone of democracy and as an object of theoretical study. Taking the latter viewpoint, a *voting rule* is a function that maps individual preferences into a collective decision. In the prevalent model, voters' preferences are represented as *rankings* over a set of alternatives, so the input to a voting rule is a vector of rankings (one per voter), known as a *preference profile*. Since ties are possible, the output can technically be a subset of alternatives.

Starting with Arrow (1951) — even, in a sense, with Condorcet (1785) — researchers in *social choice theory* have evaluated voting rules through the axiomatic lens. This means defining formal properties that voting rules should satisfy, and determining whether they do. For example, *unanimity* — the *sine qua non* of social choice — requires that the voting rule selects alternative a when given a profile where all voters rank a first. A stiffer condition is the *reinforcement* axiom, which requires that if the voting rule, given two profiles $\mathbf{R}_1$ and $\mathbf{R}_2$, chooses two overlapping sets of tied alternatives A_1 and A_2 (possibly singletons), respectively, then given the profile that consists of all votes in both $\mathbf{R}_1$ and $\mathbf{R}_2$, it should select the intersection $A_1 \cap A_2$.

The axiomatic approach is helpful in guiding the design of voting rules, and understanding their advantages and disadvantages — it is useful for the designer. However, it is equally important that voters, or users of a voting-based system, understand *why* a certain outcome was selected by a voting rule — and it has been argued that social choice axioms cannot be directly relied upon for this purpose (Procaccia, 2019). Going back to our examples, unanimity provides an obvious explanation of outcomes in profiles where all voters are in perfect agreement — but such profiles are extremely rare. Reinforcement ties several profiles together, but — in and of itself — it cannot explain the outcome in any particular profile.

Nevertheless, Cailloux & Endriss (2016) suggest that it is possible that a *chain* of axioms would do the trick. In a nutshell, given a preference profile, *intraprofile axioms* that pertain to a single profile (like unanimity) can be applied to convince a human of the outcome on certain profiles; and interprofile axioms that connect several profiles (like reinforcement) can be used to relate the outcome on these profiles to the outcome on other profiles, in a way that ultimately yields the desired outcome on the profile at hand.

For example, under the Borda rule — which dates back to the 18th Century — each voter gives $m - k$ points to the alternative she ranks in the k^{th} place, where m is the number of alternatives; winning alternatives maximize the overall score. Now consider a trivial preference profile that only includes one voter with the ranking (a, b, c). Since a has unanimous support, the unanimity axiom designates it as the winner. Another profile has two voters associated with the rankings (a, b, c) and (c, b, a) (notice that the latter ranking is the reverse of the former); by an intraprofile axiom called *cancellation*, all alternatives must be tied under this profile. Finally, on the profile that contains all three voters, Borda would select alternative a. And, indeed, due to reinforcement, the winner on the combined profile must be the intersection of the two sets of winners, namely a. Therefore, we can explain the outcome of Borda on the combined profile by chaining three axioms together. Notice that this reasoning is essentially given in natural language,

[1]Carnegie Mellon University [2]Harvard University [3]Google Research [4]Peking University. Correspondence to: Zixin Zhou <zhouzixin1998@gmail.com>.

and it is clearly possible to automatically transform it into an explanation that a layperson can understand.

Cailloux & Endriss (2016) prove that for any profile, the outcome under Borda can be explained through a chain of axioms involving six different axioms. This is an inspiring result — but it is unclear whether it is practical, as the *length* of the explanation is unknown. An explanation with a thousand steps is not something a user will sit through.

Our research challenge, therefore, is to generate the shortest possible explanations of voting outcomes — and analyze the required length — while extending the approach of Cailloux & Endriss (2016) beyond the Borda rule.

1.1. Significance for AI and Machine Learning

That explainability is crucial to modern machine learning is almost a cliché at this point, but it is worth noting that researchers in the area are increasingly turning to economic theory for answers (Štrumbelj & Kononenko, 2010; Lundberg & Lee, 2017; Chen et al., 2019). Social choice theory can provide additional, valuable tools for explainability.

Our work is specifically motivated by the *virtual democracy* paradigm — an approach to automated decision making through machine learning and social choice. In a nutshell, the idea is to collect preference data from a group of voters, and use it to learn models of their preferences over a (possibly infinite) set of alternatives. At runtime, when a specific set of alternatives is presented, the system makes a decision by applying a voting rule to the *predicted* preferences of the voters over the current alternatives. This approach has led to proof-of-concept systems that automate moral decisions faced by autonomous vehicles (Noothigattu et al., 2018) and kidney exchanges (Freedman et al., 2018). Most notably, virtual democracy is the foundation of a pilot recommendation system used to allocate food donations to recipient organizations (Lee et al., 2019). The voting rule chosen to power this system is Borda, in part because it was shown to be robust to the type of preference prediction errors that arise in virtual democracy applications (Kahng et al., 2019).

Whether virtual democracy systems directly make decisions or merely support decisions by making recommendations, they must be trusted by stakeholders in order to be adopted. Lee et al. (2019) provide evidence that the participatory nature of voting-based systems inherently encourages trust. Still, the capability to provide meaningful explanations of decisions or recommendations obtained via voting — which is what we seek to develop — would amplify that trust. Taking a broader perspective, explainable voting is applicable to the myriad environments where voting is employed, from ensemble learning through online services like RoboVote.org all the way to — dare we say it? — political elections.

1.2. Our Results

In Section 3 we focus on the Borda rule, which — as we mentioned above — has special significance. Building on the work of Cailloux & Endriss (2016), we introduce seven natural axioms that characterize Borda. We then prove (Theorem 1) that, in this framework, Borda outcomes can always be explained in $O(m^2)$ steps, where m is the number of alternatives. This result gives an algorithm for automatically generating explanations that is practical in settings where the number of alternatives is small.

In Section 4 we lay the groundwork for, and prove, our main result — a general lower bound on the length of explanations. The key idea behind our approach is to *embed* voting rules into linear spaces, which allows us to apply linear algebra machinery. For example, Borda outcomes can be determined purely from the fraction of voters who prefer one alternative to another for every pair of alternatives, and therefore Borda can be embedded into $\mathbb{Q}^{\binom{m}{2}}$. The lower bound, Theorem 2, depends on the dimension of the linear space, as well as on another measure that we call the *sensitivity* of the embedding. A notable aspect of this result is that it holds not just in the worst case, but for almost every profile. As corollaries we get asymptotically tight lower bounds for Borda as well as two other prominent rules, plurality and approval. Our lower bounds guide the way towards voting rules whose outcomes can be easily explained even when there are many alternatives — a point that we discuss in Section 5.

2. Preliminaries

In this section we provide some social choice terminology, and describe the framework of Cailloux & Endriss (2016) for the explanation of voting outcomes.

2.1. Basic Terminology

Let $\mathcal{A}$ be a finite set of alternatives and denote $m = |\mathcal{A}|$. Let $\mathcal{P}_\emptyset(\mathcal{A})$ be the set of non-empty subsets of $\mathcal{A}$. Preferences of voters are given by (strict) rankings over $\mathcal{A}$; let $\mathcal{A}!$ be the set of strict rankings. A *preference profile* (or simply *profile*) is a function $\mathbf{R} : \mathcal{A}! \to \mathbb{N}$ that specifies how many voters report each possible ranking.[1] Let $\mathcal{R}$ be the set of all non-empty profiles, that is, all profiles except the one that maps all $m!$ rankings to zero. A *voting rule* $f : \mathcal{R} \to \mathcal{P}_\emptyset(\mathcal{A})$ maps each profile $\mathbf{R} \in \mathcal{R}$ to a non-empty subset of $\mathcal{A}$, the set of tied winners for $\mathbf{R}$.

For two profiles $\mathbf{R}_1$ and $\mathbf{R}_2$, we define $\mathbf{R}_1 \bigoplus \mathbf{R}_2$ as the sum of the two profiles (that is, the multiplicity of each

[1] This definition makes our setting *anonymous*, so that nothing depends on the identity of specific voters. Most of our results apply without this restriction, but we adopt it for ease of exposition.

ranking is the sum of its multiplicities in the two profiles). For $k \in \mathbb{Z}_+$, we define $k\mathbf{R}$ as the sum of k copies of $\mathbf{R}$.

We will pay special attention to the Borda rule. As mentioned in Section 1, under Borda each voter awards $m - k$ points to the alternative ranked in the k^{th} position; the winner set consists of all alternatives with maximum score. For example, if the votes are $a \succ b \succ c \succ d$, $d \succ b \succ c \succ a$, and (again) $d \succ b \succ c \succ a$, then the winner set would be $\{b, d\}$, as both alternatives have 6 points.

2.2. Explainability Framework

We will produce explanations as proofs in a language of propositional logic over propositional variables (or atomic formulae) $\{[\mathbf{R} \mapsto A] : \mathbf{R} \in \mathcal{R}, A \in \mathcal{P}_\emptyset(\mathcal{A})\}$. The language $\mathcal{L}$ is the set of all formulae that can be formed using these variables and logical connectives $\neg, \wedge, \vee, \rightarrow$.

A voting rule f induces a truth assignment v_f to the propositional variables which assigns value *true* to atom $[\mathbf{R} \mapsto A]$ if $f(\mathbf{R}) = A$ and value *false* otherwise. By standard semantics of propositional connectives, this truth assignment extends to all formulae of $\mathcal{L}$. A truth assignment v *satisfies* a set of formulae if v assigns *true* to all formulae in the set.

We can translate familiar axioms for voting rules in social choice theory into the language $\mathcal{L}$. An *$\mathcal{L}$-axiom* is a set of formulae, each of which we call an *axiom instance*. For example, the unanimity axiom can be written as $\{[\mathbf{R} \mapsto \{a\}] : a \in A, \mathbf{R} \in \mathcal{R}$ and every voter in $\mathbf{R}$ ranks a top$\}$.

A basic axiom is **FUNC** which requires that f assigns exactly one set A to each profile $\mathbf{R}$. Thus, **FUNC** consists of the formulae $\bigvee_{A \in \mathcal{P}_\emptyset(\mathcal{A})} [\mathbf{R} \mapsto A]$ and $\bigwedge_{A_1 \neq A_2} \neg [\mathbf{R} \mapsto A_1] \vee \neg [\mathbf{R} \mapsto A_2]$ for each $\mathbf{R} \in \mathcal{R}$. This axiom is a background assumption, and we will not explicitly include it in axiomatizations.

A voting rule f *satisfies* an $\mathcal{L}$-axiom $\mathbf{X}$ if v_f satisfies $\mathbf{X}$ (recall that an $\mathcal{L}$-axiom is a set of formulae). An *$\mathcal{L}$-axiomatization* is a set $\mathcal{S}$ of $\mathcal{L}$-axioms. With a slight abuse of terminology, a voting rule f satisfies an $\mathcal{L}$-axiomatization $\mathcal{S}$ if f satisfies every axiom in $\mathcal{S}$. Finally, $\mathcal{S}$ *characterizes* a voting rule f if and only if f is the only voting rule satisfying every axiom in $\mathcal{S}$.

Let $\mathcal{S}_{\text{Borda}}$ be an $\mathcal{L}$-axiomatization. An *explanation* of an outcome A for profile $\mathbf{R}$ in terms of $\mathcal{S}$ is a formal proof of the formula $[\mathbf{R} \mapsto A]$ in a suitable proof system for propositional logic, using axioms in $\mathcal{S}$ as assumptions. Any sound and complete proof system will work, but for concreteness let us define a *proof of formula φ assuming $\mathcal{S}$* to be a sequence $\varphi_1, \ldots, \varphi_r = \varphi$ of propositional formulae such that for each $i = 1, \ldots r$, we have that either (i) φ_i is an instance of an axiom in $\mathcal{S}$, or (ii) φ_i is an instance of an axiom in **FUNC**, or (iii) φ_i is a tautology (a formula that is satisfied by every variable assignment), or (iv) there exist $j, k < i$ such that $\varphi_k = \varphi_j \rightarrow \varphi_i$ (*modus ponens*). The *length* of the proof is the number r of formulae in the sequence. For ease of exposition, when writing down proofs in this system, we will often skip steps that use only propositional reasoning.

3. An Upper Bound for Borda

In this section we present our upper bound on the length of explanations required by the Borda rule using a particular, natural axiomatization. Specifically, we show that an explanation of length $O(m^2)$ suffices; as we will see in Section 4, our main result implies that this is optimal.

We start by defining families of profiles that are useful in producing short proofs. The first family consists of *elementary* profiles $\mathbf{R}^A_{\text{elem}}$, for each non-empty $A \subseteq \mathcal{A}$, which have two voters. Let $A = \{x_1, \ldots, x_k\}$ and $\mathcal{A} \setminus A = \{y_1, \ldots, y_{m-k}\}$. The first voter has preferences $x_1 \succ x_2 \succ \cdots \succ x_k \succ y_1 \succ \cdots \succ y_{m-k}$. The second voter has preferences $x_k \succ \cdots \succ x_1 \succ y_{m-k} \succ \cdots \succ y_1$. For example, the elementary profile $\mathbf{R}^{\{a,b\}}_{\text{elem}}$, when $\mathcal{A} = \{a, b, c, d\}$, has two votes: $a \succ b \succ c \succ d$ and $b \succ a \succ d \succ c$. Intuitively, in the profile $\mathbf{R}^A_{\text{elem}}$, the alternatives in A are similar to each other (since the preferences over A 'cancel'), and stronger than the other alternatives, so a sensible voting rule should select the alternatives in A. The second family consists of *cyclic* profiles $\mathbf{R}^T_{\text{cyc}}$, where T is an m-cycle over alternatives, which is composed of all rankings generated by T. For example, the cyclic profile $\mathbf{R}^{\langle a,b,c,a \rangle}_{\text{cyc}}$ contains the rankings (a, b, c), (b, c, a) and (c, a, b). Intuitively, by symmetry, a sensible voting rule should declare a tie between all alternatives in a cyclic profile.

We also require the notion of the *delta* vector $\delta^{\mathbf{R}}$ of a profile $\mathbf{R}$, which is a vector with $\binom{m}{2}$ coordinates, where $\delta^{\mathbf{R}}_{ab}$ is the number of voters who prefer alternative a to alternative b minus the number of voters who prefer b to a. The delta vector consists of the majority margins of the profile $\mathbf{R}$. For example, if $\delta^{\mathbf{R}}_{ab} > 0$, then a majority of voters prefers a to b.

An important observation is that the delta vector is a sufficient statistic for computing the outcome under Borda. Indeed, for an alternative $a \in \mathcal{A}$, one can check that the Borda score of a is equal to $\frac{1}{2} \sum_{b \in \mathcal{A}} \delta^{\mathbf{R}}_{ab} + n(m-1)/2$, where n is the total number of voters in $\mathbf{R}$. In any fixed $\mathbf{R}$, the second term is constant, and so we can determine Borda scores and winners by inspecting only the delta vector.

We are now ready to define the axioms we need for our upper bound. We use the axiomatization proposed by Cailloux & Endriss (2016). The first three axioms give single-step proofs for "base profiles." These are the *intraprofile* axioms.

1. **ELEM**: For an elementary profile $\mathbf{R}^A_{\text{elem}}$, the only reasonable set of winners is A. Formally, $\left[\mathbf{R}^A_{\text{elem}} \mapsto A\right]$.

2. **CYCL**: For a cyclic profile $\mathbf{R}_{\text{cyc}}^T$, the reasonable set of winners is all of $\mathcal{A}$. Formally, $\left[\mathbf{R}_{\text{cyc}}^T \mapsto \mathcal{A}\right]$.
3. **CANC**: If for all pairs of alternatives a, b, a is preferred to b exactly the same number of times b is preferred to a then the set of winners is $\mathcal{A}$. Formally, $\forall \mathbf{R}$ such that $\forall a, b \in \mathcal{A}, \delta_{ab}^{\mathbf{R}} = 0$, $[\mathbf{R} \mapsto \mathcal{A}]$.

The remaining axioms are *interprofile* axioms, linking outcomes between different profiles. The first axiom captures reinforcement. The others capture consequences of reinforcement; making them separate axioms gives us convenient shortcuts in the generated explanations.

4. **REINF**: For any two profiles $\mathbf{R}_1$ and $\mathbf{R}_2$, and any two subsets of alternatives A_1 and A_2 with $A_1 \cap A_2 \neq \emptyset$, it holds that $([\mathbf{R}_1 \mapsto A_1] \wedge [\mathbf{R}_2 \mapsto A_2]) \rightarrow [\mathbf{R}_1 \bigoplus \mathbf{R}_2 \mapsto A_1 \cap A_2]$.
5. **REINF-SUB**: Subtracting a profile with a full winner set does not change the outcome. Formally, for all $\mathbf{R}, \mathbf{R}'$, $([\mathbf{R} \bigoplus \mathbf{R}' \mapsto A] \wedge [\mathbf{R}' \mapsto \mathcal{A}]) \rightarrow [\mathbf{R} \mapsto A]$.
6. **SIMP**: A profile that is a repetition of some sub-profile should have the same set of winners as the sub-profile. Formally, $\forall k \in \mathbb{Z}_+$, $[k\mathbf{R} \mapsto A] \rightarrow [\mathbf{R} \mapsto A]$.
7. **MULT**: If a profile $\mathbf{R}$ has winner set A, then the profile that repeats $\mathbf{R}$ k times has the same winner set. Formally, $\forall k \in \mathbb{Z}_+$, $[\mathbf{R} \mapsto A] \rightarrow [k\mathbf{R} \mapsto A]$.

The last axiom is not used by Cailloux & Endriss (2016); we add it for convenience. Let us refer to the $\mathcal{L}$-axiomatization consisting of Axioms 1–7 listed above as $\mathcal{S}_{\text{Borda}}$. Cailloux & Endriss (2016) show that $\mathcal{S}_{\text{Borda}}$ characterizes the Borda rule (based on a result of Young (1974)) and that for any profile $\mathbf{R}$, the outcome of Borda can be explained using $\mathcal{S}_{\text{Borda}}$ (and no other outcome can be so explained). Technically, this means that given a profile $\mathbf{R}$, it is possible to give a proof that the atomic formula $\varphi = [\mathbf{R} \mapsto f(\mathbf{R})]$ is such that $v_f(\varphi) = \textit{true}$ for all voting rules f satisfying $\mathcal{S}_{\text{Borda}}$ (and Borda is the only such rule). Our first theorem strengthens this existence result by bounding the length of the required explanation. We only give a rough proof sketch here to outline the strategy, and leave the details to Appendix A.1.

Theorem 1. *For any profile $\mathbf{R}$ with m alternatives, the outcome of the Borda rule can be explained in $O(m^2)$ steps assuming the $\mathcal{L}$-axiomatization $\mathcal{S}_{\text{Borda}}$.*

Proof sketch. The linear space $\mathbb{Q}^{\binom{m}{2}}$ of delta vectors is spanned by the delta vectors induced by elementary and cyclic profiles. Given a profile $\mathbf{R}$, we can find another profile $\mathbf{R}'$ which is a linear combination of at most $O(m^2)$ different elementary and cyclic profiles, satisfying $k\delta^{\mathbf{R}} = \delta^{\mathbf{R}'}$ for some $k \in \mathbb{Z}_+$. By the latter equality, $\mathbf{R}$ and $\mathbf{R}'$ have the same set of Borda winners. Using **ELEM**, **CYCL**, and interprofile axioms, we can show that f must elect the Borda winners at $\mathbf{R}'$. Using **CANC** and interprofile axioms, we can show that since $k\delta^{\mathbf{R}} = \delta^{\mathbf{R}'}$, we must have $f(\mathbf{R}) = f(\mathbf{R}')$, which together gives an explanation of the Borda outcome at $\mathbf{R}$. The length of the explanation is determined by the length of the decomposition of $\mathbf{R}'$, which is in $O(m^2)$. □

4. A General Lower Bound

In this section we prove our main result: a general lower bound on the required explanation length, which applies to a broad class of axiomatizations. We begin by defining some important concepts, and then proceed to the main proof.

4.1. Mathematical Framework

The Borda rule depends only on the delta vector, and the explanations constructed in Theorem 1 exploit the linear algebra of delta vectors. Our result applies more generally to voting rules that can be embedded into a linear space, and axiomatizations based on the embedding. (All definitions in this section are novel, to the best of our knowledge.)

Definition 1. A voting rule $f : \mathcal{R} \to \mathcal{P}_\emptyset(\mathcal{A})$ can be *embedded* into a linear space V over $\mathbb{Q}$ via $h : \mathcal{R} \to V$ and $g : V \to \mathcal{P}_\emptyset(\mathcal{A})$ if the following properties are satisfied:

1. $h(\mathbf{R} \bigoplus \mathbf{R}') = h(\mathbf{R}) + h(\mathbf{R}'), \forall \mathbf{R}, \mathbf{R}' \in \mathcal{R}$.[2]
2. $f(\mathbf{R}) = g(h(\mathbf{R})), \forall \mathbf{R} \in \mathcal{R}$.
3. $\{h(\mathbf{R}) : \mathbf{R} \in \mathcal{R}\}$ spans V.

We say g *admits operation* $\circ : \mathcal{P}_\emptyset(\mathcal{A}) \times \mathcal{P}_\emptyset(\mathcal{A}) \to \mathcal{P}(\mathcal{A})$ if $g(v + v') = g(v) \circ g(v')$ $\forall v, v' \in V$ s.t. $g(v) \circ g(v') \neq \emptyset$.

For example, any (anonymous) voting rule f can be trivially embedded into a linear space of dimension $m!$: the vector $h(\mathbf{R})$ shows how often each preference ranking appears in $\mathbf{R}$, and g maps each possible such vector to a set of winners. The Borda rule can be embedded into $\mathbb{Q}^{\binom{m}{2}}$ by $h(\mathbf{R}) = \delta^{\mathbf{R}}$ and $g(\delta) = \arg\max_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} \delta_{ab}$. One could also embed the Borda rule into $\mathbb{Q}^m$, with h returning the vector of Borda scores, and g selecting the alternatives with highest score.

In all these examples, if the rule f satisfies reinforcement, then the embedding admits the operation $\cap$. Intuitively, the operation $\circ$ describes how to combine two voting outcomes. When $A_1 \circ A_2 = \emptyset$, we cannot combine A_1 and A_2.

We now describe an $\mathcal{L}$-axiomatization for an embedded voting rule. This axiomatization has several abstract components, which we will discuss further after the definition.

[2] All results in this section still hold when we replace the operation $\bigoplus$ with any binary operation $\mathcal{R} \times \mathcal{R} \to \mathcal{R}$.

Definition 2. Let f be a voting rule that can be embedded into a linear space V by h and g, and assume that g admits operation $\circ$, which is commutative. Let $S \subseteq \mathcal{R}$ be a set of *base profiles*, such that S can be written as a finite union of sets of profiles, $S = \bigcup_{k=1}^{N} S_i$, for some N, where each $S_i \subseteq \mathcal{R}$ is a possibly infinite set of profiles, and $h(S_i)$ lies in a *one-dimensional* subspace of V. Let T be a (possibly infinite) set of linear functions from V to $\mathbb{Q}$; we refer to these functions as *linear predicates*. Then the $\mathcal{L}$-axiomatization $\mathcal{S}(f, h, g, \circ, V, S, T)$ consists of the following four axioms:

1. **ADD**: $\forall \mathbf{R}_1, \mathbf{R}_2 \in \mathcal{R}, A_1 \circ A_2 \neq \emptyset$,
   $[\mathbf{R}_1 \mapsto A_1] \wedge [\mathbf{R}_2 \mapsto A_2] \rightarrow [\mathbf{R}_1 \bigoplus \mathbf{R}_2 \mapsto A_1 \circ A_2]$.
2. **EMB**: $\forall \mathbf{R}_1, \mathbf{R}_2 \in \mathcal{R}$ such that $h(\mathbf{R}_1) = h(\mathbf{R}_2)$,
   $[\mathbf{R}_1 \mapsto A_1] \rightarrow [\mathbf{R}_2 \mapsto A_2]$.
3. **INIT**: $\forall \mathbf{R} \in S$, $[\mathbf{R} \mapsto f(\mathbf{R})]$.
4. **PRED**: $\forall \mathbf{R} \in \mathcal{R}, t_i \in T$,
   $t_i(h(\mathbf{R})) = 0 \rightarrow \bigvee_{A \in g(\mathcal{K}(t_i))} [\mathbf{R} \mapsto A]$, where $\mathcal{K}(t_i)$ is the kernel of t_i.

If $\circ = \cap$ then **ADD** is simply reinforcement. The **INIT** axioms are intraprofile axioms that prescribe the outcome on the set S of base profiles; this is similar to the intraprofile axioms we saw for Borda, where S would consist of elementary, cyclic, and cancellation profiles. **EMB** encodes the fact that if a voting rule f is embedded into V by an embedding h, g, then f must have the same outcome on $\mathbf{R}_1$ and $\mathbf{R}_2$ if $h(\mathbf{R}_1) = h(\mathbf{R}_2)$. For example, when embedding Borda using delta vectors, this axiom would say that two profiles with the same delta vector must yield the same outcome.

Finally, the **PRED** axiom is an intraprofile axiom which does not have an analogue among the axioms discussed in Section 3. We include it to give our framework more expressive power, especially for encoding neutrality-type axioms (which require that similar alternatives need to be treated identically). **PRED** encodes the fact that if a profile satisfies some condition (given by t_i) then its outcome should reflect it. For example, if f is the Borda rule embedded into $V = \mathbb{Q}^m$ by its scoring function, we can let $T = \{t_{ij} : t_{ij}(v) = v_i - v_j\}$, where v_i is the score of alternative i. Then $\mathcal{K}(t_{ij})$ is the set of vectors v with $v_i = v_j$, and thus $g(\mathcal{K}(t_{ij}))$ is the set of voting outcomes A which satisfy $i \in A$ if and only if $j \in A$. In other words, **PRED** would require that any two alternatives with the same Borda score are either both winners or both losers.

Our lower bound is stated in terms of the dimension of the space V and a complexity measure of the set T of linear predicates. The definition of that measure is admittedly unwieldy, but directly related to the length of explanations.

Definition 3. An outcome A is *uniquely determined* by a subset of linear predicates $T' \subseteq T$ if there is a set of rational numbers $C \subseteq \mathbb{Q}$ such that $A = \bigcap_{t_i \in T'} g(t_i^{-1}(c_i))$, where $t_i^{-1}(c_i) = \{v : t_i(v) = c_i\}$ and $c_i \in C$. The *sensitivity* of g with respect to T, $\mathrm{sen}(g, T)$, is the minimum size of $T' \subseteq T$ such that there is an outcome A (of f) that is uniquely determined by T'. If there is no such T', $\mathrm{sen}(g, T) = +\infty$.

For the embedding of Borda into $\mathbb{Q}^m$ with $T = \{t_{ij} : t_{ij}(v) = v_i - v_j\}$, we can take the winner set $A = \mathcal{A}$. The function g outputs $\mathcal{A}$ only for vectors v with $v_k = v_\ell$ for all k, ℓ. Now, $t_{ij}^{-1}(0)$ is the set of vectors v with $v_i = v_j$. Thus, the smallest set T' such that $\bigcap_{t_i \in T'} g(t_i^{-1}(0)) = \mathcal{A}$ is $T' = \{t_{12}, t_{13}, \ldots, t_{1m}\}$. It follows that $\mathrm{sen}(g, T) = m - 1$.

In the next section, we will first obtain lower bounds for axiomatizations in the form of Definition 2. In fact, we are more interested in axiomatizations like the one for Borda in Section 3. The following definition describes a class of axiomatizations generalizing those of Definition 2, to which our lower bound immediately applies.

Definition 4. An axiomatization $\mathcal{S}$ of a voting rule f is *asymptotically weaker* than $\mathcal{S}_{\text{emb}} = \mathcal{S}(f, h, g, \circ, V, S, T)$ if there is $c \geq 1$ such that for every axiom instance φ in $\mathcal{S}$, there exists a proof of φ assuming $\mathcal{S}_{\text{emb}}$ that uses at most c axiom instances of form **INIT** or **PRED**, as well as an unlimited number of **ADD** and **EMB** axiom instances.

For example, for any $k \in \mathbb{Z}_+$, $\mathcal{S}$ could include the axiom

$$[\mathbf{R} \mapsto A] \rightarrow \Big[k\mathbf{R} \mapsto \underbrace{A \circ A \circ \ldots \circ A}_{k \text{ times}}\Big]$$

since it can be deduced by repeatedly using **ADD**.

4.2. Theorem Statement and Proof

In this section we give a lower bound on the length of explanation required for random profiles. Specifically, assume that the preferences of voters are independent, and each voter picks their ranking over all $m!$ possibilities uniformly at random. This common assumption is known as the *impartial culture assumption* in social choice theory (Tsetlin et al., 2003). Let $\mathbf{R}^n$ denote the random profile generated this way for the case of n voters. When we say that $\mathbf{R}^n$ satisfies a property "with high probability," we mean that the probability converges to 1 as n goes to infinity.

Theorem 2. *Let f be a voting rule that can be embedded into a linear space V of finite dimension d by h and g. Consider an axiomatization $\mathcal{S}$ of f that is asymptotically weaker than some axiomatization $\mathcal{S}(f, h, g, \circ, V, S, T)$ based on operation $\circ$, base profiles S, and linear predicates T, satisfying the conditions of Definition 2. Then, with high probability, every explanation of the outcome $f(\mathbf{R}^n)$ at the random profile $\mathbf{R}^n$ using $\mathcal{S}$ consists of $\Omega(\min(d, \mathrm{sen}(g, T)))$ steps.*

The impartial culture assumption is debatable as a model of voter preferences, but in our case that objection is essentially irrelevant: because the theorem's conclusion holds with high probability, it provides a lower bound with respect to *almost every profile*. Moreover, the proof can be adapted to work for any $\mathcal{D}$ over $\mathcal{R}$ such that $h(\mathrm{supp}(\mathcal{D}))$ spans V; we focus on impartial culture for ease of exposition.

The high-level idea of the theorem's proof is as follows. We are going to show that for large enough n, if we only use a set $\mathcal{B}$, $|\mathcal{B}| = d-1$, of axioms from **INIT**, and a set $\mathcal{C}$, $|\mathcal{C}| = \mathrm{sen}(g,T)-1$, of axioms from **PRED** (and as many axioms as we want from **ADD** and **EMB**) in the explanation of $[\mathbf{R}^n \mapsto f(\mathbf{R}^n)]$, then with high probability, for every such $\mathcal{B}$ and $\mathcal{C}$ there is another voting rule f', which satisfies all axioms in **ADD**, **EMB**, $\mathcal{B}$ and $\mathcal{C}$, such that f' disagrees with f on $\mathbf{R}^n$, i.e. $f'(\mathbf{R}^n) \neq f(\mathbf{R}^n)$. This, in turn, implies that $\neg[\mathbf{R}^n \mapsto f(\mathbf{R}^n)]$ satisfies **ADD**, **EMB**, $\mathcal{B}$ and $\mathcal{C}$, which is a contradiction to the soundness of propositional logic. Thus, any proof of $[\mathbf{R}^n \mapsto f(\mathbf{R}^n)]$ using the axiomatization $\mathcal{S}_{\text{emb}} = \mathcal{S}(f,h,g,\circ,V,S,T)$ will use at least $\min(d,\mathrm{sen}(g,T))$ axiom instances of types **INIT** and **PRED**. Now, any proof of $[\mathbf{R}^n \mapsto f(\mathbf{R}^n)]$ in an asymptotically weaker axiomatization can be translated into a proof in $\mathcal{S}_{\text{emb}}$ with at most a constant factor blow-up in length. Thus, the proof using the asymptotically weaker axiomatization must have length $\Omega(\min(d,\mathrm{sen}(g,T)))$.

Turning to the proof itself, define the random vector $\xi_i = h(R_i)$, where R_i is the ranking (or single-voter profile) associated with the i^{th} voter (which is selected independently and uniformly at random from all $m!$ possible rankings). Fix an arbitrary basis $B \subseteq V$, and let $c_B(v)$ be the coordinates of a vector v under B. Define $X_i = c_B(b)^{\intercal} c_B(\xi_i)$, for some arbitrary non-zero vector $b \in V$. $X_1,\ldots,X_n$ are i.i.d. random variables with mean μ and variance σ^2.

Lemma 1. *For any vector $b \neq 0$, the random variable $c_B(b)^{\intercal} c_B(\xi_i)$ has non-zero mean or non-zero variance.*

Proof. We prove that if $c_B(b)^{\intercal} c_B(\xi_i)$ is a random variable with zero mean then it cannot be deterministically zero (and thus has non-zero variance). Towards a contradiction, $\Pr[c_B(b)^{\intercal} c_B(\xi_i) = 0] = 1$ implies that $c_B(b)^{\intercal} c_B(h(\mathbf{R})) = 0$ for every $\mathbf{R} \in \mathcal{R}$. But, by the definition of an embedding, $\{h(\mathbf{R}) : \mathbf{R} \in \mathcal{R}\}$ spans V. Therefore $c_B(b)$ must be the all zeros vector; a contradiction. □

Lemma 2. *For any non-zero vector b it holds that $c_B(b)^{\intercal} c_B(h(\mathbf{R}^n)) \neq 0$ with high probability.*

Proof. To prove the lemma we need to show that $\sum_{i=1}^n X_i \neq 0$ with high probability. By Lemma 1 either $\mu \neq 0$ or $\sigma \neq 0$. If $\sigma = 0$, then the X_i's are identical non-zero constants (note that this does not imply that the $\xi_i = h(R_i)$ is a constant vector). We trivially get that $\Pr[\sum_{i=1}^n X_i = 0] = 0$, for all n. If $\sigma \neq 0$ then the central limit theorem gives us that

$$\lim_{n\to\infty}\sup_{z\in\mathbb{R}}\left|\Pr\left[\sqrt{n}\left(\frac{\sum_{i=1}^n X_i}{n}-\mu\right)\le z\right]-\Phi\left(\frac{z}{\sigma}\right)\right| = 0.$$

Therefore, for any given $\epsilon > 0$, we have

$$\begin{aligned}
\Pr\left[\left|\sum_{i=1}^n X_i\right| \le \epsilon\right] &= \Phi\left(\frac{\epsilon-\mu n}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{-\epsilon-\mu n}{\sqrt{n}\sigma}\right) + o(1) \\
&= \frac{1}{\sqrt{2\pi}}\int_{(-\epsilon-\mu n)/(\sqrt{n}\sigma)}^{(\epsilon-\mu n)/(\sqrt{n}\sigma)} e^{-x^2/2}\,\mathrm{d}x + o(1) \\
&\le \frac{1}{\sqrt{2\pi}}\int_{(-\epsilon-\mu n)/(\sqrt{n}\sigma)}^{(\epsilon-\mu n)/(\sqrt{n}\sigma)} 1\,\mathrm{d}x + o(1) \\
&= \frac{2\epsilon}{\sigma\sqrt{2\pi n}} + o(1), \qquad (1)
\end{aligned}$$

For any $\delta > 0$ we can pick ϵ small enough, and n large enough, so that both terms on the right hand side of Equation (1) are smaller than $\delta/2$. It then holds that

$$\delta > \Pr\left[\left|\sum_{i=1}^n X_i\right| \le \epsilon\right] > \Pr\left[\sum_{i=1}^n X_i = 0\right]. \qquad \square$$

Lemma 3. *With high probability $h(\mathbf{R}^n)$ does not lie in the subspace spanned by any $d-1$ elements in $h(S)$.*

Proof. We can assume that $h(S)$ spans V, since otherwise we can add vectors from V to make it so (and the lemma still holds). We start by proving the lemma for finite $h(S)$. We are going to show that, with high probability, the coordinates of $h(\mathbf{R}^n)$ under any basis $B' \subseteq h(S)$ of V have no zero entries. This implies that $h(\mathbf{R}^n)$ is not in any subspace spanned by $d-1$ elements in $h(S)$. To see why this is the case, notice that if $h(\mathbf{R}^n)$ lied in the space spanned by some $B' \subseteq h(S)$, with $|B'| = d-1$, we could add one more $v \in V$ to B' and make it a basis for V. Then the coordinates of $h(\mathbf{R}^n)$ with respect to v under this basis would be zero, leading to a contradiction; thus, such a B' cannot exist.

Recall that we have already fixed one a basis for V, the basis B. For every basis $B' \subseteq h(S)$, there is a unique non-singular matrix $P_{B'}$ such that $B = B'P_{B'}$, and for any $v \in V$, $c_{B'}(v) = P_{B'}c_B(v)$. Thus, it is sufficient to prove that for every basis $B' \subseteq h(S)$ all entries of $P_{B'}c_B(h(\mathbf{R}^n))$ are non-zero with high probability.

For any B', by the union bound, the probability that $P_{B'}c_B(h(\mathbf{R}^n))$ has a zero entry is at most $\sum_{i=1}^d \Pr[P_{B'}^i c_B(h(\mathbf{R}^n)) = 0]$, where $P_{B'}^i$ is the i^{th} row of $P_{B'}$. Due to its non-singularity, each row of $P_{B'}$ must be non-zero. By Lemma 2 it holds that $\Pr[P_{B'}^i c_B(h(\mathbf{R}^n)) = 0]$ converges to zero as n goes to infinity. By applying the union bound again, we conclude that with high probability $P_{B'}c_B(h(\mathbf{R}^n))$ has no zero entries for every basis $B' \subseteq h(S)$. This concludes the proof for finite $h(S)$.

When $h(S)$ is infinite, we use that $h(S)$ is a union of finitely many one-dimensional subsets, $h(S) = \bigcup_{k=1}^N h(S_i)$. Pick an arbitrary non-zero vector b_i from each $h(S_i)$ and let $\mathbf{B}^* = \bigcup_{k=1}^{N}\{b_i\}$. Notice that $\mathbf{B}^*$ spans V, since each $h(S_i)$ is one dimensional. Therefore, we can use the finite version of this lemma for $\mathbf{B}^*$ and get that $h(\mathbf{R}^n)$ does not lie in the subspace spanned by any $d-1$ elements in $\mathbf{B}^*$.

We claim that $h(\mathbf{R}^n)$ does not lie in the subspace spanned by any $d-1$ elements in $h(S)$ either. Towards a contradiction, assume that $h(\mathbf{R}^n)$ lies in the subspace spanned by $B' \subseteq h(S)$, with $|B'| = d-1$. Without loss of generality we assume that the vectors in B' are linearly independent. Then, every element of B' must come from a different $h(S_i)$ (since each $h(S_i)$ is one-dimensional). Let $B' = \{b'_{i_1}, b'_{i_2}, \dots, b'_{i_{d-1}}\}$ and $h(\mathbf{R}^n) = \sum_{j=1}^{d-1} c_j b'_{i_j}$ where b'_{i_j} is an element of $h(S_{i_j})$, and the c_j are rational numbers. Since each $h(S_i)$ is one dimensional, $b'_{i_j} = q_{i_j} b_{i_j}$ where q_{i_j} is rational and b_{i_j} is the vector we included in $\mathbf{B}^*$. We immediately have that $h(\mathbf{R}^n) = \sum_{j=1}^{d-1} c_j q_{i_j} b_{i_j}$, which implies that $h(\mathbf{R}^n)$ lies in a subspace spanned by $d-1$ elements in $\mathbf{B}^*$ — a contradiction. □

We are now ready to complete the theorem's proof.

Proof of Theorem 2. Let $\mathcal{B} \subseteq S$, $|\mathcal{B}| = d-1$, be a set of **INIT** axiom instance and let $\mathcal{C}$, $|\mathcal{C}| = \operatorname{sen}(g,T) - 1$, be a set of **PRED** axiom instances. We show that there exists no proof of $[\mathbf{R}^n \mapsto f(\mathbf{R}^n)]$ using the axiomatization $\mathcal{S}_{\text{emb}} = \mathcal{S}(f,h,g,\circ,V,S,T)$ that uses only **INIT** axiom instances in $\mathcal{B}$ and only **PRED** axiom instance in $\mathcal{C}$.

Slightly abusing notation, let $h(\mathcal{B}) = \{v_1, v_2, \dots, v_{d-1}\}$, and assume without loss of generality that these $d-1$ vectors are linearly independent. Also assume that $h(\mathbf{R}^n)$ is not in any subspace spanned by $d-1$ elements of $h(S)$, which happens with high probability by Lemma 3. Therefore, $\{v_1, v_2, \dots, v_{d-1}, h(\mathbf{R}^n)\}$ forms a linear basis of V.[3] Furthermore, since there are at most $\operatorname{sen}(g,T) - 1$ linear predicates in $\mathcal{C}$, by the definition of sensitivity, no winning set A of f can be uniquely determined by the linear predicates in the axioms of $\mathcal{C}$. Thus, we can find a vector b such that $t_i(b) = t_i(h(\mathbf{R}^n))$ for all $t_i \in \mathcal{C}$ and $g(b) \neq g(h(\mathbf{R}^n))$.

For a profile $\mathbf{R}$ we can write $h(\mathbf{R}) = k_1 v_1 + k_2 v_2 + \cdots + k_{d-1} v_{d-1} + k_d h(\mathbf{R}^n)$, for rational $k_1, \dots, k_d$. Define a new embedding $h' : \mathcal{R} \to V$ by $h'(\mathbf{R}) = k_1 v_1 + k_2 v_2 + \cdots + k_{d-1} v_{d-1} + k_d b$, where b is as above. Due to the uniqueness of this decomposition, h' is well-defined.

Now consider the voting rule f' that outputs $g(h'(\mathbf{R}))$ on a profile $\mathbf{R}$. First, notice that f and f' disagree on $\mathbf{R}^n$, as $f'(\mathbf{R}^n) = g(h'(\mathbf{R}^n)) = g(b) \neq f(\mathbf{R}^n)$ by the choice of b. Second, for each profile $\mathbf{R}$ in $\mathcal{B}$, we have $h(\mathbf{R}) = h'(\mathbf{R})$, and therefore $f'(\mathbf{R}) = f(\mathbf{R})$. Third, for two profiles $\mathbf{R}_1, \mathbf{R}_2$ such that $h(\mathbf{R}_1) = h(\mathbf{R}_2)$ we have $h'(\mathbf{R}_1) = h'(\mathbf{R}_2)$, which implies $f'(\mathbf{R}_1) = f'(\mathbf{R}_2)$. Fourth, if $f'(\mathbf{R}) \circ f'(\mathbf{R}') \neq \emptyset$ then $f'(\mathbf{R} \bigoplus \mathbf{R}') = g(h'(\mathbf{R}) + h'(\mathbf{R}')) = f'(\mathbf{R}) \circ f'(\mathbf{R}')$. Finally, $t_i(h(\mathbf{R})) = 0$ implies $t_i(h'(\mathbf{R})) = 0$.

The above facts imply that the new rule f' satisfies **ADD**, **EMB**, $\mathcal{B}$ and $\mathcal{C}$, but that $f(\mathbf{R}^n) \neq f'(\mathbf{R}^n)$. Thus, by the consistency of $\mathcal{L}$ and the soundness of propositional logic, $\mathbf{R}^n$'s outcome cannot be explained assuming $\mathcal{S}_{\text{emb}}$ without using **INIT** axioms outside $\mathcal{B}$ or **PRED** axioms outside $\mathcal{C}$. Since this holds for any $\mathcal{B}$ and $\mathcal{C}$, every explanation of $f(\mathbf{R}^n)$ assuming $\mathcal{S}_{\text{emb}}$ must contain at least $\min(d, \operatorname{sen}(g,T))$ axiom instances of type **INIT** and **PRED**.

Consider a proof of $[\mathbf{R}^n \mapsto f(\mathbf{R}^n)]$ assuming $\mathcal{S}$ of length r, which is formally a sequence $\varphi_1, \dots, \varphi_r$ of formulae. By assumption, $\mathcal{S}$ is asymptotically weaker than $\mathcal{S}_{\text{emb}}$. Thus, for each φ_i in the proof which is an axiom instance of $\mathcal{S}$, we can replace φ_i by a proof of φ_i assuming $\mathcal{S}_{\text{emb}}$. After these replacements, we have obtained a proof of $[\mathbf{R}^n \mapsto f(\mathbf{R}^n)]$ assuming $\mathcal{S}_{\text{emb}}$; let s be the number of intraprofile axiom instances (**INIT** and **PRED**) in this proof. Because we have obtained this proof by performing at most r replacements, each time introducing at most c intraprofile axiom instances, we have $s \leq c \cdot r$. From above, we know that $s \geq \min(d, \operatorname{sen}(g,T))$. Thus, $r \geq \frac{1}{c}\min(d, \operatorname{sen}(g,T))$. Hence, any explanation of the outcome $f(\mathbf{R}^n)$ using the axiomatization $\mathcal{S}$ requires $\Omega(\min(d, \operatorname{sen}(g,T)))$ steps. □

4.3. Implications for Prominent Voting Rules

In this section, we apply the general bound of Theorem 2 to several existing axiomatizations of important voting rules. We start with the axiomatization of the Borda rule that we used in Theorem 1. There, we showed that the outcome of the Borda rule at any profile can be explained in $O(m^2)$ steps. We can now show that this is asymptotically tight.

Corollary 1. *With high probability, the outcome of the Borda rule on a random profile $\mathbf{R}^n$ requires $\Omega(m^2)$ steps to explain assuming the $\mathcal{L}$-axiomatization $\mathcal{S}_{\text{Borda}}$.*

Proof sketch. Let $\mathcal{S}(f,h,g,\circ,V,S,T)$ be the $\mathcal{L}$- axiomatization of the Borda rule, where $V = \mathbb{Q}^{\binom{m}{2}}$, $h(\mathbf{R}) = \delta^{\mathbf{R}}$ and $g(\delta) = \arg\max_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} \delta_{ab}$. Because Borda satisfies reinforcement, g admits intersection. We let S be the set consisting of elementary profiles, of cyclic profiles (as defined in Section 3), and of cancellation profiles: a cancellation profile is one in which for each pair a, b, there is an equal number of voters preferring a to b, and preferring b to a. Note that S is made up of a finite number of elementary and cyclic profiles, plus an infinite number of cancellation profiles which are all mapped to the all-zero delta vector by

[3] In the case that $h(S)$ spans $V' \subsetneq V$, the vectors in $h(\mathcal{B})$ cannot be linearly independent. But, we can still create a basis for V that includes a maximal subset of linearly independent vectors from $h(\mathcal{B})$, $h(\mathbf{R}^n)$ and other vectors from V.

h. Hence, S satisfies the condition of being a finite union of sets whose image under h is contained in a one-dimensional subspace. Finally, we set $T = \emptyset$.

We need to show that this axiomatization is asymptotically weaker than the axiomatization described in Section 3. Then Theorem 2 implies the desired result, since the dimension of V is $\Theta(m^2)$, and $\mathrm{sen}(g, T) = +\infty$.

Each **ELEM**, **CYCL**, and **CANC** axiom is an **INIT** axiom. Each **REINF** axiom is an **ADD** axiom. Each **MULT** axiom can be deduced by repeatedly applying an **ADD** axiom. Each **REINF-SUB** axiom can be inferred by combining several **ADD** axioms using propositional reasoning; similarly **SIMP** axioms can be inferred by combining **ADD** axioms. The details of the formal deductions are in Appendix A.2. □

Next, we consider the *plurality rule*. Under this rule, the winning alternatives are those that are ranked in first position by the largest number of voters. Sekiguchi (2012) has given a characterization of this rule (based on an earlier result of Yeh (2008)), using the axioms anonymity, neutrality, reinforcement, faithfulness, and tops-only. Inspecting the proof, we see that the full neutrality axiom is not needed, and only the *orbit axiom*[4] (Brandt & Geist, 2016) is required. Faithfulness requires that if there is only one voter, then the only winning alternative is the top choice of that voter. Tops-only requires that if in two different profiles $\mathbf{R}_1$ and $\mathbf{R}_2$ defined on the same voters, each voter ranks the same alternative as top choice in both profiles, then $f(\mathbf{R}_1) = f(\mathbf{R}_2)$. Using appropriate fomalizations of the axioms, is possible to translate Sekiguchi's (2012) proof into an explanation of the plurality outcome in our propositional language, and the resulting proofs will be of length $O(m)$. Using Theorem 2, we can show that this is tight. The proof of the following corollary can be found in Appendix B. We can strengthen this lower bound by adding additional axioms, and so we add the strong axiom of *equal support*, which says that in a profile where each alternative has either plurality score 1 or 0, the alternatives with score 1 are elected.

Corollary 2. *With high probability, the outcome of the plurality rule on a random profile $\mathbf{R}^n$ requires $\Omega(m)$ steps to explain, under the axioms of anonymity, reinforcement, orbit, faithfulness, equal support and tops-only.*

Throughout this paper, we have discussed voting rules whose input is specified by profiles of rankings. An alternative paradigm is used by *approval voting* (Brams & Fishburn, 2007), which allows voters to indicate, for each alternative, whether they approve or disapprove it. Then, formally, the input to the voting rule is a profile of subsets (of approved alternatives), rather than rankings. Approval voting declares that those alternatives that have been approved by the highest number of voters are winners. This rule has been axiomatically characterized among voting rules with this input format by Fishburn (1978; 1979). He gives two axiomatizations, both using axioms similar to previous examples. After redefining the set $\mathcal{R}$ of profiles to use approval ballots, our general lower bound (Theorem 2) can be proven verbatim, and implies an $\Omega(m)$ lower bound for explanations of approval voting obtained using Fishburn's axiomatizations; details are relegated to Appendix C.

Corollary 3. *With high probability, the outcome of approval voting on a random profile $\mathbf{R}^n$ of approval ballots requires $\Omega(m)$ steps to explain, under the axioms of anonymity, reinforcement, orbit, faithfulness, disjoint equality, cancellation.*

5. Discussion

We wrap up with a discussion of the practical implications of our theoretical results.

First, we wish to emphasize that our main result, Theorem 2, holds with respect to almost every profile. Focusing on Borda as an example, this means that the (computationally-efficient) explanation-generating algorithm given by Theorem 1 not only constructs the (asymptotically) shortest possible explanations in the worst case — it constructs the (asymptotically) shortest possible explanation with respect to almost every profile. That said, since these results are asymptotic, there might be some benefit in designing a search algorithm that computes the absolutely shortest explanation for any given profile. This appears to be a very difficult computational problem; preliminary experiments suggest that standard heuristic search or mathematical programming techniques cannot be used to directly tackle it.

Second, whether our results should be seen as positive or negative depends on the application. Most group decisions — such as those made through online services like `Robovote.org` — involve only a few alternatives: restaurants, movies, vacation spots, best paper awards (from a shortlist), or even prototypes to develop. In these cases, an explanation of linear or quadratic length is perfectly reasonable, so our results should be viewed in a positive light.

By contrast, some settings involve many alternatives. For example, in the work of Lee et al. (2019) — where the Borda rule is used to aggregate predicted preferences — the set of alternatives consists of hundreds of nonprofit organizations that may receive an incoming food donation. In this case, an explanation of quadratic length is a nonstarter, although linear-length explanations (such as those available for plurality and approval) may be viable. The good news is that Theorem 2 can help identify new axiomatizations that lead to short explanations, by providing necessary conditions; to

[4] Informally, the orbit axiom says that if in a profile alternatives a and b are symmetric (in the sense that swapping their names does not change the profile) then either both are winners or neither are.

explain Borda outcomes when there are hundreds of alternatives, we must find a new axiomatization that does not take the form of the axioms in Section 4.1. Our hope is that these insights will lead to new results in social choice theory, which could, in turn, be used to design explainable AI systems that are currently beyond reach.

References

Arrow, K. *Social Choice and Individual Values*. Wiley, 1951.

Ben-Ari, M. *Mathematical Logic for Computer Science*. Springer, 3rd edition, 2012.

Brams, S. J. and Fishburn, P. C. *Approval Voting*. Springer, 2nd edition, 2007.

Brandl, F. and Peters, D. Simple characterizations of approval voting. Technical report, Working Paper, 2019.

Brandt, F. and Geist, C. Finding strategyproof social choice functions via SAT solving. *Journal of Artificial Intelligence Research*, 55:565–602, 2016.

Cailloux, O. and Endriss, U. Arguing about voting rules. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 287–295, 2016.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. L-Shapley and C-Shapley: Efficient model interpretation for structured data. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

Condorcet, M.-J.-A.-N. Essai sur l'application de l'analyse à la probabilité de décisions rendues à la pluralité de voix. Imprimerie Royal, 1785. Facsimile published in 1972 by Chelsea Publishing Company, New York.

Fishburn, P. C. Axioms for approval voting: Direct proof. *Journal of Economic Theory*, 19(1):180–185, 1978.

Fishburn, P. C. Symmetric and consistent aggregation with dichotomous voting. In Laffont, J. J. (ed.), *Aggregation and Revelation of Preferences*. North-Holland, 1979.

Freedman, R., Schaich Borg, J., Sinnott-Armstrong, W., Dickerson, J. P., and Conitzer, V. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1636–1643, 2018.

Kahng, A., Lee, M. K., Noothigattu, R., Procaccia, A. D., and Psomas, C.-A. Statistical foundations of virtual democracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3173–3182, 2019.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., Noothigattu, R., See, D., Lee, S., Psomas, C.-A., and Procaccia, A. D. WeBuildAI: Participatory framework for fair and efficient algorithmic governance. In *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, article 181, 2019.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 4768–4777, 2017.

Noothigattu, R., Gaikwad, S. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., and Procaccia, A. D. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1587–1594, 2018.

Procaccia, A. D. Axioms should explain solutions. In Laslier, J.-F., Moulin, H., Sanver, R., and Zwicker, W. S. (eds.), *The Future of Economic Design*. Springer, 2019.

Sekiguchi, Y. A characterization of the plurality rule. *Economics Letters*, 116(3):330–332, 2012.

Štrumbelj, E. and Kononenko, I. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18, 2010.

Tsetlin, I., Regenwetter, M., and Grofman, B. The impartial culture maximizes the probability of majority cycles. *Social Choice and Welfare*, 21:387–398, 2003.

Yeh, C.-H. An efficiency characterization of plurality rule in collective choice problems. *Economic Theory*, 34(3): 575–583, 2008.

Young, H. P. An axiomatization of Borda's rule. *Journal of Economic Theory*, 9(1):43–52, 1974.

A. Details for Borda

A.1. Proof of Theorem 1

It will be useful to define the *beta score* of an alternative a in profile $\mathbf{R}$ as $\beta^{\mathbf{R}}_a = 2b_a \cdot m - m(m-1)$ where b_a is the Borda score of a. The beta score is a monotonically increasing linear function of the Borda score. Therefore, selecting the top alternatives based on beta scores or Borda scores defines the same voting rule. For any profile $\mathbf{R}$ the *beta vector* $\beta^{\mathbf{R}}$ maps alternatives to their beta score. Note that $\beta^{\mathbf{R}}$ is a linear transformation of $\delta^{\mathbf{R}}$. More precisely, define $\hat{\beta}(\delta^{\mathbf{R}})$ as the following beta vector: $\hat{\beta}(\delta^{\mathbf{R}})_a = \sum_{b \in \mathcal{A} \setminus \{a\}} \delta^{\mathbf{R}}_{ab}$.

We begin by establishing two lemmas that relate the length of a explanation in a profile $\mathbf{R}$ to the length of the explanation in a profile with a similar delta vector. Recall that delta vectors are a sufficient statistic to compute Borda outcomes. Therefore, if two profiles $\mathbf{R}_1$ and $\mathbf{R}_2$ have identical delta vectors then they have the same set of winners under Borda. The following lemma shows that given such profiles $\mathbf{R}_1$ and $\mathbf{R}_2$, and the set of winners of one of the two, we can produce a proof of constant length that the other profile has the same set of winners.

Lemma 4. *Let $\mathbf{R_1}$ and $\mathbf{R_2}$ be two profiles with the same delta vector. Given that $[\mathbf{R_1} \mapsto A]$, then $[\mathbf{R_2} \mapsto A]$ can be explained by* **CANC**, **REINF** *and* **REINF-SUB** *in $O(1)$ steps.*

Proof. Let $\overline{\mathbf{R}_1}$ be the profile with the same voters as $\mathbf{R}_1$, but reversed preferences. Clearly, for all alternatives $a, b \in \mathcal{A}$, $\delta_{ab}^{\overline{\mathbf{R}_1}} = -\delta_{ab}^{\mathbf{R}_1}$. We have the following explanation:

1. $[\mathbf{R}_1 \mapsto A]$
2. $[\mathbf{R}_2 \bigoplus \overline{\mathbf{R}_1} \mapsto \mathcal{A}]$ (**CANC**)
3. $(1) \wedge (2) \to [\mathbf{R}_2 \bigoplus \overline{\mathbf{R}_1} \bigoplus \mathbf{R}_1 \mapsto A]$ (**REINF**)
4. $[\mathbf{R}_2 \bigoplus \overline{\mathbf{R}_1} \bigoplus \mathbf{R}_1 \mapsto A]$ (propositional reasoning from 1–3)
5. $[\mathbf{R}_1 \bigoplus \overline{\mathbf{R}_1} \mapsto \mathcal{A}]$ (**CANC**)
6. $(4) \wedge (5) \to [\mathbf{R}_2 \mapsto A]$ (**REINF-SUB**)
7. $[\mathbf{R_2} \mapsto A]$ (propositional reasoning from 4–6) □

The next lemma shows a similar fact about sums of profiles.

Lemma 5. *Let $\mathbf{R}$, $\mathbf{R}_E$ and $\mathbf{R}_C$ be profiles such that $k_1\delta^{\mathbf{R}} = k_2\delta^{\mathbf{R}_E \bigoplus \mathbf{R}_C}$ for integers k_1, k_2, and assume that $[\mathbf{R}_E \mapsto A]$ and $[\mathbf{R}_C \mapsto \mathcal{A}]$. Then $[\mathbf{R} \mapsto A]$ can be explained in $O(1)$ steps.*

Proof. The explanation works as follows.

1. $[\mathbf{R}_E \mapsto A]$
2. $[\mathbf{R}_C \mapsto \mathcal{A}]$
3. $(1) \wedge (2) \to [\mathbf{R}_E \bigoplus \mathbf{R}_C \mapsto A]$ (**REINF**)
4. $[\mathbf{R}_E \bigoplus \mathbf{R}_C \mapsto A] \to [k_2(\mathbf{R}_E \bigoplus \mathbf{R}_C) \mapsto A]$ (**MULT**)
5. $[k_2(\mathbf{R}_E \bigoplus \mathbf{R}_C) \mapsto A]$ (propositional reasoning from 1–4)
6. $[k_1\mathbf{R} \mapsto A]$ (5 and Lemma 4)
7. $[k_1\mathbf{R} \mapsto A] \to [\mathbf{R} \mapsto A]$ (**SIMP**)
8. $[\mathbf{R} \mapsto A]$ (propositional reasoning from 6–7)

where the sixth step contains the constant length explanation of Lemma 4. □

The remainder of the proof focuses on the following task: given a profile $\mathbf{R}$ with A the set of Borda winners, construct and explain two profiles $\mathbf{R}_E$ and $\mathbf{R}_C$ such that (1) $k_1\delta^{\mathbf{R}} = k_2\delta^{R_E \bigoplus \mathbf{R}_C}$, (2) $[\mathbf{R}_E \mapsto A]$, and (3) $[\mathbf{R}_C \mapsto \mathcal{A}]$. Specifically, $\mathbf{R}_E$ will be a sum of elementary profiles whose winner sets have a non-empty intersection, and $\mathbf{R_C}$ will be a sum of cyclic profiles. Our approach borrows ideas and facts from the analysis of the algorithm *Borda-expl* presented by Cailloux & Endriss (2016).

To construct $\mathbf{R}_E$, label the alternatives as $a_1, a_2, \ldots, a_m$ in order of decreasing beta scores, so $\beta_{a_1}^{\mathbf{R}} \geq \ldots \geq \beta_{a_m}^{\mathbf{R}}$. Let

$$\mathbf{R}_E = \bigoplus_{i=1}^{m-1} \mathbf{R}_i,$$

where

$$\mathbf{R}_i = \begin{cases} \frac{\beta_{a_i}^{\mathbf{R}} - \beta_{a_{i+1}}^{\mathbf{R}}}{2} \mathbf{R}_{\text{elem}}^{\{a_1,\ldots,a_i\}} & \text{if } \beta_{a_i}^{\mathbf{R}} - \beta_{a_{i+1}}^{\mathbf{R}} > 0, \\ \mathbf{R}_{\text{elem}}^{\mathcal{A}} & \text{if } \beta_{a_i}^{\mathbf{R}} - \beta_{a_{i+1}}^{\mathbf{R}} = 0. \end{cases}$$

Note that beta scores are always even, so $(\beta_{a_i}^{\mathbf{R}} - \beta_{a_{i+1}}^{\mathbf{R}})/2$ is a non-negative integer.

Lemma 6. *$[\mathbf{R}_E \mapsto A]$ can be explained in $O(m)$ steps.*

Proof. If $\beta_{a_i}^{\mathbf{R}} = \beta_{a_{i+1}}^{\mathbf{R}}$ for all $i = 1, \ldots, m-1$, then $\mathbf{R}_E$ is composed of copies of the profile $\mathbf{R}_{\text{elem}}^{\mathcal{A}}$. Hence by **MULT** and **ELEM**, we obtain $[\mathbf{R}_E \mapsto \mathcal{A}]$, as required, since in this case $A = \mathcal{A}$.

Otherwise, let k be the smallest index with $\beta_{a_k}^{\mathbf{R}} - \beta_{a_{k+1}}^{\mathbf{R}} > 0$. For each $i = 1, \ldots, m$, an **ELEM** axiom gives

$$\left[\mathbf{R}_{\text{elem}}^{\{a_1,\ldots,a_i\}} \mapsto \{a_1, \ldots, a_i\}\right].$$

If $\beta_{a_i}^{\mathbf{R}} - \beta_{a_{i+1}}^{\mathbf{R}} > 0$, by **MULT** we have

$$\left[\frac{\beta_{a_i}^{\mathbf{R}} - \beta_{a_{i+1}}^{\mathbf{R}}}{2} \mathbf{R}_{\text{elem}}^{\{a_1,\ldots,a_i\}} \mapsto \{a_1, \ldots, a_i\}\right].$$

Note that $\{a_1, \ldots, a_i\} \subseteq \{a_1, \ldots, a_{i+1}\} \subseteq \mathcal{A}$. Therefore we can inductively apply **REINF** to combine the first i terms and the $(i+1)^{\text{th}}$ term in the $\bigoplus$-summation in the definition of $\mathbf{R}_E$.

Thus, the outcome of $\mathbf{R}_E$ is

$$\bigcap_{i:\beta_{a_i}^{\mathbf{R}} - \beta_{a_{i+1}}^{\mathbf{R}} > 0} \{a_1, \ldots, a_i\} = \{a_1, \ldots, a_k\}.$$

By choice of k, the selected outcome $\{a_1, \ldots, a_k\}$ is the set of alternatives with the highest beta scores, i.e. the set of Borda winners A. □

A useful fact, following from the discussion of Young (1974), is that $\mathbf{R}_E$ and $m\mathbf{R}$ have the same beta scores.

Lemma 7 (Young 1974)**.** *For all alternatives $a \in \mathcal{A}$, $\beta_a^{\mathbf{R}_E} = \beta_a^{m\mathbf{R}}$.*

It remains to construct $\mathbf{R}_C$, and bound the length of its explanation. Lemma 7 implies that $\left(\delta^{\mathbf{R}_E} - m\delta^{\mathbf{R}}\right) \in \mathcal{K}(\hat{\beta})$, where $\mathcal{K}(\hat{\beta})$ is the kernel space of the linear map $\hat{\beta}$ defined above. Cailloux & Endriss (2016) show that the set of delta vectors of all cyclic profiles spans $\mathcal{K}(\hat{\beta})$.

Lemma 8 (Cailloux & Endriss 2016)**.** *There exists a set of m-cycles $\mathcal{S}, |\mathcal{S}| = \binom{m-1}{2}$, such that $\rho = \{\delta^{\mathbf{R}^S_{\text{cyc}}} : S \in \mathcal{S}\}$ spans $\mathcal{K}(\hat{\beta})$.*

We now have the machinery in place to prove that the profile $\mathbf{R}_C$ has the desired properties.

Lemma 9. *There exists a profile $\mathbf{R}_C$ such that $\delta^{\mathbf{R}_C} = k\left(\delta^{\mathbf{R}_E} - m\delta^{\mathbf{R}}\right)$, for some integer k, and $\mathbf{R}_C$ is the sum of cyclic profiles. Furthermore, $[\mathbf{R}_C \mapsto \mathcal{A}]$ can be explained in $O(m^2)$ steps.*

Proof. By Lemma 8, there exists a basis ρ for $\mathcal{K}(\hat{\beta})$. Let $\delta^{\mathbf{R}_i}$ be the i^{th} base vector in ρ, with $\mathbf{R}_i$ its corresponding cyclic profile ($i \in [\binom{m-1}{2}]$), where a profile $\mathbf{R}_i$ that corresponds to $\delta^{\mathbf{R}_i}$ is guaranteed to exist by Lemma 8.[5] One can therefore decompose the target delta vector as

$$\delta^{\mathbf{R}_E} - m\delta^{\mathbf{R}} = \sum_{i \in [\binom{m-1}{2}]} c_i \delta^{\mathbf{R}_i},$$

where the coefficients c_i are all rationals (since the delta vectors are integer vectors). If there is a negative c_i in this decomposition, we can substitute $\mathbf{R}_i$ by $\overline{\mathbf{R}_i}$, the profile where every voter's preference is reversed; the delta vector changes sign and therefore $c_i\delta^{\mathbf{R}_i} = -c_i\delta^{\overline{\mathbf{R}_i}}$. Thus, without loss of generality, we can assume that all c_i are non-negative.

Next, because all the coefficients are rational, there must be an integer k such that $k \cdot c_i$ is a non-negative integer for all $i \in [\binom{m-1}{2}]$. Let

$$\mathbf{R}_C = \bigoplus_{i=1}^{\binom{m-1}{2}} k \cdot c_i \mathbf{R}_i.$$

We can see that $\delta^{\mathbf{R}_C} = k\left(\delta^{\mathbf{R}_E} - m\delta^{\mathbf{R}}\right)$ as desired.

Towards bounding the length of the explanation, since the profiles $\mathbf{R}_i$ are all cyclic, we can use **CYCL** and **MULT** to show $[kc_i\mathbf{R}_i \mapsto \mathcal{A}]$, for $i \in [\binom{m-1}{2}]$. We can then apply **REINF** $O(m^2)$ times, in any order, to combine these profiles. We conclude that $\mathbf{R}_C$ can be explained in $O(m^2)$ steps. □

Theorem 1 now follows directly from Lemmas 5, 6 and 9. □

A.2. Proof of Corollary 1

We finish our proof that the $\mathcal{L}$-axiomatization in the proof is asymptotically weaker than $\mathcal{S}_{\text{Borda}}$.

For convenience, in the following proofs we use the *deduction theorem*, which can be easily proved for this system: if we have given a proof of φ_2 using φ_1 as an assumption, then the deduction theorem states that there exists a proof of $(\varphi_1 \to \varphi_2)$ (see Ben-Ari, 2012, Thm. 3.14).

Let $\mathbf{R}, \mathbf{R}'$ be profiles and let $A \subseteq \mathcal{A}$. Consider the **REINF-SUB** axiom instance $([\mathbf{R} \bigoplus \mathbf{R}' \mapsto A] \wedge [\mathbf{R}' \mapsto \mathcal{A}]) \to [\mathbf{R} \mapsto A]$. We show that this axiom instance can be proven using **ADD** axioms:

1. $[\mathbf{R} \bigoplus \mathbf{R}' \mapsto A]$ (assumption)
2. $[\mathbf{R}' \mapsto \mathcal{A}]$ (assumption)
3. For each $B \subseteq \mathcal{A}$ where $B \neq A$:
   (a) $([\mathbf{R} \mapsto B] \wedge [\mathbf{R}' \mapsto \mathcal{A}]) \to [\mathbf{R} \bigoplus \mathbf{R}' \mapsto B]$ (**ADD**)
   (b) $[\mathbf{R} \mapsto B] \to [\mathbf{R} \bigoplus \mathbf{R}' \mapsto B]$ (propositional reasoning from 2 and (a))
   (c) $\neg[\mathbf{R} \bigoplus \mathbf{R}' \mapsto A] \vee \neg[\mathbf{R} \bigoplus \mathbf{R}' \mapsto B]$ (**FUNC**)
   (d) $\neg\,[\mathbf{R} \bigoplus \mathbf{R}' \mapsto B]$ (propositional reasoning from 1 and (c))
   (e) $\neg[\mathbf{R} \mapsto B]$ (propositional reasoning from (b) and (d))
4. $\bigvee_{C \in \mathcal{P}_\emptyset(\mathcal{A})} [\mathbf{R} \mapsto C]$ (**FUNC**)
5. $[\mathbf{R} \mapsto A]$ (propositional reasoning from 3(e) and 4)
6. $([\mathbf{R} \bigoplus \mathbf{R}' \mapsto A] \wedge [\mathbf{R}' \mapsto \mathcal{A}]) \to [\mathbf{R} \mapsto A]$ (deduction theorem from 1, 2, 5)

Let $\mathbf{R}$ be a profile, let $k \in \mathbb{Z}_+$, and consider the **SIMP** axiom instance $[k\mathbf{R} \mapsto A] \to [\mathbf{R} \mapsto A]$. We prove that this axiom can be proven using the **MULT** axiom, which is easy to deduce from **ADD**.

1. $[k\mathbf{R} \mapsto A]$ (assumption)
2. For each $B \subseteq \mathcal{A}$ where $B \neq A$:
   (a) $[\mathbf{R} \mapsto B] \to [k\mathbf{R} \mapsto B]$ (**MULT**)
   (b) $\neg[k\mathbf{R} \mapsto A] \vee \neg[k\mathbf{R} \mapsto B]$ (**FUNC**)
   (c) $\neg[k\mathbf{R} \mapsto B]$ (propositional reasoning from 1 and (b))
   (d) $\neg[\mathbf{R} \mapsto B]$ (propositional reasoning from (a) and (c))

[5] The proof of Cailloux & Endriss (2016) gives an explicit construction of ρ, thus we can find the profiles $\mathbf{R}_i$ by solving a linear system.

3. $\bigvee_{C \in \mathcal{P}_\emptyset(\mathcal{A})} [\mathbf{R} \mapsto C]$ **(FUNC)**

4. $[\mathbf{R} \mapsto A]$ (propositional reasoning from 2(d) and 3)

5. $[k\mathbf{R} \mapsto A] \to [\mathbf{R} \mapsto A]$ (deduction theorem from 1 and 4)

B. Details for Plurality

B.1. An Upper Bound for Plurality

There are multiple ways of rendering Sekiguchi's (2012) proof in our formal system, where the details depend on the exact formal axioms used. Here we give an axiomatization that leads to particularly simple explanations.

We define a family of base profiles for our axiomatization, consisting of *lollipop profiles* $\mathbf{R}^A_{\text{lolli}}$, for each non-empty $A \subseteq \mathcal{A}$, which has $|A|$ voters. Write $A = \{x_1, \ldots, x_k\}$ and $\mathcal{A} \setminus A = \{y_1, \ldots, y_{m-k}\}$. The first voter has preferences $x_1 \succ x_2 \succ \cdots \succ x_k \succ y_1 \succ \cdots \succ y_{m-k}$, the second voter has preferences $x_2 \succ x_3 \succ \cdots \succ x_k \succ x_1 \succ y_1 \succ \cdots \succ y_{m-k}$, and so on. For example, the profile $\mathbf{R}^{\{a,b,c\}}_{\text{lolli}}$ for $\mathcal{A} = \{a, b, c, d, e\}$ has three votes: $a \succ b \succ c \succ d \succ e$, $b \succ c \succ a \succ d \succ e$ and $c \succ a \succ b \succ d \succ e$. Intuitively, in the profile $\mathbf{R}^A_{\text{lolli}}$, the alternatives in A are symmetric under the cyclic permutation $(x_1\, x_2\, \ldots\, x_k)$, and are all stronger than the other alternatives. Thus, a symmetric and efficient voting rule should select the alternatives in A. Note that in $\mathbf{R}^A_{\text{lolli}}$, the alternatives in A each have plurality score 1, and other alternatives have plurality score 0.

We now define our axioms.

1. **LOLLI**: For a lollipop profile $\mathbf{R}^A_{\text{lolli}}$, the set of winners should be A. Formally, $\forall k \in \mathbb{Z}_+ \left[k\mathbf{R}^A_{\text{lolli}} \mapsto A\right]$.

2. **TOPS**: If the plurality score vectors of two profiles are same, they should select the same winners. Formally, for any profiles $\mathbf{R}_1, \mathbf{R}_2$ with $\alpha^{\mathbf{R}_1} = \alpha^{\mathbf{R}_2}$, $[\mathbf{R_1} \mapsto \mathcal{A}] \to [\mathbf{R_2} \mapsto \mathcal{A}]$.

3. **REINF**: For any two profiles $\mathbf{R}_1$ and $\mathbf{R}_2$, and any two subsets of alternatives A_1 and A_2 with $A_1 \cap A_2 \neq \emptyset$, it holds that $([\mathbf{R}_1 \mapsto A_1] \wedge [\mathbf{R}_2 \mapsto A_2]) \to [\mathbf{R}_1 \bigoplus \mathbf{R}_2 \mapsto A_1 \cap A_2]$.

Let us refer to the $\mathcal{L}$-axiomatization consisting of Axioms $1-3$ listed above as $\mathcal{S}_{\text{Plu}}$.

Theorem 3. *For any profile* $\mathbf{R}$ *with* m *alternatives, the outcome of the Plurality rule can be explained in* $O(m)$ *steps assuming the* $\mathcal{L}$*-axiomatization* $\mathcal{S}_{\text{Plu}}$.

Proof. Suppose, without loss of generality, that $\alpha_{b_1} \leq \alpha_{b_2} \leq \ldots \leq \alpha_{b_m}$. Then $\mathbf{R}$ can be decomposed into subprofiles as follows:

$$\mathbf{R} = \bigoplus_{k=1}^{m} \mathbf{R}_i,$$

where $\mathbf{R}_i$ is a profile in which alternatives $b_i, b_{i+1}, \ldots, b_m$ each have plurality score $\alpha_{b_i} - \alpha_{b_{i-1}}$ and the other alternatives have plurality score 0. Write $A = \{b_i, b_{i+1}, \ldots, b_m\}$. Then $\mathbf{R}_i$ has the same plurality score vector as the profile $(\alpha_{b_i} - \alpha_{b_{i-1}})\mathbf{R}^A_{\text{lolli}}$. If $\alpha_{b_i} - \alpha_{b_{i-1}} > 0$, then by **TOPS** and **LOLLI** we have $[\mathbf{R}_i \mapsto \{b_i, b_{i+1}, \ldots, b_m\}]$. By applying **REINF** repeatedly, we then obtain $[\mathbf{R} \mapsto f_P(\mathbf{R})]$ in $O(m)$ steps. □

We can obtain other similar axiomatizations and upper bounds by replacing the **LOLLI** axiom by other axioms that imply the **LOLLI** axiom. For instance, we can use **ORB** and

- **EFF**: for every profile $\mathbf{R}$ in which each voter ranks a higher than b, $\bigvee_{A \in \mathcal{P}_\emptyset(\mathcal{A} \setminus \{b\})} [\mathbf{R} \mapsto A]$

which says that a Pareto-dominated alternative should not be elected. It is easy to check that each axiom instance of **LOLLI** can be deduced from **ORB** and **EFF** using a proof with $O(m)$ steps. Since the explanations in the proof of Theorem 3 contain $O(m)$ instances of **LOLLI**, we can thus produce an explanation of the plurality rule in $O(m^2)$ steps. Similarly, we can deduce instances of **LOLLI** by using **ORB**, **FAITH**, and **MULT** (the latter as defined in Section 3), following the arguments in Sekiguchi (2012, Lemmas 1 and 2); this again takes $O(m)$ steps per instance of **LOLLI**, giving an overall explanation length of $O(m^2)$. Applying our framework gives a lower bound of $\Omega(m)$ on the proof length both for the axiomatization based on efficiency, and for the one based on faithfulness. Thus, it is conceivable that a different strategy could give shorter explanations under these axiomatizations.

B.2. Proof of Corollary 2

First we formally define new axioms: orbit, faithfulness, equal support, and tops-only. The *plurality score* $\alpha^{\mathbf{R}}_c$ of an alternative $c \in \mathcal{A}$ in profile $\mathbf{R}$ is the number of voters in $\mathbf{R}$ who rank c in top position. For a bijection $\sigma : \mathcal{A} \to \mathcal{A}$ and a strict order $\succ \in \mathcal{A}!$, write $\sigma(\succ)$ for the strict order obtained from $\succ$ by relabeling alternatives according to σ, so that $\sigma(a)\sigma(\succ)\sigma(b)$ if and only if $a \succ b$. Given a profile $\mathbf{R}$, write $\sigma(\mathbf{R})$ for the profile with $\sigma(\mathbf{R})(\sigma(\succ)) = \mathbf{R}(\succ)$ obtained from $\mathbf{R}$ by relabeling alternatives according to σ. Then we say that a profile $\mathbf{R}$ is *invariant* under σ if $\mathbf{R} = \sigma(\mathbf{R})$.

- **ORB**: If a profile $\mathbf{R}$ is invariant under a bijection $\sigma : \mathcal{A} \to \mathcal{A}$, and $\sigma(i) = j$, we have $\bigvee_{A \in \alpha_{i,j}} [\mathbf{R} \mapsto A]$ where $\alpha_{i,j} \subseteq \mathcal{P}_\emptyset(\mathcal{A})$ is the set of outcomes such that $\{i, j\} \subseteq A$ or $\{i, j\} \cap A = \emptyset$.

- **FAITH**: If a profile $\mathbf{R}$ contains only a single voter who ranks alternative a first, we have $[\mathbf{R} \mapsto \{a\}]$.
- **EQUAL**: Let $\mathbf{R}$ be a profile in which $\alpha_a^{\mathbf{R}} \in \{0,1\}$ for all $a \in \mathcal{A}$. Then the alternatives with score 1 are elected, so we have $\left[\mathbf{R} \mapsto \{a \in \mathcal{A} : \alpha_a^{\mathbf{R}} = 1\}\right]$.
- **TOPS**: If the plurality score vectors of two profiles are same, they should select the same winners. Formally, for any profiles $\mathbf{R}_1, \mathbf{R}_2$ with $\alpha^{\mathbf{R}_1} = \alpha^{\mathbf{R}_2}$, $[\mathbf{R_1} \mapsto \mathcal{A}] \to [\mathbf{R_2} \mapsto \mathcal{A}]$.

Formally speaking, Corollary 2 claims a lower bound for the axiomatization consisting of **REINF**, **ORB**, **FAITH**, **EQUAL**, and **TOPS**. Note that the axiomatization does not contain a formal version of anonymity; that axiom is implicit in our formal setup and the definition of $\mathcal{R}$.

Proof of Corollary 2. We embed profiles into the linear space $V = \mathbb{Q}^m$, using h which maps a profile $\mathbf{R}$ to the plurality score vector $\alpha^{\mathbf{R}} = (\alpha_a^{\mathbf{R}})_{a\in\mathcal{A}}$ and $g = \arg\max$. The set S consists of all single-voter profiles and of all profiles with $\alpha_a^{\mathbf{R}} \in \{0,1\}$ for each $a \in \mathcal{A}$. The set S is finite. Further, we use predicates $T = \{t_{ij} : t_{ij}(v) = v_i - v_j\}$. With these predicates, a **PRED** instance requires that in a profile $\mathbf{R}$ in which alternatives i and j have the same plurality score, either both are winners or both are losers. Now assume that the profile $\mathbf{R}$ is invariant under the permutation σ with $\sigma(i) = j$. Then we automatically have $\alpha_i = \alpha_j$. Hence, each instance of **ORB** can be inferred from an instance of **PRED**. To calculate the sensitivity $\operatorname{sen}(g,T)$ consider for instance the size of the smallest $T' \subseteq T$ that uniquely identifies the outcome $\mathcal{A}$. Outcome $\mathcal{A}$ only occurs in profiles in which all alternatives have the same plurality score. Note that $t_{ij}^{-1}(0)$ is the set of vectors v with $v_i = v_j$. A smallest set T' such that $\bigcap_{t_i\in T'} g(t_i^{-1}(0)) = \{\mathcal{A}\}$ is $T' = \{t_{12}, t_{13}, \dots, t_{1m}\}$. Similarly one can show that at least $m-1$ predicates are required to uniquely determine any other outcome. It follows that $\operatorname{sen}(g,T) = m-1$. The axiomatization stated in the corollary is asymptotically weaker than the axiomatization derived from the embedding: **FAITH** and **EQUAL** are implied by **INIT**, **ORB** is implied by **PRED**, and **TOPS** is implied by **EMB**. □

Corollary 2 applies to the axiomatization S_{plu} used in Theorem 3, since **EQUAL** prescribes the output at any lollipop profile, so any **LOLLI** axiom instance can be deduced from **EQUAL** and **REINF**. Thus, S_{plu} is asymptotically weaker than the axiomatization in Corollary 2. Hence, explanations using S_{plu} require $\Theta(m)$ steps.

C. Details for Approval Voting

C.1. Proof of Corollary 3

Let us redefine $\mathcal{R}$ to be the set of functions $\mathbf{R} : \mathcal{P}_\emptyset(\mathcal{A}) \to \mathbb{N}$ of profiles of approval ballots; the function $\mathbf{R}$ specifies how many voters submit a given set of approved candidates. With this alternative definition, we can define notions like voting rules $f : \mathcal{R} \to \mathcal{P}_\emptyset(\mathcal{A})$ and our language $\mathcal{L}$ exactly as before. Also, everything in Sections 4.1 and 4.2, and in particular the main lower bound of Theorem 2, continues to apply with the new $\mathcal{R}$. For the distribution over $\mathcal{R}$ used in the definition of "with high probability" for Theorem 2, we can take any distribution $\mathcal{D}$ over $\mathcal{R}$ as long as $h(\operatorname{supp}(\mathcal{D}))$ spans V, for example impartial culture for approval profiles (which selects each voter's approval set i.i.d. uniformly at random).

Now let us define axioms appropriate for the approval-based setting. Given a profile $\mathbf{R}$, the *approval score* of an alternative a is the number of voters who approve a.

1. **REINF**: For any two profiles $\mathbf{R}_1$ and $\mathbf{R}_2$, and any two subsets of alternatives A_1 and A_2 with $A_1 \cap A_2 \neq \emptyset$, it holds that $([\mathbf{R}_1 \mapsto A_1] \wedge [\mathbf{R}_2 \mapsto A_2]) \to [\mathbf{R}_1 \bigoplus \mathbf{R}_2 \mapsto A_1 \cap A_2]$. (Note that this is identical to the previous definition for strict orders.)
2. **ORB**: If a profile $\mathbf{R}$ is invariant under a bijection $\sigma : \mathcal{A} \to \mathcal{A}$, and $\sigma(i) = j$, we have $\bigvee_{A\in\alpha_{i,j}} [\mathbf{R} \mapsto A]$ where $\alpha_{i,j} \subseteq \mathcal{P}_\emptyset(\mathcal{A})$ is the set of outcomes such that $\{i,j\} \subseteq A$ or $\{i,j\} \cap A = \emptyset$.
3. **FAITH-AV**: If a profile $\mathbf{R}$ contains only a single voter with approval set A, we have $[\mathbf{R} \mapsto A]$.
4. **DE**: If a profile $\mathbf{R}$ contains exactly two voters, one with approval set A and one with approval set B where $A \cap B = \emptyset$, we have $[\mathbf{R} \mapsto A \cup B]$.
5. **CANC-AV**: If in profile $\mathbf{R}$ all alternatives have the same approval score, then $[\mathbf{R} \mapsto \mathcal{A}]$.

The voting rule *Approval Voting* (AV) selects the set of alternatives with maximum approval score. To prove our lower bound, similarly to plurality, we embed AV into $V = \mathbb{Q}^m$ using h which maps a profile $\mathbf{R}$ to the vector of approval scores, and $g = \arg\max$. The set S consists of all single-voter profiles, all two-voter profiles with disjoint approval sets, and all profiles in which all alternatives have the same approval score. Then S satisfies the conditions of Theorem 2, because the first two parts are finite, and the third part maps to a one-dimensional subspace of V. For the set of predicates, we again take $T = \{t_{ij} : t_{ij}(v) = v_i - v_j\}$ with sensitivity $m-1$.

The axiomatization with axioms 1–5 above is asymptotically weaker than the axiomatization from Theorem 2: **REINF** follows from **ADD**, **ORB** follows from **PRED**, and **FAITH-AV**, **DE**, **CANC-AV** all follow from **INIT**.

To obtain an upper bound on the length of explanations for AV, one can follow the proofs by Brandl & Peters (2019).