Author Response: Explainable Voting #8353

Review 1

We discuss the impact of our results for machine learning applications in lines 28–48 of the submission. The reflections by Reviewer 4 in points 2 and 8 are a good expression of how we hope to enrich techniques for explainable ML.

Review 2

A major question raised by your review is: how much flexibility do we have in choosing our axiom systems? The answer is that we are heavily constrained. It is possible to write down trivial axiom systems such as your INIT′ example, but of course these axioms are not normatively appealing. For other axiom systems, it is necessary to establish that they *uniquely characterize* the voting rule in question, since otherwise there will be instances where the algorithm fails to find an explanation. Very few characterization results are known for voting, even for rules other than Borda. (Worse, most known characterizations rely on limit arguments, which does not lead to explanations in our sense.) Arguably, in our paper, we have captured most of the important ones in a single framework.

Given this background, our choice of axioms for the Borda rule is not arbitrary: the axiom system we use (up to minor variations) is the only one known to characterize Borda (without limit arguments), and luckily the axioms used have considerable normative appeal. While, as you note, the INIT axioms could be seen as arbitrary, in the Borda case, they seem well-motivated via simple symmetry arguments.

Regarding using the simple embedding of Borda into space $\mathbb{N}^m$ to get shorter Borda explanations: indeed this would be possible, but in our framework the EMB axiom would then require that $f(R_1) = f(R_2)$ whenever profiles R_1 and R_2 have the same Borda scores, and this seems too specific. While we don't think this specific embedding will lead to convincing explanations, the thought process you engaged in as a reader is an example of a hope we express in the discussion: "The good news is that Theorem 2 can help identify new axiomatizations that lead to short explanations." In other words, the possibility of new embeddings that would lead to new characterizations and simpler explanations is a clear strength of our general framework, not a weakness.

Regarding the definition of "asymptotically weaker": you are correct that there is a natural version of this definition for general axiom systems. However our definition specific to $\mathcal{S}_{\text{emb}}$ exploits some additional freedom we have in our specific set-up (namely, that we are allowed to use an unlimited number of ADD and EMB axioms), and we need this freedom in our application of the framework to Borda.

Regarding "trade-off theorems" and measuring the "strength" of classes of axiom systems: this is an intriguing idea, but it is not clear how to formalize the notion of a "class" of axiom systems. For the voting context, we are again limited by the small number of axiom systems (with normative appeal) that are known to characterize common voting rules.

Review 3

You mention that we do not include empirical examples. For illustration purposes, we did include (in the supplementary material) a sample explanation of Borda applied to the mayoral election in Burlington, VA. Since by their nature the generated explanations follow a common pattern, we did not think it would be instructive to give many more examples. Quantitatively speaking, we did not see sufficiently promising avenues for empirical exploration: since our bounds provably apply to almost all instances, an empirical evaluation will not reveal that shorter explanations are possible in practice. [And this prediction was confirmed by preliminary experiments we have run.]

Review 4

We were amazed by your in-depth review, full of great suggestions. Thanks for taking the time to think deeply about our paper. Responding (much too) briefly to some of your points: We will reference a logic textbook as suggested; our (overpowered) proof system generalizes standard Hilbert systems and is thus indeed complete. Yes, "satisfies" should have been "consistent with". Yes, the pairwise equivalence should receive more space in a user-facing explanation; we shortened this too much due to page limit. Agreed, we will avoid using "$\forall$" in the metalanguage. The standard homogeneity axiom is equivalent to the conjunction of MULT and SIMP; we introduced new names since we handle the two parts separately. Thanks for pointing out the resemblence to structures from compilation complexity; we will reference and think more deeply about this. We agree that the tops-only axiom for plurality is unappealing. There are some nicer characterizations of plurality using independence of Pareto-dominated alternatives, but extending our lower-bound technique to the variable-agenda setting will require more work. Your proposed axiomatization of AV is great, and we think it fits into our framework! In particular, along the lines you sketch, it should be possible to derive instances of the cancellation axiom from BASIC-TIE and COMBINE.