Replies to Reviewers

Title: Causal Learning With Occam’s Razor

Author: Oliver Schulte

Journal: Studia Logica

To the Editor:

Thank you for your work on my manuscript. I appreciate your initiative in mounting the special on-line issue. I have gone through the reviewer comments carefully as instructed. Both reviewers make references to annotated pdfs of the original submission, for typographical and stylistic comments. I could not find those attachments in the editorial manager. I have gone through the paper several times again for orthography, grammar and style. I have also addressed the content suggestions from the reviewers, see my detailed replies below. To make it simple to find changes, I have marked previous writing that is removed with strikeout (~~like this~~) and highlighted newly written passages in blue (like this).

Reviewer quotes are in blue, replies are in red.

Replies to Reviewer 1:

Thank you for your time and the positive comments.

The results are profitably applied to learning causal graphs from a stream of true conditional dependence facts, and to learning probability estimation diagrams. Clearly, there is some idealization here: conditional independence facts are not learned infallibly, but are subject to statistical error. A few sentences defending the idealization may be helpful.

The original submission addressed this and related assumptions about the data in section 3.2.1. We have expanded this discussion of this and other assumptions into a clearly marked subsection {Assumptions and Limitations}, now Section 3.4 This section also contains a number of pointers to relevant literature.

Additional comments, typographical and otherwise, are included in the attached.

Thank you. I’m sorry for the errors. I edited the passages you indicated. To make it simple to find changes, I have marked previous writing that is removed with strikeout (~~like this~~) and highlighted newly written passages in blue (like this).

Referee 2.

Referee report on Causal Learning with Occam's Razor

Thank you for your time and the constructive comments.

This is a nice paper, and I recommend publication. There are some detailed stylistic suggestions in the accompanying markup that I think would improve the paper's impact, but I leave them to the authors' judgment.

First, I should say that the paper's value lies mainly in providing, as it says, a "gentle" introduction to learning theoretic work on occam's razor, to probabilistic work on causal discovery, and to demonstrating a strong connection between the two. It takes some doing to set all of that up for a general audience of non-specialists.

The paper also breaks some new ground. Standard causal diagrams do not determine which values of one variable cause which values of another. That extra structure is represented by "probability estimation diagrams (PEDs), which delve into the deeper structure of which values of a variable cause which values of another. The paper easily extends Schulte's approach to to PEDs.

You specifically asked me to comment on the relevance of the paper to the topic of Ockham's razor. From the preceding description, it is clear that the paper is on topic.

Here are some of the main issues that emerge in the markup:

The assumed approach to simplicity covers only very tidy problems in which every world in the simpler hypothesis is a boundary point of the more complex hypothesis. But the examples considered call for nothing more, so no harm is done, given the intended role of the paper as a gentle introduction for a general audience.

The paper does not fully solve the problem of causal inference, because it resorts to the artifice of an infallible oracle for dependence and non-identity of probabilities, rather than using statistical tests. There is a close analogy, but close is no cigar— there is a great deal that has to be done just right to connect topology in the right way with statistics, and that work is currently being done (e.g., Genin at CMU). It would be best to be candid about the gap.

The original submission addressed this and related assumptions about the data in section 3.2.1. We have expanded this discussion of this and other assumptions into a clearly marked subsection {Assumptions and Limitations}, now Section 3.4 This section also contains a number of pointers to relevant literature.

Extending the preceding point, I found it puzzling that the authors consider both oracles for probabilistic independence and oracles for probabilistic dependence. Given the way the authors set up their logical learning model, the only plausible analogy to statistics is that the oracle is for dependence, since failure to detect independence at a given sample size does not imply that it will not be detected at a much larger sample size. These things are not a mere matter of convention, or of trying harder, etc. It's the way the problem of probabilistic inference is. To make the matter sound conventional turns a victory for the authors into a surrender.

We definitely agree. Please note that the model presented in this paper takes as evidence items dependencies, not independencies. Your comments seem to address our discussion of related work (not this paper), where oracles for independencies are assumed and statistical tests are used to draw independence assumptions. It does seem appropriate to mention the use of independence conclusions, because this is a wide-spared if questionable practice (e.g. the PC algorithm by Spirtes, Glymour, and Scheines uses an independence oracle). To avoid confusion between the model of this paper, based on observing dependencies only, and related work based on observing independencies, we have rewritten in two ways. 1) We now describe our dependency-based model explicitly as a contrast to the previous independence work. 2) We go into more detail about statistical testing and what conclusions one should and should not draw from it (as also requested by referee 1). These new sections are marked in blue under Section 4 Assumptions and Limitations.

Again, the authors' discussion of competing approaches like AIC and BIC is vitiated to some extent by the fact that those methods have to strike a balance between simplicity and fit, whereas the authors essentially assume that no such balance is even necessary. Of course, any tests employed to implement the assumed oracle will strike some such sort of balance, which will probably be different from both. The relevant point (which is also missing from that discussion), is that the authors' underlying argument for Ockham's razor rests on an entirely different foundation from either AIC or BIC. That is something that could very usefully have been explained to a general audience. The silence on that crucial point at the crucial place is almost deafening.

Well, AIC and BIC were discussed in two places as selecting models with fewer parameters as mind-change minimality for context-sensitive independencies does. We read your comments as saying that while the recommendations to minimize parameters are similar, the *justification* for doing so is very different. Given that you view this point as crucial, we discuss it immediately in the introduction, along with the related question of the difference between score-based causal graph learning (e.g. maximize BIC) and constraint-based causal graph learning, as in our learning-theoretic model. The new discussion section 4 elaborates on the difference to statistical model selection, and cites Genin’s work. We also added emphasize to this point in our concluding discussion in Section 9.

Assuming that the paper is a gentle introduction for a general audience, it was a tactical mistake, in my opinion, to include the fundamental motivating examples for the causal formalism only in a technical appendix. That puts the general reader in the position of accepting a lot of formalism entirely on faith that it matters for something. Also, there is no discussion of how to use a causal diagram to infer the effects of novel policies, which is the whole point of discovering them.

We see the point, so we have expanded on the causal interpretation and use of causal graphs. See blue text in Sections 3.1 and 3.3. We would rather not move the details of the appendix into the main body because:

1) We aim for a gentle introduction to applications of learning theory, not a gentle introduction to causal theory. These are available elsewhere.

2) The results for graph structure learning are a review of our predecessor paper Schulte et al. 2010. So our goal here is to summarize the main principles and results, and trust that readers who seek more detail can go to that paper. By summarizing some details, we can get faster to the new results in this paper about learning from context-sensitive dependencies.

Similarly, the main motivation for the PED formalism occurs only after the definitions and results for them have been presented. I would again put the motivational stuff first.

The motivation is representing the context-sensitive dependencies described at the beginning of Section 8. We have added a sentence explaining the strong points of PEDs, especially compared to probability estimation trees.

The main technical result is that the length of the longest refinement path to a partition in a lattice is the cardinality of the partition. The authors prove it directly by induction. But it seems to be an easy consequence of the familiar fact that finite partition lattices are graded posets https://en.wikipedia.org/wiki/Graded\_poset. A graded poset has a rank function defined directly on the poset that respects the partial order and that increments once for immediate successors. The incrementing property gives a lower bound on maximum path length, and the fact that the ranking function is defined directly on partitions, rather than on paths, provides the upper bound. In any event, including a proof of such an elementary result in a technical appendix is an affectation that distracts from the genuine merits of the paper, which are principally expository. That the underlying mathematical ideas are so easy and natural is a feature, rather than a bug, in my opinion.

Nice, thank you! We agree that connections with classic previously known theorems are a feature not a bug: Learning theory motivates focusing on inclusion depth, and inclusion depth is often given by fundamental facts about partial orderings. The corresponding proposition 5 for the inclusion depth of graph structures similarly relies on the previous Meek-Chickering theorem. We have now added a reference to the graded poset result right after Proposition 7 is stated. This should make clear that we are not claiming credit for it, and that this is an example of how inclusion depth relates to classic results. That said, for readers who want to understand for themselves why this is true, it is a lot to ask to work through the general definitions of rank, ordering relation etc. and the often very abstract proofs in textbooks. So we kept our proof in the appendix, which as you say is quite simple.

There are some little typos here and there as noted in the attached PDF markup.

Thank you. I edited the passages you indicated. To make it simple to find changes, I have marked previous writing that is removed with strikeout (~~like this~~) and highlighted newly written passages in blue (like this).