Open Problem:

Toward a Topological Foundation for Hard Learning Problems

Hanti Lin

ika@ucdavis.edu

UC DAVIS
UNIVERSITY OF CALIFORNIA

Presented on February 2, 2025 at the Data Science/Machine Learning Forum at UC Davis

To Predict or Explain?

- Machine learning textbooks still focus on learning models that can **predict**.
- Yet there is increasing attention to the task of learning models that can **explain** (Spirtes et al. 1993, Pearl 2000, Shmueli 2010):
  - giving causal explanations,
  - thereby making counterfactual predictions, which is required for policy making
- That is,
  - We first acquire data D generated from an unknown probability distribution P.
  - In light of D, we want to learn a causal model M.
  - Then we want to use M to predict what would (most likely) happen under a counterfactual probability distribution P^* **rather than** the actual one P, where P^* is the distribution that P would be changed to if we were to manipulate some variables in M in a certain way.

But Learning Causal Models Is Hard

- Reason: the problem of **non-identifiability**.
  Very often, there are two different parameter values θ and θ' in Θ such that
  - causal models M_θ and $M_{\theta'}$ are distinct, making distinct counterfactual predictions and recommending distinct policies,
  - but $P_\theta = P_{\theta'}$, making it impossible to distinguish between the two models from observational data.
- Upshot: it is impossible to achieve
  - (model selection) consistency
  - i.e., convergence in probability to the true model at **every** parameter value in Θ.
- Good News: To restore consistency, we only need to make an assumption to rule out "almost no" parameter values:
  - in the topological sense of "nowhere dense"
  - or in the measure-theoretic sense of "Lebesgue-measure zero" if the parameter space is small enough to be finite dimensional.

An Old Solution for the Hardness

- There is an old, standard solution:
  - There is an old, standard solution (Spirtes et al. 1993).
  - That is, when we have two causal models M_θ and $M_{\theta'}$ with non-identifiability $P_\theta = P_{\theta'}$, let's rule out the **more complex** model *a priori* and design a learning algorithm that sacrifice consistency for that model, using **Ockham's razor** (in jargon, making the Causal Faithfulness Assumption).
- But that raises an issue:
  - Why use Ockham's razor?
  - That is, why sacrifice consistency at the parameter values that correspond to relatively complex causal models

An New Solution for the Hardness

- Jiji Zhang and I propose a new solution (Lin & Zhang 2020).
- Think about a hierarchy of evaluative standards:
  - *High*: consistency at **every** parameter value
    (too strong to be achievable)
  - *Middle*: consistency at **almost all** parameter values + some **robustness**
  - *Low*: consistency at **almost all** parameter values
    (too weak to tell where to sacrifice consistency)
- They prove that, for learning causal Bayes nets with categorical variables without ruling out models *a priori*,
  - it is possible for a learning algorithm to achieve the middle standard,
  - any learning method achieving it **must** sacrifice consistency at the parameter values that correspond to relatively complex causal models.

"Almost all" = all but a nowhere dense set.
"Robust" = preservation of good learning performance under perturbation of parameter values.

Extension of the New Solution?

- The extension to real-valued variables might be problematic. Crux:
  - The above assumes that we can learn/test conditional independence without facing the problem of non-identifiability.
- Bad News 1
  - Shah and Peters (2020) show that it is hard to test the conditional independence between two variables X and Y given a **real-valued** variable Z, when there is no assumption on the joint probability density function over X, Y, and Z. That is, if we require that the chance of Type I error be bounded from above by a small α (which is a sort of uniform consistency over just the null hypothesis of conditional independence), then the worst-case chance of Type II error must be high, as high as $1 - \alpha$.
- Bad News 2
  - I am already able to strengthen the result: it is impossible to achieve consistency at **every** parameter value.
  - I conjecture (with high confidence) that, **even if** we assume that the joint dentist is **smooth**, it is impossible to achieve consistency at **almost all** parameter values (in a rigorous, topological sense).

Open Problem

- Think about a Hierarchy of Modes of Convergence as Evaluative Criteria:

  (1) uniform consistency (at every parameter value)
  (2) pointwise consistency at every parameter value
  (3) pointwise consistency at almost all parameter values + some robustness
  (4) pointwise consistency at almost all parameter values

- The idea is that, in each learning problem, we ought to strive for the highest achievable criterion.

- **Open Problem**: For each evaluative criterion *C*, characterize the class of learning problems in which *C* is achievable (i.e., achieved by at least one learning algorithm).
  - Progress for (1) in classification: Vapnik & Chervonenkis (1971) and Valiant (1984), known as the Fundamental Theorem of Statistical Learning
  - Progress for (2) in hypothesis testing: Dembo & Peres (1994), A Topological Criterion for Hypothesis Testing
  - I would love to have more a more comprehensive, systematic result:
    - for a variety of different tasks: classification, regression, hypothesis testing, model selection, etc.
    - for each of those modes of convergence (1)-(4), and possibly more.

Thank You!