# Learning Causal Models Online

Chen Zhao

October 2020

## 1 Introduction

Predictive models – learned from observational data not covering the complete data distribution – can rely on spurious correlations in the data for making predictions.These correlations make the models brittle and hinder generalization. One solution for achieving strong generalization is to incorporate causal structures in the models;such structures constrain learning by ignoring correlations that contradict them.However, learning these structures is a hard problem in itself. Moreover, it's not clear how to incorporate the machinery of causality with online continual learning.In this work, we take an indirect approach to discovering causal models. Instead of searching for the true causal model directly, we propose an online algorithm that continually detects and removes spurious features. Our algorithm works on the idea that the correlation of a spurious feature with a target is not constant over-time.As a result, the weight associated with that feature is constantly changing. We show that by continually removing such features, our method converges to solutions that have strong generalization. Moreover, our method combined with random search can also discover non-spurious features from raw sensory data. Finally, our work highlights that the information present in the temporal structure of the problem –destroyed by shuffling the data – is essential for detecting spurious features online.

## 2 Motivations

- Oftentimes we want to learn from one part of the data-distribution and generalize to another. This could be due to two reasons: First, the real data distribution could be so large that it is infeasible to collect data covering the complete distribution. Second, it could be hard to access parts of the distribution – such as collecting data for testing a parachute for landing a rover on Mars. These cases require an extreme form of generalization: systematic zero-shot generalization.It is unlikely that we would achieve such generalization by minimizing empirical risk on a small part of the data distribution.
- A causal structure can constrain the dependence between variables of the world, weeding out spurious correlations. Unfortunately,learning the

causal structure in the general case requires collecting data covering the complete distribution.

- Even though deep neural networks can represent a potential causal model in their parameters, training a neural network by minimizing empirical risk on a small part of the data-distribution is unlikely to recover the true causal model. However, neural networks combined with the right learning algorithms and data might be sufficient for discovering the true causal model.

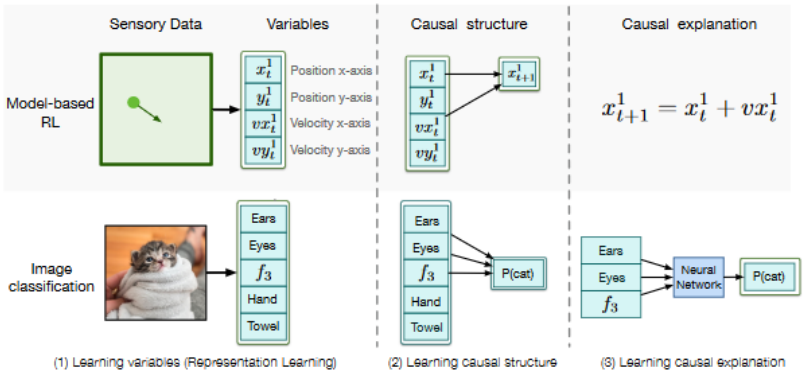


Figure 1: We look at two examples of potential causal models: predicting the position of a bouncing ball, and predicting the class of an image. For both cases, a causal model has three components: (a) extracting a set of abstract variables – features – from the raw sensory data; (b) removing the spurious variables from the list of variables; (c) capturing the exact relation between the causal variables and the target. The variables don't necessarily correspond to interpretable aspects of the world. They can represent uninterpretable abstractions, as $f_3$ for the cat image.

# 3 An online algorithm for identifying spurious features

To approximate variance online, we keep two exponentially decayed sums for each feature. First,we keep track of the running mean $u_i$ of the weight $w_i$ as the agent learns in the environment. We only update $u_i$ when $f_i^t = 1$. This is important because we only care about our estimate when a feature is active. The second metric, $v_i$, accumulates the variance of $w_i$ around the running mean $u_i$. Again, we only update $v_i$ when $f_i = 1$. The update rule of both statistics is given by:

$$u_i^t = \alpha u_i^{t-1} + (1-\alpha) w_i^t f_i^t + (1-\alpha)(1-f_i^t) u_i^{t-1}$$

$$v_i^t = \beta v_i^{t-1} + (1-\beta)(w_i^t - u_i^t)(w_i^t - u_i^{t-1}) f_i^t + (1-\beta)(1-f_i^t) v_i^{t-1}$$

where $0 < \alpha, \beta < 1$ and $\beta < \alpha$.