ON CAUSALITY IN MACHINE LEARNING

by

Charles Halberg

A Honors thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Honors Bachelor of Science

in

Mathematics

Department of Mathematics

The University of Utah

December 2022

ABSTRACT

Although the studies of causal inference and machine learning arose independently, there are emerging intersections which are proving fruitful for both fields. One of the major modern challenges in AI is to develop robust, generalizable models which can perform across a number of different tasks with minimal need for re-training and exposure to new data. Developing such models suggests a need for inference beyond prediction in settings with independent identically distributed data. Rather, the model should be able to leverage causal information in settings where interventions on a system can change the joint distribution of the data and classical statistical guarantees no longer apply. Causal inference has thus become a useful framework to ground research towards models capable of generalizing between tasks and answering questions about interventions on systems. In the other direction, much of the research in causal inference assumes knowledge of causal variables, and so machine learning is also becoming useful to causal inference for extracting high level causal information from low level data. In this paper, I begin by reviewing the foundational assumptions and concepts of causal inference, with an emphasis on those that relate to machine learning. I will cover open problems in machine learning, and how incorporating ideas from causal inference has led to progress in these areas. Then, I will discuss some machine learning approaches to answering causal questions, and perform experiments regarding causal effect estimation. Finally, I will finish with a discussion of future avenues for research in the intersection of causal inference and machine learning.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Successes and Limitations of Machine Learning

In an age of abundant data, machine learning has proven to be an essential tool in empirical science and industry applications. With powerful function approximation capabilities, high performance computing systems, and massive amounts of data for training, machine learning methods have achieved success in tasks once thought impossible for computer systems. Prominent examples include AlphaGo [30], which was able to beat the world champion in Go using deep reinforcement learning, and AlphaFold [33] which can correctly predict the three dimensional structure of a protein given only its amino acid sequence. Even without the large scale computational infrastructure required for such models, machine learning has numerous other applications like high dimensional regression, classification, and feature discovery in unstructured data.

A common thread tying the different successes of machine learning together is its unprecedented ability to learn highly non-trivial statistical dependencies in observed data. That is, when observing data from some potentially complicated joint distribution $\mathbf{X} \sim \mathbb{P}(\mathbf{X}, Y)$, where Y could be labels of the data, supervised learning methods can very successfully approximate the conditional likelihood $\mathbb{P}(Y|\mathbf{X})$. This success is in part due to the underlying assumption that the observed data is independently identically distributed (iid) from some distribution $\mathbb{P}(\mathbf{X}, Y)$. This assumption is in many cases appropriate, and so powerful statistical guarantees, such as the central limit theorem and maximum likelihood estimation, can be leveraged.

However, statistical models are only valid in the iid setting, outside of which they may fail in variety of ways. As a toy example, consider a model that attempts to predict the ratings of a movie from a labeled dataset consisting of ticket sales, X, and critic ratings, Y.

One may see a strong positive correlation between X and Y, and so it could be reasonable to build a statistical model of $\mathbb{P}(Y|X)$. However, it is clear that large audiences do not *cause* high ratings from expert critics, and so if more people went to see a movie because of a sale on movie tickets or an increase in advertising then the model would incorrectly interpret this increase in X as necessitating an increase in Y. Additionally, it may be the case that the reverse is true, where ratings at least partially cause ticket sales, in which case a predictive model of ticket sales given ratings, $\mathbb{P}(Y|X)$, would be more appropriate. From this data alone, the statistical relationship between X and Y does not give any information about which direction the arrow of causation points. In the case of only two variables, such a distinction cannot be made without prior knowledge of the context or data from experimental settings [8]. Without having a way of incorporating this information, statistical models fail to predict in situations involving a change in the underlying distribution.

More generally, this example highlights the failure of statistical models to perform inference outside of the regular iid setting. An interesting real world example of this is the phenomena of ”adversarial vulnerability” in neural networks [32]. It was empirically shown that for many baseline deep classification models, adversarial examples generated by a carefully crafted small perturbation in training data can cause the model to misclassify the example, despite the fact that the perturbation is essentially imperceptible to humans in the case of image classification [32]. So by cleverly shifting the distribution from which a training example is generated, even sophisticated deep learning models fail to predict accurately. In high-stakes situations such as relying on Artificial Intelligence (AI) systems in autonomous vehicles to detect stop signs and pedestrians, the appearance of an adversarial example would be disastrous. The conditions on which a model is trained can also change naturally over time, such as in the case of consumer sentiment, and so distribution shifts of this kind also pose a challenge for building robust, trustworthy, and generalizable AI systems. To overcome these limitations, machine learning models must be able to work outside of the commonly assumed, but often inappropriate, setting of independent identically distributed data. A key missing ingredient is incorporating ideas about causation beyond statistical relationships.

1.2 Epistemic Barriers in Causal Inference

In the study of causal inference, there are distinct 'layers' of inference which correspond to different types of queries about the environment a learner is capable of answering. Formally defined in the Pearl Causal Hierarchy [24], these tiers correspond to associational, interventional, and counterfactual settings. These tiers are colloquially thought of as "seeing", "doing", and "imagining," respectively.

The first layer, "associational," corresponds to settings in which the learner observes data $\mathcal{X}$ from known variables $\mathbf{X}$, but has no knowledge of the underlying data generating system beyond the statistical relationships found between variables [1]. This level of inference corresponds to classical statistical methods, where data is observed from some system, and inference about relationships $\mathbb{P}(Y|X_1, \ldots, X_n)$ is performed, where $Y, X_1, \ldots, X_n \in \mathbf{X}$. Recalling the movie ratings example, one observational question might be: "How well do ticket sales explain the variance in movie ratings?" As alluded to above, information in this setting is not sufficient for answering queries about the system when changes to the experimental setting occur.

The second layer, "interventional," corresponds to settings where the learner can perform interventions on a system and observe the outcomes. This is typically done by fixing a subset of variables $\mathbf{X}_0 \subset \mathbf{X}$ to be constants, allowing the learner to observe outcomes from the resulting interventional distribution [1]. Here genuinely more causal information can be gleaned from the environment based on the interventions performed since doing so changes the underlying data generating mechanisms. An interventional question one might ask is: "How will ticket sales change if ratings increase ten points?" Despite the ability to perform experiments in real world settings, there are still causal queries which experimentation alone is not capable of answering.

The third and final layer in the Pearl Causal Hierarchy is the "counterfactual" layer, in which a learner reasons about how the environment would have been different had counter-to-fact interventions occurred. Since in a real world experiment, only one outcome Y can be observed, it is reasonable to consider what other outcomes (termed "counterfactuals") would have been under different interventions. One example of a counterfactual question is: "If ratings had not increased 10 points, would ticket sales still have remained the same?" Counterfactual inference involves thinking retrospectively about what the data generating mechanism underlying observations would have looked like after performing alternative interventions. It is in this tier of causal inference where the rigorous study of causal models has a major advantage, since no observational or experimental data alone can answer counterfactual queries.

Expressions in the observational layer give information about the joint distribution of a set of variables $\mathbb{P}(\mathbf{Y} = \mathbf{y})$, and expressions in the interventional layer give information about conditional expressions of the joint distribution $\mathbb{P}(\mathbf{Y}_\mathbf{x} = \mathbf{y})$, where $\mathbf{Y}_\mathbf{x}$ denotes outcomes having performed intervention $\mathbf{x}$. Then expressions in the counterfactual layer allow for conjunctions of conditional (interventional layer) expressions $\mathbb{P}(\mathbf{Y}_\mathbf{x} = \mathbf{y}, \dots, \mathbf{Z}_\mathbf{w} = \mathbf{z})$. Informally stated, the Causal Hierarchy Theorem says that layers are separated in the measure theoretic sense, such that for almost all possible structural causal models induced by a set of observable variables, the layers remain distinct [1]. The proof involves defining symbolic languages $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ which correspond to increasingly complex probability statements available in each layer. For a given hierarchy, the layers remain distinct if there are statements in $\mathcal{L}_j$ which cannot be implied by statements in $\mathcal{L}_i$ whenever $i < j$. The full proof of this is complex and notationally dense, so for a precise description see [24]. These layers form a hierarchy in the sense that the layer above contains genuinely richer information about the causal structure of a system, which means that the knowledge required to perform inference at each layer is greater than in the layers preceding it. The Causal Hierarchy Theorem makes it explicit that statistical associations are not sufficient for answering arbitrary interventional questions, and experimental data is not sufficient for answering arbitrary counterfactual questions. The importance of recognizing these distinctions when considering causal inference in empirical science cannot be understated.

1.3 Causal Thinking in Machine Learning

Initially, the causal hierarchy theorem may make causal inference from data seem like a hopeless endeavour. Without incorporating prior information about the data collection method, this is in fact true for interventional and counterfactual questions. However, by exploiting knowledge of the interventions involved in data collection, inference at interventional and even counterfactual levels may be possible. For example, some data may be generated under known and unknown interventions, such as in large scale electronic health data, in which case the data may help to answer interventional level questions instead of simply observational ones. Having prior information about the causal meaning of data (called ”structured” data) or the setting in which it was generated allows for higher levels of inference [29]. Additionally, data generated by known interventions can allow for better estimation of counterfactual quantities such as the impact of a treatment on an outcome of interest [27]. In particular, reinforcement learning has seen a benefit from incorporating counterfactual thinking since at an intuitive level retrospectively thinking about what would have been optimal in the past can allow an agent to make better decisions in the future [2].

Machine learning excels at learning highly complicated relationships from data, and so ML methods may help ground causal inference in practical applications by learning high level causal mechanisms from low level data. Often in causal inference, at least partial knowledge of the causal structure is assumed. Clearly this assumption is unrealistic for most applications, so learning an approximation of the structure from data is an important area of research for bringing causal inference into practical applications. From observational data, equivalence classes of causal diagrams can be learned based on satisfying the statistical relationships between variables. From these classes, identifying a specific structure is more difficult, but several approaches have been introduced for doing so [14][1]. However in this paper, emphasis will be place on methods for finding approximate causal representations which are useful for problems in machine learning.

CHAPTER 2

FUNDAMENTAL CONCEPTS IN CAUSAL INFERENCE

In this section, the major concepts of causal inference are discussed, with an emphasis on those that appear in the budding intersection with machine learning. Although causal inference has not been studied with mathematical rigour until relatively recently, there is a strong body of literature which formalizes previously qualitative notions of causation in a way that allows for quantitative inference in both natural and surprising ways. Structural causal models are the language through which most causal inference literature is expressed, and it allows ideas from the Pearl Causal Hierarchy to become tractable. Structural causal models and other important concepts are covered briefly in this section, and for a comprehensive resource see [24].

2.1 Connecting Statistics and Causal Inference

It has long been known that statistical associations alone do not effectively describe causal relationships. However, it was first argued that any statistical association is the result of an underlying causal structure by Reichenbach in 1956 [25]. This is articulated in the following assumption.

Common Cause Principle: [25]

> If X and Y are two observable variables such that $\mathbb{P}(X, Y) \neq \mathbb{P}(X)\mathbb{P}(Y)$, then there exists a Z such that $\mathbb{P}(X, Y) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)$.

In other words, the common cause principle states that if two observables are statistically dependent, then there exists an unobserved Z which causally influences both and explains their dependence. This also shows that conditional independence relations are a key idea in identifying causal relationships. Indeed, a major advance in the understanding of causality came from the axiomatic description of independence relations $I(X, Z, Y)$, read ″X is independent of Y, given Z,″ and showing that many other descriptions of conditional independence outside of probability statements satisfy these axioms [9]. Most notably, Geiger and Pearl defined the following conditional independence relations: [9]

$$I(X,Z,Y)_P \iff \mathbb{P}(X,Y|Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)$$

$$I(X,Z,Y)_G \iff Z \text{ separates } X \text{ from } Y \text{ in the undirected graph } G.$$

Although $I(X,Z,Y)_P$ defines a quantitative, probabilistic interpretation of conditional independence and $I(X,Z,Y)_G$ defines a qualitative conditional independence relation using the language of graphs, there is a powerful connection between the two given by the following theorem.

Theorem 1: [8]

For every undirected graph G, there exists a non-extreme distribution P, and a 1-1 correspondence between the variables in P and the nodes of G such that for every three disjoint sets of nodes X, Y, and Z the following holds: $I(X,Z,Y)_G \iff I(X,Z,Y)_P$.

This theorem justifies the use of undirected graphs for representing probabilistic dependencies such that dependencies in the graph actually correspond with the dependencies in the variables. This idea can be extended to directed graphs for denoting direct causation using the notion of d-separation to determine conditional independence relations on directed graphs [1]. However, since statistical associations alone are not sufficient for representing causation, further conditions are needed for creating a truly causal model.

2.2 Structural Causal Models

The one to one correspondence between graphical models and non-extreme distributions given by independence relations provided the logical foundation for representing distributions of data with graphs. Representing causal relationships using graphs allows for genuinely more structural information to be represented than in statistical associations by explicitly encoding the direction of causality using directed edges. In a directed acyclic graph (DAG), a directed edge $A \to B$ denotes direct causation, where A is the cause and B is the effect. However, simply denoting this relationship is still not enough to represent the data generating process, since it only qualitatively describes causal relationships between observed variables. In addition, there should be some quantitative mechanism on this structure representing the process assigning value to each variable. Qualitative causal graphical models and quantitative value assignment mechanisms are combined in the foundational concept of structural causal models.

Definition: Structural Causal Model (SCM) [23]

A SCM is a 4-tuple $\mathcal{M} = \langle \mathbf{X}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$ where,
$\mathbf{X} = \{X_1, \ldots, X_n\}$ is a set of observable, or endogenous, variables determined by other variables in the model;
$\mathbf{U}$ is the set of background, or exogenous, variables which are determined by factors outside the model, and $|\mathbf{U}| = |\mathbf{X}|$;
$\mathcal{F} = \{f_1, \ldots, f_n\}$ is the set of deterministic functions $f_i \colon (\mathbf{U} \cup \mathbf{X} \setminus \{X_i\}) \to X_i$ assigning values to the observable variables X_i such that

$$X_i = f_i(\mathbf{Pa}_i, U_i),$$

where $\mathbf{Pa}_i$ denotes the parents of X_i in the directed acyclic graph (DAG);
$\mathbb{P}(\mathbf{U})$ is a joint probability distribution on the exogenous variables $\mathbf{U}$, effectively describing the state of the world outside the model.

The assignment $X_i = f_i(\mathbf{Pa}_i, U_i)$ indicates that each observable X_i is a deterministic function of the observable variables which directly affect it, $\mathbf{Pa}_i$, and an unexplained noise variable, U_i, encompassing randomness not captured by the model. These assignment processes $\mathcal{F}$ are often referred to as the *causal mechanisms* which comprise the data generating process that results in the data we may observe. Thus, much like in statistics, causal inference is seen as making queries about these mechanisms. It is said that an SCM is *causally sufficient* if all exogenous and endogenous variables required by the causal mechanisms $\mathcal{F}$ are present in the model [1].

Using the language of SCMs, inference is done by evaluating distributions of outcomes in the SCM under various circumstances. In general, a valuation of a joint distribution of observables $\mathbf{Y} \subset \mathbf{X}$ using an SCM $\mathcal{M}$ involves obtaining outcomes $\mathbf{y} = \mathbf{Y}$ as determined by the causal mechanisms, then accumulating the probability of observing such an outcome given possible instances of the unobserved $\mathbf{U}$ using their distribution $\mathbb{P}(\mathbf{U} = \mathbf{u})$. At the associational level, this manifests in the following way.

Definition: Layer 1 Valuation [23]

An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{X}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$ induces a joint probability distribution $\mathbb{P}^{\mathcal{M}}(\mathbf{y})$ such that for any subset of observables $\mathbf{Y} \subset \mathbf{X}$,

$$\mathbb{P}^{\mathcal{M}}(\mathbf{y}) = \sum_{\{\mathbf{u} \colon \mathbf{Y}(\mathbf{u}) = \mathbf{y}\}} \mathbb{P}(\mathbf{u}),$$

where $\mathbf{Y}(\mathbf{u})$ is the solution for outcomes of $\mathbf{Y}$ given by the causal mechanisms $\mathcal{F}$ when $\mathbf{U} = \mathbf{u}$.

So on an associational level, the causal mechanisms serve as a mapping from the distribution of the external state $\mathbb{P}(\mathbf{U})$ to the joint distribution induced by the SCM, $\mathbb{P}^{\mathcal{M}}(\mathbf{y})$. Interventions can be modelled as altering a subset of the causal mechanisms by (a) changing the distribution $\mathbb{P}(\mathbf{U})$, (b) fixing some f_i's to be constant, or (c) partially changing some f_i's to alter the dependence on causal variables $\mathbf{Pa}_i$ [29]. In any of these cases, an intervention $\mathbf{x}$ on $\mathcal{M}$ induces new causal mechanisms $\mathcal{F}_{\mathbf{x}}$ which determine the dynamics of the "interventional" SCM $\mathcal{M}_{\mathbf{x}}$. In particular, given a subset of observables $\mathbf{V} \subset \mathbf{X}$, the causal mechanisms resulting from fixing the values of the $\mathbf{V}$ at constants $\mathbf{v}$ are given by

$$\mathcal{F}_{\mathbf{v}} = \{f_i \colon X_i \notin \mathbf{V}\} \cup \{\mathbf{V} \leftarrow \mathbf{v}\},$$

where "$\leftarrow$" indicates the operation of reassigning the random variables $\mathbf{V}$ as deterministic constants $\mathbf{v}$ [1]. This type of intervention is also called the "do" operator, denoted $\mathrm{do}(\mathbf{x})$ [24]. From this, the interventional joint distribution induced by a SCM follows naturally.

Definition: Layer 2 Valuation [23]

Given an SCM $\mathcal{M}$ and intervention $\mathbf{x}$, the interventional joint distribution for any $\mathbf{Y} \subset \mathbf{X}$ is given by,

$$\mathbb{P}_{\mathbf{x}}(\mathbf{Y}) = \sum_{\{\mathbf{u} \colon \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}\}} \mathbb{P}(\mathbf{u}),$$

where $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is the solution for outcomes of $\mathbf{Y}$ given by the causal mechanisms $\mathcal{F}_{\mathbf{x}}$ induced by $\mathbf{x}$ when $\mathbf{U} = \mathbf{u}$ [23].

The intervention $\mathbf{x}$ essentially creates a new mapping $\mathcal{F}_{\mathbf{x}}$ from the distribution of the environment $\mathbb{P}(\mathbf{U})$ to the observed joint distribution after performing the intervention, $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$. In a counterfactual query, we are interested in comparing situations in which more than one intervention occurred simultaneously. Although this scenario is impossible in the real world, formulating such a distribution is useful for inference and follows easily from the interventional case.

Definition: Layer 3 Valuation [23]

An SCM $\mathcal{M}$ induces a family of joint distributions over counterfactual events $\mathbf{Y}_{\mathbf{x}}, \ldots, \mathbf{Z}_{\mathbf{w}}$ for any subsets $\mathbf{Y}, \ldots, \mathbf{Z} \subset \mathbf{X}$ and interventions $\mathbf{x}, \ldots, \mathbf{w}$ such that

$$\mathbb{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}) = \sum_{\{\mathbf{u} \colon \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}, \ldots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u}) = \mathbf{z}\}} \mathbb{P}(\mathbf{u}),$$

where $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}), \ldots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u})$ are the solutions for outcomes of $\mathbf{Y}, \ldots, \mathbf{Z}$ given by the causal mechanisms $\mathcal{F}_{\mathbf{x}}, \ldots, \mathcal{F}_{\mathbf{w}}$ induced by interventions $\mathbf{x}, \ldots, \mathbf{w}$ when $\mathbf{U} = \mathbf{u}$.

The above joint distribution describes different ”worlds” in which a different intervention was performed in each. Since only one intervention can be performed at a single moment in the real world, it is impossible to gather ground truth counterfactual data through experimentation. However, the advantage of this framework is that if the causal structure and mechanisms are known, then counterfactual inference is possible using the induced distributions. Obviously, this requires a huge amount of prior information about the underlying SCM, which puts major constraints on the current practical applications of causal inference. To deal with this constraint, machine learning methods for approximating the SCM from data have been paramount in applying causal inference to practical problems.

2.3 Independent Causal Mechanism Principle

An interesting property of structural causal models arises when we apply the common cause principle to the exogenous variables **U** and assume causal sufficiency. We assume the set **U** to be jointly independent since otherwise there should exist at least one other U_i not included in the model which explains the dependence, but because we assume the SCM is causally sufficient there is no such U_i. Because of the independence of the set **U**, the joint distribution $\mathbb{P}(\mathbf{X})$ satisfies the *causal Markov condition*, which states that conditioned on its parents in the DAG, each X_i is independent of its non-descendants [29]. Joint independence of noises **U** results in a unique decomposition of the joint distribution into a product of its causal mechanisms, referred to as the *causal (or disentangled) factorization*: [29]

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i | \mathbf{Pa}_i).$$

The disentangled factorization describes how observables are conditionally independent given their parents, and the notion of independent causal factorization can be extended to more general interventional distributions. A more general idea is that underlying causal systems should often have mechanisms that are invariant when other mechanisms are subject to distribution shifts, and that these mechanisms behave independently of each other. These ideas are summarized in the following assumption.

Independent Causal Mechanisms (ICM): [24][29]

The causal generative process of a system's variables is composed of autonomous modules that do not inform of influence each other. In the probabilistic case, this means:
(a) performing an intervention on one mechanism $\mathbb{P}(X_i, \mathbf{Pa}_i)$ does not change any other mechanism $\mathbb{P}(X_j, \mathbf{Pa}_j)$, and
(b) knowing other mechanisms $\mathbb{P}(X_i, \mathbf{Pa}_i)$ does not inform $\mathbb{P}(X_j, \mathbf{Pa}_j)$, $(i \neq j)$.

Rigorously describing the ICM assumption is beyond the scope of this paper, but [15] provides a detailed formulation based on differences in algorithmic mutual information. In the setting of structural causal models, statement (a) of the ICM principle suggests a nice property of causal mechanisms in that the effect of interventions should be localized in the mechanism it acts on. This leads to the following hypothesis for interventions on systems.

Sparse Mechanism Shift (SMS) hypothesis: [24]

Small distribution changes tend to manifest themselves in a sparse or local way in the causal (or disentangled) factorization.

Again a rigorous description is beyond the scope of this paper, but it follows from the formulation of the ICM assumption in [15], and an extensive discussion about both the SMS hypothesis and ICM assumption can be found in [24]. To see why the SMS hypothesis has implications in practical applications of causal inference, consider a possible non-causal factorization of the joint distribution:

$$\mathbb{P}(\mathbf{X}) = \prod_{i=1}^{n} \mathbb{P}(X_i | \mathbf{Y} \subset \mathbf{X} \setminus \{X_i\}).$$

Each factor will include variables which are not causes of X_i, so we expect that an intervention in one of the true causal mechanisms $\mathbb{P}(X_j | \mathbf{Pa}_j)$ would result in a change in most, if not all, of the non-causal factors in this decomposition. In contrast, under the causal factorization the impact of interventions on a system will be restricted to the mechanism(s) which it directly affects, and so the impact of the intervention will be much clearer. In light of this, the SMS hypothesis also serves as a measure of how well the causal structure is represented by the model, where good representations will result in small changes under interventions. Additionally, incorporating independent causal mechanisms can help with transferring between tasks with similar or identical underlying causal structure since the interventional distribution can be evaluated using these mechanisms. Thus, reliably learning independent causal mechanisms is critical for machine learning applications outside of the iid setting. There is already evidence that doing so improves robustness, generalizability, and reliability through causal reasoning [19][28][2].

CHAPTER 3

APPLICATIONS OF CAUSAL INFERENCE IN MACHINE LEARNING

Although statistical machine learning is successful in the iid setting, performing in interventional or out of distribution settings requires the use of causal reasoning. In causal machine learning, the data generating mechanisms are formalized as structural causal models. As a result of this assumption, the ICM and SMS principles provide practical criteria for enforcing causal representation in machine learning models.

3.1 Robustness and Generalization

Intuitively, models composed of independent causal components should be more robust when one of those mechanisms is affected by an intervention. Motivated by the ICM assumption, an unsupervised method considers learning independent causal mechanisms in a task of classifying examples that underwent unknown transformations from a baseline sample. The goal is to learn independent mappings from the reference distribution to different interventional distributions, then use these mechanisms to serve as preprocessors for a standard classifier trained in the iid reference setting [22]. The data used in training is a dataset $\mathcal{X}_I$, which is drawn from a mixture of interventional distributions $P_I = I_1, \dots, I_N$ generated by N distinct, *a priori* unknown interventions on the reference set, and a dataset $\mathcal{X}_P$ sampled from the empirical distribution P of the reference set. In the learning setting, M parametric functions $E_1(\theta_1), \dots, E_M(\theta_M)$ called experts ″compete″ for examples $x \in \mathcal{X}_I$, and since the number of true causal mechanisms is unknown it is likely $M \neq N$. An expert E_{i^*} ″wins″ an example x' in the dataset if

$$i^* = \max_{j \in \{1,\dots,M\}} c(E_j(x)),$$

where c is an objective function which takes high-values in the support of the empirical distribution P and low values outside [22]. If an expert wins, its parameters are updated to maximize $c(E_j(x'))$ and all others are unchanged. The idea is that only updating the best function for a specific data point will promote specialization to learning that mapping from the interventional distribution to the reference distribution. Thus the optimization problem is as follows: [22]

$$\theta_1^* \ldots, \theta_M^* = \max_{\theta_1 \ldots, \theta_M} \mathbb{E}_{x \sim P_I} \left[\max_{j \in \{1, \ldots, M\}} c(E_j(x)) \right].$$

Typically the experts $E_1(\theta_1), \ldots, E_M(\theta_M)$ and c will be neural networks which can be trained using efficient optimizers. Since this is an unsupervised method, convergence criteria can be specified as halting after a fixed number of iterations or until examples are not reassigned to a new expert for a fixed number of iterations in a row.

The approach described above was successful in an image classification task in which images from the benchmark MNIST dataset underwent transformations such as noise addition, translations, and color inversion. In most cases, each expert specialized in learning a single transformation. Additionally, a baseline MNIST classifier had roughly 40% accuracy on the transformed dataset, but this quickly approached performance on the original dataset ($\approx$ 99%) when given the output of the experts after only being trained on a third of the total interventional dataset [22]. The trained experts were also able to achieve similar success on an entirely new dataset consisting of letters from different alphabets transformed by the same unknown functions [22]. By specifically structuring the learning problem around causal mechanisms, the model was able to identify the interventions changing the data and generalize this outside of the domain it was trained on. This approach was clearly successful in learning interventions on a baseline sample, but it does not necessarily correspond to a true underlying causal structure. Regardless of whether underlying causal mechanisms were discovered or not, structuring the learning problem around the ICM and SMS principles boosted generalizability in interventional settings.

3.2 Anticausal Learning

An interesting result from the theory of causal inference is that some learning problems can be made easier when predicting the cause from the effect, rather than effect from cause. Prediction in the direction opposite the data generating process is referred to as *anticausal* [28]. Machine learning often works in the causal direction, in which labels (effects) are predicted from examples (causes). Recall the example of ”adversarial vulnerability" in neural networks for classifying images [32]. The models in this case predict in the causal direction, since images X are thought to cause the label Y. Assuming the ICM principle and the relationship $X \to Y$, the joint distribution admits the factorization $\mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y|X)$, where $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$ are autonomous mechanisms. This means $\mathbb{P}(Y|X)$ is the distribution induced by the underlying causal mechanism which represents the classifier, and this mechanism remains the same for any input distribution $\mathbb{P}(X)$ [28]. More precisely, Let $\mathbf{X}$ be a set of images with associated labels $\mathbf{Y}$, and let $\mathbf{X}'$ be another set of unlabeled images that are generated from outside the support of $\mathbb{P}(X)$, with the joint distribution denoted as $\mathbb{P}'(\mathbf{X})$. Then given training points from the joint $\mathbb{P}(X, Y)$ and out of distribution examples from $\mathbb{P}'(X)$, we wish to estimate the out of distribution classifier $\mathbb{P}'(Y|X)$. From the causal decomposition it follows that

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)}.$$

By the ICM, the mechanism $\mathbb{P}(Y|X)$ remains the same for all input distributions $\mathbb{P}(X)$, so intervening on $\mathbb{P}(X)$ does not change $\mathbb{P}(Y|X)$. This results in the identity

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}'(X, Y)}{\mathbb{P}'(X)} = \mathbb{P}'(Y|X) \quad [28].$$

In this light, it is unsurprising that appropriately modified examples drawn from outside of the distribution of training data are able to 'deceive' the classifier since the classifier receives no additional information to make adjustments. One fix is to add the examples $\mathbf{X}'$ to the training data and update the model of $\mathbb{P}(Y|X)$ to fit the new training data better [32]. However, retraining in this way is not always possible and does not solve the root problem of robustness outside of the support of the training data.

If we approach the problem from the anticausal direction by generating images from labels, the impact of adversarial examples should be greatly reduced. The factorization in the anticausal direction, $\mathbb{P}(X, Y) = \mathbb{P}(X|Y)\mathbb{P}(X)$, is not a causal decomposition, meaning the generating mechanism inducing $\mathbb{P}(X|Y)$ is not necessarily independent of $\mathbb{P}(Y)$ whenever the causal relationship is $X \to Y$ [17]. If $\mathbb{P}(X)$ is intervened on to produce $\mathbb{P}'(X)$, then by Bayes' rule we see

$$\mathbb{P}'(Y|X) = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}'(X)}.$$

So by shifting $\mathbb{P}(X)$, the mechanism for determining labels could be changed resulting in $\mathbb{P}'(Y)$, the generative mechanism could be changed to $\mathbb{P}'(X|Y)$, or both occur. In any of these cases, it is clear that estimating the generative process $\mathbb{P}'(X|Y)$ is useful in coming up with a robust classifier $\mathbb{P}'(Y|X)$.

Assuming the conditional $\mathbb{P}'(X|Y)$ is invertible and injective, [28] provides methods for estimating $\mathbb{P}'(Y|X)$ which are guaranteed to come up with an informative estimate in the sense that $\mathbb{P}'(Y|X) \neq \mathbb{P}(Y|X)$. Another approach with more relaxed assumptions uses and encoder-decoder architecture to model the generative process and reconstruct the input before using it as input into a classifier [11]. In both of these approaches, considering the anticausal direction by modeling the generative process is a key idea which leads to both theoretical and empirical improvements in robustness.

There is an interesting connection in how modeling $\mathbb{P}'(Y|X)$ to improve robustness is similar to the idea of learning interventions on a reference set as in [22]. In [22], the goal was to learn invariant functions which represented possible transformations of the base data so that these mappings could be used predict in interventional settings. Similarly, a central goal of anticausal learning is to discover a generative mechanism $\mathbb{P}(X|Y)$ which is invariant across possible distributions of inputs $\mathbb{P}(Y)$. In each of these approaches, learning an invariant generative mechanism is central to performance on out of distribution data. With some assumptions on the generative process, [28] showed anticausal learning to be theoretically more robust to out of distribution examples, and although not explicitly, [22] showed how the ideas of anticausal learning are useful in practical learning tasks.

3.3 Reinforcement Learning

Reinforcement Learning (RL) is an interesting field for applications of causal reasoning since data from every level of inference may be available to the agent. For example, an RL algorithm for playing chess may have access to data from multiple scenarios in which an agent chose different actions at a given instant. This type of data is impossible to obtain in the real world, but in a virtual environment we may observe multiple outcomes by taking different actions at the same time using identical RL agents. Because of this, the use of counterfactual information in RL algorithms has been used to facilitate solutions to difficult tasks.

One such approach incorporates counterfactual information into policy search by interpreting Partially Observable Markov Decision Processes (POMDPs) as structural causal models. In the paradigm of reinforcement learning, the problem setup is formulated with an agent, state space, immediate reward function, and state transition function [31]. Time is discrete, where at each time step an agent exists in a state and takes an action. For each state, there is an associated reward assigned by some function, and an agent accumulates rewards by taking actions and traversing the state space. This results in a sequence of states, actions, and rewards:

$$s_1, a_1, r_1, s_2, a_2, r_2, \ldots.$$

This can either continue indefinitely or stop when an agent reaches some state in the state space. The sequence of actions taken by an agent is referred to as a *policy*, often denoted π. The goal of reinforcement learning is to find a policy which accumulates as much reward as possible. Also because transitions between states are typically probabilistic, the goal is to find an optimal policy which maximizes the expected reward [31]. The simplest setting to consider is a Markov decision process.

Definition: Markov Decision Process [31]

A Markov Decision Process (MDP) is a 4-tuple $\langle \mathbf{S}, \mathbf{A}, R_a, P_a \rangle$ where,
$\mathbf{S}$ is the set of possible states in the environment;
$\mathbf{A}$ is the set of actions the agent can take;
$R_a(s, s')$ is the immediate reward function with discount parameter γ when transitioning from s to s' as a result of action a;
$P_a(s, s') = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a)$ is the probability of the agent transitioning from state s to s' given they took action a.

This is the simplest setting because the probability transitions under an action $a \in \mathbf{A}$ are known, so an optimal policy can be directly computed [31]. However assuming each of these are known is often too unrealistic since in most applications an agent will not have complete prior information about the dynamics of its environment, in the form of known transitions between states, or about the set of states itself. Thus Partially Observable Markov Decision Processes (POMDPs) are more fitting for practical applications. In a POMDP a finite state space is still assumed to exist, but the agent only has access to a set of observations generated by the states [31]. So instead of directly observing the current state, the state gives an observation which provides a ”hint” about what state the agent is in. The observations can be probabilistic, so an observation function also needs to be specified. This observation function simply gives the probability of each observation for each state in the model, and it can also be dependent on the action [31].

Causal Inference has an interesting application with reinforcement learning in that given a policy π, any POMDP can be represented by a SCM [2]. This is done by considering transition kernels $\mathbb{P}(S_{t+1}|S_t, A_t)$ to be deterministic functions with independent noise variables $\mathbf{U}$ such that $S_{t+1} = f_{st}(S_t, A_t, U_{st})$ [2]. Then, choosing a different policy, μ, is seen as an intervention where the functions $A_t = f_\pi(H_t, U_{at})$ are replaced by $A_t = f_\mu(H_t, U_{at})$ [2]. For each policy, the resulting SCM induces a distribution over trajectories (a_i, s_i, r_i). Figure 3.1 shows the DAG interpretation of a POMDP given a policy π.

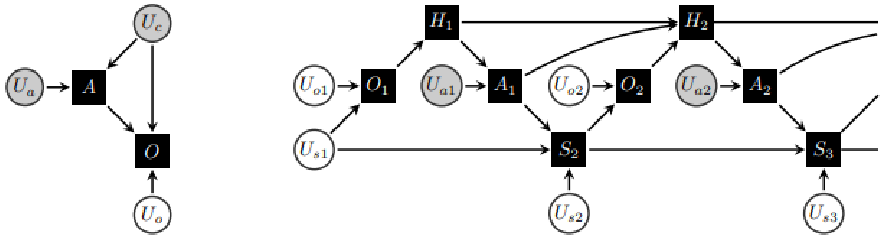

Figure 3.1. POMDP interpreted as a SCM [2]

For any scenario U_{s_t} there are many actions $A_t = (a_1, a_2, \ldots)$ which an agent may take, and so we wish to reason about the possible outcomes resulting from different actions taken under the same scenario. The advantage of formulating POMDPs in this way is that is allows for unbiased estimation of interventional distributions:

$$\mathbb{E}_{\hat{x}_o \sim p}\left[p^{do(I)|\hat{x}_o}(x)\right] = p^{do(I)}(x),$$

where $\hat{x}$ are observations from the density induced by the SCM [2]. So unbiased samples can be obtained from any interventional distribution by sampling scenarios U_i from the prior and inferring the rest from data $\hat{x}_o$ [2].

In off-policy policy evaluation, the expected return of a policy is determined without running the policy itself, and instead by inferring from simulation of the prior model $\mathcal{M}$ [31]. In model based policy evaluation, any bias in the model $\mathcal{M}$ will propagate to the estimate of expected return since the simulation of the policy is completely dependent on the prior of the model. However in counterfactual policy evaluation, scenarios from the posterior are used whenever available and the rest are simulated from the prior, thus leading to reduced bias [2]. Indeed, it was shown experimentally that error is reduced dramatically as more off policy data is made available, supporting the idea that counterfactual policy evaluation is less biased when data is available [2]. Unsurprisingly, counterfactual policy evaluation's better utilization of data for estimation in off policy settings carries over to better performance in policy search tasks, with counterfactually-guided policy search outperforming baseline model-based policy search algorithms [2].

CHAPTER 4

MACHINE LEARNING IN CAUSAL INFERENCE

The independent causal mechanism assumption is important to both the theory and practice of causal inference. The problem of learning causal structure is relaxed when ICM is assumed, since the problem can then be broken down into sub-components which can even be learned independently in the case of finitely many known interventions [22]. As a result, learning algorithms exploiting the ICM will aim to learn autonomous modules representing the physical data generating mechanisms. This is especially relevant in situations where data is generated from distinct domains which either have very similar underlying causal structure or are the result of interventions. In these cases, the learned modules are more likely to be re-useable across domains and allow prediction under interventions [29].

4.1 Disentangling Factors of Variation

One method that is similar in theory to learning independent causal mechanisms is the deep learning method of disentangling factors of variation. In particular, the unsupervised model InfoGAN[5] learns disentangled representations of the data generating mechanism through an information theoretic extension to generative adversarial networks. A Generative Adversarial Network (GAN) aims to learn a data generating distribution $\mathbb{P}_{\mathcal{G}}$ which approximates the true distribution $\mathbb{P}_{data}$. This is done by learning a generator network G which transforms independent noise variables $\mathbf{u}$ into samples $G(\mathbf{u}) \sim \mathbb{P}_{\mathcal{G}}$. G is trained by playing a minimax game against a discriminator network D which attempts to classify samples into the true distribution (given by ground truth data) or the approximate distribution (given by generated samples). In standard GANs the minimax game is formulated as follows:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim \mathbb{P}_{data}} [\log D(x)] + \mathbb{E}_{x \sim \mathbb{P}_{\mathcal{G}}} [\log(1 - D(G(x)))],$$

where $D(x) = \frac{\mathbb{P}_{data}(x)}{\mathbb{P}_{data}(x)+\mathbb{P}_{\mathcal{G}}(x)}$ is the optimal discriminator [5]. To make the generator network capture disentangled (or independent) representations in the data generating process, additional noise variables $\mathbf{c}$ are added which are meant to correspond with these invariant mechanisms. To ensure the network actually learns such representations, a regularization term based on the mutual information between $\mathbf{c}$ and $G(\mathbf{c}, \mathbf{z})$ is added such that the mutual information $I(\mathbf{c}, G(\mathbf{c}, \mathbf{z})$ is high [5]. This is motivated by the fact that if variables X and Y are related by a deterministic, invertible function, then maximal mutual information is attained [15]. So by adding the regularization term, a generative network G is rewarded for learning mappings which are highly dependent on the independent "factors of variation," $\mathbf{c}$, and also contribute to a good approximation of the true generating distribution. Adding this regularization term gives the following formulation of the minimax game:

$$\min_G \max_D V_I(G, D; \lambda) = V(G, D) + \lambda I(\mathbf{c}, G(\mathbf{z}, \mathbf{c})),$$

where λ is a hyperparamter determining how much the dependence of G on $\mathbf{c}$ is weighted [5]. In general the information term $I(\mathbf{c}, G(\mathbf{z}, \mathbf{c}))$ is difficult to maximize without this knowledge of the posterior so a variational inference approach was developed in [5] to train based on the quality of an approximate posterior.

Although the method just described for disentangling factors of variation does not explicitly talk about causality, on a conceptual level it is remarkably similar to learning independent causal mechanisms. Recall that the rigorous basis for describing the ICM principle used the idea of algorithmic mutual information for determining the extent to which causal mechanisms are similar, with mechanisms being independent if their mutual information vanishes [15]. This implies that for independent mechanisms y and x, the conditional mechanism $y|x$ is similar in complexity to the unconditioned mechanism y. The same principle is used, but in reverse, to ensure dependence of the generating mechanism $G(\mathbf{c}, \mathbf{z})$ on the underlying independent factors of variation $\mathbf{c}$ by maximizing their mutual information. So while it does not use the information metric to directly check if the learned mechanisms are independent, it does ensure that the mechanisms in the generative model preserve dependence on these independent factors. This is an important distinction to make in that untangling independent factors of variation is only achieving a special case of the ICM principle. This corresponds to the case in which causal mechanisms are a function only of the exogenous noise variables and not on other variables observed in the data, implicitly assuming that in the causal graph, $\mathbf{Pa}_i = \emptyset$ for all nodes i. For some tasks, this assumption is clearly inappropriate and so more general approaches to learning causal mechanisms would be needed.

4.2 Structural Autoencoders for Learning Causal Representations

Another approach for causal representation is to use an encoder-decoder architecture to learn a latent causal structure from high dimensional data [19]. This approach is most appropriate when the observable data is likely not reflective of causal variables. A typical example of this is image data, where the high dimensional input consists of the values for each pixel, which probably do not correspond with causal variables themselves. Instead, from the d-dimensional input we wish to discover causal variables $S_1, \ldots, S_n$ ($n \ll d$) and mechanisms

$$S_i = f_i(\mathbf{Pa}_i, U_i) \quad i \in \{1, \ldots, n\},$$

which model the causal relationships among the latent causal variables. First, an encoder $p \colon \mathbb{R}^d \to \mathbb{R}^n$ is applied to the input data which reduces the input dimension to the ”bottleneck” representation which comprises the noise variables $U_1, \ldots, U_n$ [19]. From this, the mappings $\mathcal{F}$ determined by the causal structure are applied to the noises to get the outcomes of the latent causal variables $\mathbf{S}$, and finally a decoder $q \colon \mathbb{R}^n \to \mathbb{R}^d$ is applied to return to the dimension of the data [19]. The system is trained using reconstruction error, similar to the objective in the competitive learning method, to ensure the low dimensional representation retains most of the information in the data. In practice the causal mechanisms $\mathcal{F}$ will not be known *a priori*, in which case the decoder learns the composition of the encoder and the causal mechanisms, $q \circ f$, and is thus referred to as a *structural decoder* [19]. This means that the causal structure is also implicitly learned by the decoder, and the quality of this representation is dependent on which distribution shifts and interventions are available in the data [29]. The encoder can be viewed as learning the anticausal association of causal variables from the resulting data, and the decoder as a causal generator that maps the latent structural mechanisms to the resulting output.

Since the causal mechanisms are often assumed unknown, the encoder-decoder architecture can be used in conjunction with the idea of disentangling factors of variation in the latent space to promote the development of independent mechanisms in the latent space. In the case of encoder-decoder architectures, using information based regularization is less effective in practice [6]. Instead, a hybrid sampling technique based explicitly on the independence of the latent sample space, effectively enforcing the ICM on the learned causal mechanisms, was shown to be more effective [19]. Since the noises $U_1, \ldots, U_n$ are assumed independent and not the causal variables, **S**, the goal of finding a disentangled representation of the latent **U** is achieved by enforcing an hierarchy on $U_1, \ldots, U_n$ such that they are used in the evaluation of the causal mechanisms sequentially and not simultaneously [19]. Again, the extent to which the latent noises can be disentangled is highly dependent on the interventions available in the data. In experiments using images of 3D objects subjected to transformations, incorporating this structural architecture was shown to reduce error throughout the generative process, without the need for regularization in the latent space [19]. Furthermore, the structural models were able to generalize to new datasets more successfully without negatively impacting performance on the initial set, suggesting that they do not over fit the causal mechanisms based on the original training set [19]. Much like in the anticausal and interventional problems seen above, learning autonomous representations of an underlying generative process is a key component to success outside of the iid setting.

4.3 Predicting the Impact of Treatments

Being able to understand the specific impact of interventions is a crucial part of the general scientific method. Yet, the preferred strategy of identifying causal effect is to create an intervention on the processes of interest, which is often limited by practical or ethical constraints. The importance of understanding causal relationships, and the existence of major practical constraints for doing so, has catalyzed the development of tools for determining causation from observational data with known interventions, such as predicting outcomes of medical treatments on patients without the need for long, expensive clinical trials.

An often sought after metric is the individual treatment effect (ITE) which is defined as the expected difference in outcome having received treatment, $Y_{T=1}$, and not having received treatment, $Y_{T=0}$: [27]

$$ITE = \mathbb{E}[Y_{T=1}|\mathbf{X}] - \mathbb{E}[Y_{T=0}|\mathbf{X}].$$

The difficulty in determining the ITE is that the true difference in outcomes between one treatment or another cannot be known for an individual patient. This constraint is known as *the fundamental problem of causal inference*, and a major area of research is devoted to estimating counterfactual outcomes which cannot be observed [24]. To be able to identify a causal effect, one typically assumes *ignorability* and *positive overlap*. Ignorability assumes that the treated and untreated groups have identical distributions such that the treatment assignment is independent of the effect [27]. Positive overlap assumes that the probability of receiving any level of treatment is positive for every individual, regardless of the individual's characteristics [27]. This translates to the fact that there is no individual for whom receiving the treatment is impossible, and similarly each individual has a positive probability of not receiving treatment. Together these assumptions constitute *strong ignorability*, which allows for efficient estimation of the individual treatment effect using the potential outcomes framework [27]. However, most methods to correct for potential bias require that the second assumption, positive overlap, is strongly met, which is impossible to test and rarely satisfied in practice [7]. Recent methods for estimating the impact of treatments have focused on avoiding these strong assumptions.

4.3.1 Neural Stochastic Differential Equations

Instead of trying to estimate the ITE directly with the assumptions outlined above, one method called CF-ODE uses a Bayesian approach for predicting the potential outcomes with embedded uncertainty quantification based on data coverage in the covariate space [7]. By leveraging neural ODEs, this approach also works better than other methods when data is irregularly sampled since it can output predictions at arbitrary times [4]. Additionally, this method predicts the impact of treatments over time by formulating the problem using time series and continuously modeling the dynamics of a latent hidden state used for prediction.

The problem of predicting the impact of a treatment over time is well formulated in the potential outcomes framework [27], which can be expressed as follows. Consider a set of N multivariate time series $\mathcal{X}$ sampled at times $\mathbf{t}_i = \{t_0, \dots, t_{k_i}\}$ where k_i is the number of observations of the time series $X_i(t) \in \mathcal{X}$. Here it is not assumed that $\mathbf{t}_i$ is evenly spaced, meaning the data are possibly irregularly sampled. For an individual, denote the observable history before the time of treatment t^* as $S_{t'}(X_i) = \{X_i(t) \colon t < t'\}$. The factual (observed) treatment assignment is denoted as $T_i^* \in \{0, 1\}$, and t_i^* denotes the time of treatment. For each time series, the goal is to predict the potential outcome trajectories after treatment time t_i^* given the history before treatment $S_{t_i^*}(X_i)$.

Instead of assuming ignorability and positive overlap as highlighted above, this method makes the following assumptions about the data generating system.

Assumption: Common Dynamical System [7]

All observations and potential outcomes $Y(t)$ are driven by a common dynamical system characterized by an *a priori* unknown ODE:

$$\begin{aligned} \frac{dh_T}{dt} &= f(h_T(t), u_T(t - t^*)) \\ X(t) &\sim g(h_{T^*}(t)) \\ Y_T(t) &\sim g(h_T(t)), \quad \forall t \geq t^*, \end{aligned}$$

where $u_T(\cdot)$, $g(\cdot)$, and $f(\cdot)$ are unknown functions.

This assumption asserts the idea that the latent causal mechanism is some continuous process $h(t)$ with dynamics characterized by the above ODE. We can then get observations $X(t)$ by mapping the latent process through an emission function $g(\cdot)$. This also follows the work of [12] which models the impact of interventions on a dynamical system as a continuous input $u_T(t)$, where T indicates the treatment given. Because $u_T(t)$ is supposed to represent the causal impact of treatment, we restrict $u_T(t) = 0$ for all times before treatment ($t < t^*$) since cause must precede effect [7]. With this perspective of the data generating process, ignorability is also assumed as follows.

Assumption: Ignorability [7]

Conditioned on a treatment assignment time t^*, the probability of treatment assignment is a function of the hidden state, $T(t^*) \sim \tau(h(t^*))$. Also, there exists a map $\phi(\cdot)$ between the available measurements at the time of treatment $S_{t^*}(X)$ and the unobserved latent process at the time of treatment $h(t^*)$ such that for all observed time series $X(t)$ we have

$$\phi(S_{t^*}(X)) = h(t^*).$$

Intuitively, this assumption formalizes the idea that the assigned treatment T is dependent on the past observations via the latent process $h(t^*)$, and that the observed covariates are sufficient for correcting this bias. Since we cannot observe the true hidden state $h(t)$ at any time t, we additionally require that a mapping from the hidden state to the covariates exists, and that the observed data is sufficient to control for any confounding variables [7]. This assumption corresponds to the traditional *strong ignorability* assumption, which entails that there are no confounding variables outside of what we observe. If ignorability is not held and there is an unobserved confounder, then without these further assumptions, unbiased causal effect estimation is impossible [24].

Although positive overlap is not explicitly assumed, it is still desired to understand how "certain" the estimators are depending on data availability. For the a time series $X(t)$, the certainty of potential outcome estimates can vary based on how often the treatment T is assigned for similar time series [7]. So by equipping the model with uncertainty estimates through a Bayesian framework, we can characterize uncertainty based on overlap between observed time series.

Under these assumptions, the dynamics of the data generating system is learned in two parts: recovering the hidden state and integrating this state forward with treatment inputs. First, the mapping from observations prior to the time of treatment t^* to the hidden state are learned using a RNN Φ with parameters ϕ:

$$h(t^*) = \Phi_\phi(S_{t^*}(X)).$$

The dimension of the hidden state $h(t)$ is assumed to be less than the dimension of the covariate space S_{t^*}, so Φ acts as an encoder [7]. Once the hidden state at time of treatment, $h(t^*)$, is recovered, it is evolved based on the dynamics outlined in the common dynamical system assumption and used to compute predictions for the outcomes $Y(t)$. To do this, a neural differential equation [4] is used to model the dynamics using the neural networks $f_{\theta_f}(\cdot)$, $g_{\theta_g}(\cdot)$, and $u_{T,\theta_u}(\cdot)$ such that

$$\begin{aligned} \frac{dh_T}{dt} &= f_{\theta_f}\left(h_T(t), u_{T,\theta_u}(t-t^*)\right), \\ Y(t) &= g_{\theta_g}\left(h_T(t)\right) [7]. \end{aligned}$$

Here g serves as a decoder, taking the hidden state back to the covariate space. Following the work of [12], the treatment effect, u_T, is modeled as an external intervention on the dynamical system $\frac{dh(t)}{dt}$.

To incorporate uncertainty in estimates, the Bayesian idea of estimating the posterior distribution of model parameters given the data, $\mathbb{P}(\theta_f|S_{t^*}, Y)$, is adopted [16]. This method involves assuming a prior on the process $\mathcal{H} = h_T(t)|h_T(t^*)$, written $p_0(\mathcal{H}|S_{t^*})$, and optimizing a variational approximation of the posterior, $q_\theta(\mathcal{H}|S_{t^*})$, by maximizing the evidence lower bound (ELBO), written: [16]

$$\log(\mathbb{P}(S_{t^*}, Y)) \leq \mathbb{E}_{q_\theta(\mathcal{H}|S_{t^*})}\left[\ln(\mathbb{P}(Y|\mathcal{H}))\right] - KL_{q_\theta(\mathcal{H}|S_{t^*})}\left(q_\theta(\mathcal{H}|S_{t^*})\|p_0(\mathcal{H}|S_{t^*})\right).$$

In general, the KL-divergence term is very difficult to compute since it involves the evaluation of multiple integrals, but it was shown that if the variational approximation $q_\theta(\mathcal{H}|S_{t^*})$ and the prior $p_0(\mathcal{H}|S_{t^*})$ are both assumed to be diffusion processes with the same diffusion parameter, then the KL divergence becomes tractable and only requires the evaluation of a single integral [7]. Thus, the ELBO can be computed using a stochastic differential equation solver [20] and gradients for efficient optimization of model parameters are obtained using the adjoint method for neural differential equations introduced in [4]. By modeling the treatment effect as an external intervention on the underlying dynamics, this framework can estimate potential outcome trajectories $Y_{T=1}$ and $Y_{T=0}$ by using different u_T to impact the underlying dynamics, with embedded uncertainty estimates based on availability of data. The individual treatment effect is thus computed at a time t by comparing $Y_{T=1}(t)$ and $Y_{T=0}(t)$.

4.3.2 Variational Autoencoders for Causal Effect Inference

The method described in the previous section, CF-ODE, determined the individual treatment effect by comparing how the latent state evolved after being impacted by different treatments. A similar approach, called CEVAE (causal effect variational auto-encoder), adopts the idea of using latent variables to represent unobserved confounders, $\mathbf{Z}$ [21]. This method assumes the following causal structure, which is nearly identical to that in CF-ODE:

The only difference between the causal structure from CF-ODE is that the latent variables here are *proxies*, meaning they influence the observed data and the outcome, but are not influenced themselves by the observed data [24]. Assuming this type of causal structure is useful when there are many potential proxies, and so they can be discovered and their impact on outcomes and treatments can be inferred [21]. CEVAE aims to estimate the

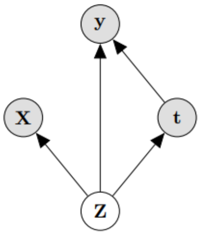

Figure 4.1. Causal diagram for CEVAE. [21]

joint interventional distribution of the data and proxies, so that the ITE can be computed directly using Pearl's backdoor adjustment formula [24]. Given an estimate of the joint distribution $\mathbb{P}(\mathbf{X}, \mathbf{Z}, \mathbf{t}, Y)$, the assumed causal structure and known facts of *do*-calculus [24] imply that the conditional distributions in the formula for ITE can be computed: [21]

$$\begin{aligned}\mathbb{P}(Y|\mathbf{X}, do(\mathbf{t}=i)) &= \int_{\mathbf{Z}} \mathbb{P}(Y|\mathbf{X}, do(\mathbf{t}=i), \mathbf{Z})\mathbb{P}(\mathbf{Z}|\mathbf{X}, do(\mathbf{t}=i))\, d\mathbf{Z} \\ &= \int_{\mathbf{Z}} p(Y|\mathbf{X}, do(\mathbf{t}=i))\mathbb{P}(\mathbf{Z}|\mathbf{X})\, d\mathbf{Z}.\end{aligned}$$

Given this fact, it remains to estimate the joint distribution $\mathbb{P}(\mathbf{X}, \mathbf{Z}, \mathbf{t}, Y)$ from observations $(\mathbf{X}, \mathbf{Y}, \mathbf{t})$. Rather than making assumptions about the dynamics of a hidden state $h(t)$, CEVAE only assumes that the joint distribution $\mathbb{P}(\mathbf{X}, \mathbf{Z}, \mathbf{t}, Y)$ can be approximately recovered soley from observation triplets of covariates, outcomes, and observed treatment, $(\mathbf{X}, Y, t)$. The joint distribution is approximately recovered by using variational auto-encoders, which are used to learn the complex relationships between observed covariates $\mathbf{X}$ and latent triplets $(\mathbf{Z}, Y, t)$ [21]. This estimation approach differs slightly from CF-ODE in that it uses the variational inference framework for estimating the posterior of an inference network modeling $\mathbb{P}(\mathbf{Z}, Y, t|\mathbf{X})$. Details of the exact implementation and design choices can be found in [21].

CHAPTER 5

EXPERIMENTS

Many of the models introduced in chapter 4 (including CF-ODE [7] and CEVAE [21]) have efficient implementations in open access python libraries [3][7]. These off the shelf packages are used to perform the following experiments. Code for producing the figures and models in the following experiments can be found here: `https://github.com/chalberg/causalML_experiments.git`

5.1 Dexamethasone Experiment

The gold standard for causal knowledge of a system is a dynamical system which completely describes the interactions between variables. Given a dynamical system, one can perfectly understand the impact that changing one variable will have on the system. This also means that inferring the impact of an intervention on a system is reducible to computing the factual and counterfactual outcomes using the deterministic equations. In the following experiment, data is synthetically generated from a dynamical system so that ground truth counterfactual outcomes can be used to the model's ability to estimate from only observable (non-counterfactual) data.

To demonstrate predicting the impact of treatments over time, CF-ODE is used to learn the impact of dexamethasone, a drug used in treatment against COVID-19, on the body's innate immune response. The dynamical system, presented in the following equation is adapted from DeBrouwer et. al. [7]. Variables z_1 and z_5 represent the innate and adaptive immune response, z_2 and z_3 the concentration of dexamethasone in the lung tissue and plasma, and z_4 represents the viral load [7].

$$\begin{aligned}
\frac{dz_1}{dt} &= k_{IR} \cdot z_4 + k_{PF} \cdot z_4 \cdot z_1 - k_O \cdot z_1 + \frac{E_{\max} \cdot z_1^{hP}}{EC_{50}^{hP} + z_1^{hP}} - k_{Dex} \cdot z_1 \cdot z_2 \\
\frac{dz_2}{dt} &= -k_2 \cdot z_2 + k_3 \cdot z_3 \\
\frac{dz_3}{dt} &= -k_3 \cdot z_3 \\
\frac{dz_4}{dt} &= k_{DP} \cdot z_4 - k_{IIR} \cdot z_4 \cdot z_1 - k_{DC} \cdot z_4 \cdot z_5^{hC} \\
\frac{dz_5}{dt} &= k_1 \cdot z_1
\end{aligned}$$

It is assumed that only variables z_1 and z_5 can be measured, so only those two variables are included in the observable time series $\mathbf{X}(t)$ [7]. The intervention is modelled by simulating a constant injection of dexamethasone ($\frac{dz_3}{dt} = 10$). Confounding is introduced by modeling a dependence of the treatment assignment on the factor k_{Dex}, that also impacts the effect of dexamethasone on the immune response.

The after simulating $1,000$ individuals, with half receiving treatment and half not, the model was trained on 750 individuals with the rest being used as a validation set. Figure 5.1 is an example of the factual and counterfactual outcomes for an individual after treatment. In this case, the individual did receive dexamethasone, so the predicted individual treatment effect on the innate immune response is increasingly negative after 4 hours.

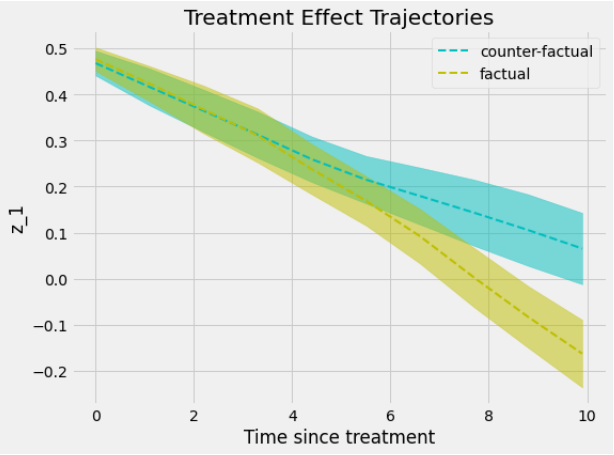

Figure 5.1. Predicted factual and counterfactual trajectories after time of treatment. In this case, the individual received Dexamethasone, so the counterfactual measures the patient's innate immune response had they not received the drug. Shaded regions represent the standard deviation of the estimate at any given time.

Also notice that the counterfactual estimate has a greater standard deviation, and consequently more uncertainty, when it is far from the factual estimate. This demonstrates the principle that estimates should be more uncertain when factual data is less available. To further see how out of distribution data impacts uncertainty in practice, a new set of covariates was randomly generated for each individual and the original treatment was applied for the randomly generated data. Figure 5.2 shows that average uncertainty in estimates of outcomes is indeed higher for out of distribution data.

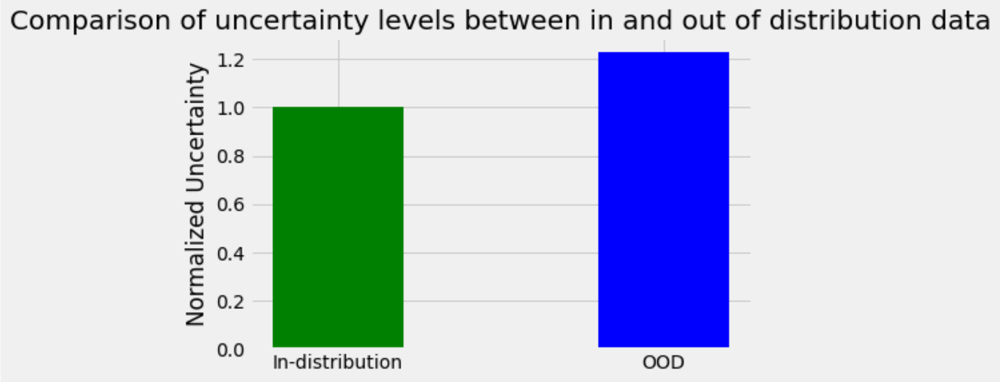

Figure 5.2. Comparison of uncertainties in estimates of innate immune response between in distribution data and out of distribution data. Clearly, the model has a more uncertain estimate when estimating from an example it has not been trained on before.

Lastly, figure 5.3 shows that on average, counterfactual estimates had slightly lower accuracy than factual estimates. However, the Precision in Estimation of Heterogeneous Effect (PEHE), defined as [13]

$$PEHE = \sqrt{(Y_{i,T} - Y_{i,T'})^2 - (\hat{Y}_{i,T} - \hat{Y}_{i,T'})^2},$$

is more useful in determining how accurately a model predicts the size of a treatment effect. CF-ODE had a relatively low average PEHE on the validation set at 0.0462, meaning the difference in size between true and predicted treatment effect was small. Additionally, since the RMSE for both the factual and counterfactual estimated outcomes were low, the model likely predicted the individual treatment effect accurately in most cases.

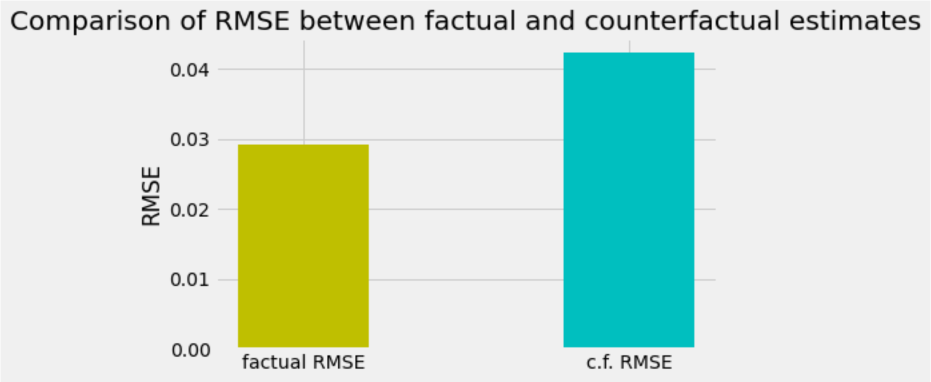

Figure 5.3. Comparison of Root Mean Square Error (RMSE) between average factual and counterfactual estimates of innate immune response (z_1). The counterfactual estimate is on average more uncertain, which corresponds with the previous example of the model being worse in out of distribution settings.

5.2 IHDP Experiment

The Infant Health and Development (IHDP) dataset originates from a randomized experiment looking at the effect of home visits by specialists on future cognitive scores in infants [10]. The IHDP simulation is considered the de-facto standard benchmark for neural network treatment effect estimation methods [3]. In this task, the performance of CEVAE is compared against a baseline S Learner, both of which are implemented using the *causalML* library [3]. In particular, the S Learner implemented here uses elastic net regression as the base model, and estimates individual treatment effect as the difference between the predicted outcomes with treatment and without treatment [18]. In both models, the Average Treatment Effect (ATE) is estimated as the mean of individual treatment effects. The performance of each of the models on the validation set is summarized in the following table.

	ATE	MSE	Abs % Error of ATE	KL Divergence	AUUC
Actual	4.775	0	0	0	NaN
S Learner	4.691	79.589	0.018	0.313	0.684
CEVAE	2.024	151.297	0.576	0.755	0.538

Notice the S Learner had a far better estimate than CEVAE in this example. This is likely because the individual treatment effect is very heterogeneous, and since CEVAE depends on the existence of informative proxies for estimating the impact of a treatment, it performed poorly in this low dimensional setting. On the other hand, S Learners flexibly estimate the factual and counterfactual distributions directly using elastic net regression, so low dimensional covariates are not an issue for predicting treatment effects as long as the influence of unmeasured confounders is minimal.

CHAPTER 6

DISCUSSION

Although causal inference is a relatively young field, it has done much to improve understanding about the mechanisms underlying the data we observe. The Pearl Causal Hierarchy gives strict limits on what can be inferred from different sources of data. Recognizing these limitations has led to a more precise study of inference in these settings, with an emphasis on interventional inference when data is generated from distributions that are the result of distinct interventions. Progress in this area has led to more robust and generalizable machine learning methods that are able to perform in out of distribution settings resulting from transformations of the baseline data. Still, most methods focus on dealing with a relatively small number or unspecified, but simple interventions. It is not clear whether approaches like competitive learning would be able to scale to settings with many thousands of possible interventions. Being able to perform well in complex, dynamic settings remains a very difficult learning problem.

Methods for learning disentangled representations and causal structure of interventional data have also been successful. The independent causal mechanism and sparse mechanism shift hypotheses are extremely useful assumptions in practice because they provide a clear goal which learning problems may be centered around. Many of the methods covered in this paper use this approach and have both theoretical and empirical success in learning problems. Additionally, the ideas from anticausal learning and examples covered in this paper provide support for adopting generative approaches to learning problems which have typically been seen as purely classification tasks. Further research towards more reliable methods for explicitly enforcing the discovery of autonomous generative mechanisms could result in better causal representation, and thus greater robustness and generalizability. Causality in machine learning is proving to be a necessary component for building reliable AI systems which can perform well in out of distribution settings.

As mentioned earlier, reliability is hugely important in high consequence applications of AI such as autonomous vehicles, and so working towards causal representation in these systems is an important area for further research.

Many methods in this paper focus on applications in image classification to demonstrate how causal ideas improve robustness and transferability, but there are many more interesting applications which have yet to be explored. Some of the most important and difficult areas for causal machine learning are in medical applications. Estimating the impact of a treatment is a widely researched problem, but many estimation techniques often rely on restrictive and untestable assumptions. Developing reliable techniques for treatment effect estimation which avoid these assumptions is an increasingly important area of research, which is becoming more feasible with the greater volume of electronic health data. There is still a significant need for further research in expanding causal machine learning to handle informative assessment bias in data collection, sparseness in longitudinal data, and multiple categorical or continuous treatments. These issues are common in electronic health record databases, so finding solutions based in causal inference would allow for much greater utilization of this abundant observational data in answering clinical questions without the need for expensive trials. With the rapid growth in electronic health data, research in causal machine learning is becoming increasingly important for developing reliable tools for individual level treatment decisions and care.

REFERENCES

[1] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, ”On Pearl’s Causal Heirarcy and the Foundations of Causal Inference,” *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Feb. 2022. https://doi.org/10.1145/3501714.3501743

[2] L. Buesing, T. Weber, Y. Zwols, et al. ”Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search” arXiv:1811.06272. 15 Nov. 2018.

[3] H. Chen, T. Harinen, J. Lee, M. Yung, and Z. Zhao, ”CausalML: Python Package for Causal Machine Learning,” arXiv:2002.11631. 25 Feb 2020. `https://github.com/uber/causalml.git`

[4] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, ”Neural Ordinary Differential Equations,” in *Advances in Neural Information Processing Systems 31* (NeurIPS 2018).

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abeel, ”InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets,” arXiv:1606.03657. 12 Jun. 2016.

[6] B. Dai, and D. Wipf, ”Diagnosing and Enhancing VAE Models,” arXiv:1903.05789. (2019)

[7] E. De Brouwer, J. G. Hernández, and S. Hyland, ”Predicting the Impact of Treatments Over Time with Uncertainty Aware Neural Differential Equations,” arXiv:2202.11987. 24 Feb. 2022.

[8] D. Geiger and J. Pearl, ”Logical and Algorithmic properties of Conditional Independence and Qualitative Independence,” UCLA, Cognitive Systems Laboratory, Technical Report R-97 (March 1988)

[9] D. Geiger and J. Pearl, ”Logical And Algorithmic Properties of Independence and their Application to Bayesian Networks,” *Annals of Mathematics and Artificial Intelligence* Vol. 2 (1990) pp. 165-178.

[10] T.R. Gross, ”Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight, Premature Infants in the United States, 1985-1988.” *Inter-university Consortium for Political and Social Research*, 1993-10-03. `https://doi.org/10.3886/ICPSR09795.v1`

[11] S. Gu and L. Rigazio, ”Towards Deep Neural Network Architectures Robust to Adversarial Examples,” arXiv:1412.5068, (2014)

[12] D. Gwak, G. Sim, M. Poli, S. Massaroli, J. Choo, and E. Choi, ”Neural Ordinary Differential Equations for Intervention Modeling,” arXiv:2010.08304. 16 Oct. 2020

[13] J. L. Hill, ”Bayesian nonparametric modeling for causal inference.” *Journal of Computational and Graphical Statistics*, 20(1):217–240. (2011)

[14] A. Jaber, J. Zhang, E. Bareinboim, ”Identification of Conditional Causal Effects under Markov Equivalence,” *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 11516–11524, Dec. 2019.

[15] D. Janzing and B. Schölkopf, ”Causal inference using the algorithmic Markov condition,” *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[16] D. P. Kingma and M. Welling, ”Auto-Encoding Variational Bayes,” arXiv:1312.6114. 1 May 2014.

[17] N. Kilbertus, G. Parascandolo, and B. Schölkopf, ”Generalization in anti-causal learning.” arXiv:1812.00524. 3 Dec. 2018

[18] S. R. Künzel et. al. ”Metalearners for estimating heterogeneous treatment effects using machine learning.” *Proceedings of the National Academy of Sciences* 116.10 (2019). pp. 4156-4165.

[19] F. Leeb, Y. Annadani, S. Bauer, and B. Schölkopf, ”Structural Autoencoders Improve Representations for Generation and Transfer,” arXiv:2006.07796. 14 Jun 2020.

[20] X. Li, T.-K. L. Wong, R. T. Q. Chen, and D. Duvenaud, ”Scalable Gradients for Stochastic Differential Equations,” In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR (2020).

[21] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, ”Causal Effect Inference with Deep-Latent Variable Models,” arXiv:1705.08821. 6 Nov. 2017.

[22] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf, ”Learning Independent Causal Mechanisms,” arXiv: 1712.00961. 8 Sep. 2018.

[23] J. Pearl, ”Causal diagrams for empirical research,” *Biometrika* (1995), pp. 669-710

[24] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009

[25] H. Reichenbach. *The Direction of Time*. University of California Press, Berkeley, CA, 1956.

[26] N. Rosemary Ke, O. Bilaniuk, A. Goya, S. Bauer, H. Larochelle, B. Schölkopf, M. C. Mozer, C. Pal, and Y. Bengio, ”Learning Neural Causal Models from Unknown Interventions,” arXiv: 1910.01075. 23 Aug 2020.

[27] D. B. Rubin, ”Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of educational Psychology*, 66(5):688. (1974)

[28] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. ” On Causal and Anticausal Learning,” *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pg. 1255–1262 (2012)

[29] B. Schölkopf, F. Locatello, S. Bauer, N. Rosemary Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, ”Towards Causal Representation Learning,” arXiv:2102.11107, 24 Feb. 2021.

[30] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. ”Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature*, 529(7587): pp. 484–489, 2016.

[31] R. Sutton and A. Barto, *Reinforcement Learning: Second Edition*. MIT Press, ISBN: 9780262364010. 2018.

[32] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, ”Intriguing Properties of Neural Networks,” arXiv preprint 1312.6199, 2013.

[33] Tunyasuvunakool et. al. ”Highly accurate protein structure prediction for the human proteome,” *Nature* 2021 Aug;596(7873):590-596. doi: 10.1038/s41586-021-03828-1. Epub 2021 Jul 22. PMID: 34293799; PMCID: PMC8387240.