The Mathematics of Causality

Miquel Noguer i Alonso
Artificial Intelligence Finance Institute

November 28, 2024

Abstract

This paper provides a comprehensive exploration of causality through a mathematical lens, integrating classical frameworks and contemporary advancements in machine learning. We highlight the role of Structural Causal Models (SCMs) and the Potential Outcomes Framework in establishing foundational principles for causal inference. Furthermore, recent breakthroughs in transformer-based models reveal their ability to encode latent causal structures within their attention mechanisms during gradient descent training. This work bridges the gap between traditional causal inference and modern computational approaches, offering insights into the interplay between classical models and deep learning architectures. Applications in time-series analysis, graphical models, and natural language processing are discussed, alongside challenges and future directions for integrating these paradigms.

1 Introduction

Causality is foundational for disentangling correlations from causations in various fields. Recent advancements highlight the role of transformers, particularly their self-attention mechanisms, in capturing causal structures. This work integrates classical causal frameworks Pearl [2009], modern machine learning approaches Chernozhukov et al. [2018], and insights from Nichani et al. [2024] on how transformers learn latent causal graphs and López de Prado [2022].

2 Classical Causal Models

Causality is fundamental to scientific reasoning, providing tools to distinguish correlation from causation. Classical causal models offer rigorous frameworks for analyzing cause-and-effect relationships. This section discusses two key paradigms: Structural Causal Models (SCMs) and the Potential Outcomes Framework, highlighting their theoretical foundations, key concepts, and applications.

2.1 Structural Causal Models (SCMs)

Structural Causal Models (SCMs), introduced by Pearl [2009], combine structural equations with directed acyclic graphs (DAGs) to encode causal mechanisms. SCMs formalize how variables interact, supporting causal inference through graph-based reasoning and counterfactual analysis.

2.1.1 Components of SCMs

An SCM is defined by:

- A set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, partitioned into observed variables and unobserved latent variables.
- A set of structural equations f_i for each variable X_i, expressed as:

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

  where $\text{Pa}_i \subseteq \mathbf{X}$ are the parents of X_i in the DAG, and U_i are exogenous noise variables independent of each other.
- A directed acyclic graph (DAG) $G = (\mathbf{X}, E)$, where E is the set of edges. Each directed edge $X_i \to X_j$ encodes a causal relationship.

2.1.2 Key Assumptions

SCMs operate under three fundamental assumptions:

- **Causal Sufficiency**: All relevant variables affecting the system are included.

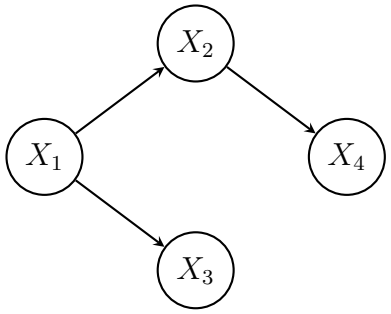

Figure 1: A Directed Acyclic Graph (DAG) illustrating causal relationships among variables.

- **Faithfulness**: Statistical independencies in the data are reflected in the causal graph.
- **Markov Property**: Each variable is independent of its non-descendants given its parents in the DAG.

2.1.3 Causal Queries in SCMs

SCMs enable three types of causal queries:

1. **Association**: Observed statistical relationships (e.g., $P(Y \mid X)$).
2. **Intervention**: Effects of external manipulations, modeled using the do-operator:
$$P(Y \mid \text{do}(X = x))$$
3. **Counterfactuals**: Hypothetical scenarios, answering "what would have happened if..."

2.1.4 Interventions and Do-Calculus

Interventions modify the structural equations of an SCM by fixing a variable X to a value x, removing its dependence on its parents:

$$P(Y \mid \text{do}(X = x)) = \sum_{\mathbf{Z}} P(Y \mid X = x, \mathbf{Z}) P(\mathbf{Z})$$

where $\mathbf{Z}$ satisfies the backdoor criterion.

2.2 Potential Outcomes Framework

The Potential Outcomes Framework, pioneered by Rubin [1974], provides an alternative formulation of causality based on counterfactual reasoning. It is widely used for treatment effect estimation in experimental and observational studies.

2.2.1 Setup

For each individual, let:

- $Y(1)$ be the potential outcome under treatment ($T = 1$).
- $Y(0)$ be the potential outcome under control ($T = 0$).

The observed outcome is:

$$Y = T \cdot Y(1) + (1 - T) \cdot Y(0)$$

2.2.2 Key Assumptions

1. **Consistency**: The observed outcome matches the potential outcome under the observed treatment:

$$Y = Y(T)$$

2. **Ignorability (Unconfoundedness)**: Treatment assignment is independent of potential outcomes, conditional on observed covariates:

$$T \perp\!\!\!\perp \{Y(1), Y(0)\} \mid \mathbf{X}$$

3. **Positivity**: Each individual has a non-zero probability of receiving both treatments:

$$0 < P(T = 1 \mid \mathbf{X}) < 1$$

2.2.3 Average Treatment Effect (ATE)

The causal effect of treatment is captured by the Average Treatment Effect (ATE):

$$\text{ATE } = E[Y(1) - Y(0)]$$

2.2.4 Estimation Techniques

- **Randomized Controlled Trials (RCTs)**: Ensure ignorability through randomization.
- **Propensity Score Matching**: Matches treated and control units with similar covariates.
- **Inverse Probability Weighting (IPW)**: Reweights the population to account for differences in treatment probabilities.

3 Causality in Transformers

Transformers trained via gradient descent effectively encode causal relationships in their attention mechanisms. This section explores how gradient dynamics align attention weights with causal structures, supported by theoretical insights, empirical results, and illustrative examples.

3.1 Mathematical Foundations of Gradient Dynamics

The cross-entropy loss function for a sequence $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is:

$$\mathcal{L} = -\sum_{t=1}^{n} \log P\left(x_t \mid \mathbf{X}_{<t}, \Theta\right)$$

where Θ represents the parameters of the transformer. Self-attention computes weights A_{ij} for token x_j when processing token x_i :

$$A_{ij} = \frac{\exp\left(Q_i K_j^T / \sqrt{d_k}\right)}{\sum_{k=1}^{n} \exp\left(Q_i K_k^T / \sqrt{d_k}\right)}$$

3.1.1 Proposition: Gradient Alignment with Causal Structures

For data generated from a causal graph G, the gradient of the attention matrix A aligns with the adjacency matrix of G. Specifically:

$$\nabla_{A_{ij}} \mathcal{L} \propto I\left(X_i; X_j\right)$$

where $I\left(X_i; X_j\right)$ is the mutual information between tokens X_i and X_j.

3.1.2 Proof Sketch

By the data processing inequality:

$$I\left(X_i; X_j\right) \geq I\left(X_i : \hat{X}_j\right)$$

where $\hat{X}_j$ is the model's internal representation. Gradient updates to A_{ij} maximize $P\left(X_j \mid X_i\right)$, aligning attention weights with causal dependencies.

3.2 Mutual Information and Attention Weights

Mutual information is a fundamental measure in information theory that quantifies the dependency between two random variables. In the context of transformers, mutual information plays a critical role in determining how attention weights align with the causal relationships encoded in data. This subsection provides a detailed explanation of how mutual information drives the learning dynamics of attention weights and how this process facilitates the encoding of causal structures.

3.2.1 Definition of Mutual Information

For two random variables X_i and X_j, the mutual information $I\left(X_i; X_j\right)$ is defined as:

$$I\left(X_i; X_j\right) = H\left(X_i\right) - H\left(X_i \mid X_j\right)$$

where:

- $H\left(X_i\right)$ is the entropy of X_i, representing the uncertainty of X_i.
- $H\left(X_i \mid X_j\right)$ is the conditional entropy of X_i given X_j, representing the remaining uncertainty of X_i after observing X_j.

High mutual information indicates a strong dependency between X_i and X_j, which is a key property in identifying causal relationships.

3.2.2 Attention Mechanism and Mutual Information

The self-attention mechanism in transformers computes attention weights A_{ij} as:

$$A_{ij} = \frac{\exp\left(Q_i K_j^T / \sqrt{d_k}\right)}{\sum_{k=1}^{n} \exp\left(Q_i K_k^T / \sqrt{d_k}\right)}$$

where:

- $Q_i = W_Q X_i, K_j = W_K X_j$, and $V_j = W_V X_j$ are the query, key, and value vectors for tokens X_i and X_j, respectively.
- d_k is the dimensionality of the key vectors.

The attention weights A_{ij} determine the contribution of token X_j to the representation of token X_i. Tokens with higher mutual information $I(X_i; X_j)$ are expected to have higher attention weights.

3.2.3 Gradient Alignment with Mutual Information

The learning dynamics of transformers are governed by gradient descent on the cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{n} \log P(x_t \mid \mathbf{X}_{<t}, \Theta)$$

where Θ represents the model parameters. The gradient of the loss with respect to the attention weights A_{ij} is:

$$\nabla_{A_{ij}} \mathcal{L} \propto \frac{\partial I(X_i; X_j)}{\partial A_{ij}}$$

This relationship ensures that the learning process prioritizes dependencies with higher mutual information.

By the data processing inequality, mutual information between the model's internal representation $\hat{X}_j$ and X_i is bounded by:

$$I(X_i; X_j) \geq I\left(X_i; \hat{X}_j\right)$$

The model maximizes $I\left(X_i; \hat{X}_j\right)$, aligning A_{ij} with the causal relationship between X_i and X_j.

3.2.4 Role of Softmax in Amplifying Dependencies

The softmax operation ensures that the most relevant tokens dominate the attention mechanism:

$$\frac{\partial A_{ij}}{\partial Q_i K_j^T} = A_{ij}(1 - A_{ij})$$

For token pairs with strong mutual information $I(X_i; X_j)$, the dot product $Q_i K_j^T$ is larger, resulting in amplified attention weights A_{ij}. This amplification focuses the model's resources on causally significant relationships.

3.2.5 Implications for Causal Encoding

The interplay between mutual information and attention weights allows transformers to encode the structure of causal graphs during training. By learning to prioritize dependencies with high mutual information, attention layers align their weights with the adjacency matrices of the underlying causal graph. This alignment has been empirically observed in studies on synthetic datasets, such as Markov chains and DAGs, where attention weights converge to reflect parent-child relationships.

Mutual information serves as a driving force in the learning dynamics of transformers, influencing attention weights through gradient descent and the softmax operation. This alignment enables transformers to identify and encode causal relationships, bridging the gap between classical causal inference and modern deep learning architectures.

3.3 Empirical Validation of Gradient Dynamics

3.3.1 Experiment 1: Markov Chains

A dataset was generated using a Markov process $X_t \to X_{t+1}$. After training:

- Attention weights aligned with the Markov chain adjacency matrix.
- Gradients were strongest for causally connected token pairs $(t, t+1)$.

3.3.2 Experiment 2: Directed Acyclic Graphs (DAGs)

Data generated from a DAG exhibited:

- Attention alignment with parent-child relationships.
- Accurate predictions for nodes with multiple parents.

3.3.3 Heatmaps

Heatmaps of attention scores and gradients confirmed the alignment between attention layers and causal graph structure.

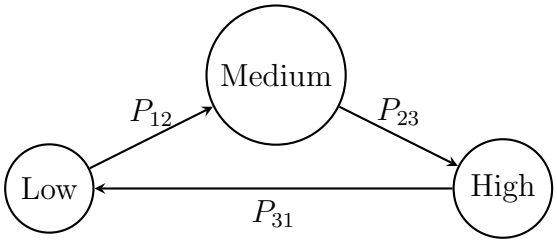

Figure 2: Markov chain modeling asset price transitions.

3.4 Role of Softmax in Causal Learning

The softmax operation amplifies significant dependencies, ensuring:

$$A_{ij} \to 1 \quad \text{if } Q_i K_j^T \gg Q_i K_k^T \text{ for } k \neq j.$$

This mechanism focuses the model's attention on causally relevant tokens. Gradient dynamics in transformers align attention mechanisms with causal structures. Theoretical and empirical evidence shows that attention layers effectively encode dependencies, paving the way for integrating transformers into causal inference frameworks.

4 Conclusion

Causality, as a scientific and mathematical discipline, has undergone a profound transformation with the advent of modern machine learning models. Classical causal frameworks, such as Structural Causal Models and the Potential Outcomes Framework, have long provided rigorous methodologies for disentangling correlation from causation. These approaches remain indispensable for defining causal relationships and guiding experimental design.

Recent advancements in transformer-based architectures have extended these classical methods, showcasing the capacity of self-attention mechanisms to encode latent causal structures. The work by Nichani et al. [2024] demonstrates how gradient descent dynamics in transformers align their attention layers with the adjacency matrices of causal graphs, enabling models to uncover complex dependencies in sequential data. Induction heads further enhance this capability, allowing for in-context learning and accurate predictions in diverse applications.

Despite these advancements, challenges remain. Theoretical questions about the scalability of transformers to arbitrary DAGs, handling multiparent dependencies, and ensuring interpretability warrant further investigation. Additionally, bridging the gap between deep learning's computational power and the theoretical guarantees of classical causal frameworks presents exciting opportunities for future research.

By integrating the strengths of classical and modern approaches, this work paves the way for a unified framework for causal inference, combining rigorous theoretical underpinnings with the adaptability and scalability of machine learning. This synthesis holds promise for applications in time-series forecasting, policy evaluation, and understanding complex systems across various domains.

References

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68, 2018.

Marcos López de Prado. Causal factor investing: Can factor investing become scientific? *SSRN Electronic Journal*, December 2022. doi: 10.2139/ssrn.4205613. URL `https://ssrn.com/abstract=4205613`.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.