SYSEN 6000: Foundations of Complex Systems

Assignment 2: Causal Loops & Emergent Behavior

Nick Kunz [NetID: nhk37] nhk37@cornell.edu

September 14, 2022

Casual Loops: Prediction Systems

The goal of this analysis was to broadly illustrate the causal relationships between system level considerations for the most important theoretical bounds of prediction systems, as they relate to synthetic data in machine learning. It is meant to serve as a basis for future research in that regard.

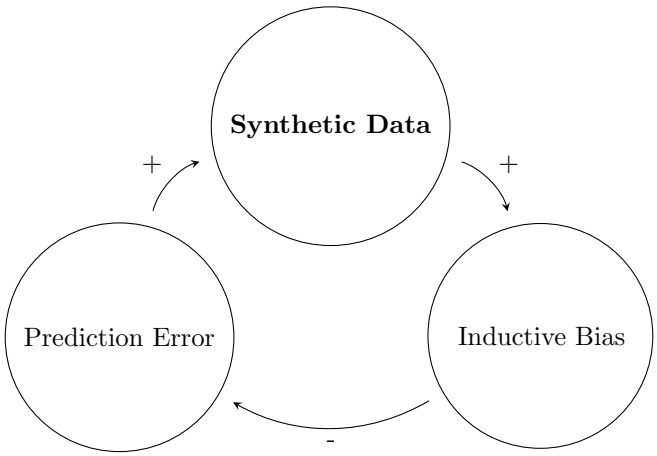

Fig. 1: Causal Loop Diagram (CLD) of synthetic data in prediction systems.

There are 3 components (nodes) contained within the CLD in Fig. 1, they are: Synthetic Data, Inductive Bias, and Prediction Error. These were selected because they capture the most basic theoretical components of many prediction systems with the exception of synthetic data, which was introduced as the domain of interest. The relationships (edges) between nodes were labeled with polarity (+/−), a metric used later to classify the type of behavior likely to emerge as a result of the causal loop.

When explaining the CLD in Fig. 1, it is best to begin at the Synthetic Data node. Take note of the polarity in relationship to the Inductive Bias node. Notice that a *positive* polarity was specified, assuming that when synthetic data increases, inductive bias also increases. Moving from the Inductive Bias node to the Prediction Error node, notice that the polarity is *negative*, suggesting that as inductive bias increases, prediction errors decrease (the limitation of this assumption is addressed later). Finally and perhaps most importantly, *positive* polarity was specified between the Prediction Error node and the Synthetic Data node. This assumes that as prediction errors increase, sample complexity of the synthetic data increases. In other words, more synthetic data is generated when prediction errors are high.

Emergent Behavior: Balancing Loops

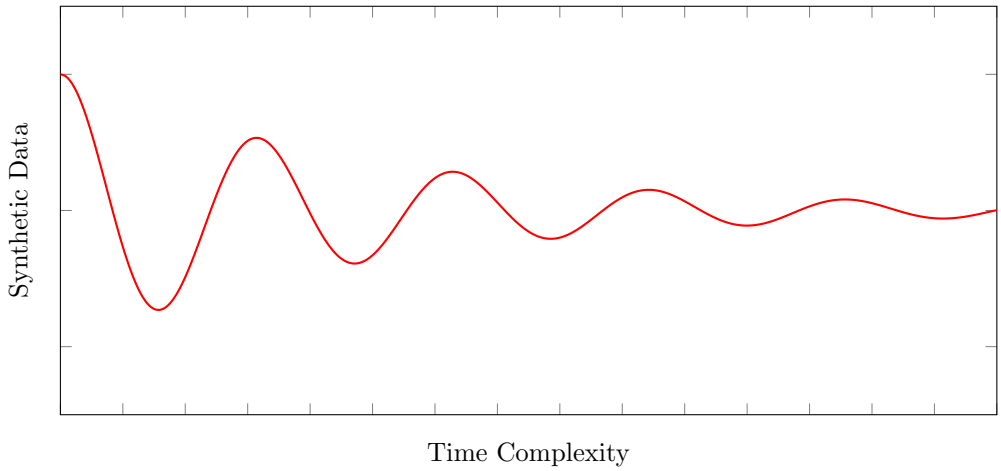

Fig. 2: Synthetic data complexity convergence over time as a B_1 loop.

When taken together, the overall polarity of the CLD in Fig. 1 can be classified as a Balancing Loop, B_1. This is in contrast to a Reinforcing Loop, R_1, which would likely emerge in the case that the sum of polarities equal zero. In this case, the overall polarity is non-zero, which indicates that the behavior of the CLD is likely to undulated more aggressively toward the beginning of the sequence and as the time complexity increases around the causal loop, convergence is likely to occur as it approaches infinity. Again, this type of behavior is considered a B_1 and is exhibited in Fig. 2. Here, the sample complexity of the generated synthetic data is in response to the time complexity over the sequence of the causal loop. It is also worth mentioning that not only would the Synthetic Data node likely converge in this way, but all of the nodes would likely exhibit a similar behavior.

Limitations: Linearity & Practical Considerations

There are two major limitations in this analysis. The first is that the polarity between corresponding nodes is assumed to be linear. The second is that the sample complexity of the synthetic data would continue in an infinite loop.

The first limitation arises by introducing inductive bias for reducing prediction errors. Introducing *small* amounts of bias to reduce prediction errors does often work in practice. Yet, the prevailing and well accepted machine learning theory states that introducing *too* much bias into prediction systems will often increase its variance, therefore increasing prediction error (1). In other words, the relationship between bias and variance is not often linear. However, it can be modeled as such, if it is assumed that the bias-variance trade off can be optimized to an approximate minima, where the causal loop would terminate.

The second limitation suggested by the diagrams is that time complexity would continue to increase beyond what would be useful for reducing prediction errors. Practically, this would not likely be the case. Generating more synthetic data would likely be halted when it was beyond what was useful in reducing prediction errors. However, it is worth mentioning that if the system were to persist with increased time complexity into infinity, we would perhaps still observe a reduction in prediction error - although a higher one than what was observed with lower time and sample complexity near the beginning of the sequence.

Although there are obvious limitations with many more to mention beyond the scope of this analysis, it still stands to reason that prediction errors could be useful inputs into the generation and application of synthetic data in machine learning to improve the quality of our prediction systems across a wide range of domains.

References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R.* Springer, 2013.