Transfer Objective for Multivariate Causal Structure Learning

Rohan Virani
Stanford University

Robathan Harries
Stanford University

Arvind Sridhar
Stanford University

1 Extended Abstract

Among the biggest outstanding challenges in machine learning involves constructing models that generalize to data beyond the training distribution. In the real world, interventions on the variables and assumptions of the model are commonplace: if some set of variables are intervened on, the underlying data generating distribution is no longer the same as the training distribution. Ideally, our model has factorized the information from the training distribution in such a way that, given an intervention, it can quickly update its internal representation to incorporate the new data. This goal motivates the need of a causal structure to underlie the factorization of the training data. Moreover, since causal modules in the real world are typically invariant, it is likely that most of the causal structure need not be updated after an intervention, making the update step easier to compute.

In this paper, we explore the use of meta-learning to learn a model that is capable of detecting causal direction in a trivariate setting, by comparing the likelihood of causal graphs under a transfer distribution. By comparing how quickly learners parametrizing different hypotheses adapt to a transfer distribution (after some sparse change from the training distribution), one can create a training signal to meta-learn the true causal structure. The first theoretical step we make is explaining how and why the Pairwise Binary Transfer Objective originally suggested by Bengio et al. [2019] might successfully be applied to determine the structure of multivariate causal graphs (specifically trivariate cases), by applying the parameter counting argument under certain assumptions. We also explain why the Causal Parent Multivariate Transfer Objective originally suggested by Bengio et al. [2019] has fundamental issues. Concretely, because this method does not assume a directed acyclic graph and learns parents of nodes independently, it learns correlation rather than causation.

To verify our hypotheses and theoretical results, we implement both the Pairwise Binary Objective and Causal Parent Multivariate Transfer Objective and experiment with a range of bivariate and trivariate causal [1]. To our knowledge, this experimentation suite is a novel contribution that has not been done before in the literature. This suite allows us to generate synthetic data from arbitrary multivariate functional causal models (i.e. a trivariate chain structure with the causal connections $0 \rightarrow 1$ and $1 \rightarrow 2$). We use this data generation tool to train both proposed models over multiple bivariate and trivariate graph structures, attempting to learn the correct causal directions. The criteria for a successful model is that it must behave in distinct and predictable ways on different causal graphs, implying the potential for accurate causal structure learning.

[1] `https://github.com/RobathanH/A-Meta-Transfer-Objective-For-Learning-To-Disentangle-Causal-Mechanisms`

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

In the bivariate case, we consider two variables that are either causally linked and independent. In both cases, the Pairwise Binary method successfully identifies causal direction, as expected, given these experiments were also conducted in Bengio et al. [2019]. By contrast, the Causal Parent method only identifies association rather than causal direction, assigning high likelihood to both directions of the causal pairing. In the trivariate case, we find that the Pairwise Bivariate method successfully identifies causal direction in most, but not all, settings. We also disprove our hypothesis that this method assigns likelihood near 50% to independent and confounded variables, as it instead varies within a large range. For independent variables, this result is workable, and may be bypassed using existing independence tests, but it becomes problematic when confounders also exhibit this behavior. Further, we find that the Causal Parent method correctly distinguishes directly connected variables from independent nodes, confounded connections, and mediated connections in most cases. However, for reasons unknown, Causal Parent is not able to distinguish direct from independent pairings in the trivariate collider (two variables causally impacting a third) case.

Overall, we comprehensively characterize, both theoretically and empirically, two potential methods for identifying causal direction in trivariate and multivariate graphs: Pairwise Binary and Causal Parent. Our experiments with synthetic data showcase the strengths and limitations of each method for various trivariate graph structures, motivating future work to further improve on these methods and develop algorithms to effectively learn the underlying causal structure of real-world datasets.

2 Introduction and Background

The need for machine learning models to learn the underlying causal structure of training data has been well-documented [Bengio et al., 2019], as it is an essential component for creating models robust to realistic distribution changes, and agents which can interact with and plan over the huge variety of real-world situations. In order to take into account interventions on variables while planning, it must be possible to imagine a change to the joint distribution of all variables due to an intervention that may not necessarily have been observed. This requires a model richer than pure correlation but grounded in causal learning and causal reasoning.

Historically, randomized controlled trials would be conducted to isolate cause and effect from confounding variables. However, we also aim to learn causal structure from purely observational data. Current research for tackling this problem is broadly split into traditional constraint and score based methods. Constraint based methods assume a conditional independence oracle and can identify a causal graph up to a Markov equivalence class. The best known method for achieving this is the PC Algorithm by Le et al. [2019]. Score based methods focus on testing graph structures by their ability to fit the data, with one of the most common methods being the Bayesian Information Criterion by Peters et al. [2014].

Applying meta learning to the problem of causal discovery is a new research direction but there has been some work in this area. Dasgupta et al. [2019] demonstrated that a recurrent neural network trained with model-free meta reinforcement learning can solve a range of problems containing causal structure. Bengio et al. [2019] released a paper that applied meta learning to learning causal direction for the first time. The key concept in this paper was that given two competing causal models, one can define a meta learning objective that measures their speed of adaptation to a transfer distribution. The best causal model was shown to adapt the fastest. This approach interestingly takes what is usually a problem in machine learning (changes in the training distribution) and turns it into a training signal to find the optimal causal structure. There is one other paper on meta learning causal mechanisms entitled "Meta Learning Causal Direction" by Ton et al. [2021]. This paper focuses on scenarios with less data and uses a combination of methods in reproducing kernel hilbert spaces and conditional mean embeddings.

The critical issue with the above papers however is their methods only work in the case of two variables A and B, and determining bivariate causal directions. However this is not sufficient to have a truly accurate causal model. If an agent wants to check a localized change in the distribution due to some new data from variables A or B, it is not necessarily true that the causal structure is that A causes B or B causes A. Instead, there may be a third variable C that is a confounder causing both, and A and B are conditionally independent given C.

A situation in which simply bivariate causal discovery breaks down is the following: consider A = **Ice Cream Sales**, B = **Shark Attacks** and C = **Weather**. The true causal graph underlying these variables is **C** -> **A** and **B** -> **A**. It is possible that in the initial training dataset, there is insufficient data on these three variables to conclude that C causes A or B, so we end up concluding A causes B or B causes A since these events are correlated. Upon incoming data from a transfer distribution, we would like to delete the edge between A and B and see if we can learn the true causal edges. Therefore it is critical that we explore if meta learning can be applied to causal structure learning in trivariate and multivariate scenarios. To the best of our knowledge, this does not exist in the literature.

Our approach builds mainly on the work of Bengio by considering a pairwise extension of their binary transfer objective. We also implement for the first time a method that scales in the multivariate setting that doesn't evaluate whole causal graphs but learns the parents of each node in parallel, which is more computationally tractable.

3 Causal Graph Terminology

We begin by defining a graph and its related notation Peters et al. [2017]. Suppose for a given dataset, there exist a finite number of features $\mathbf{X} = (X_1, ..., X_n)$. We define the number of features by an index set $\mathbf{D} = (1, ..., n)$. Then a graph $\mathbf{G} = (\mathbf{D}, \epsilon)$ is composed of both nodes $\mathbf{D}$ and edges ϵ. We say that ϵ is a subset of $\mathbf{D}^2$ such that for any $d \in \mathbf{D}$, $(d, d) \notin \epsilon$.

A parent of a node k is any l such that $(l, k) \in \epsilon$ and $(k, l) \notin \epsilon$. A child of a node k is any l such that $(k, l) \in \epsilon$ and $(l, k) \notin \epsilon$. Two nodes are denoted adjacent if both $(k, l) \in \epsilon$ and $(l, k) \in \epsilon$. We call three nodes an immorality if there are two nodes themselves not adjacent that are parents of another node. We call an edge undirected if both $(k, l) \in \epsilon$ and $(l, k) \in \epsilon$. Any edge which is not undirected is directed. The skeleton of a graph is a set of nodes and edges that does not include any directionality i.e. the graph $(\mathbf{D}, \varepsilon)$ where $(k, l) \in \varepsilon$ if either $(k, l) \in \epsilon$ or $(l, k) \in \epsilon$. A directed acyclic graph (DAG) is where there are no directed cycles i.e. for any couple of nodes (k, l) there is no directed path from k to l and from l to k, and all edges are directed.

4 Data Generation

Let us define a causal graph G with m variables with nodes V_i and edges E_i. This leads to $O(2^{m^2})$ potential causal graphs since each node V_j may or may not be a direct cause of another node V_i, thus there are m^2 decisions that can be taken to construct the causal graph. The ground truth causal graph is always a directed acyclic graph in this paper, however this information is not available to the agent learning causal structures (i.e. no properties of the graph beyond directed edges are assumed). For each variable V_j, we define the probability that it is a cause of variable V_i as p_{ij}. We then define an adjacency matrix **B** such that each entry B_{ij} is a sample from the distribution $Bernoulli(p_{ij})$. Using the graph terminology from the previous section, we can therefore say that the parents of V_i are the variables V_j such that $B_{ij} = 1$. In order to enforce the directed acyclicity of **B**, we enforce that B is lower triangular. This results in the following condition: a variable V_j cannot be a parent of variable V_i if j>i. In the trivariate scenario that we mainly focus on in this paper, the following matrices **B** would be used:

Having defined the adjacency matrix, we then approach generating samples using the following method. Each variable V_i is drawn from a categorical distribution where Vi can take N possible values. This categorical distribution is parameterised as a 2 layer neural network:

$$V_i = f_i(\theta_i, B, V) \tag{1}$$

Essentially, f_i is a multi-layer perceptron that is an independent mechanism of variable V_i, determining the conditional probability of the N possible choices for V_i given the parents V_j which are defined by the matrix **B**. We loop through all V_i starting with $i = 0$. For each variable, we mask its parents using the matrix **B** thus setting all V_j where $B_{ij} = 0$ to 0. We then pass this input through the 2 layer neural network f_i. The input is M concatenated N-dimensional one-hot vectors since V_j is expressed as a one hot vector over the N possible values it can take. The multi layer perceptron has on hidden layer of size $4M$ where M is the number of total variables, a ReLU non-linear activation function and a final layer of size N. We pass the final layer through a softmax activation function to represent the probabilites of each of the N values, and apply a one hot encoding to this vector.

Given the lower triangular condition applied to B and the fact that we iteratively generate the samples for each V_i, any potential parent of V_i, i.e. V_j for $0 <= j < i$ has been generated before we attempt to sample V_i. We enforce the samples generated for each training loop to be 1024 and the categorical distribution is parameterized by 10 (i.e. V_i can take 10 possible values at any point). The above method is sufficient for generating the initial training data to pretrain the neural networks before attempting any meta-learning due to transfer distributions.

In order to define a transfer distribution, we perform an intervention of the root nodes of the causal graph. A root node of the causal graph is any node V_i for which the sum of $B_{ij} = 0$ for all j. We perform this intervention on the root nodes by changing the categorical distribution from which they are sampled by re-initializing the parameters θ_i of the the neural network f_i. We can then resample all other variables in the causal graph.

5 Pairwise Extension of the Binary Transfer Objective

5.1 Algorithm

This method builds on the original binary transfer objective defined by Bengio et al. [2019], applying it on each pair of variables in a multivariate causal graph. For each pair of variables in the causal graph, this approach compares two possible causal hypotheses, each with a different conditional factorization of sample likelihood:

$$\begin{aligned} P_{A\to B}(A,B) &= P_{A\to B}(A)P_{A\to B}(B \mid A) \\ P_{B\to A}(A,B) &= P_{B\to A}(B)P_{B\to A}(A \mid B) \end{aligned}$$

The binary transfer objective approach predicts causal direction using a Structure Model, which meta-learns the likelihoods that each causal hypothesis is true. The Structure Model contains two Hypothesis Models, each outputting sample likelihood by assuming one causal hypothesis, storing trainable parameters for that hypothesis's sample likelihood factorization. For the case of categorical variables with N possible values, the Hypothesis Model $P_{A\to B}$ stores a parameter table $P_{A\to B}(A;\theta_{A\to B})$ containing N probabilities, and a parameter table $P_{A\to B}(B \mid A;\theta_{A\to B})$ containing N^2 probabilities.

Both Hypothesis Models are pretrained to maximize likelihood on the original distribution (prior to intervention on the root nodes of the true causal graph). The algorithm then fits these pretrained hypothesis models repeatedly to different transfer distributions (each with a new intervention on root nodes). By comparing how quickly each Hypothesis Model adapts to transfer distributions, the Structural Model can predict the causal relationship between the pair of graph nodes.

The Structure Model uses meta-learning to perform this hypothesis comparison, updating a structure parameter after each transfer episode to maximize expected adaptation speed by minimizing the following regret function:

$$\begin{aligned} R &= -\log \mathbb{E}[\mathcal{L}] \\ R &= -\log\left(\sigma(\gamma)\mathcal{L}_{A\to B} + (1-\sigma(\gamma))\mathcal{L}_{B\to A}\right) \end{aligned}$$

γ refers to the meta-learned structure parameter which, after a sigmoid activation, defines the likelihood that the true causal direction is $A \to B$.

$\mathcal{L}$ refers to online likelihood, the product of predicted sample probabilities over the course of Hypothesis Model training, which conveys the model's adaptation speed:

$$\begin{aligned}\mathcal{L}_{A\to B} &= \prod_t P_{A\to B}(X^{(t)}; \theta^{(t)}_{A\to B}) \\ \mathcal{L}_{B\to A} &= \prod_t P_{B\to A}(X^{(t)}; \theta^{(t)}_{B\to A})\end{aligned}$$

$P_{A\to B}$ gives the predicted likelihood of transfer samples $X^{(t)}$ as parameterized by Hypothesis Model $A \to B$ at inner training step t. Transfer samples $X^{(t)}$ are resampled from the transfer distribution at each inner training step t. In the original binary case from Bengio et al. [2019], $X^{(t)}$ contains only two variables, A and B. In the multivariate case, $X^{(t)}$ contains an arbitrary number of variables, and a Structure Model is created for each pair of variables, extracting only those two from each $X^{(t)}$ and applying the binary transfer objective to them alone.

By comparing the transfer adaptation speed of opposing binary causal hypotheses, this method estimates the most likely causal direction for each pair of nodes in the graph, though it makes naive assumptions that each node pair is an isolated system with a direct causal link.

5.2 Parameter Counting Argument

Bengio et al. [2019] uses a parameter counting argument to explain this adaptation speed difference, which goes as follows. Given a functional causal model for the case $A \to B$, if both Hypothesis Models train to maximum likelihood on the original distribution, then one model has learned the true values of $P(A)$, $P(B \mid A)$, and the other the true values of $P(B)$, $P(A \mid B)$. In order to simulate an intervention, transfer distributions only change the value of $P(A)$, preserving causal mechanisms. Thus, when adapting to a transfer distribution, the Hypothesis Model for $A \to B$ need only change its value of $P(A)$, whereas the other Hypothesis Model must change both $P(B)$ and $P(A \mid B)$ to maximize likelihood on the transfer distribution. As such, the Hypothesis Model corresponding to the correct causal direction will tend to adapt more quickly to transfer distributions.

5.3 Hypotheses for Multivariate Behavior

We will now extend the above parameter counting argument to justify applying this method to multivariate causal graphs, and predict behavior in the specific trivariate cases of colliders, confounders, and chains. We assume for the below argumentation that each variable is parameterized by a categorical distribution with N possible values. Each marginal distribution contains N marginal probabilities, with $N-1$ adjustable parameters. Each conditional distribution contains N^2 conditional probabilities, with $N(N-1)$ adjustable parameters.

5.3.1 Collider Graph Hypotheses

A trivariate collider causal graph has the form $A \to B$ and $C \to B$. Thus the factorized joint distribution for this graph will be $P(A, B, C) = P(A)P(C)P(B \mid A, C)$, and under transfer distributions only $P(A)$ and $P(C)$ will change.

We hypothesize that the binary transfer objective will successfully detect the causal relationships $A \to B$ and $C \to B$. Since both situations are equivalent, the following argument focuses on $A \to B$ and the two Hypothesis Models $P_{A\to B}$ and $P_{B\to A}$. We consider the effects of each root node intervention one at a time. First, we change $P(A)$, which changes maximum likelihood values of $P(A)$, $P(B)$, and $P(A \mid B)$, but not $P(B \mid A)$. This would allow $P_{A\to B}$ to adapt more quickly than $P_{B\to A}$, following from the proven binary causal graph case. Next, we change $P(C)$, which further changes the maximum likelihood values of $P(A \mid B)$ and $P(B \mid A)$. Though all parameter tables - $P(A)$, $P(B)$, $P(A \mid B)$, $P(B \mid A)$ - will now need to adapt, we propose that the two sources of change are additive, and that changes in $P(C)$ will affect $P(A \mid B)$ and $P(B \mid A)$ evenly, with the result that $P(A \mid B)$ will have changed more extensively than $P(B \mid A)$, preserving the adaptation advantage which allows the binary transfer objective to determine the true causal direction $A \to B$.

We also hypothesize that the binary transfer objective will output even probabilities between the Hypothesis Models $P_{A\to C}$ and $P_{C\to A}$. Since A and C are independent, $P(A \mid C) = P(A)$, and $P(C \mid A) = P(C)$. Thus, when transfer episodes change the root distributions $P(A)$, $P(C)$, the maximum likelihood values for $P(A)$ and $P(A \mid C)$ will change by equivalent amounts, as will those for $P(C)$ and $P(C \mid A)$. As such, Hypothesis Models $P_{A \to C}$ and $P_{C \to A}$ will adapt at very similar rates for every transfer episode, and $\frac{\partial R}{\partial \gamma}$, the structure parameter gradient, will remain near 0, resulting in unchanging hypothesis likelihoods near 0.5. We expect this behavior to occur whenever the binary transfer objective is applied to any pair of independent variables.

5.3.2 Confounder Graph Hypotheses

A confounder causal graph has the form $A \to B$ and $A \to C$. Thus the factorized joint distribution for this graph will be $P(A, B, C) = P(A)P(B \mid A)P(C \mid A)$, and under transfer distributions only $P(A)$ will change.

We hypothesize that the binary transfer objective will successfully detect the causal relationships $A \to B$ and $A \to C$. These cases reduce to the original binary graph case, as the third variable (C or B respectively) has no effect on the distributions of the target variables.

We also hypothesize that the binary transfer objective will output even probabilities between Hypothesis Models $P_{B \to C}$ and $P_{C \to B}$. This would occur because changes in $P(A)$ affect the maximum likelihood values of all Hypothesis Model parameter tables: $P(B)$, $P(C)$, $P(C \mid B)$, $P(B \mid C)$. If all parameter tables change by similar magnitudes, neither Hypothesis Model should have an adaptation advantage. Thus the structure parameter gradient, $\frac{\partial R}{\partial \gamma}$, will be near 0, and the overall Structure Model will output hypothesis likelihoods near 0.5.

5.3.3 Chain Graph Hypotheses

A chain causal graph has the form $A \to B$ and $B \to C$. Thus the factorized joint distribution for this graph will be $P(A, B, C) = P(A)P(B \mid A)P(C \mid B)$, and under transfer distributions only $P(A)$ will change.

We hypothesize that the binary transfer objective will successfully detect the causal relationships $A \to B$ and $B \to C$. Changes in $P(A)$ affect the maximum likelihood values for $P(B)$, $P(C)$, $P(A \mid B)$, and $P(B \mid C)$, but not for $P(B \mid A)$ or $P(C \mid B)$. This should provide an adaptation advantage to Hypothesis Models which use those unchanged parameter tables, namely $P_{A \to B}$ and $P_{B \to C}$.

We also hypothesize that the binary transfer objective will incorrectly detect the causal relationship $A \to C$, which in the graph is a causal connection mediated by a third variable. Since the maximum likelihood value of $P(C \mid A)$ remains unchanged by changes to $P(A)$, the hypothesis model $P_{A \to C}$ should have an adaptation advantage over $P_{C \to A}$.

6 Causal Parent Multivariate Transfer Objective

6.1 Algorithm

The previous method compares pairs of nodes and attempts to determine pairwise causal directionality. This method, which is detailed but not tested in appendix F of Bengio et al. [2019], expands the binary transfer objective regret function to the case of arbitrarily large graphs, then factors the problem of determining full causal graphs into independent sub-problems of determining the causal parents for each individual node, vastly reducing the space of hypotheses from $O(2^{m^2})$ possible full causal graphs to $O(2^{m-1})$ possible causal parent sets.

First, we begin with the same high-level regret function from the binary transfer objective, and expand the expectation over the space of full causal graphs, using B as the definition of any particular hypothesis graph, where $B_{ij} = 1$ if node j directly causes node i:

$$\begin{aligned} R &= -\log \mathbb{E}[\mathcal{L}] \\ R &= -\log \sum_{B} P(B) \mathcal{L}_B \end{aligned}$$

Next, we factor online likelihood into a product of output sample probabilities over training (as done in the binary method), and then further factor each full-graph sample probability into the sample probabilities of each graph node conditional on the values of its parents. $P_{B_i}(X^{(t)})$ is a model which computes the probability of node i values in $X_i^{(t)}$, given the values of its causal parents $\{X_j^{(t)} \mid B_{ij} = 1\}$. We can also use this to define a form of online likelihood which considers only one node at a time, $\mathcal{L}_{B_i}$.

$$\begin{aligned}
\mathcal{L}_B &= \prod_t P_B(X^{(t)}; \theta_B^{(t)}) \\
&= \prod_t \prod_i P_{B_i}(X_i^{(t)}; \theta_{B_i}^{(t)}) \\
&= \prod_i \mathcal{L}_{B_i}
\end{aligned}$$

Next, we make the critical assumption that the likelihood of a particular hypothesis graph $P(B)$ can be factorized into independent likelihoods for each node's set of direct causal parents $P(B_i)$. This allows the regret function to be fully separated into independent regret functions for each node and its expected set of parents.

$$\begin{aligned}
P(B) &= \prod_i P(B_i) \\
R &= -\log \sum_B P(B)\mathcal{L}_B \\
R &= -\log \sum_B \prod_i P(B_i)\mathcal{L}_{B_i} \\
R &= -\log \sum_{B_1} \sum_{B_2} ... \sum_{B_M} \prod_i P(B_i)\mathcal{L}_{B_i} \\
R &= -\log \prod_i \left(\sum_{B_i} P(B_i)\mathcal{L}_{B_i} \right) \\
R &= \sum_i -\log \left(\sum_{B_i} P(B_i)\mathcal{L}_{B_i} \right) \\
R &= \sum_i -\log \mathbb{E}_{B_i}[\mathcal{L}_{B_i}] \\
R &= \sum_i R_i
\end{aligned}$$

Finally, we define structure parameters γ_{ij} which smoothly parametrize the expected structure of the full graph as a product of directed edge likelihoods.

$$\begin{aligned}
B_{ij} &\sim \text{Bernoulli}(\sigma(\gamma_{ij})) \\
P(B) &= \prod_{ij} P(B_{ij}) \\
P(B) &= \prod_{ij} \begin{cases} \sigma(\gamma_{ij}) & B_{ij} = 1 \\ 1 - \sigma(\gamma_{ij}) & B_{ij} = 0 \end{cases}
\end{aligned}$$

Since node i's causal parent structure parameters γ_i only affect the value of $P(B_i)$, each node-specific regret function is completely independent. We can meta-learn each node's causal parents in parallel, and each independent regret function need only compute an expectation over $O(2^{M-1})$ possible causal parent hypotheses, rather than $O(2^{M^2})$ full causal graph hypotheses (where M is the number of nodes in the graph). This can be seen by computing the partial derivative of the total regret function with respect to the structure parameters for the causal parent of node i:

$$\begin{aligned}
\frac{\partial R}{\partial \gamma_i} &= -\frac{\sum_{B_i} P(B_i)\mathcal{L}_{B_i} \frac{\partial \log P(B_i)}{\partial \gamma_i}}{\sum_{B_i} P(B_i)\mathcal{L}_{B_i}} \\
\frac{\partial R}{\partial \gamma_i} &= -\frac{\mathbb{E}_{B_i}[\mathcal{L}_{B_i} \frac{\partial \log P(B_i)}{\partial \gamma_i}]}{\mathbb{E}_{B_i}[\mathcal{L}_{B_i}]} \\
\frac{\partial R}{\partial \gamma_i} &= -\frac{\mathbb{E}_{B_i}[\mathcal{L}_{B_i} \frac{\partial}{\partial \gamma_i}(B_i \log \sigma(\gamma_i) + (1 - B_i)\log(1 - \sigma(\gamma_i)))]}{\mathbb{E}_{B_i}[\mathcal{L}_{B_i}]} \\
\frac{\partial R}{\partial \gamma_i} &= -\frac{\mathbb{E}_{B_i}[\mathcal{L}_{B_i}(B_i - \sigma(\gamma_i))]}{\mathbb{E}_{B_i}[\mathcal{L}_{B_i}]}
\end{aligned}$$

As a result, for each node i, a Causal Parent Structure Model meta-learns the structure parameters γ_i by comparing the online likelihood (transfer adaptation speed) of Hypothesis Models which each assume one of the possible sets of causal parents for node i.

Rather than using probability tables as in the binary case, this method parametrizes each Hypothesis Model as a two-layer neural network, with input size $M * N$ (M variables, with a length N one-hot value vector for each) and output size N, containing the logits for a softmax prediction of output node probabilities. To compute node likelihood conditional on some set of causal parents, input samples are masked using the value of B_i, setting the one-hot sample vectors for non-parent nodes (with $B_ij = 0$) to be zeros only.

Since $O(2^{M-1})$ Hypothesis Models is still an intractable number of models to train and compare, Bengio et al. [2019] suggests replacing explicit expectation over all Hypothesis Models with a biased approximation of the expectation, done by sampling and testing a fixed number of graph hypotheses based on the current values of $P(B_i)$, resampling these hypotheses for each transfer episode.

6.2 Hypotheses

We hypothesize that by factoring full-graph structure learning into independent causal-parent structure learning problems, the Causal Parent Multivariate approach will fail to distinguish causal direction between variables, and instead only compute associations between variables.

Specifically, the assumption that $P(B_i)$ is independent from $P(B_k)$ entirely decouples the values of γ_{ik} and γ_{ki}, allowing structure likelihoods for both opposing causal edges to increase in value together.

In the binary transfer objective, opposing causal edge hypotheses are mutually exclusive, with structure likelihoods parametrized as a sigmoid and its complement. The corresponding regret function displays this as an expectation of online likelihood over both opposing Hypothesis Models, and thus updates the structure parameter based on comparative adaptation speed between Hypothesis Models $P_{A \to B}$ and $P_{B \to A}$.

In the Causal Parent Multivariate transfer objective, the regret function for node i contains an expectation of online likelihood for node i over the possible parent sets for node i. As such, structure parameter γ_ij is updated based on the comparative adaptation speed between Hypothesis Models (parent set assignments) which contain the edge $j \to i$ and those that don't. We expect that this approach will successfully determine the set of parent nodes required to best predict the value of node i, regardless of causal direction.

This hypothesis implies that the Causal Parent approach will not be able to distinguish causal direction, but will be able to distinguish direct association from independence (as in the collider graph), association through shared cause (as in the confounder graph), and association through mediator (as in the chain graph).

7 Experiments and Results

For our experiments, we tested both methods, Pairwise Binary Transfer Objective and Causal Parent Multivariate Transfer Objective, over five small graphs cases: Two bivariate cases (independence and dependence), and three trivariate cases (collider, confounder, and chain).

Though we tested our methods over multiple hyperparameters, the following experiments all share one set of hyperparameter values. Each categorical variable has $N = 5$ possible values. The inner learning rate (applied to Hypothesis Models training within each transfer episode) is 0.001, and the outer learning rate (applied to structure parameter training across multiple transfer episodes) is 0.01. Model pretraining included 200 gradient steps, each applied on 1000 datapoints repeatedly sampled from the original distribution. Each transfer episode included 20 gradient steps, each applied on 500 datapoints repeatedly sampled from the transfer distribution. The Causal Parent Multivariate Transfer Objective sampled 20 hypotheses per transfer episode.

7.1 Bivariate: Independence

To begin, we tested both methods on the simplest possible graph. Two completely independent root variables. Interestingly, results do not support our hypothesis that the Pairwise Binary Transfer Objective outputs even likelihoods for pairs of independent variables, as instead it outputs varying unpredictable likelihoods. This behavior is theoretically distinguishable from causal behavior, but still reduces the approach's applicability for distinguishing between causally-connected and independent variables.

Results for the Causal Parent Multivariate Transfer Objective show that it detects independence as a near-50% edge likelihood.

Figure 1 shows examples of these two behaviors, and Appendix A.1 contains the full test results.

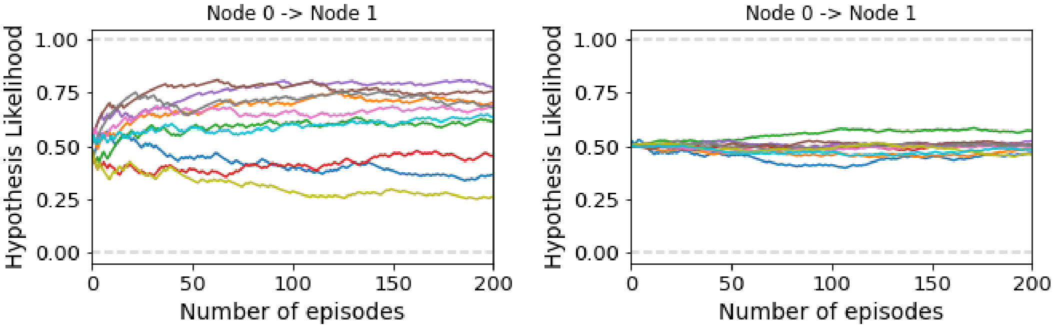


Figure 1: Structure parameter meta-training for Pairwise Binary (left) and Causal Parent Multivariate (right) Transfer Objectives on a graph with two independent variables. Full test results are in Appendix A.1.

7.2 Bivariate: Dependence $(0 \rightarrow 1)$

Next, we tested both methods on the simple binary causal graph tested in Bengio et al. [2019]. Unsurprisingly, the Pairwise Binary Transfer Objective replicates the results from the original paper, correctly determining the true causal direction. On the other hand, the Causal Parent Multivariate Transfer Objective detects 'causation' in both directions, confirming our hypothesis that the causal parent approach measures association without causal direction. Figure 2 shows examples of these behaviors, and Appendix A.2 contains the full experiment results.

7.3 Trivariate: Collider $(0 \rightarrow 2, 1 \rightarrow 2)$

We then extended our causality identification algorithms to the 3 variable (trivariate) case, starting with a collider structure. Figure 3 presents our results, and Appendix A.3 displays them at a larger scale. The Pairwise Binary Transfer algorithm was successfully able to identify the correct direction

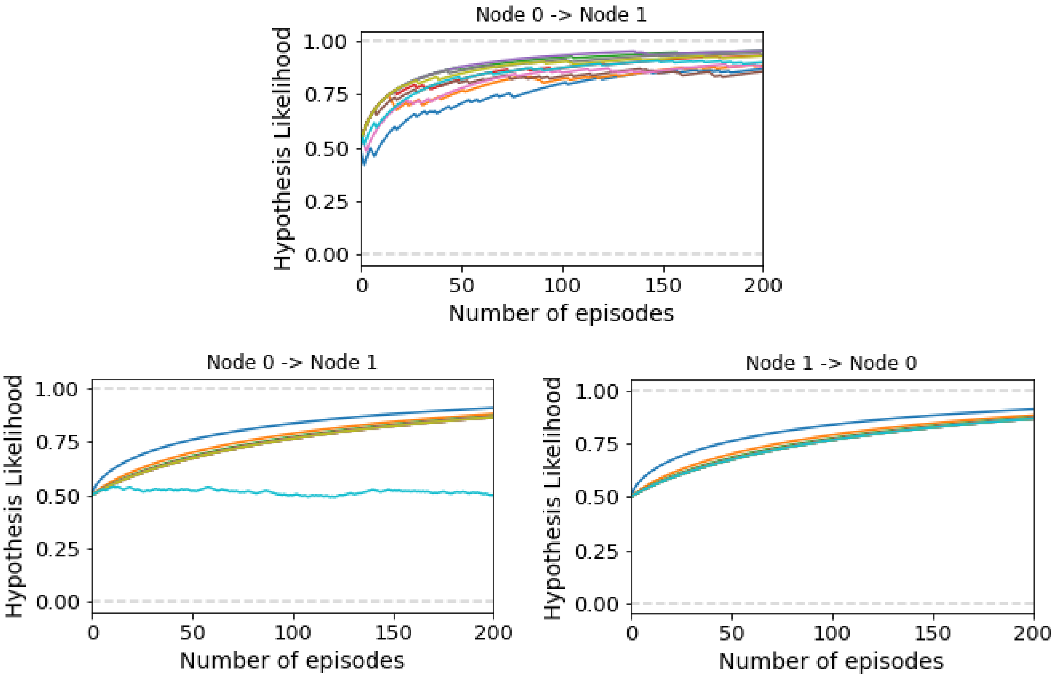


Figure 2: Structure parameter meta-training curves for the Pairwise Binary (top) and Causal Parent Multivariate (bottom) Transfer Objectives on a simple binary causal graph. Full test results are in Appendix A.2.

of causality, assigning high likelihood to the correct causal relationships ($0 \rightarrow 2$ and $1 \rightarrow 2$), supporting part of our hypothesis about detecting existing causal relationships. However, as in the bivariate independent case, it gives unpredictable results somewhat near 50% confidence for independent variables ($0 \leftrightarrow 1$), which is mild evidence against our hypothesis that the pairwise objective will output even probabilities for independent variables, reducing the approach's applicability for distinguishing between causally-connected and independent variables.

On the other hand, the Causal Parent Multivariate Transfer Objective behaves unexpectedly, and cannot successfully distinguish between direct connections ($0 \rightarrow 2$ and $1 \rightarrow 2$) and independence ($0 \leftrightarrow 1$), outputting a high likelihood for all edges. This disproves part of our hypothesis, revealing unexpected shortcomings for this method.

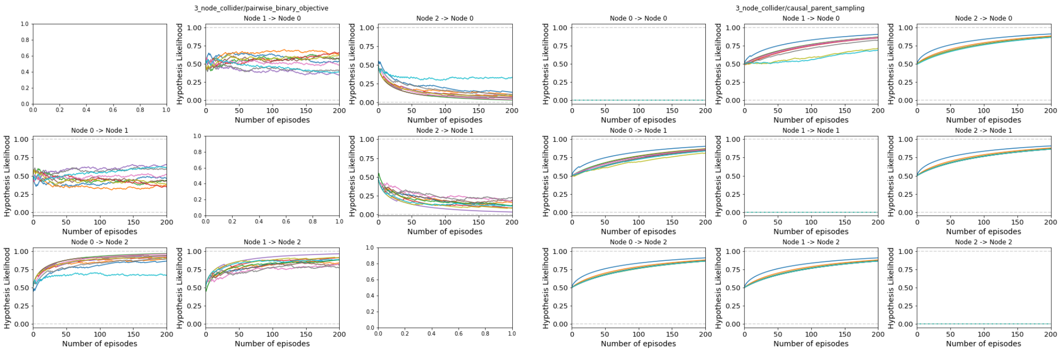


Figure 3: Structure parameter meta-training curves for the Pairwise Binary (left) and Causal Parent (right) objectives on a trivariate collider graph. The Pairwise Binary algorithm correctly identifies the correct causal direction of ($0 \rightarrow 2$) and ($1 \rightarrow 2$), assigning high likelihood to only those directions and fluctuating around 0.5 for the independent pair ($0 \rightarrow 1$). By contrast, the Causal Parent algorithm does not even identify association correctly, assigning high likelihood to all pairings. Larger versions of these results are in Appendix A.3.

7.4 Trivariate: Confounder $(0 \to 1, 0 \to 2)$

Results for the confounder graph for the Pairwise Binary Transfer Objective do not support our hypothesis that it stays balanced for confounded connections. Though it can correctly identify causal direction for directly-connected nodes, it gives wildly varying outputs for the pair of nodes which share a single cause. These wildly varying results are likely due to imbalances in how the transfer changes in $P(0)$ impact the maximum likelihood distributions of $P(1)$, $P(2)$, $P(1 \mid 2)$, and $P(2 \mid 1)$. This confounded output is clearly distinguishable from true causal output when tested over multiple instantiations of the graph functions (randomizing not just root node distributions, but conditional edge distributions). Though we do this for our tests, it is not possible to randomize every causal mechanism in reality, and as such the pairwise method may identify confounded connections as weak causality in either direction.

Results for the Causal Parent Multivariate Transfer Objective support our hypothesis that it can distinguish direct connections from confounded connections. The method converges towards 100% edge likelihood for nodes which have a direct connection $(0 \leftrightarrow 1,\ 0 \leftrightarrow 2)$, and behaves clearly differently for the confounded connection $(1 \leftrightarrow 2)$, staying stable at 50% edge likelihood.

Figure 4 displays test results for this graph, with larger versions in Appendix A.4.

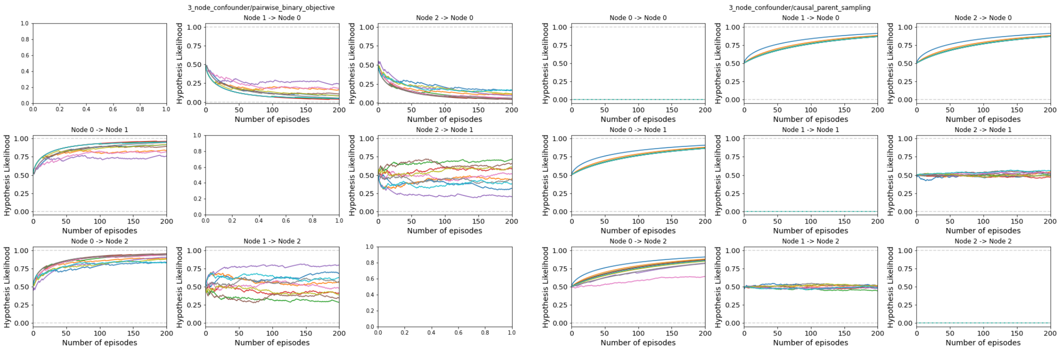


Figure 4: Structure parameter meta-training curves for the Pairwise Binary (left) and Causal Parent Multivariate (right) Transfer Objectives on a trivariate confounder graph. The pairwise method can detect causal direction, but is unreliable for confounded variables. The causal parent method cannot determine causal direction, but can identify confounded variables. Larger versions of test results are in Appendix A.4.

7.5 Trivariate: Chain $(0 \to 1 \to 2)$

On the chain graph (Figure 5), the Pairwise Binary Transfer Objective fails to identify the correct direction of causation, contrary to our hypothesis and unexpected given the results with the other trivariate graphs. Supporting our expectations, it assigns high likelihood to the correct $(0 \to 1)$ pairing as well as the incorrect mediated connection $(0 \to 2)$. It differs from our expectations in its outputs for the $(1 \to 2)$ connection, which vary wildly, similar to independent and confounded cases. This case represents an unexpected limitation of the Pairwise Binary Transfer Objective.

The Causal Parent Multivariate Transfer Objective, similar to the trivariate confounder case, distinguishes direct association from mediated association, assigning high likelihood to direct connections and stabilizing around 50% for the mediated connection $(0 \to 2)$. However, it again fails to detect causal direction, assigning high likelihood to both directions of the correct pairings. Larger versions of our test results are in Appendix A.5.

8 Discussion and Future Work

There are multiple opportunities for further exploration of algorithms for the trivariate case. We initially wanted to incorporate an independence hypothesis into the set of hypotheses over the pairwise binary objective. In the current formulation, we only compare two causal directions by comparing their respective model decompositions. Due to Bengio's parameter counting argument, this method

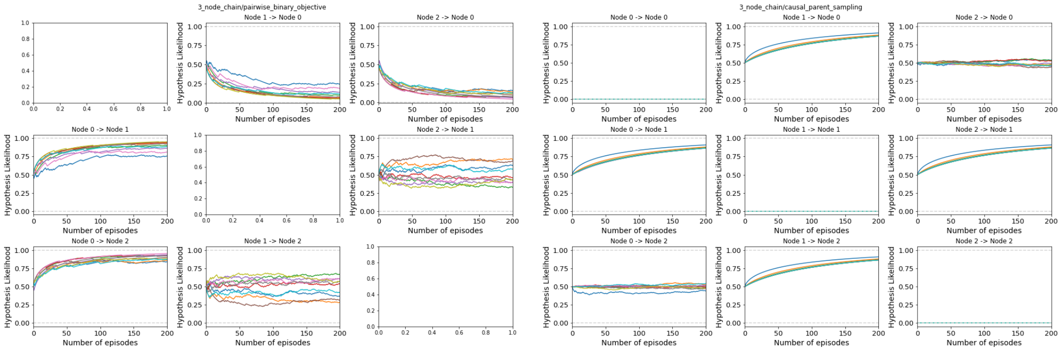


Figure 5: Structure parameter meta-training curves for the Pairwise Binary (left) and Causal Parent Multivariate (right) Transfer Objectives on a trivariate chain graph. The Pairwise Binary algorithm fails to detect the correct causal direction ($0 \rightarrow 1$ and $1 \rightarrow 2$), instead assigning high likelihood to the $0 \rightarrow 2$ pairing and fluctuating wildly with respect to the $1 \rightarrow 2$ pairing. As with the confounder case, Causal Parent correctly detects association but fails to detect causation, assigning high likelihood to both directions of the correct pairings. Larger versions of test results are in Appendix A.5.

works since one model $O(N^2)$ time to retrain while the other takes $O(N)$ if parameterized as categorical distributions with N possible values. than the other model. However, this would not be the case if we factored $P(A, B)$ as $P(A)P(B)$ and added it as a third hypothesis. We assume without loss of generality that A causes B is the true causal model and perform interventions on A for the transfer distribution. Thus for both the factorizations $P(A)P(B|A)$ and $P(A)P(B)$, only $P(A)$ has to be relearned under the transfer distribution so there is no noticeable difference in adaptation speed (they both converge in $O(N)$ time).

The second option we considered was intractable in the trivariate case. With pairwise causal directions, there are only three potential cases, A causes B, B causes A or A and B are independent. However, in the trivariate case, you have 27 candidate models because each node pair has 3 potential relationships. This is not a scalable approach to tackling multivariate causal structure learning as for a graph with M variables, one would have to build $O(M^3)$ models to compare full causal graphs, and therefore we did not consider it applicable in this paper.

The most important takeaway from this paper is that the Causal Parent Multivariate Transfer Objective, which was hypothesized to work by Bengio, does not work. The reason for this is that, since each node learns its parents independently, a directed acyclic graph is not assumed! Therefore the approach can simultaneously assign high likelihood to both causal directions, since the two hypotheses are never directly compared.

There are many potential alterations to the Causal Parent Multivariate Transfer Objective which remove the mistaken independence assumption. One can add regularization terms which penalize assigning high likelihood to both opposing causal directions; one can add constraints which limit the sum of opposing causal direction likelihoods to 1, projecting structure parameter gradients onto that constraint surface; One can even directly replace sigmoid structure parameters for edge likelihood with three-channel softmax structure parameters, parametrizing for each node pair the likelihood of forward edge, backward edge, and no edge. All of these potential methods attempt to reintroduce the competition between opposing causal hypotheses, but by breaking the original independence assumption they implicitly expand the space of considered hypotheses from the comparatively feasible $O(2^{M-1})$ space of causal parent hypotheses to the incredibly non-feasible $O(2^{M^2})$ space of possible full causal graph hypotheses. Nonetheless, these approaches are important avenues for future research into causal structure discovery algorithms.

Based on the general failure modes of both methods tested in this paper, another promising potential algorithm could use both methods in concert. Apart from its puzzling behavior on the trivariate collider case, the Causal Parent Multivariate Transfer Objective can accurately distinguish between node pairs with some direct causal connection and node pairs without. It cannot, however, determine the causal direction of any direct edge. Conversely, the Pairwise Binary Transfer Objective can, apart from a notable exception on the trivariate chain graph, accurately determine the causal direction for existing edges, but gives inconsistent outputs for independent and confounded node pairs. If further work can remove or account for the inconsistent failure cases of each, one could then use the Causal Parent method to determine direct edges, and then apply the Pairwise Binary method to determine the causal direction for each identified direct edge.

Over the course of this project we have implemented categorical data generation for arbitrary multivariate causal graphs, implemented two proposed meta-learning-based multivariate causal structure learning models, and explored their success and failure cases over binary and trivariate causal graphs. Our results indicate that there is still much future work to be done extending meta-learning to multivariate causal structure learning, but also show that the foundational concept of meta-learning structure parameters to maximize transfer adaptation speed holds promise for the goal of robust causal structure learning.

9 Contributions

Over the course of this project, we experienced a few coordination difficulties, and had trouble balancing the workload.

As of the project milestone, Rohan Virani and Robathan Harries were responsible for derivation of Bengio's theorem for the trivariate case; Robathan Harries was responsible for creating an environment to produce synthetic data consistent with a given causal structure and parametrization, as well as creating the base interface of the model itself, for use testing different internal models. All three group members were responsible for running experiments on different causal structures and types of synthetic data, and analyzing the effects of changing particular test parameters.

As of the final submission, detailed contributions are as follows:

- Rohan Virani
  - Wrote the Project Proposal
  - Identified and analyzed the multivariate structure models to implement and test (with Robathan Harries)
  - Completed a detailed literature review of relevant causal structure learning methods and their overlap with our planned approach.
  - Created the majority of the Project Poster (60%)
  - Contributed significantly to the Final Report
- Arvind Sridhar
  - Contributed a few sentences to the Project Milestone Report
  - Collected a suite of test results, some of which were used in the Project Poster
  - Contributed to part of the Project Poster (20%)
  - Contributed significantly to the Final Report
- Robathan Harries
  - Identified and analyzed the multivariate structure models to implement and test (with Rohan Virani)
  - Implemented the synthetic data generation algorithm
  - Implemented a framework for testing models on various causal graphs
  - Collected an initial suite of test results for the milestone report
  - Wrote the Milestone Report (with a few sentences added by Arvind Sridhar)
  - Implemented the Pairwise Binary Transfer Objective method
  - Implemented the Causal Parent Multivariate Transfer Objective method
  - Contributed to part of the Project Poster (20%)
  - Ran an intensive suite of tests for the Final Report
  - Contributed significantly to the Final Report

Our project code repository can be found at `https://github.com/RobathanH/A-Meta-Transfer-Objective-For-Learning-To-Disentangle-Causal-Mechanisms`

References

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *CoRR*, abs/1901.10912, 2019. URL `http://arxiv.org/abs/1901.10912`.

Ishita Dasgupta, Jane X. Wang, Silvia Chiappa, Jovana Mitrovic, Pedro A. Ortega, David Raposo, Edward Hughes, Peter W. Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *CoRR*, abs/1901.08162, 2019. URL `http://arxiv.org/abs/1901.08162`.

Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, 16(5):1483–1495, 2019.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schoelkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, JUN 2014.

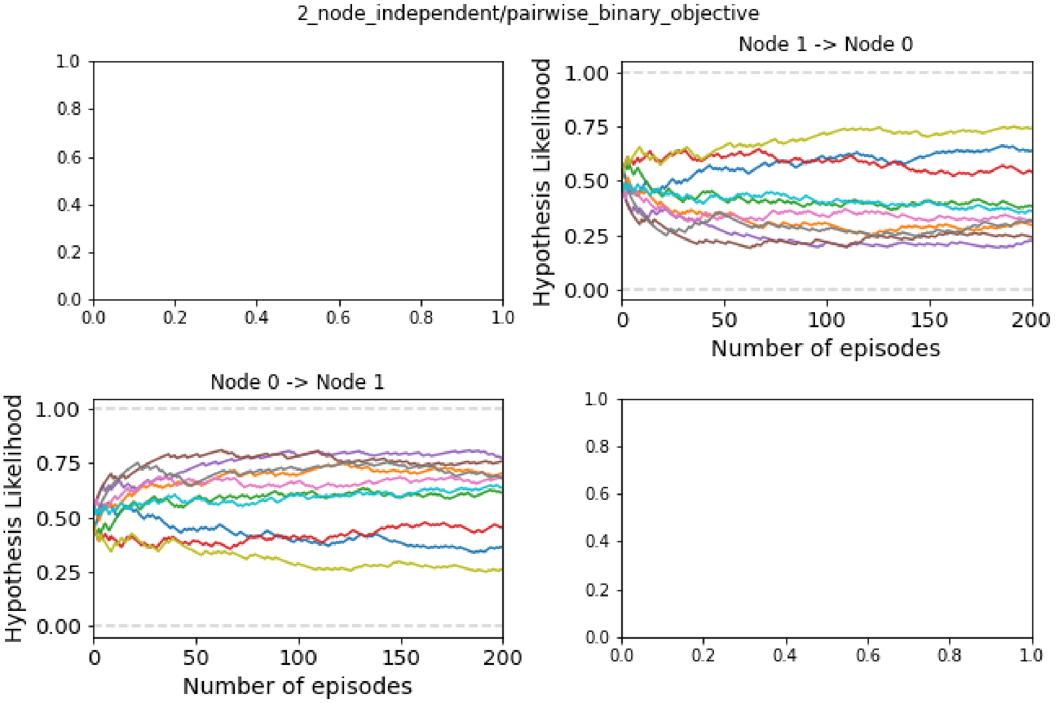
Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

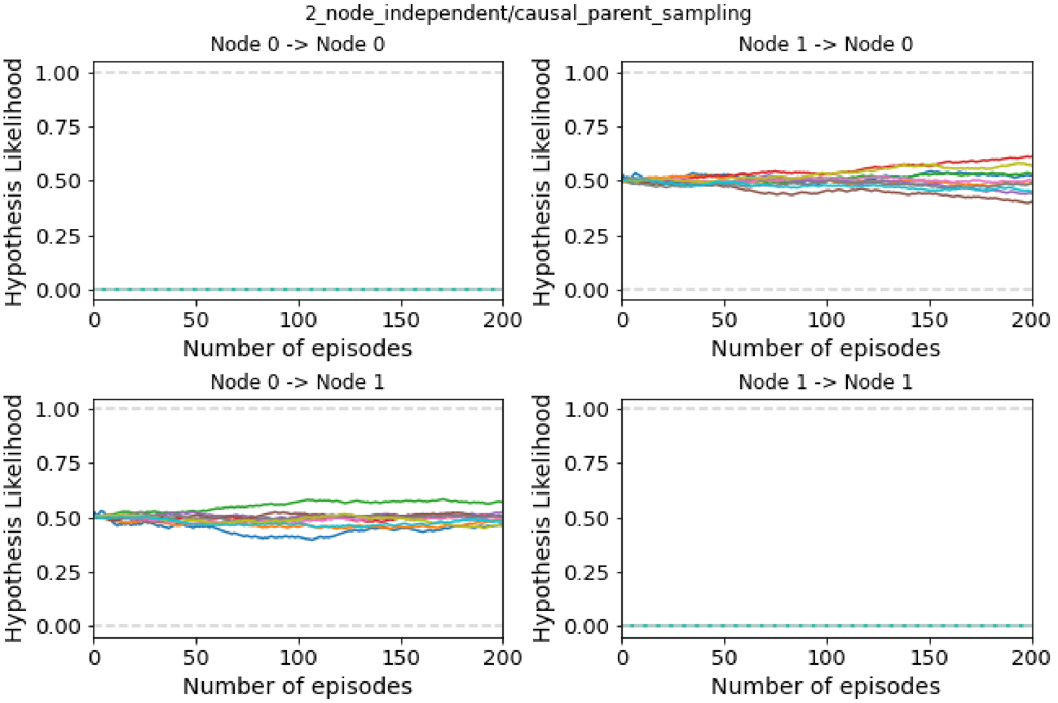
Jean-Francois Ton, Dino Sejdinovic, and Kenji Fukumizu. Meta learning for causal direction, 2021.

A Full Test Result Figures

A.1 Bivariate: Independence

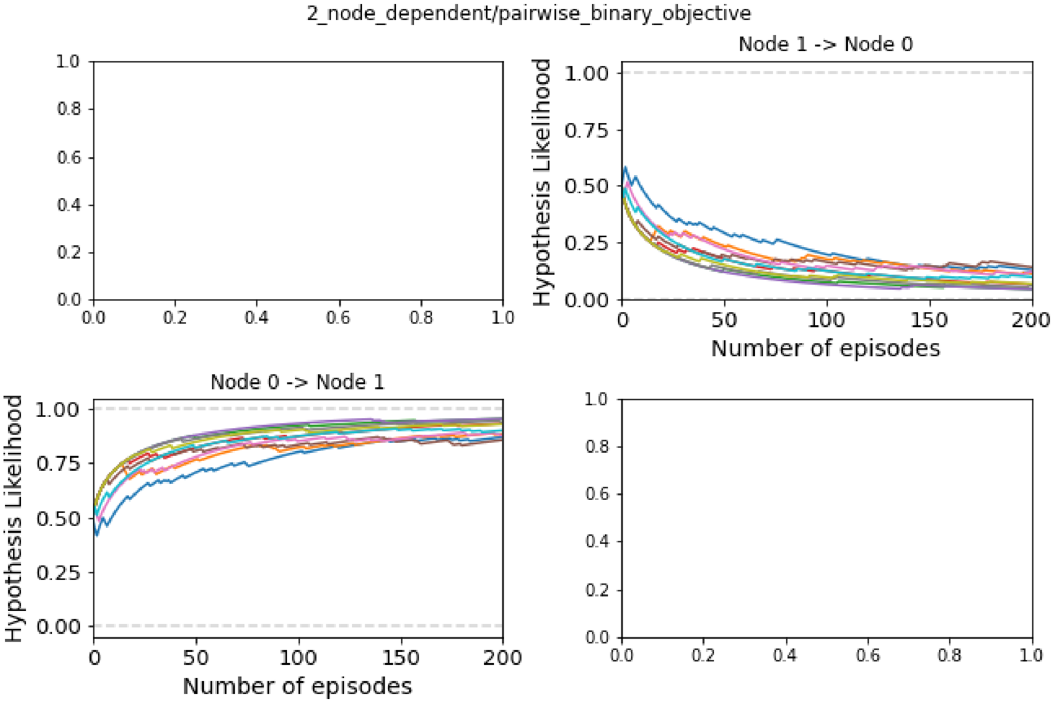
Full results for the bivariate causal graph with two independent variables.
The top figure contains results for Pairwise Binary Transfer Objective, and bottom figure contains results for Causal Parent Multivariate Transfer Objective.

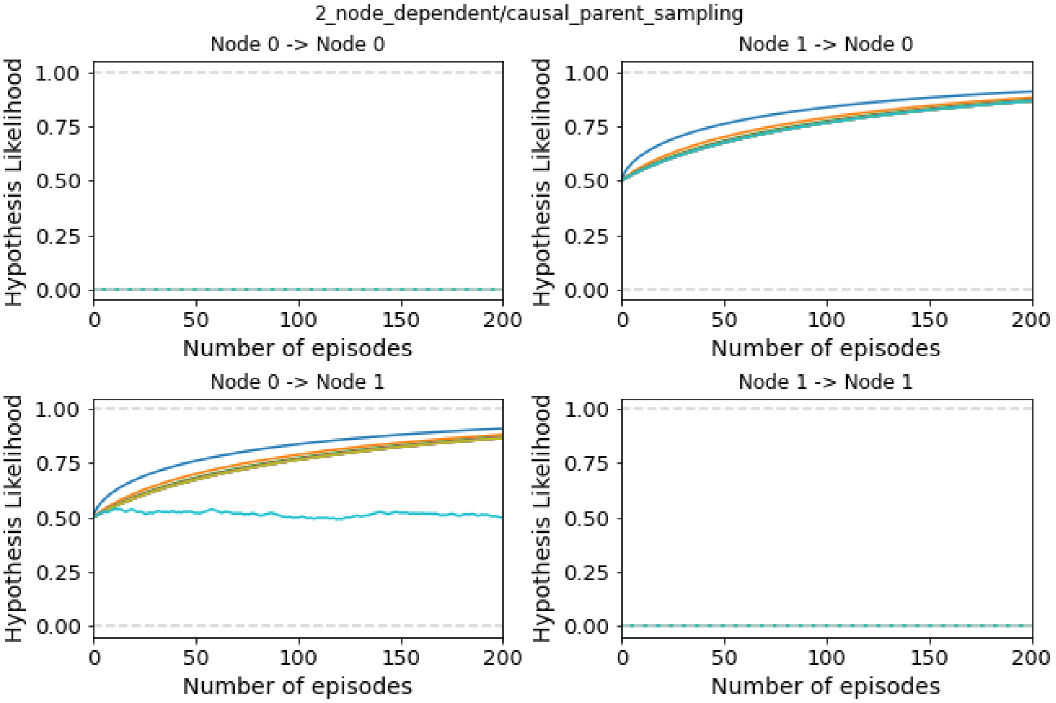




A.2 Bivariate: Dependence $(0 \rightarrow 1)$

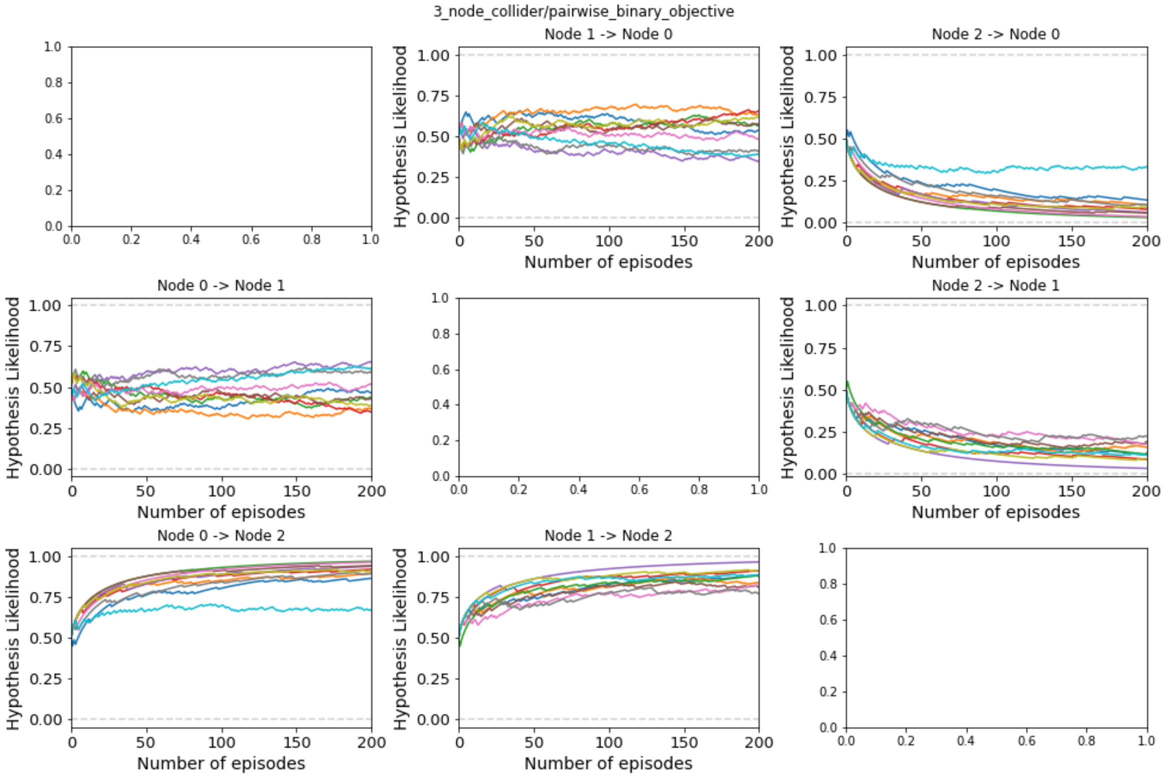
Full results for the bivariate dependence case, in which node *0* causes node *1*.
The top figure contains results for Pairwise Binary Transfer Objective, and bottom figure contains results for Causal Parent Multivariate Transfer Objective.

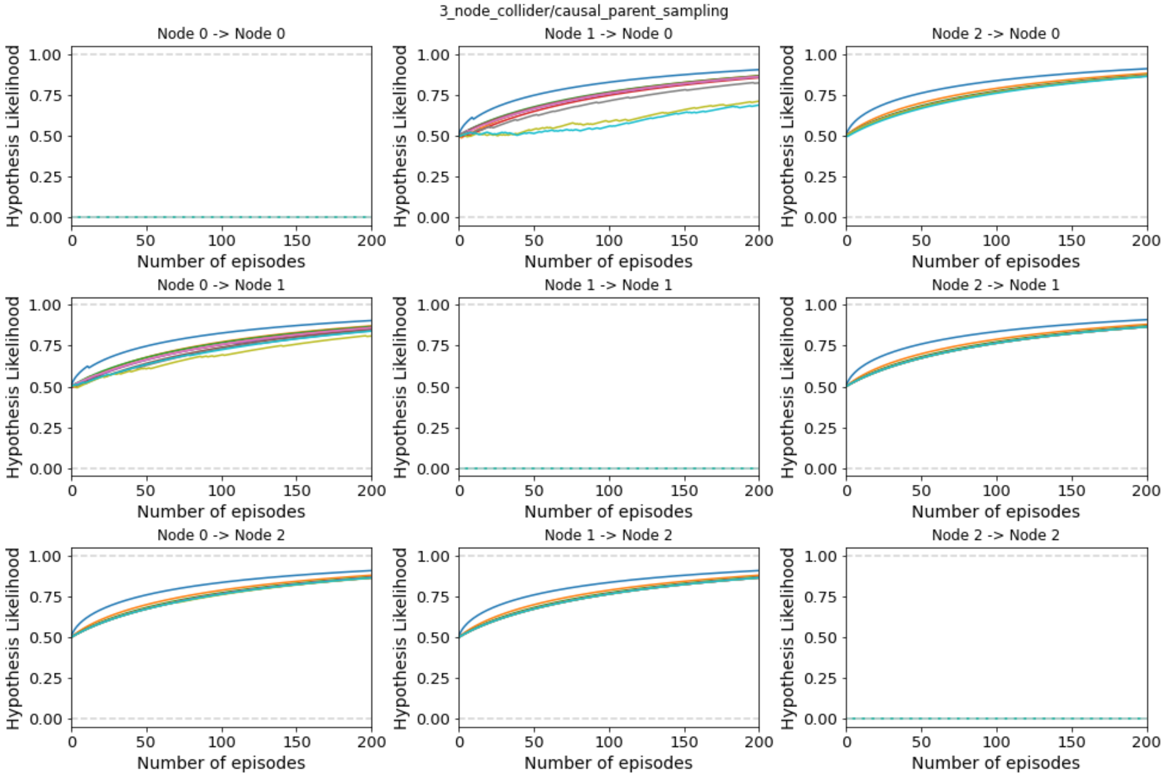




A.3 Trivariate: Collider $(0 \rightarrow 2, 1 \rightarrow 2)$

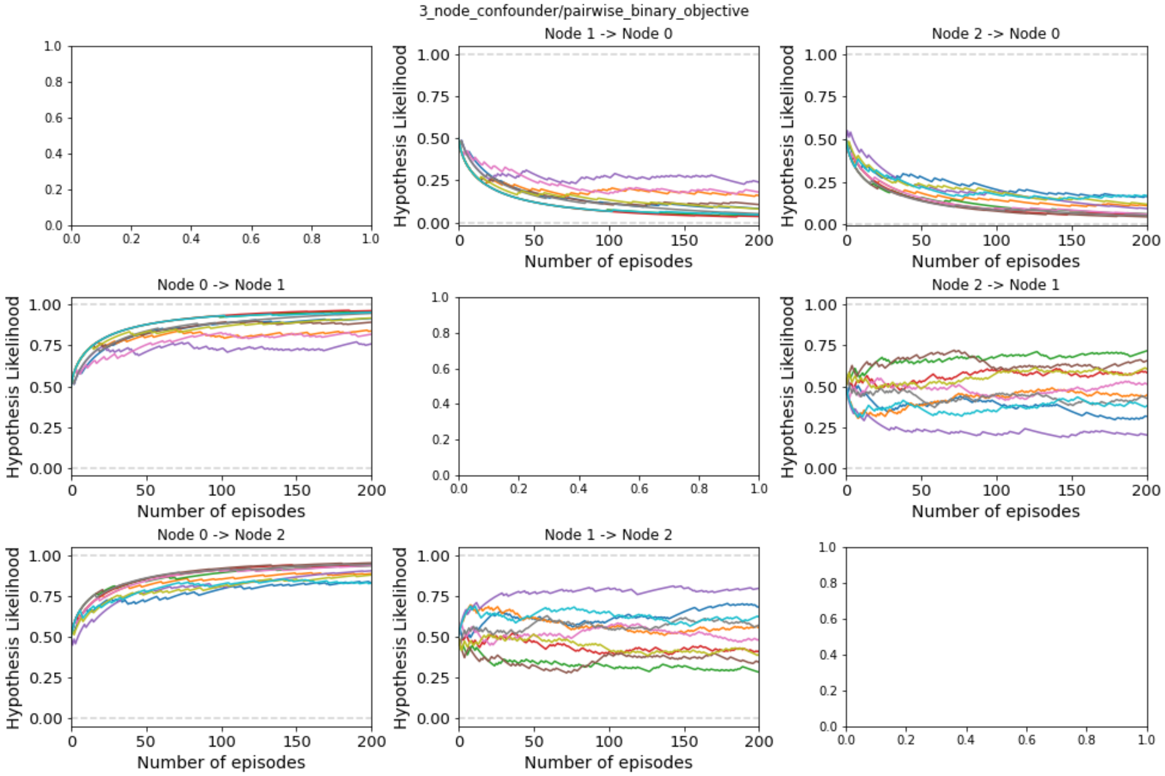
Full results for the trivariate collider case, in which nodes *0* and *1* both cause node *2*.
The top figure contains results for Pairwise Binary Transfer Objective, and bottom figure contains results for Causal Parent Multivariate Transfer Objective.

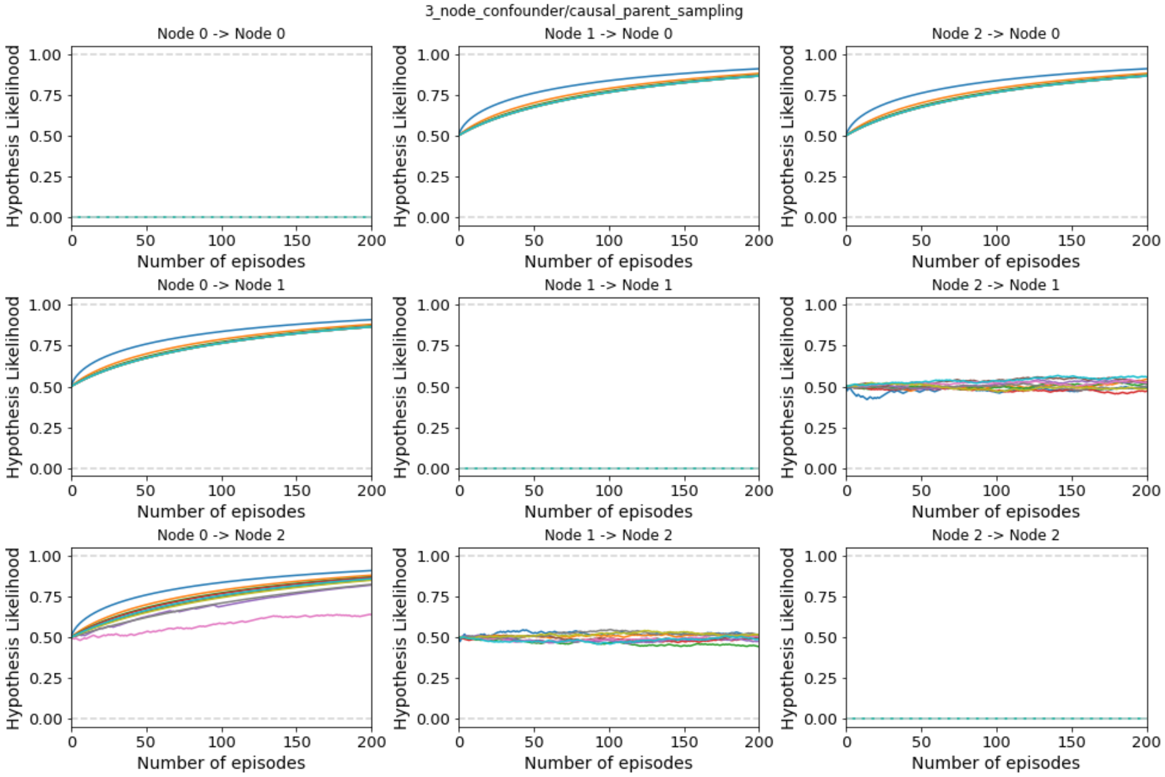




A.4 Trivariate: Confounder ($0 \rightarrow 1$, $0 \rightarrow 2$)

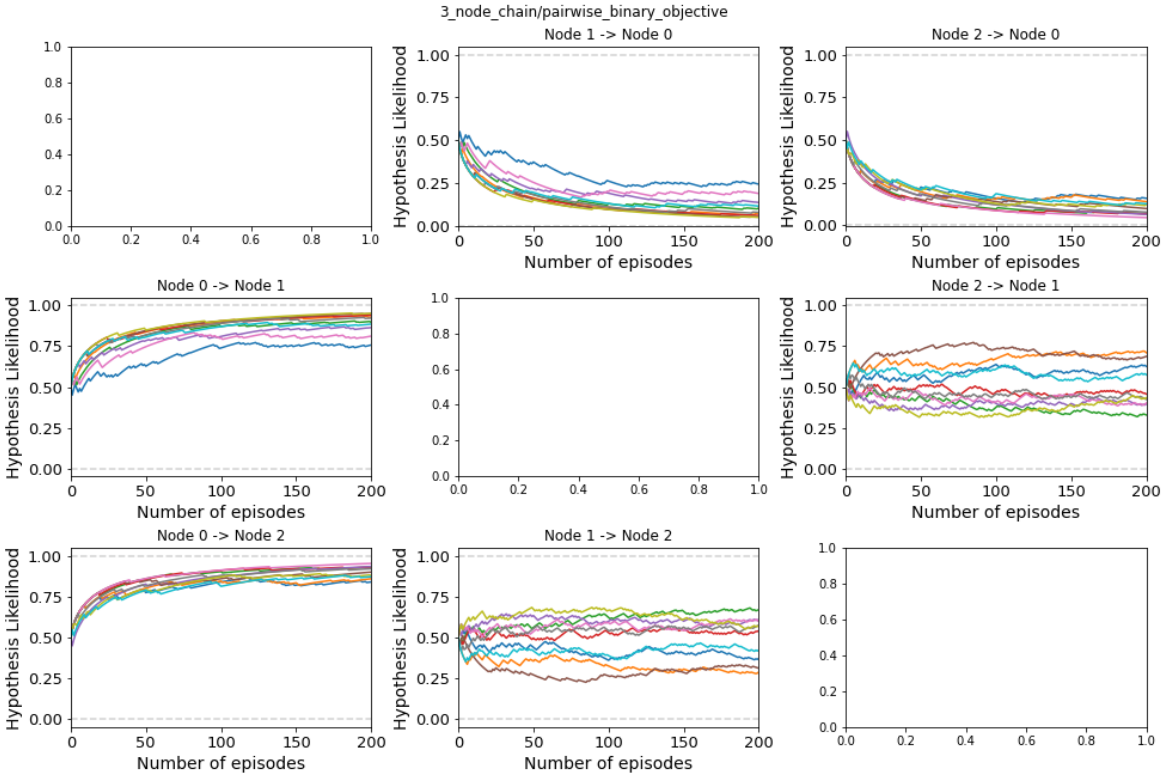
Full results for the trivariate confounder case, in which node *0* causes both node *1* and node *2*. The top figure contains results for Pairwise Binary Transfer Objective, and bottom figure contains results for Causal Parent Multivariate Transfer Objective.

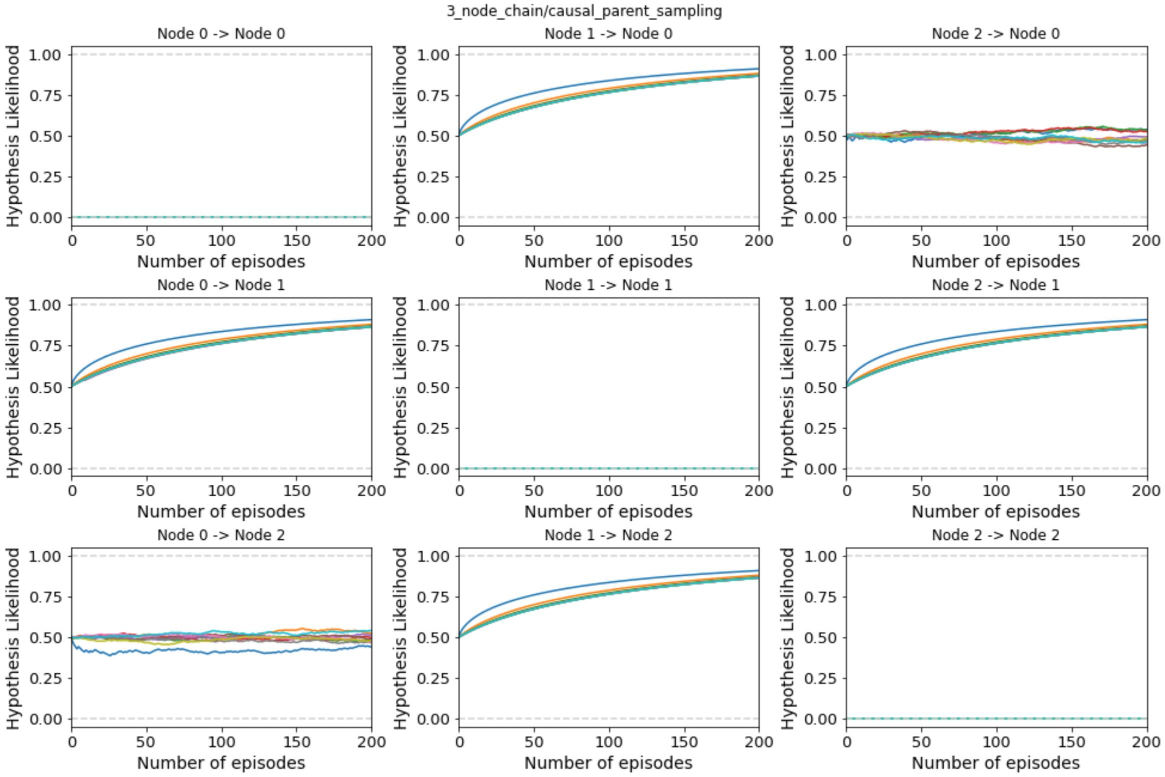




A.5 Trivariate: Chain $(0 \to 1 \to 2)$

Full results for the trivariate chain case, in which nodes *0* causes node *1*, which in turn causes node *2*. The top figure contains results for Pairwise Binary Transfer Objective, and bottom figure contains results for Causal Parent Multivariate Transfer Objective.


3_node_chain/pairwise_binary_objective
Node 1 -> Node 0
Node 2 -> Node 0
Node 0 -> Node 1
Node 2 -> Node 1
Node 0 -> Node 2
Node 1 -> Node 2
Hypothesis Likelihood
Number of episodes


3_node_chain/causal_parent_sampling
Node 0 -> Node 0
Node 1 -> Node 0
Node 2 -> Node 0
Node 0 -> Node 1
Node 1 -> Node 1
Node 2 -> Node 1
Node 0 -> Node 2
Node 1 -> Node 2
Node 2 -> Node 2
Hypothesis Likelihood
Number of episodes