A Meta Learning Approach to Discerning Causal Graph Structure

Justin Wong [*1] **Dominik Damjakob** [*1]

Abstract

We explore the usage of meta-learning to derive the causal direction between variables by extending previous work in Bengio et al. (2019) which focused on learning the causal direction between two variables as well as a simple extension into using rotational encoder-decoder structure in representation learning. We discuss difficulties in extending their encoder-decoder structure and limitations of the approach.

As an improvement, we choose to adopt the Functional Causal Model (FCM) graph representation similar to Goudat et al. (2018). This allows for a direct representation of the variable observations as a function of its causal parents, and we demonstrate improved performance in learning ability for causal direction. Using this method, we are able to extend the methodology to learning causal directions for observed variables given independent hidden effects, a direct extension of the rotational encoder-decoder from Bengio et al. (2019).

A common assumption of related work work often is that causality between variables is known and only direction needs to be determined. However, the question of which direction is correct is badly specified in the case where no causal relationship exists. We further extend the derivation of Bengio et al. (2019) to the related question of whether two variables have a causal relationship or not. This allows us to introduce a secondary meta-parameter that measures the confidence in the existence of a causal relationship between variables versus the two variables being independent of one another.

We continue to improve previous results by introducing the effects of latent confounding effects. While other work such as Goudat et al. (2018) assumes the distribution of these confounders is known and given, we find that it is more amenable to real world applications to model the distribution of latent effects. This is done by assuming the existence of a proxy variable through which we can measure them and so we reformulate the optimization problem in terms of a lower bound for the likelihoods for the latent variables.

We then conduct experiments using synthetic data distributions and measure our model's performance in terms of speed of convergence and magnitude of confidence. We notice that in our complex model using latent confounders, we are able to achieve over ninety percent confidence of the correct causal direction quickly. This outperforms the baseline approach of Bengio et al. (2019) which was found on a much simpler problem involving no latent confounders which converged to less than seventy percent confidence over a similar training span. Similarly, our new meta-parameter measuring confidence of the existence of a causal relationship also can converge in the desired direction.

We then perform a series of robustness tests on our models. First, we demonstrate the robustness of our method when some distributional assumptions are violated. In these cases, the confidence path converges more slowly than otherwise. However, we still are able to recover the causal direction with a high degree of confidence. Next, we consider the effects of a limited dataset as in real-world scenarios on the model convergence. We find that with varying degrees of certainty depending on the size of the dataset, we are still able to find the correct causal relations. These checks demonstrate that despite the ideas and models being developed for ideal situations, they are still applicable for non-ideal settings in practice.

[*]Equal contribution [1]Department of Statistics, Stanford University, Palo Alto, United States. Correspondence to: Justin Wong <juswong@stanford.edu>, Dominik Damjakob <damjakob@stanford.edu>.

Not included in the proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020.

1. Introduction

Background

When it comes to inferring causal direction, the most popular tool is proper trial designs of experiments. More specifically, the randomized control trial (RCT) is a popular tool of choice since it allows us to easily separate treatment results from confounding variables so we are able to retrieve statistics such as average treatment effects that can inform the causality and strength of a given treatment. However, such methods are not globally applicable since randomized trials can not be conducted in many scenarios: for example, they can be too costly, unethical or simply infeasible due to the complexity of real world systems. Furthermore, the RCT only informs a method for some prospective studies and is not at all helpful for retrospective studies - which is a large source of data to analyze. Applications in complex clinical trials or questions in implementation of social, economic and political sciences require more specialized tools to assist in discerning causality in the slew of data generated from less than ideal conditions in the modern computing era.

Machine learning, most notably deep learning, is a powerful tool that has allowed for state-of-the-art performance in both discriminative and generative tasks and has enjoyed huge amounts of growth in recent years as a result. However, canonical learning techniques often are likelihood based optimizations which converge regardless of causal direction. That is, given a joint distribution $p(x, y)$ the formulations $p(x|y)p(y) = p(y|x)p(x)$ are functionally equivalent parameterizations. For most distributions which are well behaved, this will lead to no notable difference in convergence of the learning model irregardless of the true causal direction despite the two parameterizations implicitly defining an opposing causal direction. As such, we require specialized learning methods and importantly, additional assumptions that allow deep learning models to discern the subtle difference between parameterizations.

We begin with the philosophy of Occam's razor - if multiple answers are correct, the best answer is the simplest one. Applied to our problem, this suggests that given the parameterizations $p(x|y)p(y)$ and $p(y|x)p(x)$ we assume that the "true" parameterization is the one that yields the simpler pair of conditional and marginal distributions. Specifically, the measure of simplicity that we will use is the speed at which we are able to adapt a transfer distribution. Thus, under the assumptions, we expect that the correct causal direction will allow for faster transfer learning of the distribution under which we develop our methodology.

Related Work

Bengio et al. (2019) use meta-learning to derive the causal direction between variables in an optimization-based framework. In their work, they apply learned models assuming different causal directions to data with a changed transfer distribution. As the correct causal model will only have to adjust its transfer distribution, and thus adapt faster, this allows it to extract the underlying causal directions. Bengio et al. (2019) further apply this model to the Representation Learning domain, in which information from underlying variables has to be extracted. For this, they only consider a single freedom rotation framework as defined in Bengio et al. (2013), though more general models could provide better generalisation. Such new representation learning models have emerged in recent literature, among them SimCLR (Chen et al., 2020) and Deep Cluster (Caron et al., 2018; Zhan et al., 2020).

Goudat et al. (2018) leverage a series of generative models called Causal Generative Neural Networks (CGNN) in order to model each of the observable states of the graph. This allows it to resample a dataset distribution by sampling in topological order of the graph and using the parents as inputs to the children. Given a starting causal structure, CGNN iteratively refines the direction by an iterative method of resampling and computing a Maximum Mean Discrepancy (MMD) statistic that serves as a heuristic measurement as to the similarity of the ground truth distribution to the current iteration's resampled distribution. It also features a cycle correcting method that allows it to resolve conflicting causal structure with the best orientation by using a non-convex hill climbing technique. The large difference here is that they are using MMD as a heuristic to inform their decision making while we will use meta-learning to optimize a structural meta-parameter to denote the confidence of the direction. Furthermore, CGNN makes distributional assumptions about confounding distributions that we look to generalize through variational inference.

Ton et al. (2018) describe a meta-learning technique that is based on the CGNN technique in Goudat et al. (2018). Noting some issues of CGNN, the authors propose a meta-CGNN method that frames a different meta-learning task to leverage similarity between datasets to learn the causal direction of a new dataset more quickly. Given some cause and effect database where the causal direction is known, they train a resusable generative model to learn to determine the causal direction on the meta-training tasks and attempt to do the same at meta-test time.

Our novel contribution is twofold. First, we introduce a new meta-parameter β to discern the existence of a causal relationship using a similar structure as the α parameter from Bengio et al. (2019). This allows us to frame the casual direction question correctly and protect from miss-specified inferences being concluded. Second, we introduce variational inference techniques inspired from Madras et al. (2018) alongside structural models from Goudat et al. (2018)

to give an explicit representation for latent variables. This is an improvement upon prior works Bengio et al. (2013) and Bengio et al. (2019) since it allows us to have a more direct optimization problem which we will see yields better performance. Furthermore, it generalizes results from Goudat et al. (2018) since we are no longer assuming that we know the exact distribution of latent confounders and instead allow stochasticity by way of inference through a proxy variable. We further study the effects that these changes have on robustness tests such as deviation from distributional assumptions and finite dataset sizes to find that our work extends well into practical applications.

2. Methodology

Meta-Learning Causal Directions

To learn the joint distribution of two variables X and Y we can use their conditional distributions $p_{x|y}$ and $p_{y|x}$ alongside their marginal distributions p_x and p_y. Thus, Bengio et al. (2019) use a Bayesian framework with

$$P_{X\to Y}(X,Y) = P_{X\to Y}(X)P_{X\to Y}(Y|X)$$
$$P_{Y\to X}(X,Y) = P_{Y\to X}(Y)P_{Y\to X}(X|Y)$$

where both parameterizations can be learned by Bayesian networks. As in their experiments $X \to Y$, a training distribution $p_0(x,y) = p_0(x)p(y|x)$ can be used. Thereafter, the distribution is changed to the transfer distribution $p_1(x,y) = p_1(x)p(y|x)$. Both networks are meta-trained to the transfer distribution for T steps with resulting likelihoods

$$L_{X\to Y} = \prod_{t=1}^{T} P_{X\to Y,t}(x_t, y_t)$$
$$L_{Y\to X} = \prod_{t=1}^{T} P_{Y\to X,t}(x_t, y_t)$$

which is trained in the following two step process:

1. The relationship between X and Y is learned using two models: one assumes X causes Y, the other the opposite causal direction.
2. The distribution of X is changed to a transfer distribution. Both models are retrained on the new data and the resulting likelihoods are recorded.

Here, $P_{X\to Y,t}$ denotes the trained Bayesian network after step t. Next, the loss statistic

$$R(\alpha) = -ln\left(\sigma(\alpha)L_{X\to Y} + (1-\sigma(\alpha))L_{Y\to X}\right)$$

is computed with α denoting a structural parameter defining the causal direction and $\sigma(\cdot)$ the sigmoid transformation. In this methodology, α is now optimized to minimize $R(\alpha)$. The loss statistic's gradient is

$$\frac{\partial R}{\partial \alpha} = \sigma(\alpha) - \sigma(\alpha + ln(L_{X\to Y}) - ln(L_{Y\to X}))$$

such that $\frac{\partial R}{\partial \alpha} > 0$ if $L_{X\to Y} < L_{Y\to X}$, that is if $P_{X\to Y}$ is better at explaining the transfer distribution than $P_{Y\to X}$. Bengio et al. (2019) show that if

$$E_{D_{transfer}}[ln(L_{X\to Y})] > E_{D_{transfer}}[ln(L_{Y\to X})]$$

where $D_{transfer}$ is the data drawn from the transfer distribution, stochastic gradient descent on $E_{D_{transfer}}[R]$ will converge to $\sigma(\alpha) = 1$ and $\sigma(\alpha) = 0$ if $E_{D_{transfer}}[ln(L_{X\to Y})] < E_{D_{transfer}}[ln(L_{Y\to X})]$. As the loss function modeling the correct direction - this is, if X causes Y, $L_{X\to Y}$ - only needs to update its estimate for the unconditional distribution $P_{X\to Y}(X)$ from $p_0(x)$ to $p_1(x)$ while the reverse direction networks needs to change both $P_{Y\to X}(Y)$ and $P_{Y\to X}(X|Y)$, it holds that indeed the loss statistic for the correct direction has a lower expected value and we can recover the causal direction.

Existence of Causal Relationships

There is a limitation in the α structural parameter in that it answers the question of which causal direction is more likely. However, such a question is malformed when there is no causal direction. In a fully determined scenario, we may expect that the likelihoods are approximately equal.

To solve this, we introduce a second meta-learning parameter β to learn if there exists a causal relationship. We introduce another training distribution $p_\beta(x,y) = p_\beta(x)p_\beta(y)$. Then we consider the stronger causal relation $P_{causal}(X,Y) = max(P_{X\to Y}(X,Y), P_{Y\to X}(X,Y))$ and the independent marginals $P_{independent}(X,Y) = P(X)P(Y)$. Now we can similarly consider the likelihood

$$L_{independent} = \prod_{t=1}^{T} P_{independent}(X_t, Y_t)$$
$$L_{causal} = max(L_{X\to Y}, L_{Y\to X})$$

and compute the familiar loss statistic $R(\beta) = -[ln(\sigma(\beta))L_{independent} + (1-\sigma(\beta))L_{causal}]$. Similarly we can minimize $R(\beta)$ where $\sigma(\beta) \to 1$ if independence is better at explaining the transfer distribution and $\sigma(\beta) \to 0$ otherwise.

By similar derivation as Bengio et al. (2019), if we have that the $P_{independent}$ is better at explaining the transfer distribution than P_{causal}

$$E_{D_{transfer}}[ln(L_{independent})] > E_{D_{transfer}}[ln(L_{causal})]$$

Suppose that the ground truth is that there is no causality between the variables. Then the two likelihoods will share a marginal likelihood term, but causal likelihood additionally has its conditional probability misspecified while the independent likelihood just needs to update other marginal distribution. The reverse is true when there exists causality between the two variables. Then since the misspecified model averages larger out of sample error, it holds that the loss statistic for the existence or non-existence of the causal direction (whichever is correct) indeed has lower expected value and we can recover this binary decision.

Latent Variables Structure

The previous results make the assumptions that the observed X and Y are independent of other hidden effects. However, this is unlikely to hold in practice. Thus, Bengio et al. (2019) extended their causality direction model to a simple version of confounders, namely a rotation model. Specifically, they believe that instead of the true variables A and B, we can only observe $X, Y = Ro_\theta(A, B)$, where

$$Ro_\theta(A, B) = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}$$

is a rotational decoder parameterized by a single variable θ that rotates the 2 vectors by matrix multiplication with the matrix. To account for this, they introduce the rotational inverse encoder $U, V = Encoder_\phi(X, Y)$ and continue their causal analysis with U and V. Importantly, they train the encoder in the same loop as their Bayesian models and attempt to recover the parameter ϕ as a sub-problem while they learn α.

A notable shortcoming of this approach is that it can not be extended to a more general encoder-decoder structure. The natural extension is to assumes that these two are simple linear models such that

$$Encoder(X, Y) = \begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$$

. However, optimizing over this choice of encoder causes the encoder parameters to converge such that U and V become linearly dependent vectors. This can be seen as some form of reward hacking as this transformation trivially allows for a perfect prediction of V given U. We attempt to resolve this problem by introducing various forms of regularization such as penalization of deviation of the determinant from the desired values or rewarding linear independence. Notice that imposing a hard constraint on linear independence returns a scaled version of the rotational encoder-decoder setup. However, we were not able to pick soft penalization terms strict enough to encourage a non-trivial solution. The problem of choosing such a term is also not obvious and seems not to generalize well for a larger number of variables.

Hence, we deviate from Bengio et al. (2019) by adopting a different graph model that allows for expressing latent variables and effects more explicitly. Noting the success of Goudat et al. (2018) in determining causal relations in graphs, we attempt to adopt a similar representation of variable observations by using Functional Causal Models which suggests that observations are formed as tuples

$$X_i = (\{P_i\}, f_i, \epsilon_i)$$

where i indexes the vertex on the causal graph, $\{P_i\}$ denotes the set of causal parents of X_i, ϵ_i is independent noise modelling latent effects on X_i, and f_i is a learned function.

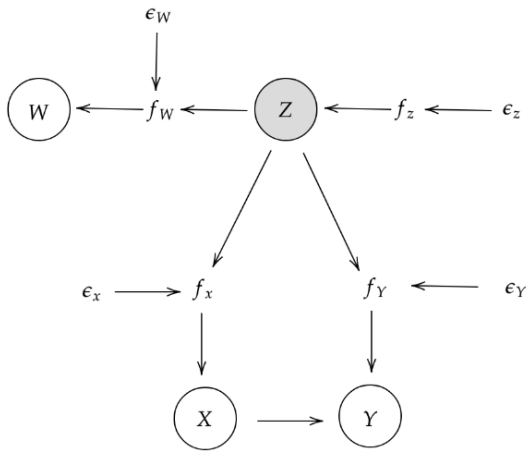


Figure 1. Example causal graph featuring observation variables X and Y, latent confounder Z, and proxy variable W.

As an example, in Figure 1, we can consider the FCM function for Z to be $f_Z(\epsilon_Z)$ since it only has a causal parent in its unobserved, independent latent effects. In this case $\{P_Z\}$ is the empty set. To contrast, the FCM function for Y would be $f_Y(X, Z, \epsilon_Y)$ since it has causal parents X, the latent Z, and its own independent hidden effects ϵ_Y. In this case $\{P_Y\} = \{X, Z\}$.

In the canonical definition of FCM, these functions look to predict the realization of the observable $\hat{X}_i = f_i(\{P_i\}, \epsilon_i)$.

However, we find it more useful to use the FCM structure to predict model parameters for the distribution of the observable

$$f_i : (\{P_i\}, \epsilon_i) \rightarrow (\{\pi_j\}_i, \{\mu_j\}_i, \{\sigma_j\}_i)$$

which we assume to be a Gaussian mixture with j Gaussians. Then in our previously defined language, we have $p_{X|Y}(X|Y) = f_X(Y, \epsilon_X), P(Y) = f_Y(\epsilon_Y)$ which is the conditional and marginal distributions relevant for the $Y \rightarrow X$ direction and similarly $P_{Y|X}(Y|X) = f_Y(X, \epsilon_Y), P(X) = f_X(\epsilon_X)$ are the relevant distributions for the $X \rightarrow Y$ direction. Given realizations of the ground truth observations, we then have a closed form for the likelihoods given each of the learned distributions.

Modelling Confounding Factors

The basic FCM structure allows us to generalize the idea of an encoder-decoder structure by modelling hidden effects as different distributions ϵ_i. Fortunately, it is also easy to extend to modelling latent confounders between variables. In particular, in the style of (Madras et al., 2018), we introduce the latent variable Z which can affect both X and Y. Since Z is not an independent hidden effect, it cannot be absorbed into either ϵ_X or ϵ_Y. Instead, we must append Z as an input to both f_X and f_Y.

While Goudat et al. (2018) assumes that each confounder follows a known distribution, this is perhaps an overly ideal scenario that we are unlikely to encounter in practice. Instead, we choose to infer the values of Z. Let us consider the proxy variable W that is also impacted by Z such that $W|Z \perp\!\!\!\perp X|Z$ and $W|Z \perp\!\!\!\perp Y|Z$ as depicted in the setup from Figure 1. While Z and W can in principle be sets of variables of arbitrary length, for simplicity we restrict both to a single variable in this analysis. Further, we will model all causal causal effects of Z on X, Y and W as additive effects.

To incorporate this variational inference of latent variables into the causal direction methodology, we follow (Madras et al., 2018) and assume that Z, $W|Z$ and $Y|X, Z$ are continuous, normally distributed variables with assumed

$$\begin{aligned} p(Z) &\sim N(0, 1) \\ p(W|Z) &\sim N(\mu_W(Z), \sigma^2_W(Z)) \\ p(X|Z) &\sim N(\mu_X(Z), \sigma^2_X(Z)) \\ p(Z|W) &\sim N(\mu_Z(X), \sigma^2_Z(X)) \end{aligned}$$

We can estimate a lower bound for the combined probability of all variables by the Evidence Lower Bound (ELBO) which is defined by

$$L = \sum_{i=1}^{N} E_{p(z_i|x_i)}[lnp(w_i|z_i) + lnp(x_i|z_i) + lnp(y_i|x_i, z_i) + lnp(z_i) - lnp(z_i|x_i)]$$

We generate the ELBO for both causal directions such that they approximate $L_{X\rightarrow Y}$ and $L_{Y\rightarrow X}$. To compute the expected value inside the ELBO we use Monte Carlo simulation. Further, we use variational encoders to model $\mu_W(Z), \sigma^2_W(Z)$ and their alike parameters for X and Z.

As a further deviation from (Bengio et al., 2019) we will split the algorithm into 2 parts: First, the inference models are estimated for the two causal directions and their latent variable representations and thereafter the causal direction is inferred by optimizing α on the two models. The difference to the former structure is that we no longer run these two steps iteratively, but instead separate them to allow for a more independent deduction of latent variables. While this change my appear like a trivial deviation, it fundamentally changes the nature of the meta-learner's usage. Instead of configuring causal models and their best direction at the same time, it now simply checks which of two models is more likely to yield the correct causal direction.

3. Experiments & Results

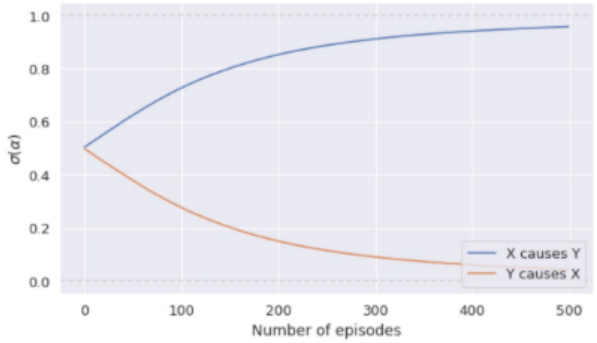
Unless otherwise specified, we modelled all data by the above process. Specifically, we modelled Y as $Y \sim f(X) + Z$ where f is a cubic spline function. For our experiments, we use 1000 observations for each draw of the training and transfer distribution, as this is also the number used by (Bengio et al., 2019), and 300 iterations of the Monte Carlo method. The trained models are meta-learned for 5 steps and the FCM uses 12 Gaussians. Note that due to the split optimization procedure that trains α and β for otherwise fully optimised models our method is particularly robust such that repetitions of the same scenario tend to converge to the same solution; in fact often the estimator paths for α and β are sheer indistinguishable over different repetitions.

Moreover, we model X by $X \sim N(\mu_X, 2) + Z$ with $\mu_X = 0$ for the training distribution and $\mu_X \sim U(-4, 4)$ under the transfer distribution. For our inference analysis we assumed that all variables are normally distributed. All models were trained for 500 meta iterations and the results for $\sigma(\hat{\alpha})$ and $\sigma(\hat{\beta})$ were extracted in the end.

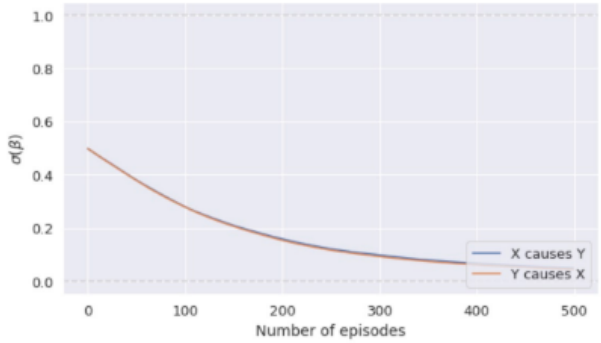
Normality Results

To show the functionality of our methodology a plot of the estimation path for $\sigma(\alpha)$ and $\sigma(\beta)$ is provided in Figure 5. The graph depicts two distinct scenarios: in the first our true model states that X causes Y while in the second we model

that Y causes X. This allows us to evaluate the properties of our optimisation method without falling for a potential fallacy if α is biased in a certain direction.



(a) Results for $\sigma(\alpha)$



(b) Results for $\sigma(\beta)$

Figure 2. $\sigma(\alpha)$ and $\sigma(\beta)$ estimates for the the standard model. We generate two models with opposite directions. In blue X causes Y while in orange Y is the true cause of X

The sigmoid transformation of α shows that in both cases we are able to infer the correct causal direction. For the model that specifies X causing Y we can observe that $\sigma(\alpha) \to 1$ for larger epochs, while for the reverse causal direction the parameter goes to 0. Further, in both cases $\sigma(\beta)$ correctly tends to 0, such that according to the model there exists a causal relationship.

It can also be seen that for both parameters the estimators follow a strikingly smooth curve. Also, for $\sigma(\beta)$ the results are almost identical for the two scenarios. Both of these observations are an artifact of the unusually robust nature of our estimator caused by the 2 step estimation procedure in combination with our large number of observations.

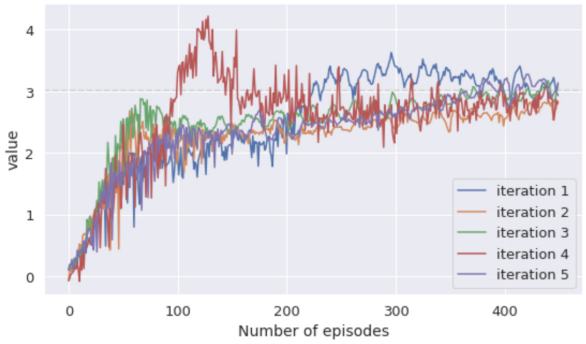
Analysis of the FCM

To analyze the correctness of our results we can investigate the output of the FCM, specifically the estimate for the mean of the predicted variable. If our assumed causal direction is that X causes Y and the FCM models k Gaussian variables then we can use

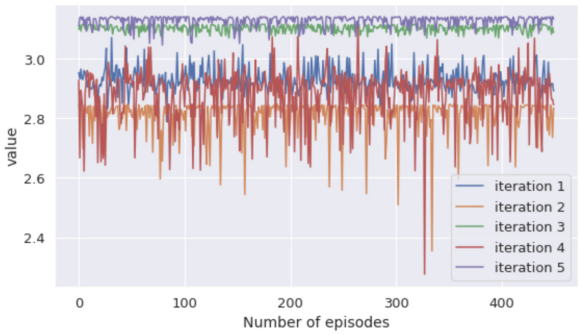
$$\hat{y}(x,z) = \frac{\sum_{i=1}^{k} \pi_{i|x,z}\mu_{i|x,z}}{\sum_{i=1}^{k} \pi_{i|x,z}}$$

as the prediction of the mean of Y given X and Z. The FCM mean of X can be inferred in a likewise fashion. For this analysis, we use $X = 0$, $Y = 0$ and $Z = 0$ as input variables to predict the conditioned mean.

Plots for this experiment for 5 repetitions can be seen in Figure 3. For the causal direction $X \to Y$ in Figure 3a the dashed line indicates the actual mean of $Y|X = 0, Z = 0$ for our spline function. As the spline does not have a proper inverse function, no such line is included in Figure 3b.



(a) Results for the model assuming X causes Y



(b) Results for the model assuming Y causes X

Figure 3. output of the FCM indicating the mean of Y given X and Z for $X = 0, Z = 0$. The dashed line shows the actual value of Y at this point

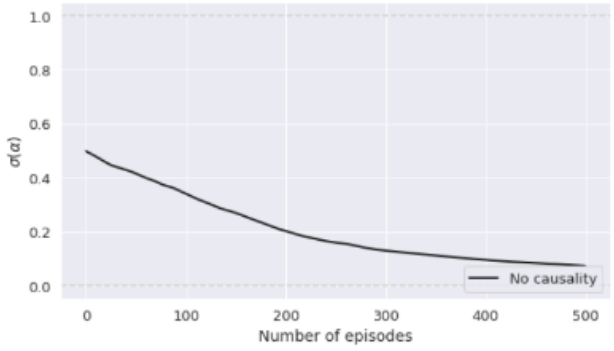
As Figure 3a shows, the FCM mean converges to the correct conditioned mean for all five repetitions. For all iterations this happens within the first 300 episodes, though the fourth iteration shows some stronger fluctuations before. As we only use our converged FCM when training α and β, this also explains why the prior training paths for the two parameters look strikingly similar; as the FCM models have already

converged in all cases, the model training the two parameters will be more robust to random sampling and gradient descent perturbations.

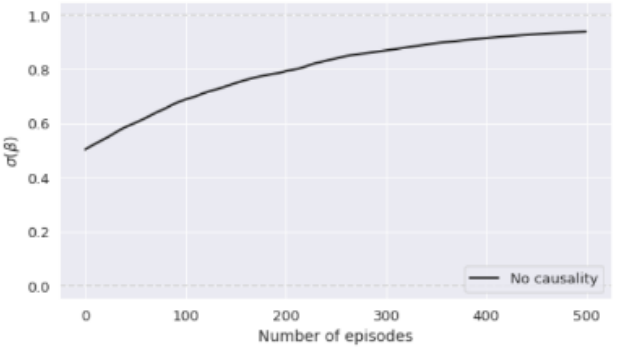
In models that train correctly, we expect that the total variation after convergence of the FCM should be due to the ϵ noise. Seeing that the FCM models for $X \to Y$ converge to the same variance after many iterations adds evidence towards that causal direction. In contrast, the variance of $Y \to X$ FCM models fail to consistently attain the same variance near the end of training which suggests that indeed this is the wrong causal direction.

For the causal direction $Y \to X$ the graph shows that the FCMs converge to two distinct conditional means for X. The reason for this behavior is the aforementioned non-invertibility of the spline function. Yet, the fourth iteration also shows some strong fluctuation seven for larger episodes, indicating that the FCM is less stable for this causal direction. An additional interesting difference between the two plots is that the starting positions differ for the two causal direction. While for $X \to Y$ all conditional means start around 0 and then gradually improve towards the true value, they start at separate values for $Y \to X$ and remain at these values throughout the optimization process.

We conjecture that the cause of the different initial starting values is most likely due to the random effect of realization of ϵ noise input to the FCMs modelling the independent hidden effects to the variables. In particular, since the spline is not one-to-one, the inverse function can take on multiple values and the initial state will change preference depending on the random noise. Moreover, the unstable learning is not enough to move from one local extrema to another.



(a) Results for $\sigma(\alpha)$



(b) Results for $\sigma(\beta)$

Figure 4. $\sigma(\alpha)$ and $\sigma(\beta$ estimates for a model with $X|Z \perp\!\!\!\perp Y|Z$

Detection of no Causality

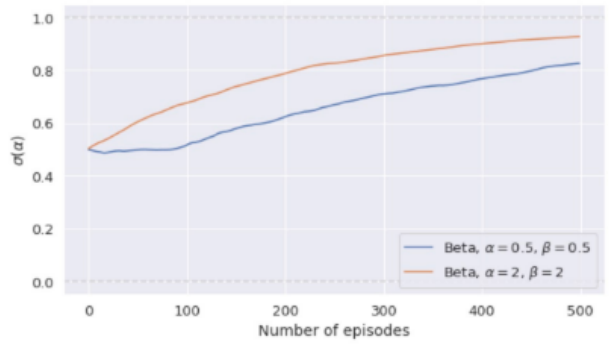
To demonstrate the issues with the case where there is no causal relationship between X and Y variables, we consider independent ground truth relationship after conditioning on the latent variable Z. In Figure 4 we plot the paths for $\sigma(\alpha)$ and $\sigma(\beta)$ in this problem setup.

The sigmoid transformation for β is converging towards 1 which suggests that the model is confident that there is no causal relationship. We can see in the previous section that the reverse process holds otherwise, namely that the sigmoid transformation for β converges towards 0 when the model has some causal structure.

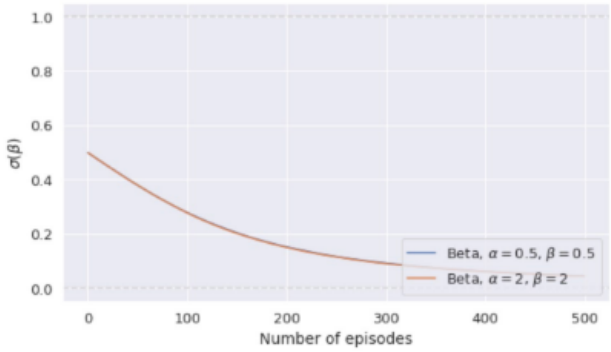
It is notable that the sigmoid transformation for α does indeed converge towards a specific direction, in this case 0. However, recall that since the ground truth is independence, the question that α is answering is malformed. That is, given the choices $X \to Y$ and $Y \to X$, it is simply trying to find which transfer distribution is better. Due to the abundance of randomness from the FCM as well as that from inferring the latent variable, it is easy to fit to the noise and prefer one direction over the other despite. However, this is reasonable since neither is the correct answer. This suggests that in cases where we would like to query the existence of a causal relationship along with the direction if it exists, we should consult β first to determine if α is relevant in answering a well-defined question.

Robustness for Non-normality

In our model we correctly assume that Z is normally distributed when estimating its parameter for our causal inference. As this assumption of a correct causal model is not always applicable in practice it is of interest if our inference results keep their predictive power if this is not the case. Hence, we model Z as a Beta distributed variable with parameters a and b and choose 2 variable combinations for the two variables. The first is $a = 2, b = 2$, which has a more similar probability density function (pdf) to the Normal distribution. The second is $a = 0.5, b = 0.5$, which has a U-shaped pdf and is therefore strongly dissimilar to a Normal distribution. The results for the two distributions are shown in Figure 5. As can be seen especially for the



(a) Results for $\sigma(\alpha)$



(b) Results for $\sigma(\beta)$

Figure 5. $\sigma(\alpha)$ and $\sigma(\beta$ estimates if $X \sim Beta(0.5, 0.5)$ or $Beta(2, 2)$.

Beta distribution with the smaller hyperparameters we need longer to predict the causal direction and only achieve a lower value of $\sigma(\alpha)$. However, the optimizer nonetheless predicts the correct causal direction for both scenarios. Additionally, the estimators for $\sigma(\beta)$ follow the same path in both scenarios and correctly state that there exists a causal relationship despite the misspecified latent variable model. In total this non-normality analysis shows that our model is robust to the distribution of the latent variable.

Robustness for Limited Sample Data

In previous results, we sample data from well defined ground truth distributions. In particular, for each iteration step of the training process we sample new data from the specified distributions and - in case of the training distribution - randomly change the mean of X's probability distribution to derive the transfer distribution. However, in real world applications, there will only exist a finite dataset - some number of realizations from the ground truth distribution which is inaccessible and only a limited number of transfer distributions. In effect, by sampling from the ground truth directly, we are assuming some infinite size dataset from which we draw data. Instead, a more realistic approach is to mimic a finite dataset environment by first creating a static dataset. Afterwards, we can draw data samples from it during training and compute our parameter estimates in the usual way. For such a more realistic approach it is also important to account for the effect that in practice there exists only a limited number of transfer distributions.

Therefore, to define these finite datasets we use four hyperparameters:

1. The total number of observations in the training distribution
2. The total number of observations in each sample of the transfer distribution
3. The number of observations used in each training episode
4. The number of distinct transfer distributions

When sampling our data from the predefined datasets we only use a portion of that data during each training episode. Specifically, we will use the Bootstrap to sample our data, such that we draw samples with replacement from the datasets.

Table 1 depicts mean results for $\sigma(\alpha)$, $\sigma(\beta)$ and the FCM mean over 5 iterations for different combinations of the four hyperparameters. To analyze effects for datasets of different size we used 100 and 1000 total observations for the training samples and 500 and 100 for each of the transfer distributions. For each combination, we used a fixed amount of observations per training episode and 50, 10, 3, and 1 transfer distributions.

The table shows that for restricted preselected data samples the predictive accuracy of our methodology decreases. For example, for 100 samples with 10 transfer distributions we only reach a mean value of 0.775 for $\sigma(\alpha)$ and 0.216 for $\sigma(\beta)$. While other mean values from the table support this notion, some outliers appear to reject a generalization of this deduction; for example, for 100 observations and 1 transfer distribution we reach the usual mean value of 0.955 for $\sigma(\alpha)$. However, it should be noted that these mean observations suffer from two shortcomings: first, we only computed five iterations due to constraints on our computing time and architecture. And second, mean values for a low number of transfer distributions are particularly prone to random effects. For example, if we only sample one transfer distribution, during training of α and β we keep evaluating the trained causal models on the same small dataset, thus strongly exposing the values of α and β to random effects.

Despite the strong effect of random effects on our mean

Total observations for training sample	Total observations for transfer samples	Observations per episode	Total transfer distributions	Mean of $\sigma(\alpha)$	Mean of $\sigma(\beta)$	Mean of FCM
1000	500	100	50	0.778	0.058	2.415
1000	500	100	10	0.957	0.044	2.967
1000	500	100	3	0.416	0.045	2.359
100	100	50	10	0.775	0.216	2.735
100	100	50	3	0.949	0.202	2.69
100	100	50	1	0.955	0.108	2.894

Table 1. Mean results for predefined datasets. For all results we used the average of 5 repetitions.

results, they indicate that for restricted samples our methodology performs worse. This especially applies to our β parameter, as especially values for $\sigma(\beta)$ deviate further from 0 for more restricted data samples. Additionally, we can also note that our estimates for the FCM mean have worsened in comparison to our results in previous sections. Nonetheless, our methodology appears to work better than pure-chance guesses. Thus, we conclude this section by stating that albeit our meta-learning algorithm performs worse for predefined datasets, it can still generate valuable solutions.

4. Conclusions

In our experimentation, we have improved the framework of Bengio et al. (2019) to express the causal graph structure more explicitly by using FCM. This innovation yields great improvements in the results as we are able to demonstrate faster and better convergence to higher confidence of the correct causal direction in more difficult problem setups as compared to prior works. In particular, we have shown that for generalized independent hidden effects and with single latent confounders, we are able to recover the correct causal direction with near sure confidence. The framework that we use is easily extendable to larger number of confounding factors as well as more observational variables as they can be fit into our structure by introducing more inputs to the relevant FCM networks and more FCM networks to learn respectively.

We also introduce a novel parameter β to determine the existence of a causal relationship. This is useful in conjunction with the previously developed α to determine whether learning a causal direction is well-specified or reasonable. In similar cases with confounding effects, we demonstrate the ability of this parameter to also learn the correct relationship with good confidence.

Importantly, we consider the mechanics of the FCM models that are novel for this meta-learning approach and demonstrate that they are able to converge to recover the distributional parameters (under the model assuming the correct causal relation) as desired. These results also suggest some rationale behind the smoothness of the α and β paths and their convergence properties.

Finally, we show that in cases where model assumptions are violated, we are still able to learn the causal relations with some effectiveness. When we have distributional deviation from normality, the α paths have more difficulty converging but still tend towards the correct direction and result in the correct conclusion. When we constrain the model to finite datasets, the probability of converging to the correct causal relation decreases as a function of dataset size, but we empirically still see that we can often recover the correct relation regardless.

Directions for Further Research

In order to generalize assumptions about the distribution of confounding effects, we introduce a variational inference framework for the latent variables that we measure through some proxy variable. However, this also necessitates the replacement of the exact likelihood with the ELBO since the exact form is no longer tractable. Unfortunately, the ELBO is just a bound for the log likelihood and is not necessarily sharp. In particular, it is possible for $L_{X\to Y} > L_{Y\to X}$ while also having $ELBO_{X\to Y} < ELBO_{Y\to X}$ which would be problematic for the optimization of $R(\alpha)$ and cause it to step in the wrong direction. Luckily, it seems that the distributions we use do not exhibit this property but its possibility solicits additional work to find guarantees on the ELBO inequalities with respect to each other or some other similar means to resolve this uncertainty.

We also note some related issues with the β meta parameter. It seems that the ELBO bounds for the marginal distributions are empirically tighter than the ELBO bounds for the conditional distributions. This causes the independence hypothesis to be implicitly favored over the causal directions. We suggest a simple and natural solution - introduce a prior confidence of causal existence to factor into the computation which allows for a balancing of this inequality. This seems like a strong assumption, but we note that it is at least less restrictive than other related works which often assume that there exists causality and the direction only remains to be determined. However, we concede that there may be a better method of balancing or resolving the ELBO inequality issue

between the marginal and conditional distributions.

Furthermore, we may be able to extend this work to larger causal graphs at the cost of large amounts of compute power. The current model requires a neural network to learn a FCM for each variable in the graph. The current method that we have for inferring causal direction on larger graphs would be to assume a starting orientation and iteratively perform a two step process similar to the method of Goudat et al. (2018) to perform hill climbing.

1. Update causal direction scheme by relearning the transfer distributions in topological order using the current orientation to decide the causal parents of variables.
2. Resolve any formed cycles by reorienting violating causal arrows with the smallest confidence

However, we notice the ability of meta-CGNN in Ton et al. (2018) to use dataset and causal direction pairs as meta-tasks to train networks that can leverage dataset similarities to find causal directions during meta-test time more quickly. It may be possible to combine these two works in order to consolidate the computational complexity of extending work to larger graph sizes.

Finally, we explore only some perturbations to the distributional assumptions. However, there are a larger number of possibilities to explore as well as extensions to more complex datasets. Improvements on these fronts will allow for the development of agents with improved robustness and usefulness in practice.

References

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

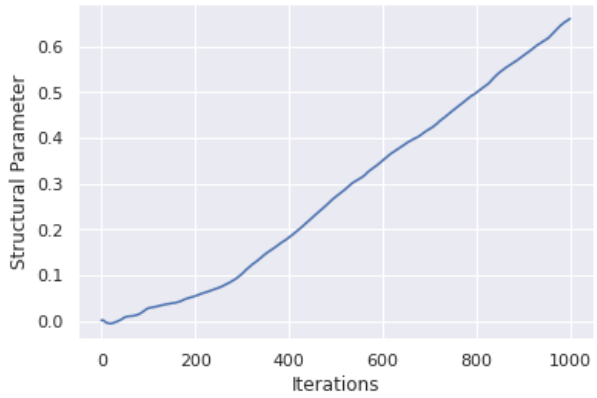
Goudat, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. Causal generative neural networks. *arXiv preprint arXiv:1711.08936*, 2018.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Fairness through causal awareness: Learning causal latent-variable models for biased data. *arXiv preprint arXiv:1809.02519*, 2018.

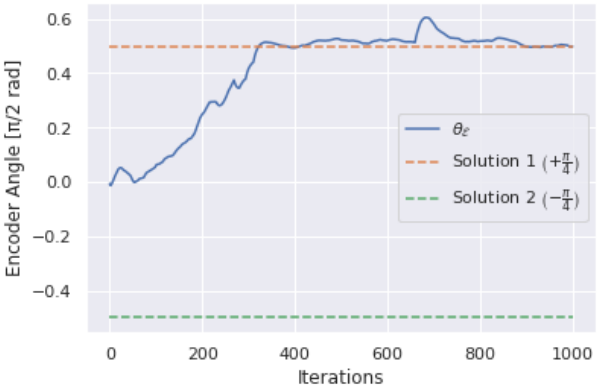
Ton, J.-F., Sejdinovic, D., and Fukumizu, K. Meta learning for causal direction. *arXiv preprint arXiv:1711.08936*, 2018.

Zhan, X., Xie, J., Liu, Z., Ong, Y.-S., and Loy, C. C. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6688–6697, 2020.

A. Appendix



(a) α parameter from Bengio et al. (2019) encoder-decoder structure.

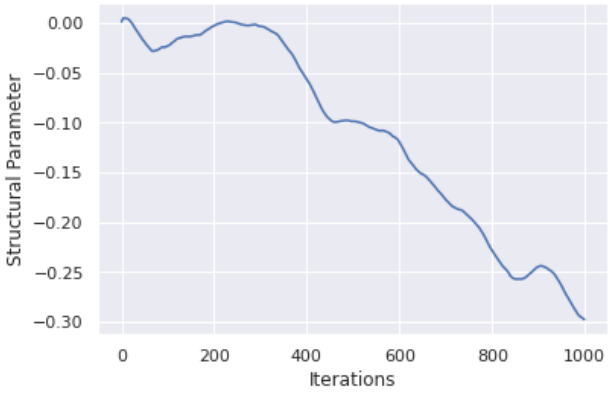


(b) θ parameter from Bengio et al. (2019) encoder-decoder structure.

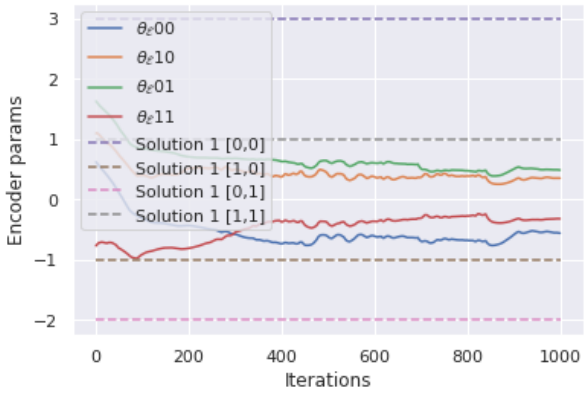
Figure 6. Recovering results from Bengio et al. (2019) encoder-decoder structure.

Notice in Figure 6 that with double the iteration count, $\alpha \rightarrow 0.6$ or equivalently $\sigma(\alpha) \rightarrow\approx 0.66$ which is much less than the confidence our FCM-utilizing method achieves.

However, they are able to recover the rotational parameter for the encoder nicely.



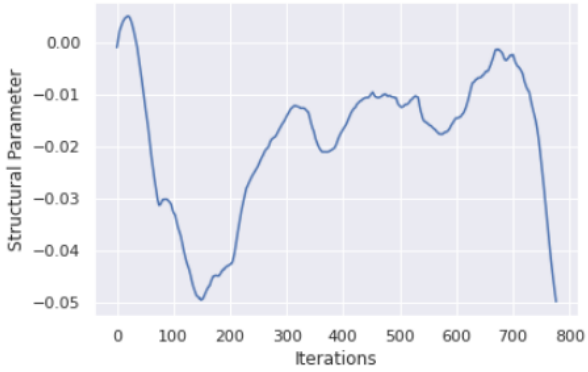
(a) α parameter from generalized encoder-decoder structure.



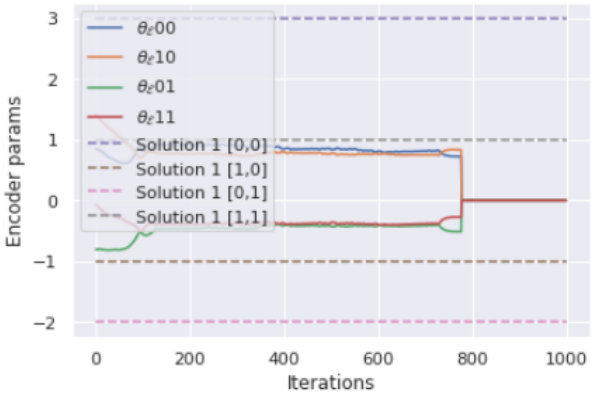
(b) θ parameters from generalized encoder-decoder structure.

Figure 7. Generating results from generalized encoder-decoder structure.

In contrast, extending the encoder-decoder structure to be a linear layer yields deficient solutions. We can see that the encoder parameters stagnate towards common coefficients that encourage colinearity of output in Figure 7 as well a demonstrating that the α parameter does not converge strongly in either direction. Similarly, in Figure 8 we see that soft-penalization to encourage non-degenerate solutions does not help the issue and we continue to see bad behavior in both the parameters of the encoder and the structural parameter.



(a) α parameter from generalized encoder-decoder structure with soft regularization.



(b) θ parameters from generalized encoder-decoder structure with soft regularization.

Figure 8. Generating results from generalized encoder-decoder structure with soft regularization.