

# A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

Yoshua Bengio<sup>1,2,5</sup>, Tristan Deleu<sup>1</sup>, Nasim Rahaman<sup>4</sup>, Nan Rosemary Ke<sup>3</sup>, Sébastien Lachapelle<sup>1</sup>, Olexa Bilaniuk<sup>1</sup>, Anirudh Goyal<sup>1</sup> and Christopher Pal<sup>3,5</sup>  
Mila, Montréal, Québec, Canada

<sup>1</sup> Université de Montréal

<sup>2</sup> CIFAR Senior Fellow

<sup>3</sup> École Polytechnique Montréal

<sup>4</sup> Ruprecht-Karls-Universität Heidelberg

<sup>5</sup> Canada CIFAR AI Chair

## Abstract

We propose to meta-learn causal structures based on how fast a learner adapts to new distributions arising from sparse distributional changes, e.g. due to interventions, actions of agents and other sources of non-stationarities. We show that under this assumption, the correct causal structural choices lead to faster adaptation to modified distributions because the changes are concentrated in one or just a few mechanisms when the learned knowledge is modularized appropriately. This leads to sparse expected gradients and a lower effective number of degrees of freedom needing to be relearned while adapting to the change. It motivates using the speed of adaptation to a modified distribution as a meta-learning objective. We demonstrate how this can be used to determine the cause-effect relationship between two observed variables. The distributional changes do not need to correspond to standard interventions (clamping a variable), and the learner has no direct knowledge of these interventions. We show that causal structures can be parameterized via continuous variables and learned end-to-end. We then explore how these ideas could be used to also learn an encoder that would map low-level observed variables to unobserved causal variables leading to faster adaptation out-of-distribution, learning a representation space where one can satisfy the assumptions of independent mechanisms and of small and sparse changes in these mechanisms due to actions and non-stationarities.

## 1. Introduction

Current machine learning methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice. It is not enough to obtain good generalization on a test set sampled from the same distribution as the training data, we would also like what has been learned in one setting to generalize well in other related distributions. These distributions may involve the same concepts that were seen previously by the learner, with the changes typically arising because of actions of agents. More generally, we would like what has been learned previously to form a rich base from which very fast adaptation to a new but related distribution can take place, i.e., obtain good transfer. Some new concept may have to be learned but because most of the other relevant concepts have already been captured by the learner (as well as how they can be composed), learning can be very fast on the transfer distribution.

Short of any assumption, it is impossible to have a successful transfer to an unrelated distribution. In this paper we focus on the assumption that the changes are sparse when the knowledge is represented in an appropriately modularized way, with only one or a few of the modules having changed. This is especially relevant when the distributional change is due to actions by one or more agents, such as the interventions discussed in the causality literature (Pearl, 2009; Peters et al., 2017), where a single causal variable is clamped to a particular value. In general, it is difficult for agents to influence many underlying causal variables at a time, and although this paper is not about agent learning as such, this is a property of the world that we propose to exploit here, to help discovering these variables and how they are causally related to each other.

To motivate the need for inferring causal structure, consider that interventions may be actually performed or may be imagined. In order to properly plan in a way that takes into account interventions, one needs to imagine a possible change to the joint distribution of the variables of interest due to an intervention, even one that has never been observed before. This goes beyond good transfer learning and requires causal learning and causal reasoning. For this purpose, it is not sufficient to learn the joint distribution of the observed variables. One also should learn enough about the underlying high-level variables and their causal relations to be able to properly infer the effect of an intervention. For example,  $A=\text{Raining}$  causes  $B=\text{Open Umbrella}$  (and not vice-versa). Changing the marginal probability of **Raining** (say because the weather changed) does not change the mechanism that relates  $A$  and  $B$  (captured by  $P(B|A)$ ), but will have an impact on the marginal  $P(B)$ . Conversely, an agent’s intervention on  $B$  (**Open umbrella**) will have no effect on the marginal distribution of  $A$  (**Raining**). That asymmetry is generally not visible from the  $(A, B)$  training pairs alone, until a change of distribution occurs, e.g. due to an intervention. This motivates the setup of this paper, where one learns from a set of distributions arising from not necessarily known interventions, not simply to capture a joint distribution but to discover the some underlying causal structure.

Machine learning methods are often exploiting some form of assumption about the data distribution (or else, the no free lunch theorem tells us that we cannot have any confidence in generalization). In this paper, we are considering not just assumptions on the data distribution but also on how it changes (e.g., when going from a training distribution to a transfer distribution, possibly resulting from some agent’s actions). We propose to rely on the assumption that, *when the knowledge about the distribution is appropriately represented, these changes would be small*. This arises because of an **underlying assumption** (but more difficult to verify directly) **that only one or few of the ground truth mechanisms have been changed**, due to some generalized form of intervention leading to the modified distribution.

How can we exploit this assumption? As we explain theoretically and verify experimentally here, if we have the right knowledge representation, then we should get fast adaptation to the transfer distribution when starting from a model that is well trained on the training distribution. This arises because of our assumption that the ground truth data generative process is obtained as the composition of independent mechanisms and that, very few ground truth mechanisms and parameters need to change when going from the training distribution to the transfer distribution. A model capturing a corresponding factorization of knowledge would thus require just a few updates, a few examples, for this adaptation to the transfer distribution. As shown below, the expected gradient on the unchanged parameters would be near 0 (if the model was already well trained on the training distribution), so the effective search space during adaptation to the transfer distribution would be greatly reduced, which tends to produce fast adaptation, as found experimentally.

Thus, based on the assumption of small change in the right knowledge representation space, we can define a meta-learning objective that measures the speed of adaptation, i.e., a form of regret, in order to optimize the way in which knowledge should be represented, factorized and structured. *This is the core idea presented in this paper*. Note that a stronger signal can be obtained when there are more non-stationarities, i.e., many changes in distribution, just like in meta-learning we get better results with more meta-examples.

In this way, we can take what is normally considered a nuisance in machine learning (changes in distribution due to non-stationarity, uncontrolled interventions, etc.) and turn that into a training signal to find a good way to factorize knowledge into components and mechanisms that match the assumption of small change. Thus, we end up optimizing in an end-to-end way the very thing we care about at the end, i.e. fast transfer and robustness to distributional changes. If the data was really generated from the composition of independent causal mechanisms (Peters et al., 2017), then there exists a good factorization of knowledge that mimics that structure. If in addition, at each time step, agents in the real world tend to only be able to change one or very few high-level variables (or the associated mechanisms producing them), then our assumption of small change (in the right representation) should be generally valid. Also, in addition to obtaining fast transfer, we may be able to recover a good approximation of the true causal decomposition into independent mechanisms (to the extent that the observations and interventions can reveal those mechanisms).

In this paper, we begin exploring the above ideas with specific experiments on synthetically generated data in order to validate them and demonstrate the existence of simple algorithms to exploit them. However it is clear to us that much more work will be needed to evaluate the proposed approach in a diversity of settings and with different specific parametrizations, training objectives, environments, etc. We begin with what are maybe the simplest possible settings and evaluate whether the above approach can be used to

learn the direction of causality. We then study the crucial question of obtaining a training signal about how to transform raw observed data into a representation space where the latent variables can be modeled by a sparse causal graph with sparse distributional changes and show results that confirm that the correct encoder leads to a better value of our expected regret meta-learning objective.

## 2. Which is Cause and Which is Effect?

To anchor ideas and show an example of application of the above-proposed meta-objective for knowledge decomposition, we consider in this section the problem of determining if variable  $A$  causes variable  $B$  or vice-versa. The learner observes training samples  $(a, b)$  from a pair of related distributions, which by convention we call the training distribution and the transfer distribution. Note that based only on samples from a single (training) distribution, in general both the  $A \rightarrow B$  model ( $A$  causes  $B$ ) and the  $B \rightarrow A$  model (vice-versa, see Equation (1) below) tend to perform as well in terms of ordinary generalization (to a test set sampled from the training distribution), see also a theoretical argument and simulation results in Appendix A. To highlight the power of the proposed meta-learning objective, we consider the situation where lots of examples are available for the training distribution but very few for the transfer distribution. In fact, as we will argue below, the training signal that will allow us to infer the correct causal direction will be stronger if we have access to many short transfer adaptation episodes. Short episodes are most informative because after having seen a lot of data from the transfer distribution, it will not matter much whether  $A$  causes  $B$  or vice-versa (when there is enough training data compared to the number of free parameters, both models converge towards an optimal estimation of the joint). However, in order to generalize quickly from very few examples of the transfer distribution, it does matter to have made the correct choice of the causal direction. Let us now justify this in more detail below and then demonstrate this by simulations.

### 2.1 Learning a Causal Graph with two Discrete Variables

Let both  $A$  and  $B$  be discrete variables each taking  $N$  possible values and consider the following two parametrizations (the  $A \rightarrow B$  model and the  $B \rightarrow A$  model) to estimate their joint distribution:

$$\begin{aligned} P_{A \rightarrow B}(A, B) &= P_{A \rightarrow B}(A)P_{A \rightarrow B}(B | A) \\ P_{B \rightarrow A}(A, B) &= P_{B \rightarrow A}(B)P_{B \rightarrow A}(A | B) \end{aligned} \tag{1}$$

Each of these two graphical models (denoted  $A \rightarrow B$  and  $B \rightarrow A$ ) decomposes the joint into two separately parametrized modules, each corresponding to a different causal mechanism associated with the probability of a variable given its parents in the graph. This amounts to four modules:  $P_{A \rightarrow B}(A)$ ,  $P_{A \rightarrow B}(B | A)$ ,  $P_{B \rightarrow A}(B)$  and  $P_{B \rightarrow A}(A | B)$ . We will train both models independently. Since we assume in this section that the pairs  $(A, B)$  are completely observed, we can use a simple maximum likelihood estimator to independently train all four modules (the log-likelihood of the joint decomposes into separate objective functions, one for each conditional, in a directed graphical model with fully observed variables). In the discrete case with tabular parametrization, the maximum likelihood estimator can be computed analytically, and corresponds to the appropriately normalized relative frequencies. Let  $\theta$  denote the parameters of all these models, split into sub-vectors for each module, e.g.,  $\theta_{A|B}$  for the  $N^2$  conditional probabilities for each possible value of  $B$  and each possible value of  $A$ . In our experiments, we parametrized these probabilities via softmax of unnormalized quantities.

#### 2.1.1 THE ADVANTAGE OF THE CORRECT CAUSAL MODEL

First, let us consider simply the likelihood of the training data only (i.e., no change of distribution) for the different causal models considered. Both models have  $O(N^2)$  parameters, and maximum likelihood estimation leads to indistinguishable test set performance (where the test set is sampled from the training distribution). See Appendix A for a demonstration that both models would have the same likelihood, and associated experimental results. These results are not surprising in light of the existing literature on non-identifiability of causality from observations (Pearl, 2009; Peters et al., 2017), but they highlight the importance of using changes in distribution to provide a signal about the causal structure.

Now instead let us compare the performance of our two hypotheses ( $A \rightarrow B$  vs  $B \rightarrow A$ ) in terms of how fast the two models adapt on a transfer distribution after having been trained on the training distribution. We will assume simple stochastic gradient descent on the parameters for this adaptation but other procedures could be used, of course. Without loss of generality, let  $A \rightarrow B$  be the correct causal model. To make the case stronger, let us consider that the change between the two distributions amounts to a random change in the parameters of the true  $P(A)$  for the cause  $A$  (because this will have an impact on the effect  $B$ , which can be picked up and reveal the causal direction). We do not assume that the learner knows what intervention was performed, unlike in more common approaches to causal discovery and controlled experiments. We only assume that some change happened and we try to exploit that to reveal structural causal information.

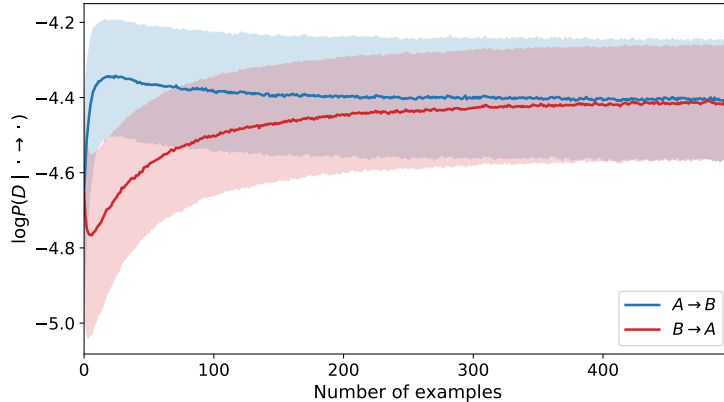


Figure 1: Adaptation to the transfer distribution, as more transfer distribution examples are seen by the learner (horizontal axis), in terms of the log-likelihood on the transfer distribution (on a large test set from the transfer distribution, tested after each update of the parameters). Here the model is discrete, with  $N = 10$ . Curves are the median over 10 000 runs, with 25-75% quantiles intervals, for both the correct causal model (blue, top) and the incorrect one (red, bottom). We see that the correct causal model adapts faster (smaller regret), and that the most informative part of the trajectory (where the two models generalize the most differently) is in the first 10-20 examples.

## 2.2 Experiments on Adaptation to the transfer distribution

We present experiments comparing the learning curve of the correct causal model on the transfer distribution vs the learning curve of the incorrect model. The adaptation with only a few gradient steps on data coming from a different, but related, transfer distribution is critical in getting a signal that can be leveraged by our meta-learning algorithm. To show the effect of this adaptation, and motivate our use of only a small amount of data from the transfer distribution, we experimented with a model on discrete random variables taking  $N = 10$  possible values.

In this experiment, we fixed the underlying causal model to be  $A \rightarrow B$ , and trained the modules for each marginal and conditional distributions with maximum likelihood on a large amount of data from some training distribution, as explained in Appendix A. See also Appendix G.1 and Table G.1 for details on the definitions of these modules.

We then adapt all the modules on data coming from a transfer distribution, corresponding on an intervention on the random variable  $A$  (i.e., the marginal  $P(A)$  of the ground truth model is modified, while leaving  $P(B | A)$  fixed). We used RMSprop for the adaptation, with the same learning rate. For assessing reproducibility and statistical robustness, the experiment was repeated over 100 different training distributions, and over 100 transfer distributions for each training distributions, leading to 10 000 experiments overall. The procedure to acquire different training/transfer distributions is detailed in Appendix G.1.

In Figure 1, we report the log-likelihoods of both models, evaluated on a large test set of 10 000 from the transfer distribution. We can see that as the number of examples from the transfer distribution (equal to the number of adaptation steps) increases, the two models eventually reach the same log-likelihood, reflecting

our observation from Appendix A. However the causal model  $A \rightarrow B$  adapts faster than the other model  $B \rightarrow A$ , with the most informative part of the trajectory (where the difference is the largest) is within the first 10 to 20 examples.

### 2.2.1 PARAMETER COUNTING ARGUMENT

A simple parameter counting arguments helps us understand what we are observing in Figure 1. First, consider the expected gradient on the parameters of the different modules, during the adaptation phase to the transfer distribution, which we designate as adaptation episode, and corresponds to learning from a meta-example.

**Proposition 1** *The expected gradient over the transfer distribution of the regret (accumulated negative log-likelihood during the adaptation episode) with respect to the module parameters is zero for the parameters of the modules that (a) were correctly learned in the training phase, and (b) have the correct set of causal parents, corresponding to the ground truth causal graph, if (c) the corresponding ground truth conditional distributions did not change from the training distribution to the transfer distribution.*

The proof is given in Appendix B. The basic justification for this proposition is that for the modules that were correctly learned in the training distribution and whose ground truth conditional distribution did not change with the transfer distribution, the parameters already are at a maximum of the log-likelihood over the transfer distribution, so the expected gradient is zero.

As a consequence, the effective number of parameters that need to be adapted, when one has the correct causal graph structure, is reduced to those of the mechanisms that actually changed from the training to the transfer distribution. Since sample complexity - the number of training examples necessary to learn a model - grows approximately linearly (Ehrenfeucht et al., 1989) with VC-dimension (Vapnik and Chervonenkis, 1971), and since VC-dimension grows approximately linearly in the number of parameters in linear models and neural networks Shalev-Shwartz and Ben-David (2014), the learning curve on the transfer distribution will tend to improve faster for the model with the correct causal structure, for which fewer parameters need to be changed. Interestingly, we do not need to have the whole causal graph correctly specified before getting benefits from this phenomenon. If we only have part of the causal graph correctly specified and we change our causal hypothesis to include one more correctly specified mechanism, then we will obtain a gain in terms of the adaptation sample complexity (which shows up when the change in distribution does not touch that mechanism). This nice property also shows up in Proposition 4 (Appendix F), showing a decoupling of the meta-objective across the independent mechanisms.

Let us consider the special case we have been studying up to now. We have four modules, two of which ( $P_{A \rightarrow B}(A)$  and  $P_{B \rightarrow A}(B)$ ) are marginal discrete distributions over  $N$  values, which require each  $N - 1$  free parameters. The other two modules are conditional probability tables that have  $N$  rows each with  $N - 1$  free parameters, i.e., a total of  $N(N - 1)$  free parameters. If  $A \rightarrow B$  is the correct model and the transfer distribution only changed the true  $P(A)$  (the cause), and if  $P(B | A)$  had been correctly estimated on the training distribution, then for the correct model only  $N - 1$  parameters need to be re-estimated. On the other hand, because of Bayes' rule, under the incorrect model ( $B \rightarrow A$ ), a change in  $P(A)$  leads to new parameters for both  $P(B)$  and  $P(A | B)$ , i.e., all  $N(N - 1) + (N - 1) = N^2 - 1$  parameters must be re-estimated. In this case we see that sample complexity may be  $O(N^2)$  for the incorrect model while it would be  $O(N)$  for the correct model (assuming linear relationship between sample complexity and number of free parameters). Of course, if the change in distribution had been over  $P(B | A)$  instead of  $P(A)$ , the advantage would not have been as great. This would motivate information gathering actions generally resulting in a very sparse change in the mechanisms.

### 2.3 Smooth parameterization of the causal structure

In the more general case with many more than two hypotheses for the structure of the causal graph, there will be an exponentially large set of possible causal structures explaining the data and we won't be able to enumerate all of them (and pick the best one after observing episodes of adaptation). However, we can parameterize our belief about an exponentially large set of hypotheses by keeping track of the probability for each directed edge of the graph to be present, i.e., specify for each variable  $B$  whether some variable  $A$  is a

direct causal parent of  $B$  (for all pairs  $(A, B)$  in the graph). We will develop such a smooth parametrization further in Appendix F, but it hinges on gradually changing our belief in the individual binary decisions associated with each edge of the causal graph, so we can jointly do gradient descent on all these beliefs at the same time.

In this section, we study the simplest possible version of this idea, representing that edge belief via a structural parameter  $\gamma$  with  $\text{sigmoid}(\gamma) = \text{sigmoid}(\gamma)$ , our believed probability that  $A \rightarrow B$  is the correct choice. For that single pair of variables scenario, let us consider two explanations for the data (as in the above sections, for models  $A \rightarrow B$  and  $B \rightarrow A$ ), one with probability  $p(A \rightarrow B) = \text{sigmoid}(\gamma)$  and the other with probability  $p(B \rightarrow A) = 1 - \text{sigmoid}(\gamma)$ . We can write down our transfer objective as a log-likelihood over the mixture of these two models. Note this is different from the usual mixture models, which assume separately for each example that it was sampled from one component or another with some probability. Here, we assume that all of the observed data was sampled from one component or the other. The transfer data regret (negative log-likelihood accumulated along the online adaptation trajectory) under that mixture is therefore as follows:

$$\mathcal{R} = -\log [\text{sigmoid}(\gamma)\mathcal{L}_{A \rightarrow B} + (1 - \text{sigmoid}(\gamma))\mathcal{L}_{B \rightarrow A}] \quad (2)$$

where  $\mathcal{L}_{A \rightarrow B}$  and  $\mathcal{L}_{B \rightarrow A}$  are the online likelihoods of both models respectively on the transfer data. They are defined as

$$\begin{aligned} \mathcal{L}_{A \rightarrow B} &= \prod_{t=1}^T P_{A \rightarrow B}(a_t, b_t; \theta_t) \\ \mathcal{L}_{B \rightarrow A} &= \prod_{t=1}^T P_{B \rightarrow A}(a_t, b_t; \theta_t), \end{aligned}$$

where  $\{(a_t, b_t)\}_t$  is the set of transfer examples for a given episode and  $\theta_t$  aggregates all the modules' parameters as of time step  $t$  (since the parameters could be updated after each observation of an example  $(a_t, b_t)$  from the transfer distribution).  $P_{\text{model}}(a, b; \theta)$  is the likelihood of example  $(a, b)$  under some *model* that has parameters  $\theta$ .

The quantity of interest here is  $\frac{\partial \mathcal{R}}{\partial \gamma}$ , which is our training signal for updating  $\gamma$ . In the experiments below, after each episode involving  $T$  transfer examples we update  $\gamma$  by doing one step of gradient descent, to reduce the transfer negative log-likelihood or regret  $\mathcal{R}$ . **What we are proposing is a meta-learning framework in which the inner training loop updates the module parameters (separately) as examples are seen (from either distribution being currently observed), while the outer loop updates the structural parameters (here it is only the scalar  $\gamma$ ) with respect to the transfer negative log-likelihood.**

The gradient of the transfer log-likelihood with respect to the structural parameter  $\gamma$  is pushing  $\text{sigmoid}(\gamma)$  towards the posterior probability that the correct model is  $A \rightarrow B$  and  $(1 - \text{sigmoid}(\gamma))$  towards the posterior probability that the correct model is  $B \rightarrow A$ :

**Proposition 2** *The gradient of the negative log-likelihood of the transfer data in Equation (2) wrt. the structural parameter  $\frac{\partial \mathcal{R}}{\partial \gamma}$  is given by*

$$\frac{\partial \mathcal{R}}{\partial \gamma} = \sigma(\gamma) - P(A \rightarrow B \mid D_2), \quad (3)$$

where  $D_2$  is the transfer data, and  $P(A \rightarrow B \mid D_2)$  is the posterior probability of the hypothesis  $A \rightarrow B$  (when the alternative is  $B \rightarrow A$ ). Furthermore, this can be equivalently written as

$$\frac{\partial \mathcal{R}}{\partial \gamma} = \sigma(\gamma) - \sigma(\gamma + \Delta), \quad (4)$$

where  $\Delta = \log \mathcal{L}_{A \rightarrow B} - \log \mathcal{L}_{B \rightarrow A}$  is the difference between the log-likelihoods of the two hypotheses on the transfer data  $D_2$ .



The proof is given in Appendix D. Note how this posterior probability is basically measuring which hypothesis is better explaining the episode transfer data  $D_2$  overall along the adaptation trajectory.  $D_2$  is a meta-example for updating the structural parameters like  $\gamma$ . **Larger  $\Delta$  of one hypothesis over the other leads to moving meta-parameters faster towards the favoured hypothesis.** This difference in online accumulated log-likelihoods  $\Delta$  also relates to log-likelihood scores in score-based methods for structure learning of graphical models (Koller and Friedman, 2009)<sup>1</sup>.

To find where SGD converges, note that the actual posterior depends on the prior  $\text{sigmoid}(\gamma)$  and thus keeps changing after each gradient step. We are really doing SGD on the expected value of  $\mathcal{R}$  over transfer sets  $D_2$ . Equating the gradient of this expected value to zero to look for the stationary convergence point, we thus see  $\text{sigmoid}(\gamma)$  on both sides of the equation, and we obtain convergence when the new value of  $\text{sigmoid}(\gamma)$  is consistent with the old value, as clarified in this proposition.

**Proposition 3** *Stochastic gradient descent (with appropriately decreasing learning rate) on  $E_{D_2}[\mathcal{R}]$  with steps from  $\frac{\partial \mathcal{R}}{\partial \gamma}$  converges towards  $\text{sigmoid}(\gamma) = 1$  if  $E_{D_2}[\log \mathcal{L}_{A \rightarrow B}] > E_{D_2}[\log \mathcal{L}_{B \rightarrow A}]$ , or  $\sigma(\gamma) = 0$  otherwise.*

The proof is given in Section E of the Appendix, and shows that optimizing  $\gamma$  will end up picking the correct hypothesis, i.e., the one that has the smallest regret (or fastest convergence), measured as the accumulated log-likelihood as adaptation proceeds on the transfer distributions sampled from the distribution  $D_2$ , which we can think of like a distribution over tasks, in meta-learning. This analogy with meta-learning also appears in our gradient-based adaptation procedure, which is linked to existing methods like the first-order approximation of MAML (Finn et al., 2017), and its related algorithms (Nichol et al., 2018). Algorithm 1 (Appendix C) illustrates the general pseudo-code for the proposed meta-learning framework.

### 2.3.1 EXPERIMENTAL RESULTS

To illustrate the convergence result from Proposition 3, we experiment with learning the structural parameter  $\gamma$  in a bivariate model, with discrete random variables, each taking  $N = 10$  and  $N = 100$  possible values. In this experiment, we assume that the underlying causal model (unknown to the algorithm) is fixed to  $A \rightarrow B$ , so that we want the structural parameter to eventually converge to  $\sigma(\gamma) = 1$ . The details of the experimental setup can be found in Appendix G.1.

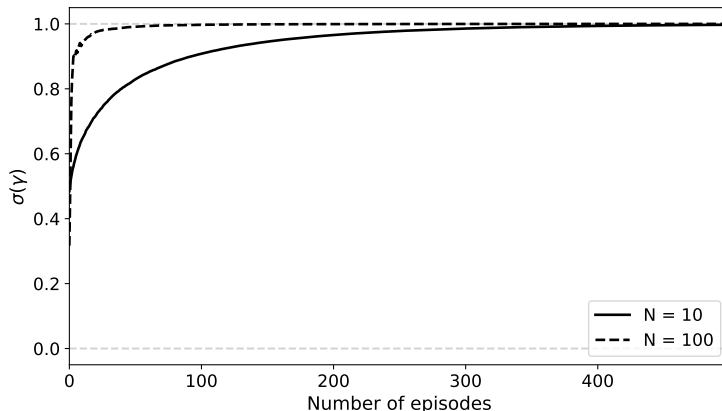


Figure 2: Evolution of model’s belief  $p(A \rightarrow B) = \sigma(\gamma)$  on a bivariate model, with discrete random variables ( $N = 10$  &  $N = 100$ ). The horizontal axis represents the number of episodes (i.e., meta-examples) seen during meta-training, which corresponds to the number of SGD updates of the structural parameter  $\gamma$ .

1. One can see  $\log \mathcal{L}_{A \rightarrow B}$  as a score attributed to graph  $A \rightarrow B$ , analogously for  $\log \mathcal{L}_{B \rightarrow A}$ . The gradient is then pushing toward the graph with the highest score.

In Figure 2, we show the evolution of  $\sigma(\gamma)$  (which is the model’s belief of  $A \rightarrow B$  being the correct causal model) as the number of episodes increases. Starting from an equal belief for both  $A \rightarrow B$  and  $B \rightarrow A$  to occur ( $\sigma(\gamma) = 0.5$ ), the structural parameter converges to  $\sigma(\gamma) = 1$  within 500 episodes.

This observation is consistent across a range of domains, including models with multimodal or multivariate continuous variables, and different parametrizations of the models. In Appendix G.2, we present results for two discrete variables but using MLPs to parametrize the conditional distributions, and where there are more causal hypotheses: we consider one binary choice for each directed edge in the graph, to decide whether one variable is a direct causal parent or not. Figure 3 shows that the correct causal graph is quickly recovered. To estimate the gradient, we use a generalization of the regret loss (introduced above, Equation (2)) and its gradient, described in Appendix F.

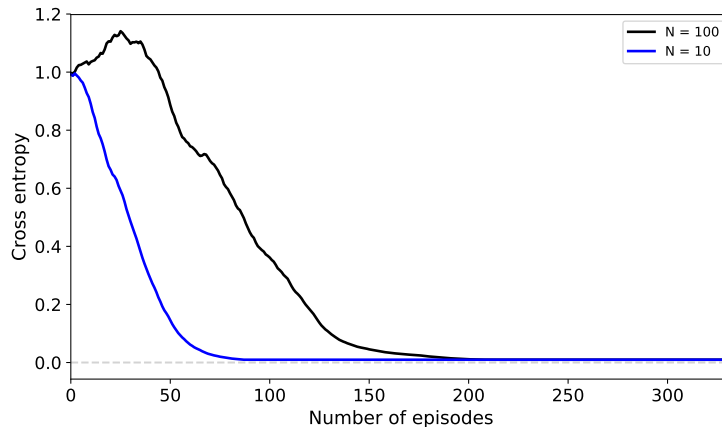


Figure 3: Cross entropy between the ground-truth SCM structure and the learned SCM structure. Each MLP is trained to predict the conditional distribution associated with each discrete variable with  $N$  categories (10 and 100 in this case), given its parents. Between 50 and 100 meta-examples are sufficient to recover the causal structure.

In Appendix G.3, we consider the case of continuous scalar multimodal variables. The ground truth joint distribution is obtained by making the effect  $B$  a non-linear function  $f(A)$  of cause  $A$ , where  $f$  is a randomly generated spline. Figure G.1 shows an example of a resulting joint distribution. We model the conditionals with mixture density networks (Bishop, 1994) and the marginals by a Gaussian mixture. We obtain results that are similar to the discrete case, with the correct causal interpretation being recovered quickly, as illustrated in Figure G.2.

We also show in Appendix G.4 results on models with two continuous random variables, each being distributed as a multivariate Gaussian, with  $N = 10$  dimensions. Similar to the experiment with discrete random variables, the same argument about parameter counting mentioned in Section 2.2.1 holds here. Again, we obtain results consistent with the previous examples, where the structural parameter  $\gamma$  converges to 1, effectively recovering the correct causal model  $A \rightarrow B$ .

### 3. Representation Learning

So far, we have assumed that the system has unrestricted access to the true underlying causal variables,  $A$  and  $B$ . However in many realistic scenarios for learning agents, the observations available to the learner might not be instances of the true causal variables but sensory-level data instead, like pixels and sounds. If this is the case, our working assumption – that the correct causal graph will be sparsely connected, made of independent components, and affected sparsely by distributional shifts – can not be expected to hold true in general in the space of observed variables. To tackle this, we propose to follow the deep learning objective of disentangling the underlying causal variables (Bengio et al., 2013), and learn a representation in which these properties hold. In the simplest form of this setting, the learner must map its raw observations to a hidden



representation space  $H$  via an encoder  $\mathcal{E}$ . The encoder is trained such that the hidden space  $H$  helps to optimize the meta-transfer objective described above, i.e., we consider the encoder, along with  $\gamma$ , as part of the set of structural or meta-parameters to be optimized with respect to the meta-transfer objective.

To study this simplified setting, we consider that our raw observations  $(X, Y)$  originate from the true causal variables  $(A, B)$  via the action of a ground truth decoder  $\mathcal{D}$  (or generator network) that the learner is not aware of but is implicitly trying to invert, as illustrated in Figure 4. The variables  $A$ ,  $B$ ,  $X$  and  $Y$  are assumed to be scalars, and we first consider  $\mathcal{D}$  be a rotation matrix such that:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = R(\theta_{\mathcal{D}}) \begin{bmatrix} A \\ B \end{bmatrix} \quad (5)$$

The encoder is set to another rotation matrix, one that maps the observations  $X, Y$  to the hidden representation  $U, V$  as follows:

$$\begin{bmatrix} U \\ V \end{bmatrix} = R(\theta_{\mathcal{E}}) \begin{bmatrix} X \\ Y \end{bmatrix} \quad (6)$$

The causal modules are now to be trained on the variables  $U$  and  $V$  in the same way as detailed in Section 2, as if they were observed directly. Indeed, if the encoder is valid one would obtain either  $(U, V) = (A, B)$  or  $(U, V) = (B, A)$  up to a negative sign, but we say in that case and without loss of generality that  $(U, V)$  recovered  $(A, B)$ , corresponding to the solution  $\theta_{\mathcal{E}} = -\theta_{\mathcal{D}}$ . In this case, the model  $U \rightarrow V$  is causal and should therefore have an advantage over the anticausal model  $V \rightarrow U$ , as far as adaptation speed on the transfer distribution is concerned. However, if the encoder is not valid, one would obtain superpositions of the form:

$$U = \cos(\theta)A - \sin(\theta)B \quad (7)$$

$$V = \sin(\theta)A + \cos(\theta)B \quad (8)$$

where  $\theta = \theta_{\mathcal{E}} + \theta_{\mathcal{D}}$ . In the extremum where  $\theta = \frac{\pi}{4}$ , it is clear that the model  $U \rightarrow V$  will not have an advantage over the model  $V \rightarrow U$  in terms of regret on the transfer distribution. However, the question we are interested in is whether it is possible to learn the encoder  $\theta_{\mathcal{E}}$ . We verify this experimentally using Algorithm 1, but where the meta-parameters are now both  $\gamma$  (choosing between cause and effect which is which) and the parameters of the encoder (here the angle of a rotation matrix). The details of that experiment are provided in Appendix H, which illustrates – see Figure 5 – how the proposed objective can disentangle (here in a very simple setting) the ground truth variables (up to permutation).

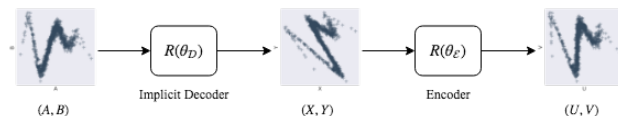


Figure 4: The complete computational graph. The variables  $(A, B)$  are assumed to originate from the true underlying causal distribution, but the observations available to the learner are  $(X, Y)$  samples, which are obtained from  $(A, B)$  via the action of an implicit (a priori unknown) decoder  $R(\theta_{\mathcal{D}})$ . The encoder  $R(\theta_{\mathcal{E}})$  must be learned to undo the action of the (unknown) decoder and thereby recover the true causal variables.

## 4. Related Work

Although this paper focuses on the causal graph, the proposed objective is motivated by the more general question of discovering the underlying causal variables (and their dependencies) that explain the environment of a learner and make it possible for that learner to plan appropriately. The discovery of underlying explanatory variables has come under different names, in particular the notion of disentangling underlying variables (Bengio et al., 2013). As stated already by Bengio et al. (2013) and clearly demonstrated by Locatello et al. (2018), assumptions, priors or biases are necessary to identify the underlying explanatory

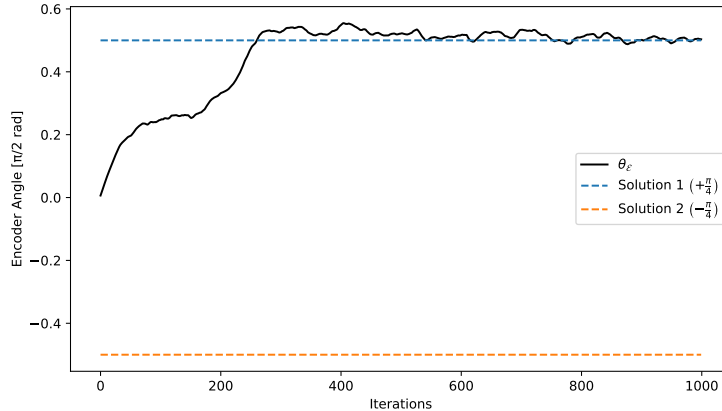


Figure 5: Evolution of the encoder parameter  $\theta_{\mathcal{E}}$  as training progresses. We set the parameter  $\theta_{\mathcal{D}}$  of the implicit decoder to  $-\frac{\pi}{4}$ : this corresponds to two valid solutions for the encoder, namely  $+\frac{\pi}{4}$  and  $-\frac{\pi}{4}$ . For the former, we obtain  $\theta = 0$  corresponding to the correct causal graph  $U \rightarrow V$ ; and for the latter,  $\theta = \frac{\pi}{2}$ , corresponding to the correct causal graph  $V \rightarrow U$ . The encoder parameter  $\theta_{\mathcal{E}}$  is trained jointly with the structural parameter, and we find that it converges to one of the two valid solutions. Further details and the corresponding evolution of the structural parameter can be found in Appendix H.

variables. The latter paper (Locatello et al., 2018) also reviews and evaluates recent work on disentangling, and discusses different metrics that have been proposed. An extreme view of disentangling is that the explanatory variables should be marginally independent, and many deep generative models (Goodfellow et al., 2016) and independent component analysis models (Hyvärinen et al., 2001; Hyvärinen et al., 2018) are built on this assumption. However, the kinds of high-level variables that we manipulate with natural language are not marginally independent: they are related to each other through statements that are usually expressed in sentences (e.g., a classical symbolic AI fact or rule), involving only a few concepts at a time. This kind of assumption has been proposed to help discover such linguistically relevant high-level representations from raw observations, such as the consciousness prior (Bengio, 2017), with the idea that humans focus at any particular time on just a few concepts that are present to our consciousness. The work presented here could provide an interesting meta-learning objective to help learn such encoders as well as figure out how the resulting variables are related to each other. In that case, one should distinguish two important assumptions: the first one is that the causal graph is sparse (has few edges, as in the consciousness prior (Bengio, 2017) and in some methods to learn Bayes net structure, e.g. (Schmidt et al., 2007)); and the second one is that it changes sparsely due to interventions (which is the focus of this work).

Approaches for Bayesian network structure learning based on discrete search over model structures and simulated annealing are reviewed in Heckerman et al. (1995). There, it has been common to use Minimum Description Length (MDL) principles to score and search over models Lam and Bacchus (1993); Friedman and Goldszmidt (1998), or the Bayesian Information Criterion (BIC) to search for models with high relative posterior probability Heckerman et al. (1995). Prior work such as Heckerman et al. (1995) has also relied upon purely observational data, without the possibility of interventions and therefore focused on learning likelihood or hypothesis equivalence classes for network structures. Since then, numerous methods have also been devised to infer the causal direction from purely observational data (Peters et al., 2017), based on specific, generally parametric assumptions, on the underlying causal graph. Pearl’s seminal work on do-calculus Pearl (1995, 2009); Bareinboim and Pearl (2016) lays a foundation for expressing the impact of interventions on probabilistic graphical models – we use it in our work. In contrast, here we are proposing a meta-learning objective function for learning causal structure, not requiring any specific constraints on causal graph structure, only on the sparsity of the changes in distribution in the correct causal graph parametrization.

Our work is also related to other recent advances in causation, domain adaptation and transfer learning. Magliacane et al. (2018) have sought to identify a subset of features that lead to the best predictions for a variable of interest in a source domain such that the conditional distribution of the variable of interest given these features is the same in the target domain. Johansson et al. (2016) examine counterfactual inference and formulate it as a domain adaptation problem. Shalit et al. (2017) propose a technique called counterfactual regression for estimating individual treatment effects from observational data. Rojas-Carulla et al. (2018) propose a method to find an optimal subset that makes the target independent from the selection variables. To do so, they make the assumption that if the conditional distribution of the target given some subset is invariant across different source domains, then this conditional distribution must also be the same in the target domain. Parascandolo et al. (2017) propose an algorithm to recover a set of independent causal mechanisms by establishing competition between mechanisms, hence driving specialization. Alet et al. (2018) proposed a meta learning algorithm to recover a set of specialized modules, but did not establish any connections to causal mechanisms. More recently, Dasgupta et al. (2019) adopted a meta-learning approach to draw causal inferences from purely observational data.

## 5. Conclusion and Future Work

We have established in very simple bivariate settings that the rate at which a learner adapts to sparse changes in the distribution of observed data can be exploited to select or optimize causal structure and disentangle the causal variables. **This relies on the assumption that with the correct causal structure, those distributional changes are localized and sparse.** We have demonstrated these ideas through theoretical results as well as experimental validation. See <https://github.com/authors-1901-10912/A-Meta-Transfer-Objective-For-Learning-To-Disentangle-Causal-Mechanisms> for source code of the experiments.

This work is only a first step in the direction of optimizing causal structure based on **the speed of adaptation to modified distributions**. On the experimental side, many settings other than those studied here should be considered, with different kinds of parametrizations, richer and larger causal graphs, different kinds of optimization procedures, etc. Also, much more needs to be done in exploring how the proposed ideas can be used to learn good representations in which the causal variables are disentangled, since we have only experimented at this point with the simplest possible encoder with a single degree of freedom. Scaling up these ideas would permit their application towards improving the way in which learning agents deal with non-stationarities, and thus improving sample complexity and robustness of learning agents.

## Acknowledgements

We would like to acknowledge support for this project from NSERC, CIFAR and Canada Research Chairs, as well as the feedback from Rémi Le Priol, Isabelle Lacroix, Alexandre Piché, and Akram Erraqabi. AG would like to thank Sergey Levine, Chelsea Finn, Michael Chang, Abhishek Gupta for useful discussions.

## References

- Ferran Alet, Tomás Lozano-Pérez, and Leslie P Kaelbling. Modular meta-learning. *arXiv preprint arXiv:1806.10166*, 2018.
- Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.
- Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1798–1828, 2013.
- Christopher M Bishop. Mixture density networks. Technical report, Citeseer, 1994.

- Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal Reasoning from Meta-reinforcement Learning. *arXiv preprint arXiv:1901.08162*, 2019.
- Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3), 1989.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *International Conference on Machine Learning (ICML)*, 2017.
- Nir Friedman and Moises Goldszmidt. Learning bayesian networks with local structure. In *Learning in graphical models*, pages 421–459. Springer, 1998.
- Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL <http://deeplearningbook.org>.
- David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.
- Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley-Interscience, May 2001. ISBN 047140540X.
- Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *CoRR*, arXiv:1805.08651, 2018.
- Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.
- Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.
- Wai Lam and Fahiem Bacchus. Using causal information and local measures to learn bayesian networks. In *Proceedings of the Ninth international conference on Uncertainty in artificial intelligence*, pages 243–250. Morgan Kaufmann Publishers Inc., 1993.
- Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *CoRR*, arXiv:1811.12359, 2018.
- Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10869–10879, 2018.
- Alex Nichol, Joshua Achiam, and John Schulman. On First-Order Meta-Learning Algorithms. *arXiv:1803.02999*, 2018.
- Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. *arXiv preprint arXiv:1712.00961*, 2017.
- Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.
- Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.
- Mark W. Schmidt, Alexandru Niculescu-Mizil, and Kevin P. Murphy. Learning graphical model structure using l1-regularization paths. In *AAAI*, 2007.

- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - from Theory to Algorithms*. Cambridge University Press, 2014.
- Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.
- V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

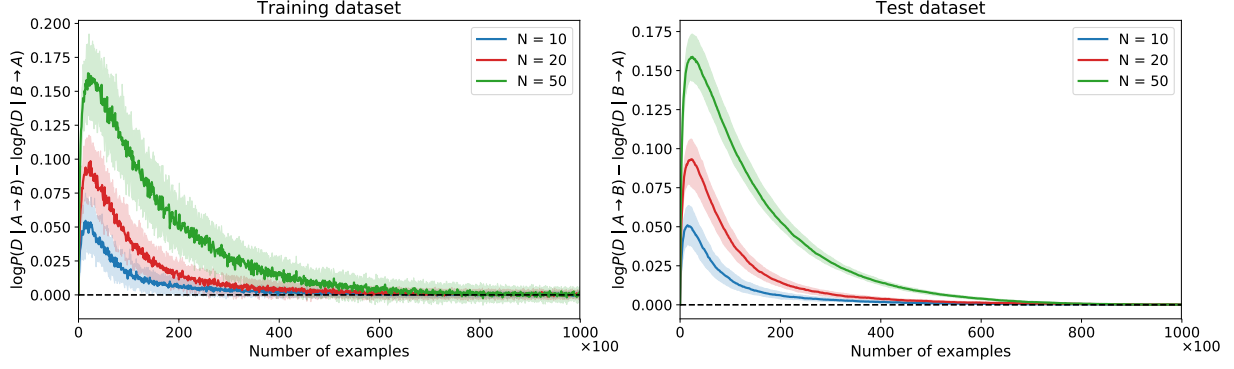


Figure A.1: Difference in log-likelihoods between the two models  $A \rightarrow B$  and  $B \rightarrow A$  on training and test data from the same distribution on discrete data, for different values of  $N$ , the number of discrete values per variable. Once fully trained, both models become indistinguishable from their log-likelihoods only, even on test data. The solid curves represent the median values over 100 different runs, and the shaded areas their 25-75 quantiles.

## Appendix A. Results on Non-Identifiability of Causal Structure

We show here that the maximum likelihood estimation of both models specified in Equation (1) yields the same estimated distribution over  $A$  and  $B$ , i.e., the joint likelihood on the training distribution is not sufficient to distinguish the  $A \rightarrow B$  and  $B \rightarrow A$  causal models, in the non-parametric case (no assumption at all on the family of distributions). Let

$$\begin{aligned} \theta_i &= P_{A \rightarrow B}(A = i) & \theta_{j|i} &= P_{A \rightarrow B}(B = j \mid A = i) \\ \eta_j &= P_{B \rightarrow A}(B = j) & \eta_{i|j} &= P_{B \rightarrow A}(A = i \mid B = j). \end{aligned}$$

We now state the maximum likelihood estimators for each models:

$$\begin{aligned} \hat{\theta}_i &= n_i/n & \hat{\theta}_{j|i} &= n_{ij}/n_i \\ \hat{\eta}_j &= n_j/n & \hat{\eta}_{i|j} &= n_{ij}/n_j \end{aligned} \tag{9}$$

where  $n$  is the total number of observations,  $n_i$  the number of times we observed  $A = i$ ,  $n_j$  the number of times we observed  $B = j$  and  $n_{ij}$  the number of times we observed  $A = i$  and  $B = j$  jointly. We can now compute the likelihood for each model:

$$\begin{aligned} \hat{P}_{A \rightarrow B}(A, B) &= \hat{\theta}_i \hat{\theta}_{j|i} = n_{ij}/n \\ \hat{P}_{B \rightarrow A}(A, B) &= \hat{\eta}_j \hat{\eta}_{i|j} = n_{ij}/n \end{aligned} \tag{10}$$

which is what we intended to show. To illustrate this result, we also experiment with learning the modules for both models  $A \rightarrow B$  and  $B \rightarrow A$  with SGD. In Figure A.1, we show the difference in log-likelihoods between these two models, evaluated on training and test data sampled from the same distribution, during training. We can see that while the model  $A \rightarrow B$  fits the data faster than the other model (corresponding to a positive difference in Figure A.1), both models achieve the same log-likelihoods on both models at convergence. This shows that the two models are indistinguishable based on data sampled from the same distribution, even on test data.

## Appendix B. Proof of the Zero-Gradient Proposition

Let us restate more formally and prove Proposition 1.



**Proposition 1** Consider conditional probability modules  $P_{\theta_i}(V_i|\text{pa}(i, V, B_i))$  where  $B_{ij} = 1$  indicates that  $V_j$  is among the parents  $\text{pa}(i, V, B_i)$  of  $V_i$  in a directed acyclic causal graph. Consider ground truth training distribution  $P_1$  and transfer distribution  $P_2$  over these variables, and ground truth causal structure  $B$ . The joint log-likelihood  $\mathcal{L}(V)$  for a sample  $V$  with respect to the module parameters  $\theta$  decomposed into module parameters  $\theta_i$  is  $\mathcal{L}(V) = \sum_i \log P_{\theta_i}(V_i|\text{pa}(i, V, B_i))$ . If (a) a model has the correct causal structure  $B$ , and (b) it been trained perfectly on  $P_1$ , leading to estimated parameters  $\theta$ , and (c) the ground truth  $P_1$  and  $P_2$  only differ from each other only for some  $P(V_i|\text{pa}(i, V, B_i))$  for  $i \in C$ , then  $E_{V \sim P_2}[\frac{\partial \mathcal{L}(V)}{\partial \theta_i}] = 0$  for  $i \notin C$ .

**Proof** Let  $V_{-i}$  be the subset of  $V$  excluding  $V_i$ . We can simplify the expected gradient as follows.

$$\begin{aligned}
E_{V \sim P_2} \left[ \frac{\partial \mathcal{L}(V)}{\partial \theta_i} \right] &= \\
&\sum_V P_2(V) \sum_k \frac{\partial}{\partial \theta_i} \log P_{\theta_k}(V_k|\text{pa}(k, V, B_k)) \\
&= \mathbb{1}_{i \in C} \sum_{V_{-i}} P_2(V_{-i}) \sum_{V_i} P_2(V_i|\text{pa}(i, V, B_i)) \\
&\quad \frac{\partial}{\partial \theta_i} \log P_{\theta_i}(V_i|\text{pa}(i, V, B_i)) + \\
&\mathbb{1}_{i \notin C} \sum_{V_{-i}} P_2(V_{-i}) \sum_{V_i} P_1(V_i|\text{pa}(i, V, B_i)) \\
&\quad \frac{\partial}{\partial \theta_i} \log P_{\theta_i}(V_i|\text{pa}(i, V, B_i))
\end{aligned} \tag{11}$$

where the second equality is obtained because  $\theta_i$  does not influence module  $k \neq i$ , and  $P_2$  is the same  $P_1$  for conditionals with  $i \notin C$  (assumption (c)). Now for the special case of  $i \notin C$ , we obtain

$$\begin{aligned}
E_{V \sim P_2} \left[ \frac{\partial \mathcal{L}(V)}{\partial \theta_i} \right] &= \sum_{V_{-i}} P_2(V_{-i}) \sum_{V_i} P_1(V_i|\text{pa}(i, V, B_i)) \\
&\quad \frac{\partial}{\partial \theta_i} \log P_{\theta_i}(V_i|\text{pa}(i, V, B_i)) \\
&= \sum_{V_{-i}} P_2(V_{-i}) \sum_{V_i} P_{\theta_i}(V_i|\text{pa}(i, V, B_i)) \\
&\quad \frac{\partial}{\partial \theta_i} \log P_{\theta_i}(V_i|\text{pa}(i, V, B_i)) \\
&= 0
\end{aligned} \tag{12}$$

where the second equality arises from assumption (b), and the last line from zeroing the inner sum via the general identity

$$\sum_v p_\theta(v) \frac{\partial}{\partial \theta} \log p_\theta(v) = \frac{\partial}{\partial \theta} \sum_v p_\theta(v) = \frac{\partial 1}{\partial \theta} = 0.$$

■

## Appendix C. Pseudo-Code

```

Draw initial meta-parameters of learner
Draw a training set from training distr.
Set causal structure to include all edges
Initialize learner parameters for this model
Pre-train the learner's parameters on the training set
Repeat  $J$  times
| Draw a transfer distr.
| Draw causal structure(s) according to meta-parameters
| Repeat  $T$  times
| | Sample minibatch from transfer distribution
| | Accumulate online log-likelihood of minibatch
| | Update the model parameters accordingly
|
Compute the meta-parameters gradient estimator
Update the meta-parameters by SGD
Optionally reset parameters to pre-training value

```

**Algorithm 1:** Meta-Transfer Learning of Causal Structure

## Appendix D. Proof of the Structural Parameter Gradient Proposition

Let us restate more formally and prove Proposition 2.

**Proposition 2** *The gradient of the negative log-likelihood regret of the transfer data*

$$\mathcal{R} = -\log [\text{sigmoid}(\gamma)\mathcal{L}_{A \rightarrow B} + (1 - \text{sigmoid}(\gamma))\mathcal{L}_{B \rightarrow A}]$$

*with respect to the structural parameter  $\gamma$  (where  $\sigma(\gamma) = P(A \rightarrow B)$ ) is given by*

$$\frac{\partial \mathcal{R}}{\partial \gamma} = \sigma(\gamma) - P(A \rightarrow B \mid D_2), \quad (13)$$

*where  $D_2$  is the transfer data, and  $P(A \rightarrow B \mid D_2)$  is the posterior probability of the hypothesis  $A \rightarrow B$  (when the alternative is  $B \rightarrow A$ ), defined by applying Bayes rule to  $P(D_2 \mid A \rightarrow B) = \prod_{t=1}^T P(a_t, b_t \mid A \rightarrow B, \theta_t) = \mathcal{L}_{A \rightarrow B}$ . Furthermore, this can be equivalently written as*

$$\frac{\partial \mathcal{R}}{\partial \gamma} = \sigma(\gamma) - \sigma(\gamma + \Delta), \quad (14)$$

*where  $\Delta = \log \mathcal{L}_{A \rightarrow B} - \log \mathcal{L}_{B \rightarrow A}$  is the difference between the log-likelihoods of the two hypotheses on transfer data  $D_2$ .*

**Proof** First note that, using Bayes rule,

$$\begin{aligned}
P(A \rightarrow B \mid D_2) &= \frac{P(D_2 \mid A \rightarrow B)P(A \rightarrow B)}{P(D_2 \mid A \rightarrow B)P(A \rightarrow B) + P(D_2 \mid B \rightarrow A)P(B \rightarrow A)} \\
&= \frac{\mathcal{L}_{A \rightarrow B}\sigma(\gamma)}{\mathcal{L}_{A \rightarrow B}\sigma(\gamma) + \mathcal{L}_{B \rightarrow A}(1 - \sigma(\gamma))} \\
&= \frac{\sigma(\gamma)\mathcal{L}_{A \rightarrow B}}{M},
\end{aligned} \quad (15)$$

where  $M = \sigma(\gamma)\mathcal{L}_{A \rightarrow B} + (1 - \sigma(\gamma))\mathcal{L}_{B \rightarrow A}$  is the online likelihood of the transfer data under the mixture, so that the regret is  $\mathcal{R} = -\log M$ . For the second line above, note that

$$\begin{aligned} P(D_2 | A \rightarrow B) &= \prod_{t=1}^T P(a_t, b_t | A \rightarrow B, \{(a_s, b_s)\}_{s=1}^{t-1}) \\ &= \prod_{t=1}^T P(a_t, b_t | A \rightarrow B, \theta_t) = \mathcal{L}_{A \rightarrow B} \end{aligned} \quad (16)$$

where  $\theta_t$  encapsulates the information about  $\{(a_s, b_s)\}_{s=1}^{t-1}$  (through some adaptation procedure). Since we only consider the two hypotheses  $A \rightarrow B$  and  $B \rightarrow A$ , we also have  $P(B \rightarrow A | D_2) = \frac{(1 - \sigma(\gamma))\mathcal{L}_{B \rightarrow A}}{M} = 1 - P(A \rightarrow B | D_2)$ . Then

$$\begin{aligned} \frac{\partial \mathcal{R}}{\partial \gamma} &= -\frac{\sigma(\gamma)(1 - \sigma(\gamma))\mathcal{L}_{A \rightarrow B} - \sigma(\gamma)(1 - \sigma(\gamma))\mathcal{L}_{B \rightarrow A}}{M} \\ &= \sigma(\gamma)P(B \rightarrow A | D_2) \\ &\quad - (1 - \sigma(\gamma))P(A \rightarrow B | D_2) \\ &= \sigma(\gamma) + \sigma(\gamma)P(A \rightarrow B | D_2) \\ &\quad - P(A \rightarrow B | D_2) - \sigma(\gamma)P(A \rightarrow B | D_2) \\ &= \sigma(\gamma) - P(A \rightarrow B | D_2) \end{aligned} \quad (17)$$

which concludes the first part of the proof. Moreover, in order to prove the equivalent formulation in Equation (14), it is sufficient to prove that  $P(A \rightarrow B | D_2) = \sigma(\gamma + \Delta)$ . Using the logit function  $\sigma^{-1}(z) = \log \frac{z}{1-z}$ , and the expression in Equation (15), we have

$$\begin{aligned} \sigma^{-1}(P(A \rightarrow B | D_2)) &= \log \frac{\sigma(\gamma)\mathcal{L}_{A \rightarrow B}}{M - \sigma(\gamma)\mathcal{L}_{A \rightarrow B}} \\ &= \log \frac{\sigma(\gamma)\mathcal{L}_{A \rightarrow B}}{(1 - \sigma(\gamma))\mathcal{L}_{B \rightarrow A}} \\ &= \underbrace{\log \frac{\sigma(\gamma)}{1 - \sigma(\gamma)}}_{=\gamma} + \underbrace{\log \frac{\mathcal{L}_{A \rightarrow B}}{\mathcal{L}_{B \rightarrow A}}}_{=\Delta} \\ &= \gamma + \Delta \end{aligned} \quad (18)$$

■

## Appendix E. Proof of the Proposition on the Convergence Point of Gradient Descent on the Structural Parameter

We use the same notation as in the above proof and statement.

**Proposition 3** *Stochastic gradient descent (with appropriately decreasing learning rate) on  $E_{D_2}[\mathcal{R}]$ , with  $\mathcal{R} = -\log [\text{sigmoid}(\gamma)\mathcal{L}_{A \rightarrow B} + (1 - \text{sigmoid}(\gamma))\mathcal{L}_{B \rightarrow A}]$  and with steps following  $\frac{\partial \mathcal{R}}{\partial \gamma}$  converges towards  $\text{sigmoid}(\gamma) = 1$  if  $E_{D_2}[\log \mathcal{L}_{A \rightarrow B}] > E_{D_2}[\log \mathcal{L}_{B \rightarrow A}]$ , or  $\sigma(\gamma) = 0$  otherwise.*

**Proof** We are going to consider the fixed point of gradient descent when the gradient is zero, since we already know that SGD converges with an appropriately decreasing learning rate. Let us introduce some notation to simplify the algebra:  $p = \text{sigmoid}(\gamma)$ ,  $M = p\mathcal{L}_{A \rightarrow B} + (1 - p)\mathcal{L}_{B \rightarrow A}$ , so  $\mathcal{R} = \log M$ , and define

$P_1 = \frac{p\mathcal{L}_{A \rightarrow B}}{M} = P(A \rightarrow B \mid D_2)$ , and  $P_2 = \frac{(1-p)\mathcal{L}_{B \rightarrow A}}{M} = 1 - P_1$ . Framing the stationary point in terms of  $p$  rather than  $\gamma$  gives us the inequality constraints  $-p \leq 0$  and  $p - 1 \leq 0$  and no equality constraint. Applying the KKT conditions with constraint functions  $-p$  and  $p - 1$  gives us

$$\begin{aligned} E_{D_2} \left[ \frac{\partial \mathcal{R}}{\partial p} \right] &= -\mu_1 + \mu_2 \\ \mu_i &\geq 0 \\ \mu_1 p &= 0 \\ \mu_2 (p - 1) &= 0 \end{aligned} \tag{19}$$

We already see from the last two equations that if  $p \in (0, 1)$  (i.e. excluding 0 and 1), we must have  $\mu_1 = \mu_2 = 0$ , i.e.,  $E[\frac{\partial \mathcal{R}}{\partial p}] = 0$  (with drop the  $D_2$  subscript on  $E$  when it is clear from context). Let us study that case first and show that it leads to an inconsistent set of equations (thus forcing the solution to be either  $p = 0$  or  $p = 1$ ). Let us rewrite the gradient to highlight  $p$  in it:

$$\begin{aligned} \frac{\partial \mathcal{R}}{\partial p} &= P(A \rightarrow B \mid D_2) - p \\ &= \frac{p\mathcal{L}_{A \rightarrow B}}{p\mathcal{L}_{A \rightarrow B} + (1-p)\mathcal{L}_{B \rightarrow A}} - p \\ &= \frac{p\mathcal{L}_{A \rightarrow B} - p(p\mathcal{L}_{A \rightarrow B} + (1-p)\mathcal{L}_{B \rightarrow A})}{M} \\ &= \frac{p(1-p)\mathcal{L}_{A \rightarrow B} - p(1-p)\mathcal{L}_{B \rightarrow A}}{M} \\ &= p(1-p) \frac{\mathcal{L}_{A \rightarrow B} - \mathcal{L}_{B \rightarrow A}}{M} \end{aligned} \tag{20}$$

The KKT conditions with the above two inequality constraints for  $0 \leq p \leq 1$  give

$$E \left[ \frac{\partial \mathcal{R}}{\partial p} \right] = \mu_2 - \mu_1. \tag{21}$$

If we consider the solutions  $p \in (0, 1)$  (i.e.,  $\mu_1 = \mu_2 = 0$ ) we now show that we get a contradiction. First note that to satisfy the above equation with  $\mu_1 = \mu_2 = 0$  means that either  $p = 0$  or  $p = 1$  (which is inconsistent with the assumption that  $p \in (0, 1)$ ) or that  $E \left[ \frac{\mathcal{L}_{A \rightarrow B} - \mathcal{L}_{B \rightarrow A}}{M} \right] = 0$ . Let us consider that equation, and since  $p \neq 0$  and  $p \neq 1$  we can either multiply by  $p$  or by  $1 - p$  on both sides. Assuming  $p \neq 0$  and multiplying by  $p$  gives

$$\begin{aligned} 0 &= E \left[ \frac{p(\mathcal{L}_{A \rightarrow B} - \mathcal{L}_{B \rightarrow A})}{M} \right] = E \left[ P_1 - \frac{p\mathcal{L}_{B \rightarrow A}}{M} \right] \\ &= E \left[ P_1 - \frac{\mathcal{L}_{B \rightarrow A} - \mathcal{L}_{B \rightarrow A} - p\mathcal{L}_{B \rightarrow A}}{M} \right] \\ &= E \left[ P_1 + \frac{\mathcal{L}_{B \rightarrow A}}{M} - P_2 \right] \\ &= E \left[ P_1 + \frac{\mathcal{L}_{B \rightarrow A}}{M} - (1 - P_1) \right] = E \left[ \frac{\mathcal{L}_{B \rightarrow A}}{M} - 1 \right]. \end{aligned} \tag{22}$$

For this equation to be satisfied, we need  $\mathcal{L}_{B \rightarrow A} = M$  all the time, since  $\mathcal{L}_{B \rightarrow A} \leq M$  by construction. This would however correspond to  $p = 0$ . Similarly, assuming  $p \neq 1$  we can multiply the stationarity equation by  $1 - p$  and get

$$\begin{aligned} 0 &= E \left[ \frac{(1-p)(\mathcal{L}_{A \rightarrow B} - \mathcal{L}_{B \rightarrow A})}{M} \right] \\ &= E \left[ \frac{(1-p)\mathcal{L}_{A \rightarrow B}}{M} - P_2 \right] \\ &= E \left[ \frac{\mathcal{L}_{A \rightarrow B}}{M} - P_1 - (1 - P_1) \right] = E \left[ \frac{\mathcal{L}_{A \rightarrow B}}{M} - 1 \right] \end{aligned} \tag{23}$$

Again, this can only be 0 if  $\mathcal{L}_{A \rightarrow B} = M$  all the time, i.e.,  $p = 1$ . We conclude that the solutions  $p \in (0, 1)$  are not possible because they would lead to inconsistent conclusions, which leaves only  $p = 0$  or  $p = 1$ . When  $p = 0$  we have  $E[\mathcal{R}] = E[\log \mathcal{L}_{A \rightarrow B}]$ , and when  $p = 1$  we have  $E[\mathcal{R}] = E[\log \mathcal{L}_{B \rightarrow A}]$ . Thus the minimum will be achieved at  $p = 1$  when  $E_{D_2}[\log \mathcal{L}_{A \rightarrow B}] > E_{D_2}[\log \mathcal{L}_{B \rightarrow A}]$ , or  $p = 0$  otherwise. ■

## Appendix F. More Than Two Causal Hypotheses

In this section, we consider one approach to generalize to more than two causal structures. We consider  $m$  variables, corresponding to  $O(2^{m^2})$  possible causal graphs, since each variable  $V_j$  could be (or not) a direct cause of any variable  $V_i$ , leading to  $m^2$  binary decisions. Note that a causal graph can in principle have cycles (if time is not measured with sufficient precision), although having a directed acyclic graph allows a much simpler sampling procedure (ancestral sampling). In our experiments the ground truth graph will always be directed, to make sampling easier and faster, but the learning procedure will not directly assume that. Motivated by the mechanism independence assumption, we propose a heuristic to learn the causal graph in which we independently parametrize the binary probability  $p_{ij}$  that  $V_j$  is a parent (direct cause) of  $V_i$ . As was the case for Section 2, we parametrize this Binomial distribution via binary edges  $B_{ij}$  that specify the graph structure:

$$\begin{aligned} B_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ P(B) &= \prod_{ij} P(B_{ij}). \end{aligned} \tag{24}$$

where  $p_{ij} = \text{sigmoid}(\gamma_{ij})$ . Let us define the parents of  $V_i$ , given  $B$ , as the set of  $V_j$ 's such that  $B_{ij} = 1$ :

$$\text{pa}(i, V, B_i) = \{V_j \mid B_{ij} = 1, j \neq i\} \tag{25}$$

where  $B_i$  is the bit vector with elements  $B_{ij}$  (and  $B_{ii} = 0$  is ignored). Similarly, we could parametrize the causal graph with a structural causal model where some of the inputs (from variable  $j$ ) of each function (for variable  $i$ ) can be ignored with some probability  $p_{ij}$ :

$$V_i = f_i(\theta_i, B_i, V, N_i) \tag{26}$$

where  $N_i$  is an independent noise source to generate  $V_i$  and  $f_i$  parametrizes the generator (as in a GAN), while not being allowed to use variable  $V_j$  unless  $B_{ij} = 1$  (and of course not being allowed to use  $V_i$ ). We can consider that  $f_i$  is a kind of neural network similar to the denoising auto-encoders or with dropout on the input, where  $B_i$  is a binary mask vector that prevents  $f_i$  from using some of the  $V_j$ 's (for which  $B_{ij} = 0$ ).

The conditional likelihood  $P_{B_i}(V_i = v_{ti} \mid \text{pa}(i, v_t, B_i))$  measures how well the model that uses the incoming edges  $B_i$  for node  $i$  performs for example  $v_t$ . We build a multiplicative (or exponentiated) form of regret by multiplying these likelihoods as  $\theta_t$  changes during an adaptation episode, for node  $i$ :

$$\mathcal{L}_{B_i} = \prod_t P_{B_i}(V_i = v_{ti} \mid \text{pa}(i, v_t, B_i)). \tag{27}$$

The overall exponentiated regret for the given graph structure  $B$  is  $\mathcal{L}_B = \prod_i \mathcal{L}_{B_i}$ . Similarly to the bivariate case, we want to consider a mixture over all the possible graph structures, but where each component must explain the whole adaptation sequence, thus we define as a loss for the generalized multi-variable case

$$\mathcal{R} = -\log E_B[\mathcal{L}_B] \tag{28}$$

Note the expectation over the  $2^{m^2}$  possible values of  $B$ , which is intractable. However, we can still get an efficient stochastic gradient estimator, which can be computed separately for each node of the graph (with samples arising only out of  $B_i$ , the incoming edges into  $V_i$ ):

**Proposition 4** *The overall regret (Equation (28)) rewrites*

$$\mathcal{R} = - \sum_i \log \sum_{B_i} P(B_i) \mathcal{L}_{B_i} \quad (29)$$

and if we are willing to consider multiple samples of  $B$  in parallel, a biased but asymptotically unbiased (as the number  $K$  of these samples  $B^{(k)}$  increases to infinity) estimator of the gradient of the overall regret with respect to meta-parameters can be defined:

$$g_{ij} = \frac{\sum_k (\sigma(\gamma_{ij}) - B_{ij}^{(k)}) \mathcal{L}_{B_i}^{(k)}}{\sum_k \mathcal{L}_{B_i}^{(k)}} \quad (30)$$

where the  $^{(k)}$  index indicates the values obtained for the  $k$ -th draw of  $B$ .

**Proof** Recall that  $\mathcal{L}_B = \prod_i \mathcal{L}_{B_i}$  so we can rewrite the regress loss as follows:

$$\begin{aligned} \mathcal{R} &= -\log E_B[\mathcal{L}_B] \\ &= -\log \sum_B P(B) \mathcal{L}_B \\ &= -\log \sum_{B_1} \sum_{B_2} \dots \sum_{B_M} \prod_i P(B_i) \mathcal{L}_{B_i} \\ &= -\log \prod_i \left( \sum_{B_i} P(B_i) \mathcal{L}_{B_i} \right) \\ &= - \sum_i \log \sum_{B_i} P(B_i) \mathcal{L}_{B_i} \end{aligned} \quad (31)$$

So the regret gradient on meta-parameters  $\gamma_i$  of node  $i$  is

$$\begin{aligned} \frac{\partial \mathcal{R}}{\partial \gamma_i} &= - \frac{\sum_{B_i} P(B_i) \mathcal{L}_{B_i} \frac{\partial \log P(B_i)}{\partial \gamma_i}}{\sum_{B_i} P(B_i) \mathcal{L}_{B_i}} \\ &= - \frac{E_{B_i}[\mathcal{L}_{B_i} \frac{\partial \log P(B_i)}{\partial \gamma_i}]}{E_{B_i}[\mathcal{L}_{B_i}]} \end{aligned} \quad (32)$$

Note that with the sigmoidal parametrization of  $P(B_{ij})$ ,

$$\log P(B_{ij}) = B_{ij} \log \text{sigmoid}(\gamma_{ij}) + (1 - B_{ij}) \log(1 - \text{sigmoid}(\gamma_{ij}))$$

as in the cross-entropy loss. Its gradient can similarly be simplified to

$$\begin{aligned} \frac{\partial \log P(B_{ij})}{\partial \gamma_{ij}} &= \frac{B_{ij}}{\text{sigmoid}(\gamma_{ij})} \text{sigmoid}(\gamma_{ij})(1 - \text{sigmoid}(\gamma_{ij})) \\ &\quad - \frac{(1 - B_{ij})}{(1 - \text{sigmoid}(\gamma_{ij}))} \text{sigmoid}(\gamma_{ij})(1 - \text{sigmoid}(\gamma_{ij})) \\ &= B_{ij} - \text{sigmoid}(\gamma_{ij}) \end{aligned} \quad (33)$$

A biased but asymptotically unbiased estimator of  $\frac{\partial \mathcal{R}}{\partial \gamma_{ij}}$  is thus obtained by sampling  $K$  graphs (over which the means below are run):

$$g_{ij} = \sum_k (\sigma(\gamma_{ij}) - B_{ij}^{(k)}) \frac{\mathcal{L}_{B_i}^{(k)}}{\sum_{k'} \mathcal{L}_{B_i}^{(k')}} \quad (34)$$

where index  $^{(k)}$  indicates the  $k$ -th draw of  $B$ , and we obtain a weighted sum of the individual binomial gradients weighted by the relative regret of each draw  $B_i^{(k)}$  of  $B_i$ , leading to Equation (30). ■



This decomposition is good news because the loss is a sum of independent terms, one per node  $i$ , depending only of  $B_i$  and and similarly  $g_{ij}$  only depends on  $B_i$  rather than the full graph structure. We use the estimator from Equation (30) in the general pseudo-code for meta-transfer learning of causal structure displayed in Algorithm 1.

## Appendix G. Results on Learning which is Cause and which is Effect

In order to assess the performance of our meta-learning algorithm, we applied it on generated data from three different domains: discrete random variables, multimodal continuous random variables and multivariate gaussian-distributed variables. In this section, we describe the setups for all three experiments, along with additional results to complement the results described in the main text. Note that in all these experiments, we fix the structure of the ground-truth to be  $A \rightarrow B$ , and only perform interventions on the cause  $A$ .

### G.1 Discrete variables and Two Causal Hypotheses

We consider a bivariate model, where both random variables are sampled from a categorical distribution. The underlying ground-truth model can be described as

$$\begin{aligned} A &\sim \text{Categorical}(\pi_A) \\ B \mid A = a &\sim \text{Categorical}(\pi_{B|a}), \end{aligned} \quad (35)$$

with  $\pi_A$  is a probability vector of size  $N$ , and  $\pi_{B|a}$  is a probability vector of size  $N$ , which depends on the value of the variable  $A$ . In our experiment, each random variable can take one of  $N = 10$  values. Since we are working with only two variables, the only two possible models are:

- *Model  $A \rightarrow B$* :  $P(A, B) = P(A)P(B \mid A)$
- *Model  $B \rightarrow A$* :  $P(A, B) = P(B)P(A \mid B)$

We build 4 different modules, corresponding to the model of each possible marginal and conditional distribution. These modules' definition and their corresponding parameters are shown in Table G.1.

	Distribution	Module	Parameters	Dimension
<i>Model <math>A \rightarrow B</math></i>	$P(A)$	$P(x_A = i; \theta_A) = [\text{softmax}(\theta_A)]_i$	$\theta_A$	$N$
	$P(B \mid A)$	$P(x_B = j \mid x_A = i; \theta_{B A}) = [\text{softmax}(\theta_{B A}(i))]_j$	$\theta_{B A}$	$N^2$
<i>Model <math>B \rightarrow A</math></i>	$P(B)$	$P(x_B = j; \theta_B) = [\text{softmax}(\theta_B)]_j$	$\theta_B$	$N$
	$P(A \mid B)$	$P(x_A = i \mid x_B = j; \theta_{A B}) = [\text{softmax}(\theta_{A B}(j))]_i$	$\theta_{A B}$	$N^2$

Table G.1: Description of the 2 models, with the parametrization of each module, for a bivariate model with discrete random variables. *Model  $A \rightarrow B$*  and *Model  $B \rightarrow A$*  both have the same number of parameters  $N^2 + N$ .

In order to get a set of initial parameters, we first train all 4 modules on a training distribution. This training distribution corresponds to a fixed choice of  $\pi_A^{(1)}$  and  $\pi_{B|a}$  (for all  $N$  possible values of  $a$ ). Note that the superscript in  $\pi_A^{(1)}$  emphasizes the fact that this defines the distribution prior to intervention, with the mechanism  $P(B \mid A)$  being unchanged by the intervention. These probability vectors are sampled randomly from a uniform Dirichlet distribution

$$\begin{aligned} \pi_A^{(1)} &\sim \text{Dirichlet}(\mathbf{1}_N) \\ \pi_{B|a} &\sim \text{Dirichlet}(\mathbf{1}_N) \quad \forall a \in [1, N]. \end{aligned} \quad (36)$$

Given this initial training distribution, we can sample a large dataset of training examples  $\{(a_i, b_i)\}_{i=1}^n$  from the ground-truth model, using ancestral sampling.

$$\begin{aligned} a &\sim \text{Categorical}(\pi_A^{(1)}) \\ b &\sim \text{Categorical}(\pi_{B|a}). \end{aligned} \quad (37)$$

Using this large dataset from the training distribution, we can train all 4 modules using gradient descent, or any other advanced first-order optimizer, like RMSprop. The parameters  $\theta_A$ ,  $\theta_{B|A}$ ,  $\theta_B$  &  $\theta_{A|B}$  of the different modules found after this initial training will be used as the initial parameters for the adaptation on a new transfer distribution.

Similar to the way we defined the training distribution, we can define a transfer distribution as a soft intervention on the random variable  $A$ . In this experiment, this accounts for changing the distribution of  $A$ , that is with a new probability vector  $\pi_A^{(2)}$ , also sampled randomly from a uniform Dirichlet distribution

$$\pi_A^{(2)} \sim \text{Dirichlet}(\mathbf{1}_N) \quad (38)$$

To perform adaptation on the transfer distribution, we also sample a smaller dataset of *transfer* examples  $D_2 = \{(a_i, b_i)\}_{i=1}^m$ , with  $m \ll n$  the size of the training set. In our experiment, we used  $m = 20$  transfer examples. We also used ancestral sampling on this new transfer distribution to acquire samples, similar to Equation (37) (with  $\pi_A^{(2)}$  instead of  $\pi_A^{(1)}$ ).

Starting from the parameters estimated after the initial training on the training distribution, we perform a few steps of adaptation on the modules parameters  $\theta_A$ ,  $\theta_{B|A}$ ,  $\theta_B$  &  $\theta_{A|B}$  using  $T$  steps of gradient descent based on the transfer dataset  $D_2$ . The value of the likelihoods for both models is recorded as well, and computed as

$$\begin{aligned} \mathcal{L}_{A \rightarrow B} &= \prod_{t=1}^T P(\mathbf{a}_t \mid \theta_A^{(t)}) P(\mathbf{b}_t \mid \mathbf{a}_t; \theta_{B|A}^{(t)}) \\ \mathcal{L}_{B \rightarrow A} &= \prod_{t=1}^T P(\mathbf{b}_t \mid \theta_B^{(t)}) P(\mathbf{a}_t \mid \mathbf{b}_t; \theta_{A|B}^{(t)}), \end{aligned} \quad (39)$$

where  $(\mathbf{a}_t, \mathbf{b}_t)$  represents a mini-batch of examples from  $D_2$ , and the superscript  $t$  on the parameters highlights the fact that these likelihoods are computed after  $t$  steps of adaptation. This product over  $t$  ensures that we monitor the progress of adaptation along the whole trajectory. In this experiment, we used  $T = 2$  steps of gradient descent on mini-batch of size 10 for the adaptation.

Finally, in order to update the structural parameter  $\gamma$ , we can use Proposition 2 to compute the gradient of the loss  $L$  with respect to  $\gamma$ :

$$\mathcal{R}(\gamma) = -\log[\sigma(\gamma)\mathcal{L}_{A \rightarrow B} + (1 - \sigma(\gamma))\mathcal{L}_{B \rightarrow A}] \quad (40)$$

$$\frac{\partial \mathcal{R}}{\partial \gamma} = \sigma(\gamma + \Delta) - \sigma(\gamma), \quad (41)$$

where  $\Delta = \log \mathcal{L}_{A \rightarrow B} - \log \mathcal{L}_{B \rightarrow A}$ . The update of  $\gamma$  can be one step of gradient descent, or using any first-order optimizer like RMSprop. We perform multiple interventions over the course of meta-training by sampling multiple transfer distributions, and following the same steps of adaptation and update of the structural parameter  $\gamma$ .

In Figure 2, we report the evolution of the structural parameter  $\gamma$  (or rather,  $\sigma(\gamma)$ ) as a function of the number of meta-training steps or, similarly, the number of different interventions made on the causal model. The model's belief  $P(A \rightarrow B) = \sigma(\gamma)$  indeed converges to 1, proving that the algorithm was capable of recovering the correct causal direction  $A \rightarrow B$ .

## G.2 Discrete Variables with MLP Parametrization

We consider a bivariate model similar to the ones defined above, where each random variable is sampled from a categorical distribution. Instead of expressing probabilities in a tabular form, we train  $M = 2$  simple feed-forward neural networks (MLP), one per conditional variable. MLP  $i$  is the independent mechanism of causal variable  $i$  that determines the conditional probability of the  $N$  discrete choices for variable  $i$ , given its parents.

Each MLP receives  $M$  concatenated  $N$ -dimensional one-hot vectors, masked appropriately according to the chosen causal structure  $B$ , i.e., with the  $j$ -th input of the  $i$ -th MLP being multiplied by  $B_{ij}$ . Each

directed edge presence or absence is thus indicated by  $B_{ij}$ , with  $B_{ij} = 1$  if variable  $j$  is a direct causal parent of variable  $i$ . The MLP maps the  $MN$  input units through one hidden layer that contains  $H = 4M$  hidden units and a ReLU non-linearity, and then maps the  $H$  hidden units to  $N$  output units and a softmax representing a predicted categorical distribution.

The causal structure belief is specified by an  $M \times M$  matrix  $\gamma$ , with  $\sigma(\gamma_{ij})$  the estimated probability that variable  $i$  is directly caused by variable  $j$ . The causal structure  $B_{ij}$  is drawn from  $\text{Ber}(\gamma_{ij})$ , as per Algorithm 1. We generalize the estimator introduced for the 2-hypotheses case as per Appendix F, i.e., we use the gradient estimator in Equation 30.

To evaluate the correctness of the structure being learnt, we measure the cross entropy between the ground-truth SCM and the learned SCM. In Figure 3 we show this cross-entropy over different episodes of training for bivariate discrete distributions with either 10 categories or 100 categories. Both models are first pretrained for 100 examples with fully connected edges before starting training on the transfer distributions.

### G.3 Continuous Multimodal Variables

Consider a family of joint distributions  $P_\mu(A, B)$  over the causal variables  $A$  and  $B$  sampled from the structural causal model (SCM):

$$\begin{aligned} A &\sim P_\mu(A) = \mathcal{N}(\mu, \sigma^2 = 4) \\ B &:= f(A) + N_B \quad N_B \sim \mathcal{N}(\mu = 0, \sigma^2 = 1) \end{aligned} \quad (42)$$

where  $f$  is a randomly generated spline and  $N_B$  is sampled i.i.d from the unit-normal distribution.

To obtain the spline, we sample the  $K$  points  $\{x_k\}_{k=1}^K$  uniformly spaced from the interval  $[-R_A, R_A]$ , and another  $K$  points  $\{y_k\}_{k=1}^K$  uniform randomly from the interval  $[-R_B, R_B]$ . This yields  $K$  pairs  $\{(x_k, y_k)\}_{k=1}^K$ , which make the knots of a second-order spline. We set  $K = 8$ ,  $R_A = R_B = 8$  for our experiments. In Figure G.1, we plot samples from one such SCM for the training distribution ( $\mu = 0$ ) and two transfer distributions ( $\mu = \pm 4$ ).

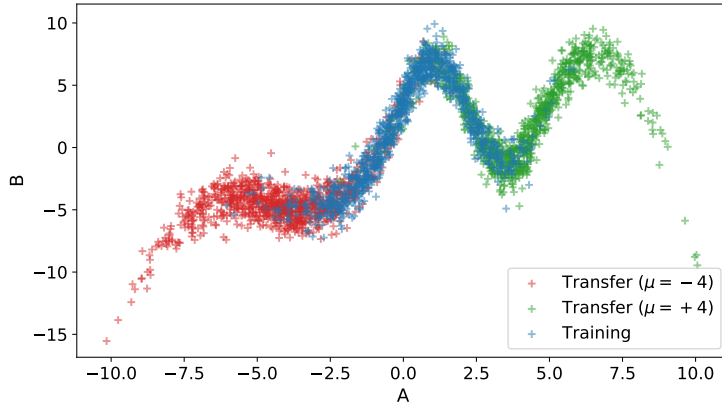


Figure G.1: Train (red) and transfer (green and blue) samples from an SCM generated with the procedure described in Equation (42). The green data-points are sampled from  $P_{\mu=(-4)}(A, B)$ , whereas the blue data-points are samples from  $P_{\mu=(+4)}(A, B)$  and the red data points (training set) are from  $P_{\mu=0}(A, B)$ .

The conditionals  $P(B | A; \theta_{B|A})$  and  $P(A | B; \theta_{A|B})$  are parameterized as 2-layer Mixture Density Networks Bishop (1994) with 32 hidden units and 10 components. The marginals  $P(A | \theta_A)$  and  $P(B | \theta_B)$  are parameterized as Gaussian Mixture Models, also with 10 components. The training now follows as described below.

Similar to Appendix G.1, we first pre-train the modules corresponding to the conditionals and marginals on the training distribution. To that end, we select  $P_{\mu=0}(A, B)$  as the training distribution, sample a

(large) training dataset  $\{(a_i, b_i)\}_{i=1}^n$  from it using ancestral sampling, and solve the following two problems independently until convergence:

$$\max_{\theta_A, \theta_{B|A}} \sum_{i=1}^n \log P(a_i | \theta_A) P(b_i | a_i; \theta_{B|A}) \quad (43)$$

$$\max_{\theta_B, \theta_{A|B}} \sum_{i=1}^n \log P(b_i | \theta_B) P(a_i | b_i; \theta_{A|B}) \quad (44)$$

The adaptation performance of  $A \rightarrow B$  and  $B \rightarrow A$  models can now be evaluated on transfer distributions. For a  $\mu$  sampled uniformly in  $[-4, 4]$ , we select  $P_\mu(A', B')$  as the transfer distribution, and denote with  $(A', B')$  samples from it. Both models are fine-tuned on  $P_\mu(A', B')$  for  $T = 10$  iterations (see Algorithm 1), and the area under the corresponding negative-log-likelihood curves becomes the regret:

$$\mathcal{R}_{A \rightarrow B} = - \sum_{t=1}^T \log P(B'|A'; \theta_{A \rightarrow B}^{(t)}) P(A' | \theta_{A \rightarrow B}^{(T)}) \quad (45)$$

and likewise for  $\mathcal{R}_{B \rightarrow A}$ . In these experiments, the modules corresponding to the marginals (ie. GMM) are learned *offline* via Expectation Maximization, and we denote with  $P(A' | \theta_{A \rightarrow B}^{(T)})$  the trained model. These can now be used to define the following meta-objective for the structural meta-parameter  $\gamma$ :

$$\mathcal{R}(\gamma) = \log[\sigma(\gamma)e^{\mathcal{R}_{A \rightarrow B}} + (1 - \sigma(\gamma))e^{\mathcal{R}_{B \rightarrow A}}] \quad (46)$$

The structural regret  $\mathcal{R}(\gamma)$  is now minimized with respect to  $\gamma$  for 200 iterations (updates of  $\gamma$ ). Figure G.2 shows the evolution of  $\sigma(\gamma)$  as training progresses. This is expected, given that we expect the causal model to perform better on the transfer distributions, i.e. we expect  $\mathcal{R}_{A \rightarrow B} < \mathcal{R}_{B \rightarrow A}$  in expectation. Consequently, assigning a larger weight to  $\mathcal{R}_{A \rightarrow B}$  optimizes the objective.

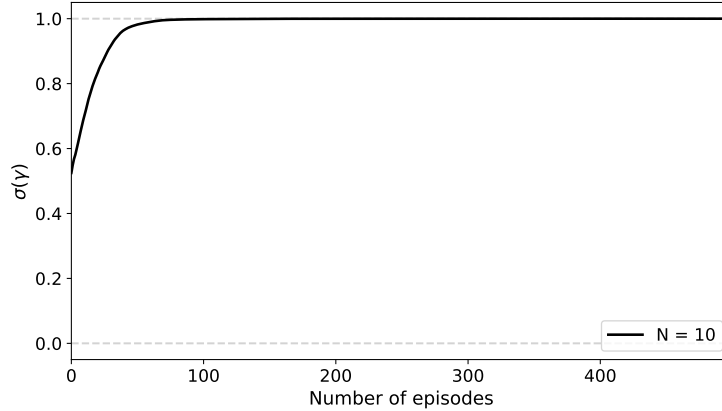


Figure G.2: Evolution of the sigmoid of structural meta-parameter  $\sigma(\gamma)$  with training iterations. It is indeed expected to increase if  $A \rightarrow B$  is the true causal graph (see Equation (46)).

#### G.4 Linear Gaussian Model

In this experiment, the two variables we consider are vectors (i.e.  $A \in \mathbb{R}^d$  and  $B \in \mathbb{R}^d$ ). The ground truth causal model is given by

$$\begin{aligned} A &\sim \mathcal{N}(\mu_A, \Sigma_A) \\ B &:= \beta_1 A + \beta_0 + N_B \quad N_B \sim \mathcal{N}(0, \Sigma_B) \end{aligned} \quad (47)$$

where  $\mu_A \in \mathbb{R}^d$ ,  $\beta_0 \in \mathbb{R}^d$  and  $\beta_1 \in \mathbb{R}^{d \times d}$ .  $\Sigma_A$  and  $\Sigma_B$  are  $d \times d$  covariance matrices<sup>2</sup>. In our experiments,  $d = 100$ . Once again, we want to identify the correct causal direction between  $A$  and  $B$ . To do so, we consider two models:  $A \rightarrow B$  and  $B \rightarrow A$ . We parameterize both models symmetrically:

$$\begin{aligned} P_{A \rightarrow B}(A) &= \mathcal{N}(A; \hat{\mu}_A, \hat{\Sigma}_A) \\ P_{A \rightarrow B}(B \mid A = a) &= \mathcal{N}(B; \hat{W}_1 a + \hat{W}_0, \hat{\Sigma}_{A \rightarrow B}) \\ P_{B \rightarrow A}(B) &= \mathcal{N}(B; \hat{\mu}_B, \hat{\Sigma}_B) \\ P_{B \rightarrow A}(A \mid B = b) &= \mathcal{N}(A; \hat{V}_1 b + \hat{V}_0, \hat{\Sigma}_{B \rightarrow A}) \end{aligned} \quad (48)$$

Note that each covariance matrix is parameterized using the Cholesky decomposition. Unlike previous experiments, we are not conducting any pre-training on actual data. Instead, we fix the parameters of both models to their exact values according to the ground truth parameters introduced in Equation 47. For model  $A \rightarrow B$ , this can be done trivially. For the second model, we can compute its exact parameters analytically. Once the exact parameters are set, both models are equivalent in the sense that  $P_{A \rightarrow B}(A, B) = P_{B \rightarrow A}(A, B) \forall A, B$ .

Each meta-learning episode starts by initializing the parameters of both models to the values identified during the pre-training. Afterward, a transfer distribution is sampled (i.e.  $\mu_A \sim \mathcal{N}(0, I)$ ). Then, both models are trained on samples from this distribution, for 10 iterations only. During this adaptation, the log-likelihoods of both models are accumulated in order to compute  $\mathcal{L}_{A \rightarrow B}$  and  $\mathcal{L}_{B \rightarrow A}$ . At this stage, we compute the meta objective estimate  $\mathcal{R} = -\log[\text{sigmoid}(\gamma)\mathcal{L}_{A \rightarrow B} + (1 - \text{sigmoid}(\gamma))\mathcal{L}_{B \rightarrow A}]$ , compute its gradient w.r.t.  $\gamma$  and update  $\gamma$ .

Figure G.3 shows that, after 200 episodes,  $\sigma(\gamma)$  converges to 1, indicating the success of the method on this particular task.

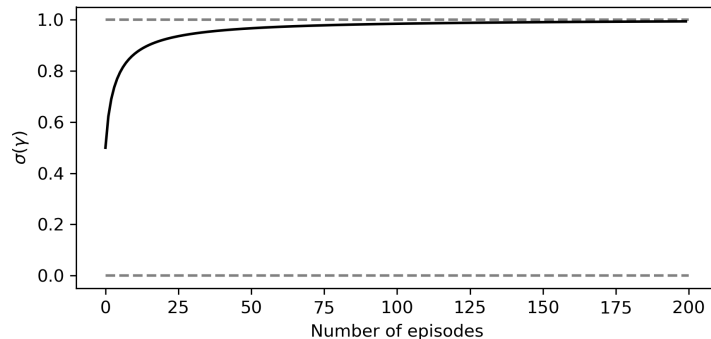


Figure G.3: Convergence of the causal belief (to the correct answer) as a function of the number of meta-learning episodes, for the linear Gaussian experiments.

## Appendix H. Results on Learning the Correct Encoder

The causal variables  $(A, B)$  are sampled from the distribution described in Eqn 42, and are mapped to observations  $(X, Y) \sim P_\mu(X, Y)$  via a hidden (and a priori unknown) decoder  $\mathcal{D} = R(\theta_{\mathcal{D}})$ , where  $R$  is a rotation matrix. The observations are then mapped to the hidden state  $(U, V) \sim P_\mu(U, V)$  via the encoder  $\mathcal{E} = R(\theta_{\mathcal{E}})$ . The computational graph is depicted in Figure 4.

<sup>2</sup>. Ground truth parameters  $\mu_A$ ,  $\beta_1$  and  $\beta_0$  are sampled from a Gaussian distribution, while  $\Sigma_A$  and  $\Sigma_B$  are sampled from an inverse Wishart distribution.

Analogous to Equation 46 in Appendix G.3, we now define the regret over the variables  $(U, V)$  instead of  $(A, B)$ :

$$\mathcal{R}(\gamma, \theta_{\mathcal{E}}) = \log[\sigma(\gamma)e^{\mathcal{R}_{U \rightarrow V}} + (1 - \sigma(\gamma))e^{\mathcal{R}_{V \rightarrow U}}] \quad (49)$$

where the dependence on  $\theta_{\mathcal{E}}$  is implicit in  $(U, V)$ . In every meta-training iteration, the  $U \rightarrow V$  and  $V \rightarrow U$  models are trained on the training distribution  $P_{\mu=0}(U, V)$  for  $T' = 20$  iterations. Subsequently, the regrets  $\mathcal{R}_{U \rightarrow V}$  and  $\mathcal{R}_{V \rightarrow U}$  are obtained by a process identical to that described in Equation 45 of Appendix G.3 (albeit with variables  $(U, V)$  and  $T = 5$ ). Finally, the gradients of  $\mathcal{R}(\gamma, \theta_{\mathcal{E}})$  are evaluated and the meta-parameters  $\gamma$  and  $\theta_{\mathcal{E}}$  are updated. This process is repeated for 1000 meta-iterations, and Figure 5 shows the evolution of  $\theta_{\mathcal{E}}$  as training progresses (where  $\theta_{\mathcal{D}}$  has been set to  $-\frac{\pi}{4}$ ). Further, Figure H.1 shows the corresponding evolution of the structural parameter  $\gamma$ .

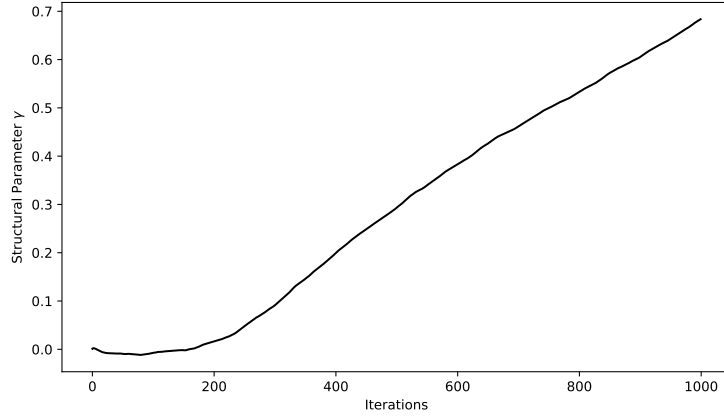


Figure H.1: Evolution of the structural parameter  $\gamma$  as training progresses with the encoder. The corresponding evolution of the encoder parameter  $\theta_{\mathcal{E}}$  is shown in Figure 5. Observe that the system converges to  $\theta = 0$ , implying that the correct causal direction is  $U \rightarrow V$  and the parameter  $\gamma$  should increase with meta-training iterations.