

# Probability trees and the value of a single intervention

Tue Herlau

Technical University of Denmark, 2800 Lyngby, Denmark  
tuhe@dtu.dk

## Abstract

The most fundamental problem in statistical causality is determining causal relationships from limited data. Probability trees, which combine prior causal structures with Bayesian updates, have been suggested as a possible solution. In this work, we quantify the information gain from a single intervention and show that both the anticipated information gain, prior to making an intervention, and the expected gain from an intervention have simple expressions. This results in an active-learning method that simply selects the intervention with the highest anticipated gain, which we illustrate through several examples. Our work demonstrates how probability trees, and Bayesian estimation of their parameters, offer a simple yet viable approach to fast causal induction.

## Introduction

Perhaps the most fundamental problem in statistical causality is the problem of *causal induction* itself, namely how to obtain the general causal relationships from limited observations (?).

Humans are able to infer causal relationships from samples that are too small for statistical tests to produce significant results (?). This ability is seemingly at odds with assumptions that are at the center of many popular algorithms for learning causal graphical models, which often require statistical testing of independence claims using large samples of observations (??).

Bayesian causal induction and probability trees offer an intuitively appealing possible answer to the question of how humans can learn rapidly from such small samples, namely as a combination of prior knowledge in the form of observations and Bayesian updating (?). This is accomplished by encoding alternative causal hypothesis directly in the probability tree, thereby making no distinction between random variables representing data and the causal hypothesis itself (??).

In this work, we show how probability trees and information theory can predict circumstances under which causal induction will occur quickly. We do this by defining the expected information gain of an intervention, and demonstrating how it can be estimated prior to performing the interven-

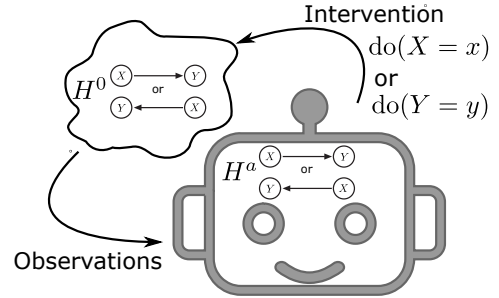


Figure 1: Either  $X$  is the cause of  $Y$  or  $Y$  is the cause of  $X$ , and we distinguish between the Agents hypothesis  $H^a$  about the causal orientation and the actual orientation  $H^0$ . Our overall goal is to quantify how much information can be gained about  $H^0$  on average from a single intervention.

tion, thereby leading to a natural active-learning method for causal induction.

We will focus on arguably the most basic form of causal learning (and certainly the problem which has garnered the most attention in psychology (?)), namely learning a single causal relationship  $X \rightarrow Y$  or  $Y \rightarrow X$  between two variables  $X \in \{1, \dots, K_X\}$  and  $Y \in \{1, \dots, K_Y\}$  in the very small sample limit.

We focus on two questions: (i) how much information will the agent *anticipate* from an intervention? and (ii) how much information will an agent actually *gain* from an intervention?. To quantify these expressions we distinguish between the agent's belief about the causal orientation  $H^a \in \{h_{X \rightarrow Y}^a, h_{Y \rightarrow X}^a\}$  and the true causal orientation governing the world  $H^0 \in \{h_{X \rightarrow Y}^0, h_{Y \rightarrow X}^0\}$  (see fig. 1). This is necessary, since a reduction in uncertainty about  $H^a$  is by itself meaningless, and we want to quantify the information gain about the true orientation  $H^0$  *regardless* of what it is.

While both  $H^0$  and  $H^a$  have simple expressions, the anticipated information gain is a Jeffrey divergence. Intuitively, what makes an intervention  $\text{do}(X = x)$  appealing is that there are  $Y = y$ -values such that  $P(y|x) \neq P(y)$ , however with *rare* events being more informative than common events. Priors will therefore be expected to play a more significant role in causal induction, a point we return to in

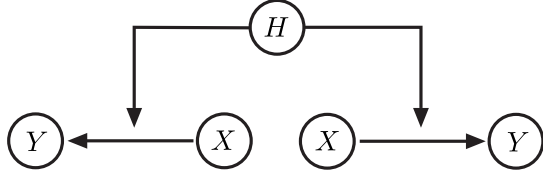


Figure 2: A probability tree for the problem involving  $X$  and  $Y$ . The variable  $H = h$  or  $H = \neg h$  determines the causal orientation of  $X$  and  $Y$ . Thus, given  $H$ , the effect of interventions is computed from the correct branch.

Example 3. Simulated examples<sup>1</sup> show that the gains can be estimated from small samples.

**Related work** The effect of a cause has been quantified using information theory (?), however, without considering learning in a Bayesian setting. Entropic causal inference (see (?)) specifies circumstances where the causal direction between categorical variables can be determined from observational data under assumptions of limited entropy. Information geometry has been proposed as a means to infer causal orientation by relying on distributional assumptions (?), however both settings are different from the Bayesian learning setting considered here. Targeted intervention selection has also been considered by (??) but using a different measure to quantify the goodness of an intervention.

## Methods

A probability tree describes the relationships between the causal hypothesis  $H$  and all variables (in this case  $X$  and  $Y$ ). In the tree, each internal node represents a causal mechanism, such that a path from the root node to a leaf corresponds to a realization of a causal mechanism (?).

The resolution order of the variables will typically depend on the path taken in the tree. In the left-most branch,  $X$  precedes  $Y$  under the hypothesis  $h_{X \rightarrow Y}$ , whereas in the right-most branch  $Y$  precedes  $X$ . Thus, the probability tree allows the agent to model different potential mechanisms for the data-generating procedure.

Using Pearls shorthand  $\hat{x} = \text{do}(X = x)$  for an intervention, we can easily predict the effect of an intervention on  $y$  from their joint distribution  $P(X, Y)$ . For instance,  $P(y|\hat{x}, h_{X \rightarrow Y}) = P(p|x)$  (?). Therefore, we can compute the posterior probability of  $h$  given an intervention  $\hat{x}$  resulted in observing  $Y = y$ :

$$\begin{aligned}
 P(h|\hat{x}, y) &= \frac{P(y|h, \hat{x})P(\hat{x}|h)P(h)}{\sum_{h \in \{h_{X \rightarrow Y}, h_{Y \rightarrow X}\}} P(y|h, \hat{x})P(\hat{x}|h)P(h)} \\
 &= \frac{P(y|x)P(h)}{P(y|x)P(h) + P(y|\neg h)P(\neg h)} \quad (1)
 \end{aligned}$$

It is thus clearly evident that interventions change the posterior belief about  $H$  as long as  $P(y|x) \neq p(y)$ , but that purely observational data does not (?).

<sup>1</sup>Code to reproduce all plots in this paper can be found at [https://gitlab.compute.dtu.dk/tuhe/probability\\_tree\\_learning](https://gitlab.compute.dtu.dk/tuhe/probability_tree_learning)

## Bayesian learning

Let's now consider the case in which the probabilities such as  $P(y|\hat{x}, h)$  are not known, but must be inferred from data.

The data will be given as a  $K_X \times K_Y$  matrix  $n$  such that  $n_{ij}$  is the number of times  $(X = i, Y = j)$  has been jointly observed. For clarity, we will consistently use  $Q$  to denote probabilities that are learned (from  $n$ ) and  $P$  for their actual value, as dictated by whatever mechanism governs the world. In that case, the Agent's belief about the causal orientation  $H^a$ , conditional on  $n$ , is given in eq. (1) as long as we condition all probabilities on  $n$ . In this case, the left-hand side is  $Q(h^a|\hat{x}, y, n)$  and we therefore need to compute expressions such as:

$$Q(h|n), \quad Q(y|h, \hat{x}, n), \quad Q(y|n). \quad (2)$$

To specify a suitable prior, we assume that observational data alone should remain uninformative about  $h$ , i.e. that  $Q(h|x, y, n) = Q(h)$ <sup>2</sup>. To ensure this, we must have

$$Q(y|h, x, n)Q(x|n, h) = Q(y|\neg h, x, n)Q(x|n, \neg h) \quad (3)$$

The natural way to ensure this is to specify the prior  $Q(x, y|h)$  without reference to  $h$ , and then define the above expressions as the marginals and conditionals of the posterior. Specifically, we define

$$(\theta_{x,y})_{x,y=1}^{K_X, K_Y} \sim \text{Dir}((\theta)_{xy} | \alpha 1_{K_X K_Y}), \quad Q(x, y|h) = \theta_{x,y}.$$

A standard application of Bayes' theorem gives us the standard posterior as  $Q_n(x, y) = Q(x, y|n) = \frac{n_{xy} + \alpha}{N + K_X K_Y \alpha}$  from which all relevant quantities follow. For instance:

$$Q(y|h_{X \rightarrow Y}, n, \hat{x}) = Q_n(y|x) = \frac{n_{yx} + \alpha}{\sum_y n_{xy} + K_Y \alpha} \quad (4a)$$

$$Q(y|h_{Y \rightarrow X}, n, \hat{x}) = Q_n(y) = \frac{\sum_x n_{xy} + K_X \alpha}{N + K_X K_Y \alpha}. \quad (4b)$$

## Information gain

If we suppose the agent performs an intervention  $\text{do}(X = x)$  and observes  $y$ , and assume the agent is a-priori uninformed  $P(H^a = h) = \frac{1}{2}$ , then the change in the agent's belief about the causal orientation  $h_{X \rightarrow Y}^a$ , measured in bits, is given by eq. (1) and eq. (5):

$$\begin{aligned}
 I(\hat{x}|y) &= \log \frac{Q(h_{X \rightarrow Y}^a|n, \hat{x}, y)}{Q(h_{Y \rightarrow X}^a|n, \hat{x}, y)} = \log \frac{Q(y|\hat{x}, n, h_{X \rightarrow Y}^a)}{Q(y|\hat{x}, n, h_{Y \rightarrow X}^a)} \\
 &= \log \frac{Q_n(y|x)}{Q_n(y)}. \quad (5)
 \end{aligned}$$

Recall that the true causal orientation is  $H^0$ . When  $H^0 = h_{X \rightarrow Y}^0$ , the above expression captures the gain in evidence for  $H^0$ , otherwise if  $H^0 = h_{Y \rightarrow X}^0$  the gain in evidence for

<sup>2</sup>This assumption is not innocent, as it is violated in work which assumes the causal orientation can be identified from observational data (?). This could be implemented by specifying  $(P(x|y, h), P(y|h)P(x|y, h))$  and  $(P(x|y, \neg h), P(x|\neg h))$  separately.

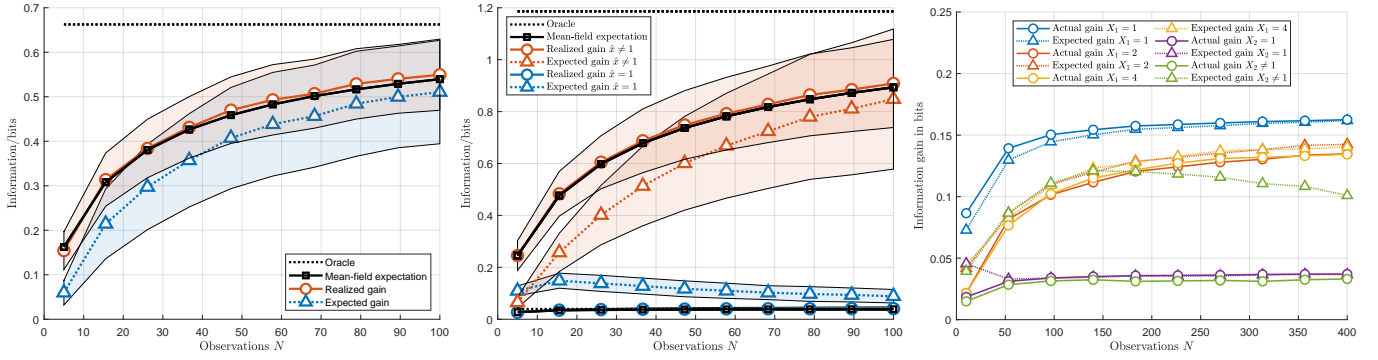


Figure 3: The plots show both the expected information gain (triangles) and the realized gain (circles) for Example 1 (left), Example 2 (middle) and Example 3 (right). The expected gain, eq. (7), signifies the agent’s estimate of the information gain from a given intervention  $\hat{x}$  or  $\hat{y}$ , i.e. the agent will decide based on this value (by symmetry, all possible interventions are represented in the figures). The realized gain (computed using eq. (6)), is the gain the agent will actually obtain. Each simulation is run 1000 times, using a varying number of observations  $N$ , and the shaded region indicates one standard deviation.

the *true* direction is  $-I(\hat{x}|y)$ . Thus, the information gain in the *true* direction is

$$\begin{aligned}\Delta_{\text{do}(X=x)} &= \sum_{h^0} P(h^0) \sum_{y|h^0} P(y|h^0, \hat{x}) (-1)^{\delta_{h^0, h^0_{X \rightarrow Y}} - 1} I(\hat{x}|y) \\ &= \frac{1}{2} \sum_y [P(y|x) - P(y)] I(\hat{x}|y).\end{aligned}\quad (6)$$

We refer to this as the *realized information gain*, as it represents the change in evidence for the true orientation the agent would *in fact* experience after performing  $\hat{x}$ .

Prior to the intervention  $\hat{x}$ , the agent can reason about the expected information gain. This can be computed by simply inserting  $Q$  instead of  $P$  in eq. (6). Upon simplification, we see that the information gain is the Jeffrey divergence  $D_J(p, q) = \sum_x (p(x) - q(x)) \log \frac{p(x)}{q(x)}$

$$\Delta_{\text{do}(X=x)}^{\text{Expected}} = D_J(Q_n(Y|x), Q_n(Y)). \quad (7)$$

We refer to this as the *expected information gain*, as it is the gained information about the true causal orientation the agent expects prior to performing  $\hat{x}$ .

**Example 1: Two correlated variables** We first consider the simplest possible case where  $K_X = K_Y = 2$ , the true causal orientation is  $H^0 = h_{X \rightarrow Y}^0$ , and  $P(x, y) = P_{xy}$ , considered as a  $2 \times 2$  matrix, is  $P = \begin{bmatrix} \rho & (1-\rho) \\ (1-\rho) & \rho \end{bmatrix}$  where  $\rho$  controls the correlation of  $X$  and  $Y$ . We first consider a *mean-field case* where  $n = \frac{N}{2} P$ . The realized information gain will be given by eq. (6)

$$\Delta_{\text{do}(X=x)} = \frac{1}{2} (\rho - \frac{1}{2}) \log \frac{N\rho + 2\alpha}{N(1-\rho) + 2\alpha}. \quad (8)$$

As expected, in the uncorrelated case  $\rho = \frac{1}{2}$  the information gain is 0. On the other hand, if the problem is deterministic  $\rho = 1$  the information gain is  $\Delta_{\text{do}(X=x)} = \frac{1}{4} \log \frac{N+2\alpha}{2\alpha}$ . In the case of two variables, the expected amount of information gained about the *true* causal orientation for a single intervention will therefore scale as  $\log(N)$ : the more we know

about the system, the more information can be gained from a single intervention.

A more important question is how this expression behaves when  $n$  is randomly generated. We examine this by sampling 1000 realizations of  $n$  using  $P$ , and use these to compute the average realized gain and expected gain, eq. (6) and eq. (7) respectively. We plot these along with the Bayesian update, which we obtain using the true probabilities as in eq. (1) (see fig. 3 (a)). The simulations use  $\alpha = 2$  and  $\rho = 0.9$ . The shaded area corresponds to one standard deviation; by symmetry both  $\hat{x}$  and  $\hat{y}$  behave the same. The realized gain is in this case larger than the expected gain due to the prior term  $\alpha$ . This can be understood by noting that the Jeffrey divergence, due to the log-term, is quite unstable for low probability events. This means that the prior for low-probability events can generally be expected to play a large role in causal inference, and an under-estimation in particular may lead to very over-confident updates. We return to this point in Example 3.

**Example 2: Exploration** This example illustrates how the method can guide exploration. Assume  $K_X = K_Y = 4$  and  $P_{11} = \rho$  and otherwise  $P_{xy} = \frac{1-\rho}{K^2-1}$ . By symmetry, we only need to consider  $\text{do}(X=1)$  and  $\text{do}(X=2)$ . The result, using the same settings as in Example 1, can be found in fig. 3 (b). Although  $\hat{x} = 1$  is by far the most likely event, it is not informative about the causal orientation, since  $P(y|x=1)$  is very nearly equal to  $P(y)$ . Note that in this case the expected gain is larger than the realized gain. Interestingly, with as little as 20 samples, the method will suggest an optimal intervention.

**Example 3: A single good intervention** To illustrate a case where a single intervention is better than all the other, consider a problem where  $P_{1,4} = P_{x,1} = \frac{\rho}{5}$  and otherwise  $P_{xy} = \frac{1-\rho}{11}$ . Both realized and expected gain of all interventions (that are not similar by symmetry) are visualized in fig. 3 (c), and  $\hat{x} = 1$  constitutes the (single) optimal intervention (the standard deviations are not shown for visual clarity). Since the variables are nearly independent, more sam-

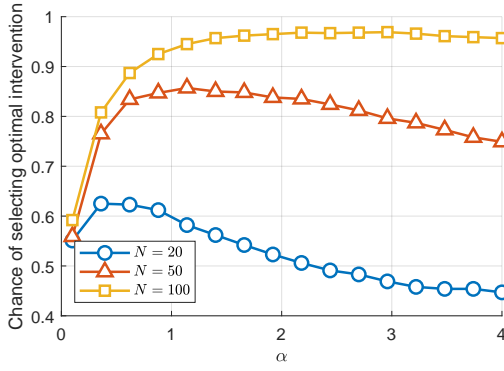


Figure 4: The problem described in Example 3, shown in fig. 3 (right), but where we consider the chance of selecting the single best intervention ( $\hat{x} = 1$ ) as a function of prior  $\alpha$  and for different number of observations. The prior is necessary to obtain stable estimates of the expected information gain, however, it will impact the estimate of the expected information gain slightly differently and therefore a large prior may change the ordering. The method selects the best intervention with a much higher probability than chance even for very low counts.

ples have been included, in order to show that the expected gain eventually converges to the mean-field value.

We note that the expected information gain can be overestimated in the small-sample limit (see e.g.  $\hat{y} \neq 1$ ) where the prior term  $\alpha$  will be more important. To gain more insight into this, we consider the same example, but now show the probability that the expected gain will be the largest for the optimal intervention  $\hat{x} = 1$ , i.e. the chance that the agent will actually select the optimal intervention (see fig. 4). This probability is plotted as a function of the regularization parameter  $\alpha$  for three representative numbers of samples. Although the ability to select the optimal intervention is more impacted by the prior in the small-sample limit, we note that even for just 20 samples, it is much higher than chance ( $\frac{1}{8}$ ).

**Example 4: Active learning** This example will consider a concrete Active learning setting and show that Bayesian causal induction can learn quicker when actions are selected based on eq. (7), compared to random selection. We consider the ground-truth as fixed at  $H_0 = h_{X \rightarrow Y}^0$ , and generate larger random problems by setting  $K_X = K_Y = 8$  and setting  $p_{xy} = \frac{u_{xy}}{\sum_{x,y} u_{xy}}$ , where  $u_{xy}$  are i.i.d. uniform random variables in  $[0, 1]$ . Given  $P_{xy}$ , we sample  $n$  as in the previous examples, and compute the information gain in favor of  $h_{X \rightarrow Y}^0$  based on an intervention  $\hat{x}$  or  $\hat{y}$  and corresponding observations of  $y$  and  $x$  using the Bayesian update given in eq. (5). We then consider the case where interventions are selected randomly, as well as the case where they are selected using the maximal expected information gain computed using eq. (7). The results are averaged over  $10^5$  simulations using  $\alpha = 2$ . In both cases, Bayesian causal induction gains information about the true causal orientation, however, the informative action selections result in about twice as large gain in evidence on average.

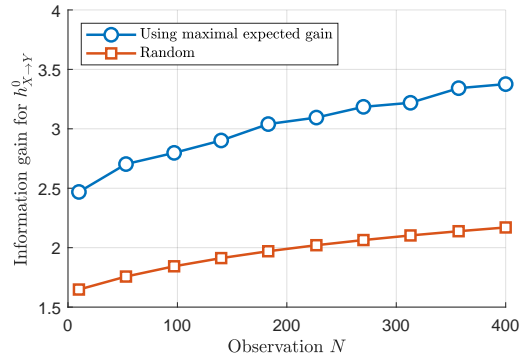


Figure 5: Evaluations of method for an actual intervention-selection problem. The truth is considered fixed as  $H^0 = h_{X \rightarrow Y}^0$ . From this,  $10^5$  random joint distributions  $P(X, Y)$  are generated (see text), and the information gains towards  $h_{X \rightarrow Y}^0$  are computed when interventions are either selected randomly, or when using the maximum anticipated gain eq. (7). As shown, the information gain is in both cases positive, but about twice as large when interventions are selected using our method.

## Discussion and conclusion

Probability trees are conceptually the simplest possible approach to causal inference: consider the causal orientation as an event, specify a prior, and compute the posterior. Although recent work has demonstrated how probability trees can represent both interventions and counterfactual, as well as context-dependent causal statements that cannot be expressed in a directed acyclic causal model (?), their practical use has remained limited.

In this work, we have highlighted another aspect of probability trees and Bayesian causal induction, namely the ability to make predictions about the information gain *prior* to performing interventions. We have illustrated this in the simplest possible situation, and shown how to express both the the expected gain *before* making an intervention, and the realized gain *after* making an intervention.

In making statements such as this, it is important to emphasize that a reduction in uncertainty is by itself meaningless, rather, what matters is the information gain in favor of the true hypothesis, a distinction we make using  $H^0$  and  $H^a$ .

In experiments, we have shown these measures can quantify the information gain and distinguish between different interventions. An active-learning example (fig. 5) shows an increased information gain when our method is used to select optimal interventions.

Many interesting avenues remain unexplored, such as the generalization to larger graphs, and concrete concentration bounds on the expected and realized information gains in terms of  $P$ .

Ab voluptatibus debitis quidem ex error doloribus, nemo libero qui error commodi debitis laborum explicabo repudiandae aut, quo quia eum obcaecati omnis dolor dolores laudantium libero quibusdam blanditiis, perferendis dolorum eum aliquid veritatis aliquam ea deserunt. Corrupti teneatur atque quis voluptatum pariatur a recusandae, cum ratione

molestiae commodi facilis porro minima repudiandae consequatur accusamus cupiditate.Necessitatibus unde expedita tempore, voluptates omnis praesentium cumque