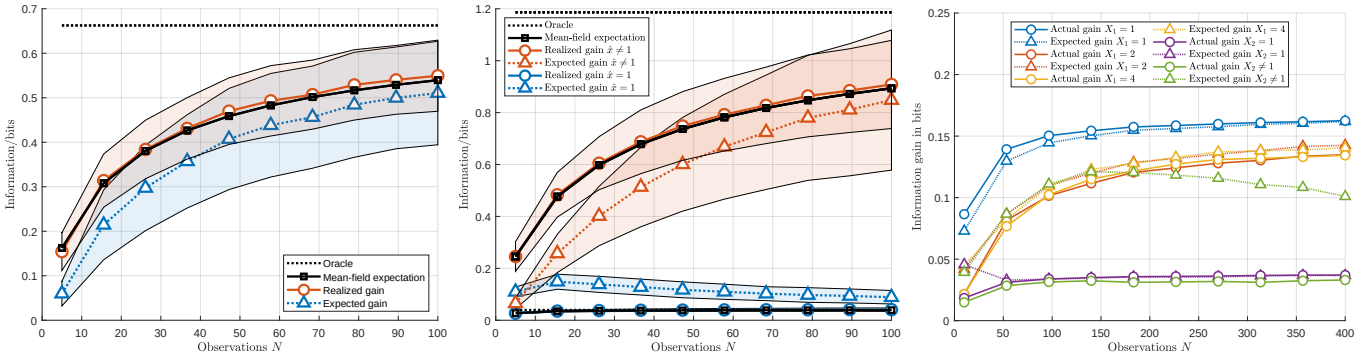

Figure 3: The plots show both the expected information gain (triangles) and the realized gain (circles) for Example 1 (left), Example 2 (middle) and Example 3 (right). The expected gain, eq. (7), signifies the agent's estimate of the information gain from a given intervention $\hat{x}$ or $\hat{y}$, 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

$$\Delta_{\text{do}(X=x)} = \sum_{h_0} P(h^0) \sum_{y|h^0} P(y|h^0, \hat{x})(-1)^{\delta_{h^0, h^0_{X\to Y}}{}^{-1}} I(\hat{x}|y)$$
$$= \frac{1}{2} \sum_{y} \left[P(y|x) - P(y)\right] I(\hat{x}|y). \tag{6}$$

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 $D_J(p,q) = \sum_x (p(x) - q(x)) \log \frac{p(x)}{q(x)}$

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

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 $H^0 = h^0_{X\to Y}$, and $P(x,y) = P_{xy}$, considered as a 2×2 matrix, is $P = \begin{bmatrix} \rho & (1-\rho) \\ (1-\rho) & \rho \end{bmatrix}$ where ρ 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)

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

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 α. 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-

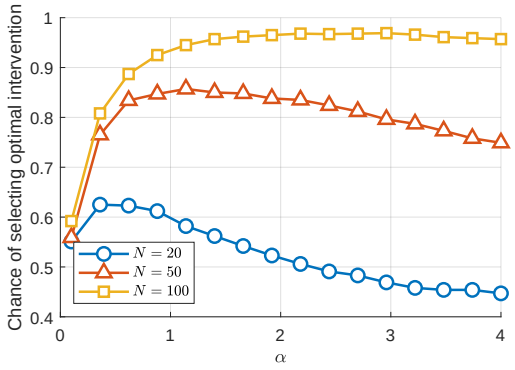

Figure 4: The problem described in Example 3, shown in fig. 3 (right), but where we consider the chance of selecting the single best intervention ($\hat{x} = 1$) as a function of prior α and for different number of observations. The prior is necessary to obtain stable estimates of the expected information 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.

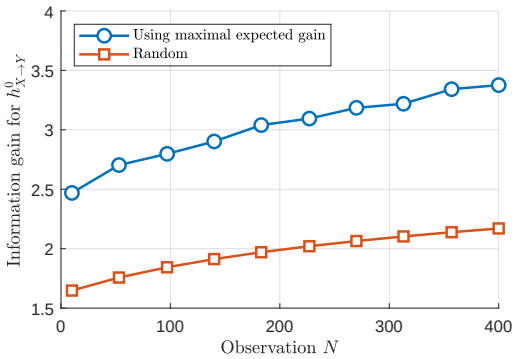

Figure 5: Evaluations of method for an actual intervention-selection problem. The truth is considered fixed as $H^0 = h^0_{X\to Y}$. From this, 10^5 random joint distributions $P(X, Y)$ are generated (see text), and the information gains towards $h^0_{X\to Y}$ are computed when interventions are either 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.

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 over-estimated in the small-sample limit (see e.g. $\hat{y} \neq 1$) where the prior term α 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 α 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 $H_0 = h^0_{X\to Y}$, 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 of $h^0_{X\to Y}$ 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.

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 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 tenetur 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