Causality with Gates

John Winn
Microsoft Research, Cambridge, UK

Abstract

An intervention on a variable removes the influences that usually have a causal effect on that variable. Gates [1] are a general-purpose graphical modelling notation for representing such context-specific independencies in the structure of a graphical model. We extend d-separation to cover gated graphical models and show that it subsumes *do* calculus [2] when gates are used to represent interventions. We also show how standard message passing inference algorithms, such as belief propagation, can be applied to the gated graph. This demonstrates that causal reasoning can be performed by probabilistic inference alone.

1 Introduction

Gates were recently introduced by Minka and Winn [1, 3] as a notation for representing context-specific independence in factor graphs [4] that also allow inference to be performed by message-passing on the gated graph. Gates can be used for such tasks as representing mixtures and Bayesian model comparison, as well as structure learning tasks, such as detecting if a particular edge is present or not. In this paper, we show that gates can also be used to do causal reasoning, when we have a data set in which interventions have been performed on particular variables in the graph.

Causality is an important topic in machine learning, because detecting causal relationships between variables allows the consequences of interventions on those variables to be predicted. This is essential in domains such as healthcare or climate modelling, where the cost of an inappropriate intervention can be very high. Understanding the causal relationships between variables also provides more insight into the mechanisms involved in the system being modelled, which is crucial for scientific applications where the goal is to deduce these mechanisms from observed data.

To model interventions requires the ability to represent context-specific independence: in the context of an intervention on a variable, any influences that normally have a causal effect on that variable are removed. Bayesian networks and factor graphs lack the ability to represent such context-specific independence and so are unable to represent interventions in sufficient detail to reason about conditional independence properties. Pearl's innovative *do* calculus [2, 5] was proposed as an additional mechanism outside of probabilistic inference which allows for reasoning about interventions and hence causality. However, we argue that a modelling notation that allows for context-specific independence renders such additional mechanisms unnecessary. With such a notation, interventions can be represented sufficiently well within the graphical model to reason about causality using only the tools of probabilistic inference.

Dawid proposed the use of decision nodes in a graphical model to represent interventions, forming an influence diagram [6]. However, decision nodes hide the context-specific independence properties so that they are not represented in the graph, but are "introduced as an implicit, externally specified, constraints" (quotation from [6]). Also, because the context-specific independence is not explicit in the graphical data structure, it cannot be exploited by an inference algorithm.

In this work, we show how to use gates to capture the context-specific independence of interventions. We demonstrate that the context-specific independence properties of gated graphs give rise to the same rules as *do* calculus. We hence provide a mechanism for reasoning about interventions and causality using only probabilistic inference in a graphical model of both the underlying data set and the data about interventions that were performed. We also show that inference about interventions can be performed by applying message passing algorithms, such as belief propagation, on the gated graph.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

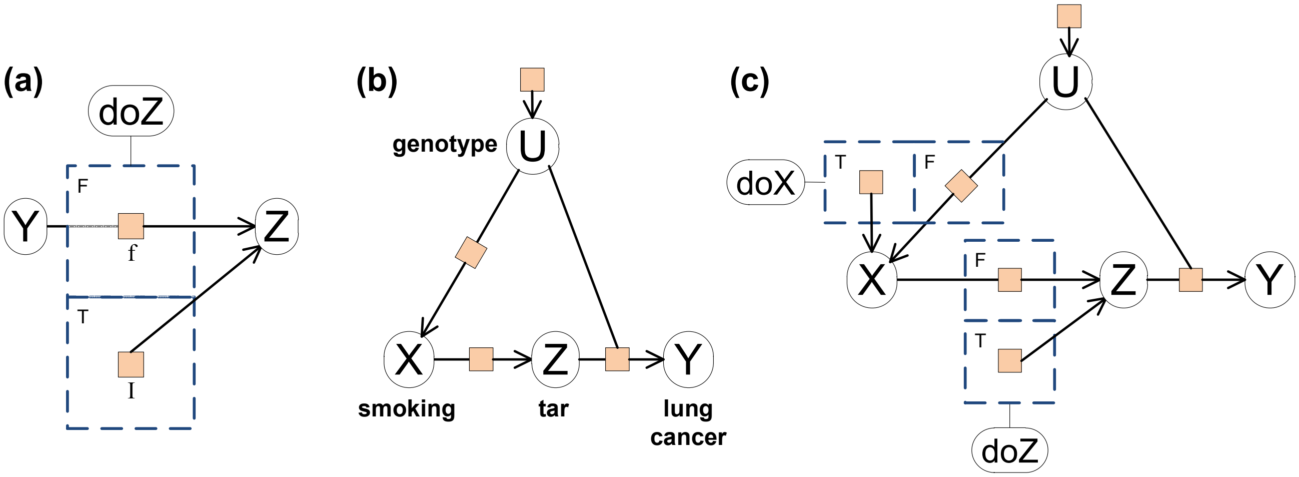

Figure 1: **(a) Using a pair of gates to represent an intervention** The *doZ* variable switches between the natural state, where Z is influenced through f by parent variable Y, and the intervention state, where Z is influenced solely by the intervention factor I. **(b) Directed factor graph for a smoking/lung cancer model** This models the relationship between smoking X and lung cancer Y via the quantity of tar in the lungs Z under the influence of genetic factors U. **(c) Gated graph for the smoking/lung cancer model with interventions** The graph of (b) modified to use the gate structure of (a) to represent interventions on X and Z.

2 Modelling Interventions using Gates

From [1], a gate is a container which encloses part of a factor graph and switches it on or off depending on the state of a selector variable. The gate is on when the selector has a particular value called the key and off for all other values. A gate allows context-specific independencies to be made explicit in the graphical model: the dependencies represented by any factors inside the gate are present only in the context of the selector variable having the key value.

Mathematically, a gate represents raising the contained factors to the power zero if the gate is off, or one if it is on: $\left(\prod_i f_i(x)\right)^{\delta(c=key)}$ where c is the selector variable and f_i are the factors contained within the gate.

Frequently, gates are used in blocks with one gate for each possible value of a selector variable, so that one gate is on and all other gates are off for any value of the selector. For example, figure 1a shows a block of two gates switched by a binary selector variable *doZ*. When *doZ* is false (F), the top gate is on and the bottom gate is off, so that the variable Y is a parent of Z via a factor f. When *doZ* is true (T), the bottom gate is on and the top gate is off, so that Z has no parents and is connected to a prior factor I. In this paper, we mark factors as directed by adding an arrow pointing to the child of the factor – a less cluttered notation than [7] where arrows are also added to the other edges pointing towards the factor. The arrow indicates that the factor sums to 1 across all values of the child variable, for any configuration of the parent variables – in other words it takes the form of a conditional probability distribution.

We can use the gated factor graph of figure 1a as a model of a variable which may or may not be set by an intervention. The variable *doZ* controls whether there is an intervention or not. If *doZ* is true then Z is set according to the nature of the intervention, defined by $I(Z)$. For example, $I(Z)$ may be a 1 at a particular value of Z and zero otherwise, which would represent an intervention that constrains Z to have that value. Alternatively, $I(Z)$ could be a distribution, which would represent a randomized intervention. An example of this would be if Z is binary and $I(Z)$ is as a uniform Bernoulli distribution, which would represent setting Z according to a coin flip. If *doZ* is false, then no intervention occurs and Z is influenced by its parent Y, as normal. It is also possible for I to have other forms (for example, to have parent variables) allowing various forms of imperfect intervention to be represented. This will be discussed in section 6.

It is possible to take a factor graph representing a particular system and modify it by introducing the structure of figure 1a wherever interventions are to be performed in that system. Alternatively, one could consider constructing a graph of both the system and the interventions, without going via the interim model. For example, figure 1b shows the graph of the smoking/lung cancer model from [5] and figure 1c shows a modified version of the graph where gates have been added to allow interventions on smoking X or amount of tar Z. We can consider the latter graph to be modelling the original data $\{X, Y, Z, U\}$ as well as the additional data that intervention actions were taken, recorded as $\{doX, doZ\}$. We would argue that there is nothing to distinguish the modelling of interventions from modelling of any other aspect of the system and no particular reason to consider *doZ* as a variable different from any other. The gate structure of figure 1a just encodes the fact that when we observe that an intervention action was taken, we know a priori that it has a direct and overriding influence on the quantity represented by Z.

The idea of representing an intervention as a variable within the graph was first suggested by Spirtes et al. [8] and further explored by Pearl [5, 9] who represented interventions as variables in an augmented Bayesian network. In this augmented network, a binary intervention variable is added as an additional parent to each variable that can be the target of an intervention, with each conditional probability function being modified appropriately. Similarly, Dawid used decision nodes in influence diagrams [6] to represent variables which have interventions. Because neither Bayesian networks nor influence diagrams show context-specific independencies, these representations hide the independence structure of the conditional probability function associated with the intervention. In contrast, gated factor graphs expose this structure and so allow us to do reasoning about context-specific independence, as shall be shown in section 3. The use of gates to make this conditional independence structure apparent, without recourse to any causality-specific notation, is the main contribution of this paper.

3 Conditional Independence in Gated Graphs

In order to reason about causal relationships between variables, we need to be able to test for conditional independence between variables in our gated graphs. To do this, we will define an extended form of *d-separation* [10] which can be applied to directed factor graphs that include gates. We will first recall the rules for d-separation in directed factor graphs previously defined by Frey in [7]. We will then extend this definition to cover gated graphs. We consider the restricted case where there are only factors inside of a gate, but no variables (whereas [1] does allow variables to be inside gates). This restriction simplifies the following definition and is sufficient for reasoning about interventions (since the gated graph of figure 1a does not have any variables inside a gate).

d-separation in directed factor graphs

Let a *directed factor graph* be a factor graph where all factors are directed, as defined in section 2, so that each factor is connected to a single child variable and zero or more parent variables. We also constrain each variable node to be the child of at most one factor, similar to Bayesian networks.

Given three disjoint variable sets X, Y and Z in a directed factor graph, we can say that X and Y are conditionally independent given Z if the nodes in Z *d-separate* (block) all paths from nodes in X to nodes in Y. A path in a factor graph is any connected sequence of variable-to-factor or factor-to variable edges. Note that it differs slightly from a path in a Bayesian network, in that a path can pass from one parent variable to another through the intervening factor, without passing through the child variable node (such as in figure 1c where a path can pass from U to Z without passing through Y). From [7], a path is d-separated by the nodes in Z if:

1. the path contains a variable node that is in Z *or*
2. the path passes through a directed factor f from one parent variable to another, and neither the child variable of f nor any of its descendants are in Z.

The first of these criteria is exactly the conditional independence condition for undirected factor graphs. The second specifies any additional independencies that can be deduced from knowing that the factors are directed.

d-separation in gated graphs

A *gated* directed factor graph is a directed factor graph that also includes gates. When using gates, we allow a variable to be the child of multiple factors in different gates, provided that only one of these factors is on for any specific configuration of the gate selector variables.

For d-separation in gated graphs, we extend the definition of a path. We now also allow a path to pass from the selector variable of a gate to any factor in that gate (or vice-versa) and consider the selector variable to be a parent of any such factor. This extension arises because we can transform the gated graph into an ungated graph by adding the selector variable as an additional parent to each contained factor, and extend the factor function appropriately. For example, if the original factor function was $f(X \mid Y)$ and the selector variable C, the extended factor function would be:

$$\begin{aligned} f'(X \mid Y, C) &= f(X \mid Y), \text{if } C = \text{key} \\ &= 1, \text{otherwise.} \end{aligned}$$

The advantage of using the gated graph over the ungated one, is that we can detect additional independencies. In a gated graph, a path is d-separated if:

1,2. either of the above two d-separation criteria apply *or*

3. the path passes through two or more disjoint gates (that is, gates with the same selector variable but different key values) whose selector variable is in Z.

The new third criterion gives us the additional independencies. When it applies, the path will pass through at least one gate that is off for any configuration of Z and so renders the path d-separated. For example, if a variable has different parents through two disjoint gates, then the path from one parent through the child variable to the other parent is blocked when the selector is in Z.

In gated graphs, we also need to update slightly the definition of a descendant in light of this third criterion. The usual definition is that W is a descendant node of X if there is a directed path from X to W. For gated graphs, we add the additional restriction that this path must not pass through disjoint gates.

context-specific d-separation in gated graphs

A further advantage of using gated graphs is that we can also detect context-specific independence. In other words, we can detect that X and Y are independent for specific configurations of the observed variables Z, even if they are not independent for all configurations of Z. We can test for context-specific independence by modifying criterion 3 of our definition of d-separation. Suppose the variables Z take on a configuration z, we say that a path is d-separated in the context $Z{=}z$ if:

1,2. either of the first two above d-separation criteria apply *or*

3. the path passes through a gate which is turned off. A gate is turned off if its selector variable is in Z and its key is not equal to the value of the selector variable in z.

In other words, we first remove from the graph any factors contained in gates that are turned off and then apply the usual d-separation criteria for ungated graphs. Again, we need to update our definition of a descendant node in light of this new criterion. We add the restriction that the path from a node to any descendant must not pass through any gates which are off. So, for example, observed nodes which lie on a path through an off gate do not count as descendants when assessing criterion 2.

4 Equivalence to *do* calculus

We will now show that context-specific d-separation in gated graphs of perfect interventions gives rise to the same rules as *do* calculus. We consider each rule in turn by (i) redefining the equality statement by rewriting Pearl's notation (where a variable with an intervention is marked with a hat, such as $\hat{x}$) in terms of separate intervention variables (such as doX) (ii) deriving the condition for equality in terms of d-separation in the gated graph (iii) showing the equivalence of this condition with Pearl's original condition in the ungated graph. Throughout, we follow Pearl and assume that the intervention factor function (I from section 2) is a point mass at a particular value.

Rule 1: Insertion/deletion of observations

> $P(y\,|\,\hat{x}, z, w) \;=\; P(y\,|\,\hat{x}, w)$ if $(Y \perp\!\!\!\perp Z)\,|\,X, W$ in a graph where the parent edges of X have been removed.

In a gated graph model, we want to determine when $P(y\,|\,doX{=}T, Z{=}z, W{=}w)$ is equal to $P(y\,|\,doX{=}T, W{=}w)$ for any z. This will be true if Y is conditionally independent of Z in the context where $W{=}w$ and $doX{=}T$. Because doX is observed to be true, all factors connecting X to any parent of X will be inside a gate which is turned off. This situation is illustrated in figure 2a where the top gate is turned off (shown in gray), so paths from X to its parents are blocked according to criterion 3 of context-specific d-separation. Hence, the condition is equivalent to evaluating the standard d-separation criteria in an ungated graph with parent edges of X removed. So Rule 1 arises from d-separation in the gated graph and can be rewritten as follows:

> $P(y\,|\,doX = T, z, w) \;=\; P(y\,|\,doX = T, w)$ if $(Y \perp\!\!\!\perp Z)\,|\,doX{=}T, W$ in the gated graph.

Rule 2: Action/observation interchange

> $P(y\,|\,\hat{x}, \hat{z}, w) = P(y\,|\,\hat{x}, z, w)$ if $(Y \perp\!\!\!\perp Z)\,|\,X, W$ in a graph where the parent edges of X have been removed and the child edges of Z have been removed.

In a gated graph model, we want to determine when $P(y\,|\,doX = T, Z = z, doZ = T, W = w)$ is equal to $P(y\,|\,doX{=}T, Z{=}z, doZ{=}F, W{=}w)$ where we have included the redundant $Z{=}z$ in the first expression to aid the proof. This equality will hold if Y is conditionally independent of doZ in the context where $Z{=}z, W{=}w$ and $doX{=}T$. Because doX is observed to be true, the above argument for removing parent edges of X applies

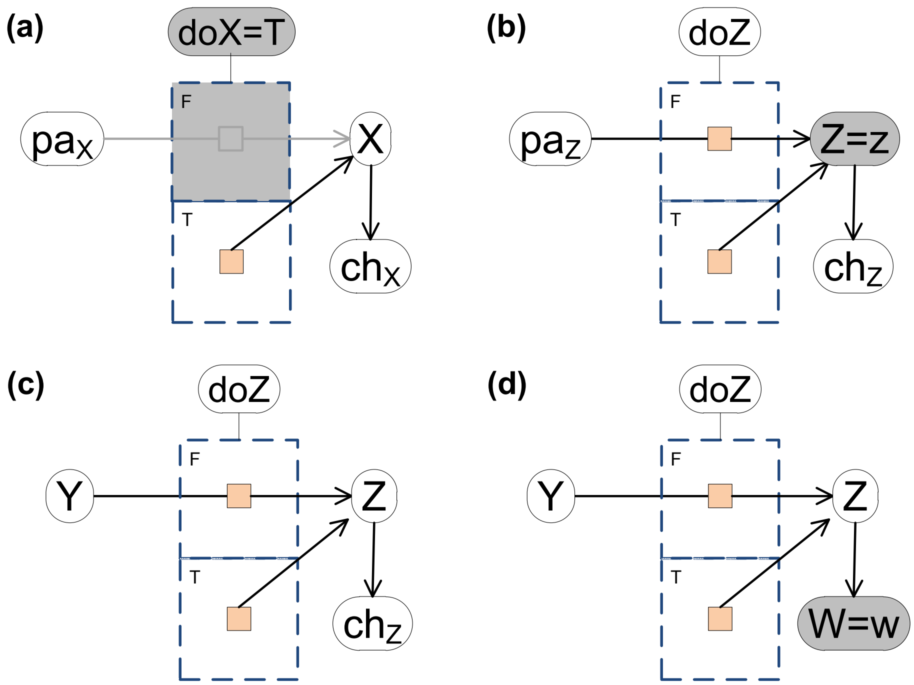

Figure 2: **Illustration of the rules of *do* calculus using gated graphs (a) Rule 1**: the edges from X to its parents are turned off when $doX{=}T$, due to being in a switched off gate (shown shaded gray). **(b) Rule 2**: paths from doZ to the children of Z are blocked by the observation at Z **(c,d) Rule 3**: Y is independent of doZ but not of Z in (c) whereas in (d) Y is not independent of either Z or doZ. See text for a complete explanation.

again here. Because Z is observed, it blocks any paths from doZ to the children of Z (figure 2b), according to criterion 1 of d-separation. However, also because Z is observed, it unblocks the paths from doZ to the parents of Z according to criterion 2 of d-separation. Hence, this is equivalent to standard d-separation between Y and Z in the ungated graph with the child edges from Z removed, but the parent edges of Z left in place. Thus Rule 2 also arises from d-separation in the gated graph and can be rewritten as:

> $P(y \,|\, doX = T, z, doZ, w) = P(y \,|\, doX = T, z, w)$ if $(Y \perp\!\!\!\perp doZ) \,|\, doX = T, Z, W$ in the gated graph.

Rule 3: Insertion/deletion of actions

> $P(y \,|\, \hat{x}, \hat{z}, w) = P(y \,|\, \hat{x}, w)$ if $(Y \perp\!\!\!\perp Z) \,|\, X, W$ in a graph where the parent edges of X have been removed and the parent edges have been removed for those nodes in Z which are not ancestors of any node in W, in the graph with the parent edges of X removed.

In a gated graph model, we want to determine when $P(y \,|\, doX{=}T, doZ{=}T, W{=}w)$ is equal to $P(y \,|\, doX{=}T, doZ = F, W = w)$. This equality will hold if Y is conditionally independent of doZ in the context where $W{=}w$ and $doX{=}T$. This differs from Rule 2 because now Z is unobserved.

To map this condition back into the ungated graph, we must now consider in what circumstances Y might be conditionally independent of doZ but not of Z. Consider figure 2c, which shows a graph Y where is not independent of Z but is independent of doZ, because of criterion 2 of d-separation. Conversely, in the graph of figure 2d, Y is not independent of either Z or doZ. In general, Y will be independent of doZ but not of Z where there is a path from Y to Z that passes through the gate controlled by doZ and no descendant of Z is observed. Hence, if we remove the parent edges of Z only where no descendant of Z is observed, then standard d-separation between Y and Z will be equivalent to d-separation between Y and doZ in the gated graph. So Rule 3 also arises from d-separation in the gated graph and can be written as:

> $P(y \,|\, doX{=}T, doZ, w) = P(y \,|\, doX{=}T, w)$ if $(Y \perp\!\!\!\perp doZ) \,|\, doX{=}T, W$ in the gated graph.

As shown, the three rules of *do* calculus can be rewritten as tests of context-specific independence in the particular gated graph where the gate structure of figure 1a is used to represent an intervention.

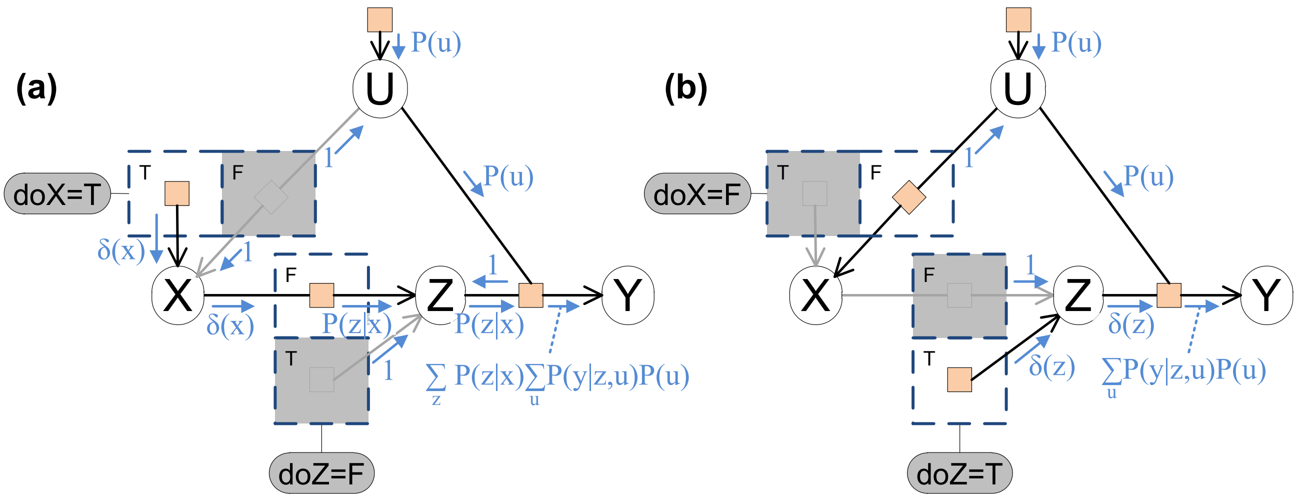

Figure 3: **Belief propagation for causal inference (a) Intervention on X** – when there is an intervention on X but not on Z, *doX* is true and *doZ* is false. Hence, gates shown shaded gray are switched off and the belief propagation messages are as shown in blue. **(b) Intervention on Z** similar diagram as (a) for when the intervention is on Z and not X.

5 Inference about Interventions

The primary purpose of *do* calculus has been to allow inference calculations that involve interventions (like $\hat{x}$ or $\hat{z}$) to be transformed into calculations involving ordinary observations (like x or z). To show that gated graphs can be used in place of *do* calculus, we must show that we can perform inference in these graphs to compute marginal probabilities of interest. Happily, Minka and Winn [1] showed that several common inference algorithms can be straightforwardly extended to apply to gated graphs, including expectation propagation [11], variational message passing [12] and Gibbs sampling. Furthermore, these algorithms do not require the two-stage approach of *do* calculus, since there is no need to first transform the problem to remove interventions.

Example 1: Inference in the Smoking and Lung Cancer Model

In [2], Pearl demonstrated how to apply *do* calculus to the model of figure 1b to infer various marginal probabilities, conditioned on interventions. Pearl considered three tasks for this model: (1) compute $P(z\,|\,\hat{x})$, (2) compute $P(y\,|\,\hat{z})$ and (3) compute $P(y\,|\,\hat{x})$.

We will now show how to compute these marginals directly by applying belief propagation to the gated graph. We use the expectation propagation algorithm defined in [1] simplified to the case where all variables are discrete (so expectation propagation becomes belief propagation), there are no nested gates, and all gate selector variables are observed. In this case all messages are exactly as for belief propagation, except that we set to uniform all messages on edges passing out of an off gate (shaded grey in figure 3).

So, the message from a factor f to a variable x_i is:

$$\begin{aligned} m_{f\to x_i} &= 1 \quad \text{if passing out of an off gate} \\ &= \prod_{j\neq i} m_{x_j\to f}\, f(x) \quad \text{otherwise.} \end{aligned}$$

The message from a variable to a factor never passes out of a gate (since variables do not lie inside gates) and so takes the standard form $m_{x_i\to f} = \prod_{g\neq f} m_{g\to x_i}$.

Figure 3a and b show the relevant belief propagation messages passed when conditioning on $\{doX{=}T, doZ{=}F\}$ and $\{doZ{=}T, doX{=}F\}$ respectively. In each case the graph has no loops, due to one gate in the main loop being switched off, and so belief propagation is exact. The marginal for a variable can then be computed as the product of all messages incoming to the variable's node:

1. $P(z\,|\,\hat{x})$ from figure 3a: $P(z\,|\,doX{=}T, doZ{=}F) = 1\times 1\times P(z\,|\,x) = P(z\,|\,x)$
2. $P(y\,|\,\hat{z})$ from figure 3b: $P(y\,|\,doZ{=}T, doX{=}F) = \sum_u P(y\,|\,u,z)P(u)$
3. $P(y\,|\,\hat{x})$ from figure 3a: $P(y\,|\,doX{=}T, doZ{=}F) = \sum_z P(z\,|\,x)\sum_u P(y\,|\,u,z)P(u)$

These may be seen to be identical to the results from [2], noting that $\sum_u P(y\,|\,u,z)P(u) = \sum_x P(y\,|\,x,z)P(x)$.

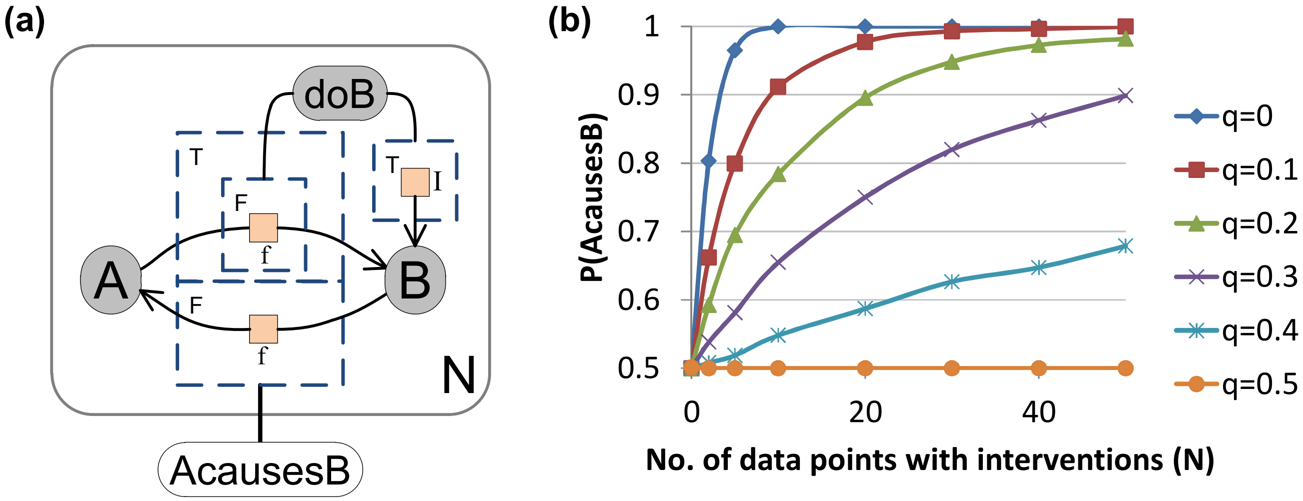

Figure 4: **(a) A model for inferring whether A causes B or vice-versa** Pairs of gates are used both to model an intervention and to determine the direction of the edge. **(b) Posterior probability that A causes B** Plot of the posterior probability that A causes B in the model of (a) given data with interventions sampled from a true model where A does cause B. The probability of the true model increases with the size of the data set and decreases as the noise increases.

Example 2: Inferring the Causal Direction of an Edge

A common application of causal reasoning is to infer the direction of an edge in a graph, in other words, to infer whether A causes B or B causes A. Figure 4a shows a gated graph which allows the direction of the edge between A and B to be inferred, through interventions on B. This example also demonstrates that gates can be used for other purposes in addition to modelling interventions – here we use one pair of gates to model the intervention on B and a second pair of gates to determine the direction of the arrow between A and B. When we have an intervention on B, the factor between A and B is only switched off in the case where A causes B. Hence, the factor for the case where B causes A is not switched by doB. In both cases, the priors on the parent variable are assumed to be uniform and so the factors for these priors have been omitted from graph. If they were non-uniform they would have to be added to the model and included inside the appropriate gates.

The plate of size N means that model represents N measurements of A and B where the intervention on B may or may not have occurred in each case, as recorded by doB. Because the variable $AcausesB$ is outside of the plate, it encodes the assumption that A causes B (or vice-versa) for all measurements. Note that this model considers only whether A causes B or that B causes A – it does not consider, for example, a third variable C causing both A and B.

To make this example concrete, we will assume that all variables are binary and that the factor functions are $I(X) = 0.5$ and $f(X \,|\, Y) = 1 - q$ if $X{=}Y$ or q otherwise. This definition of f means that the child variable X takes the value of the parent variable Y and flips it with probability q. So setting $q{=}0$ would make the variables equal whereas $q{=}0.5$ would make them independent. This definition of I means that, when intervening, we set the value of B according to a coin flip.

Suppose the true model is that A causes B through factor f with noise q. We generated data sets of $\{A_n, B_n, doB_n\}_{n=1}^{N}$ from this true model both with and without interventions – that is, with doB_n either all true or all false. We also varied the data set size N and the amount of noise q. Given these data sets, we applied expectation propagation to infer the posterior distribution over $AcausesB$. In each case, we assumed that the true value of q was known. Where we applied the model to data without interventions, $P(AcausesB)$ was always exactly 0.5, indicating that we cannot infer the causal direction of the edge without performing an intervention. Figure 4b shows the results when we applied the model to data with interventions, for varying values of N and q[1]. To account for variability in the sampled data sets, the posterior probability of $AcausesB$ was averaged over 1,000 runs. The results show that gated graphs have allowed us to infer a precise probability of the true model (A causes B) and determine how it increases as the number of data points increases or as the noise decreases.

[1] The source code for this experiment is available for download from `http://bit.ly/causality_with_gates`.

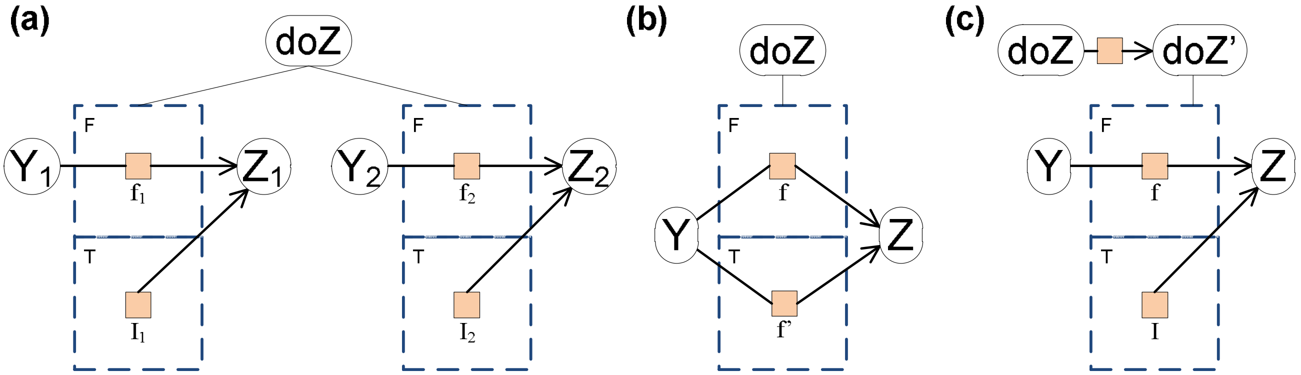

Figure 5: **Imperfect interventions (a) 'Fat hand' intervention** where an intervention on a variable Z_1 has the unintended effect of intervening on Z_2. **(b) Mechanism change** where an intervention softens or changes how a variable is affected by its parents, rather than removing the effect entirely. **(c) Partial compliance** where an intervention is not always successful (the latent variable doZ' represents success).

6 Discussion

Gated graphs are advantageous because of their:

Generality – the criteria for d-separation in gated graphs and the extended inference algorithms apply to all gated graphs, not just those intended for reasoning about interventions and causality. For example, we can use the same d-separation criteria to assess context-specific independence in a mixture model. They would show, for example, that the parameters of one mixture component are conditionally independent of the parameters of another component given the data and the indicator variables for all data points, but not given the data alone. In short, there is no need to learn a new notation or use a separate algorithm for causal reasoning problems – probabilistic inference in gated graphs is sufficient.

Flexibility – the gated structure of figure 1 models a perfect intervention which affects only the variable Z, completely overrides any influence of the normal influences Y and where the intervention function I and the choice of intervention doZ do not depend on any other variables. For real interventions, one or more of these may not hold, as discussed in [13]. When using gates, such imperfect interventions can be modeled using the appropriately modified gate structure (figure 5).

For example, introducing a receptor blocker in a cell may directly affect other receptors than the intended target. This 'fat hand' intervention can be modelled by having the intervention be a selector variable for multiple gates, one for each affected receptor (figure 5a). One can also have the intervention function change rather than eliminate the dependency on the normal parent variables (figure 5b), which has been called "mechanism change" [14]. Another type of imperfect intervention, discussed in [15], is that of imperfect compliance. For example, in a medical trial, participants instructed not to smoke on a particular day may do so anyway due to factors such as stress. We can model this by having a latent selector variable representing whether the person actually followed instructions with the intervention as a parent of this variable (figure 5c).

Imperfect interventions can also be more complex than any of these examples. When analyzing large scale 'natural experiments', such as all health records for patients in a particular region, interventions will have been performed for particular reasons decided by the doctors involved based on the information available to them at the time. Such a decision process could be extremely complex. We could potentially model this decision process as a separate sub-model connected to the selector variable for the gate, where the sub-model has complex dependencies on other variables which may have influenced the decision. As in figure 5c, we can also reason about *whether* an intervention took place, for example, if data about treatments is missing from the health records.

These two reasons suggest that gated graphs may be broadly applicable to inferring causal structure in many practical applications. Assessing which gated structures are most appropriate for representing the various forms of imperfect intervention is a promising direction for future research.

Acknowledgements

The author would like to thank Tom Minka for many insightful suggestions and also John Guiver, Kevin Murphy and Nevena Lazic for valuable feedback on the paper.

References

[1] T. Minka and J. Winn. Gates. In *Advances in Neural Information Processing Systems 21*, 2009.

[2] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–688, 1995.

[3] T. Minka and J. Winn. Gates: A graphical notation for mixture models. Technical Report MSR-TR-2008-185, Microsoft Research Ltd, 2008.

[4] B. Frey, F. Kschischang, H. Loeliger, and N. Wiberg. Factor graphs and algorithms. In *Proc. of the 35th Allerton Conference on Communication, Control and Computing*, 1998.

[5] J. Pearl. *Causality – models, reasoning, and inference*. Cambridge University Press, second edition, 2000, 2009.

[6] A. P. Dawid. Influence diagrams for causal modelling and inference. *Intl. Stat. Review*, 70:161–189, 2002. Corrections p437.

[7] B. Frey. Extending factor graphs so as to unify directed and undirected graphical models. In *Proc. of the 19th conference on Uncertainty in Artificial Intelligence*, pages 257–264, 2003.

[8] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction and Search*. Springer-Verlag, 1993.

[9] J. Pearl. Comment: Graphical models, causality, and intervention. *Statist. Sci.*, 8:266–9, 1993.

[10] D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.

[11] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

[12] J. Winn and C. M. Bishop. Variational Message Passing. *JMLR*, 6:661–694, 2005.

[13] D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proc. of the 8th workshop on Artificial Intelligence and Statistics*, 2007.

[14] J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17th conference on Uncertainty in Artificial Intelligence*, 2001.

[15] A. P. Dawid. Causal inference using influence diagrams: the problem of partial compliance. In *Highly structured stochastic systems*, number 27 in Oxford Statistical Science Series. Oxford University Press, 2003.