Knowledge Intensive Learning: Combining Qualitative Constraints with Causal Independence for Parameter Learning in Probabilistic Models

Shuo Yang, Sriraam Natarajan

School of Informatics and Computing, Indiana University, USA

Abstract. In Bayesian networks, prior knowledge has been used in the form of causal independencies between random variables or as qualitative constraints such as monotonicities. In this work, we extend and combine the two different ways of providing domain knowledge. We derive an algorithm based on gradient descent for estimating the parameters of a Bayesian network in the presence of causal independencies in the form of Noisy-Or and qualitative constraints such as monotonicities and synergies. Noisy-Or structure can decrease the data requirements by separating the influence of each parent thereby reducing greatly the number of parameters. Qualitative constraints on the other hand, allow for imposing constraints on the parameter space making it possible to learn more accurate parameters from a very small number of data points. Our exhaustive empirical validation conclusively proves that the synergy constrained Noisy-OR leads to more accurate models in the presence of smaller amount of data.

1 Introduction

Human advice or input is generally provided in learning Bayesian networks using the structure of the Bayesian network [1]. Given this network structure, most methods use some form of optimization to learn the parameters of the models. Initially, advice giving methods simply served to constrain the structure of the network. While the use of prior structure does reduce the number of examples required to learn a reasonable network, learning parameters can still require certain amount of examples to converge to a reasonable estimate. However many domains, such as medicine, can be data poor (for example, number of positive examples of a disease can possibly be quite low) but knowledge rich due to several decades of research. This domain knowledge is mostly about the influential relationships between the random variables of interest in the domain.

One of the most prominent methods of providing domain knowledge to a probabilistic learner is to provide the set of causal independencies that exist in the domain [2]. Also called as Independence of Causal Influences (ICI) [3–6], this form of knowledge identifies *sets of parents* that are independent of each other when affecting the target random variable. The effects of these sets of random

variables can typically then be combined using a function such as Noisy-Or. The key advantage of such knowledge is that these lead to a drastic reduction in the number of parameters associated with the conditional distributions (from exponential in the total number of parents to exponential in the size of these sets). This reduction can greatly affect the number of examples required for training an accurate model. While this is very attractive, ICI can be very restrictive and easily violated in many domains.

An equally alternative and more recent method of providing advice to learners is based on qualitative influences [7–11]. Qualitative influence (QI) statements essentially outline how the change of one variable affects the change of another variable. A classical example of such QI statements is monotonicity [7, 8, 12] where an increase in value of one random variable (say cholesterol) increases the probability that another variable (say risk of heart attack) takes a higher value. Another direction has been in combining context-specific independencies [13] with QI statements [9] and showing that learning with such constraints is a special case of isotonic regression [14].

In this work, we extend and combine these different methods of specifying domain knowledge. More precisely, we extend the research in two directions – First, current methods for QI can handle monotonicity statements while we extend these directions by allowing for synergistic interactions [7] between random variables. While monotonicities model the qualitative dependency between two random variables, synergistic interactions allow for richer influence relationships. For instance, with synergies, it is possible to specify statements such as "Increase in blood sugar level increases the risk of heart attack in high cholesterol level patients more than it does in low cholesterol level patients". This statement explains how sugar level and cholesterol level interact when influencing heart attack. Second, we use such synergistic and monotonicity statements and combine them with the concept of ICI i.e., we treat each "set" of monotonicity and synergistic interaction as independent of each other and their influences are combined with a combining rule. In this work we employ Noisy-Or [5] for combining the independent influences. While previous work has used context-specific independences, we generalize them to using ICI.

Following prior work [8], we convert the monotonicity and synergy statements to constraints on the parameter space of the conditional distributions. We then combine the different conditional distributions using Noisy-Or and derive the overall objective function. We adopt a gradient descent algorithm with exterior penalty method to optimize the objective function and outline the algorithm for learning in the presence of qualitative and conditional influences.

To summarize, we make the following contributions: first, we extend the qualitative influences to include synergies. Second, we combine these qualitative influences with the independence of causal influence (ICI) such as Noisy-Or and derive a new objective function that includes these influences as constraints on the parameter space. Third, we derive an algorithm for parameter learning in the presence of sparse data by exploiting these influences. Finally, we perform an exhaustive evaluation in 11 different standard domains to understand the

impacts and influences of the different types of influence statements. Our results show clearly that the use of such influences helps the learning algorithm improve its performance in the presence of sparse data.

2 Background

We provide a brief background on qualitative and conditional influences. First, we introduce some basic notations used throughout the paper. In a Bayesian network with n discrete valued nodes, we denote the parameters by θ_{ijk} ($i \in \{1, 2, ..., n\}, j \in \{1, 2, ..., v_i\}, k \in \{1, 2, ..., r_i\}$) which means the conditional probability of X_i to be its k-th value given the j-th configuration of its parents (i.e. $P(X_i^k | pa_i^j)$). r_i denotes the number of states of the discrete variable X_i; Pa_i represents the parent set of node X_i; the number of configurations of Pa_i is $v_i = \prod_{X_t \in Pa_i} r_t$; j is the index of a particular configuration of node X_i's parents pa_i^j.

2.1 Qualitative Influences - Monotonic Constraints

Qualitative influence, specifically monotonicity has been explored extensively in previous work [7–9]. Specifically, Altendorf et al. [8] used monotonicities in the context of learning Bayesian networks. Monotonic influence means that stochastically, higher values of a random variable, say X result in higher (or lower) values of another variable Y, and is denoted as $X \underset{\succ}{}^{M+} Y$ (or $X \underset{\succ}{}^{M-} Y$). The interpretation is that increasing values of X shifts the cumulative distribution function of Y to the right (i.e., higher values of Y are more likely). This means that $P(Y \leq y | X = x_1) \leq P(Y \leq y | X = x_2)$ (where $x_1 \geq x_2$). Note that the same definition can be extended in the presence of multiple parents by fixing the values of the other parents. If one of X_i's parents (denoted by X_c) has monotonic constraint on X_i, this relationship still stands given the same configuration of other parents, the general form of monotonic constraints of X_c on X_i is

$$P(X_i \leqslant k_c | X_c^m, C_i^n) \geqslant P(X_i \leqslant k_c | X_c^{m+1}, C_i^n) \tag{1}$$

where $k_c \in \{1, 2, ..., r_i - 1\}, m \in \{1, 2, ..., r_c - 1\}, X_c^m \leqslant X_c^{m+1}$, C_i^n represents all possible configurations of X_i's parents other than X_c, n is the index.

Altendorf et al. [8] used these qualitative constraints to learn the parameters of a Bayes net by introducing a penalty to the objective function when the constraints are violated. Assume there is a monotonic constraint: $P(X_i \leq k_c | pa_i^{j_2}) \leq P(X_i \leq k_c | pa_i^{j_1})$. The constraint function δ with margin ϵ is defined as:

$$\delta = P(X_i \leq k_c | pa_i^{j_2}) - P(X_i \leq k_c | pa_i^{j_1}) + \epsilon \tag{2}$$

The corresponding penalty function is $P_{j_1, j_2}^{i, k_c} = I_{(\delta > 0)} \delta^2$ (where I=1 when $\delta > 0$ and I=0 when $\delta \leqslant 0$). In order to rule out the need for the simplex constraints ($\sum_{k=1}^{r_i} \theta_{ijk} = 1$), Altendorf et al. defined μ_{ijk} such that

$$\theta_{ijk} = \frac{exp(\mu_{ijk})}{\sum_{k'=1}^{r_i} exp(\mu_{ijk'})} \tag{3}$$

They then derived the gradient of the objective function wrt μ and used exterior penalty method to learn from data. We refer to their work for more details. Tong et al. [11] and de Campos et al. [10], considered the problem of facial recognition from images and applied qualitative constraints to learning for recognizing these faces. They took an EM approach for learning the parameters of these qualitative constraints (that possibly could include synergy). We on the other hand, take a gradient descent approach that allows for including conditional influences such as Noisy-Or.

2.2 Noisy-Or

The term *independence of causal influence* was first used by Heckerman and Breese [3] to model the situation where there are several variables that influence a random variable independently but their collective influence can be computed using some deterministic or stochastic combination function. Typical examples of ICI include Noisy-Or, Noisy-And, Noisy-Min, Noisy-Max, Noisy-Add etc. Representing and learning with such ICI relationships have long been explored in the context of Bayes nets [3–6]. In this work, we consider a particular type of ICI, the most popular one – Noisy-Or. Simply put, if there are n independent causes $\{X_1, ..., X_n\}$ for a random variable Y and assuming for simplicity that Y is binary, then the target distribution $P(Y = 1|X_1 = x_1, ..., X_n = x_n)$ is given by

$$P(Y = 1|X_1 = x_1, ..., X_n = x_n) = 1 - \prod_i P(Y = 0|X_i = x_i) \tag{4}$$

The interpretation is that if any parent, say X_i takes value x_i, Y will take the value 1, unless there is an effect of inhibition. This inhibition has a probability of $P(Y = 0|X_i = x_i)$ [6] and these inhibitory effects are assumed to be independent ($1 - q_i$ for $i^{th} parent$).

3 Qualitative Constraints - Synergies

In this section, we extend the previous work on montonicities [8] by allowing for synergistic interactions. After presenting the definition of synergies, we derive the gradient for learning with such knowledge from data.

In the presence of a small amount of training data, the parameters in conditional probability tables (CPT) estimated only based on the data are most likely inaccurate, and in some cases can result in even uniform distributions due to the lack of data about certain configurations of the parents. Fortunately, in many of the real world problems, domain experts can provide sufficient prior knowledge about the influences that exist in the domain. We consider a particular type of the domain knowledge namely, qualitative influence statements that allow us to apply some constraints on the CPTs. These constraints aid in obtaining more accurate estimates of the parameters of the CPTs. More specifically, we propose to exploit the monotonicity and synergy constraints and combine them with a rule such as Noisy-Or when learning the parameters.

When multiple parents influence the resultant independently, we can simply employ monotonicities as presented in the previous section. Synergies on the other hand, allow for richer interactions between the parents where the set of parents can influence the resultant variable dependently. We use Wellman's definition on qualitative synergy [7]. Assume that two variables X_1 and X_2 affect a third variable Y synergistically (where each of them has the $X_1 \overset{M+}{\succ} Y$ and $X_2 \overset{M+}{\succ} Y$ relationship with the target). This is denoted as $X_1, X_2 \overset{S+}{\succ} Y$ (sub-synergy is denoted as S-)[1]. In simple terms, this means that increasing X_1 has a greater (lesser for sub-synergy) effect on Y for high values of X_2 than low values of X_2; likewise for increasing X_2 with fixed X_1. Note that two causes having the same monotonic influence is the premise of their synergistic interaction, which means by our definition, there can not be a synergy or sub-synergy relation between X_1 and X_2 if $X_1 \overset{M+}{\succ} Y$ while $X_2 \overset{M-}{\succ} Y$ i.e., the parents in the synergy relationship cannot have different types of monotonic influences to the target.

Consider for example a medical diagnosis problem and assume that the target of interest is heart attack. An example of a synergistic statement in the domain is, *cholesterol and blood pressure interact synergistically when influencing heart attack*. In simpler terms, the above statement simply means that hypertension increases the risk of heart attack in high cholesterol level patients more than it does in low cholesterol level patients. This defines how the two risk factors (cholesterol and blood pressure) interact with heart attack. Note that each of the cholesterol level and blood pressure has a monotonic relationship with heart attack when considered individually (i.e. $Chl \overset{M+}{\succ} HA$ and $BP \overset{M+}{\succ} HA$). A classic example of a sub-synergy in medical research is that coronary heart disease (CHD) is markedly more common in men than in women; CHD risk increases with age in both sexes, but the increase is sharper in women [15]. Hence gender and age interact sub-synergistically when influencing CHD.

Formally, based on the definition above, assume $x_i^j \leqslant x_i^{j+1}$ where x_i^j is the j^{th} value of variable X_i. Since $P(Y \leq k_c | X_1^i, X_2^j) \geqslant P(Y \leq k_c | X_1^{i+1}, X_2^j)$ (implied by $X_1 \overset{M+}{\succ} Y$), X_1's effect on Y at low level of X_2 is

$$P(Y \leq k_c | X_1^i, X_2^j) - P(Y \leq k_c | X_1^{i+1}, X_2^j)$$

and similarly, at high level of X_2 is

$$P(Y \leq k_c | X_1^i, X_2^{j+1}) - P(Y \leq k_c | X_1^{i+1}, X_2^{j+1})$$

The synergistic constraint on conditional probability distribution can be mathematically represented as:

$$P(Y \leq k_c | X_1^i, X_2^j) - P(Y \leq k_c | X_1^{i+1}, X_2^j) \leq$$
$$P(Y \leq k_c | X_1^i, X_2^{j+1}) - P(Y \leq k_c | X_1^{i+1}, X_2^{j+1})$$

[1] Note that what we use the terminology of sub-synergy due to Wellman. This same concept is also called as anti-synergy in the literature

where $i \in \{1, 2, ..., r_1 - 1\}$ and $j \in \{1, 2, ..., r_2 - 1\}$. Note the above inequation is essentially X_1's effect on Y with fixed X_2. Similarly the synergistic constraint of X_2 on Y with fixed $X_1$is

$$P(Y \leq k_c|X_1^i, X_2^j) - P(Y \leq k_c|X_1^i, X_2^{j+1}) \leq$$
$$P(Y \leq k_c|X_1^{i+1}, X_2^j) - P(Y \leq k_c|X_1^{i+1}, X_2^{j+1})$$

Note that by definition, both of the above inequations need to be satisfied to make X_1 and X_2 a synergistic pair. We can generalize the above two inequalities into one inequality constraint by simply moving the subtractors to the other side of the inequality, which is the general form of synergy that we consider.

$$P(Y \leq k_c|X_1^i, X_2^j) + P(Y \leq k_c|X_1^{i+1}, X_2^{j+1}) \leq$$
$$P(Y \leq k_c|X_1^{i+1}, X_2^j) + P(Y \leq k_c|X_1^i, X_2^{j+1}) \tag{5}$$

Assume Y is binary, X_1 and X_2 are both ternary. Now, the synergy constraints between X_1 and X_2 that affect Y are as presented in Table 1. The key

Synergy Constraints of X_1 and X_2 on CPT of Y
$P(Y=0\|x_1^1, x_2^1) + P(Y=0\|x_1^2, x_2^2) \leqslant P(Y=0\|x_1^1, x_2^2) + P(Y=0\|x_1^2, x_2^1)$
$P(Y=0\|x_1^1, x_2^2) + P(Y=0\|x_1^2, x_2^3) \leqslant P(Y=0\|x_1^1, x_2^3) + P(Y=0\|x_1^2, x_2^2)$
$P(Y=0\|x_1^2, x_2^1) + P(Y=0\|x_1^3, x_2^2) \leqslant P(Y=0\|x_1^2, x_2^2) + P(Y=0\|x_1^3, x_2^1)$
$P(Y=0\|x_1^2, x_2^2) + P(Y=0\|x_1^3, x_2^3) \leqslant P(Y=0\|x_1^2, x_2^3) + P(Y=0\|x_1^3, x_2^2)$

Table 1: Synergy Constraints.

difference to monotonicity is that the constraints are on a set of parents (two in our example) rather than a single parent.

3.1 Derivation of the gradient for the synergy qualitative influence

We now derive the gradients by extending the prior work [8]. Let us redefine the parameters of the conditional distributions as shown in Equation 3. This allows us to define a constraint function δ for the synergistic constraints:

$$P(X_i \leq k_c|pa_i^{j_1}) + P(X_i \leq k_c|pa_i^{j_4}) \leq P(X_i \leq k_c|pa_i^{j_2}) + P(X_i \leq k_c|pa_i^{j_3})$$

The constraint function for the above definition is:

$$\delta = P(X_i \leq k_c|pa_i^{j_1}) + P(X_i \leq k_c|pa_i^{j_4}) - P(X_i \leq k_c|pa_i^{j_2}) - P(X_i \leq k_c|pa_i^{j_3}) + \epsilon \tag{6}$$

The above definition is similar to the monotonicity case. Then the gradient of the penalty function can be computed as:

$$\frac{\partial}{\partial \mu_{ijk}} P^{i,k_c}_{j_1,j_2,j_3,j_4} = \frac{\partial}{\partial \mu_{ijk}} I_{(\delta \geq 0)} \delta^2$$

$$= 2I_{(\delta \geq 0)}\delta(\frac{\partial}{\partial \mu_{ijk}}\frac{Z^i_{j_1k_c}}{Z^i_{j_1}} + \frac{\partial}{\partial \mu_{ijk}}\frac{Z^i_{j_4k_c}}{Z^i_{j_4}} - \frac{\partial}{\partial \mu_{ijk}}\frac{Z^i_{j_2k_c}}{Z^i_{j_2}} - \frac{\partial}{\partial \mu_{ijk}}\frac{Z^i_{j_3k_c}}{Z^i_{j_3}} + \frac{\partial}{\partial \mu_{ijk}}\epsilon)$$

$$\begin{aligned}
&= 2I_{(\delta \geq 0)}\delta[\frac{Z^i_{j_1}I_{(j=j_1 \wedge k \leq k_c)}exp(\mu_{ijk}) - Z^i_{j_1k_c}I_{(j=j_1)}exp(\mu_{ijk})}{(Z^i_{j_1})^2} \\
&+ \frac{Z^i_{j_4}I_{(j=j_4 \wedge k \leq k_c)}exp(\mu_{ijk}) - Z^i_{j_4k_c}I_{(j=j_4)}exp(\mu_{ijk})}{(Z^i_{j_4})^2} \\
&- \frac{Z^i_{j_2}I_{(j=j_2 \wedge k \leq k_c)}exp(\mu_{ijk}) - Z^i_{j_2k_c}I_{(j=j_2)}exp(\mu_{ijk})}{(Z^i_{j_2})^2} \\
&- \frac{Z^i_{j_3}I_{(j=j_3 \wedge k \leq k_c)}exp(\mu_{ijk}) - Z^i_{j_3k_c}I_{(j=j_3)}exp(\mu_{ijk})}{(Z^i_{j_3})^2}] \\
&= 2I_{(\delta \geq 0)}\delta\, exp(\mu_{ijk})(I_{(j=j_1)} + I_{(j=j_4)} - I_{(j=j_2)} - I_{(j=j_3)})\frac{I_{(k \leq k_c)}Z^i_j - Z^i_{jk_c}}{(Z^i_j)^2} \quad (7)
\end{aligned}$$

where I is the indicator function, $Z^i_{jk_c} = \sum_{k=1}^{k_c} exp(\mu_{ijk})$, and $Z^i_j = Z^i_{jr_i}$. This gradient is very similar to the one obtained by Altendorf et al. [8]. The key difference is that in their formalism, every constraint inequality applied to two parameters, but our constraint inequality is applied to four parameters (assuming two parents of a random variable and all the variables are binary valued). This is due to the fact that monotonicities are associated with a single parent but synergies exist in a set of parents where each of the parent has a monotonic relationship with the target. Note that while we define these gradients with only two parents for brevity, they can be easily extended to sets of variables.

It should be mentioned that the definition of synergy we focus in this paper is different from the definition of Xiang and Jia [16]. It can be easily proved that the reinforcement in their work is equivalent with the positive monotonicity we defined in our paper and as defined by Altendorf et al. [8]. We clearly show this relationship in the appendix A. While their work focuses on the representation of monotonicities using ICI, we go further and combine synergies with Noisy-Or. In addition, we also derive the gradient for this combination and develop a learning algorithm in the next section.

4 Learning Parameters in Presence of Qualitative and Independence Knowledge

In the previous section, we presented the idea of using monotonicities and synergies as domain knowledge that makes it possible to learn from sparse data. However, it is a tedious work to list all constraints inequalities when there are a large number of parents and unnecessary when the qualitative constraint sets are independent with each other. If there are totally n parents and one of them (say X_c) has monotonic constraints on the resultant CPT, then the number of the

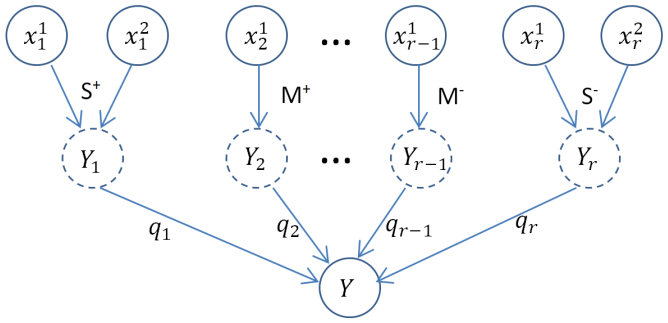

Fig. 1: Noisy-Or Bayesian Network with Qualitative Constraints.

constraints inequalities is proportional to $\prod_{X_j \in Pa_i \setminus X_c} r_j$, which is exponential in the number of total parents. The Noisy-Or structure, however, can make the number of constraints linear in the number of parent sets. In this work, we propose to use Noisy-Or to separate the influence of the different *sets* of qualitative constraints. So the inherent assumption is that the different sets of influences are independent of each other and the final structure is simply a Noisy-Or of the resulting distributions. It can be shown that introducing an extra layer of hidden nodes can still preserve the qualitative constraints of the features on the output, which exist between them in the original BN (see appendix B).

An example scenario is presented in Figure 1 where x_i^j represents the j^{th} nodes in i^{th} constraint set. As can be seen the sets of parents can have a synergistic effect ($S+$), sub-synergistic effect (S-), monotonic ($M+$) or anti-monotonic (M-) effect. Each of these parent sets yields a distribution over the target (which is essentially a hidden node Y_i that is not observed in the data). These different distributions are then combined using the Noisy-Or combining rule where each of these can have an inhibition probability $(1 - q_i)$. In this work, we learn the parameters of the conditional distributions and the inhibition parameters.

Algorithm 1 presents the process of learning the parameters of conditional distributions and inhibitions given these qualitative statements and conditional influences (where α and β indicate the descent step size of CPT parameter and Noisy-Or parameter). The qualitative constraints are only applied on the CPTs of hidden node. So, the objective function of Noisy-Or parameters q_i is the log-likelihood function while the objective function J of CPT parameters is log-likelihood function minus the sum of all involved penalty functions times a penalty weight ω. It is an iterative procedure where we first learn the inhibition probabilities of the different combinations. Then using these *estimated* probabilities, we estimate the parameters of the conditional distributions subjected to the appropriate qualitative influences. This procedure is continued till convergence. It is possible that the algorithm sometimes may not converge to a feasible solution that satisfies all the constraints. In such cases, we increase the weight of the penalty so that the solution does not go outside the feasible region. It must

Algorithm 1 Parameter Learning in Noisy-Or BN Combining Qualitative Constraints

1. Initialize the parameters μ_{ijk} and q_i randomly
2. Repeat untill convergence:
   for $i = 1 \rightarrow r_l$ **do**
   Noisy-Or Parameter Gradient Step:
   Compute the gradient of Noisy-Or parameters $\frac{\partial LL}{\partial q_i}$ for all the q_i.
   Noisy-Or Parameter Update Step:
   Update each q_i by $q_i = q_i + \beta \frac{\partial LL}{\partial q_i}$
   for $j = 1 \rightarrow v_i$ **do**
   CPT Parameter Gradient Step:
   Compute the gradient of CPT parameters
   $$\frac{\partial J}{\partial \mu_{ijk}} = \frac{\partial LL}{\partial \mu_{ijk}} - \omega \sum \frac{\partial P_j^{i,k_c}}{\partial \mu_{ijk}}$$
   for each hidden node Y_i given every possible configuration of its parents
   CPT Parameter Update Step:
   Update each μ_{ijk} by $\mu_{ijk} = \mu_{ijk} + \alpha \frac{\partial J}{\partial \mu_{ijk}}$
   end for
   end for
3. If outside the feasible region, increase the penalty weight ω and repeat step 2

be mentioned that we are *not learning* the qualitative relationships but assume that these are given.

We use e_l to indicate the l^{th} training example, r_l to denote the number of qualitative constraints sets the l^{th} instance have, $X_{l,i}$ to represent the input vector of i^{th} constraints set in l^{th} training example, q_i as the conditional probability $P(Y = 1|Y_i = 1)$. The loglikelihood function in Noisy-Or BN combining multiple constraints sets is given by $LL = \sum_l log(P(y_l|e_l))$ where $P(y = 1|e_l)$ is

$$\begin{aligned} P(y=1|e_l) &= 1 - \prod_{i=1}^{r_l} [P_i(y=0|\mathbf{X}_{l,i}) + (1-q_i)P_i(y=1|\mathbf{X}_{l,i})] \\ &= 1 - \prod_{i=1}^{r_l} [P_i(y=0|\mathbf{X}_{l,i}) + (1-q_i)(1-P_i(y=0|\mathbf{X}_{l,i}))] \\ &= 1 - \prod_{i=1}^{r_l} [1 - q_i + q_i P_i(y=0|\mathbf{X}_{l,i})] \end{aligned} \tag{8}$$

Substitute Equation 3, we get:

$$P(y=1|e_l) = 1 - \prod_{i=1}^{r_l} [1 - q_i + q_i \frac{exp(\mu_{ij0})}{exp(\mu_{ij0}) + exp(\mu_{ij1})}] \tag{9}$$

Taking the derivative of the loglikelihood function with respect to μ_{ijk}, we get

$$\frac{\partial LL}{\partial \mu_{ijk}} = \sum_l \frac{1}{P(y_l|e_l)} \frac{\partial P(y_l|e_l)}{\partial \mu_{ijk}} = \sum_l [\frac{1}{P(y_l|e_l)}(-1)^y(-1)^k \varphi(e_l)] \tag{10}$$

Where,

$$\begin{aligned}\varphi(e_l) &= q_i[\frac{\partial}{\partial \mu_{ij0}}(\frac{exp(\mu_{ij0})}{exp(\mu_{ij0}) + exp(\mu_{ij1})})] \prod_{i' \neq i}[1 - q_{i'} + q_{i'} P_{i'}(y = 0|\mathbf{X}_{l,i'})] \\ &= q_i[\frac{exp(\mu_{ij0})exp(\mu_{ij1})}{(exp(\mu_{ij0}) + exp(\mu_{ij1}))^2}] \prod_{i' \neq i}[1 - q_{i'} + q_{i'} P_{i'}(y = 0|\mathbf{X}_{l,i'})] \end{aligned} \tag{11}$$

The gradient of loglikelihood function with respect to q_i is given by:

$$\frac{\partial LL}{\partial q_i} = \sum_l \frac{1}{P(y_l|e_l)} \frac{\partial P(y_l|e_l)}{\partial q_i} = \sum_l [\frac{1}{P(y_l|e_l)}(-1)^y \phi(e_l)] \tag{12}$$

$$\phi(e_l) = (P_i(y = 0|\mathbf{X}_{l,i}) - 1) \prod_{i' \neq i}[1 - q_{i'} + q_{i'} P_{i'}(y = 0|\mathbf{X}_{l,i'})] \tag{13}$$

Once this gradient is obtained, we perform the iterative update of the Noisy-Or parameters and the CPT parameters as shown in Algorithm 1.

The natural question to ask is, where does the knowledge come from? We believe that, in many domains such as medicine, obtaining this knowledge is natural – for instance, there exists published research in understanding interactions of risk factors when predicting a disease, say heart attack. From this perspective, our proposed work here can be considered as enabling domain experts to provide more information to guide the algorithms in their search through the space of parameters. In addition, our algorithms can provide a method to evaluate the extent to which the domain knowledge is correct – it can determine the violations of the constraints in the training data. So we provide a method by which the domain experts can include some knowledge that is fully satisfied by the data and their best guesses at other relationships. Our algorithms can naturally fit the true knowledge and determine how much of the guesses are true. As we show in our experiments, there are some cases, where the independence between the sets of relationships may not be always true and in some cases, the monotonicities are as valuable as synergies. We aim to understand the interplay between the qualitative constraints and Noisy-Or and aim to determine if the combination is indeed a powerful method to exploit prior knowledge.

5 Experimental Evaluation

In this section, we present the results of evaluating our algorithm on 11 different standard machine learning domains from the UCI repository. The key questions that we seek to ask in our experiments are:

Q1: How does the use of qualitative constraints compared to not using any influence statements?

Q2: How valid is the independence assumption (i.e., how good is only using Noisy-Or)?

Q3: How does synergy compare to monotonicity?

Q4: How does the addition of the conditional influences with qualitative constraints help?

For each dataset, we learn the parameters by implementing six algorithms: (i) learning merely from data, (ii) with monotonic constraints, (iii) with synergy constraints, (iv) learning with Noisy-Or structure, (v) monotonic constraints plus Noisy-Or and (vi) synergy constraints plus Noisy-Or. All used features are discretized into two states under the following rules: i) nominal variables such as sex, race are assigned a class based on their qualitative relationships with their children nodes; ii) ordinal variables (e.g. {small, med, big}) are divided into two classes based on their distributions; iii) continuous values such as blood pressure, thyroxine are discretized according to practical thresholds in corresponding domains. The AUC-ROC and P values are calculated to compare their performances. We perform 10-fold cross-validation on all the domains for parameter selection and present the results on test set.

Domain	Target Attribute	Num of Parents	Num of Samples
Heart Disease	Diagnosis of HD	5	297
Breast Cancer	BC recurrence	5	286
Credit Approval	Card Approval	5	300
Car Evaluation	Acceptable or not	6	300
Pima Indian Diabetes	Diabetes status	4	300
Census Income	$> 50K$ or $\leqslant 50K$	7	300
Iris	Versicolour or Virginica	4	100
Glass Identification	float or non_float(Building_windows)	5	146
Ecoli	Protein localization	5	284
Thyroid Disease	TD status	5	185
Hepatitis	Death of hepatitis	5	144

Table 2: Details of the experimental domains.

Table 2 presents the target attribute of interest in each domain in the second column. The third column lists the number of parents and the final column presents the number of all instances (sum of training set and test set whose proportion is about 10:1 in every domain). For the different domains, we provide prior knowledge– synergies and monotonicities whenever applicable. An example of such a network is presented in Figure 2. As can be seen, this is in the breast cancer domain where the goal is to predict recurrence of breast cancer based on 5 different attributes {age, menopause, tumor-size, deg-malig, irradiation}. In its Noisy-Or model, irradiation has a negative monotonic effect on the probability

of recurrence and the others have positive monotonic effects. Parent set {age, menopause} has a synergistic interaction while {tumor-size, deg-malig} is sub-synergistic.

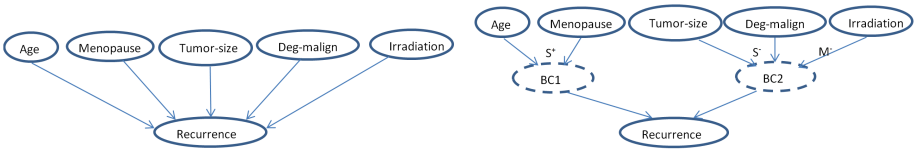

Fig. 2: An example domain without (left) and with (right) qualitative influences and Noisy-Or.

The results of using the different algorithms are presented in Figure 3 where the consolidated AUC-ROC over all the domains is presented. The first bar graph of every domain is a simple inverted naive Bayesian network where every feature is considered to be a parent of the target and the parameters are learned. The subsequent ones are (in that order) – *Noisy-Or*, *monotonicity* constraints [8], *synergies*, *monotonicity with Noisy-Or* and *synergy with Noisy-Or*. Hence, the last three bar graphs are the algorithms presented in this work. As can be seen, very clearly, in all the domains, the use of qualitative constraints and qualitative constraints with Noisy-Or outperform simply learning the conditional distributions from data. Hence **Q1** can be answered affirmatively.

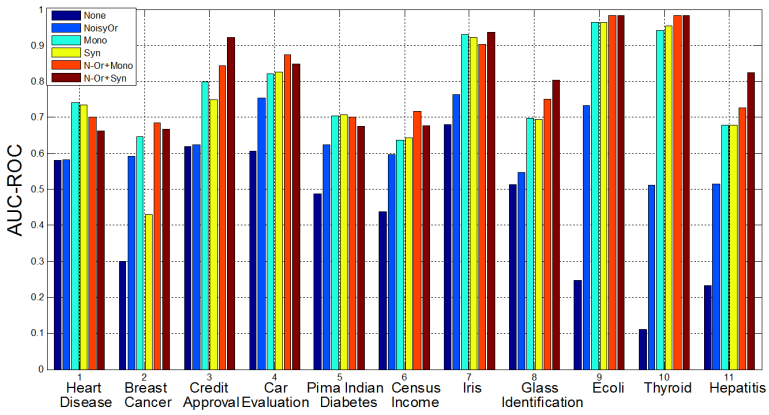

Fig. 3: Performance of the different algorithmic settings in several domains. Best viewed in color.

The interesting observation is that Noisy-Or assumption seems to be a strong one in several domains. In many domains, the use of Noisy-Or is better than

assuming no knowledge in almost all the domains. But the use of only qualitative statements such as monotonicity and synergy yield significantly better results in 9 domains when compared to only using Noisy-Or. Hence, it is clear that the answer to **Q2** is that only using Noisy-Or is not sufficient for a majority of domains. Comparing monotonicity and synergies, it is clear that there is not much difference in several domains – except for breast cancer domain where synergy seems to be significantly worse than monotonicity and Noisy-Or. Hence, in answer to **Q3**, there is no significant difference between using monotonocity and synergy constraints. It remains interesting to understand the situations in which the use of synergy is more useful than the monotonicities.

Finally, the combination of qualitative and conditional influences seems to perform the best in most of the domains. The results are comparable to or better than simple qualitative constraints in all the domains. In 7 domains, the use of conditional influences seems to improve upon the use of only qualitative constraints. While in 3 others, there is no significant change in performance by adding conditional influences. Only in one domain (Breast Cancer), there is a very small dip (that is **not** statistically significant) in the AUC-ROC values. Hence, to answer **Q4**, we can affirmatively state that the use of conditional influences improves the performance of qualitative influence relationships in a majority of domains. Interestingly, the use of Noisy-Or with synergies improves upon Noisy-Or with monotonicity in three domains while in other domains the results are comparable. This is very similar to the observation about **Q3** where synergies and monotonicities exhibit comparable performance in most domains. All the significance results reported here are the results of using t-test with p-values < 0.05. Figure 4 presents the learning curves comparing the use of qualitative influence plus Noisy-Or against simple Noisy-Or and using no knowledge in two sample domains. The performance using prior knowledge has a jump start and faster convergence in both the domains, justifying the use of qualitative and conditional influence statements in these domains.

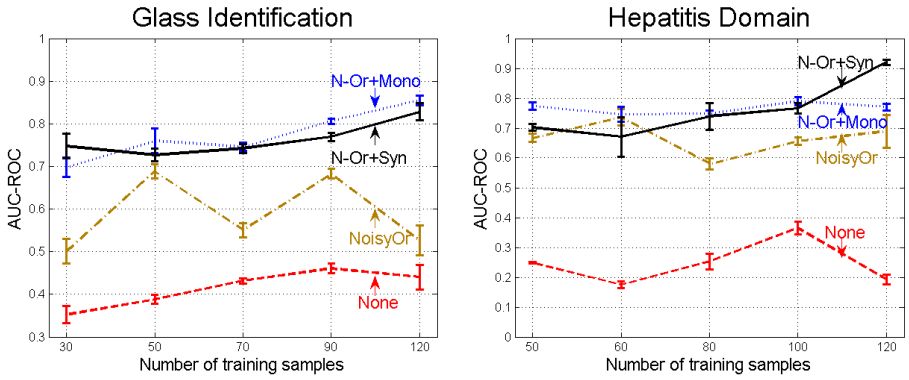

Fig. 4: Learning Curve in two domains with no knowledge, Noisy-Or and qualitative influences + Noisy-Or.

6 Conclusion

We presented a framework for combining qualitative and conditional influence statements when biasing probabilistic learners. We formalized the notion of synergistic interactions and derived the gradients for learning in the presence of such statements. We then extended our model to include conditional influences such as Noisy-Or and derived an algorithm for learning in presence of these two types of constraints. We evaluated our algorithms on 11 different domains and the results conclusively proved that the use of qualitative influences when combined with conditional influences yields a better performance in a majority of the domains. Our goal is to next understand the different types of conditional influences and their interactions with qualitative constraints. Exploring the use of such constraints in learning the structure of a full Bayesian network remains a very interesting direction for the future research.

Acknowledgments Sriraam Natarajan gratefully acknowledges support of the DARPA DEFT Program under the Air Force Research Laboratory (AFRL) prime contract no. FA8750-13-2-0039. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

A Equivalence of Reinforcement to positive monotonicity

If variable Y is resulted from a set of causes $\mathbf{X}$, the causes in $\mathbf{X}$ are said to *reinforce* each other if for any subset $\mathbf{X'} \subset \mathbf{X}$ the following holds [16]: $P(Y \quad is \quad true|\mathbf{X'} \quad is \quad true) \leqslant P(Y \quad is \quad true|\mathbf{X} \quad is \quad true)$

Proof Assume variable Y has three parents $\{x_1, x_2, x_3\}$ all of which have positive monotonic influence on Y ($x_{1\succ}^{M+}Y$, $x_{2\succ}^{M+}Y$, $x_{3\succ}^{M+}Y$)and all variables are binary. The monotonic constraints of $x_{1\succ}^{M+}Y$ at the context of C=($\{x_2 = 1, x_3 = 1\}, \{x_2 = 1, x_3 = 0\}, \{x_2 = 0, x_3 = 1\}$) is

$$P(Y = 1|x_1^1, x_2^1, x_3^1) \geqslant P(Y = 1|x_1^0, x_2^1, x_3^1) \tag{14}$$

$$P(Y = 1|x_1^1, x_2^1, x_3^0) \geqslant P(Y = 1|x_1^0, x_2^1, x_3^0) \tag{15}$$

$$P(Y = 1|x_1^1, x_2^0, x_3^1) \geqslant P(Y = 1|x_1^0, x_2^0, x_3^1) \tag{16}$$

$x_{2\succ}^{M+}Y$ at the context of C=($\{x_1 = 1, x_3 = 1\}, \{x_1 = 1, x_3 = 0\}$) is

$$P(Y = 1|x_1^1, x_2^1, x_3^1) \geqslant P(Y = 1|x_1^1, x_2^0, x_3^1) \tag{17}$$

$$P(Y = 1|x_1^1, x_2^1, x_3^0) \geqslant P(Y = 1|x_1^1, x_2^0, x_3^0) \tag{18}$$

$x_{3\succ}^{M+}Y$ at the context of $C = (\{x_1 = 1, x_2 = 1\})$ is

$$P(Y = 1|x_1^1, x_2^1, x_3^1) \geqslant P(Y = 1|x_1^1, x_2^1, x_3^0) \tag{19}$$

Inequ.18 and **Inequ.19** $\Rightarrow P(Y = 1|x_1^1, x_2^1, x_3^1) \geqslant P(Y = 1|x_1^1, x_2^0, x_3^0)$
Inequ.15 and **Inequ.19** $\Rightarrow P(Y = 1|x_1^1, x_2^1, x_3^1) \geqslant P(Y = 1|x_1^0, x_2^1, x_3^0)$

Inequ.16 and **Inequ.17**$\Rightarrow P(Y=1|x_1^1,x_2^1,x_3^1) \geqslant P(Y=1|x_1^0,x_2^0,x_3^1)$

The inequalities above can be presented as the probability of Y is true given all the causes $\{x_1,x_2,x_3\}$are activated is no less than that of only part of the causes $(\{x_1\},\{x_2\},\{x_3\},\{x_2,x_3\},\{x_1,x_3\},\{x_1,x_2\})$ is activated, which is the definition of reinforce.

B Sub-synergy and synergy constraints can be preserved in Noisy-Or structure.

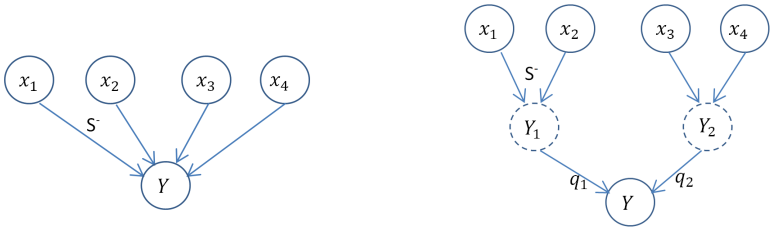

Fig. 5: Sub-synergy in One Layer BN (left) and Noisy-Or BN (right)

Assume variable Y has four parents $\{x_1,x_2,x_3,x_4\}$, all the variables are binary and $x_{1\succ}^{M+}Y$, $x_{2\succ}^{M+}Y$ sub-synergistically (as shown in Figure 5). Sub-synergy constraints of x_1 and x_2 on variable Y given the context $\{x_3^i,x_4^j\}$ is given by:

$$\begin{aligned}&P(Y=0|x_1^1,x_2^1,x_3^i,x_4^j)+P(Y=0|x_1^0,x_2^0,x_3^i,x_4^j)\geqslant\\&P(Y=0|x_1^1,x_2^0,x_3^i,x_4^j)+P(Y=0|x_1^0,x_2^1,x_3^i,x_4^j)\end{aligned}\tag{20}$$

In the Noisy-Or structure, we can introduce two hidden nodes Y_1 and Y_2 the sub-synergy constraint of x_1 and x_2 on hidden node Y_1 is given by:

$$P(Y_1=0|x_1^1,x_2^1)+P(Y_1=0|x_1^0,x_2^0)\geqslant P(Y_1=0|x_1^1,x_2^0)+P(Y_1=0|x_1^0,x_2^1)\tag{21}$$

Based on Equation 8, we have

$$\begin{aligned}P(Y=0|x_1^1,x_2^1,x_3^i,x_4^j)&=[1-q_1+q_1P(Y_1=0|x_1^1,x_2^1)]\times[1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]\\P(Y=0|x_1^0,x_2^0,x_3^i,x_4^j)&=[1-q_1+q_1P(Y_1=0|x_1^0,x_2^0)]\times[1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]\\P(Y=0|x_1^1,x_2^0,x_3^i,x_4^j)&=[1-q_1+q_1P(Y_1=0|x_1^1,x_2^0)]\times[1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]\\P(Y=0|x_1^0,x_2^1,x_3^i,x_4^j)&=[1-q_1+q_1P(Y_1=0|x_1^0,x_2^1)]\times[1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]\end{aligned}$$

Since q_1 is a probability which is no less than zero, multiply q_1 to Inequality 21 we get

$$q_1P(Y_1=0|x_1^1,x_2^1)+q_1P(Y_1=0|x_1^0,x_2^0)\geqslant q_1P(Y_1=0|x_1^1,x_2^0)+q_1P(Y_1=0|x_1^0,x_2^1)$$

Add $(1-q_1)$ to every item,

$$\begin{aligned}&[1-q_1+q_1P(Y_1=0|x_1^1,x_2^1)]+[1-q_1+q_1P(Y_1=0|x_1^0,x_2^0)]\geqslant\\&[1-q_1+q_1P(Y_1=0|x_1^1,x_2^0)]+[1-q_1+q_1P(Y_1=0|x_1^0,x_2^1)]\end{aligned}$$

Since $1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)=1-q_2P(Y_2=1|x_3^i,x_4^j)$, which is no less than zero, multiply it with the above inequality,

$$\begin{aligned}
&[1-q_1+q_1P(Y_1=0|x_1^1,x_2^1)][1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]\\
&+[1-q_1+q_1P(Y_1=0|x_1^0,x_2^0)][1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]\geqslant\\
&[1-q_1+q_1P(Y_1=0|x_1^1,x_2^0)][1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]\\
&+[1-q_1+q_1P(Y_1=0|x_1^0,x_2^1)][1-q_2+q_2P(Y_2=0|x_3^i,x_4^j)]
\end{aligned}$$

which is equivalent with Inequality 20.

It is easy to prove the transitivity of monotonic constraints in the proposed model. The process is similar as this one, which will not be shown here.

References

1. Heckerman, D., Geiger, D., Chickering, D.: Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning **20** (1995) 197–243
2. Vomlel, J.: Noisy-or classifier. International Journal of Intelligent Systems **21**(3) (March 2006) 381398
3. Heckerman, D., Breese, J.: A new look at causal independence. In: UAI. (1994)
4. Srinivas, S.: A generalization of the Noisy-Or model. In: UAI, San Francisco, CA, Morgan Kaufmann (1993) 208–215
5. Pearl, J.: Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann Publishers Inc. (1988)
6. Vomlel, J.: Exploiting functional dependence in Bayesian network inference. In: UAI, Edmonton, Canada (2002) 528–535
7. Wellman, M.: Fundamental concepts of qualitative probabilistic networks. Artificial Intelligence **44** (1990) 257303
8. Altendorf, E., Restificar, A., Dietterich, T.: Learning from sparse data by exploiting monotonicity constraints. In: UAI. (2005) 18–26
9. Feelders, A., van der Gaag, L.: Learning Bayesian network parameters with prior knowledge about context-specific qualitative influences. In: UAI. (2005) 193–200
10. Campos, C., Tong, Y., Ji, Q.: Constrained maximum likelihood learning of Bayesian networks for facial action recognition. In: ECCV. (2008) 168–181
11. Tong, Y., Ji, Q.: Learning Bayesian networks with qualitative constraints. In: CVPR. (2008)
12. Brunk, H.: Maximum likelihood estimates of monotone parameters. The Annals of Mathematical Statistics **26**(4) (Dec 1955) 607–616
13. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: UAI. (1996) 115–123
14. Robertson, T., Wright, F., Dykstra, R.: Order Restricted Statistical Inference. Wiley Chichester (1988)
15. Pekka, J., Erkki, V., Jaakko, T., Pekka, P.: Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14786 middle-aged men and women in finland. Circulation **99** (1999) 1165–1172
16. Xiang, Y., Jia, N.: Modeling causal reinforcement and undermining for efficient cpt elicitation. IEEE Transactions on Knowledge and Data Engineering **19**(12) (2007) 1708–1718