Variational hybridization and transformation for large inaccurate noisy-or networks

Yusheng Xie  Nan Du  Wei Fan  Jing Zhai  Weicheng Zhu
Baidu Research
Sunnyvale, CA 94089
{xieyusheng,dunan,fanwei03,zhaijing01,zhuweicheng}@baidu.com

Abstract

Variational inference provides approximations to the computationally intractable posterior distribution in Bayesian networks. A prominent medical application of noisy-or Bayesian network is to infer potential diseases given observed symptoms. Previous studies focus on approximating a handful of complicated pathological cases using variational transformation. Our goal is to use variational transformation as part of a novel hybridized inference for serving reliable and real time diagnosis at web scale. We propose a hybridized inference that allows variational parameters to be estimated without disease posteriors or priors, making the inference faster and much of its computation recyclable. In addition, we propose a transformation ranking algorithm that is very stable to large variances in network prior probabilities, a common issue that arises in medical applications of Bayesian networks. In experiments, we perform comparative study on a large real life medical network and scalability study on a much larger (36,000x) synthesized network.

1 Introduction

Noisy-or Bayesian network (NOBN) is a popular class of statistical models in modeling observable events and their unobserved potential causes. One of the best known medical applications of NOBN is Quick Medical Reference (QMR-DT) ([1991]). QMR-DT describes expert-assessed relationships between 4,000+ observable binary symptom variables (collectively denoted as $S$) and 500+ binary latent disease variables (collectively denoted as $D$) as illustrated in Figure 1 (a).

We improve variational inference for a large QMR-DT style NOBN in areas of scalability, stability, and accuracy to previously unattainable or untested levels. As part of a medical messaging bot, the inference goal is to perform reliable real time diagnosis at web scale. Figure 1(b) shows the messaging bot’s interface. The ongoing project aims to serve a substantial portion of Internet users who experience health issues (e.g., 3 to 8 million daily active users[1]) with reliable disease diagnosis that is more accurate and accessible than text-based web searches, web searches that emphasize retrieval similarity but lack clinical technicality (e.g., 38.5 $^\circ$C fever lasting 3 days and 39.5 $^\circ$C fever lasting 8 days. The latter could be 20x more fatal in probability). The developing bot has completed 1,000+ organic, non-scripted dialogues with 100+ qualified human testers. Assessed by 50+ licensed doctors, the network plans to cover all conceivable human diseases and health conditions[2] approximately 40,000 (80x that of QMR-DT) according to the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). To the best of our knowledge, the aforementioned scales make it the largest medical application of noisy-or Bayesian networks.

[1]: Assume an average person is sick 2-4 days per year and our reachable population is 600 to 800 million.
[2]: A sub-network focusing on maternal and infant care is completed and used in our experiments.
Recent advances in modern machine learning and artificial intelligence quickly proliferate far beyond the traditional Bayesian framework. But for mission critical applications such as medical diagnosis, one prefers Bayesian network-based approach for reasoning instead of entirely data driven approach. The reasons are due to traceable outcome, easy debuggability, and provenance. Data source unreliability and scarcity also prevent some medical applications from taking full advantage of the large body of data driven algorithms that can be quickly accelerated by larger datasets (e.g., machine translation \cite{luong2015better}, speech recognition \cite{amodei2015deep}). For example, the Caroli disease\footnote{Caroli disease is a type of congenital dilatation of intrahepatic bile duct. It has the code Q44.6 in ICD-10.} has fewer than 250 recorded cases worldwide, making it almost impossible to “gather/label more data points”. On the other hand, no disease should be too rare to deserve attention From an ethical perspective, even a 1-in-1,000,000 chance (technically \textit{extremely rare}) translates into over 6,000 suffering individuals worldwide. From an academic perspective, understanding rare diseases brings irreplaceable medical knowledge.

The expert-assessed probability of observing symptom $f$ given only disease $d$ is denoted as $P(f^+ | d^+)$.

We use $\pi(f)$ to denote $\{d \mid d \in D, P(f^+ | d^+) > 0\}$, the set of diseases that could cause $f$ with non-zero probability. Like QMR-DT, we assumed $P(d^+)$ for each $d \in D$: the prior probability of having disease $d$ without observing any symptoms. We further define $P(f^-)$ and $P(d^-)$ notations as $P(f^-) = 1 - P(f^+)$ and $P(d^-) = 1 - P(d^+)$, respectively.

In a typical diagnosis session, the user first inputs her positive and negative findings: $F^+ = \{f_1^+, f_2^+, \ldots\} \subset S$, $F^- = \{f_1^-, f_2^-, \ldots\} \subset S$. Then the model performs inference to calculate $P(F^+, F^-)$, which is the crux in deriving the conditional $P(d_i^+ | F^+, F^-)$ for each $d \in D$.

\subsection{Background on variational inference}

The exact inference for $P(F^+, F^-)$ is intractable \cite{cooper1990do} and intractability motivates investigations into approximation inference algorithms. The variational method \cite{jordan1999introduction, jaakkola1999dynamic} and the mean field local approximation \cite{ng2000efficient} are both hybrid approximation algorithms.

To describe the variational approximation, let $\theta_{j_i} \equiv - \log P(f_j^- | d_i^+)$. \cite{jordan1999introduction, jaakkola1999dynamic} show that

\begin{equation}
    P\left(f_j^+ | \pi(f_j)^+\right) = e^{\sum_{i=1}^{\pi(f_j)} \theta_{j_i}} \leq e^{\sum_{i=1}^{\pi(f_j)} \xi_{j_i} - f^*(\xi_j)} \equiv P(f_j^+ | \pi(f_j), \xi_j)
\end{equation}

and

\begin{equation}
    P(f_j^+ | \xi_j) = \prod_{d_i \in \pi(f_j)} \left[ P(f_j^+ | d_i^+, \xi_j) \cdot P(d_i^+) + P(f_j^+ | d_i^-, \xi_j) \cdot P(d_i^-) \right] \geq P(f_j^+),
\end{equation}

\footnote{Without loss of generality, the leak probabilities \cite{jordan1999introduction} are omitted in our discussion.}
where $\xi_j$ is the free variational parameter, $f(x) \equiv \log (1 - e^{-x})$, and $f(x)$’s convex conjugate function takes the form $f^*(\xi) = -\xi \log \xi + (\xi + 1) \log(\xi + 1)$, for $\xi > 0$. Equation 2 transforms $P(f_j^+) \rightarrow$ into its variational upper bound $P(f_j^+ \mid \xi_j)$ using the inequality from conjugate duality.

Breaking $F^+ = \{f_1^+, f_2^+, \ldots\}$ into the partition $F_1^+ \cup F_2^+$ allows exact inference on $F_1^+$ and variational inference on $F_2^+$. \cite{Jaakkola:1999} (JJ99) calculates the joint variational posterior as

$$P_{JJ99}(F_1^+, F_2^+, F^- \mid \Xi^{\text{min}}) = e^{-\sum_{j=1}^n \xi_j f_j^+} \prod_{d_i \in F_2^+} \left[ \frac{\sum_{\xi_j \in F_j^+} \xi_j \theta_{ij}}{\sum_{\xi_j \in F_j^+} \theta_{ij}} \cdot \left( P(d_i^+ \mid F_1^+, F^-) + P(d_i^- \mid F_2^+, F^-) \right) \right],$$

where $\Xi = \{\xi_1, \xi_2, \ldots\}$. Finding $\arg \min_{\Xi} P(F^+ \mid \Xi)$ can be relaxed to finding $\arg \min_{\xi_j} \log P(f_j^+ \mid \xi_j)$ for each $\xi_j \in \Xi$. The $\xi$-convexity permits second order optimization methods (CVX) to find each $\xi_j$. From Equation 3, the 1st order partial derivatives are

$$\frac{\partial}{\partial \xi_j} \log P(f_j^+ \mid \xi_j) = \log \frac{\xi_j}{1 + \xi_j} + \sum_{d_i \in \xi_j(f_j^+)} \theta_{ij} e^{-\xi_j \theta_{ij} + 1}.$$  

where $p_i \equiv P(d_i^+) / P(d_i^-)$ is the inverse prior odds for the $i$th disease. The 2nd order partial derivatives are derived mechanically. Figure 1(c) illustrates the complexity of exact and variational inference in real application. More discussion on existing inference algorithms are in related works section (see Table 2).

### 2 Inaccuracy in widely-ranged disease priors

Inaccurate hidden variable prior is a recognized \cite{Jernite:2013, Mansinghka:2006} but often avoided \cite{Cheng:2002, Liao:2009, Riggelsen:2006} issue in NOBN. Inaccuracy in disease prior is among the most likely errors in constructing a NOBN for medical applications. Real life disease priors can span several orders of magnitude. For example, acne (ICD-10 code: L70.0) affects 80% to 90% teenagers in the western world \cite{Dawson:2013} while syndromes like the Caroli disease have historical infection rates less than 0.00001%. It is very likely, even for medical experts or statistical estimators, to misjudge the prior probability by an order of magnitude relative to other very rare or very common diseases. So it is beneficial to obtain fast and accurate variational algorithms that are resistant to the large variances in disease priors.

In the following two sections, we propose inference algorithms that can greatly immunize the current variational inference against inaccuracy in disease priors.

### 3 Variational-first hybridization and joint hybridization

The $F^+ = F_1^+ \cup F_2^+$ partition employed in JJ99 is a realization of the classic hybrid paradigm: balancing accuracy and runtime over the entire $F^+$ by 1) applying different posterior estimators (variational, exact, MCMC, etc.) to $F_1^+ \cup F_2^+$, and 2) controlling their cardinalities. But JJ99 has two main drawbacks that prevent it from fulfilling the scalability and stability requirements in building a web diagnostic bot.

First, Equation 3 estimates $\Xi^{\text{min}}$ by using the exactly treated disease posterior $P(d_1^+ \mid F_2^+, F^-)$. The $\Xi^{\text{min}}$ estimations need be recalculated for every case of $F_1^+ \cup F_2^+ \cup F^-$ since each case would produce different disease posteriors that affect the gradients in Equation 2. Second, in order to pass confident posteriors to its variational step, JJ99 basically “primes” the potentially inaccurate disease priors with evidences from $F_2^+ \cup F^-$. Since the hybridized complexity decreases exponentially w.r.t. to $|F_1^+|$, $F_1^+$ usually contains less evidence than $F_2^+ \cup F^-$ in practice (i.e., $|F_1^+| < |F_2^+ \cup F^-|$). In other words, JJ99 uses a substantial portion of the evidence in priming the unaudited priors first and then refines the posterior probabilities using the smaller leftover portion of evidence.

We propose the variational-first hybridization (VFH) that can fix both issues. Described in Algorithm 1 VFH performs inference on $F_1^+$ first (to prime the unaudited priors) and on $F_2^+$ and $F^-$ later (to refine the posteriors). Calculating $\Xi^{\text{min}}$ in VFH relies on disease priors instead of posteriors.
Algorithm 1: the proposed variational-first hybridization (VFH) algorithm.

**Input:** $F_1^+$, list of positive findings to be inferred variationally, $F_2^+$, list of positive findings to be inferred exactly, $F^-$, list of negative findings to be inferred exactly, $\theta_{j}i$ for each $f_j \in S$ and $d_i \in D$, $P(d_i)$, disease prior probability for each $d_i \in D$.

**Output:** The joint variational evidence of given findings $F_1^+$, $F_2^+$, and $F^-$.  
1. Calculate $P(F_1^+|\Xi)$ as a function of $\Xi$ from Equation 6.
2. $\Xi^{\min} \leftarrow \arg\min_{\Xi} P(F_1^+|\Xi)$ using Newton’s method on its derivatives (shown in Equation 4).
3. for each $d_i \in D$ do
4. Calculate $P(d_i^+|F_1^+, \Xi^{\min})$ from Equation 6 and $P(F_1^+|\Xi)$.
5. $P(d_i) \leftarrow P(d_i^+|F_1^+, \Xi^{\min})$ (update disease priors with posteriors).
6. $P(F_2^+, F^-) \leftarrow$ Quickscore($F_2^+, F^-$).
7. return $P(F_2^+, F^-)$

Therefore, the calculation is invariant to the findings that make up $F_1^+ \cup F_2^+ \cup F^-$. Invariant $\Xi^{\min}$ allows caching $\Xi^{\min}$ values and leads to faster inference as summarized in Table I.

Equation 6 explicitly expresses the joint variational evidence of given findings using VFH:

$$P_{VFH}(F_1^+, F_2^+, F^-|\Xi^{\min}) = \sum_{P' \in 2^{F_2^+}} (-1)^{|P'|} \prod_{i=1}^{|P'|} P(f_j^-|d_i^+) P(d_i^+|F_1^+, \Xi^{\min}) + P(d_i^-|F_1^+, \Xi^{\min}),$$

where $2^{F_2^+}$ denotes the power set of $F_2^+$ and the $P(d_i^+|F_1^+, \Xi^{\min})$ terms are calculated from

$$P(F^+|\Xi) = e^{-\sum_{j=1}^{|F^+|} f_j(\xi_j)} \prod_{d_i \in (F^+)} \left[ e^{\sum_{j=1}^{|F^+|} \xi_j \theta_{j}i} \cdot P(d_i^+) + P(d_i^-) \right].$$

Besides VFH, we can also hybridize the exact evidence $P(F_2^+, F^-)$ and the variational evidence $P(F_1^+|\Xi)$ jointly (JH):

$$P_{JH}(F_1^+, F_2^+, F^-|\Xi^{\min}) = \sum_{P' \in 2^{F_2^+}} (-1)^{|P'|} \prod_{i=1}^{|P'|} P(f_j^-|d_i^+) \prod_{k=1}^{|P'|} P(f_j^+|d_i^+, \xi^{\min}_k) P(d_i^+) + \prod_{k=1}^{|P'|} P(f_j^-|d_i^+, \xi^{\min}_k) P(d_i^-).$$

Like VFH, JH has the same advantages over JJ99 when $|F_1^+| < |F_2^+ \cup F^-|$.

3.1 Estimate $\xi^{\min}_j$ without disease prior or posterior

If we solve $\xi^{\min}_j$ from $\arg\min_{\xi_j} \log P(f_j^+|\xi_j, \pi(f_j)^+)$ instead of $\arg\min_{\xi_j} \log P(f_j^+|\xi_j)$, the resulting $\xi^{\min}_j$ has a closed form solution. To see this, take the equality in Equation 11 and let $x_j \equiv \sum_{i=1}^{|\pi(f_j)^+|} \theta_{j,i}$. The equality $e^{f(x_j)} = e^{x_j \tau_j} f^*(\xi_j)$ holds if and only if $\xi^{\min}_j = \arg\min_{\xi_j} \xi_j x_j = f^*(\xi_j)$. Simple algebra gives the closed form $\xi^{\min}_j = (e^{\tau_j} - 1)^{-1}$. Conceptually, $\arg\min_{\xi_j} \log P(f_j^+|\xi_j, \pi(f_j)^+)$ would surely result in suboptimal $\xi^{\min}_j$ due to its lack of prior knowledge. However, we find this approach competitive for a certain range of disease priors (shown in experiments). The prior/posterior-free (PPF) estimator of $\Xi^{\min}$ is independent of disease prior or posterior and allows $\xi^{\min}_j$ to be pre-computed and cached regardless of JJ99, VFH or JH.

3.2 N-scalability of JJ99, VFH, and JH

The ability to process a large number ($N$) of diagnosis with low latency is quintessential for web scalability. The variational step in JJ99-CVX (baseline) is $O(N)$, which would put increasing strain.
on the server as $N$ grows. On the other hand, the proposed VFH and JH perform the variational step in constant time w.r.t. $N$. With the proposed PPF estimator of $\Xi_{\text{min}}$, all hybridization schemes can execute variational transformation in constant time w.r.t. $N$. Table 1 summarizes the practical efficiency of the proposed variational hybridization when used with either CVX or PPF estimator of $\Xi_{\text{min}}$. Note that Table 1 only compares the cost of the variational step. We evaluate the overall inference cost for different inferencers in the Experiments section.

Table 1: Detailed temporal complexities for the proposed variational parameter estimation in terms of $|D|$, $|F_i^+|$, $|F_i^-|$, $c$, $|S|$, and $N$. All entries are Big-O complexity. Note that although JH is equivalent to VFH in variational parameter estimation, JH will have higher overall inference complexity due to difference between Equation 5 and 7.

| $\Xi_{\text{min}}$ solver | # of queries | JJ99 | VFH | JH |
|---------------------------|-------------|------|-----|----|
| CVX                       | $\infty$    | $N$  | $N \cdot |D| \cdot |F_i^+| \log \log \frac{1}{\epsilon}$ | $|D| \cdot |F_i^-| \log \log \frac{1}{\epsilon}$ |
| PPF                       | $\infty$    | $N$  | $N \cdot |D| \cdot |F_i^+| \log \log \frac{1}{\epsilon}$ | $|D| \cdot |S| \log \log \frac{1}{\epsilon}$ |

4 Variational transformation with uncertain disease priors

In addition to the inference formula (JJ99, VFH, or JH) and the $\Xi_{\text{min}}$ solver (CVX or PPF), there is a third component in variational inference that is critical to the posterior accuracy: the transformation ranking algorithm that partitions $F^+$ into $F_i^+$ and $F_i^-$, given fixed $|F_i^+|$. (Jaakkola and Jordan 1999) and (Ng and Jordan 2000) use a simple greedy heuristic ordering (GDO) algorithm to rank the order of transformation based on the greedy local optimum for further minimizing the overall variational upper bound (which is firstly minimized by setting $\Xi = \Xi_{\text{min}}$). Minimizing the overall variational upper bound is, naturally, a commendable goal. But given the inaccuracy in widely-ranged disease priors, is there an ordering algorithm that can fender off that uncertainty more effectively than GDO?

To simplify the discussion, we assume uniform $\theta_i = c$ for any $j, i$ pair such that $P(f_j^- | d_i^+) < 1$, where $c \in (0, +\infty)$. Let the random variable (r.v.) $P = \frac{1}{m} \sum_{k=1}^{m} U_k$, where $U_k \sim \text{iid Unif}(0, \frac{2}{1+p})$ for $k = 1, 2, \ldots, m$. We further assume that the inverse disease prior odds: $P(d_i^-) / P(d_i^+) = p_i$ for any $i \in \{1, \ldots, |D|\}$ are drawn independently from $P$. The choice of $m$ is rather inconsequential in our discussion. For a reasonable $m$ (e.g., $5 < m < 1,000$), the uniform mean distribution $P$ introduces Gaussian-like variance without breaking the positive definite constraint on $p_i$’s.

We desire to establish an ordering algorithm that minimizes the variance in posterior predictions due to $P$. The first step is to show its existence. Formally, it is stated and proved in Proposition 1.

Proposition 1. Fix $p \in [0, +\infty)$, $c \in (0, +\infty)$, and $n \in \{1, 2, \ldots, |F^+|\}$. Then there exists a $F_i^+ \subset F^+$ such that $|F_i^+| = n$ and $\text{Var} \left[ \log P \left( d_i | F_i^+, P, \Xi_{\text{min}} \right) : P \right]$ is approximately minimized for every $d_i \in D$.

Proof. Let the r.v. $Q_i$ denote $P \left( d_i | F_i^+, P, \Xi_{\text{min}} \right) : P$. And let $\gamma > 1$ denote the expected value of $\exp \left[ c \sum_{j=1}^{|F_i^+|} \epsilon_j^\min 1_{j_i} \right]$, where the r.v. $1_{j_i}$ models the likelihood of whether $d_i \in \pi(f_j)$. Now we can express $Q_i$ as $Q_i = \frac{\gamma}{\gamma P^+ + 1}$ and reduce $\text{Var} \left[ \log Q_i \right]$ to simple functions of $E \left[ P \right]$ and $\text{Var} \left[ P \right]$, which are known quantities of the uniform mean (Bates) distribution.

\[
\text{Var} \left[ \log Q_i \right] = \text{Var} \left[ \log \left( \frac{\gamma}{(\gamma - 1)P^+ + 1} \right) \right] = \text{Var} \left[ \log \left( (\gamma - 1)P^+ + 1 \right) \right] \approx \frac{\text{Var} \left[ (\gamma - 1)P^+ + 1 \right]}{E \left[ (\gamma - 1)P^+ + 1 \right]^2} = \frac{1}{12n} \left( \frac{2\gamma - 1}{\gamma + 1} \right)^2 = \frac{1}{3n} \left( \frac{1}{\gamma + 1} \right)^2.
\]

The “$\approx$” in Equation 8 is the result of Taylor series expansion on $\log \left( (\gamma - 1)P^+ + 1 \right)$, a common resort to approximate the moments of a (log-)$\text{transformed}$ random variable (van der Vaart 1998).
We refer to this strategy as finding-degree order (FDO).

Approximately, \( \text{Var} \left[ \log Q_i \right] \propto \gamma \). Observe that, for fixed \( n \), choosing the \( n \) smallest \( \xi_j^{\min} E[1_{ij}] \)'s will guarantee the smallest \( \gamma \). We show the existence of \( F_i^+ \subset F^+ \) by the following construction: consecutively selecting the \( f_j^+ \)'s associated with the \( n \) smallest \( \xi_j^{\min} E[1_{ij}] \)'s.

Proposition 1 states the existence and the construction of \( F_i^+ \) to FDO without calculating \( \gamma \)'s or \( \xi_j^{\min} \)'s. For a wide range of practical parameter settings we are interested in (e.g., Figure 2 subfigures), we notice that \( \gamma \) will guarantee the smallest \( \text{Var} \left[ \log Q_i \right] \). Proposition 2 shows that \( \text{Var} \left[ \gamma \right] \leq 0, \gamma \rightarrow \infty \), for each \( n \). However, the construction of \( F_i^+ \) involves calculating \( \gamma \) for each \( Q_i \) and \( \xi_j^{\min} \) for all \( f_j \in F^+ \), which makes the ordering algorithm slower than the actual variational transformation (so is GDO).

Approximately, Var \( \left[ \log Q_i \right] \propto \gamma \). Observe that, for fixed \( n \), choosing the \( n \) smallest \( \xi_j^{\min} E[1_{ij}] \)'s will guarantee the smallest \( \gamma \). We show the existence of \( F_i^+ \subset F^+ \) by the following construction: consecutively selecting the \( f_j^+ \)'s associated with the \( n \) smallest \( \xi_j^{\min} E[1_{ij}] \)'s.

Proposition 2. Fix \( p \in (0, +\infty), c \in (0, +\infty) \). Then for any \( j \in \{ j \mid \pi \left( f_j^+ \right) \neq \emptyset \} \), its variational parameter \( \xi_j^{\min} \) decreases monotonically on \( (0, +\infty) \) as \( E \left[ \pi \left( f_j^+ \right) \right] \) increases.

Proof. \( \xi_j^{\min} \) can be solved from either \( \arg \min_{\xi_j} P \left( f_j^+ \mid \xi_j, \pi(f_j^+) \right) \) or \( \arg \min_{\xi_j} P \left( f_j^+ \mid \xi_j \right) \). Since \( \xi_j^{\min} = \arg \min_{\xi_j} P \left( f_j^+ \mid \xi_j, \pi(f_j^+) \right) \) can be seen as the special case when \( p = 0 \), our argument below applies to both cases.

For fixed \( p, c \), we can solve for \( \xi_j^{\min} \) by letting \( \frac{\partial}{\partial \xi_j} \text{log} P(f_j^+ \mid \xi_j) = 0 \). We have \( E \left[ \pi \left( f_j^+ \right) \right] = \frac{1}{\xi_j^{\min}} \right) \left( pe^{c\xi_j^{\min}} - 1 + 1 \right) . \) Taking derivative of \( E \left[ \pi \left( f_j^+ \right) \right] \) w.r.t. \( \xi_j^{\min} \) gives:

\[
\frac{dE \left[ \pi \left( f_j^+ \right) \right]}{d\xi_j^{\min}} = -\frac{e^{c\xi_j^{\min}} + pe^{c\xi_j^{\min} \left( \xi_j^{\min} + 1 \right) + 1} \log(1 + \frac{1}{\xi_j^{\min}})}{e^{c\xi_j^{\min}} \xi_j^{\min} \left( \xi_j^{\min} + 1 \right)} < 0 \text{ for } \xi_j^{\min} > 0. \tag{9}
\]

Since the same strategy minimizes \( \text{Var} \left[ \log Q_i \right] \) for every \( d_i \in D_i \), it must be the most stable globally as well. Therefore, we arrive at an extremely simple variational transformation algorithm: sort \( f_i^+ \in F^+ \) by ascending rank of \( \pi \left( f_j^+ \right) \) and let that order be the order of variational transformation. We refer to this strategy as finding-degree order (FDO).

5 Related work

Exact inference on NOBN is fundamentally intractable (Cooper, 1990). Brute force inference on NOBN is \( O \left( |F| \cdot 2^{|D|} \right) \) as it calculates \( P(F) \) by summing up \( P(F \mid D') \cdot P(D') \), where \( D' \) can be
the combination of the presence or the absence of any subsets of $D$. Junction tree algorithms (Pearl 1988) can be more efficient in practice at $O(2^{|M|})$, where $|M|$ is the maximal clique size of the moralized network.

Quickscore (Heckerman 1990) reduces the temporal complexity to some exponential function of a quantity substantially smaller than $|D|$ or $|M|$ and make the inference practical for common usage. Quickscore (Heckerman 1990) achieves $\tilde{O}(|D| \cdot 2^{|F|})$ by exploiting marginal and conditional independence.

Table 2: Overall temporal complexities for exact and variational inferences on NOBN in terms of $|D|$, $|M|$, $|F|$, and $|F'|$ (note that all results are independent of $|S|$). In practical applications like QMR-DT, $|D| = 534$, $|M| \approx 151$, and $|F| \approx 43$. Jordan et al. 1999, Jaakkola and Jordan 1999.

|                | Brute force | Junction tree | Quickscore | Variational |
|----------------|-------------|---------------|------------|-------------|
| $O(|F| \cdot 2^{|D|})$ | $O(|D| \cdot 2^{|M|})$ | $\tilde{O}(|D| \cdot 2^{|F|})$ | $\tilde{O}(|D| \cdot [|F'| + 2^{|F|-|F'|}])$ |

Various approximate inference methods are proposed in place of Quickscore when processing expensive inference cases in NOBN (particularly QMR-DT). Variational inference for NOBN developed in (Jaakkola and Jordan 1999) reduces the cost in computing $P(F)$ by applying variational transformation to a subset of $F \subset F$. The variational evidence is incorporated as posterior probability when performing quickscore on the remaining findings. The running time is then $\tilde{O}(|D| \cdot [|F'| + 2^{|F|-|F'|}])$.

Other general approximation methods that can be applied to NOBN include loopy belief propagation (Murphy et al. 1999), mean field approximation (Ng and Jordan 2000), and importance sampling based sampling methods (Gogate and Domingos 2010). Some have also considered processing each finding in $F$ sequentially (Bellala et al. 2013), which is arguably more similar to the style of a realistic patient-to-doctor diagnosis.

6 Experiments

We evaluate the proposed inference algorithms on a real-world symptom-disease NOBN called F120. F120 is a QMR-like medical NOBN constructed from multiple reliable medical knowledge sources and is amended by medical experts. Unlike QMR-DT, F120 focuses on symptoms and diseases related to maternal and infant care. Due to the anonymous submission, the authors refrain from discussing F120’s details other than listing its vital statistics in Table 3.

Due to the unavailability of the proprietary QMR-DT network (Mansinghka et al. 2006), an anonymized version (aQMR) is available (Halpern and Sontag 2013). However, aQMR anonymizes the symptom and disease node names and randomizes QMR-DT’s $P(f^+ \mid d^+)$ probabilities. With the medical connotation removed, it is difficult to confidently generate user queries (a user query is a tuple $(F^+, F^-, d_i)$, where $d_i$ is the label disease: the most likely disease given the symptoms according to medical experts). Previous works working with aQMR do not face this issue since they do not require use-cases. For example, (Halpern and Sontag 2013, Jernite et al. 2013) focus on recovering the network structure and parameters; (Gogate and Domingos 2010) focuses on the inference time and the relative divergence between approximate inference outcome and the exact inference outcome.

We also evaluate the algorithm’s scalability on the artificially generated S1 that is much larger in scale than F120 and QMR-DT. S1 has 40,000 hidden disease nodes, which is approximately the total number of diseases in ICD-10 classification. Figure 5 compares various inference algorithms against the baseline in (Jaakkola and Jordan 1999) (JJ99+CVX). The proposed variational-first hybridization (VFH) is consistently faster than other methods. Despite having the same variational cost as VFH (shown in Table 4), Joint hybridization (JH) is the slowest due to its repeated negative evidence computation of Equation 7. JJ99+PPF is significantly faster than JJ99+CVX due to the simplified $\sum_{min}$ estimation.

Figure 4 compares the inference accuracies on F120. To simulate the wide-ranged inaccuracy in the disease priors $P(d^+)$’s, we scramble them with samples drawn from the uniform mean (Bates)

\[ the \text{soft-O bound is derived from } O\left(|D| \cdot |F^-| \cdot 2^{|F^+|}\right) \text{ given in (Heckerman 1990).} \]
VFH+CVX+FDO performs better than the JJ99+CVX+GDO baseline across the wide range of performance. (e, f, g, h) measure the corresponding top-3 accuracies. We introduce novel algorithms for variational hybridization.

In this work, we study the important problem of approximate inference on noisy-or Bayesian networks (specifically, their medical applications). We introduce novel algorithms for variational hybridization.
and variational transformation. The proposed algorithms greatly immunize the current variational inference algorithms against the inaccuracies in widely-ranged hidden prior probabilities, a common issue that arises in modern medical applications of Bayesian networks. In the future, we plan to investigate the applicability of the proposed algorithms to more general Bayesian networks.

References

D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. C. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu. Deep speech 2: End-to-end speech recognition in english and mandarin. CoRR, abs/1512.02595, 2015. URL http://arxiv.org/abs/1512.02595

G. Bellala, J. Stanley, S. K. Bhavnani, and C. Scott. A rank-based approach to active diagnosis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(9):2078–2090, Sept 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.30.

J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning bayesian networks from data: An information-theory based approach. Artificial Intelligence, 137(1 - 2):43 – 90, 2002. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/S0004-3702(02)00191-1. URL http://www.sciencedirect.com/science/article/pii/S0004370202001911

G. F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks (research note). Artif. Intell., 42(2-3):393–405, Mar. 1990. ISSN 0004-3702. doi: 10.1016/0004-3702(90)90060-D. URL http://dx.doi.org/10.1016/0004-3702(90)90060-D

A. L. Dawson and R. P. Dellavalle. Acne vulgaris. BMJ, 346, 2013. doi: 10.1136/bmj.f2634. URL http://www.bmj.com/content/346/bmj.f2634

V. Gogate and P. M. Domingos. Formula-based probabilistic inference. In P. Grunwald and P. Spirtes, editors, UAI, pages 210–219. AUAI Press, 2010.

Y. Halpern and D. Sontag. Unsupervised learning of noisy-or bayesian networks. In Proceedings of the 29th Conference, UAI, pages 272–281, 2013.

D. Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In UAI, pages 163–172, 1990.

T. S. Jaakkola and M. I. Jordan. Variational probabilistic inference and the qmr-dt network. Journal of artificial intelligence research, pages 291–322, 1999.

Y. Jernite, Y. Halpern, and D. Sontag. Discovering hidden variables in noisy-or networks using quartet tests. In Advances in Neural Information Processing Systems 26, pages 2355–2363, 2013.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Machine learning, 37(2):183–233, 1999.

W. Liao and Q. Ji. Learning bayesian network parameters under incomplete data with domain knowledge. Pattern Recognition, 42(11):3046 – 3056, 2009. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2009.04.006. URL http://www.sciencedirect.com/science/article/pii/S0031320309001472

T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. In ACL, 2015. URL http://arxiv.org/pdf/1410.8206v4.pdf

V. K. Mansinghka, C. Kemp, T. L. Griffiths, and J. B. Tenenbaum. Structured priors for structure learning. In Proceedings of the 22nd Conference, UAI, 2006.

B. Middleton, M. Shwe, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the internist-i/qmr knowledge base. Methods of information in medicine, 30:241–255, 1991.

K. Murphy. Bayes Net Toolbox. https://github.com/bayesnet/bnt 2002. [Online Lecture notes; accessed Jan-2016].

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In UAI, pages 467–475, 1999.

A. Y. Ng and M. I. Jordan. Approximate inference algorithms for two-layer bayesian networks. In Advances in Neural Information Processing Systems, pages 533–539, 2000.
J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

C. Riggelsen. Learning parameters of bayesian networks from incomplete data via importance sampling. *International Journal of Approximate Reasoning*, 42(1-2):69 – 83, 2006. ISSN 0888-613X. doi: http://dx.doi.org/10.1016/j.ijar.2005.10.005. URL: [http://www.sciencedirect.com/science/article/pii/S0888613X05000654](http://www.sciencedirect.com/science/article/pii/S0888613X05000654).

A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.