The Parameterized Complexity of Approximate Inference in Bayesian Networks

Johan Kwisthout

j.kwisthout@donders.ru.nl
http://www.dcc.ru.nl/johank/

Donders Center for Cognition
Department of Artificial Intelligence

PGM 2016
8th September 2016
- Exact Bayesian inference is known to be PP-complete [and thus NP-hard] in general (Littman et al, 1998)
- We know quite well when exact inference can be tractable:
  - Limit treewidth of the network and cardinality of the variables
- We know that approximating Bayesian inference is also intractable* in general
  * Standard complexity disclaimers: $P \neq NP, BPP \neq NP$
- **When is approximate Bayesian inference tractable?**
Agenda

- Introduction
- Fixed-error randomized tractability
- Parameters of interest
- Example results
- De-randomization result
- Conclusion
Problem definitions

**MProb**

**Input:** A Bayesian network $\mathcal{B}$ with designated subset of variables $H$ and a corresponding joint value assignment $h$ to $H$.

**Output:** $\Pr(h)$.

**CProb**

**Input:** A Bayesian network $\mathcal{B}$ with designated non-overlapping subsets of variables $H$ and $E$ and corresponding joint value assignments $h$ to $H$ and $e$ to $E$.

**Output:** $\Pr(h | e)$. 
Approximation

Additive approximation

**AA-MProb**

**Input:** As in MProb, in addition, error bound $\epsilon < 1/2$.

**Output:** $q(h)$ such that $\Pr(h) - \epsilon < q(h) < \Pr(h) + \epsilon$.

**AA-CProb**

**Input:** As in CProb, in addition, error bound $\epsilon < 1/2$.

**Output:** $q(h | e)$ such that $\Pr(h | e) - \epsilon < q(h | e) < \Pr(h | e) + \epsilon$. 
Approximation

Relative approximation

**RA-MProb**

*Input:* As in MProb, in addition, error bound $\epsilon$.

*Output:* $q(h)$ such that $\frac{\Pr(h)}{1+\epsilon} < q(h) < \Pr(h) \times (1 + \epsilon)$.

**RA-CProb**

*Input:* As in CProb, in addition, error bound $\epsilon$.

*Output:* $q(h \mid e)$ such that $\frac{\Pr(h \mid e)}{1+\epsilon} < q(h \mid e) < \Pr(h \mid e) \times (1 + \epsilon)$.
Approximation

- These approximation problems are deterministic: They are guaranteed to always give results within these bounds.

- Stochastic (randomized) approximations: Add confidence bound $0 < \delta < 1$.

- Approximate result is within error bounds $\epsilon$ with probability at least $\delta$.

- If $\delta$ is polynomially bounded away from $1/2$, then the probability of answering correctly can be boosted arbitrarily close to 1 while still requiring only polynomial time.
PP and BPP

- PP and BPP are defined as classes of decision problems decidable by a **probabilistic Turing machine**

- PP: Yes-instances are accepted with probability \( \delta = \frac{1}{2} + \frac{1}{c^n} \)

- BPP: Yes-instances are accepted with probability polynomially bounded away from \( \frac{1}{2} \) (i.e., \( \delta = \frac{1}{2} + \frac{1}{n^c} \))

- We know that NP \( \subseteq \) PP yet assume that BPP = P

- If \( \Pi \in \text{BPP} \) than \( \Pi \) has an efficient randomized (Monte Carlo) algorithm and \( \delta \) can be boosted close to 1 in poly time
• A (possibly NP-hard) problem Π is called **fixed-parameter tractable** for parameter κ if it can be solved in time, exponential *only* in κ and polynomial in the input size n, that is, in \(O(f(\kappa) \cdot n^c)\)

• In practice, this means that instances of Π can be solved **efficiently**, even when the problem is NP-hard in general, if κ is known to be small

• if Π remains NP-hard for all but finitely many values of the parameter κ, then \(\{\kappa\}\)-Π is para-NP-hard: bounding κ **does not render** the problem tractable
• Analogously, the class FERT (Kwisthout, 2015) characterizes problems $\{\kappa\}$-\Pi that can be **efficiently** computed with a randomized algorithm if $\kappa$ is bounded.

• Formally, $\{\kappa\}$-\Pi $\in$ FERT if there is a probabilistic Turing machine that accepts Yes-instances of \Pi with probability $1/2 + \min(f(\kappa), 1/n^c)$ for a constant $c$ and $f : \mathbb{R} \rightarrow [0, 1/2]$.

• If \Pi remains PP-hard even for bounded $\kappa$, then $\{\kappa\}$-\Pi is para-PP-hard.

• Corollary: If \Pi is NP-hard for bounded $\kappa$, then $\{\kappa\}$-\Pi $\notin$ FERT.
Overview
Introduction
Fixed-error randomized tractability
Parameterized results
De-randomization result
Conclusion

Inclusion diagram
## Parameters of interest

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_e$</td>
<td>dependence value of the evidence</td>
</tr>
<tr>
<td>$B$</td>
<td>local variance bound of the network</td>
</tr>
<tr>
<td>$P_h$</td>
<td>posterior probability</td>
</tr>
<tr>
<td>$</td>
<td>E</td>
</tr>
<tr>
<td>$P_e$</td>
<td>prior probability of the evidence</td>
</tr>
<tr>
<td>$d$</td>
<td>maximum indegree of the network</td>
</tr>
<tr>
<td>$l$</td>
<td>maximum path length of the network</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>absolute or relative error</td>
</tr>
</tbody>
</table>
Parameters of interest

- The *dependence value of the evidence* \( (D_e) \) is a measure of the cumulative strength of the dependencies in the network, given a particular joint value assignment to the evidence variables.

- We can define for any node \( X_i \) its dependence strength (given evidence \( e \)) \( \lambda_i \) as \( u_i/l_i \), where \( l_i \) and \( u_i \) are the greatest, resp. smallest numbers such that

\[
\forall x_i \in \Omega(X_i) \forall p \in \Omega(\pi(X_i)) l_i \leq \Pr(X_i = x_i \mid p, e) \leq u_i
\]

- The dependence value of the network (given evidence \( e \)) is then given as \( D_e = \prod_i \lambda_i \)
Parameters of interest

- The *local variance bound* \((B)\) of a network is a measure of the representational expressiveness and the complexity of inference of the network.
- It is defined similarly as the dependence value, but is not conditioned on the evidence, and a maximization rather than a product of the local values is computed.
- Let \(\lambda_i = u_i/l_i\), where \(l_i\) and \(u_i\) are the greatest, resp. smallest numbers such that
  \[
  \forall x_j \in \Omega(x_i) \forall p \in \Omega(\pi(x_i)) l_i \leq \Pr(X_i = x_i \mid p) \leq u_i
  \]
- The local variance bound of the network is then given as \(B = \max_i(\lambda_i)\).
Example results (1/3)

- For any fixed $\epsilon < \frac{1}{2^{n+1}}$ absolute approximate inference is PP-hard (rounding to the nearest $\frac{1}{2^n}$ solves $\text{MajSat}$)
- For $\epsilon \geq \frac{1}{n^c}$ we can efficiently approximate the marginal inference problem absolutely by simple forward sampling
- As the number of samples needed to approximate $\text{aa-MProb}$ depends (only) on $\epsilon$, this yields the following results:

**Corollary (Kwisthout, 2009)**

$\{d\}$-$\text{aa-MProb}$ *is para-PP-hard*.

**Result (Henrion, 1986)**

$\{\epsilon\}$-$\text{aa-MProb} \in \text{FERT}$.
Example results (2/3)

• For relative approximation, however, it is impossible to approximate \( \text{ra-MProb} \) in polynomial time with \( \text{any} \) bound \( \epsilon \).

• This would directly indicate whether \( \Pr(h) = 0 \) or not, and hence by a corollary of Cooper’s (1990) result, would decide \text{Satisfiability}. We thus have that:

\begin{center}
\textbf{Result (Cooper, 1990)}
\end{center}

\[ \{\epsilon\}\text{-ra-MProb is para-NP-hard}. \]

\begin{center}
\textbf{Corollary}
\end{center}

\[ \{\epsilon\}\text{-ra-MProb} \not\in \text{FERT}. \]
• More interesting to consider computing posterior probabilities conditioned on evidence
• Dagum & Luby (1993) showed that there cannot be any polynomial time absolute approximation algorithm for any $\epsilon < \frac{1}{2}$ because such an algorithm would decide $3\text{Sat}$
• Their proof uses a singleton evidence variable ($|E| = 1$) and at most three incoming arcs per variable ($d = 3$)

Result (Dagum & Luby, 1993)

$\{\epsilon, d, |E|\}$-AA-CPROB is para-NP-hard.

Corollary

$\{\epsilon, d, |E|\}$-AA-CPROB $\not\in$ FERT.
More interesting results

Result (Dagum & Chavez, 1993)
\{D_e\}\text{-AA-CProb} is para-PP-hard.

Result (Dagum & Chavez, 1993)
\{D_e, \epsilon\}\text{-RA-CProb} \in \text{FERT}.

Result (follows from Park & Darwiche, 2004)
\{B\}\text{-RA-CProb} \notin \text{FERT} for any \epsilon < 1.

Result (Dagum & Luby, 1997)
\{B, |E|, \epsilon\}\text{-RA-CProb} \in \text{FERT}. 
De-randomization

Result (Henrion, 1986)
\{\epsilon\}\text{-aa-MProb} \in \text{FERT}.

- We can de-randomize this stochastic algorithm by simulating all possible sequences of random coin flips and taking a majority decision.
- If there are \(O(n)\) such bits, this will yield an exponential-time algorithm (like exact computation).
- Do we actually need so many random bits?
De-randomization

- If $O(\log n)$ *distinct* random bits would suffice we could effectively simulate the randomized algorithm in poly time.
- $O(\log n)$ random bits can be amplified in polynomial time to a $k$-wise independent $O(n)$-bit random string for every fixed $k$ (Luby, 1988).
- Do all the random bits need to be fully independent of each other?
- **No!** For every variable, the random bits should be independent of the random bits needed to select its parents.
- If the number of parents is bounded (i.e., bounded in-degree $d$) then $d$-wise independence suffices.
FPT result for approximate inference

Result (Henrion, 1986)

\( \{\epsilon\}-\text{aa-MProb} \in \text{FERT} \).

Result

\( \{d, \epsilon\}-\text{aa-MProb} \in \text{FPT} \).

Theoretical question

Whether \( \text{BPP} \equiv \text{P} \) is a crucial open question in theoretical computer science; its parameterized analog would be whether \( \text{FERT} \equiv \text{FPT} \). Here, we need to bound \( d \) as well as \( \epsilon \): We have an efficient randomized algorithm for every \( d \) but no known deterministic one.
In this paper we investigated the parameterized complexity of approximate Bayesian inference, both by re-interpretation of known results in the formal framework of fixed-error randomized complexity theory, and by adding a few new results. This gives an insight in what constraints can, or cannot, render approximation tractable. These constraints are notably different from the traditional constraints (treewidth and cardinality) that are necessary and sufficient to render exact computation tractable. In future work we wish to extend this line of research to other problems in Bayesian networks, notably the MAP problem.