Intractable Bayesian Models and Approximation

"Approximation is like an ordinary medicine:
by itself, it's neither placebo nor panacea;
to be of use, you need to *prove* that it works!"

Johan Kwisthout, Radboud University Nijmegen

"If I have seen further than others, it is by standing on the shoulders of giants" – Sir Isaac Newton

Joint work with Iris van Rooij (Radboud University Nijmegen)

Previously presented (condensed presentation) at ICCM'12

Outline

- Bayesian models of cognition
- Computational Intractability
  - Intermission: Socrates and Cebes (I)
- On *optimists*, *pessimists*, and *realists*
  - Intermission: Socrates and Cebes (II)
- Approximation – placebo vs. panacea
- A cook-book recipe for provable tractability
- Case study: most simple explanations
  - Intermission: Socrates and Cebes (III)
- Further work and conclusions

Cognitive modeling

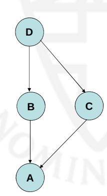
- Computational models of cognition
  - Understand how the mind works
  - Predict human behavior (HCI)
  - Artificial intelligence / robotics / BCI
- Marr's hierarchy of analysis:
  - Computational level (what)
  - Algorithmic level (how)
  - Implementational level (realisation)
- Study, experiment, simulate, predict

Bayesian models of cognition

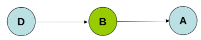
- Many computational models nowadays are based on **Bayesian abduction**
- Bayesian abduction = inferring the *most probable explanation* of a set of observed phenomena
  - What are this person's intentions given what I observe as his actions?
  - What does she want to communicate here?
  - What is the object that is partially occluded in my line of vision?

Bayesian networks

- **Bayesian network**: models a set of stochastic variables and the independency relations among them
- Directed acyclic graph with nodes and arrows; probability distribution for every node
- **A** (directly) depends on **B** and **C**
- The probability distribution of A is conditioned on the values of **B** and **C**
- **B** and **C** (directly) depend on **D**
- Other dependencies between variables depend on observations in the network

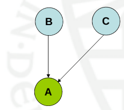


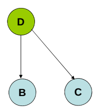
Conditional dependencies



- If **B** is observed, **A** is independent of **D** as there is no direct link: **A**'s probability distribution is governed by **B** only

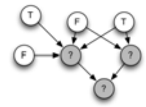
- If **A** is observed, **B** and **C** become dependent on each other as information on **B** 'explains away' **C** vice versa

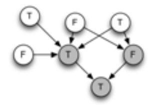




- If **D** is observed, **B** and **C** become independent from each other as **D** is a common cause of **B** and **C**

Bayesian abduction in a nutshell





- **Input**: A Bayesian network partitioned in two sets **H** and **E**, and an observation **e** for the variables in **E**
- **Output**: The most probable joint value assignment **h** to **H** with **E = e**, or *argmax*$_h$ Pr(**H = h**, **E = e**)
- Bayesian abduction or *Inference to the Most Probable Explanation* happens to be computationally intractable (*NP*-hard) in general

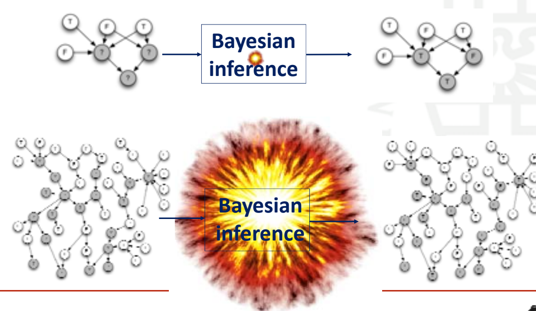
Why does it matter?

- Recall: *NP*-hard means: no polynomial worst-case algorithm possible unless *P* = *NP*

N	**10**	**50**	**100**	**300**	**1000**
$\log_2 N$	3	5	6	8	9
$5 \cdot N$	50	250	500	1,500	5,000
$N \cdot \log_2 N$	33	282	665	2,469	9,966
N^2	100	250	10,000	90,000	1,000,000
N^3	1,000	125,000	1,000,000	$2.7 \cdot 10^7$	$1.0 \cdot 10^9$
2^N	1,024	$1.1 \cdot 10^{15}$	$1.3 \cdot 10^{30}$	$2.0 \cdot 10^{90}$	$1.0 \cdot 10^{301}$
$N!$	3,628,800	$3.0 \cdot 10^{64}$	$9.3 \cdot 10^{157}$	$3.1 \cdot 10^{614}$	$4.0 \cdot 10^{2567}$

- Not polynomial = intractable for all but very small inputs

Bayesian Inference is NP-Hard



So what?

- *NP*-hardness means that in *general* there cannot exist a polynomial-time algorithm for solving *arbitrary* instances of Bayesian abduction (proven by reduction from SAT)
- Consequently, there are instances ***that are valid model instances*** that cannot be computed in polynomial time – and thus, the validity of the computational model of the cognitive task is at stake
- "Hey, my model does not encode SAT formulas and the like, that is not a real world problem!"

Begging the question

- ***Socrates:*** "Your model assumes *NP*-hard computations!"
- ***Cebes:*** "*NP*-hardness doesn't say anything. Of course there are instances that my model doesn't compute in polynomial time. But these are unrealistic instances. My model does well on reasonable instances"
- ***Socrates:*** "Fine. Which are those reasonable instances?"
- ***Cebes:*** "Well, those instances that my model computes in polynomial time, of course!"

Why care about tractability in cognitive modeling at all?

- If we make computational models of cognitive processes, we want the properties of these models to reflect reality – if it works in practice (in the brain), it should work in theory (in the model)
- Obviously, the brain cannot perform *unconstrained* NP-hard computations – so our models need to be constrained as well in a plausible way

One could be too hard on oneself…

- ***Socrates:*** "Your model assumes *NP*-hard computations!"
- ***Cebes:*** "O dear! Then my model apparently is fatally flawed! Bayesian models cannot explain human cognition!"

… or not hard enough…

- ***Socrates:*** "Your model assumes *NP*-hard computations!"
- ***Cebes:*** "Well, never mind, let's assume that the brain does not compute exactly, but just *approximates* Bayesian computations!"

Now what?

- As Socrates pointed out, we cannot just ignore NP-hardness – this issue needs to be addressed (remember our goal was to **understand** how the mind works!)
- **"Now what?" – three ways of dealing with intractability**
  - The *doomsday* approach
  - The *hand-waving* approach
  - The *analytical* or *rational* approach

The doomsday approach

Bayesian abduction is *NP*-hard

Bayesian models are no good models of the brain

- The **pessimists** throw away the baby with the bath-water: because Bayesian abduction is *NP*-hard, that doesn't rule out that *many* instances of abduction problems can be solved tractably

The hand-waving approach

Bayesian abduction is *NP*-hard

OK, fine; we'll just assume that the mind *approximates* Bayesian abduction then…

- The **optimists** try to solve the problem by asserting that approximation, satisficing, and using heuristics will be sufficient to overcome intractability. However, approximating Bayesian abduction and satisficing is *NP*-hard as well (Kwisthout, 2011; Kwisthout, van Rooij and Wareham, 2011)!

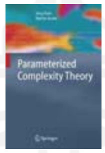
The analytical approach

Bayesian abduction is NP-hard

How can we *constrain* our model such that abduction becomes tractable again?

- The **realists** see the strength of Bayesian models but acknowledge that they are too broad and need to be constrained in order to overcome intractability. They will look for ***problem parameters*** that – when bounded – render the problem tractable

Parameterized complexity theory



- Even when a problem Π is NP-hard in general, it may be the case that there exist particular problem parameters, such that the problem can be solved tractably if the parameter is low.
- Formally, a problem with input size n may have parameters $k_1, k_2, ..., k_n$ and an algorithm solving the problem in time $O(f(k_1,k_2,...,k_n) \cdot n^c)$ for an arbitrary computable function f and a constant c
- Hence, when $k_1, k_2, ..., k_n$ are small (enough), the running time of the algorithm is dominated by the $O(n^c)$ factor

~~Begging~~ Answering the question

- ***Socrates:*** "Your model assumes NP-hard computation!"
- ***Cebes:*** "NP-hardness doesn't say anything. Of course there are instances that my model doesn't compute in polynomial time. But these are unrealistic instances. My model does well on reasonable instances"
- ***Socrates:*** "Fine. Which are those reasonable instances?"
- ***Cebes:*** "Well, those instances in which parameters $k_1, k_2, ..., k_n$ are small!"

Parameterized complexity analysis of MPE

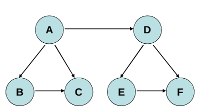
- Parameters that – when small – render Bayesian abduction tractable:
  - One minus the probability of the most probable explanation (i.e., when the probability of the MPE is high)
  - The *treewidth* of the network *and* the number of possible values per variable (both need to be small)
- Parameters that – even when small – do *not* render Bayesian abduction tractable:
  - The degree of the network, i.e., the number of incoming/outgoing arcs
  - The number of possible values per variable alone
- Other parameters are yet undecided
  - Treewidth alone, range of the probability distribution, …

Treewidth

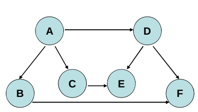
- The *treewidth* of a graph is a theoretical concept that loosely correlates to a measure on the *localness of the connections* in the graph
- If connections tend to be clustered in small sub-networks, with few connections between them, treewidth often is low
- If connections are scattered all over the place, treewidth may be high
- Many NP-hard graph problems are tractable when the treewidth of the graph is small

Hans Bodlaender

Two examples



- Two distinct clusters with only one connection
- Treewidth happens to be 2



- No distinct clusters, connections all over the place
- Treewidth happens to be 4
- Intuitive idea: computations are easier when they are localized

Bayesian approximation: placebo nor panacea

- We can't simply **assume** tractability due to approximation
  - In general, approximation is *as intractable* as exact computation
- We ***also*** can't assume that approximation plays no role
  - There are parameters in Bayesian computations that, when constrained, render an approximate problem tractable while leaving the exactly computed problem intractable
- Bottom line: approximation is like a regular medicine: it *might* work, but we must **prove** that it works under well-defined conditions
- We contributed a ***cook-book recipe*** for such proofs

What is approximation?

- "The sun is **roughly** 100 times as large as the moon"
- "There is a mountain range on Mars that **looks roughly like** a human face"
- "**Roughly speaking**, height determines shoe size"
- Three ways of approximating Bayesian Inference
  - Value-approximation (almost as likely as..)
  - Structure-approximation (very similar to..)
  - Expectation-approximation (very likely to be..)

  .. the most probable explanation
- Still other notions of approximation may exist

A cook-book recipe

- To be valid, claims of "tractible due to approximation" should be supported by:

a) A precise **definition** of "approximation"

b) In case the formal approximation problem is *NP*-hard (as well), a set of **problem parameters** that are believed to be constrained

c) A **formal proof** that the thus constrained approximation problem becomes tractable

d) Arguments or (empirical) evidence **supporting** the assumptions in b)

Case study: Most Simple Explanation

Most Simple Explanation

- Intuitive idea: in real life, only few variables are relevant in determining the most probable explanation.
- Instead of computing full posterior distributions, we seek to find the explanation that is most probable in the majority of possible worlds
- [Proposed by K. (2010)]

Case study: Most Simple Explanation

- MSE gives an intuitive explanation of how we solve abduction problems
- Solving the MSE problem is, like many problems in Bayesian networks, *NP*-hard in general
- However, under particular constraints, MSE can serve as a adequate **approximate model of abduction** in real-world situations
- We will illustrate this with the cook-book recipe shown earlier

A cook-book recipe

- To be valid, claims of "tractible due to approximation" should be supported by:

a) A precise **definition** of "approximation"

b) In case the formal approximation problem is *NP*-hard (as well), a set of **problem parameters** that are believed to be constrained

c) A **formal proof** that the thus constrained approximation problem becomes tractable

d) Arguments or (empirical) evidence **supporting** the assumptions in b)

MSE case study – step a)

- *To instantiate any claim of tractable approximation, it is required to give an explicit and precise definition of the approximation that is used*
- We use the following ***expectation-approximation***: we take N random samples and decide on the most probable explanation in the majority of worlds
- This allows for a particular **error**
- To decrease this error to reasonable proportions, we may need an exponential number of samples

MSE case study – step b)

- *Whenever this approximation is intractable in general, define problem parameters that are hypothesized to be constrained in the real world*
- Among many possible **parameters**, we choose:
  1. The treewidth (measure on localness of the connections) of the Bayesian network
  2. The cardinality of the variables in the network
  3. The skewedness of the probability distribution (probability that two random possible worlds yield different most probable explanations

MSE case study – step c)

- *Give a formal proof that expectation-approximation becomes tractable when the values of these parameters are constrained*
- We give a **sampling algorithm** [see the paper] that can compute MSE in polynomial time, with only a small (fixed) possibility of error, if the treewidth and cardinality are low and the probability distribution is biased towards a particular explanation

MSE case study – step d)

- *Provide support for the hypothesized constraints in b) with arguments or empirical findings*
- From the machine learning literature it is known that bounded treewidth prevents ***overfitting***. There is reason to believe that human knowledge structures try to avoid overfitting as well (to be able to adapt to new situations)
- **Few samples suffice** to make decisions almost as good as decisions based on full Bayesian computations, if the probability distribution is biased towards a particular explanation [Vul et al, 2009, "One and done."]

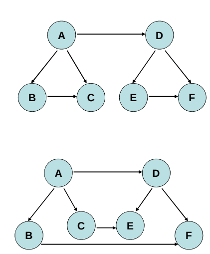
Socrates revisited

- ***Socrates:*** "Which are those reasonable instances?"
- ***Cebes:*** "Well, those instances in which parameters k_1, k_2, ..., k_n are small"
- ***Socrates***: "Ah, but are they small in practice?"
- ***Cebes:*** "I don't know, but let's ask a cognitive scientist to see whether she thinks that it is plausible that k_1, k_2, ..., k_n are typically small in cases where humans perform the cognitive task easily"

Socrates revisited

- ***Cebes:*** "Dear cognitive scientist, do you think that k_1, k_2, ..., k_n are typically small in cases where humans perform these cognitive tasks easily?"
- ***Cognitive Scientist***: "Hmm, well, I'm pretty sure that k_1, k_2, ..., k_{n-1} are, but I'm not sure about k_n really..."
- ***Socrates***: "So, Cebes, how could you verify whether k_n is indeed small in practice and thus that your model is a good description of reality?"
- ***Cebes:*** "Well, er, … let's design an experimental setting with two comparable scenarios in which a cognitive task is measured, *that differs only in* k_n, and measure reaction times and error rates. If my model is right, performance will lower significantly when k_n increases!"

Possible setup for an experiment



- These networks differ *only* in their treewidth!
- Can we **design experiments** that employ, e.g., scenarios in which the knowledge is structured according to these networks?
- If so, since treewidth is the only variable that is manipulated, indeed treewidth is a **source of complexity** in the model

Conclusion

- Despite intractability in general, Bayesian abduction is still a very **useful framework** for computational cognitive models, but we need to **constrain the input** to make it tractable
- Approximation is **neither panacea nor placebo** – it *may* help to render the model tractable in some cases, but we need formal proofs!
- We provided a **cookbook recipe** for studying the interplay between approximation and constraints on input domains

References

1. Mark Blokpoel, Johan Kwisthout, and Iris van Rooij (2012). When can predictive brains be truly Bayesian? *Frontiers in Theoretical and Philosophical Psychology*,3 (406).
2. Johan Kwisthout (2011). *Most Probable Explanations in Bayesian Networks: Complexity and Tractability.* International Journal of Approximate Reasoning, 52 (9), 1452 - 1469.
3. Mark Blokpoel, Johan Kwisthout, Todd Wareham, Pim Haselager, Ivan Toni, and Iris van Rooij (2011). *The computational costs of recipient design and intention recognition in communication.* CogSci 2011.
4. Johan Kwisthout, Todd Wareham, and Iris van Rooij (2011). *Bayesian Intractability is not an Ailment that Approximation can Cure*. Cognitive Science, 35 (5), 779 - 784.
5. Johan Kwisthout and Iris van Rooij (2012). *Bridging the Gap between Theory and Practice of Approximate Bayesian Inference*. International Conference on Cognitive Modeling 2012.