Parameterized Hardness of Active Inference

Nils Donselaar N.DONSELAAR@DONDERS.RU.NL
Radboud University Nijmegen
Donders Institute for Brain, Cognition and Behaviour
Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Abstract

Within the field of cognitive neuroscience, predictive processing is an increasingly popular unifying account of cognitive capacities including action and perception which posits that these rely on probabilistic generative models to predict sensory input or the consequences of one's behaviour. In the corresponding literature one frequently encounters naive claims about the computational resources required to successfully employ such models, while actual complexity analyses are often lacking. In this paper we study the problem of selecting a course of action which yields the greatest reduction in prediction error between the intended outcome and the current state, known in this area as *active inference*. Modelling the problem in terms of Bayesian networks and the relative entropy (Kullback-Leibler divergence) between a target and an induced distribution, we derive parameterized (in)tractability results extending the $\mathsf{NP}^{\mathsf{PP}}$-hardness classification found in Kwisthout (2014). These show that contrary to common belief, the size of the prediction error does not determine whether active inference is tractable, not even when the number of actions and outcomes to be considered is restricted. Moreover, this conclusion appears to extend even to an approximate version of the problem. We believe these results can be of interest to both cognitive scientists seeking to evaluate the plausibility of their explanatory theories, and to researchers working on probabilistic models, as they relate to existing work on the hardness of observation selection in decision making.

Keywords: predictive processing; active inference; prediction error; relative entropy; Bayesian networks; parameterized complexity theory.

1. Preliminaries

Motivated in part by empirical findings and in part by the increasing number of successful implementations based on similar principles, the theory of *predictive processing* has in the course of the last decade received increasing attention both in the fields of cognitive psychology and neuroscience as well as in philosophy of cognitive science. At its core, predictive processing postulates that the cognitive systems governing action, perception and even higher-order beliefs are characterized by four shared traits: they operate on probabilistic information such as likelihood estimations in a mathematically valid way; this information is used to maintain a generative model which is employed to continuously predict sensory input in the immediate future; these models work to minimize the discrepancy between observation and prediction which is called *prediction error*; finally, these models are stratified into multiple hierarchies, possibly instantiated in different cortical areas.[1]

One problem which has been of specific interest in the context of predictive processing, particularly in the field of cognitive neuroscience, is that of *active inference*. This term refers to the problem of selecting a course of action which yields the greatest reduction in prediction error between some

[1] An illustration of this principle would be the visual information processing system in the brain, in which one distinguishes between several visual areas, believed to correspond to increasing levels of abstraction.

distribution representing a desirable outcome and the predicted distribution over those same events. As in Friston et al. (2015), this prediction error is typically construed in terms of the relative entropy between the two distributions, i.e. as the *Kullback-Leibler divergence* of the preferred distribution with respect to the posterior distribution associated with the selected course of action.

While one may suspect that active inference is a computationally demanding task, proving such a statement requires a formalization of the framework of predictive processing in which the problem is cast. One such formalization which is becoming more widely endorsed is in terms of Bayesian networks, for which Kwisthout (2014) demonstrates that the problem of active inference is $\mathsf{NP}^{\mathsf{PP}}$-hard. Incidentally, this is also the hardness of a closely related problem which has been the object of research on probabilistic models, namely that of observation selection or evidence gathering. Here one asks for which nodes of the network one can best perform a test to determine their current state, based on their expected value in terms of a score function which measures how informative it is to learn that a node is in a particular state. In Krause and Guestrin (2009) it is shown that in general this problem is $\mathsf{NP}^{\mathsf{PP}}$-hard as well, further highlighting the similarity between the two problems.[2]

Nevertheless, it has been suggested in the literature (see e.g. Clark (2013), pp. 25 and 31 in particular) that the problem of active inference (or more likely an approximate version thereof) is tractable when the size of the prediction error is suitably restricted. In order to assess this claim, we extend the existing analysis into the domain of parameterized complexity theory, which provides us a formal means of determining how specific aspects of the problem contribute to its overall hardness. In particular we consider multiple ways of incorporating the prediction error as a parameter, besides other parameters known or expected to be relevant in this context, such as the number of actions to choose from or even the error bound on an approximate answer.

In the next section we introduce the relevant parts of probabilistic and parameterized complexity theory with which the reader may be unfamiliar. In Section 3 we fix notational conventions and explain the reduction strategy which drives all of the hardness results in Sections 4 and 5, which cover exact and approximate versions of active inference respectively. We then discuss in the final section how these results are relevant to the discussion surrounding active inference, as well as their place in the theoretical research on Bayesian networks and other probabilistic models.

2. Computational Complexity

While we assume the reader to be familiar with the basics of classical complexity theory, we shall briefly cover the probabilistic complexity classes BPP and PP, along with the central notions from parameterized complexity theory insofar they are required to state our results.

In what follows, by a *probabilistic algorithm* we mean an algorithm which has access to random bits, i.e. bits whose values are determined according to independent uniform probability distributions. The intended technical formalization is in terms of Turing machines whose transition functions are stochastic with these same properties.

[2] However, note that the two problems are different in that for active inference the outcome of the "observation" being made is guaranteed to be the selected action, hence it does not consider expectations over nodes but rather the value of particular outcomes (though technically KL divergence is itself an expectation of relative informational content).

Definition 1 *BPP is the class of decision problems solvable in time polynomial in the size of the input by a probabilistic algorithm which gives the correct answer with probability more than $\frac{2}{3}$.*

The class BPP is commonly thought of as the class of problems efficiently solvable by a probabilistic algorithm. While its relation to NP is not known (i.e. if BPP is included in NP or vice versa), there are grounds for believing that in fact $\mathsf{P} = \mathsf{BPP}$ such as presented in Impagliazzo and Wigderson (1997), further supporting the view that membership of BPP indicates (randomized) tractability. Another important probabilistic complexity class is defined as follows:

Definition 2 *PP is the class of decision problems solvable in time polynomial in the size of the input by a probabilistic algorithm which gives the correct answer with probability more than $\frac{1}{2}$.*

Although the two classes may seem similar, not only does PP contain NP, the class P^{PP} moreover contains the entire polynomial hierarchy, as shown in Toda (1991).[3] This informs us that the PP-hardness or even $\mathsf{NP}^{\mathsf{PP}}$-hardness of a problem is as strong of an indication (and likely stronger) as to the intractability of a problem as is the notion of NP-hardness. We typically establish hardness of one of the former two kinds by giving reductions from the following problems.

MAJORITY SATISFIABILITY (MAJSAT)
Input: A propositional formula $\varphi(x_1, \ldots, x_n)$ in CNF, consisting of clauses $\{c_1, \ldots, c_m\}$.
Question: Is φ satisfied by more than half of the possible truth assignments to its variables $(x_1, \ldots, x_n)$?

EXISTENTIAL MAJORITY SATISFIABILITY (E-MAJSAT)
Input: A propositional formula $\varphi(x_1, \ldots, x_n)$ in CNF, consisting of clauses $\{c_1, \ldots, c_m\}$, along with an integer $1 \leq k \leq n$. (The value $k = 0$ would yield a degenerate case equivalent to MAJSAT.)
Question: Is there a truth assignment $(a_1, \ldots, a_k)$ such that the formula $\varphi(a_1, \ldots, a_k, x_{k+1}, \ldots, x_n)$ is satisfied by more than half of the possible truth assignments to its variables $(x_{k+1}, \ldots, x_n)$?

Proposition 3 MAJSAT *is* PP*-complete, and* E-MAJSAT *is* $\mathsf{NP}^{\mathsf{PP}}$*-complete.*[4]

In the next section we shall discuss how we can construct reductions from these problems to the ones which we will be considering, namely by drawing on methods previously used to obtain hardness results for other problems related to Bayesian networks. First, we provide a short introduction to *parameterized complexity theory*, a field extending classical complexity theory which includes Downey and Fellows (1999) amongst its foundational works.

Though largely similar in approach, we now deal with *parameterized* decision problems, which consist of pairs (x, k) where x is the input as usual and k is some parameter. While parameters are typicially natural numbers, these can encode other countable sets such as the rationals, hence those are also permissable. The counterpart to P is now the class FPT (for "fixed-parameter tractable").

Definition 4 *FPT is the class of parameterized decision problems for which there exists an algorithm that runs in time at most $f(k)x^c$, where f is a computable function in k and c is a constant.*

[3] The superscript class is an *oracle*, e.g. P^{PP} is the class of problems solvable by a deterministic Turing machine in polynomial time where moreover the machine can decide in a single step whether some input belongs to a given problem in PP. In turn, the polynomial hierarchy can be thought of as the limit of successively larger towers of NP.

[4] Proofs for these two results can be found in Gill (1977) and Wagner (1986) respectively.

As one can tell from the definition above, parameterized complexity is based on trying to identify which factors contribute to the superpolynomial time a problem is believed to require in order to solve it. One example is the NP-complete problem VERTEX COVER: when parameterizing with k the maximum size of the vertex cover, we find that k-VERTEX COVER is in FPT. We can extend this approach to other complexity classes as well, in a way explored in Flum and Grohe (2003).

Definition 5 *Let C be a complexity class and* D *be any decision problem. Then* k-D *is in paraC for some parameter k on the same conditions of* D *being in C, except the upper bound on the time or space used (whichever measure is used to characterize* D*) may now include a factor $f(k)$.*

For example, the classes paraBPP and paraPP require the existence of a probabilistic algorithm deciding the problem with probability more than $\frac{2}{3}$ and $\frac{1}{2}$ respectively, except the time taken may be of the form $f(k)x^c$ as for the class FPT. Observe that $\mathsf{paraC} = \mathsf{paraC}'$ precisely when $\mathsf{C} = \mathsf{C}'$, hence it makes sense to consider the hardness of parameterized decision problems as we do in the classical case. This brings us to the final notion to discuss: that of a parameterized reduction.

Definition 6 *An fpt-reduction from a problem* k-D *to a problem* k'-D$'$ *is a function which maps* (x, k) *to* (x', k') *in time* $f(k)|x|^c$ *such that* $(x, k) \in k\text{-D}$ *if and only if* $(x', k') \in k'\text{-D}'$*, with there being a computable function* $g(k)$ *such that* $k' \leq g(k)$*.*

We conclude by noting that such fpt-reductions behave like polynomial time (Karp) reductions in all important respects, such as closure under composition. This ensures that the definition of a parameterized problem being hard for a parameterized complexity class whenever all problems in this class fpt-reduce to the problem carries the same meaning as before.

3. Constructing the Hardness Proofs

All of the hardness results in this paper are established using the same underlying construction, which extends the usual way of encoding a propositional formula into a (discrete) Bayesian network in polynomial time. To streamline the presentation of the proofs, we therefore discuss this construction separately, allowing us to cover the notational conventions employed at the same time.

By a Bayesian network $\mathcal{B}$ we mean a connected DAG $\mathbf{G} = (\mathbf{V}, \mathbf{A})$ along with a collection of probability distributions $\mathbf{Pr}(X \mid \pi)$ for every $X \in \mathbf{V}$ and every configuration $\pi \in \Omega(\rho(X))$ of the parents $\rho(X)$ of X. With any proposition formula φ we can associate a Bayesian network $\mathcal{B}_\varphi$ in the way introduced in Cooper (1990) and expanded on in Park and Darwiche (2004) and Kwisthout (2009). The network $\mathcal{B}_\varphi$ has three essential kinds of binary nodes: uniformly distributed nodes $X_1, \ldots, X_n$ representing the variables $x_1, \ldots, x_n$, followed by nodes $C_1, \ldots, C_m$ for the clauses $c_1, \ldots, c_m$ (with X_i leading to C_j precisely when x_i or $\neg x_i$ occurs in c_j) which are *True* precisely when they would be satisfied by the assignment to the variables, and a terminal node V_φ representing their conjunction in a similar fashion. This construction ensures that $\text{Pr}(\top) = \frac{\#\varphi}{2^n}$ where $\text{Pr}(\top)$ is short for $\text{Pr}(V_\varphi = \text{True})$ and $\#\varphi$ is the number of satisfying assignments to the variables of φ.

We extend this network $\mathcal{B}_\varphi$ to $\mathcal{B}^*_\varphi$ by including a binary node Y with possible values y and $\neg y$, and another binary node Z with parents V_φ and Y which takes values z and $\neg z$ (see Figure 1). The probability distributions we associate to Y and Z vary across the different reductions, hence these details will be separately provided in each of the individual proofs.

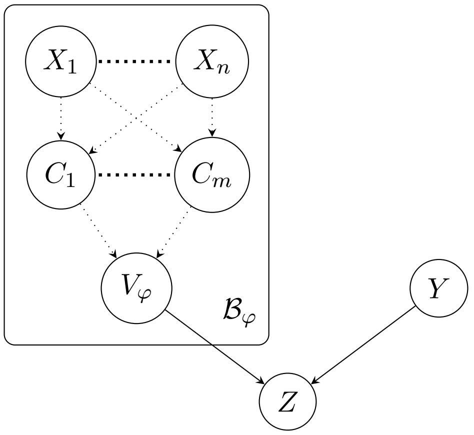

Figure 1: The network $\mathcal{B}^*_\varphi$ corresponding to a propositional formula φ.

4. Exact Active Inference

As remarked during the introduction, we follow Kwisthout (2014) in modelling the problem of active inference in terms of Bayesian networks. Apart from a particular network $\mathcal{B}$, assume we are also given two particular subsets $\mathbf{A}, \mathbf{P} \subseteq \mathbf{V}$ which make up the actions and predictions respectively, along with a target distribution $\text{Pr}_T(\mathbf{P})$ over these predictions. The problem of active inference is then to select an element $\mathbf{a} \in \mathbf{A}$, representing a compound action consisting of particular possibilities a for all $A \in \mathbf{A}$, which yields the greatest reduction in relative entropy of the target distribution with respect to the posterior likelihood of our predictions given $\mathbf{a}$. Here by the relative entropy of P with respect to Q, both distributions over some discrete variable X, we mean the *Kullback-Leibler divergence* given by $D_{KL}(P(X) \parallel Q(X)) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$. We remark that this value is non-negative, though possibly infinite or even undefined, and moreover asymmetric in general.

In order to formulate the corresponding decision problem, we introduce some notation for three important quantities involving relative entropy. First, we will use κ_I to denote the *initial* relative entropy $D_{KL}(\text{Pr}_T(\mathbf{P}) \parallel \text{Pr}(\mathbf{P}))$. In turn, $\min\{D_{KL}(\text{Pr}_T(\mathbf{P}) \parallel \text{Pr}(\mathbf{P} \mid \mathbf{a})) : \mathbf{a} \in \Omega(\mathbf{A}), \text{Pr}(\mathbf{a}) > 0\}$ will be referred to as the *optimal* relative entropy κ_O. Lastly, the *decrease* in relative entropy Δ_κ is simply $\kappa_I - \kappa_O$.[5] There are now two possible formalizations of active inference:

ACTIVE INFERENCE (THRESHOLD VERSION)
Input: A Bayesian network $\mathcal{B} = (\mathbf{G}, \mathbf{Pr})$, $\mathbf{A}, \mathbf{P} \subseteq \mathbf{V}$, a probability distribution $\text{Pr}_T(\mathbf{P})$, $q \in \mathbb{Q}$.
Question: Is $\kappa_O < q$?

ACTIVE INFERENCE (ABSOLUTE VERSION)
Input: A Bayesian network $\mathcal{B} = (\mathbf{G}, \mathbf{Pr})$, $\mathbf{A}, \mathbf{P} \subseteq \mathbf{V}$, a probability distribution $\text{Pr}_T(\mathbf{P})$, $q \in \mathbb{Q}$.
Question: Is $\Delta_\kappa > q$?

Note that whenever κ_I is known, for instance when given as a parameter, the two variants of active inference become interchangeable. This is because going from κ_I to below a threshold q is the

[5] If either κ_I, κ_O or both are infinite, we take Δ_κ to be ∞, $-\infty$ or 0 respectively, with $-\infty < q < \infty$ for all q.

same as reducing κ_I by at least $\kappa_I - q$, and vice versa reducing κ_I by at least q is the same as going below the threshold $\kappa_I - q$. However, before we present our first hardness result showing that κ_I nevertheless has little effect as a parameter, we must touch on what it means to parameterize by the typically irrational values κ_I, κ_O or Δ_κ (from here on called "relative entropy parameters").

To this end, note that there are only finitely many possible values of $D_{KL}(\Pr_T(\mathbf{P}) \parallel \Pr(\mathbf{P} \mid \mathbf{a}))$ for $\mathbf{a} \in \mathbf{A}$. It is therefore possible and sufficient to give a rational approximation $\tilde{\kappa}_I$ of κ_I such that $\tilde{\kappa}_I - D_{KL}(\Pr_T(\mathbf{P}) \parallel \Pr(\mathbf{P} \mid \mathbf{a})) > q$ precisely when $\kappa_I - D_{KL}(\Pr_T(\mathbf{P}) \parallel \Pr(\mathbf{P} \mid \mathbf{a})) > q$. Similarly, when parameterizing with κ_O or Δ_κ, we must provide rational approximations $\tilde{\kappa}_O$ and $\tilde{\Delta}_\kappa$ such that respectively $\kappa_I - \tilde{\kappa}_O > q$ precisely when $\Delta_\kappa > q$ and $\kappa_I - \tilde{\Delta}_\kappa < q$ precisely when $\kappa_O < q$. A relative entropy parameter is therefore such a rational approximation with the promise that it is accurate enough to carry the relevant information pertaining to the problem.

Proposition 7 *The problem κ_I-ACTIVE INFERENCE is $\mathsf{paraNP^{PP}}$-hard. Furthermore, we find that the problem $\{|\Omega(\mathbf{A})|, |\Omega(\mathbf{P})|, \kappa_I\}$-ACTIVE INFERENCE is still paraPP-hard.*

Proof We reduce from E-MAJSAT. Given the formula φ and $\kappa_I > 0$, construct the network $\mathcal{B}^*_\varphi$ by taking Y to be uniformly distributed and[6]

$$\Pr(z \mid V_\varphi, Y) = \begin{cases} \dfrac{1}{2} + 2^{-\kappa_I - 1} \text{ and } 3 \cdot 2^{-\kappa_I - 1} & \text{for } V_\varphi = \top, Y = y \quad \text{ and } \kappa_I \leq 1 \text{ resp. } \kappa_I > 1 \\ 3 \cdot 2^{-\kappa_I - 1} - \dfrac{1}{2} \text{ and } 2^{-\kappa_I - 1} & \text{for } V_\varphi = \top, Y = \neg y \quad \text{and } \kappa_I \leq 1 \text{ resp. } \kappa_I > 1 \\ 2^{-\kappa_I} & \text{for } V_\varphi = \bot \end{cases}$$

For $\kappa_I \leq 1$ we find $\Pr(z) = 2^{-\kappa_I}(1 - \Pr(\top)) + \frac{1}{2}\Pr(\top)(\frac{1}{2} + 2^{-\kappa_I-1} + 3 \cdot 2^{-\kappa_I - 1} - \frac{1}{2}) = 2^{-\kappa_I}$, and for $\kappa_I > 1$ we find $\Pr(z) = 2^{-\kappa_I}(1 - \Pr(\top)) + \frac{1}{2}\Pr(\top)(3 \cdot 2^{-\kappa_I - 1} + 2^{-\kappa_I - 1}) = 2^{-\kappa_I}$. Thus by letting $\Pr_T(z) = 1$, we have $D_{KL}(\Pr_T(Z) \parallel \Pr(Z)) = -\log \Pr(z) = \kappa_I$ as desired. Furthermore, note that $\Pr(z \mid y) \geq \Pr(z \mid \neg y)$, and $\Pr(z \mid y)$ is $2^{-\kappa_I} + (\frac{1}{2} - 2^{-\kappa_I - 1})\Pr(\top)$ and $2^{-\kappa_I}(1 + \frac{1}{2}\Pr(\top))$ for $\kappa_I \leq 1$ and $\kappa_I > 1$ respectively. Hence there is now an assignment $a_1, \ldots, a_k$ such that $\Pr(V_\varphi = \top \mid a_1, \ldots, a_k) > \frac{1}{2}$ precisely when $\Pr(z \mid a_1, \ldots, a_k, y) > 3 \cdot 2^{-\kappa_I - 2} + \frac{1}{4}$ for $\kappa_I \leq 1$ and $\Pr(z \mid a_1, \ldots, a_k, y) > 5 \cdot 2^{-\kappa_I - 2}$ for $\kappa_I > 1$. For the threshold version, this corresponds to $q = \kappa_I + 2 - \log(2^{\kappa_I} + 3)$ and $q = \kappa_I + 2 - \log 5$ for $\kappa_I \leq 1$ and $\kappa_I > 1$; for the absolute version, this corresponds to $q = \log(2^{\kappa_I} + 3) - 2$ and $q = \log 5 - 2$ for $\kappa_I \leq 1$ and $\kappa_I > 1$. The second statement follows by considering $k = 0$, effectively making this a reduction from MAJSAT. ■

As the initial relative entropy does not fully constrain the optimal relative entropy, one could have expected that as a parameter κ_I does not affect the hardness of ACTIVE INFERENCE. Indeed, most of the hardness of the problem is due to having to determine the value of κ_O, which is more difficult than for κ_I. Yet when given κ_O as a parameter in its place, the same hardness result for ACTIVE INFERENCE obtains, though observe that now only the absolute version is meaningful.

Proposition 8 *The problem $\{|\Omega(\mathbf{A})|, |\Omega(\mathbf{P})|, \kappa_O\}$-ACTIVE INFERENCE is paraPP-hard.*

[6] Here we face the same issue as when parameterizing by κ_I that $2^{-\kappa_I}$ is generally not a rational value: as before, this is resolved by taking a rational approximation instead such that the important properties are retained.

Proof We reduce from MAJSAT. Given the formula φ and $\kappa_O \geq 0$, construct the network $\mathcal{B}_\varphi^*$ by taking Y to be uniformly distributed and

$$\Pr(z \mid V_\varphi, Y) = \begin{cases} 2^{-\kappa_O - 1} & \text{for } V_\varphi = \top \text{ and } Y = \neg y \\ 2^{-\kappa_O} & \text{otherwise} \end{cases}$$

Now $\Pr(z) = \frac{1}{2}2^{-\kappa_O} + \frac{1}{2}\Pr(\top)2^{-\kappa_O - 1} + \frac{1}{2}(1 - \Pr(\top))2^{-\kappa_O} = 2^{-\kappa_O}(1 - \frac{1}{4}\Pr(\top))$, as well as $\Pr(z \mid y) = 2^{-\kappa_O}$ with $\Pr(z \mid y) \geq \Pr(z \mid \neg y)$. Thus by letting $\Pr_T(z) = 1$ we find that y is the best choice with $D_{KL}(\Pr_T(Z) \parallel \Pr(Z \mid y)) = \kappa_O$ as required, and $\Delta_\kappa = \kappa_O - \log(1 - \frac{1}{4}\Pr(\top)) - \kappa_O$, which is greater than $q = 3 - \log 7$ if and only if $\Pr(\top) > \frac{1}{2}$. ■

This time, the hardness of the problem lies in having to compute the value of κ_I. The same reasoning and subsequent hardness applies when we parameterize with Δ_κ instead, although this time it is only meaningful to consider the threshold version of ACTIVE INFERENCE.

Proposition 9 *The problem* $\{|\Omega(\mathbf{A})|, |\Omega(\mathbf{P})|, \Delta_\kappa\}$-ACTIVE INFERENCE *is* paraPP-*hard.*

Proof We reduce from MAJSAT. Given the formula φ and $\Delta_\kappa \geq 0$, construct the network $\mathcal{B}_\varphi^*$ by taking $\Pr(y) = \frac{2}{3}2^{-\Delta_\kappa}$ and

$$\Pr(z \mid V_\varphi, Y) = \begin{cases} 1 - \dfrac{1}{3}2^{-\Delta_\kappa} & \text{for } V_\varphi = \top \text{ and } Y = y \\ \dfrac{1}{3}2^{-\Delta_\kappa}(1 + (3 \cdot 2^{\Delta_\kappa} - 2)^{-1}) & \text{for } V_\varphi = \top \text{ and } Y = \neg y \\ 1 - \dfrac{2}{3}2^{-\Delta_\kappa} & \text{for } V_\varphi = \bot \text{ and } Y = y \\ \dfrac{1}{3}2^{-\Delta_\kappa} & \text{for } V_\varphi = \bot \text{ and } Y = \neg y \end{cases}$$

We find that $\Pr(z \mid y) = \Pr(\top)(1 - \frac{1}{3}2^{-\Delta_\kappa}) + (1 - \Pr(\top))(1 - \frac{2}{3}2^{-\Delta_\kappa}) = 1 - \frac{1}{3}2^{-\Delta_\kappa}(2 - \Pr(\top))$, and that $\Pr(z \mid y) \geq \Pr(z \mid \neg y)$. On the other hand, it is the case that

$$\begin{aligned} \Pr(z) &= \Pr(\top)\frac{2}{3}2^{-\Delta_\kappa}(1 - \frac{1}{3}2^{-\Delta_\kappa}) + \Pr(\top)(1 - \frac{2}{3}2^{-\Delta_\kappa})\frac{1}{3}2^{-\Delta_\kappa}(1 + (3 \cdot 2^{\Delta_\kappa} - 2)^{-1}) \\ &\quad + (1 - \Pr(\top))\frac{2}{3}2^{\Delta_\kappa}(1 - \frac{2}{3}2^{\Delta_\kappa}) + (1 - \Pr(\top))(1 - \frac{2}{3}2^{-\Delta_\kappa})\frac{1}{3}2^{-\Delta_\kappa} \\ &= 2^{-\Delta_\kappa}\left[\Pr(\top)(\frac{2}{3} + \frac{1}{3} - \frac{1}{3}2^{-\Delta_\kappa}(\frac{2}{3} + \frac{2}{3} - \frac{1}{3})) + (\frac{2}{3} + \frac{1}{3})(1 - \Pr(\top))(1 - \frac{2}{3}2^{-\Delta_\kappa})\right] \\ &= 2^{-\Delta_\kappa} \cdot \Pr(z \mid y) \end{aligned}$$

By letting $\Pr_T(z) = 1$, we have $D_{KL}(\Pr_T(Z) \parallel \Pr(Z)) - D_{KL}(\Pr_T(Z) \parallel \Pr(Z \mid y)) = \Delta_\kappa$, and moreover $D_{KL}(\Pr_T(Z) \parallel \Pr(Z \mid y)) < -\log(1 - 2^{-\Delta_\kappa - 1})$ if and only if $\Pr(\top) > \frac{1}{2}$. ■

We can push these results even further if we also consider the restricted local variance bound $B = \sum_{X \in A \cup P} \frac{u_X}{\ell_X}$ where $u_X = \max\{\Pr(X = x \mid \rho(X) = \pi) : x \in \Omega(X), \pi \in \Omega(\rho(X))\}$ and $\ell_X = \min\{\Pr(X = x \mid \rho(X) = \pi) : x \in \Omega(X), \pi \in \Omega(\rho(X))\}$. In all of the previous three cases, adding B as a parameter is still insufficient to guarantee tractability.

Proposition 10 *$\mathsf{PP} \subseteq \mathsf{C}$ holds for a class C whenever there is any $K \in \{\kappa_I, \kappa_O, \Delta_\kappa\}$ for which it is the case that the problem $\{|\Omega(\mathbf{A})|, |\Omega(\mathbf{P})|, K, B\}$-ACTIVE INFERENCE is in paraC.*

Proof In the reductions given in the proofs of Propositions 7, 8 and 9, taking $\kappa_I = 1$, $\kappa_O = 1$ and $\Delta_\kappa = 2 - \log 3$ respectively leads to $B = 3$. Hence we can reduce MAJSAT to ACTIVE INFERENCE in a way which sets all of $\{|\Omega(\mathbf{A})|, |\Omega(\mathbf{P})|, K, B\}$ to constant, so that any algorithm to solve the parameterized problem induces a classical algorithm for MAJSAT. ■

These results hint at the necessity of either considering a stronger set of parameters, or settling for an approximate solution to the problem. In the next section we will show that the latter approach also does not trivially lead to parameterized tractability, but not before we provide a positive result along the former lines (though note the absence of a relative entropy parameter). Here we define $\Omega(\mathbf{P})_+ = \{\mathbf{p} \in \Omega(\mathbf{P}) : \Pr_T(\mathbf{p}) > 0\}$, and use tw to denote the treewidth of the moralised graph.

Proposition 11 *The problem $\{\max_{X \in \mathbf{V}} |\Omega(X)|, |\mathbf{A}|, |\Omega(\mathbf{P})_+|, \mathrm{tw}\}$-ACTIVE INFERENCE is in FPT.*

Proof We know that CONDITIONAL INFERENCE is FPT for the parameters $\max_{X \in \mathbf{V}} |\Omega(X)|$ and tw, using the junction tree algorithm (Lauritzen and Spiegelhalter (1988)). This transforms the network into a cluster graph where each cluster contains at most tw + 1 nodes, so that marginalization (which can be computed separately within each separate cluster) takes into account at most $(\max_{X \in \mathbf{V}} |\Omega(X)|)^{\mathrm{tw}+1}$ possible instantiations at once. Thus we can compute $\Pr(\mathbf{p})$ and $\Pr(\mathbf{p} \mid \mathbf{a})$ for all $\mathbf{p} \in \Omega(\mathbf{P})_+$ and $\mathbf{a} \in \Omega(\mathbf{A})$ by doing inference for a maximum of $|\Omega(\mathbf{P})_+|(|\Omega(\mathbf{A})|+1)$ times, which is therefore FPT in $\{\max_{X \in \mathbf{V}} |\Omega(X)|, |\mathbf{A}|, |\Omega(\mathbf{P})_+|, \mathrm{tw}\}$ as $|\Omega(\mathbf{A})| \leq (\max_{X \in \mathbf{V}} |\Omega(X)|)^{|\mathbf{A}|}$. Finally, using all these probabilities, we can approximate κ_I and κ_O to within the required precision for a comparison against the threshold value q, which is in FPT for $|\Omega(\mathbf{P})_+|$ and $|\Omega(\mathbf{A})|$. This shows that (either version of) ACTIVE INFERENCE is in FPT for the given parameters. ■

5. Approximate Active Inference

Given the hardness results of the previous section, we now turn to a parameterized analysis of an approximate version of ACTIVE INFERENCE instead, with our approach following that of Marx (2008). That is, in order to capture the notion of approximation, the approximate problem also specifies an error bound ϵ as part of its input, and subsequently asks not whether $\kappa_O < q$ or $\Delta_\kappa > q$, but instead whether $\kappa_O \leq q \pm \epsilon$ or $\Delta_\kappa \geq q \pm \epsilon$ respectively. Here, $\kappa_O \leq q \pm \epsilon$ is shorthand for the question whether $\kappa_O \leq q - \epsilon$, given the promise that $\kappa_O \notin (q - \epsilon, q + \epsilon)$; the case of $\Delta_\kappa \geq q \pm \epsilon$ is defined analogously. This enables us to explicitly parameterize by ϵ^{-1}: yet taking ϵ^{-1} in place of a relative entropy parameter is still not enough to yield tractability, as evidenced by the result below.

Proposition 12 *The problem $\{|\Omega(A)|, |\Omega(P)|, \epsilon^{-1}\}$-ACTIVE INFERENCE is paraNP-hard.*

Proof We reduce from SATISFIABILITY. Given the formula φ and $\epsilon > 0$, we again construct the network $\mathcal{B}^*_\varphi$, here by taking $\Pr(y) = (2^n(2^{2\epsilon} - 1) + 1)^{-1}$ and

$$\Pr(z \mid V_\varphi, Y) = \begin{cases} 1 - (1 - 2^{-n})2^{-2\epsilon} & \text{for } V_\varphi = \top \text{ and } Y = y \\ 0 & \text{for } V_\varphi = \top \text{ and } Y = \neg y \\ 2^{-n-2\epsilon} & \text{for } V_\varphi = \bot \end{cases}$$

Then $\Pr(z) = \Pr(\top)(2^n(2^{2\epsilon}-1)+1)^{-1}(1-(1-2^{-n})2^{-2\epsilon}) + (1-\Pr(\top))2^{-n-2\epsilon} = \Pr(\top)2^{-n-2\epsilon} + (1-\Pr(\top))2^{-n-2\epsilon} = 2^{-n-2\epsilon}$, hence letting $\Pr_T(z) = 1$ leads to $\kappa_I = n+2\epsilon$. Now $\Pr(z \mid y) \geq \Pr(z \mid \neg y)$, with $\Pr(z \mid y) = \Pr(\top) - (\Pr(\top) - 2^{-n})2^{-2\epsilon} = 2^{-n}(\#\varphi - (\#\varphi - 1)2^{-2\epsilon})$, hence $D_{KL}(\Pr_T(Z) \parallel \Pr(Z \mid y)) = n - \log(\#\varphi - (\#\varphi - 1)2^{-2\epsilon})$. If φ is not satisfiable, this is $n + 2\epsilon$ again, whereas if φ is satisfiable, this value is at most n. This gives us that $\kappa_O \leq n + \epsilon \pm \epsilon$, or alternatively $\Delta_\kappa \geq \epsilon \pm \epsilon$, if and only if φ is satisfiable. ■

However, when we take both ϵ^{-1} and one of the stronger relative entropy parameters κ_O or Δ_κ, we can in fact derive a tractability result for the approximate version of ACTIVE INFERENCE.

Proposition 13 *Both* $\{|\Omega(\mathbf{P})_+|, \kappa_O, \epsilon^{-1}\}$-ACTIVE INFERENCE *and* $\{|\Omega(\mathbf{P})_+|, \Delta_\kappa, \epsilon^{-1}\}$-ACTIVE INFERENCE *are in* paraBPP*, i.e. they are randomized fixed-parameter tractable.*

Proof The main insight is that we only need to compute κ_I to sufficient precision (determined by the threshold q and the error bound ϵ) in order to correctly answer the question whether $\Delta_\kappa > q$ respectively $\kappa_O < q$. This can be done by approximating $\Pr(\mathbf{p})$ for all $\mathbf{p} \in \Omega(\mathbf{P})_+$ with sufficient accuracy using forward sampling (Henrion (1988)), which is in paraBPP for ϵ^{-1} and $|\Omega(\mathbf{P})_+|$, and in turn computing κ_I from these values is in FPT for $|\Omega(\mathbf{P})_+|$. This shows that the approximate version of ACTIVE INFERENCE is randomized fixed-parameter tractable for the given parameters. ■

The result above is of a slightly unusual nature in that the parameter κ_O or Δ_κ is not used to suitably bound the time taken by an algorithm capable of solving ACTIVE INFERENCE, but instead we take its explicit value in order to avoid having to compute it separately. This suggests that there is no analogue of Proposition 13 for κ_I, and that there should in fact be a corresponding hardness result. We did not succeed in showing this: observe that one cannot extend the approach used in the proof of Proposition 12 to include κ_I, as necessarily $\kappa_I \geq n$ for this strategy to work, but the SATISFIABILITY problem is trivially in FPT for the parameter n.

However, we have been able to extend the classic result by Dagum and Luby (1993) by showing that approximate conditional inference is NP-hard under randomized reduction, even when it is parameterized by ϵ^{-1}, κ_I and $|\Omega(\mathbf{P})_+|$, under a certain interpretation of the latter two. Specifically, we consider $D_{KL}(\Pr(\mathbf{H} \mid \mathbf{e}) \parallel \Pr(\mathbf{H}))$ for κ_I and $\{\mathbf{h} \in \Omega(\mathbf{H}) : \Pr(\mathbf{h} \mid \mathbf{e}) > 0\}$ for $\Omega(\mathbf{P})_+$. The problem from which we present a reduction is SATISFIABILITY WITH UNIQUENESS PROMISE, which was shown to be NP-hard under randomized reduction by Valiant and Vazirani (1986). We provide its description below, along with that of CONDITIONAL INFERENCE for completeness' sake.

SATISFIABILITY WITH UNIQUENESS PROMISE
Input: A propositional formula φ.
Promise: The formula φ has at most one satisfying truth assignment.
Question: Is φ satisfiable?

CONDITIONAL INFERENCE (APPROXIMATE VERSION)
Input: A Bayesian network $\mathcal{B} = (\mathbf{G}, \mathbf{Pr})$, two sets of variables $\mathbf{H}, \mathbf{E} \subseteq \mathbf{V}$, joint value assignments $\mathbf{h} \in \Omega(\mathbf{H}), \mathbf{e} \in \Omega(\mathbf{E})$, and rational values $0 \leq q \leq 1, 0 < \epsilon \leq \frac{1}{2}$.
Question: Is $\Pr(\mathbf{h} \mid \mathbf{e}) \geq q \pm \epsilon$?

Proposition 14 *The problem* $\{\epsilon^{-1}, \kappa_I, |\Omega(\mathbf{H})_+|\}$-CONDITIONAL INFERENCE *is* paraNP*-hard under randomized reduction, i.e. under polynomial time reduction with one-sided polynomial error.*

Proof We reduce from SATISFIABILITY WITH UNIQUENESS PROMISE. Given the formula φ and $0 < \epsilon \leq \frac{1}{2}$, once more we construct the network $\mathcal{B}^*_\varphi$ by taking Y to be uniformly distributed and

$$\Pr(z \mid V_\varphi, Y) = \begin{cases} 2\epsilon(1 - 2^{-n}) & \text{for } V_\varphi = \top \text{ and } Y = y \\ 0 & \text{for } V_\varphi = \top \text{ and } Y = \neg y \\ 2^{-n}(\dfrac{1}{2} - \epsilon)^2 & \text{for } V_\varphi = \bot \text{ and } Y = y \\ 2^{-n}(\dfrac{1}{2} - \epsilon)(\dfrac{1}{2} + \epsilon) & \text{for } V_\varphi = \bot \text{ and } Y = \neg y \end{cases}$$

Let $\mathbf{h} = y$ and $\mathbf{e} = z$. We now find that

$$\Pr(y \mid z) = \frac{\Pr(\top)2\epsilon(1 - 2^{-n}) + (1 - \Pr(\top))2^{-n}(\frac{1}{2} - \epsilon)^2}{\Pr(\top)2\epsilon(1 - 2^{-n}) + (1 - \Pr(\top))2^{-n}(\frac{1}{2} - \epsilon)}.$$

This means that $\Pr(y \mid z) = \frac{1}{2} - \epsilon$ whenever $\#\varphi = 0$, whereas $\Pr(y \mid z) = \frac{1}{2} + \epsilon$ whenever $\#\varphi = 1$, hence $\Pr(y \mid z) \geq \frac{1}{2} \pm \epsilon$ if and only if φ is satisfiable. Moreover, we find in general that $D_{KL}(\Pr(Y \mid z) \parallel \Pr(Y)) = 1 + \Pr(y \mid z) \log \Pr(y \mid z) + \Pr(\neg y \mid z) \log \Pr(\neg y \mid z)$, which evaluates to $1 + (\frac{1}{2} + \epsilon) \log(\frac{1}{2} + \epsilon) + (\frac{1}{2} - \epsilon) \log(\frac{1}{2} - \epsilon)$ both when $\#\varphi = 0$ and $\#\varphi = 1$. Thus if the formula φ satisfies the uniqueness promise, the approximation of $1 + (\frac{1}{2} + \epsilon) \log(\frac{1}{2} + \epsilon) + (\frac{1}{2} - \epsilon) \log(\frac{1}{2} - \epsilon)$ will satisfy the promise that it corresponds to $\kappa_I = D_{KL}(\Pr(Y \mid z) \parallel \Pr(Y))$.[7] ■

While falling short of the hardness result we were ultimately interested in, Proposition 14 strongly suggests that ACTIVE INFERENCE is hard for these parameters, in particular ϵ^{-1} and κ_I.[8] If this is indeed true, this would provide a serious challenge to the view common in cognitive neuroscience that an approximate version of active inference is feasible when the prediction error is low. Before we return to this discussion in our conclusion, we finish this section as we did the previous one by demonstrating how a known algorithm for inference can be used to obtain a parameterized tractability result for ACTIVE INFERENCE, though again one without a relative entropy parameter.

Proposition 15 *The problem* $\{|\Omega(\mathbf{A})|, |\Omega(\mathbf{P})|, B, \epsilon^{-1}\}$-ACTIVE INFERENCE *is in* paraBPP*, i.e. it is randomized fixed-parameter tractable.*

Proof By Dagum and Luby (1997), $\{|\mathbf{H}|, |\mathbf{E}|, B, \epsilon^{-1}\}$-CONDITIONAL INFERENCE is randomized fixed-parameter tractable, as one needs to do likelihood weighting at most $\mathcal{O}(B^{|\mathbf{H}|+|\mathbf{E}|}\epsilon^{-2})$ passes to obtain an accurate approximation of $\Pr(\mathbf{h} \mid \mathbf{e})$. Thus we can approximate $\Pr(\mathbf{p})$ and $\Pr(\mathbf{p} \mid \mathbf{a})$ for all $\mathbf{p} \in \Omega(\mathbf{P})_+$ and $\mathbf{a} \in \Omega(\mathbf{A})$ in a number of passes parameterized by $\{|\Omega(\mathbf{A})|, |\Omega(\mathbf{P})|, B, \epsilon^{-1}\}$, allowing us to probabilistically solve ACTIVE INFERENCE for this parameter set. ■

[7] Note that we can in fact force any value $\kappa_I \geq 1$ to obtain by including a third possible value $\tilde{y}$ to $\Omega(Y)$ and taking $\Pr(y) = \Pr(y') = 2^{-\kappa_I}(\frac{1}{2} + \epsilon)^{(\frac{1}{2}+\epsilon)}(\frac{1}{2} - \epsilon)^{(\frac{1}{2}-\epsilon)}$, with $\Pr(z \mid \tilde{y}) = 0$ independent of V_φ. Evaluating the expression $D_{KL}(\Pr(\mathbf{H} \mid \mathbf{e}) \parallel \Pr(\mathbf{H}))$ will then return the chosen value κ_I as desired.

[8] Although there is a straightforward reduction from CONDITIONAL INFERENCE to ACTIVE INFERENCE, because the definition of κ_I is incompatible between the two problems, this reduction does not work with κ_I as a parameter.

6. Concluding Remarks

In this paper we set out to investigate the computational hardness of the problem of active inference, based on a formalization in terms of Bayesian networks which is deemed to adequately capture the way active inference is thought of in its original setting, namely in cognitive neuroscience informed by the ideas underlying the theory of predictive processing. In the literature concerning active inference, whenever considerations of computational complexity are explicitly addressed, the common suggestion is that approximation and small prediction errors are what enables the efficient operation of cognitive systems involved with carrying out this task. To this end, we engaged in a parameterized complexity analysis of the formal problem centered around precisely these two features, along with a number of relevant additional parameters. What can we say about this view based on the hardness and tractability results which we obtained in the previous two sections?

While Proposition 11 serves as a reminder that approximation is not necessary for tractability, it is unclear whether the treewidth would be small in relevant instances. Moreover, Proposition 12 illustrates that even in contexts with a small scope, approximation alone is insufficient to guarantee tractability. Thus there is a need for additional factors to explain why active inference could be feasibly performed on the small time scales required to function in the real world, and so the question which plays a central role is whether the size of the prediction error can indeed serve this purpose.

Although it does not provide a decisive answer, we believe that Proposition 14 makes it sufficiently plausible that it is not the case that a small prediction error makes an approximate version of active inference tractable. Proposition 13 may suggest otherwise, but we claim that only the initial relative entropy κ_I truly captures the role of the prediction error in predictive processing accounts, and that the part played by κ_O and Δ_κ in the proof of this result is sufficiently far removed from this interpretation as to make the result unsuitable for this context. Accepting this assessment entails the need for a more refined explanation of how cognitive agents can routinely engage in active inference.

Furthermore, we expect most of the results and insights obtained in this paper to carry over to related problems such as observation selection, which we already touched on in the introduction. While the problem of active inference itself might not be native to the field of probabilistic models, in this way our findings may still contribute to the broader research within this area. In particular we conjecture that the KL divergence does not play an essential role in these results, so that other measures may be accommodated for as well, which would bestow further generality on these results.

Acknowledgments

The author would like to acknowledge the NWO EW TOP grant 612.001.601 which has supported his research for this paper. Furthermore, he would like to thank both Johan Kwisthout and Hans Bodlaender for helpful discussions and comments on earlier versions.

References

A. Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

G. F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60:141–153, 1993.

P. Dagum and M. Luby. An optimal approximation algorithm for Bayesian inference. *Artificial Intelligence*, 93:1–27, 1997.

R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.

J. Flum and M. Grohe. Describing parameterized complexity classes. *Information and Computation*, 187:291–319, 2003.

K. Friston et al. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

J. Gill. Computational complexity of probabilistic Turing machines. *SIAM Journal on Computing*, 6(4):675–695, 1977.

M. Henrion. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In J. F. Lemmer and L. N. Kanal, editors, *Uncertainty in Artificial Intelligence*, volume 5 of *Machine Intelligence and Pattern Recognition*, pages 149–163. 1988.

R. Impagliazzo and A. Wigderson. P = BPP if E requires exponential circuits: Derandomizing the XOR lemma. In *Proceedings of STOC '97*, pages 220–229, 1997.

A. Krause and C. Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, 35:557–591, 2009.

J. Kwisthout. *The Computational Complexity of Probabilistic Networks*. PhD thesis, Utrecht University, 2009.

J. Kwisthout. Minimizing relative entropy in Hierarchical Predictive Coding. In L. C. van der Gaag and A. J. Feelders, editors, *Proceedings of PGM 2014*, pages 254–270, 2014.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50(2):157–224, 1988.

D. Marx. Parameterized complexity and approximation algorithms. *The Computer Journal*, 51: 60–78, 2008.

J. D. Park and A. Darwiche. Complexity results and approximation strategies for MAP explanations. *Journal of Artificial Intelligence Research*, 21:101–133, 2004.

S. Toda. PP is as hard as the polynomial-time hierarchy. *SIAM Journal on Computing*, 20(5): 865–877, 1991.

L. G. Valiant and V. V. Vazirani. NP is as easy as detecting unique solutions. *Theoretical Computer Science*, 47:85–93, 1986.

K. W. Wagner. The complexity of combinatorial problems with succinct input representation. *Acta Informatica*, 23(3):325–356, 1986.