Minimizing Relative Entropy in Hierarchical Predictive Coding

Johan Kwisthout

Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands, j.kwisthout@donders.ru.nl

Abstract. The recent Hierarchical Predictive Coding theory is a very influential theory in neuroscience that postulates that the brain continuously makes (Bayesian) predictions about sensory inputs using a generative model. The Bayesian inferences (making predictions about sensory states, estimating errors between prediction and observation, and lowering the prediction error by revising hypotheses) are assumed to allow for efficient approximate inferences in the brain. We investigate this assumption by making the conceptual ideas of how the brain may minimize prediction error computationally precise and by studying the computational complexity of these computational problems. We show that each problem is intractable in general and discuss the parameterized complexity of the problems.

1 Introduction

The assumption that the brain in essence is a Bayesian inferential machine, integrating prior knowledge with sensory information such as to infer the most probable explanation for the phenomena we observe, is quite wide spread in neuroscience [19]. Recently, this 'Bayesian brain' hypothesis has merged with the hypothesis that the brain is a prediction machine that continuously makes predictions about future sensory inputs, based on a generative model of the causes of these inputs [17] and with the free energy principle as a driving force of prediction error minimization [13]; the resulting theory has been called Hierarchical Predictive Coding or Predictive Processing [7]. It is assumed to explain and unify all cortical processes, spanning all of cognition [6]. Apart from being one of the most influential current unifying theories of the *modus operandi* of the brain, it has inspired researchers in domains such as developmental neurorobotics [23], human-robot interaction [25], and conscious presence in virtual reality [26].

At the very heart of Hierarchical Predictive Coding (hereafter HPC) are the Bayesian predictions, error estimations, and hypothesis revisions that are assumed to allow for efficient approximate Bayesian inferences in the brain [7]. As Bayesian inferences are intractable in general, even to approximate [1, 9], this invites the question to what extent the HPC mechanism indeed renders these inferences tractable [3, 22]. In essence, minimizing prediction errors boils down to minimizing the relative entropy or Kullback-Leibler divergence between

the predicted and observed distributions [14]. Lowering the relative entropy between prediction and observation can be done in many ways: we can revise the hypothesized causes that generated the prediction; alternatively, we may adjust the probabilistic dependences that modulate how predictions are generated from hypotheses, or we might want to seek and include additional observations into the model in order to adjust the posterior distribution over the predictions. In contrast, we might also bring prediction and observation closer to each other by *intervention* in the world, thus hopefully manipulating the observation to better match what we predicted or expected. This is referred to as *active inference* in the HPC literature [15].

The contribution of this paper is to make these informal notions explicit and to study the computational complexity of minimizing relative entropy using these notions. We show that each conceptualization of prediction error minimization yields an intractable (i.e., NP-hard) computational problem. However, we can clearly identify where the border between tractable and intractable lies by giving fixed-parameter tractability results for all discussed problems. The remainder of this paper is structured as follows. In Section 2 we formally define HPC in the context of discrete Bayesian networks. We recall some needed preliminaries from computational complexity and discuss related work. In Section 3 we discuss the complexity of computing entropy and relative entropy in Bayesian networks. In Sections 4 and 5 we discuss *belief revision* and *model revision*, respectively, and in Section 6 we investigate the complexity of deciding which observation to make in order to decrease prediction error. In Section 7 we turn to the complexity of *active inference*, i.e., deciding which possible action to perform to decrease prediction error. We switch to the parameterized complexity of these problems in Section 8. In Section 9 we conclude this paper and sketch possible future work.

2 Preliminaries

A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr_\mathcal{B})$ is a graphical structure that models a set of stochastic variables, the conditional independences among these variables, and a joint probability distribution over these variables. $\mathcal{B}$ includes a directed acyclic graph $\mathbf{G}_\mathcal{B} = (\mathbf{V}, \mathbf{A})$, modeling the variables and conditional independences in the network, and a set of conditional probability tables (CPTs) $\Pr_\mathcal{B}$ capturing the stochastic dependences between the variables. The network models a joint probability distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ over its variables, where $\pi(V_i)$ denotes the parents of V_i in $\mathbf{G}_\mathcal{B}$. By convention, we use upper case letters to denote individual nodes in the network, upper case bold letters to denote sets of nodes, lower case letters to denote value assignments to nodes, and lower case bold letters to denote joint value assignments to sets of nodes. We use the notation $\Omega(V_i)$ to denote the set of values that V_i can take. Likewise, $\Omega(\mathbf{V})$ denotes the set of joint value assignments to $\mathbf{V}$.

HPC can be understood as a cascading hierarchy of increasingly abstract hypotheses about the world, where the predictions on one level of the hierarchy are identified with the hypotheses at the subordinate level. At any particular level,

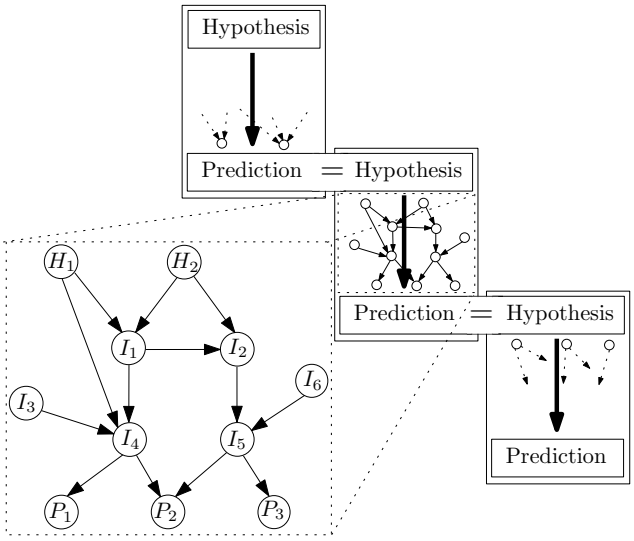

Fig. 1. An example level L of the HPC hierarchy, with hypothesis variables Hyp $= \{H_1, H_2\}$, prediction variables Pred $= \{P_1, P_2, P_3\}$, and intermediate variables Int $= \{I_1, \ldots, I_6\}$.

making a prediction based on the current hypothesis in any of the assumed levels corresponds to computing a posterior probability distribution $\Pr(\Pr_{\text{Pred}} \mid \Pr_{\text{Hyp}})$ over the space of candidate predictions, given the current estimated probability distribution over the space of hypotheses, modulated by contextual dependences. We can thus describe each level L of the HPC hierarchy as a Bayesian network $\mathcal{B}_L$, where the variables are partitioned into a set of hypothesis variables Hyp, a set of prediction variables Pred, and a set of intermediate variables Int, describing contextual dependences and (possibly complicated) structural dependences between hypotheses and predictions. We assume that all variables in Hyp are source variables, all variables in Pred are sink variables, and that the Pred variables in $\mathcal{B}_L$ are identified with the Hyp variables in $\mathcal{B}_{L+1}$ for all levels of the hierarchy save the lowest one (See Figure 1). As HPC is claimed to be a unifying mechanism describing all cortical processes [6], we do not impose additional *a priori* constraints on the structure of the network describing the stochastic relationships [16]. Motivated by the assumption that global prediction errors are minimized by local minimization [18], we will focus on the computations in a single level of the network.

Computing the prediction error at any level of the hierarchy corresponds to computing the relative entropy or Kullback-Leibler divergence

$$D_{\text{KL}}(\Pr_{(\text{Pred})} \| \Pr_{(\text{Obs})}) = \sum_{\mathbf{p} \in \Omega(\text{Pred})} \Pr_{\text{Pred}}(\mathbf{p}) \log \left(\frac{\Pr_{\text{Pred}}(\mathbf{p})}{\Pr_{\text{Obs}}(\mathbf{p})} \right)$$

between the probability distributions over the prediction Pred and the (possibly inferred) observation Obs[1]. In the remainder of this paper, to improve readability we abbreviate $D_{\mathrm{KL}}(\Pr_{(\mathrm{Pred})} \| \Pr_{(\mathrm{Obs})})$ to simply D_{KL} when the divergence is computed between $\Pr_{(\mathrm{Pred})}$ and $\Pr_{(\mathrm{Obs})}$; we sometimes include brackets $D_{\mathrm{KL}[\psi]}$ to refer to the divergence under some particular value assignment, parameter setting, or observation ψ.

The computed prediction error is used to bring prediction and observation closer to each other; either by belief revision, model revision, or by passive or active intervention. In belief revision, we lower prediction error by revising the probability distribution over the space of hypotheses $\Pr_{\mathrm{Hyp}}$; by model revision by revising some parameters in $\Pr_{\mathcal{B}}$; by passive intervention by observing the values of some of the intermediate variables; by active intervention by setting the values of some of the intermediate variables. These notions will be developed further in the remainder of the paper when we discuss the computational complexity of these mechanisms of lowering prediction error.

2.1 Computational Complexity

In the remainder, we assume that the reader is familiar with basic concepts of computational complexity theory, in particular Turing Machines, the complexity classes P and NP, and NP-completeness proofs. In addition to these basic concepts, to describe the complexity of various problems we will use the *probabilistic* class PP, oracle machines, and some basic principles from parameterized complexity theory. The interested reader is referred to [10] for more background on complexity issues in Bayesian networks, and to [12] for an introduction in parameterized complexity theory.

The class PP contains languages L that are accepted in polynomial time by a *Probabilistic Turing Machine*. This is a Turing Machine that augments the more traditional non-deterministic Turing Machine with a probability distribution associated with each state transition. Acceptance of an input x is defined as follows: the probability of arriving in an *accept state* is strictly larger than $1/2$ if and only if $x \in L$. This probability of acceptance, however, is not fixed and may (exponentially) depend on the input, e.g., a problem in PP may accept 'yes'-instances with size $|x|$ with probability $1/2 + 1/2^{|x|}$. This means that the probability of acceptance cannot in general be amplified by repeating the computation a polynomial number of times and making a decision based on a majority count, ruling out efficient randomized algorithms. Therefore, PP-complete problems are considered to be intractable. The canonical PP-complete problem is MAJSAT: given a Boolean formula ϕ, does the majority of the truth assignments satisfy ϕ? In Bayesian networks, the canonical problem of determining whether $\Pr(\mathbf{h} \mid \mathbf{e}) > q$ for a given rational q and joint variable assignments $\mathbf{h}$ and $\mathbf{e}$ (known as the INFERENCE problem) is PP-complete.

[1] Conform the definition of the Kullback-Leibler divergence, we will interpret the term $0 \log 0$ as 0 when appearing in this formula, as $\lim_{x \to 0} x \log x = 0$. The KL divergence is undefined if for any $\mathbf{p}$, $\Pr_{\mathrm{Obs}}(\mathbf{p}) = 0$ while $\Pr_{\mathrm{Pred}}(\mathbf{p}) \neq 0$.

A Turing Machine $\mathcal{M}$ has *oracle access* to languages in the class C, denoted as $\mathcal{M}^{\mathsf{C}}$, if it can decide membership queries in C ("consult the oracle") in a single state transition. For example, $\mathsf{NP}^{\mathsf{PP}}$ is defined as the class of languages which are decidable in polynomial time on a non-deterministic Turing Machine with access to an oracle deciding problems in PP.

Sometimes problems are intractable (i.e., NP-hard) in general, but become tractable if some *parameters* of the problem can be assumed to be small. Informally, a problem is called fixed-parameter tractable for a parameter k (or a set $\{k_1, \ldots, k_n\}$ of parameters) if it can be solved in time, exponential *only* in k and polynomial in the input size $|x|$, i.e., in time $\mathcal{O}(f(k) \cdot |x|^c)$ for a constant c and an arbitrary function f. In practice, this means that problem instances can be solved efficiently, even when the problem is NP-hard in general, if k is known to be small.

Finally, a word on the representation of numerical values. In the complexity proofs we assume that all parameter probabilities are rational numbers (rather than reals), and we assume that logarithmic functions are approximated when needed with sufficient precision, yet polynomial in the length of the problem instance. All logarithms in this paper have base 2.

2.2 Previous Work

The computational complexity of various problems in Bayesian networks is well studied. Interestingly, such problems tend to be complete for complexity classes with few other "real-life" complete problems. For example, deciding upon the MAP distribution is $\mathsf{NP}^{\mathsf{PP}}$-complete [24], as well as deciding whether the parameters in a network can be tuned to satisfy particular constraints [21]. Deciding whether a network is monotone is $\mathsf{co}-\mathsf{NP}^{\mathsf{PP}}$-complete [27], and computing the same-decision probability of a network has a $\mathsf{PP}^{\mathsf{PP}}$-complete decision variant [11]. Some results are known on the complexity of entropy computations: In [8] it was established $\#\mathsf{P}$-hardness of computing the (total) entropy of a Bayesian network; computing the relative entropy between two arbitrary probability distributions is PP-hard [20]. In [2] it was proved that no approximation algorithm can compute a bounded approximation on the entropy of arbitrary distributions using a polynomial amount of samples.

While concerns with respect to the computational complexity of inferences in (unconstrained) HPC models have been raised in [3] and [22], and acknowledged in [6], this paper is (to the best of our knowledge) the first to explicitly address the complexity of minimizing relative entropy in the context of HPC.

3 The Complexity of Computing Relative Entropy in HPC

The first computational problem we will discuss is the computation of the entropy of a prediction, and the relative entropy between a prediction and an

observation. While complexity results are known for the computation of the entropy of an entire network [8], respectively the relative entropy between two arbitrary distributions [20], we will here show that decision variants of both problems remain PP-complete even for *singleton* and *binary* hypothesis, prediction, and observation variables. The proof construct we introduce in this proof will be reused, with slight modifications, in subsequent proofs.

We start with defining a decision variant of ENTROPY.

ENTROPY
Instance: A Bayesian network $\mathcal{B}$ with designated variable subsets Pred and Hyp; rational number q.
Question: Is the entropy $E(\text{Pred}) = -\sum_{\mathbf{p} \in \Omega(\text{Pred})} \Pr(\mathbf{p}) \log \Pr(\mathbf{p}) < q$?

We will reduce ENTROPY from MINSAT, defined as follows:

MINSAT
Instance: A Boolean formula ϕ with n variables.
Question: Does the *minority* of truth assignments to ϕ satisfy ϕ?

Note that MINSAT is the complement problem of the PP-complete MAJSAT problem; as PP is closed under complement, MINSAT is PP-complete by a trivial reduction. In order to change as little as possible to the construct in subsequent proofs, we will sometimes reduce from MINSAT and sometimes from MAJSAT.

We will illustrate the reduction form MINSAT to ENTROPY using the example Boolean formula $\phi_{\text{ex}} = \neg x_1 \wedge (x_2 \vee \neg x_3)$; note that this is a 'yes'-instance to MINSAT as three out of eight truth assignments satisfy ϕ_{ex}. We construct a Bayesian network $\mathcal{B}_\phi$ from ϕ as follows. For every variable x_i in ϕ, we construct a binary variable X_i in $\mathcal{B}_\phi$, with values t and f and uniform probability distribution. The set of all variables $X_1, \ldots, X_n$ is denoted with $\mathbf{X}$. For each logical operator in ϕ, we create an additional variable in the network $\mathcal{B}_\phi$. The parents of this variable are the variables that correspond with the sub-formulas joined by the operator; its conditional probability table mimics the truth table of the operator. The variable associated with the top-level operator of ϕ will be denoted by V_ϕ. In addition, we include a binary hypothesis variable H, with uniformly distributed values t and f, and a binary prediction variable P, with values t and f. The parents of this variable are V_ϕ and H, and the conditional probability table of this variable mimics an *and*-operator, i.e., $\Pr(P = t \mid V_\phi, H) = 1$ if and only if both V_ϕ and H are set to t. In Figure 2 we illustrate how $\mathcal{B}_{\phi_{\text{ex}}}$ is thus constructed from ϕ_{ex}. We set Pred $= P$, Hyp $= H$, and $q = 1/2 - 3/4 \log 3/4$.

Theorem 1. ENTROPY *is* PP*-complete, even for singleton binary variables* Pred *and* Hyp*.*

Proof. Membership proof in PP follows from a trivial modification of the proof that computing the Kullback-Leibler divergence between two distributions is in PP, such as presented in [20].

To prove PP-hardness, we will reduce MINSAT to ENTROPY. Let ϕ be an instance of MINSAT and let $\mathcal{B}_\phi$ be the Bayesian network constructed from ϕ as

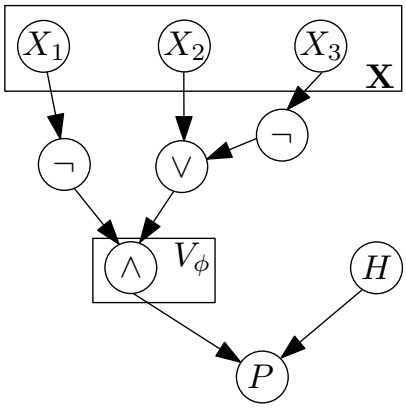

Fig. 2. The Bayesian network $\mathcal{B}_{\phi_{\text{ex}}}$ that is constructed from the MINSAT example ϕ_{ex}. Note that we here have a single hypothesis node H (a source node) and a single prediction node P (a sink node).

---

described above. Observe that in $\mathcal{B}_\phi$, the posterior probability $\Pr(V_\phi = t \mid \mathbf{X} = \mathbf{x}) = 1$ if and only if the truth assignment corresponding with the joint value assignment $\mathbf{x}$ satisfies ϕ, and 0 otherwise. In particular, if exactly half of the truth assignments satisfy ϕ, then $\Pr(V_\phi = t) = 1/2$ and consequently $\Pr(P = t) = 1/4$. The entropy then equals $E(P) = -(\Pr(P = t) \log \Pr(P = t) + \Pr(P = f) \log \Pr(P = f)) = -(1/4 \log 1/4 + 3/4 \log 3/4) = 1/2 - 3/4 \log 3/4$. The entropy ranges from $E(P) = 0$ in case ϕ is not satisfiable (and hence $\Pr(P = t) = 0$) and $E(P) = 1$ in case ϕ is a tautology (and hence $\Pr(P = t) = 1/2$). In particular, if and only if the minority of truth assignments to ϕ satisfies ϕ, then $E(P) < 1/2 - 3/4 \log 3/4 = \text{q}$. Note that the reduction can be done in polynomial time, given our assumptions on the tractable approximation of the logarithms involved; hence, ENTROPY is PP-complete. □

Computing the *relative* entropy between a prediction and an observation is defined as a decision problem as follows.

RELATIVEENTROPY
Instance: A Bayesian network $\mathcal{B}$ with designated variable subset Pred, where $\Pr_{(\text{Pred})}$ denotes the posterior distribution over Pred; an observed distribution $\Pr_{(\text{Obs})}$ over Pred; a rational number q.
Question: Is the relative entropy $D_{\text{KL}} < q$?

To prove PP-completeness, we use the same construction as above, but now we set $q = 3/4 \log 3/2 - 1/4$. In addition, we set $\Pr_{(\text{Obs})}$ to $\Pr(P = t) = 1/2$.

Theorem 2. RELATIVEENTROPY *is* PP*-complete, even for singleton binary variables* Pred *and* Hyp*.*

Proof. Membership in PP of the more general problem of computing the Kullback-Leibler divergence between two arbitrary probability distributions was established in [20]. To prove PP-hardness, we reduce MAJSAT to RELATIVEENTROPY. Let ϕ be an instance of MAJSAT and let $\mathcal{B}_\phi$ be the Bayesian network constructed from ϕ as described above. Observe that in $\mathcal{B}_\phi$ D_{KL} decreases when $\Pr(V_\phi = t)$ increases; in particular, when $\Pr(V_\phi = t) = p$ (and hence $\Pr(P = t) = p/2$), $D_{\mathrm{KL}} = p/2 \log(\frac{p/2}{1/2}) + (2-p)/2 \log(\frac{(2-p)/2}{1/2})$. Note that $\Pr(V_\phi = t) = 1/2$ if exactly half of the truth assignments to ϕ satisfy ϕ. Hence, if and only if a *majority* of truth assignments to ϕ satisfies ϕ, then $D_{\mathrm{KL}} < 1/4 \log(\frac{1/4}{1/2}) + 3/4 \log(\frac{3/4}{1/2}) = 3/4 \log 3/2 - 1/4 = q$. As the reduction can be done in polynomial time, this proves that RELATIVEENTROPY is PP-complete. □

In subsequent sections we will discuss the complexity of *lowering* the relative entropy D_{KL} by means of belief revision, model revision, or by passive or active intervention.

4 Revision of Beliefs

In this section we discuss *belief revision*, i.e., changing the probability distribution over the hypothesis variables, as a means to reduce relative entropy. We formulate two decision problems that capture this concept; the first one focuses on lowering the relative entropy *to* some threshold, the second one on lowering the relative entropy *by* some amount.

BELIEFREVISION1
Instance: A Bayesian network $\mathcal{B}$ with designated variable subsets Hyp and Pred, where $\Pr_{(\mathrm{Hyp})}$ denotes the prior distribution over Hyp, and $\Pr_{(\mathrm{Pred})}$ denotes the posterior distribution over Pred; an observed distribution $\Pr_{(\mathrm{Obs})}$ over Pred; a rational number q.
Question: Is there a (revised) prior probability distribution $\Pr_{(\mathrm{Hyp})'}$ over Hyp such that $D_{\mathrm{KL}[\mathrm{Hyp}']} < q$?

BELIEFREVISION2
Instance: As in BELIEFREVISION1.
Question: Is there a (revised) prior probability distribution $\Pr_{(\mathrm{Hyp})'}$ over Hyp such that $D_{\mathrm{KL}[\mathrm{Hyp}]} - D_{\mathrm{KL}[\mathrm{Hyp}']} > q$?

We prove that both problems are PP-hard via a reduction from MAJSAT, again using the construct that we used in the proof of Theorem 1, but we redefine the conditional probability distribution $\Pr(P \mid V_\phi, H)$ and we redefine $\Pr_{(\mathrm{Hyp})}$, $\Pr_{(\mathrm{Obs})}$, and q. Let $\Pr(P \mid V_\phi, H)$ be defined as follows:

$$\Pr(P = t \mid V_\phi, H) = \begin{cases} 3/8 \text{ if } V_\phi = t, H = t \\ 0 \quad \text{if } V_\phi = t, H = f \\ 1/8 \text{ if } V_\phi = f, H = t \\ 0 \quad \text{if } V_\phi = f, H = f \end{cases}$$

We set $\Pr_{(\text{Hyp})}$ to $\Pr(H = t) = 0$ and $\Pr_{(\text{Obs})}$ to $\Pr(P = t) = {}^{15}/_{16}$. For BELIEFREVISION1, we redefine $q = q_1 = {}^1/_4 \log(\frac{{}^1/_4}{{}^{15}/_{16}}) + {}^3/_4 \log(\frac{{}^3/_4}{{}^1/_{16}})$. For BELIEFREVISION2, we redefine $q = q_2 = 4 - {}^1/_4 \log(\frac{{}^1/_4}{{}^{15}/_{16}}) - {}^3/_4 \log(\frac{{}^3/_4}{{}^1/_{16}})$. We now claim the following.

Theorem 3. BELIEFREVISION1 *and* BELIEFREVISION2 *are* PP*-hard, even for singleton binary variables* Pred *and* Hyp*.*

Proof. To prove PP-hardness, we reduce BELIEFREVISION from MAJSAT. Let ϕ be an instance of MAJSAT and let $\mathcal{B}_\phi$ be the Bayesian network constructed from ϕ as described above. Observe that in $\mathcal{B}_\phi$ $D_{\text{KL}[\text{Hyp}]}$ is independent of $\Pr(V_\phi)$ as $\Pr(P = t \mid V_\phi, H) = 0$ (as $\Pr(H = t) = 0$) and thus $D_{\text{KL}[\text{Hyp}]} = 0 + \log(\frac{1}{{}^1/_{16}}) = 4$.

We now investigate the effect of revising the hypothesis distribution $\Pr_{(\text{Hyp})}$ to $\Pr_{(\text{Hyp})'}$. For every probability distribution $\Pr(V_\phi)$, D_{KL} increases when $\Pr(H = t)$ goes to 0, and decreases when $\Pr(H = t)$ goes to 1. That is, $D_{\text{KL}[\text{Hyp}']}$ is minimal for $\Pr_{(\text{Hyp})'} = \Pr(H = t) = 1$. In general, for $\Pr(H = t) = 1$ and $\Pr(V_\phi) = p$, $\Pr(P = t \mid V_\phi, H) = {}^{(2p+1)}/_8$ and $D_{\text{KL}[\text{Hyp}']} = {}^{(2p+1)}/_8 \log(\frac{{}^{(2p+1)}/_8}{{}^{15}/_{16}}) + {}^{(7-2p)}/_8 \log(\frac{{}^{(7-2p)}/_8}{{}^1/_{16}})$. For $\Pr(V_\phi) = {}^1/_2$ and $\Pr_{(\text{Hyp})'} = \Pr(H = t) = 1$, $\Pr(P = t \mid V_\phi, H) = {}^1/_4$ and $D_{\text{KL}[\text{Hyp}']} = {}^1/_4 \log(\frac{{}^1/_4}{{}^{15}/_{16}}) + {}^3/_4 \log(\frac{{}^3/_4}{{}^1/_{16}})$. We have in that case that $D_{\text{KL}[\text{Hyp}]} - D_{\text{KL}[\text{Hyp}']} = 4 - {}^1/_4 \log(\frac{{}^1/_4}{{}^{15}/_{16}}) - {}^3/_4 \log(\frac{{}^3/_4}{{}^1/_{16}})$.

In particular if and only if $\Pr(V_\phi) > {}^1/_2$ there exists a revised hypothesis distribution $\Pr_{(\text{Hyp})'}$ (i.e., $\Pr(H = t) = 1$) such that $D_{\text{KL}[\text{Hyp}']} < q_1$ and that $D_{\text{KL}[\text{Hyp}]} - D_{\text{KL}[\text{Hyp}']} > q_2$. Now, $\Pr(V_\phi) > {}^1/_2$ if and only if there is a majority of truth assignments to ϕ that satisfies ϕ. Given that the reduction can be done in polynomial time, this proves PP-hardness of both BELIEFREVISION1 and BELIEFREVISION2. □

Note that these problems are not known or believed to be in PP, as we need to determine a revised probability distribution $\Pr_{(\text{Hyp})'}$ as well as computing the relative entropy. In case Hyp is a singleton binary variable (as in our constrained proofs), the probability $\Pr_{(\text{Pred})}$ depends linearly on this distribution [4], but the complexity of this dependency grows when the distribution spans multiple variables. This makes a polynomial sub-computation of $\Pr_{(\text{Hyp})'}$, and thus membership in PP, unlikely. However, we can non-deterministically *guess* the value of $\Pr_{(\text{Hyp})'}$ and then decide the problem using an oracle for RELATIVEENTROPY; for this reason, the problems are certainly in the complexity class NP^{PP}.

5 Revision of Models

In the previous section we defined belief revision as the revision of the prior distribution over Hyp. We can also revise the stochastic dependences in the model, i.e., how Pred depends on Hyp (and Int). However, a naive formulation

of *model revision* will give us a trivial algorithm for solving it, yet unwanted side effects.

NaiveModelRevision
Instance: A Bayesian network $\mathcal{B}$ with designated variable subsets Hyp and Pred, where $\Pr_{(\text{Pred})}$ denotes the posterior distribution over Pred; an observed distribution $\Pr_{(\text{Obs})}$ over Pred; a rational number q.
Question: Is there a probability distribution $\Pr_{\text{new}}$ over the variables in $\mathcal{B}$ such that $D_{\text{KL}[\text{new}]} < q$?

Note that this problem can be solved rather trivially by reconfiguring the CPTs such that $\Pr(\text{Pred}) = \Pr(\text{Obs})$ and thus $D_{\text{KL}[\text{new}]} = 0$. This has of course consequences for previous experiences—we are likely to induce unexplained past prediction errors. However, we cannot assume that we have access to (all) previous predictions and observations, making it close to impossible to minimize joint prediction error over all previous predictions and observations. As we do want to constrain the revisions in some way or another, we propose to revise the current model by allowing modification only of a *designated subset* of parameters in the model. So, we reformulate model revision to decide whether we can decrease D_{KL} by a change in a subset $\mathbf{p}$ of parameter probabilities in the network.[2] As in belief revision, we define two variants of the decision problem.

ModelRevision1
Instance: A Bayesian network $\mathcal{B}$ with designated variables Hyp and Pred, where $\Pr_{(\text{Pred})}$ denotes the posterior distribution over Pred; an observed distribution $\Pr_{(\text{Obs})}$ over Pred; a subset $\mathbf{P}$ of the parameter probabilities represented by $\Pr_{\mathcal{B}}$; a rational number q.
Question: Is there a combination of values $\mathbf{p}$ to $\mathbf{P}$ such that $D_{\text{KL}[\mathbf{p}]} < q$?

ModelRevision2
Instance: As in ModelRevision1.
Question: Is there a combination of values $\mathbf{p}$ to $\mathbf{P}$ such that $D_{\text{KL}} - D_{\text{KL}[\mathbf{p}]} > q$?

We will show that these problems are $\mathsf{NP}^{\mathsf{PP}}$-complete, that is, as least as hard as Partial MAP [24] and Parameter Tuning [21]. To prove $\mathsf{NP}^{\mathsf{PP}}$-hardness, we reduce from the following $\mathsf{NP}^{\mathsf{PP}}$-complete problem:

E-Majsat
Instance: A Boolean formula ϕ with n variables, partitioned into sets $\mathbf{X_E} = x_1, \ldots, x_k$ and $\mathbf{X_M} = x_{k+1}, \ldots, x_n$ for $1 \leq k \leq n$.
Question: Is there a truth assignment $\mathbf{x_E}$ to $\mathbf{X_E}$ such that the majority of truth assignments to $\mathbf{X_M}$ together with $\mathbf{x_E}$ satisfy ϕ?

[2] One of the anonymous made the interesting observation that *changing the network structure* (i.e., removing or adding arcs) can also be seen as model revision. We do not address that aspect here.

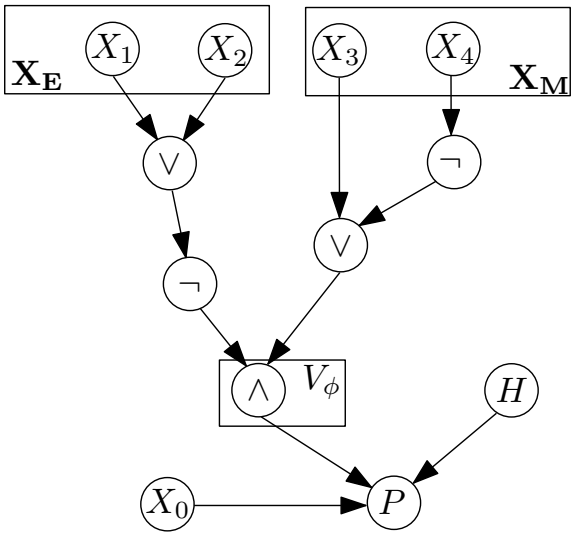

Fig. 3. The Bayesian network $\mathcal{B}_{\phi_{\text{ex}}}$ that is constructed from the E-MAJSAT example ϕ_{ex}.

We will use the following E-MAJSAT instance $(\phi_{\text{ex}}, \mathbf{X_E}, \mathbf{X_M})$ as a running example in the construction: $\phi_{\text{ex}} = (\neg(x_1 \vee x_2)) \wedge (\neg(x_3 \vee \neg x_4)), \mathbf{X_E} = \{x_1, x_2\}, \mathbf{X_M} = \{x_3, x_4\}$; note that this is a 'yes'-instance to E-MAJSAT: for $x_1 = x_2 = f$, three out of four truth assignments to $\mathbf{X_M}$ satisfy ϕ_{ex}.

We construct $\mathcal{B}_{\phi_{\text{ex}}}$ from ϕ_{ex} in a similar way as in the proof of Theorem 3, but we add another binary variable X_0 as an additional parent of P, with prior probability distribution $\Pr(X_0 = t) = 0$ (Figure 3). We define $\Pr(P \mid V_\phi, H, X_0)$ as follows:

$$\Pr(P{=}t \mid V_\phi, H, X_0) = \begin{cases} 3/8 \text{ if } V_\phi{=}t, H{=}X_0{=}t \\ 1/8 \text{ if } V_\phi{=}f, H{=}X_0{=}t \\ 0 \quad \text{otherwise} \end{cases}$$

We redefine $\Pr_{(\text{Hyp})}$ to $\Pr(H = t) = 1/2$ and $\Pr_{(\text{Obs})}$ to $\Pr(P = t) = 31/32$. In addition, we designate the sets of variables $\mathbf{X_E}$ and $\mathbf{X_M}$ in the network, and we set $\mathbf{P} = \mathbf{X_E} \cup \{X_0\}$. We set $q = q_1 = 1/8 \log(\frac{1/8}{31/32}) + 7/8 \log(\frac{7/8}{1/32})$ and $q = q_2 = 5 - 1/8 \log(\frac{1/8}{31/32}) - 7/8 \log(\frac{7/8}{1/32})$.
We now claim the following.

Theorem 4. MODELREVISION1 *and* MODELREVISION2 *are* $\mathsf{NP}^{\mathsf{PP}}$*-complete, even for singleton binary variables* Pred *and* Hyp.

Proof. Membership follows from the following algorithm: non-deterministically guess a combination of values $\mathbf{p}$ and compute the (change in) relative entropy.

This can be done in polynomial time using a non-deterministic Turing Machine with access to an oracle for problems in PP.

To prove $\mathsf{NP}^{\mathsf{PP}}$-hardness, we reduce ModelRevision from E-Majsat. Let $(\phi, \mathbf{X_E}, \mathbf{X_M})$ be an instance of Majsat and let $\mathcal{B}_\phi$ be the Bayesian network constructed from ϕ as described above. Observe that in $\mathcal{B}_\phi$, given the prior probability distribution of X_0, we have that $\Pr(P = t \mid V_\phi, H, X_0) = 0$ independent of the probability distribution of V_ϕ, and thus $D_{\mathrm{KL}} = 0 + \log(1/32) = 5$. If we revise the prior probability distribution of X_0, we observe that D_{KL} decreases when $\Pr(X_0 = t)$ goes to 1; $D_{\mathrm{KL}[\Pr(X_0=t)]}$ is minimal for $\Pr(X_0 = t) = 1$. In that case, for $\Pr(V_\phi) = p$, $\Pr(P = t \mid V_\phi, H, X_0) = (2p+1)/16$ and $D_{\mathrm{KL}[\Pr(X_0=t)=1]} = (2p+1)/16 \log(\frac{(2p+1)/16}{31/32}) + (15-2p)/16 \log(\frac{(15-2p)/16}{1/32})$.

For $\Pr(V_\phi) = 1/2$, $\Pr(P = t \mid V_\phi, H, X_0) = 1/8$ and $D_{\mathrm{KL}[\Pr(X_0=t)=1]} = 1/8 \log(\frac{1/8}{31/32}) + 7/8 \log(\frac{7/8}{1/32})$. We have in that case that $D_{\mathrm{KL}} - D_{\mathrm{KL}[\Pr(X_0=t)=1]} = 5 - 1/8 \log(\frac{1/8}{31/32}) - 7/8 \log(\frac{7/8}{1/32})$.

If there exists a truth assignment $\mathbf{x_E}$ to $\mathbf{X_E}$ such that the majority of truth assignments to $\mathbf{X_M}$ satisfies ϕ, then there exists a combination of values $\mathbf{p}$ to $\mathbf{P} = \mathbf{X_E} \cup \{X_0\}$ such that $\Pr(V_\phi) > 1/2$ and thus $D_{\mathrm{KL}[\Pr(X_0=t)=1]} < q_1$ and $D_{\mathrm{KL}} - D_{\mathrm{KL}[\Pr(X_0=t)=1]} > q_2$; namely, the combination of values to $\mathbf{X_E}$ that sets $\Pr(X_i = t)$ to 1 if $X_i \in \mathbf{X_E}$ is set to t, and $\Pr(X_i = t)$ to 0 if $X_i \in \mathbf{X_E}$ is set to f, together with setting $\Pr(X_0 = t)$ to 1. Vice versa, if we can revise $\mathbf{P}$ such that $D_{\mathrm{KL}[\Pr(X_0=t)=1]} < q_1$ and that $D_{\mathrm{KL}} - D_{\mathrm{KL}[\Pr(X_0=t)=1]} > q_2$, then there exists a truth assignment $\mathbf{x_E}$ to $\mathbf{X_E}$ such that the majority of truth assignments to $\mathbf{X_M}$ satisfies ϕ, namely, the truth assignment that sets $X_i \in \mathbf{X_E}$ to t if $\Pr(X_i = t) \geq 1/2$ and to f otherwise.

Given that the reduction can be done in polynomial time, this proves $\mathsf{NP}^{\mathsf{PP}}$-completeness of both ModelRevision1 and ModelRevision2. □

6 Adding Additional Observations to the Model

Apart from revising the probability distribution of the hypotheses and from revising the parameters in the model, we can also lower relative entropy by some action that influences either the outside world or our perception of it. By observing previously unobserved variables in the model (i.e., changing our perception of the world), the posterior probability of the *prediction* can be influenced; similarly, we can *intervene* in the outside world, thus influencing the posterior probability over the *observation*. In both cases, we will need to decide on *which* observations to gather, respectively *which* variables to intervene on. Again we assume that the set of allowed observations, respectively interventions, is designated. We will first focus on the question which candidate observations to make. As in the previous two problems, we formulate two decision problems that capture this question.

AddObservation1
Instance: A Bayesian network $\mathcal{B}$ with designated variables Hyp and Pred, where $\Pr_{(\mathrm{Pred})}$ denotes the posterior distribution over Pred; an observed distribution

$\Pr_{(\text{Obs})}$ over Pred; and rational number q. Let $\mathbf{O} \subseteq \text{Int}$ denote the set of observable variables in $\mathcal{B}$.
Question: Is there a joint value assignment $\mathbf{o}$ to $\mathbf{O}$ such that $D_{\text{KL}[\mathbf{o}]} < q$?

AddObservation2
Instance: As in AddObservation1.
Question: Is there a joint value assignment $\mathbf{o}$ to $\mathbf{O}$ such that $D_{\text{KL}} - D_{\text{KL}[\mathbf{o}]} > q$?

While these problems are *conceptually* different from the ModelRevision problems, from a *complexity* point of view they are very similar: the effect of setting a prior probability of a variable X_i in the proof construct to 1, and observing its value to be t, are identical; the same holds for setting it to 0, respectively observing its value to be f. This allows us to prove $\mathsf{NP}^{\mathsf{PP}}$-completeness of AddObservation using essentially the same construct as in the proof of Theorem 4; however, we must take care that the prior probability distribution of X_0 is such that no inconsistencies in the network emerge as a result of observing its value to t. In particular, if $\Pr(X_0 = t) = 0$, then we cannot observe X_0 to be t without creating an inconsistency in the network.

So, we redefine $\Pr(X_0 = t) = 1/2$; now, $\Pr(P = t \mid V_\phi, H, X_0)$ (and thus also D_{KL}) becomes dependent of the probability distribution of V_ϕ. In particular, for $\Pr(V_\phi) = p$ we have that $\Pr(P = t \mid V_\phi, H, X_0) = (2p+1)/32$ and consequently, $D_{\text{KL}} = (2p+1)/32 \log(\frac{(2p+1)/32}{31/32}) + (31-2p)/32 \log(\frac{(31-2p)/32}{1/32})$. We therefore redefine $q_2 = 1/16 \log(\frac{1/16}{31/32}) + 15/16 \log(\frac{15/16}{1/32}) - q_1 = 1/16 \log(\frac{1/16}{31/32}) + 15/16 \log(\frac{15/16}{1/32}) - 1/8 \log(\frac{1/8}{31/32}) - 7/8 \log(\frac{7/8}{1/32})$. We set $\mathbf{O} = \mathbf{X_E} \cup \{X_0\}$.

Theorem 5. AddObservation1 *and* AddObservation2 *are* $\mathsf{NP}^{\mathsf{PP}}$*-complete.*

Proof. Membership follows from a similar argument as for ModelRevision. To prove $\mathsf{NP}^{\mathsf{PP}}$-hardness, we again reduce from E-Majsat. Let $(\phi, \mathbf{X_E}, \mathbf{X_M})$ be an instance of E-Majsat and let $\mathcal{B}_\phi$ be the Bayesian network constructed from ϕ as described above. The probability distribution $\Pr(P = t \mid V_\phi, H, X_0)$ depends as follows on the observed value of X_0: $\Pr(P = t \mid V_\phi, H, X_0 = t) = (2p+1)/16$ and $\Pr(P = t \mid V_\phi, H, X_0 = f) = 0$. In particular, if $\Pr(V_\phi) > 1/2$, then $\Pr(P = t \mid V_\phi, H, X_0 = t) > 1/8$ and hence $D_{\text{KL}[X_0=t]} < 1/8 \log(\frac{1/8}{31/32}) + 7/8 \log(\frac{7/8}{1/32})$. Similarly, $\Pr(P = t \mid V_\phi, H, X_0 = f) = 0$ and hence $D_{\text{KL}[X_0=f]} = 5$. So, only if X_0 is observed to be t and $\Pr(V_\phi) > 1/2$ we have that $D_{\text{KL}[X_0=t]} < q_1$ and $D_{\text{KL}} - D_{\text{KL}[X_0=t]} > q_2$.

If there exists a truth assignment $\mathbf{x_E}$ to $\mathbf{X_E}$ such that the majority of truth assignments to $\mathbf{X_M}$ satisfies ϕ, then there exists a joint value assignment to $\mathbf{O} = \mathbf{X_E} \cup \{X_0\}$ such that $\Pr(V_\phi) > 1/2$ and $D_{\text{KL}[\mathbf{o}]} < q_1$ and that $D_{\text{KL}} - D_{\text{KL}[\mathbf{o}]} > q_2$. Namely, the joint value assignment that sets X_0 to t and sets the variables in $\mathbf{X_E}$ according to $\mathbf{x_E}$. And vice versa, if there exists a joint value assignment $\mathbf{o}$ to $\mathbf{O}$ such that $D_{\text{KL}[\mathbf{o}]} < q_1$ and $D_{\text{KL}} - D_{\text{KL}[\mathbf{o}]} > q_2$, then there is a truth assignment to $\mathbf{X_E}$ such that the majority of truth assignments to $\mathbf{X_M}$ satisfy ϕ, namely, the truth assignment that sets $X_i \in \mathbf{X_E}$ to t if $X_i \in \mathbf{o}$ is observed as t, and to f

otherwise. As this reduction can be done in polynomial time, this proves that AddObservation1 and AddObservation2 are $\mathsf{NP}^{\mathsf{PP}}$-complete. □

7 Intervention in the Model

We can bring prediction and observation closer to each other by changing our prediction (by influencing the posterior distribution of the prediction by revision of beliefs, parameters, or observing variables), but also by what in the HPC framework is called *active inference*: actively changing the causes of the observation to let the observation ("the real world") match the prediction ("the model of the world"). This is a fundamental aspect of the theory, which is used to explain how a desire of moving one's arm—i.e., the expectation or prediction that one's arm will be in a different position two seconds from now—can yield actual motor acts that establish the desired movement. We implement this as *intervention* in the Bayesian framework, and the problem that needs to be resolved is to decide *how* to intervene.

The predicted result of an action of course follows from the generative model, which represents how (hypothesized) causes generate (predicted) effects, for example, how motor commands sent to the arm will change the perception of the arm. So, from a computational point of view, the decision variants of the Intervention problem are identical to the decision variants of the Observation problem:

Intervention1
Instance: A Bayesian network $\mathcal{B}$ with designated variables Hyp and Pred, where $\Pr_{(\text{Pred})}$ denotes the posterior distribution over Pred; an observed distribution $\Pr_{(\text{Obs})}$ over Pred; and rational number q. Let $\mathbf{A} \subseteq$ Int denote the set of intervenable variables in $\mathcal{B}$.
Question: Is there a joint value assignment $\mathbf{a}$ to $\mathbf{A}$ such that $D_{\text{KL}[\mathbf{a}]} < q$?

Intervention2
Instance: As in Intervention1.
Question: Is there a joint value assignment $\mathbf{a}$ to $\mathbf{A}$ such that $D_{\text{KL}} - D_{\text{KL}[\mathbf{a}]} > q$?

Corollary 1. Intervention1 *and* Intervention2 *are* $\mathsf{NP}^{\mathsf{PP}}$*-complete.*

8 Parameterized Complexity

What situational constraints can render the computations tractable? From the intractability proofs above we can already infer what *does not* make prediction error minimization tractable. Even for binary variables, singleton hypothesis and prediction nodes, and at most three incoming arcs per variable, all problems remain intractable. It is easy to show that ModelRevision, AddObservation,

and INTERVENTION remain PP-hard when there is just a single designated parameter, observable or intervenable variable. The complexity of these problems is basically in the *context* that modulates the relation between hypothesis and prediction.

ADDOBSERVATION and INTERVENTION are fixed-parameter tractable for the parameter set {treewidth of the network, cardinality of the variables, size of Pred} plus the size of $\mathbf{O}$, respectively $\mathbf{A}$. In that case, the computation of D_{KL} is tractable, and we can search joint value assignments to $\mathbf{O}$, respectively $\mathbf{A}$ exhaustively. Similarly, when the computation of D_{KL} is tractable, one can use parameter tuning algorithms to decide MODELREVISION and BELIEFREVISION; these problems are fixed-parameter tractable for the parameter set {treewidth of the network, cardinality of the variables, size of Pred} plus the size of $\mathbf{P}$, respectively Hyp [5].

9 Conclusion

Hierarchical Predictive Coding (HPC) is an influential unifying theory in theoretical neuroscience, proposing that the brain continuously makes Bayesian predictions about future states and uses the prediction error between prediction and observation to update the hypotheses that drove the predictions. In this paper we studied HPC from a computational perspective, formalizing the conceptual ideas behind hypothesis updating, model revision, and active inference, and studying the computational complexity of these problems. Despite rather explicit claims on the contrary (e.g., [7, p.191]), we show that the Bayesian computations that underlie the error minimization mechanisms in HPC are *not* computationally tractable in general, even when hypotheses and predictions are constrained to binary singleton variables. Even in this situation, rich contextual modulation of the dependences between hypothesis and prediction may render successful updating intractable. Further constraints on the structure of the dependences (such as small treewidth and limited choice in which parameters to observe or observations to make) are required.

In this paper, we focused on computations within a particular level of the hierarchy and on error minimization. There is more to say about the computations that are postulated within HPC, for example how increasingly rich and complex knowledge structures are *learned* from prediction errors. We leave that for further research.

References

1. A. M. Abdelbar and S. M. Hedetniemi. Approximating MAPs for belief networks is NP-hard and other theorems. *Artificial Intelligence*, 102:21–38, 1998.
2. T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.
3. M. Blokpoel, J. Kwisthout, and I. van Rooij. When can predictive brains be truly Bayesian? *Frontiers in Theoretical and Philosophical Psychology*, 3:406, 2012.

4. E. Castillo, J.M. Gutiérrez, and A.S. Hadi. Sensitivity analysis in discrete Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 27:412–423, 1997.
5. H. Chan and A. Darwiche. Sensitivity analysis in Bayesian networks: From single to multiple parameters. In *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*, pages 67–75, 2004.
6. A. Clark. The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Theoretical and Philosophical Psychology*, 4:e270, 2013.
7. A. Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.
8. G. F. Cooper and E. Herskovitz. Determination of the entropy of a belief network is NP-hard. Technical Report KSL-90-21, Stanford University. Computer Science Dept. Knowledge Systems Laboratory, March 1990.
9. P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153, 1993.
10. A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. CU Press, Cambridge, UK, 2009.
11. A. Darwiche and A. Choi. Same-decision probability: A confidence measure for threshold-based decisions under noisy sensors. In *5th European Workshop on Probabilistic Graphical Models*, 2010.
12. R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, Berlin, 1999.
13. K.J. Friston. The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009.
14. K.J. Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.
15. K.J. Friston, J. Daunizeau, J. Kilner, and S.J. Kiebel. Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102(3):227–260, 2010.
16. T.L. Griffiths, N. Chater, C. Kemp, A. Perfors, and J.B. Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010.
17. J. Hohwy. *The Predictive Mind*. Oxford University Press, 2013.
18. J. M. Kilner, K. J. Friston, and C. D. Frith. The mirror-neuron system: A Bayesian perspective. *Neuroreport*, 18:619–623, 2007.
19. D. Knill and A. Pouget. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience*, 27(12):712–719, 2004.
20. J. Kwisthout. *The Computational Complexity of Probabilistic Networks*. PhD thesis, Faculty of Science, Utrecht University, The Netherlands, 2009.
21. J. Kwisthout and L.C. van der Gaag. The computational complexity of sensitivity analysis and parameter tuning. In D.M. Chickering and J.Y. Halpern, editors, *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 349–356. AUAI Press, 2008.
22. J. Kwisthout and I. Van Rooij. Predictive coding: Intractability hurdles that are yet to overcome [abstract]. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society., 2013.
23. J.-C. Park, J.H. Lim, H. Choi, and D.-S. Kim. Predictive coding strategies for developmental neurorobotics. *Frontiers in Psychology*, 3:134, 2012.
24. J.D. Park and A. Darwiche. Complexity results and approximation settings for MAP explanations. *Journal of Artificial Intelligence Research*, 21:101–133, 2004.

25. A.P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, and C. Frith. The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, 7(4):413–422, 2012.
26. A.K. Seth, K. Suzuki, and H.D. Critchley. An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2:e395, 2011.
27. L.C. van der Gaag, H.L. Bodlaender, and A.J. Feelders. Monotonicity in Bayesian networks. In M. Chickering and J. Halpern, editors, *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pages 569–576. Arlington: AUAI press, 2004.