Secondary Publication

Tull, Sean; Kleiner, Johannes; Smithe, Toby St Clere

Active Inference in String Diagrams : A Categorical Account of Predictive Processing and Free Energy

Date of secondary publication: 05.12.2025

Submitted Version (Preprint), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-112054x

Primary publication

Tull, Sean; Kleiner, Johannes; Smithe, Toby St Clere (2023): Active Inference in String Diagrams : A Categorical Account of Predictive Processing and Free Energy, in: arXiv, pp. 1–41, doi: 10.48550/arxiv.2308.00861.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

arXiv:2308.00861v1 [math.CT] 1 Aug 2023

Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy

Sean Tull[1,2], Johannes Kleiner[2,3,4], and Toby St Clere Smithe[5,6]

[1]Quantinuum
[2]Association for Mathematical Consciousness Science
[3]Ludwig Maximilian University of Munich
[4]University of Bamberg
[5]Topos Institute
[6]VERSES Research

Abstract

We present a categorical formulation of the cognitive frameworks of Predictive Processing (PP) and Active Inference, expressed in terms of string diagrams interpreted in a monoidal category with copying and discarding. This includes diagrammatic accounts of generative models, Bayesian updating, perception, planning, active inference, and free energy. In particular we present a diagrammatic derivation of the formula for active inference via free energy minimisation, and establish a compositionality property for free energy, allowing free energy to be applied at all levels of an agent's generative model. Aside from aiming to provide a helpful graphical language for those familiar with active inference, we conversely hope that this article may provide a concise formulation and introduction to the framework.

1 Introduction

Predictive processing (PP) is a framework for modelling cognition and adaptive behaviour in both biological and artificial systems [WM17, Hoh20]. A prominent sub-field is the programme of *Active Inference*, developed by Friston and collaborators [SFW22, PPF22, FFR$^+$17, SBPF21], which aims to provide a unified understanding of cognition and action which can be applied at many levels, from a single neuron to an entire brain or organism. More specifically, active inference gives a proposal for how a cognitive agent represents its own beliefs about the world, how it updates these beliefs in light of new observations, and how it chooses the actions it takes, with the latter ultimately leading to new observations.

Central to the framework is that an agent possesses a *generative model* which explains its observations causally in terms of both hidden states of the world and its own actions. Note that this model is internal to the agent, and typically distinct from the 'true' causal process in the world which produces the observations. After receiving an observation, the agent may *update* this generative model to determine likely hidden states which caused the observation (the process of perception) and choose its actions (the process of planning). In active inference, both forms of updating are carried out together through a form of approximate Bayesian inference, by minimising a quantity known as *free energy* [FKH06, Fri10].

While active inference seeks a principled account of cognition, at present its formalisation can seem fairly complex, and there are various aspects which do not follow immediately from simply applying the definitions to a given generative model. Conceptually clear formal accounts of the framework would be desirable to simplify the theory and address these issues, as well as for applications within AI.

One hope for such a formal account would be for it to be both *compositional* and *graphical*. Indeed the generative models in PP are highly structured, often given as 'hierarchical models' [DVF17] which are best represented diagrammatically in terms of probabilistic graphical models such as Bayesian networks. While there has been support for, and steps towards, a graphical account of active inference [FPdV17], so far the graphical aspects only formally describe the structure of a generative model, while other aspects such as

updating and free energy are still treated through traditional probabilistic calculations, and only informally in diagrams.

Recently however, fully formal diagrammatic methods have been developed for both describing Bayesian networks and carrying out probabilistic reasoning about them. These approaches are based on (monoidal) *category theory* and its associated graphical language of *string diagrams* [PZ23]. Category theory has been applied across the sciences as a general mathematics of interacting *processes*, including within probability theory [CS12, CJ19], causality [JKZ19, FK23, LT23], game theory [GHWZ18], machine learning [FST19, SGW21], quantum computing [AC04] and natural language processing [CCS08]. In particular a major ongoing development is in the study of probabilistic processes in terms of *cd-categories* (and 'Markov categories'), which allow one to carry out probabilistic reasoning entirely through string diagrams [Fri20].

In this work we give a full categorical account of predictive processing and active inference in terms of string diagrams, interpreted in cd-categories. In doing so we aim to give a conceptually clear account of the main features of the framework: generative models, Bayesian updating (including with soft observations), perception, action planning, and their combination in active inference, and both *variational* and *expected* free energy.

A highlight is a fully graphical derivation of the well-known formula for active inference in terms of minimisation of free energy. While this is a central result within active inference, its usual justification is more heuristic in nature. Here we instead derive the free energy formula purely graphically from a diagrammatic account of active inference itself, providing what we argue is the most transparent account of this result known so far.

The categorical perspective also naturally leads us to consider more novel aspects of active inference. These include the definition of *open generative models* (essentially from [LT23]) which are generative models coming with 'inputs', allowing them to serve as the building blocks of an overall generative model.

We also introduce a notion of variational free energy for open models which allows us to establish the desirable property that free energy is *compositional*. Namely, a system with an overall generative model composed from sub-models may minimise global VFE by minimising VFE locally within each component. This is a crucial fact in order to apply free energy as proposed to all levels of a system, say from a whole brain down to its individual neurons.

Overall, we hope that our diagrammatic accounts of PP can provide a conceptually clear view of the framework, and also a natural language for reasoning within it. Indeed, as argued for example in [LT23] and elsewhere [JKZ19] diagrams in cd-categories provide a natural way to both represent causal (generative) models, as well as reason about them. As we demonstrate here they are also natural for describing the structure of active inference, including free energy. Aside from aiming to provide a helpful graphical language for those familiar with active inference, we conversely hope that this article may provide a succinct introduction to PP for those already familiar with string diagrams and categorical reasoning.

Further motivations Though primarily a framework for cognition, various proposals have been put forward for how predictive processing may be related to *consciousness* [WM17]. In previous work, two of the authors developed a categorical account of the *Integrated Information Theory* of consciousness, again essentially using cd-categories [TK21, KT21] and based on the work here we hope to give a categorical account of how consciousness may be accounted for within PP [Dea21, HS20]. We also see this work as a piece of the programme of *Compositional Intelligence*, which explores how categorically structured models and processes can be applied to (artificial) intelligence. Specifically, PP may be seen as a proposal for how compositional intelligence manifests in biology; that is, how biological systems may employ compositionality to carry out intelligent and adaptive behaviour.

Active inference can also be understood as an alternate proposal to reinforcement learning (RL) for how agents can learn adaptive behaviour, and shares similar features including the role of probabliistic models and inference [TMSB20]. It differs from conventional RL by replacing an explicit reward function with the aim of maximizing evidence for a probabilistic model, where the agent's preferences are now encoded in the model's prior distribution [FDK09].

Related work This work can be seen as a part of the growing field of 'categorical cybernetics' [Smi21b, CGHR21], including previous work from one of the authors on compositional accounts of Bayesian updating [Smi20] and of active inference in terms of 'statistical games' [Smi21a, Smi22]. It differs from previous

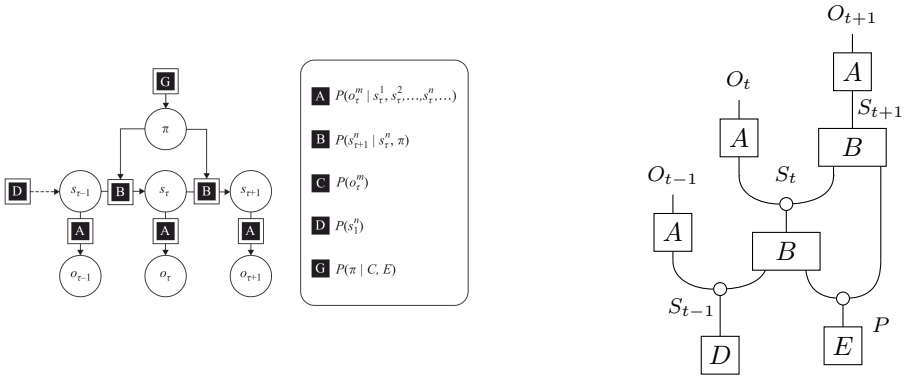


Figure 1:
A generative model diagram from the recent book *Active Inference* by Parr, Pezzulo and Friston [PPF22] (left) and the equivalent string diagram representation (right) (though replacing the informal EFE term G with the prior E).

works by directly formalising the active inference framework itself, and by working explicitly graphically within the simple string-diagrammatic setting of cd-categories, with the aim of supplying a simple abstract characterization of active inference agents.

In this way the work is a part of a general movement in applying string diagrams in cd-categories to probability theory and causal reasoning. A categorical account of Bayesian inversion was first given by Coecke and Spekkens in [CS12], and then within cd-categories by Cho and Jacobs [CJ19], with further developments in categorical probability by Fritz [Fri20]. Our diagrammatic account of generative models is precisely that given for causal models in part by one of the authors in [LT23], which builds on the earlier categorical treatments of (causal) Bayesian networks by Fong [Fon13], Jacobs et al [JKZ19] and others e.g. [FK23]. Indeed, as an agent's explanation for the observations it receives from the world, a generative model is ultimately a causal model [Pea09], though this is not often stressed in the literature.

The two forms of soft Bayesian updating treated here we first studied by Jacobs in [Jac19]. The specific treatment of conditioning in cd-categories used here is from [LT23]. Cd-categories with (non-unique) conditionals have also been recently studied as 'partial Markov categories' in [DLR23], along with both notions of updating. Our treatment of free energy refers to the KL divergence of distributions; we note that an axiomatic treatment of Markov categories coming with divergences on their morphisms has recently been given by Perrone in [Per22].

Within active inference itself, graphical aspects have been increasingly prominent, with discussion of the 'graphical brain' in [FPdV17]. In such works it is argued that one may describe models as (non-directed) Forney Factor Graphs (FFGs) [DVF17]. However, generative models are inherently directed, going from states to observations with the other direction being intractable to compute exactly. Thus it is more natural to treat models using (generalisations of) directed Bayesian networks. Nonetheless we note though that FFGs derived from a model still have a role when minimising VFE via 'message passing' algorithms [PMKF19].

Interestingly, while a Bayesian network is typically depicted as a DAG (with only the variables labelled), one may argue the diagrams in active inference have been naturally 'converging' on their string diagrammatic representation, which also includes labels on the channels; see Figure 1. We claim that the advantage of string diagrams beyond DAGs is in allowing one to both represent and reason about the model in the same formalism.

We note that the diagrams in PP are at times only semi-formal, including aspects such as the free energy which are not strictly part of the generative model. One may see this work as a step towards the shared goal of representing formally all aspects of PP within one language of diagrams.

Structure of the article We begin in Section 2 by introducing cd-categories and their diagrammatic account of probability theory. We then apply these to introduce from scratch the key aspects of PP: generative models as Bayesian networks, and their generalisation to open generative models in a cd-category (Section 3),

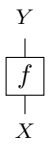
(Bayesian) updating of generative models from observations (Section 4), perception and planning (Section 5) and their combination in exact active inference (Section 6). We then discuss free energy (Section 7) and give a graphical derivation of active inference via free energy minimisation (Section 8). In Section 9 we then introduce free energy for open models using a graphical formalism of 'log-boxes' and use this to establish the compositionality property of free energy. Finally we discuss future work in Section 10.

Acknowledgements We thank Robin Lorenz for helpful discussions and development of the treatment of causal models used here. This research was supported by grant number FQXi-RFP-CPW-2018 from the Foundational Questions Institute and Fetzer Franklin Fund, a donor advised fund of the Silicon Valley Community Foundation. Sean Tull would also like to thank Quantinuum for their generous support in this research. Johannes Kleiner would like to thank the Mathematical Institute of the University of Oxford for hosting him while working on this research.

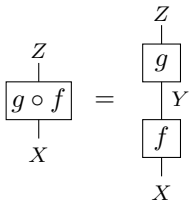
2 Categorical Setup

Let us begin by introducing the graphical treatment of probabilistic processes in terms of *string diagrams*, developed by numerous authors [CS12, CJ19, Fri20]. Formally, these correspond to working in a 'monoidal category' or more specifically a 'cd-category', but in practice one may avoid mathematical details and simply work with the diagrams themselves. Though cd-categories are very general, in this article it suffices to consider the category $\mathbf{Mat}_{\mathbb{R}^+}$ of $\mathbb{R}^+$ valued finite matrices, introduced shortly in Example 1.

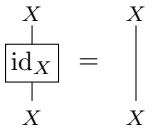
A *category* $\mathbf{C}$ consists of a collection of *objects* $X, Y, \dots$ and *morphisms* or *processes* $f\colon X \to Y$ between them, which we can compose in sequence. In string diagrams we depict an object X as a wire and a morphism $f\colon X \to Y$ as a box with lower input wire X and upper output wire Y, read from bottom to top.



Given another morphism $g\colon Y \to Z$ we can compose them to yield a morphism $g \circ f\colon X \to Z$, depicted as:



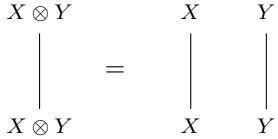
Each object X also comes with an *identity* morphism $\mathrm{id}_X\colon X \to X$ depicted as a blank wire:



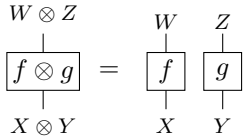
The identity leaves any morphism alone under composition, that is $\mathrm{id}_Y \circ f = f = f \circ \mathrm{id}_X$ for any $f\colon X \to Y$.

Formally, a *symmetric monoidal category* $(\mathbf{C}, \otimes, I)$ is a category $\mathbf{C}$ with a functor $\otimes\colon \mathbf{C} \times \mathbf{C} \to \mathbf{C}$, and natural transformations which express that $\otimes$ is suitably associative and symmetric, with a distinguished unit object I [Coe06]. All of these aspects however are expressed most simply in diagrams.

Firstly, the *tensor* $\otimes$ allows us to compose any pair of objects X, Y into an object $X \otimes Y$, depicted by placing wires side-by-side.

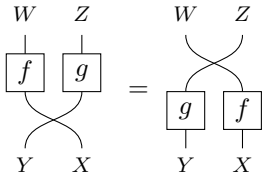


Given morphisms $f\colon X \to W$ and $g\colon Y \to Z$ we can similarly form their 'parallel composite' $f \otimes g\colon X \otimes Y \to W \otimes Z$ as below.

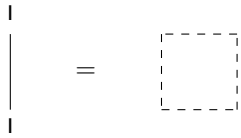


In text we will at times omit the tensor symbols and write e.g. 'f from X to Y, Z' or $f\colon X \to Y, Z$ in place of $f\colon X \to Y \otimes Z$.

The tensor is symmetric so we can 'swap' pairs of wires past each other, such that swapping twice returns the identity, and boxes carry along the swaps as below.

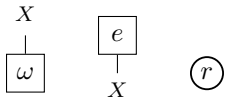


We also have a distinguished *unit object* I whose identity morphism we depict simply as empty space, and denote by 1.



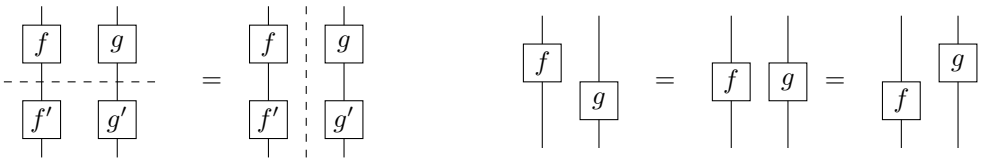
(1)

Intuitively, tensoring any object by the unit simply leaves it invariant. The unit allows us to consider morphisms with 'no inputs' and/or 'no outputs' in diagrams. A morphism $\omega\colon \mathsf{I} \to X$ is called a *state* of X, depicted with no input. An *effect* on X is a morphism $e\colon X \to \mathsf{I}$, depicted with no output. A morphism $r\colon \mathsf{I} \to \mathsf{I}$, drawn with no input or output, is called a *scalar*.



In particular the 'empty space' diagram (1) is the scalar $1 = \mathrm{id}_{\mathsf{I}}$.

The compositions $\circ, \otimes$ satisfy axioms which must be considered when working symbolically but are trivial in the graphical language. An example is associativity of composition $(h \circ g) \circ f = h \circ (g \circ f)$, which is automatic from simply drawing three boxes in sequence on the same wire. Similarly the rule $(f \otimes g) \circ (f' \otimes g') = (f \circ f') \otimes (g \circ g')$, displayed in the left identity below, and the 'interchange law' $(f \otimes \mathrm{id}) \circ (\mathrm{id} \otimes g) = f \otimes g = (\mathrm{id} \otimes g) \circ (f \otimes \mathrm{id})$, displayed in the right identity below, amount to letting us freely slide boxes along wires.

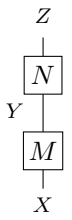


Let us now introduce our primary example category in this article.

Example 1. *In the category* $\mathbf{Mat}_{\mathbb{R}^+}$ *of positive valued matrices, the objects are finite sets* $X, Y, \ldots$ *and the morphisms* $M\colon X \to Y$ *are functions* $M\colon X \times Y \to \mathbb{R}^+$ *where* $\mathbb{R}^+ := \{r \in \mathbb{R} \mid r \geq 0\}$*. Equivalently such a function is given by an '*$X \times Y$ *matrix' with entries* $M(y \mid x) := M(x, y) \in \mathbb{R}^+$ *for* $x \in X$*,* $y \in Y$*.*

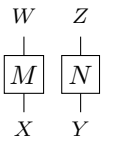
$$:: \ (x, y) \ \mapsto \ M(y \mid x)$$

Composition of $M\colon X \to Y$ and $N\colon Y \to Z$ is given by summation over Y:



$$:: \ (x,z) \ \mapsto \ \sum_{y \in Y} N(z \mid y) M(y \mid x)$$

The tensor $\otimes$ is given on objects by the Cartesian product $X \otimes Y = X \times Y$, and on morphisms by the Kronecker product, i.e. the usual tensor product of matrices:



$$:: \ ((x,y),(w,z)) \ \mapsto \ M(w \mid x) N(z \mid y)$$

The symmetry is simply the isomorphism $X \times Y \simeq Y \times X$. The unit object $\mathsf{I} = \{\star\}$ is the singleton set. A state of X is then equivalent to a positive function on X:

$$:: \ x \ \mapsto \ \omega(x)$$

where $\omega(x) := \omega(x \mid \star)$. Similarly, an effect e on X is also equivalent to a positive function on X via $e(x) := e(\star \mid x)$.

$$:: \ x \ \mapsto \ e(x)$$

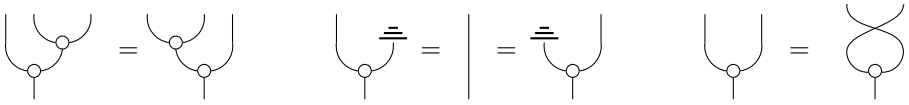
Finally, a scalar is precisely a positive real $r \in \mathbb{R}^+$.

2.1 Cd-categories

Many aspects of probability theory can be treated entirely diagrammatically, by noting that categories such as $\mathbf{Mat}_{\mathbb{R}^+}$ come with the following further structure.

Definition 2. *[CJ19] A* cd-category *(*copy-discard category*) is a symmetric monoidal category in which each object comes with a specified pair of morphisms*

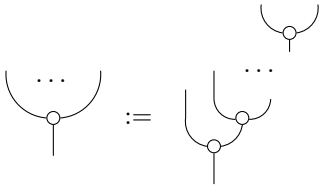
called copying *and* discarding*, respectively, which satisfy the following:*



The choice of these morphisms is moreover 'natural' in that the following hold for all objects X, Y.

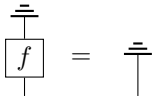
(2)

Thanks to these axioms for copying, we can unambiguously define a copying morphism with n output legs, for any $n \geq 1$, via:

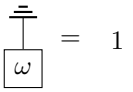


with the $n = 0$ case defined to be discarding $\overline{\top}$.

The presence of discarding allows us to identify the truly 'probabilistic' processes in a cd-category. We say that a morphism f is a *channel* when it preserves discarding, as below.



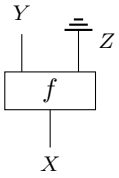
In particular, we call a state ω *normalised* when the following holds. For an explanation of why this gives the usual definition, cf. Example 3 below.



Here we will often call a normalised state ω of X a *distribution* of X, even when working in a general cd-category [1]. We also call a normalised state of $X \otimes Y$ a *joint distribution* over X, Y.

A cd-category in which every morphism is a channel, or equivalently $\overline{\top}$ is the unique effect on any object, is called a *Markov category* [Fri20]. Given any cd-category $\mathbf{C}$, its subcategory $\mathbf{C}_{\text{channel}}$ of channels always forms a Markov category.

Discarding allows us to 'ignore' certain outputs of a process. Given any morphism f from X to Y, Z, its *marginal* $X \to Y$ is the following morphism:

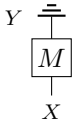


Let us see how these features describe discrete probability theory within our example category.

Example 3. $\mathbf{Mat}_{\mathbb{R}^+}$ *is a cd-category. Copying on* X *is given* $\curlyvee(y, z \mid x) = \delta_{x,y,z}$ *with value* 1 *iff* $x = y = z$ *and* 0 *otherwise. Discarding* $\overline{\top}$ *on* X *is given by the function with* $x \mapsto 1$ *for all* $x \in X$*. Hence a state* ω *is normalised, i.e. forms a distribution on* X*, precisely when it forms a* probability distribution *over* X *in the usual sense, i.e. its values sum to* 1*.*

$$= \sum_{x \in X} \omega(x) \quad = \quad 1$$

More generally, a process $M \colon X \to Y$ *is a channel iff it forms a* probability channel*, or equivalently a* stochastic *matrix, meaning that it sends each* $x \in X$ *to a normalised distribution* $M(y \mid x)$ *over* Y*. Indeed we have that:*



$$:: \quad x \;\mapsto\; \sum_{y} M(y \mid x)$$

[1] This is to avoid confusion with the usual use of the term (hidden) 'state' in PP.

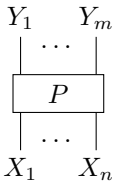
Hence M is a channel iff this effect is constant at 1, i.e. for all x we have

$$\sum_{y \in Y} M(y \mid x) = 1$$

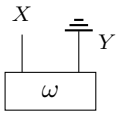
In typical probability theory, such a channel is also often called a 'conditional probability distribution' $P(Y \mid X)$ with values denoted $P(y \mid x) := P(Y = y \mid X = x)$ for $x \in X$, $y \in Y$. The subcategory of channels in $\mathbf{Mat}_{\mathbb{R}^+}$ is the Markov category $\mathbf{FStoch}$ of finite Stochastic matrices.

Let us see how a few features of probability theory appear in diagrams. Firstly, for any X, Y, a distribution ω on $X \otimes Y$ corresponds to a joint distribution over X, Y (left-hand below). In particular given a pair of distributions ϕ, σ over X, Y, the distribution $\phi \otimes \sigma$ corresponds to the resulting product distribution over $X \times Y$, with X and Y independent from each-other (right-hand below).

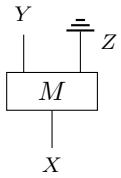
A general channel as below represents a probability channel $P(Y_1, \ldots, Y_m \mid X_1, \ldots, X_n)$.



Marginalisation of any morphism corresponds to the usual notion in probability theory, given by summation over the discarded object.



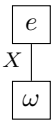
$:: x \mapsto \sum_{y \in Y} \omega(x, y)$



$:: (x, y) \mapsto \sum_{z \in Z} M(y, z \mid x)$

Finally we observe that for any effect $e \colon X \to \mathbb{R}^+$ and distribution ω the scalar $e \circ \omega$ corresponds to the expectation value of the function e according to the probability distribution ω.

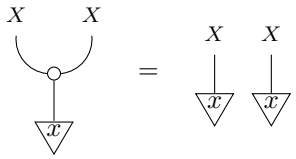
$$\mathop{\mathbb{E}}_{x \sim \omega} e(x) \quad = \quad$$



$$= \quad \sum_{x \in X} e(x)\omega(x)$$

2.2 Sharp states and caps

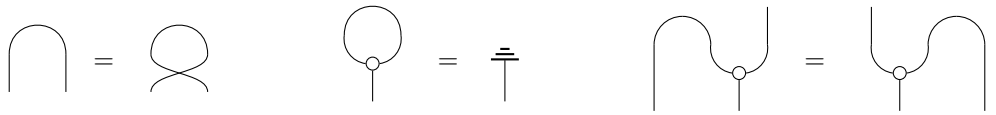
The copying morphisms in a cd-category allow us to identify those states which are really 'deterministic' [Fri20]. We call a state x *sharp*, and depict it with a triangle as below, when it is copied by $\curlyvee$, that is:



(3)

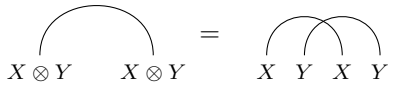
In many categories there is a corresponding effect for each state, playing an important role for sharp states, thanks to the following feature. We say that $\mathbf{C}$ has *caps* when each object comes with a distinguished effect

on $X \otimes X$ depicted $\cap$ and satisfying:



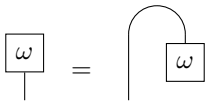
(4)

and such that the following holds for all objects X, Y:



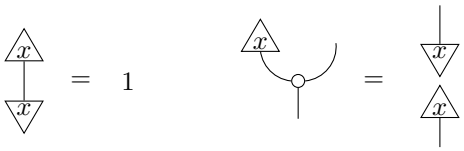
Intuitively, the cap is an effect which checks if its two input wires are 'in the same state'. The first equation in (4) expresses that this comparison is symmetric, and the remaining two that it is compatible with copying; for example the second says that each input when copied is equal to its copy.

Practically, caps allow us to 'turn outputs into inputs'. In particular, for each state ω we can define a corresponding effect by 'flipping ω upside-down':



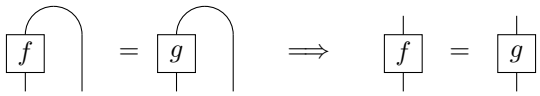
(5)

When $\omega = x$ is a sharp state, we call this effect *sharp* also. One may verify that it is the unique effect satisfying the following.



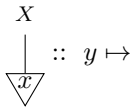
(6)

Caps are particularly useful in diagrammatic reasoning when they are *cancellative*, meaning that:



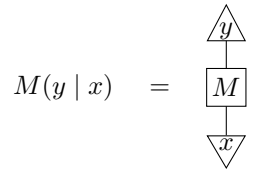
for all morphisms f, g.

Example 4. $\mathbf{Mat}_{\mathbb{R}^+}$ *has cancellative caps. Each point $x \in X$ corresponds to a normalised sharp state on X which we again denote by x, given by the point probability distribution δ_x at X.*

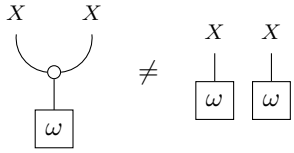


$$:: \; y \mapsto \begin{cases} 1 & x = y \\ 0 & \textit{otherwise} \end{cases}$$

The corresponding effect is given by the function δ_x also. Each cap is given by $\cap(x, y) = \delta_{x,y}$. We note a useful fact that for any morphism $M \colon X \to Y$ its values $M(y \mid x)$ can be given diagrammatically as below.



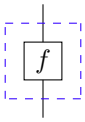
Every sharp state on X is of the above form for some $x \in X$, or else given by the zero state *0 defined by $0(x) = 0$ for all $x \in X$. The only sharp scalars are 0 and 1. Note that a general state ω, even when normalised, is not copyable.*



Indeed the left-hand side is the distribution $(x, y) \mapsto \omega(x)\delta_{x,y}$*, while the right is* $(x, y) \mapsto \omega(x)\omega(y)$*, which differ unless* ω *is zero or* $\omega = \delta_x$ *for some* $x \in X$*.*

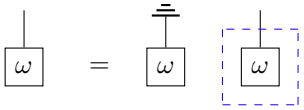
2.3 Normalisation

In graphical probabilistic reasoning it is also useful to be able to normalise states and processes. We say that a cd-category $\mathbf{C}$ has *normalisation* when it comes with a rule assigning each morphism $f\colon X \to Y$ a new morphism called the *normalisation* of f, depicted by drawing a dashed blue box:



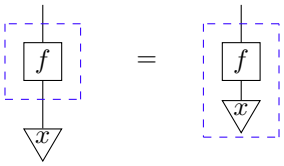
(7)

such that these normalisations satisfy various axioms, of which we sketch a few here. For a full definition see [LT23]. Firstly, a general state ω is equal to a scalar multiple of its normalisation. In particular in $\mathbf{Mat}_{\mathbb{R}^+}$ when the state is non-zero, this means that its normalisation is indeed normalised in our above sense, i.e. a distribution.



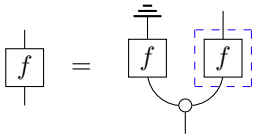
(8)

For a general morphism f its normalisation is given on each sharp state x by normalising $f \circ x$.



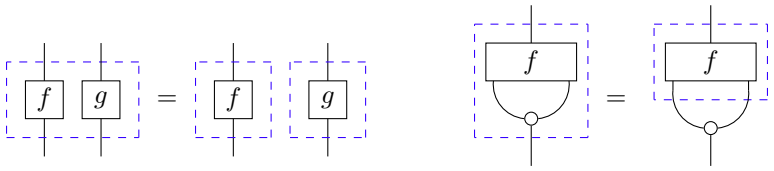
(9)

These two rules combine to give the following equation without explicit reference to states.



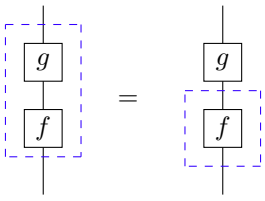
(10)

Note that if f we already a channel then it would be equal to its normalisation, as in this case we can passing the discarding through f and then the copy map above. In general normalisations satisfy a few graphical conditions including the following.



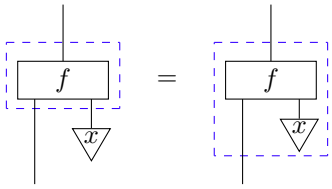
(11)

Further, for all morphisms f and channels g we have:



(12)

and for all sharp states x and morphisms f we have the following.



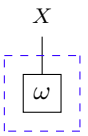
For a full account of the properties of normalisation see [LT23]. We note that for a general morphism f, its normalisation is not necessarily a channel but only a 'partial channel' [2]. In terms of states, this is because its sends each sharp state x either to a normalised state, or else to 0 if $f \circ x = 0$. However in $\mathbf{Mat}_{\mathbb{R}^+}$ it will be a channel provided f has 'full support', so that $f \circ x$ is non-zero for all non-zero sharp states x.

Throughout the article, the following notation will be useful. For any set X and function $f\colon X \to \mathbb{R}^+$ let us write

$$\underset{x}{\mathrm{Norm}}\, f(x) := \frac{f(x)}{\sum_{x' \in X} f(x')}$$

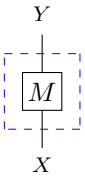
whenever this is well-defined, i.e. the denominator is finite and non-zero.

Example 5. $\mathbf{Mat}_{\mathbb{R}^+}$ *has normalisation. On each object* X *the zero state* 0*, given by* $0(x) = 0$ *for all* $x \in X$*, is defined to have normalisation* 0*. For any non-zero state* ω *we indeed have*



$$:: \; x \; \mapsto \underset{x}{\mathrm{Norm}}\, \omega(x)$$

For a general morphism $M\colon X \to Y$ *the normalisation is given by:*



$$:: \; (x, y) \mapsto \begin{cases} \mathrm{Norm}_y\, M(y \mid x) & \text{if } \sum_{y \in Y} M(y \mid x) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

As a result if M *has full support, so that* $M(y \mid x) \neq 0$ *for some* y*, for all* x*, then its normalisation is a probability channel.*

2.4 Further cd-categories

Though we will not need them here, we note that the notion of a cd-category is much more general than $\mathbf{Mat}_{\mathbb{R}^+}$, and give a few examples for those familiar with them. The category **Rel** whose objects are sets and morphisms are relations is a cd-category, as are its subcategories **PFun** of sets and partial functions and **Set** of sets and functions, with the latter forming the channels in **PFun**.

There are also many more cd-categories of a 'probabilistic' nature, see for instance [Pan98, CJ19, Fri20]. In particular to treat general probability spaces (including 'continuous probability channels') one may work in the category $\mathrm{Kl}(G)$ of measurable spaces $X = (X, \Sigma_X)$ and Markov (sub-)kernels $f\colon X \to Y$, which send each $x \in X$ to a (sub-)probability measure $f(x)$ over Y. Roughly, this means replacing all instances of summation Σ in $\mathbf{Mat}_{\mathbb{R}^+}$ above with integration $\int$. Of particular interest in PP is the following subcategory, though we will not work with it in detail in this article.

Example 6. *[Fri20, Section 6] In the category* **Gauss** *the objects are spaces* $X = \mathbb{R}^n$ *and morphisms* $M\colon X \to Y$ *are Markov kernels* $f\colon X \to Y$ *with densities of the form* $f(y \mid x) = \eta(y - Mx)$ *for some fixed Gaussian noise distribution* η *(independent of* x*) and linear map* $M\colon X \to Y$*. This category models linear processes with Gaussian noise. More general non-linear Gaussian processes are studied in PP under the so-called 'Laplace assumption'.*

[2] Such morphisms are called 'quasi-total' in [DLR23], where morphisms satisfying (10) are also called 'normalisations', though are not uniquely chosen unlike our definition.

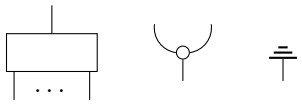
3 Generative Models

A central feature in PP is that each cognitive agent possesses a generative model which describes their internal beliefs about how the observations they receives arise from hidden states of the world[3]. In its simplest form, a generative model consists of a channel $c\colon S \to O$ describing how likely $c(o \mid s)$ a given observation $o \in O$ is for each hidden state $s \in S$, along with a distribution σ over S describing prior beliefs about how likely each state is.

However, generative models typically come with further compositional structure, relating various spaces of observations and hidden states, as formalised by a *Bayesian network* (or more precisely a *causal* Bayesian network, see later discussion), a probabilistic graphical model based on a directed acyclic graph (DAG). There is in fact a close correspondence between DAGs and cd-categories, allowing us to describe and study such models entirely in terms of string diagrams. This view also leads one to consider more general 'open generative models', which may come with 'input' variables. These open models can be used to which describe the individual components of an overall generative model in the usual sense. For more details on the approach used here, see [LT23].

We begin by relating DAGs with the following class of string diagrams.

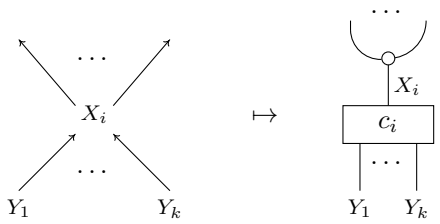
Definition 7. *[LT23] A* network diagram *is a string diagram D built from single-output boxes, copy maps and discarding:*



with labellings on the wires, such that any wires not connected by a sequence of copy maps are given distinct labels, and each label appears as an output at most once and as an input to any given box at most once.

Such diagrams are best understood by examples, which we come to shortly. Before this, we note that network diagrams are in fact equivalent to DAGs in the following sense. By an *open DAG* we mean a finite DAG G with vertices $V = \{X_1, \ldots, X_n\}$, along with subsets $I, O \subseteq V$ of *input* and *output* vertices, respectively, such that each input vertex has no parents in G.

Given any open DAG G, we may construct an equivalent network diagram featuring a box c_i with output X_i for each non-input vertex X_i. The box c_i itself has an input wire for each parent of X_i in G. In the diagram we copy the output of this box and pass it to each of the children of X_i, as well as an extra time if $X_i \in O$ i.e. X_i is an output vertex of the DAG.



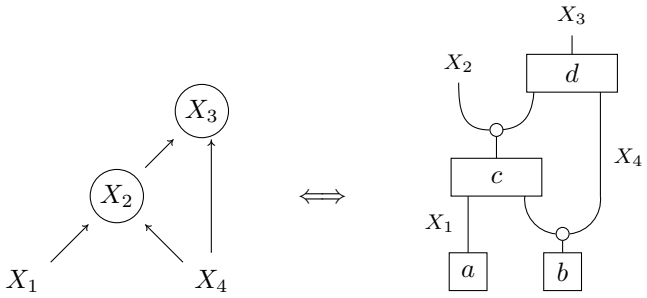
By construction, this yields a network diagram D_G from the inputs I to the outputs O of the DAG. Conversely, given any such network diagram D we define an open DAG $G_D = (G, I, O)$ with a vertex $X \in V$ for each wire X in D, and with $X \in I, O$ iff it is an input (resp. output) to the diagram.

In practice the labellings of the boxes are arbitrary, and we consider any two network diagrams equivalent when they are the same up to the equations of a cd-category and box re-labellings. Then the above yields a one-to-one correspondence between open DAGs and network diagrams [LT23, Sections 3,5].

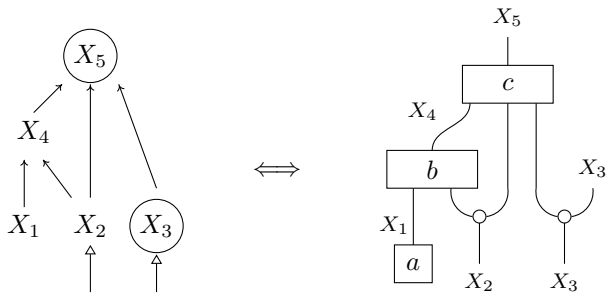
Example 8. *Consider the open DAG G over $\{X_1, X_2, X_3, X_4\}$ below, with output vertices $O = \{X_2, X_3\}$ circled, and with no input vertices. The equivalent network diagram D_G is shown to the right. Note that the*

[3]Note that is distinct from whatever 'true' external process produces the observations in reality, with the latter often called the 'generative process' to distinguish it from the agent's own 'generative model' [PPF22].

labels of the boxes are arbitrary.



Example 9. *The following depicts an open DAG over $V = \{X_1, \ldots, X_5\}$ with outputs $O = \{X_3, X_5\}$ and with inputs $I = \{X_2, X_3\}$ highlighted with special incoming arrows. To the right we show the corresponding network diagram with the same inputs and outputs.*



We may now define generative models themselves, which involve specifying actual channels corresponding to the boxes in the network diagram.

An *interpretation* $[\![-]\!]$ of a network diagram D in a cd-category $\mathbf{C}$ consists of specifying an object $[\![X_i]\!]$ for each wire X_i and channel $[\![f]\!] \colon [\![X_1]\!] \otimes \cdots \otimes [\![X_k]\!] \to [\![X]\!]$ for each box f in D with inputs $X_1, \ldots, X_k$ and output X.

Definition 10. *[LT23] Let $\mathbf{C}$ be a cd-category. An* open generative model *in $\mathbf{C}$ is given by a network diagram D along with an interpretation $[\![-]\!]$ in $\mathbf{C}$. We call the objects corresponding to output wires* observed *and the rest* hidden*. We call such a model* closed *when it has no inputs.*

Note that an object of an open model may be both an input and output. In practice, we omit the $[\![-]\!]$ symbols and for each wire X in the network diagram of a model denote the corresponding object $[\![X]\!]$ in $\mathbf{C}$ also by X. Similarly for each box c in the diagram with output X we also write c for the corresponding channel $[\![c]\!]$.

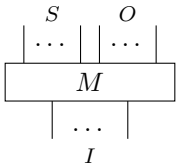
Remark 11. *Formally, an open generative model in our sense is the same as an* open causal model *in the sense of [LT23]; that is, both have the same mathematical definition. However a 'generative model' typically refers to a causal model with the extra interpretation of being possessed by a cognitive agent.*

Indeed, though not often stressed in the literature, a typical generative model in PP may be seen as a causal *Bayesian network, i.e. a* causal model *in the sense of Pearl [Pea09]. This means that the probability channels which constitute the network do not represent arbitrary relationships but in fact (beliefs about) causal ones, such as how observations are caused by (rather than merely correlated with) hidden states of the world. For more discussion see Section 10.*

Given any open generative model $\mathbb{M}$ we obtain an overall channel from its inputs to its outputs by composing the channels of the model, i.e. viewing the (interpreted) network diagram as a single channel in $\mathbf{C}$. Often it is useful to also consider the following related channel.

Definition 12. *Let $\mathbb{M} = (D, [\![-]\!])$ be an open generative model in $\mathbf{C}$ with inputs I and outputs O. Let S denote the non-input hidden (non-output) objects of the model. The* total channel *M of the model is the*

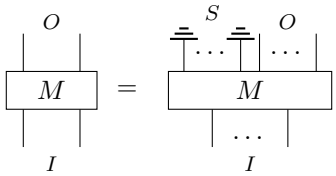
channel from I to S, O:



(13)

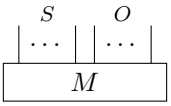
given by interpreting the network diagram D' in which we modify D by adding an extra copy morphism to each object in S, to make it an output.

Conversely, the usual channel from inputs to outputs is then simply the marginal over S:



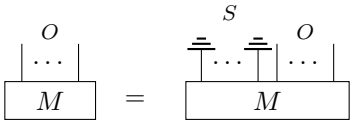
(14)

In particular for a closed generative model, with no inputs, we call the total channel the *total distribution* of the model. It is a joint distribution over the hidden objects S and observed objects O:



(15)

with the original distribution over the observed objects as its marginal.



(16)

Let us now consider generative models in our main example category.

Example 13. *A closed generative model $\mathbb{M}$ in $\mathbf{Mat}_{\mathbb{R}^+}$ is precisely a* Causal Bayesian Network (CBN)*. This consists of specifying:*

- *a finite DAG $G$with a subset $O \subseteq V$ of 'observed' vertices and the remaining $S = V \setminus O$ being 'hidden';*
- *for each vertex X_i an associated variable with a finite set of values also denoted X_i, and a mechanism c_i given by a probability channel with density:*

$$P(X_i \mid \mathsf{Pa}(X_i)) \tag{17}$$

The term 'causal' refers to the fact each such mechanism has a causal interpretation.

Indeed, as we have seen, such a DAG G with outputs O is equivalent to a network diagram D with no inputs. Specifying an interpretation of D is then the same as choosing the sets X_i of values and channels (17) *for each box in the diagram. A CBN defines a joint distribution*[4] *over all the variables $V = \{X_1, \ldots, X_n\}$ with density*

$$P(V) := \prod_{i=1}^{n} P(X_i \mid \mathsf{Pa}(X_i)) \tag{18}$$

which is precisely (15)*. The output distribution of the CBN is given by the marginal $P(O)$ over only the observed variables, corresponding to* (16)*.*

[4]Often a Bayesian network is instead defined as a distribution $P(V)$ satisfying the *Markov condition* (18) in terms of its conditionals. Since these conditionals may not be unique, and the channels c_i are an important component of the model, we instead include the latter explicitly; for more discussion see [LT23].

Example 14. *An open generative model* $\mathbb{M}$ *in* $\mathbf{Mat}_{\mathbb{R}^+}$ *is an 'open CBN', where now for the input variables no channel* (17) *is specified. This induces via* (13) *the total channel*

$$P(S, O \mid I)$$

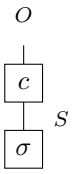
from the inputs to the non-input hidden variables S *and output variables* O*, which here we would denote (the entries of) by* $M(s, o \mid i)$*. Its marginal* $P(O \mid I)$ *on the observed variables* O *yields the channel* $M(o \mid i)$ *from* (14)*.*

In short, a (closed) generative model in $\mathbf{Mat}_{\mathbb{R}^+}$ specifies the internal structure of an output distribution $P(O)$ in terms of further variables and channels (17), while an open generative model similarly specifies the internal structure of a channel $P(O \mid I)$ from inputs I to outputs O.

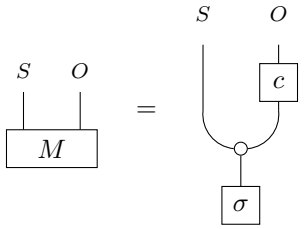
For the remainder of this section we will describe some of the common forms of (open) generative models which appear in PP.

3.1 Simple generative models

By a *generative model* $S \to O$ we mean a generative model $\mathbb{M}$ with network diagram:

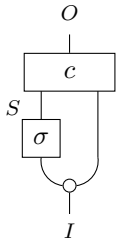


Thus $\mathbb{M}$ consists of objects S, O with O observed and S hidden, a channel $c\colon S \to O$, called the *likelihood*, and a distribution σ on S, called the *prior*. As alluded to earlier, we call S the *hidden states* and O the *observations* of the model. The total distribution of the model is given by



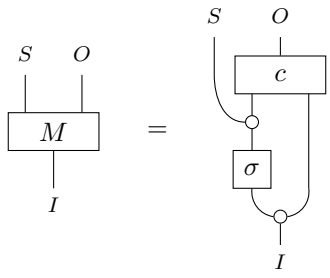
(19)

More generally, we can consider an *open* variant of such a generative model which now comes with a hidden object I of *inputs*, with the following network diagram:



(20)

Hence both the prior and likelihood now take an additional I input. The total channel is given by



(21)

Intuitively, such an open model $\mathbb{M}$ consists of specifying a particular generative model $S \to O$ for each input in I.

Example 15. *A generative model $S \to O$ in $\mathbf{Mat}_{\mathbb{R}^+}$ consists of a finite set S of hidden states, O of observations, a likelihood channel $c(o \mid s)$ and prior distribution $\sigma(s)$. Often would often write $c(o \mid s)$ as simply $P(o \mid s)$ and $\sigma(s)$ as $P(s)$. We interpret $c(o \mid s)$ as the probability of observing o when in the hidden state s. Then the total state* (19) *is the joint distribution over S, O given by*

$$M(s, o) = c(o \mid s)\sigma(s)$$

and typically simply denoted $P(s, o)$. As the notation suggests $P(o \mid s)$ is the conditional and $P(s)$ the marginal of the joint distribution $P(s, o)$.

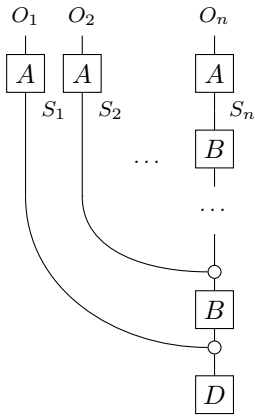
When the generative model is open as in (20) *it now comes with a finite set I of input values, with likelihood $c(o \mid i, s)$ and prior $\sigma(s \mid i)$. The total channel* (21) *is then given by*

$$M(s, o \mid i) = c(o \mid i, s)\sigma(s \mid i)$$

Thus for each input i we obtain a generative model $\mathbb{M}(i)$ of the form $S \to O$ and an induced joint distribution $M(i)$ over S, O.

3.2 Discrete time models

For a given $n \in \mathbb{N}$, a *discrete time generative model* is a closed generative model $\mathbb{M}$ of the form



Thus it consists of observed objects $O_1, \ldots, O_n$, hidden objects $S_1, \ldots, S_n$, a *prior* distribution D over S_1, *observation* channels $\{A\colon S_t \to O_t\}_{t=1}^{n}$ and *transition* channels $\{B\colon S_t \to S_{t+1}\}_{t=1}^{n-1}$. Typically we mean that there are fixed objects S, O such that $S_j = S, O_j = O$ for all j, and similarly as our notation suggests all the A and B channels for each time step are taken to be identical.

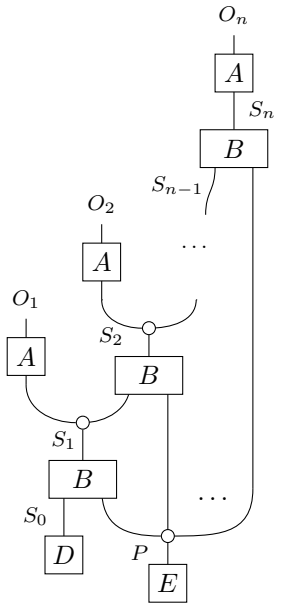
We interpret the model $\mathbb{M}$ as describing the evolution of a system over discrete time steps from $t = 1, \ldots, n$. The system begins in its initial state with prior distribution D and then evolves over each time step according to the transition channels B. Independently, we observe the system at each time t via the channel A to produce an observation in O_t.

Example 16. *A discrete time generative model in $\mathbf{Mat}_{\mathbb{R}^+}$ is also called a* Hidden Markov Model *or* partially observable Markov decision process (POMDP) *[PPF22].*

3.3 Policy models

We now introduce an explicit ingredient whereby the agent can model its own *actions*. As in reinforcement learning [TMSB20], a choice of actions or behaviour is called a *policy*. In a discrete time setting, a policy can be thought of as determining likely sequences of actions over the time steps, which in turn influence the evolution of the states over time.

An *n-time step model with policies* is a generative model $\mathbb{M}$ of the form:



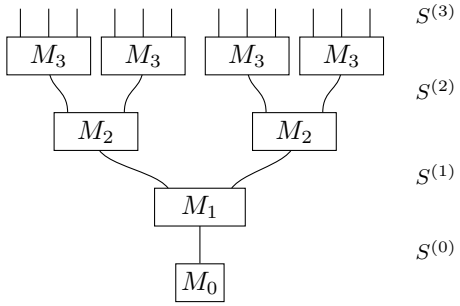
Thus it now includes a hidden object P of *policies* which forms an input to each transition channel B from S_t, P to S_{t+1}, for $t \leq n-1$. The model also comes with a prior distribution E over P, which are called the *habits* of the system. Note that here the policy the system is undertaking is considered hidden.

Again we typically take $S_j = S$ and $O_j = O$ for some fixed objects S, O, with all channels A identical and all B channels identical.

Example 17. *Models of this form, within* $\mathbf{Mat}_{\mathbb{R}^+}$*, are the central examples used in the active inference tutorial [SFW22] and book [PPF22].*

3.4 Hierarchical models

Central to much of PP is the study of *hierarchical* generative models [DVF17, PPF22], which have a natural graphical description. These are generative models given by composing various open generative models in layers, where the outputs of the open models in one layer match the inputs of the models in the next layer, such as in the example below.



(22)

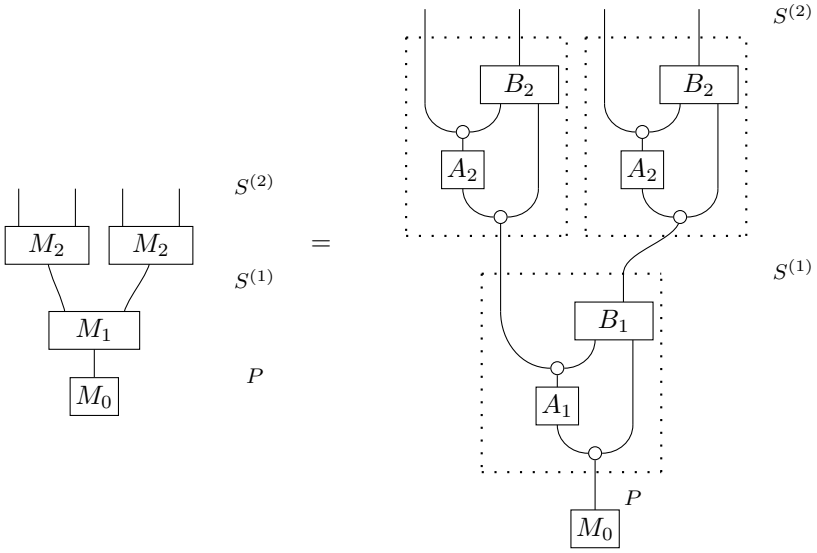
Here it is understood that each box M_j represents an open generative model, which we may decompose further in terms of its own network diagram with inputs and outputs as shown. The right-hand labels indicate that the input wire to M_1 has type $S^{(0)}$, the output wires from M_1 both have type $S^{(1)}$ etc [5].

[5]It is also common to introduce a labelling convention for the wires such as $S^{(0,1,2)}$ where the indices represent wire numbers in each layer as we read up the diagram. However this quickly becomes unwieldy, and in most cases the graphical description of the network is the most convenient.

We interpret the inputs to each (box within a) layer as a 'control' signal from the layer below. Note that because we read diagrams bottom to top, the layers further down the diagram are in fact those usually referred to as more 'high-level' or 'higher' in the hierarchy.

The structure of the model tells us that the 'high-level' features cause the generation of the 'lower-level' features. For example $S^{(0)}$ could describe an overall action policy while the $S^{(3)}$ control more fine-grained motor actions. Another common example explored in [DVF17] is where the output wires from each box denote individual time steps. In this case time runs faster in the lower-level layers (higher in the diagram). For example in the diagram above six time steps occur in layer $S^{(3)}$ for every time step in layer $S^{(1)}$.

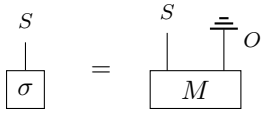
Plugging in the network diagrams for each open model corresponding to M_j yields a composite network diagram for the whole hierarchical model. For example in the following hierarchical model, the network diagrams for M_1 and M_2 are shown in the highlighted boxes below and compose to yield the diagram on the right-hand side.



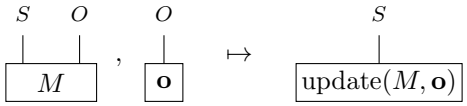
Much of the PP literature concerns such hierarchical models and the passing of these 'top-down predictions' (the flow of information up the diagram in this case) are adjusted by 'bottom-up errors' passed back down the model. The latter takes place when a model is *updated*, which we address next.

4 Updating Models

Consider an agent with be a simple generative model $\mathbb{M}$ of the form $S \to O$ as in Section 3.1. Recall that this induces a joint distribution M over S, O as in (15), whose marginal on S is the prior σ describing 'beliefs' about how likely each state in S is to occur.



Now suppose the agent receives an observation, which in general may be 'soft', given by an distribution $\mathbf{o}$ over O. The agent would like to *update* these beliefs to obtain a new *posterior* distribution over S, describing how likely each $s \in S$ now is given the observation.



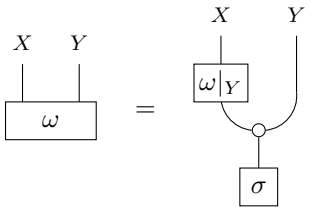
How then should the agent update the marginal on S to yield this posterior? For a general soft observation with distribution $\mathbf{o}$ over O there are at least two distinct but natural ways to carry out Bayesian-style

updating, as pointed out by Jacobs in [Jac19], which we describe in this section. When the observation **o** is sharp, however, corresponding to a single element $o \in O$, there is a canonical way to carry out this belief updating, usually simply referred to as *Bayesian* updating, which we introduce first.

4.1 Sharp Updating

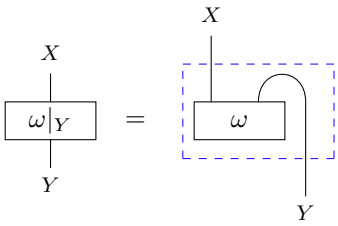
Let us begin by describing updates with respect to a sharp observation, given by (a point distribution at) an element $o \in O$. Such Bayesian updating is closely related to the notion of *conditional* probabilities, which have a nice characterisation in cd-categories. Here we follow the approach to conditioning from [LT23], building on earlier treatments [CS12, CJ19, Fri20]; see also [DLR23].

Definition 18. *Let* $\mathbf{C}$ *be a cd-category, and* ω *a joint distribution over* X, Y*. Then* a conditional *of* ω *by* Y *is a morphism* $\omega|_Y \colon Y \to X$ *such that the following holds:*



(23)

where σ *is the marginal of* ω *on* Y*. If* $\mathbf{C}$ *has normalisation and cancellative caps, we define* the (minimal) conditional *to be the morphism:*

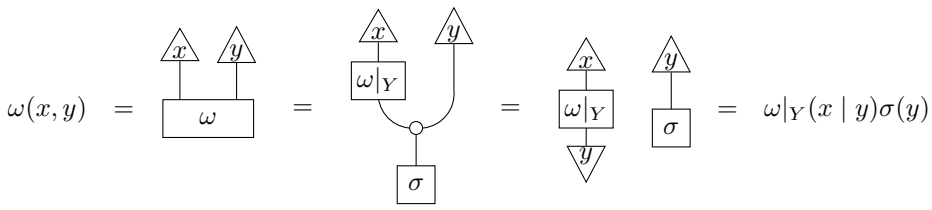


Each minimal conditional is indeed a conditional as shown in the Appendix of [LT23]. As we saw for normalisations, a conditional is only a partial channel in general, being a channel only when ω has 'full support'.

Example 19. *In* $\mathbf{Mat}_{\mathbb{R}^+}$ *the minimal conditional* $\omega|_Y$ *is given by*

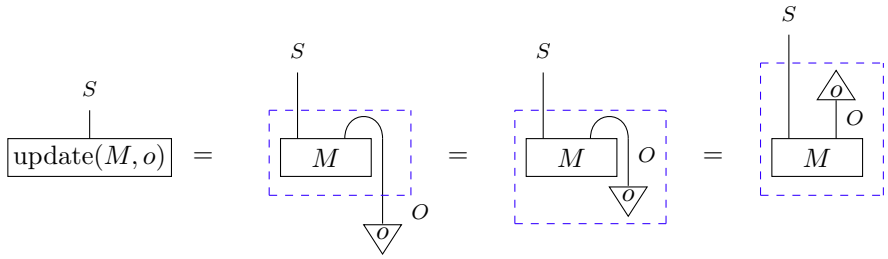
$$\omega|_Y(x \mid y) := \operatorname*{Norm}_x \omega(x, y) = \frac{\omega(x, y)}{\sum_{x'} \omega(x', y)}$$

whenever the sum in the denominator is non-zero, and $\omega|_Y(x \mid y) = 0$ *for all* x *otherwise. Thus when* ω *is normalised with density denoted* $P(X, Y)$ *this is the usual conditional* $P(X \mid Y)$*. The condition* (23) *amounts to the usual 'chain rule'* $P(x, y) = P(x \mid y)P(y)$ *for the probability distribution* $P(x, y) = \omega(x, y)$*, since* $\sigma(y)$ *is the marginal* $P(y)$ *and we have:*



For a generative model $\mathbb{M}$ of the form $S \to O$ with joint distribution M over S, O we call the minimal conditional $M|_O \colon O \to S$ the *Bayesian inverse* of the model. It specifies how to update beliefs about S for

each specific sharp observation $o \in O$. Explicitly, given a sharp distribution $\mathbf{o} = \delta_o$ over O for some $o \in O$ the updated beliefs are given by the posterior:



(24)

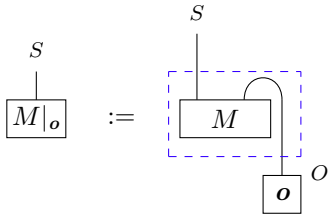
Example 20. *In* $\mathbf{Mat}_{\mathbb{R}^+}$*, for a sharp observation* δ_o *for some* $o \in O$*, the posterior is the distribution over* S *given by the usual Bayesian update:*

$$M(s \mid o) = \frac{M(s,o)}{\sum_{s'} M(s',o)} \tag{25}$$

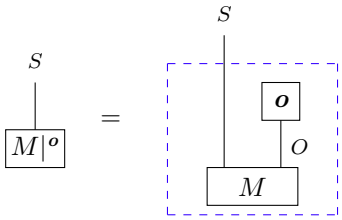
4.2 Pearl and Jeffrey Updating

There are two distinct ways to generalise updating to the case of a soft observation given by a distribution $\mathbf{o}$ over O, described in [Jac19]. Diagrammatically these correspond to generalising from either the former or latter diagrams in (24). For more on both forms of updating in cd-categories see also the treatment by Di Lavore and Román [DLR23].

Definition 21. *Let* $\mathbf{C}$ *be a cd-category with normalisation and cancellative caps, and* M *a joint distribution over* S, O*. Given a distribution* $\boldsymbol{o}$ *over* O*, the* Jeffrey update *denoted* M_J *or* $M|_{\boldsymbol{o}}$ *is given by the composite* $M|_O \circ \boldsymbol{o}$*, i.e.:*



whenever this is normalised, and more generally is given by the normalisation of the above state. The Pearl update *denoted* M_P *or* $M|^{\boldsymbol{o}}$ *is instead given by the normalisation:*



recalling that the effect $\boldsymbol{o}$ *is given by composing* $\boldsymbol{o}$ *with a cap as in* (5)*.*

Example 22. *For a generative model* $\mathbb{M}$ *from* S *to* O *in* $\mathbf{Mat}_{\mathbb{R}^+}$*, with joint distribution* M *over* S, O*, the Jeffrey update is given by*

$$M_J(s) = \mathop{\mathbb{E}}_{o \sim \boldsymbol{o}} \mathop{\mathrm{Norm}}_{s} M(s,o) = \sum_o \frac{M(s,o)\boldsymbol{o}(o)}{\sum_{s'} M(s',o)} \tag{26}$$

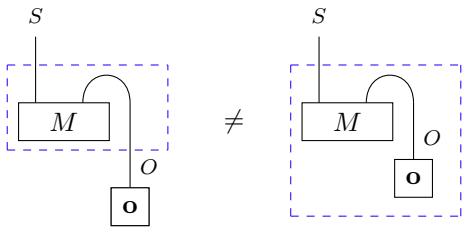
while the Pearl update is

$$M_P(s) = \mathop{\mathrm{Norm}}_{s} \mathop{\mathbb{E}}_{o \sim \boldsymbol{o}} M(s,o) = \frac{\sum_o M(s,o)\boldsymbol{o}(o)}{\sum_{s',o'} M(s',o')\boldsymbol{o}(o')} \tag{27}$$

The distinction between both update procedures is not always considered in the literature since for sharp observations they coincide with the usual Bayesian update. Indeed the following is immediate from (24).

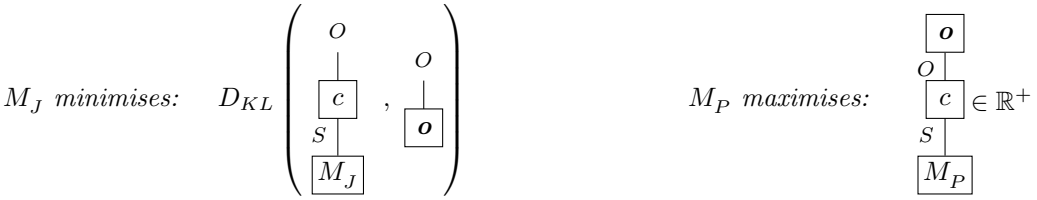
Lemma 23. *Let* $\mathbf{C}$ *be a cd-category with normalisation and cancellative caps. Then for each sharp state* o *on* O *the updates coincide:* $M_J = M_P = M|_O \circ o$.

In contrast, for a general observation $\mathbf{o}$ the two updates differ in the way they apply normalisation, amounting to whether one normalises with respect to (or separately from) the observation itself.



The Jeffrey update simply composes the observation $\mathbf{o}$ with the Bayesian inverse (partial) channel $M|_O$. If $M|_O$ is only a partial channel the result may not be normalised (such as when $\mathbf{o}$ falls outside the support), in which case the update is then further normalised. The Pearl update instead involves a single normalisation, taking place after composing with the observation, so that $\mathbf{o}$ is inside the normalisation box.

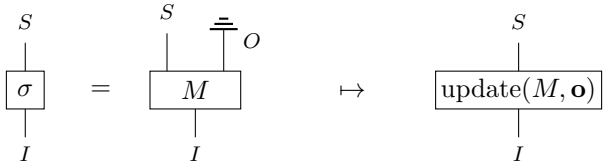
Remark 24. *Jacobs has compared the two forms of updating within* $\mathbf{Mat}_{\mathbb{R}^+}$ *in detail, noting that their inferences can differ considerably, but that both can be considered rational notions of updating [Jac19]. One difference between the updates is that by definition Jeffrey updating forms a probability channel in* O *(whenever* $M|_O$ *is a channel, i.e.* M *has full support over* O*). In contrast, the normalisation over* $\boldsymbol{o}$ *in Pearl updating means that it does not form a channel in* O*. The two update procedures can also be characterised by the following respective properties. For a generative model* $\mathbb{M}$ *over* S, O *with likelihood* c*, Jeffrey updating minimises the KL-divergence between* $\boldsymbol{o}$ *and the marginal on* O *of the updated model in which we replace the prior with the posterior (left-hand below). Pearl updating instead has the property that it maximizes the expected value of the function* $\boldsymbol{o}$ *(right-hand below).*



The PP literature has mostly focused on updating with respect to sharp observations, in which the two notions coincide. It is an interesting question for the future to determine which (if either) form of updating is most natural in Bayesian models of cognition.

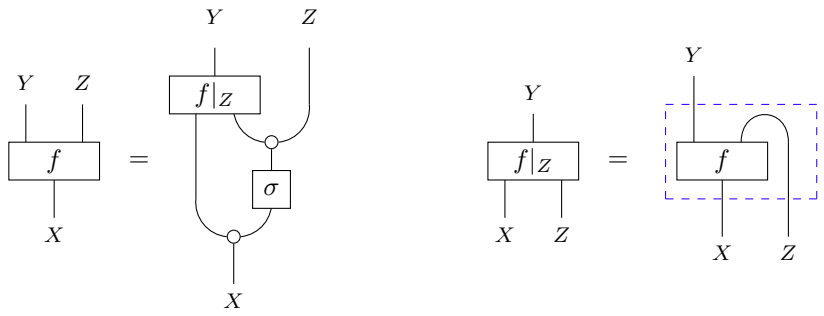
4.3 Updating Open Models

Since a typical generative model in PP is composed of various *open* generative models $\mathbb{M}$, it is also important to describe how an agent may update such open models $\mathbb{M}$, now coming with inputs I. In this case we consider the induced channel $M\colon I \to S, O$. The prior beliefs about S are now given by the marginal $\sigma\colon I \to S$, which we can think of specifying beliefs over S for each input $i \in I$. Given an observation $\mathbf{o}$ over O the agent now wishes to update this to a posterior channel of the same kind.

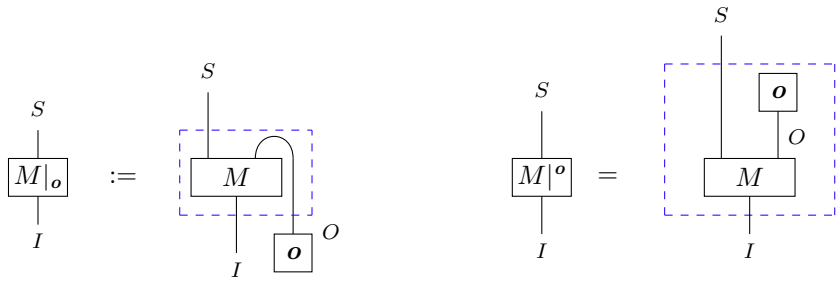


All of the treatment of updating above generalises straightforwardly to such open models, amounting to updating the corresponding closed model $M(i)$ over S, O for each input $i \in I$.

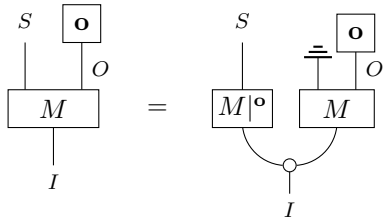
Explicitly, for any morphism $f\colon X \to Y \otimes Z$ in a cd-category, *a conditional* is any morphism $f|_Z$ satisfying the left-hand equation below, where σ is the corresponding marginal of f. In the presence of normalisation and cancellative caps, *the (minimal) conditional* is that given on the right below, as in [LT23].



Definition 25. *Let $M\colon I \to S, O$ be the channel induced by an open model $\mathbb{M}$, and $\boldsymbol{o}$ a distribution over O, in a cd-category with normalisation and cancellative caps. The* Jeffrey update *denoted M_J or $M|_{\boldsymbol{o}}$ is given by composing $M|_O$ with $\boldsymbol{o}$ as left-hand below (or more generally by its normalisation if the result is not a partial channel). The* Pearl update *denoted M_P or $M|^{\boldsymbol{o}}$ is instead given as on the right-hand side.*



By the defining property of normalisations (10) the Pearl update $M|_{\mathbf{o}}$ satisfies the following, which will be useful later.



(28)

Example 26. *In* $\mathbf{Mat}_{\mathbb{R}^+}$ *the minimal conditional of f by Z is given by*

$$f|_Z(y \mid x, z) = \operatorname*{Norm}_y f(y, z \mid x)$$

and for a probability channel $P(Y, Z \mid X)$ corresponds to the usual conditional $P(Y \mid X, Z)$. The formulae for both updates $M_J(s \mid i), M_P(s \mid i)$ are the same as (26), (27) *simply replacing each $M(s, o)$ term with $M(s, o \mid i)$, i.e.*

$$M_J(s \mid i) = \mathop{\mathbb{E}}_{o \sim \boldsymbol{o}} \operatorname*{Norm}_s M(s, o \mid i)$$
$$M_P(s \mid i) = \operatorname*{Norm}_s \mathop{\mathbb{E}}_{o \sim \boldsymbol{o}} M(s, o \mid i)$$

Again both update procedures coincide with $M(s \mid o, i)$ for sharp observations $i \in I$ and all inputs $i \in I$.

Remark 27. *Di Lavore and Román also study both forms of updating in cd-categories in which (non-chosen) conditionals exist in [DLR23], calling them 'partial Markov categories'. There updating is defined via arbitrary (non-minimal) conditionals, meaning that $M|_O$ can be arbitrarily defined outside the support on O of M. However since this arbitrary choice can impact the result of a Jeffrey update $M|_O \circ \boldsymbol{o}$ when $\boldsymbol{o}$ is also non-zero outside this support, we instead define updating via the minimal conditional $M|_O$.*

5 Perception and Planning

Let us now see how the notion of updating is applied by an agent to govern its behaviour in PP. Two fundamental uses of updating are the following.

Perception Firstly, as already alluded to, we can consider the case of an agent with a generative model $\mathbb{M}$ from S to O, interpreted as accounting for observations O in terms of hidden states of the world S. For example, O may be the space of pixel-level descriptions of images while S is a compressed representational space of possible objects which the images portray.

Given an observation encoded by a (soft or sharp) distribution $\mathbf{o}$ over O, the agent can update its prior over hidden states S to obtain a posterior describing how likely each hidden state is to have caused the observation. We refer to this general process of updating as *perception* and view the resulting distribution as the agent's specific perception of the observation $\mathbf{o}$. Intuitively perception takes the 'raw data' of the observation $\mathbf{o}$ and returns (a distribution over) representations S.

Intuitively, the update answers the question *'Given that I have received this observation, how likely is each possible world state?'*. In the literature this is often referred to as *inference*, in reference to Bayesian inference.

Planning A second application of updating by an agent is in *planning* its behaviours. Here an agent possesses a generative model $\mathbb{M}$ of the same formal structure but with objects labelled P, F and interpreted differently. Now P encodes the action policies, or behaviours, the agent may carry out, while F represents observations (or states) it may receive in the *future*. The model $\mathbb{M}$ includes a prior over policies which we can think of as the agent's habits or typical behaviours.

Here the agent possesses some *preferences* about which future observations (or states) are most desirable, encoded by a distribution C over F. Intuitively, the distribution will have highest density on the most desirable outcomes. The agent can then plan its actions by updating its habits with respect to these preferences:

The process of deriving this distribution can intuitively be called 'planning'. We can think of this update as answering the question *'Given that I will obtain my preferences in the future, how likely is each policy to have led to this outcome?'*.

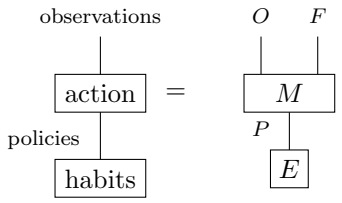
The resulting 'plan' distribution over P can be used to guide the agent's future behaviour. For example, an agent may then sample an policy to pursue from this distribution, so that the more probable policies according to the distribution are more likely to be carried out.

6 Exact Active Inference

Both uses of updating by an agent, planning and perception, come together in the concept of *active inference*, of which we are now able to present a fully formal diagrammatic account.

Consider an agent possessing a generative model describing how its actions, in the form of action policies P, bring about changes in its observations. These consist of both observations for the present time (and

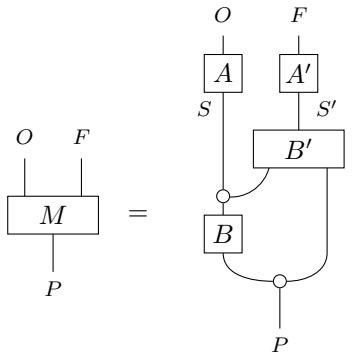
previous times) O and for future time steps F. Thus the agent has a closed generative model $\mathbb{M}$ of the following form.



(29)

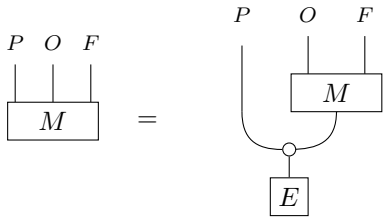
Here (abusing notation slightly) we denote by M also the channel from policies to observations induced by the model, and E is the prior over policies describing the agent's habits.

Suppose further that the agent's model explains the observations at each of these time steps through hidden states, where S denotes the hidden states in the present time and S' in the future, so that we have:



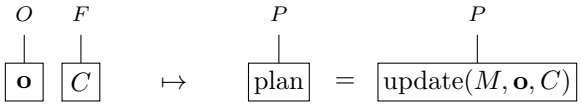
(30)

for 'observation' channels A, A' and 'transition' channels B, B'. The induced distribution on P, O, F is then given by:



(31)

The goal of active inference is then the following. The agent receives a current observation given by a distribution $\mathbf{o}$ over O, and also carries a distribution C describing its preferences for future observations F. The agent then wishes to update its prior E over policies to yield a posterior which describes its plan of action[6]:



(32)

Intuitively, the posterior over policies can be thought of as answering the question *'Given that I have received this observation* $\boldsymbol{o}$ *now, and will attain my preferences* C *in the future, which action policy am I pursuing?'*. Note that, perhaps surprisingly, the agent's own action policy is thus treated as hidden from itself, and something that it must infer.

Now, typically the objects above all decompose into further structure, as in the following example.

Example 28. *A common application of active inference is to the discrete-time models with policies given in Section 3.3, which we may view as instances of* (30) *as follows. Consider such a model featuring* N *time-steps, where* $n << N$ *is considered the current time, and all times* m *with* $n \leq m \leq N$ *as in the future.*

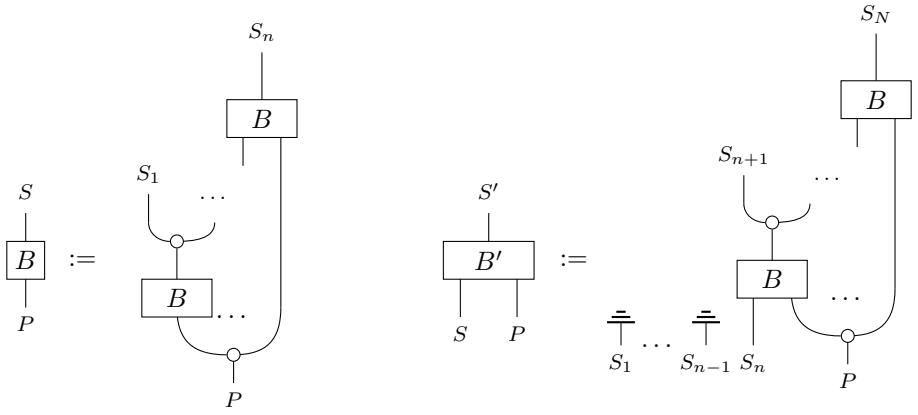
[6]Ultimately, having derived their 'plan' distribution the agent may then sample a single action policy $\pi \in P$ as in Section 5, and act accordingly. We imagine that via the true 'generative process' in the world (distinct from the agent's model) this leads to further observations in the future, to which the agent carries out further planning steps, and so on. Our focus is simply on a single step of how the agent derives their 'plan' from $\mathbf{o}$ and C.

The spaces of 'current' hidden states and observations S, O are the products over all previous time-steps $t = 1, \ldots, n$ up to and including the current time, while the future hidden states and observations S', F take the product over all future time-steps $t = n+1, \ldots, N$.

$$S := S_1 \otimes \cdots \otimes S_n \qquad S' := S_{n+1} \otimes \cdots \otimes S_N$$
$$O := O_1 \otimes \cdots \otimes O_n \qquad F := O_{n+1} \otimes \cdots \otimes O_N$$

The observation channels in the overall model (30) *would then be given by:*

while the transition channels are as follows:

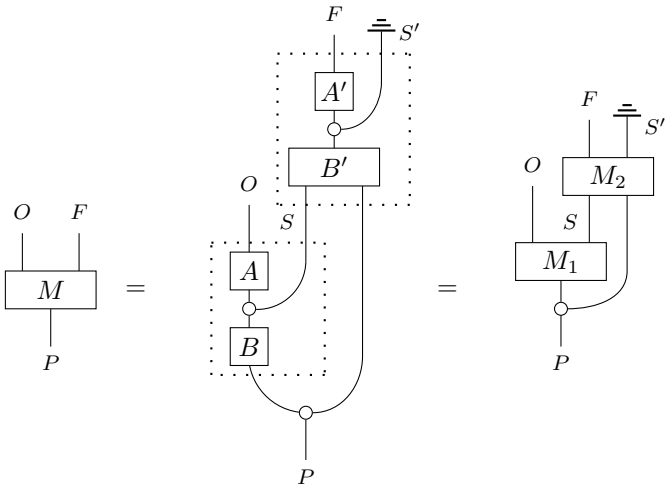


so that the composite (31) *yields the network diagram for the overall model for times $t = 1, \ldots, N$.*

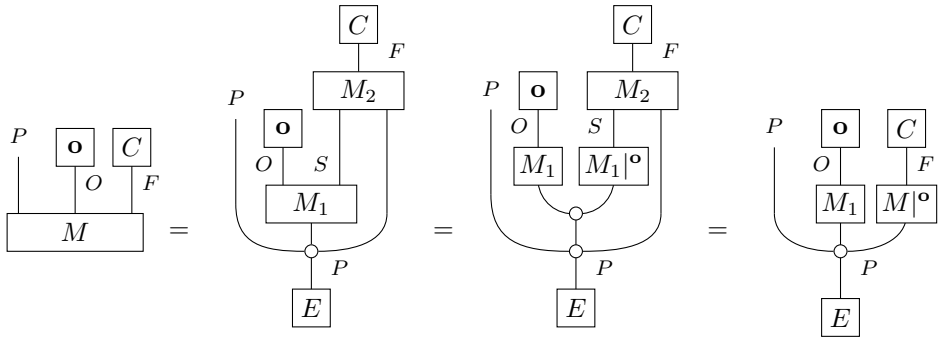
An agent may employ various update procedures, such as those discussed in Section 4, to calculate its plan of action (32). Though both forms of updating coincide for sharp inputs, and the observations $\mathbf{o}$ in the active inference literature are typically taken to be sharp, the preferences C are often not; that is, there may be multiple desirable future observations in F. Thus Pearl and Jeffrey updating can be expected to differ.

Here we will describe an exact active inference procedure based on Pearl updating, allowing both observations $\mathbf{o}$ and preferences C to be soft. We leave the exploration of Jeffrey updating in active inference for future work.

Now let us consider how the agent can in the ideal case compute its plan (32) via an exact update procedure. Firstly, let us rewrite the channel in (30) as follows.

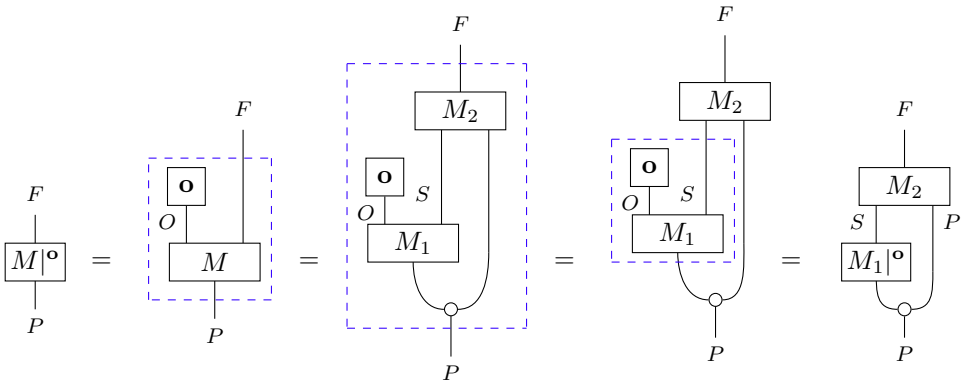


Here the channels M_1, M_2 are the compositions indicated by the highlighted boxes[7]. Now applying the property of Pearl updates (28) to M_1 we have the following:



(33)

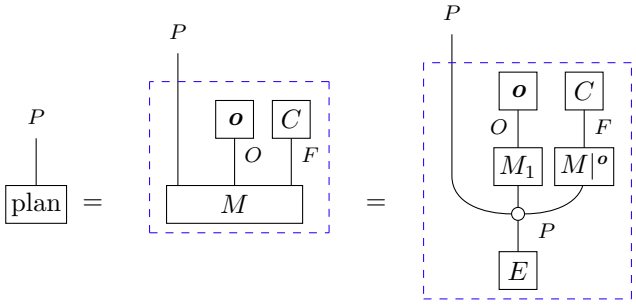
Here we have again denoted by M_1, M_2 their respective marginals on O, F, given by discarding S, S' respectively. In the last step we used associativity of copying and the following argument:



where in the middle step we used (11) and (12) to slide the channel M_2 and copying out of the normalisation box, respectively.

Thus we obtain an exact expression for active inference.

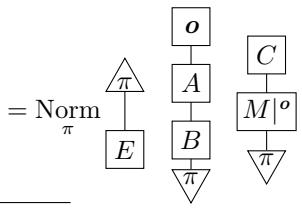
Proposition 29. *The plan over policies in Pearl-style exact active inference is given by:*



(34)

In $\mathbf{Mat}_{\mathbb{R}^+}$ *the plan has density over policies* $\pi \in P$ *given by:*

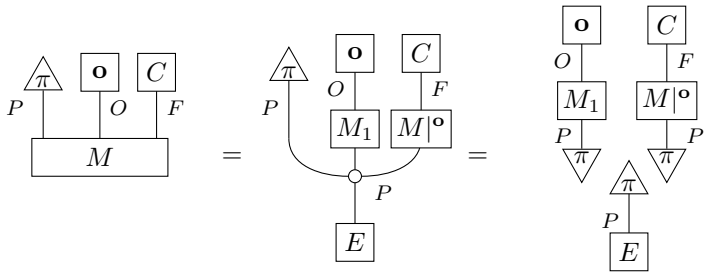
$$\text{plan}(\pi) := \underset{\pi}{\text{Norm}}\left(E(\pi)(\boldsymbol{o} \circ M_1(\pi))(C \circ M|^{\boldsymbol{o}}(\pi))\right) \tag{35}$$



(36)

[7]While we could define M_2 without S' as an output, the appearance of S' will be useful later in treating approximate active inference. Note also that the dashed boxes in this case do not denote normalisation.

Proof. The first equality holds by definition, so $\text{plan}(\pi) = \text{Norm}_\pi f(\pi)$ where f is the density of the state in (33). But this is given by:



using that π is sharp, where the three right-hand scalars are precisely the terms in (35). The last line comes from noting that the given marginal $M_1 \colon P \to A$ is precisely $B \circ A$. □

There is only one problem with this form of active inference: the quantity (35) is completely intractable to calculate. Along with the normalisation in calculating $M|^{\mathbf{o}}$, calculating the terms in (35) would involve summation (or integration) over S, O and S', F respectively, requiring us to respectively calculate:

$$\sum_{s \in S, o \in O} \mathbf{o}(o) A(o \mid s) B(s \mid \pi) \qquad\qquad \sum_{o' \in F} C(o') M|^{\mathbf{o}}(o' \mid \pi)$$

To make the calculation of these updates tractable, an agent in active inference is understood to instead use a special form of approximation scheme, to which we now turn.

7 Free Energy

We have seen that for an agent to perform exact Bayesian updating is computationally intractable. In active inference, an agent instead carries out approximate updating by minimising a quantity known as *free energy* [FKH06, Fri10, PPF22]. In this section for simplicity we work concretely in the category $\mathbf{C} = \mathbf{Mat}_{\mathbb{R}^+}$, though the same notions should be similarly defined in continuous settings.

The extra mathematical ingredient[8] needed to define free energy will be the following .

Definition 30. *For any distribution σ over X and $x \in X$ we define the* surprise *as* $\mathrm{S}(\sigma)(x) := -\log \sigma(x)$. *For another distribution ω on X we define the overall* surprise *of σ relative to ω as the expectation value:*

$$\mathrm{S}\left(\boxed{\omega}, \boxed{\sigma}\right) \quad := \quad - \mathop{\mathbb{E}}_{x \sim \omega} \log \sigma(x)$$

The entropy $\mathrm{H}(\omega)$ *of ω is its self-surprise:*

$$\mathrm{H}\left(\boxed{\omega}\right) := \mathrm{S}\left(\boxed{\omega}, \boxed{\omega}\right)$$

while the Kullback-Liebler (KL) divergence $\mathrm{D}(\omega, \sigma)$ *from σ to ω is the difference between these quantities:*

$$\mathrm{D}\left(\boxed{\omega}, \boxed{\sigma}\right) := \mathrm{S}\left(\boxed{\omega}, \boxed{\sigma}\right) - \mathrm{H}\left(\boxed{\omega}\right)$$

The KL divergence is a commonly used similarity measure on distributions, with $D(\omega, \sigma) \geq 0$ and $D(\omega, \omega) = 0$ for all distributions ω, σ.

We may now define the following general notion of free energy. Throughout we consider a distribution M over S, O, which we imagine to be induced by a generative model from S to O. In this section for simplicity given any such distribution we denote its marginals on S, O and conditional channels $M|_S, M|_O$ again simply by M.

[8] To study free energy we will move beyond a purely diagrammatic approach and make use of some probabilistic calculations, most notably to define 'surprise'. However later in Section 9 we will see how to represent surprise in diagrams (via 'log-boxes'). In future work it would be interesting to represent all of the calculations in this section using such diagrams.

Definition 31 (Free Energy)**.** *The* Free Energy *of a distribution Q over S, O relative to M is defined as:*

$$\mathrm{FE}\left(\begin{array}{cc} S & O \\ \vert & \vert \\ \boxed{Q} & \end{array}, \begin{array}{cc} S & O \\ \vert & \vert \\ \boxed{M} & \end{array}\right) := \mathrm{S}\left(\begin{array}{cc} S & O \\ \vert & \vert \\ \boxed{Q} & \end{array}, \begin{array}{cc} S & O \\ \vert & \vert \\ \boxed{M} & \end{array}\right) - \mathrm{H}\left(\begin{array}{c} S \\ \vert \\ \boxed{Q} \end{array}\right) \tag{37}$$

Explicitly then we can re-write the free energy in the following useful form.

$$\mathrm{FE}(Q, M) = \mathop{\mathbb{E}}_{(s,o)\sim Q}[\log(Q(s)) - \log(M(s,o))] \tag{38}$$

$$= \mathop{\mathbb{E}}_{(s,o)\sim Q}[\log(Q(s) - \log(M(s \mid o)) - \log M(o)] \tag{39}$$

$$= \mathop{\mathbb{E}}_{o\sim Q} \mathrm{S}\left(\begin{array}{cc} S & S \\ \vert & \vert \\ \boxed{Q} & \boxed{M} \\ \vert & \vert \\ \boxed{o} & \boxed{o} \end{array}\right) + \mathrm{S}\left(\begin{array}{cc} O & O \\ \vert & \vert \\ \boxed{Q} & \boxed{M} \end{array}\right) - \mathrm{H}\left(\begin{array}{c} S \\ \vert \\ \boxed{Q} \end{array}\right) \tag{40}$$

We now turn to two specific variants of this quantity commonly considered in active inference.

7.1 Variational Free Energy

Suppose an agent receives an observation given by a distribution $\mathbf{o}$ over O, and wishes to perform an approximate Bayesian update of its prior beliefs about S as encoded by the marginal of M on S. It may do so by finding the distribution q over S which minimises the following quantity.

Definition 32 (Variational Free Energy)**.** *Given a distribution M over S, O and distribution $\boldsymbol{o}$ over O, the* Variational Free Energy (VFE) *of a distribution q over S is defined as:*

$$\mathrm{F}\left(\begin{array}{c} S \\ \vert \\ \boxed{q} \end{array}\right) := \mathrm{FE}\left(\begin{array}{cc} S & O \\ \vert & \vert \\ \boxed{q} & \boxed{\boldsymbol{o}} \end{array}, \begin{array}{cc} S & O \\ \vert & \vert \\ \multicolumn{2}{c}{\boxed{M}} \end{array}\right)$$

An important feature of the VFE is the following. Using the expression (40) and pulling the entropy term inside the expectation we see that

$$\mathrm{F}(q) = \mathop{\mathbb{E}}_{o\sim \mathbf{o}} \mathrm{D}\left(\begin{array}{cc} S & S \\ \vert & \vert \\ \boxed{q} & \boxed{M} \\ & \vert \\ & \boxed{o} \end{array}\right) + \mathrm{S}\left(\begin{array}{cc} O & O \\ \vert & \vert \\ \boxed{\mathbf{o}} & \boxed{M} \end{array}\right) \tag{41}$$

$$\geq \mathrm{D}\left(\begin{array}{cc} S & S \\ \vert & \vert \\ \boxed{q} & \boxed{M} \\ & \vert \\ & \boxed{\mathbf{o}} \end{array}\right) + \mathrm{S}\left(\begin{array}{cc} O & O \\ \vert & \vert \\ \boxed{\mathbf{o}} & \boxed{M} \end{array}\right) \tag{42}$$

The inequality follows from concavity of the KL divergence and Jensen's inequality, which states that for any probability measure ω on X, measurable function $f\colon X \to \mathbb{R}$ and concave function ϕ on $\mathbb{R}$ we have

$$\mathop{\mathbb{E}}_{x\sim\omega}[\phi(f(x))] \leq \phi(\mathop{\mathbb{E}}_{x\sim\omega}[f(x)]) \tag{43}$$

In particular we see that the inequality (42) will be a strict equality whenever $\mathbf{o} = \delta_o$ is given by a sharp observation $o \in O$. In this case the minimum VFE value is given by the exact Bayesian inverse $M|_o$, with value $\mathrm{F} = -\log M(o)$. Hence for a sharp observation o, minimising the VFE minimises the KL-divergence between q and the Bayesian inverse $M|_o$, achieving approximate inversion $q \approx M|_o$. Moreover $\mathrm{F}(q)$ is an upper bound on the surprise of the observation o, and when $q \approx M|_o$ we have $\mathrm{F}(q) \approx S(o, M)$.

VFE Updating This process of minimising VFE to compute an approximate Bayesian update is central in active inference, but typically only considered for such sharp observations. Here we can now consider the more general minimisation of VFE for a soft observation given by a distribution $\mathbf{o}$. In fact we may view this as another notion of updating for a prior over S, in addition to the two forms of updating met in Section 4.

Firstly, observe that in the expression (41) since the surprise term is constant, the distribution q which minimises $\mathrm{F}(q)$ will be that which minimises the left-hand expected KL term, which is equal to the following.

$$\mathop{\mathbb{E}}_{s\sim q}\left[\log q(s) - \mathop{\mathbb{E}}_{o\sim\mathbf{o}} \log M(s \mid o)\right]$$

This quantity will in turn be minimised when this expression over S is equal to a constant K, so that:

$$\log q(s) = \mathop{\mathbb{E}}_{o\sim\mathbf{o}}[\log M(s \mid o)] + K$$

The distribution q will be given by normalising $q(s)$ in the above expression, allowing us to ignore the constant and yielding the following notion of updating motivated by the VFE. Recall that the *softmax* of a function $f\colon X \to \mathbb{R}^+$ is defined by $\sigma(f)(x) = \mathrm{Norm}_x\, e^{f(x)}$.

Definition 33 (VFE Update)**.** *Given a joint distribution M over S, O and distribution $\boldsymbol{o}$ over O the* VFE update *is the posterior*

$$M_F(s) = \mathop{\mathrm{Norm}}_{s} e^{\mathbb{E}_{o\sim\boldsymbol{o}} \log M(s|o)} \tag{44}$$

$$= \sigma(\mathop{\mathbb{E}}_{o\sim\boldsymbol{o}} \log M(s \mid o)) \tag{45}$$

where σ denotes a softmax over S.

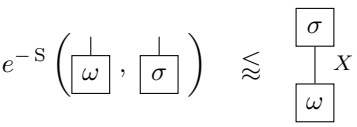
Similarly, for any channel M from P to S, O we define the VFE update *of its marginal $P \to S$ point-wise, by $M_F(s \mid \pi) = M(\pi)_F(s)$ for each $\pi \in P$.*

From the derivation above we see that $q = M_F$ is the distribution which minimises $\mathrm{F}(q)$. Note that, as for our other forms of updating, for a sharp observation $\mathbf{o} = \delta_o$ we have $M_F(s) = M(s \mid o)$.

To relate general VFE minimisation for a soft observation to expectation values, we will use the following form of approximation. Firstly, note that by Jensen's inequality, for any probability measure ω and real function f over X we have:

$$e^{\mathbb{E}_{x\sim\omega}[\log f(x)]} \leq \mathop{\mathbb{E}}_{x\sim\omega}[f(x)] \tag{46}$$

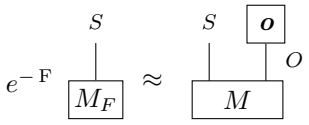
Whenever we take both sides of such an inequality to be approximately equal, let us say we are using a *log approximation*. In particular for any distributions ω, σ on X the follow holds log-approximately:



(47)

Indeed this states precisely (46) for the case $f(x) = \sigma(x)$. Such approximations can be used to relate free energy to exact expectation values, as follows.

Proposition 34. *Let M_F be the VFE update of M relative to a distribution $\boldsymbol{o}$ over O, and* F *its VFE value. Then the following holds log-approximately:*



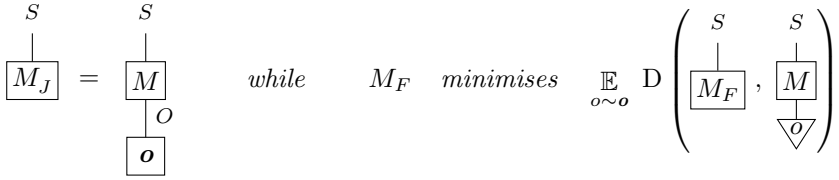
Proof. Define $f(s) := e^{\mathbb{E}_{o\sim\mathbf{o}} \log M(s|o)}$ and the normalisation constant $K = \sum_s f(s)$, so that $KM_F(s) = f(s)$. Then we have:

$$F = \mathrm{S}(\mathbf{o}, M) + \mathop{\mathbb{E}}_{s\sim q}[\log M_F(s) - \mathop{\mathbb{E}}_{o\sim\mathbf{o}} \log M(s \mid o)] \tag{48}$$

$$
\begin{aligned}
&= \mathrm{S}(\mathbf{o}, M) - \log K \\
e^{-F} M_F(s) &= e^{-\mathrm{S}(\mathbf{o},M)} K M_F(s) \\
&= e^{-\mathbb{E}_{o\sim\mathbf{o}}[\log M(o) + \log M(s|o)]} \\
&= e^{-\mathbb{E}_{o\sim\mathbf{o}}[\log M(s,o)]} \approx \mathop{\mathbb{E}}_{o\sim\mathbf{o}} M(s,o)
\end{aligned}
$$

where in the last step we used a log-approximation. □

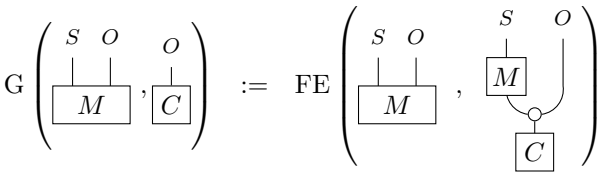
Remark 35. *Compare the formula for VFE update to the Jeffrey and Pearl updates* (26), (27). *While the Jeffrey update composes the conditional $O \to S$ with $\boldsymbol{o}$ exactly, the VFE update instead minimises the expected KL below.*



7.2 Expected Free Energy

A second form of free energy employed in active inference is used by an agent with a model featuring a space O describing observations in the future. It then has a distribution C over O modelling preferences for these future observations. Rather than updating its beliefs about future states, the agent simply want to assess how well the marginal of the model on O will fit these preferences, via the following approximation.

Definition 36. *Given a distribution M over S, O and distribution C over O, the* Expected Free Energy (EFE) *is defined as*



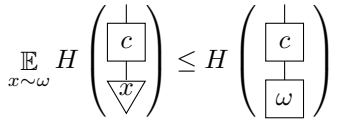
(49)

The EFE compares the given model M to the right hand generative model which perfectly attains the preferences, via its marginal C over O, whilst making use of the same inverse channel $O \to S$. Writing the EFE explicitly, and then rewriting in terms of the typically more readily computable channel $S \to O$, we have

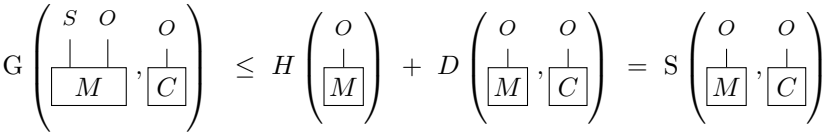
$$
\begin{aligned}
\mathrm{G}(M, C) &= \mathop{\mathbb{E}}_{(s,o)\sim M} [\log(M(s)) - \log(M(s \mid o))] - \mathop{\mathbb{E}}_{o\sim M} [\log C(o)] \\
&= \mathop{\mathbb{E}}_{\substack{s\sim M \\ o\sim M\circ s}} [-\log(M(o \mid s)] + \mathop{\mathbb{E}}_{o\sim M} [\log M(o) - \log C(o)] \\
&= \mathop{\mathbb{E}}_{s\sim M} \left[H\left(\begin{array}{c} O \\ \boxed{M} \\ \boxed{s} \end{array} \right) \right] + D\left(\begin{array}{cc} O & O \\ \boxed{M} & \boxed{C} \end{array} \right)
\end{aligned}
$$

The final line expresses the EFE in terms of a right-hand *risk* term, which assesses how well the predicted state over O matches the preferences C, and a left-hand *uncertainty* term given by the expected entropy in the observations. Thus minimising EFE requires both matching preferences and reducing uncertainty. For more interpretations of EFE see [PPF22].

Now using Jensen's inequality and the concavity of entropy, one may show that for any distribution ω and channel c we always have:



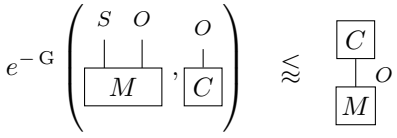
Hence the EFE is bounded above by the surprise of the preferences:



(50)

Thus minimising the EFE results in reducing the surprise of the preferences, making them more likely to be obtained according to the model. Taking the inequality to be an approximation and applying exponentials to both side along with a log-approximation then gives the following.

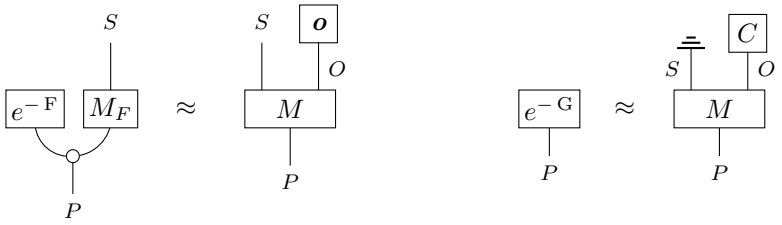
Proposition 37. *The EFE is bounded above and approximately equal to the expectation value:*



7.3 Free Energy in Active Inference

We conclude this section by noting two uses of free energy in approximate active inference, treated in the next section. For these we now consider a channel M from P to S, O, typically induced by an open model. For each $\pi \in P$ this specifies a joint distribution $M(\pi)$ over S, O, to which we may apply free energy calculations.

Corollary 38. *Let $\boldsymbol{o}$ and C be distributions over O. Let $M_F \colon P \to S$ be the VFE update of M by $\boldsymbol{o}$, and for each $\pi \in P$ set $\mathrm{F}(\pi) := \mathrm{F}(M(\pi)_F)$ to the corresponding VFE value. Similarly for each $\pi \in P$ let $\mathrm{G}(\pi) = \mathrm{G}(M(\pi), C)$. Then we have the following approximations:*



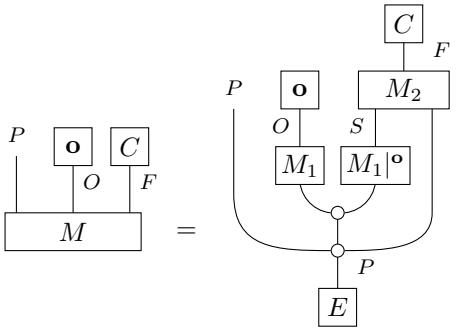
In the above the effect $e^{-\mathrm{F}}$ is given by $\pi \mapsto e^{-\mathrm{F}(\pi)}$, for $\pi \in P$, and $e^{-\mathrm{G}}$ is defined similarly.

Proof. For the first approximation, plugging in a (sharp state given by) an element $\pi \in P$ to both sides shows that this is equivalent to Proposition 34 holding for each joint distribution $M(\pi)$ over S, O with respect to the observation $\mathbf{o}$. For the second approximation, apply Proposition 37 to the joint distribution $M(\pi)$ over S, O for each $\pi \in P$. □

8 Active Inference via Free Energy

Let us now return to the situation of an agent carrying out active inference as in Section 6. As before the agent's generative model $\mathbb{M}$ in (29) consists of its habits E over policies P and a channel M from P to current and future observations O, F, factoring via current and future hidden states S, S'. Given its observation $\mathbf{o}$ and future preferences C it can now use free energy to give a viable approximation of its updated plan of

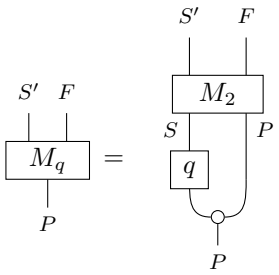
behaviour from Proposition 29, proceeding in two steps. We saw already in (33) that:



'Perception' step In the first step, the agent approximately updates the part of the model pertaining to the current time, M_1, in light of the observation $\mathbf{o}$. For each policy π it computes a distribution $q(\pi)$ with (approximately) minimal VFE $\mathrm{F}(q(\pi))$, thus obtaining a channel $q\colon P \to S$ which approximates the VFE update of M_1 by $\mathbf{o}$. For each $\pi \in P$ denote the corresponding VFE value by $\mathrm{F}(\pi)$. Explicitly:

$$\mathrm{F}(\pi) = \mathrm{F}\left(\begin{array}{c} S \\ \boxed{q(\pi)} \end{array}\right) := \mathrm{FE}\left(\begin{array}{c} S \quad O \\ \boxed{q(\pi)}\ \boxed{\mathbf{o}} \end{array}, \begin{array}{c} S \quad O \\ \boxed{M_1} \\ P \\ \overset{\pi}{\triangledown} \end{array}\right)$$

'Prediction' step In the second step, the agent uses this approximation channel q, to obtain a channel M_q which approximates the model over future states and observations, defined as follows:



For each policy π this induces a distribution $M_q(\pi)$ over S', F, for which the agent can compute the EFE with respect to the preferences:

$$\mathrm{G}(\pi) := \mathrm{G}\left(\begin{array}{c} S' \quad F \\ \boxed{M_q} \\ \overset{\pi}{\triangledown} \end{array}, \begin{array}{c} F \\ \boxed{C} \end{array}\right) \tag{51}$$

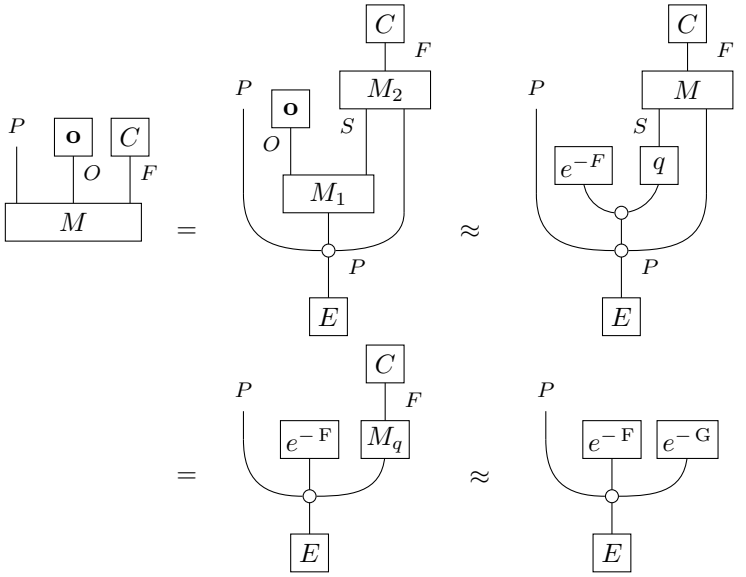
Using these free energy quantities, the agent may carry out approximate active inference. The following formula is central in the active inference literature.

Theorem 39. *The agent can carry out approximate active inference given observation $\boldsymbol{o}$ and preferences C by setting its plan to have density*

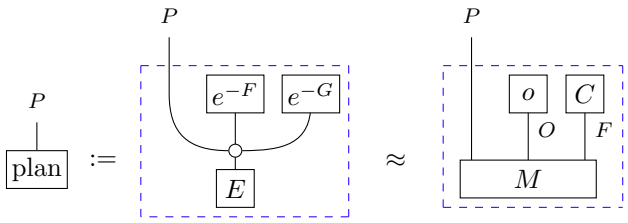
$$\mathrm{plan}(\pi) := \sigma(\log E(\pi) - F(\pi) - G(\pi)) \tag{52}$$

where σ denotes a softmax over $\pi \in P$.

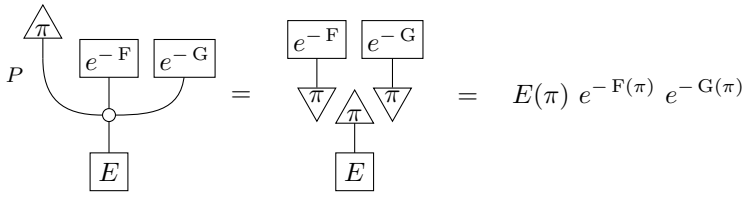
Proof. We have:



where we used Corollary 38 in both approximation steps. Thus defining our plan as on the left-hand below yields an approximate update:



But finally, note that the left-hand distribution is precisely given by (52). Indeed for each $\pi \in P$, corresponding to a sharp effect on P, we have:



Hence the normalisation of the above is precisely the softmax expression (52). □

This formula for active inference via free energy, though frequently used, is usually only justified in a fairly heuristic manner [PPF22]. Previous accounts rely on the less clear notion of treating EFE as a 'prior' to updating[9]. Here we have instead seen how the expression can be derived from a direct diagrammatic argument, directly from the structure of the generative model.

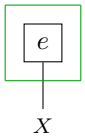
9 Compositionality of Free Energy

A crucial aspect of active inference is the idea that an agent can be understood to minimise free energy at all levels, so that it may be seen to globally minimise free energy in its generative model by minimising free energy within each component.

To formalise this idea we must first introduce a notion of free energy for open models. For this we will make use of the following graphical notation for the surprise.

[9]Despite the fact EFE is not straightforwardly a component of the generative model, and requires inference over present states S to be calculated first, rather than prior to them

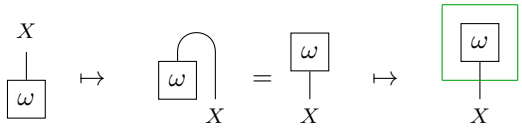
Definition 40. *Given any effect e on X in $\mathbf{Mat}_{\mathbb{R}^+}$, corresponding to a function $e\colon X \to \mathbb{R}^+$, we denote by*



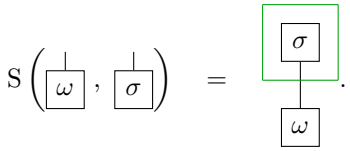
the function $-\log e(x)\colon X \to (-\infty, \infty]$.

Remark 41. *Note that a log-box is no longer an effect within $\mathbf{Mat}_{\mathbb{R}^+}$, since when $e(x) = 0$ we will have $-\log e(x) = \infty$. Here we will interpret any diagram involving log-boxes with inputs $X_1, \ldots, X_n$ and no outputs as a (formula specifying a) function $X_1 \times \cdots \times X_n \to (-\infty, \infty]$. Composing boxes in the diagram amounts to summation over wires, as for $\mathbf{Mat}_{\mathbb{R}^+}$. Given two such diagrams D_1, D_2 we write $D_1 + D_2$ for the function given by their point-wise sum as functions. In future it would be interesting to explore a formal categorical semantics for log-boxes.*

In particular we can apply a log box to any distribution ω in $\mathbf{Mat}_{\mathbb{R}^+}$ by first turning it into an effect, yielding the surprise $S(\omega)(x) = -\log \omega(x)$.



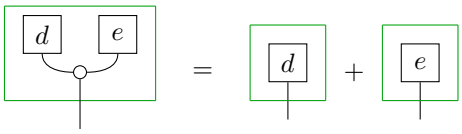
Similarly for a pair of distributions ω, σ we have

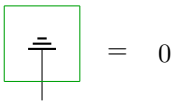


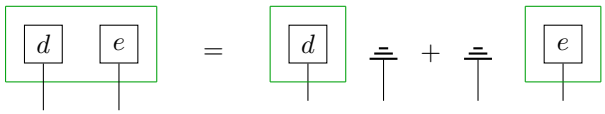
(53)

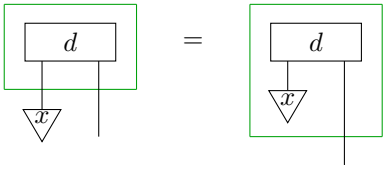
From the properties of the logarithm, one may verify that log-boxes then satisfy the following compositional properties.

Lemma 42. *For all effects d, e and sharp states x the following hold.*

1. 

2. 

3. 

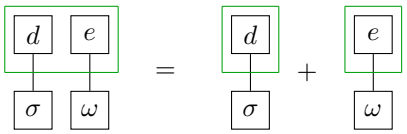
4. 

Proof. Plugging inputs x, y into each equation they reduce to the following respective properties of the logarithm. (1): $\log(d(x)e(x)) = \log(d(x)) + \log(e(x))$. (2): $\log(1) = 0$. (3): $\log(d(x)e(y)) = \log d(x) + \log e(y)$. (4) holds by definition, since both diagrams are given by $y \mapsto \log d(x, y)$. □

The following properties then follow from diagrammatic reasoning, using the relation between caps and copying.

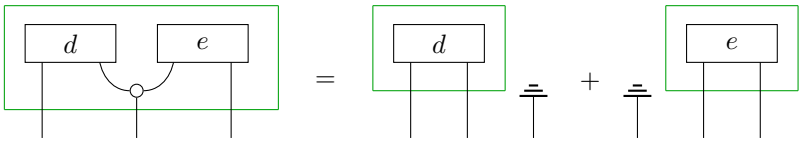
Proposition 43.

1. *For all effects d, e and normalised states σ, ω:*

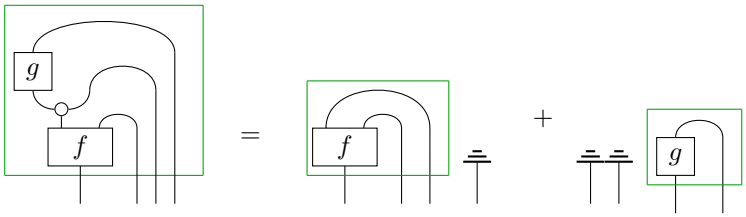


   In particular, entropy is additive across parallel composition: $H(\sigma \otimes \omega) = H(\sigma) + H(\omega)$.

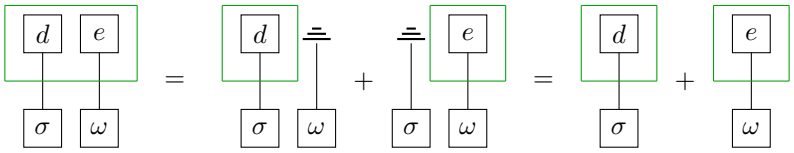
2. *For all effects d, e:*



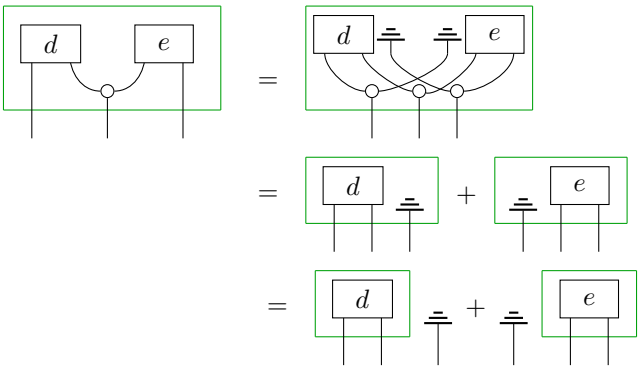
3. *For all morphisms f, g:*



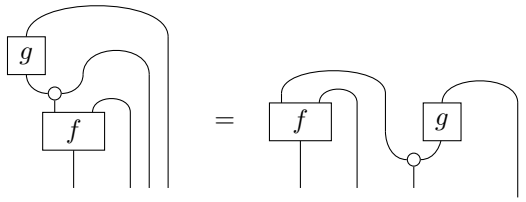
Proof. (1) follows from Lemma 42 (3) since:



(2) follows from Lemma 42 (1), 2 and 3 since:

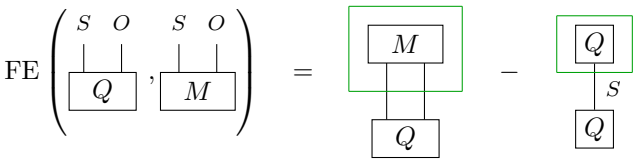


(3) is a special case of (2) where we define the effects d, e by composing f, g with caps on their output, respectively, since using the relation between caps and copying we see that:

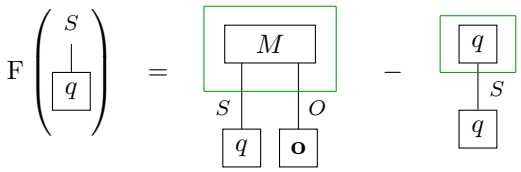


□

Now, by construction, for any joint distributions M, Q over S, O the free energy is given by:

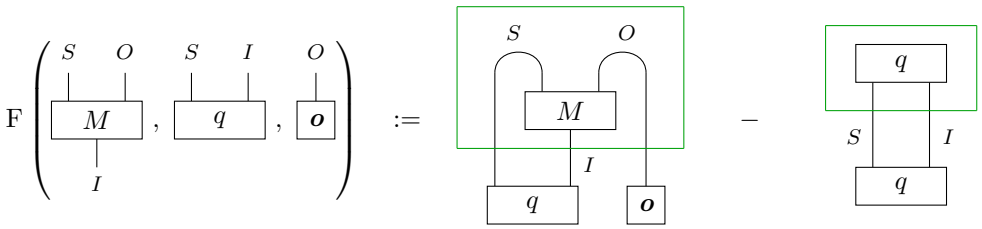


Hence for any joint distribution M over S, O and distributions $q, \mathbf{o}$ over S, O respectively, the VFE is given by:



We can use this to define a generalisation of VFE for (the channels induced by) open generative models.

Definition 44 (Open VFE)**.** *Given a channel $M\colon I \to S, O$, distribution q over I, S and distribution $\boldsymbol{o}$ over O, we define the* open Variational Free Energy *as*

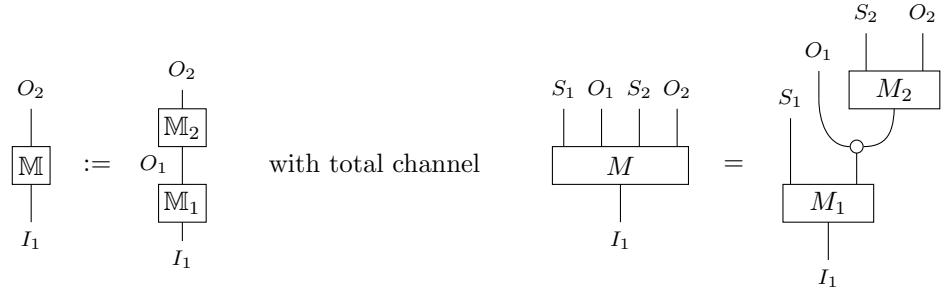


(54)

In the special case where I is trivial, the open VFE coincides with the usual VFE.

We can use the compositional properties of log boxes to show that this form of free energy is compositional, in an appropriate sense. First consider the following two ways in which we may compose open models.

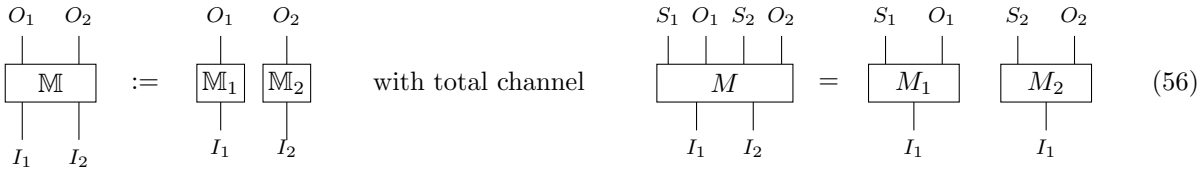
Consider a pair of open models $\mathbb{M}_1, \mathbb{M}_2$ with inputs I_1, I_2, outputs O_1, O_2 and hidden states S_1, S_2, respectively, such that $O_1 = I_2$. We can compose these in sequence into a single open model $\mathbb{M}$ from I_1 to O_2, with S_1, S_2, O_1 as its hidden states, with induced total channel M from I_1 to the remaining variables given as below.



(55)

Formally, the left-hand diagram is a composition in the category of open causal models; see [LT23, Sec. 5].

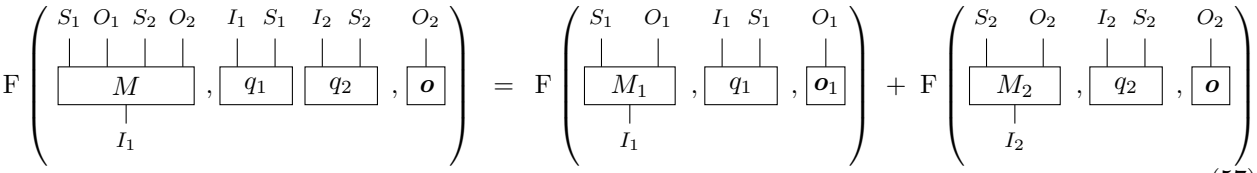
We can also compose open models in parallel. Given two open models $\mathbb{M}_i$ with inputs I_i, outputs O_i and hidden states S_i, for $i = 1, 2$ we can define an open model $\mathbb{M}$ with both sets of inputs, outputs and hidden states with induced induced total channel M from I_1, I_2 to the remaining variables given as below.



(56)

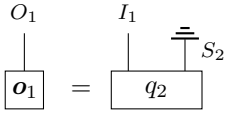
For each of these forms of composition of open models, we wish to establish that free energy is compositional in that the VFE of the open model $\mathbb{M}$ is determined from the VFE of its constituents. This ensures that locally minimising VFE (within each of sub-component) can achieve global VFE minimisation also.

Theorem 45. *For the sequential composite model $\mathbb{M}$ in (55) with $O_1 = I_2$, with total channel M, and any distribution $\boldsymbol{o}$ over O_2 and distributions q_1, q_2, the following holds:*



(57)

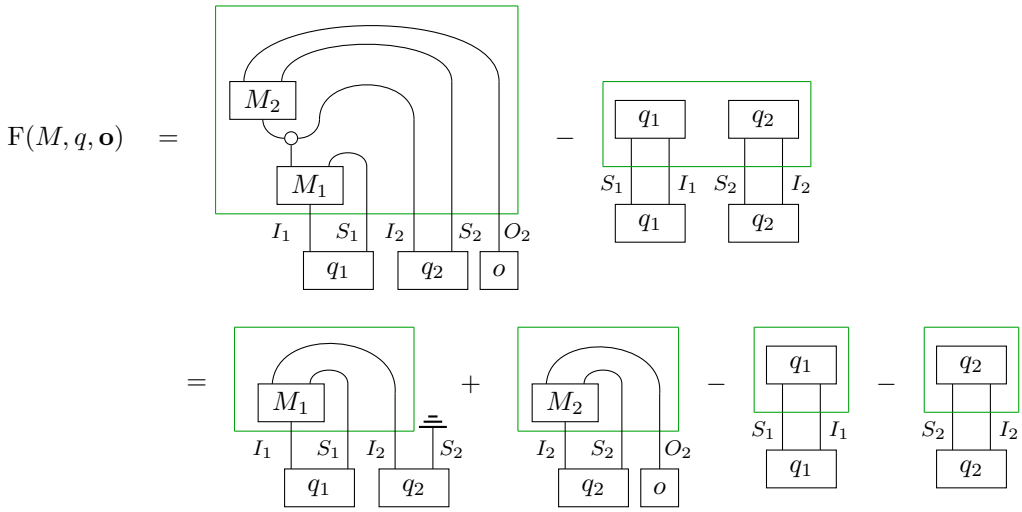
where



(58)

Intuitively, (58) expresses the way in which the beliefs about inputs I_2 in M_2 are passed down to the model $\mathbb{M}_1$ as observations in $O_1 = I_2$.

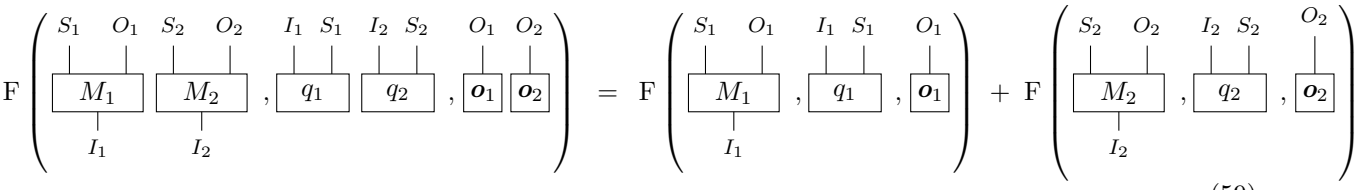
Proof. After rearranging some wires, we have that



where for the first two terms we apply Proposition 43 with $f = M_1$ and $g = M_2$, along with the fact that $\mathbf{o}$ and q_1 are normalised, and the second two terms are from Proposition 43 (1). But this is precisely the right-hand side of (57). □

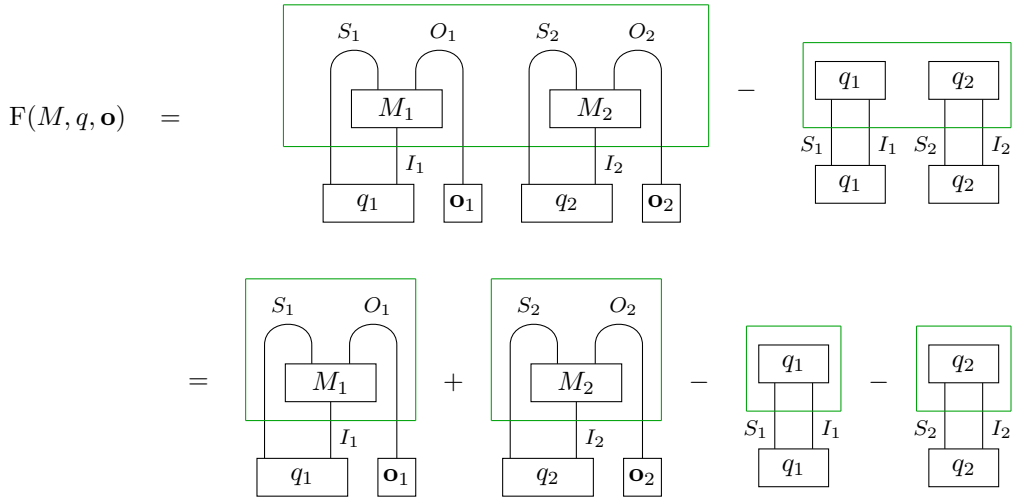
Next let us turn to the parallel composite of open models.

Theorem 46. *For any channels and distributions $M_i, q_i, \boldsymbol{o}_i$ for $i = 1, 2$, the following holds.*



(59)

Proof. The left-hand side is given by



which is precisely the right-hand side, where we applied Proposition 43 (1). □

The above results tell us that an agent with an overall generative model may minimise VFE by minimising VFE locally within each sub-model, an important property underlying the application of the free energy to all levels of a system.

10 Outlook

In this article we have aimed to give a concise formulation of active inference in terms of string diagrams interpreted in a cd-category $\mathbf{C}$, focusing on the case of finite discrete systems as described by $\mathbf{C} = \mathbf{Mat}_{\mathbb{R}^+}$. In particular we were able to derive the formula for approximate active inference via free energy minimisation purely from the high-level structure of a generative model undertaking active inference, and derived a compositionality property for free energy.

However these are just the first steps towards a fully compositional account of intelligent behaviour according to predictive processing, and there are many directions for future work.

Message passing So far we only studied active inference at a high-level, saying that an agent must, for each observation, arrive at an updated distribution q by free energy minimisation, without discussing how this is to be carried out. In PP this minimisation is normally achieved via so-called 'message passing' algorithms, such as 'variational' and 'marginal' message passing [PMKF19]. These are defined on undirected graphical models described by Forney factor graphs, induced by a generative model. In future it would be interesting to include a categorical account of message passing within our framework, to complete our description of active inference.

Continuous settings Another technical matter would be to extend the treatment of PP beyond the finite case to further cd-categories describing continuous settings, such as a suitable category of Gaussian probabilistic processes, which are widely employed in PP under the 'Laplace assumption'. One issue is

in extending our treatment of minimal conditionals to such continuous settings, where they are not as straightforwardly defined.

Causal reasoning We have here pointed out that an generative model may be seen precisely as a causal model [Pea09]. In future it would be interesting to explore how an agent may carry out causal reasoning on its model using concepts from the causal model framework such as 'interventions', as treated graphically in [LT23], and how such reasoning relates to active inference.

Approximations The treatment of active inference via free energy in Section 8 relied on applying various approximation steps from Section 7 to parts of the diagram. Certainly more could be done to set bounds on how well these approximations hold, including how they extend from part of a diagram to the whole generative model.

Updating within PP The categorical perspective led us to naturally consider soft observations (given by distributions) rather than the usual sharp ones (given by points), which come with distinct notions of Jeffrey and Pearl updating (Section 4), as well our new notion of VFE updating (Section 7). While we were able to describe active inference via the latter two forms of updating, it would be interesting to compare against Jeffrey updating and establish which form of exact updating is most naturally considered (and approximated) in PP. That is, given that both forms of updating have different goals [Jac19], which one (if either) is approximately carried out by the brain? This question was also raised in [DLR23].

We note that Pearl updating can be more generally defined with respect to any effect (see e.g. [DLR23]), i.e. any (not necessarily normalised) function. There is disagreement between active inference and reinforcement learning (RL) in whether an agent's preferences should, rather than as a distribution as in active inference, be simply modelled by a function $C\colon F \to \mathbb{R}^+$ assigning a 'value' in $\mathbb{R}^+$ to each possible future observation, i.e. as an effect C on F [FDK09, TMSB20]. In this case Pearl updating may be the most natural to treat planning. In contrast, Jeffrey updating may be most fitting for perception, with an observation $\mathbf{o}$ naturally encoded as a distribution i.e. a 'fuzzy point' in O.

Consciousness in PP Various proposals have been put forward for how PP and active inference can be related to consciousness. Continuing from previous work from two of the authors on IIT [KT21, TK21], in future we hope to account for these proposals within our graphical account of active inference.

Categorical modifications of PP Beyond simply recasting previous results in PP categorically, in future one may also study what new insights the compositional perspective may bring to PP and active inference, and to connect the work to ongoing research within categorical cybernetics [Smi21b, CGHR21] and more broadly to the research programme of compositional intelligence.

References

[AC04] Samson Abramsky and Bob Coecke. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.*, pages 415–425. IEEE, 2004.

[CCS08] Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140. Citeseer, 2008.

[CGHR21] Matteo Capucci, Bruno Gavranović, Jules Hedges, and Eigil Fjeldgren Rischel. Towards foundations of categorical cybernetics. *arXiv preprint arXiv:2105.06332*, 2021.

[CJ19] Kenta Cho and Bart Jacobs. Disintegration and bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

[Coe06] Bob Coecke. Introducing categories to the practicing physicist. In *What is category theory*, pages 45–74. Polimetrica Monza, 2006.

[CS12] Bob Coecke and Robert W Spekkens. Picturing classical and quantum bayesian inference. *Synthese*, 186:651–696, 2012.

[Dea21] George Deane. Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution. *Neuroscience of Consciousness*, 2021(2):niab024, 2021.

[DLR23] Elena Di Lavore and Mario Román. Evidential decision theory via partial markov categories. *arXiv preprint arXiv:2301.12989*, 2023.

[DVF17] Bert De Vries and Karl J Friston. A factor graph description of deep temporal active inference. *Frontiers in computational neuroscience*, 11:95, 2017.

[FDK09] Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7):e6421, 2009.

[FFR+17] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

[FK23] Tobias Fritz and Andreas Klingler. The d-separation criterion in categorical probability. *J. Mach. Learn. Res*, 24(46):1–49, 2023.

[FKH06] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of physiology-Paris*, 100(1-3):70–87, 2006.

[Fon13] Brendan Fong. Causal theories: A categorical perspective on bayesian networks. *arXiv preprint arXiv:1301.6201*, 2013.

[FPdV17] Karl J Friston, Thomas Parr, and Bert de Vries. The graphical brain: belief propagation and active inference. *Network neuroscience*, 1(4):381–414, 2017.

[Fri10] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[Fri20] Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

[FST19] Brendan Fong, David Spivak, and Rémy Tuyéras. Backprop as functor: A compositional perspective on supervised learning. In *2019 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–13. IEEE, 2019.

[GHWZ18] Neil Ghani, Jules Hedges, Viktor Winschel, and Philipp Zahn. Compositional game theory. In *Proceedings of the 33rd annual ACM/IEEE symposium on logic in computer science*, pages 472–481, 2018.

[Hoh20] Jakob Hohwy. New directions in predictive processing. *Mind & Language*, 35(2):209–223, 2020.

[HS20] Jakob Hohwy and Anil Seth. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, 1(II), 2020.

[Jac19] Bart Jacobs. The mathematics of changing one's mind, via jeffrey's or via pearl's update rule. *Journal of Artificial Intelligence Research*, 65:783–806, 2019.

[JKZ19] Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal inference by string diagram surgery. In *Foundations of Software Science and Computation Structures: 22nd International Conference, FOSSACS 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 22*, pages 313–329. Springer, 2019.

[KT21] Johannes Kleiner and Sean Tull. The mathematical structure of integrated information theory. *Frontiers in Applied Mathematics and Statistics*, 6:602973, 2021.

[LT23] Robin Lorenz and Sean Tull. Causal models in string diagrams. *arXiv preprint arXiv:2304.07638*, 2023.

[Pan98] Prakash Panangaden. Probabilistic relations. *School of Computer Science Research Reports-University of Birmingham CSR*, pages 59–74, 1998.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Per22] Paolo Perrone. Markov categories and entropy. *arXiv preprint arXiv:2212.11719*, 2022.

[PMKF19] Thomas Parr, Dimitrije Markovic, Stefan J Kiebel, and Karl J Friston. Neuronal message passing using mean-field, bethe, and marginal approximations. *Scientific reports*, 9(1):1889, 2019.

[PPF22] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

[PZ23] Robin Piedeleu and Fabio Zanasi. An introduction to string diagrams for computer scientists. *arXiv preprint arXiv:2305.08768*, 2023.

[SBPF21] Noor Sajid, Philip J Ball, Thomas Parr, and Karl J Friston. Active inference: demystified and compared. *Neural computation*, 33(3):674–712, 2021.

[SFW22] Ryan Smith, Karl J Friston, and Christopher J Whyte. A step-by-step tutorial on active inference and its application to empirical data. *Journal of mathematical psychology*, 107:102632, 2022.

[SGW21] Dan Shiebler, Bruno Gavranović, and Paul Wilson. Category theory in machine learning. *arXiv preprint arXiv:2106.07032*, 2021.

[Smi20] Toby St Clere Smithe. Bayesian updates compose optically. *arXiv preprint arXiv:2006.01631*, 2020.

[Smi21a] Toby St Clere Smithe. Compositional active inference i: Bayesian lenses. statistical games. *arXiv preprint arXiv:2109.04461*, 2021.

[Smi21b] Toby St Clere Smithe. Cyber kittens, or some first steps towards categorical cybernetics. *arXiv preprint arXiv:2101.10483*, 2021.

[Smi22] Toby St Clere Smithe. Compositional active inference ii: Polynomial dynamics. approximate inference doctrines. *arXiv preprint arXiv:2208.12173*, 2022.

[TK21] Sean Tull and Johannes Kleiner. Integrated information in process theories. *Journal of Cognitive Science*, 2021.

[TMSB20] Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*, 2020.

[WM17] Wanja Wiese and Thomas Metzinger. Vanilla pp for philosophers: A primer on predictive processing. 2017.