Learning from What's Right and Learning from What's Wrong

Bart Jacobs

Institute for Computing and Information Sciences (iCIS),
Radboud University Nijmegen, The Netherlands.

bart@cs.ru.nl

The concept of updating (or conditioning or revising) a probability distribution is fundamental in (machine) learning and in predictive coding theory. The two main approaches for doing so are called Pearl's rule and Jeffrey's rule. Here we make, for the first time, mathematically precise what distinguishes them: Pearl's rule increases validity (expected value) and Jeffrey's rule decreases (Kullback-Leibler) divergence. This forms an instance of a more general distinction between learning from what's right and learning from what's wrong. The difference between these two approaches is illustrated in a mock cognitive scenario.

1 Introduction

Intuitively, people can learn by reinforcing what goes well, or by steering away from what goes wrong: they can go even higher, or go lower. In the first case one improves a positive evaluation and in the second case one reduces a negative outcome. In this paper we shall refer to the first approach as learning from what's right, or more simply, from rightness. The second approach is described in terms of learning from what's wrong, or from wrongness.

Learning is at the heart of the current AI-revolution. In a mathematical setting learning involves adapting/updating parameters, with respect to some objective (expressed as 'objective' function). Also in such a setting one can distinguish whether this adaptation is guided by increasing what is right, or by decreasing what is wrong. Learning from rightness can be done by increasing a positive evaluation, like reward, match, likelihood or validity. Learning from wrongness happens by decreasing a negative evaluation, like an error, loss, penalty, divergence or distance. This distinction between learning from rightness/wrongness is not new and may be expressed alternatively for instance in terms of reward/error-based learning. Also, the distinction is not absolute, since what's good in one context may be bad in another.

In probabilistic learning there are two different approaches to updating, namely following Pearl [29] (and Bayes) or following Jeffrey [23], see for comparisons *e.g.* [3, 27, 7, 19]. The two approaches can give completely different outcomes, but it is poorly understood when to use which approach. For instance, it is suggested by [7] that Jeffrey's rule is most appropriate for correction after a 'surprise', but it remains vague what a surprise is. At a conceptual level, the main contribution of this paper lies in showing that Pearl's approach is learning from rightness, and Jeffrey's approach is learning from wrongness. The objective function that is used here for rightness is validity (that is, expected value), and for wrongness it is divergence (in Kullback-Leibler form). Thus, it will be shown that Pearl's update rule increases validity and Jeffrey's rule decreases divergence. The latter divergence-decrease result is the main mathematical contribution of this paper. Its proof makes use of rather heavy mathematical machinery,

Ana Sokolova (Eds.): MFPS 2021
EPTCS 351, 2021, pp. 116–133, doi:10.4204/EPTCS.351.8

taken mainly from [8]; it is relegated to the appendix. The fact that Pearl's approach increases validity is mathematically less complicated and the relevant parts of this claim have already been published *e.g.* in [18, 22].

Probabilistic learning typically involves an iterative process where each single step yields an improvement w.r.t. an objective function. This paper concentrates on these single steps, since its focus is on capturing the (mathematical) difference between Pearl and Jeffrey. What happens when these single learning steps are iterated is a topic in itself, which is not covered here.

This paper builds on the mathematical formalisations of the rules of Jeffrey and Pearl introduced in [19] — where notably Jeffrey's rule is captured via a 'dagger'. In fact, had the main result of this paper (Jeffrey reduces divergence) been known at the time of writing [19], it would have fitted perfectly there. Alas, insights come slowly, and so a separate paper is written now, as an addendum to [19]. The current addendum is much more mathematical in nature than [19], since the proof of our main result is non-trivial. In addition, this addendum explains the results in a more cognition-oriented language.

We thus start from the mathematical formalisation of [19] that uses (discrete, finite) probability distributions (also called states), fuzzy (soft) predicates, and channels (conditional distributions), together with operations such as state/predicate transformation along a channel and updating a distribution with a predicate. Within this framework the update rules of Pearl and Jeffrey are applied in a common setting, which we briefly introduce, without explaining all details yet. One starts from a distribution σ on some set X together with a channel c from X to Y, that is, with a conditional probability distribution $p(y \mid x)$, or, more categorically, with a Kleisli map of the distribution monad $\mathscr{D}$. Along this channel one can transform (push forward) the distribution σ on X to a distribution $c \gg\!= \sigma$ on Y, namely $(c \gg\!= \sigma)(y) = \sum_x \sigma(x) \cdot p(y \mid x)$. This $c \gg\!= \sigma$ can be seen as a prediction. We consider the situation where we are confronted with new information (evidence) on Y which leads us to update σ to a new distribution σ'. Pearl and Jeffrey provide two different rules for performing this update, which increase validity and decrease divergence, respectively. (In this paper we only consider updating the state σ along a channel c, but one may go further and update the mediating channel c as well, like in Expectation Maximisation, see [18]). This validity-increase and divergence-decrease is characteristic for the rules of Pearl and Jeffrey: we demonstrate that Jeffrey's rule, in general, does not give a validity increase, and similarly, that Pearl's rule need not give a divergence-decrease (see Remark 1).

Interestingly, the channel-based setting fits the neuroscientific setting that underlies predictive coding theory (also called predictive processing or free energy principle). This theory goes back to Hermann von Helmholtz in the 19th century and is described in modern terms first by [30] and in many other recent sources, *e.g.* [10, 15, 5]. Naively, humans learn by absorbing sensory information from the outside world and by building up a more or less accurate internal picture. Alternatively, the mind projects, evaluates and updates: predictive coding theory describes the human mind basically as a Bayesian prediction engine that compares its predictions to observations, leading to internal adaptations. To quote Friston [10]: "The *Bayesian brain hypothesis* uses Bayesian probability theory to formulate perception as a constructive process based on internal or generative models. [...] In this view, the brain is an inference machine that actively predicts and explains its sensations. Central to this hypothesis is a probabilistic model that can generate predictions, against which sensory samples are tested to update beliefs about their causes." We translate this to the above setting: the mind's internal state may be (partially) represented by a distribution σ on X, as used in the previous paragraph. The channel c is part of the generative model that produces the prediction $c \gg\!= \sigma$, as distribution on the external world Y. Confronted with (mismatching) sensory information (about Y), the brain updates its internal state σ (on X). This is how learning happens in the predictive model. This paper uses a running example of this kind.

An intriguiging question is: does this learning/updating happen according to Pearl or to Jeffrey? For-

mulated more abstractly, does the mind learn from what's right or from what's wrong? An (empirical) answer to that question lies far beyond this paper, but predictive coding theory suggests that our minds use Jeffrey's rule since they try to minimise prediction errors, see [10]. An additional argument in this direction is that successive Pearl-updates commute, but successive Jeffrey-updates do not, see [19] for details. It is well-known that the human mind is highly sensitive to the order in which it processes information (or: is primed/updated). The main result of this paper (Theorem 3) strengthens the mathematical basis of predictive coding theory: it extends learning from point data to learning from distributions and shows that in such learning from distributions the prediction error, expressed as Kullback-Leibler divergence, is reduced.

The structure of this paper is simple: after introducing preliminaries in Section 2, Pearl's and Jeffrey's update rules are described in Sections 3 and 4, following [19]. Section 5 contains the main new result of this paper, namely that Jeffrey's rule decreases divergence. Its relevance to predictive coding theory is explained in Section 6. Finally, the appendix contains a proof of the main result.

2 Prelimaries on states, channels and prediction

This section introduces basic concepts and fixes notation. Suppose you mix paint of different colours, say with ratio $\frac{1}{2}$ red, $\frac{1}{8}$ green and $\frac{3}{8}$ blue. In that case we can write the paint distribution as a formal sum:

$$\tfrac{1}{2}|R\rangle + \tfrac{1}{8}|G\rangle + \tfrac{3}{8}|B\rangle.$$

The letters represent the different colours; they are written between 'ket' brackets $|-\rangle$ which are borrowed from quantum physics. The kets are meaningless notation that serve to separate the items in the distribution from their frequencies (or probabilities).

In general a (finite, discrete) *probability distribution* over a set X is a finite formal sum of the form $\sum_i r_i|x_i\rangle$ where the $r_i \in [0,1]$ are probabilities that add up to one: $\sum_i r_i = 1$. The x_i are members of the set X. Such a distribution can also be written as a function $\omega\colon X \to [0,1]$ with finite support, that is with only finitely many $x \in X$ with $\omega(x) \neq 0$. We then have $r_i = \omega(x_i)$. We say that ω has *full support* when $\omega(x) > 0$ for each $x \in X$ — which implicitly requires that the set X is finite. We use 'state' as synonym for 'distribution', but the term 'multinomial' is also common in the literature. We freely switch between the formal sum notation $\omega = \sum_i r_i|x_i\rangle$ and the function notation $\omega\colon X \to [0,1]$. We write $\mathscr{D}(X)$ for the set of distributions on the set X. This $\mathscr{D}$ is the (finite, discrete) distribution monad on the category of sets.

A *channel* is a Kleisli map for this distribution monad, that is, a function of the form $c\colon X \to \mathscr{D}(Y)$. We say that c is a channel from X to Y and often write this as $c\colon X \nrightarrow Y$, with a small circle on the shaft of the arrow. Such a channel gives a distribution $c(x)$ on Y for each $x \in X$. It is thus a conditional distribution, which is commonly written as $p(y \mid x)$. Alternatively, when X, Y are finite, we can see the channel as a stochastic matrix. Each function $f\colon X \to Y$ gives rise to a 'deterministic' channel $\langle f \rangle\colon X \nrightarrow Y$, via $\langle f \rangle(x) = 1|f(x)\rangle$. Channels have a lot of algebraic structure: they can be composed sequentially and also in parallel — they form the morphisms of a symmetric monoidal (Kleisli) category, see *e.g.* [17, 16]. Channels are becoming popular in a principled, axiomatic approach to probability, see *e.g.* [12, 22, 21, 20].

Let $c\colon X \nrightarrow Y$ be a channel from X to Y. We can then define *state tranformation* along c as a function $\mathscr{D}(X) \to \mathscr{D}(Y)$. It maps a distribution σ on X to a 'transformed' distribtution $c \gg\!= \sigma$ on Y, via the

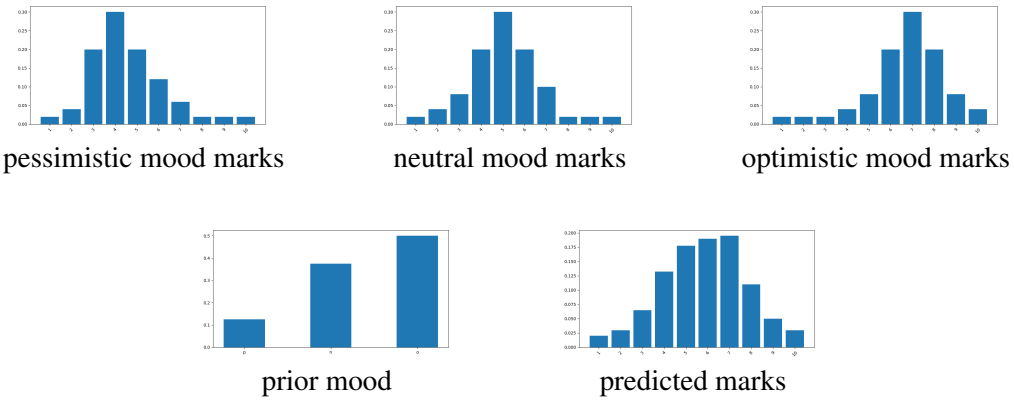


Figure 1: Distributions occurring in Example 1.

following definition: for $y \in Y$,

$$\big(c \gg\!= \sigma\big)(y) \;=\; \sum_{x \in X} \sigma(x) \cdot c(x)(y).$$

This new distribution is often called the *prediction*. This will be illustrated next in our leading example.

Example 1. We consider a very simple situation as instantiation of predictive coding. Assume we use only three possible options to describe the mood of a teacher, namely: pessimistic (p), neutral (n) or optimistic (o). We thus have a three-element probability space $X = \{p,n,o\}$. We assume an *a priori* mood distribution:

$$\sigma \;=\; \tfrac{1}{8}|p\rangle + \tfrac{3}{8}|n\rangle + \tfrac{1}{2}|o\rangle.$$

This mood thus tends towards optimism.

Associated with these different moods the teacher has different views on how pupils perform in a particular test. This performance is expressed in terms of marks, which can range from 1 to 10, where 10 is best. The probability space for these marks is written as $Y = \{1,2,\ldots,10\}$.

The view of the teacher is expressed via a channel $c \colon X \nrightarrow Y$ with:

$$\begin{aligned}
c(p) &= \tfrac{1}{50}|1\rangle + \tfrac{2}{50}|2\rangle + \tfrac{10}{50}|3\rangle + \tfrac{15}{50}|4\rangle + \tfrac{10}{50}|5\rangle + \tfrac{6}{50}|6\rangle + \tfrac{3}{50}|7\rangle + \tfrac{1}{50}|8\rangle + \tfrac{1}{50}|9\rangle + \tfrac{1}{50}|10\rangle \\
c(n) &= \tfrac{1}{50}|1\rangle + \tfrac{2}{50}|2\rangle + \tfrac{4}{50}|3\rangle + \tfrac{10}{50}|4\rangle + \tfrac{15}{50}|5\rangle + \tfrac{10}{50}|6\rangle + \tfrac{5}{50}|7\rangle + \tfrac{1}{50}|8\rangle + \tfrac{1}{50}|9\rangle + \tfrac{1}{50}|10\rangle \\
c(o) &= \tfrac{1}{50}|1\rangle + \tfrac{1}{50}|2\rangle + \tfrac{1}{50}|3\rangle + \tfrac{2}{50}|4\rangle + \tfrac{4}{50}|5\rangle + \tfrac{10}{50}|6\rangle + \tfrac{15}{50}|7\rangle + \tfrac{10}{50}|8\rangle + \tfrac{4}{50}|9\rangle + \tfrac{2}{50}|10\rangle.
\end{aligned}$$

These three outcomes are plotted in the first row in Figure 1. They clearly show that the better the mood, the better the marks.

The second row in Figure 1 describes the mood distribution $\sigma \in \mathscr{D}(X)$ and the predicted marks $c \gg\!= \sigma \in \mathscr{D}(Y)$, for this mood distribution. The latter is a convex combination of the three plots in the top row, where the weights are determined by σ.

This example will be continued below. The teacher will be confronted with the marks that the pupils actually obtain. This will lead the teacher to an update of his/her own mood, in two possible (different) ways, according to Pearl and according to Jeffrey.

3 Pearl's updating, increasing what's right

Before describing Pearl's updating, we collect the relevant notions and definitions, especially about predicates, validity, updating and predicate transformations. The new material starts in Subsection 3.2.

3.1 Predicates and updating

Probabilistic revision involves updating a distribution on the basis of evidence. Traditionally this evidence takes the form of an event, that is a subset $E \subseteq X$ of the probability space X. A useful, more general approach uses (fuzzy) predicates as evidence. They are functions of the form $p\colon X \to [0,1]$, with characteristic functions of subsets as a special 'sharp' case. A point predicate, for $x \in X$, is a special (sharp) predicate $\mathbf{1}_x\colon X \to [0,1]$ sending $x' \neq x$ to 0 and x to 1. Every predicate p on a finite set X can be described as finite sum $\sum_x p(x) \cdot \mathbf{1}_x$. We shall write $\mathrm{Pred}(X) = [0,1]^X$ for the set of predicates on X.

For a distribution $\omega \in \mathscr{D}(X)$ and a predicate $p \in \mathrm{Pred}(X)$ on the same set we write the *validity* of the predicate/evidence p in the state ω as $\omega \models p$. It is a number in $[0,1]$, defined as expected value:

$$\omega \models p \;=\; \sum_{x \in X} \omega(x) \cdot p(x).$$

When this validity is non-zero, we define the updated state $\omega|_p \in \mathscr{D}(X)$ as the normalised product:

$$\omega|_p(x) \;=\; \frac{\omega(x) \cdot p(x)}{\omega \models p}.$$

This updating satisfies some important properties, including Bayes' law and the multiple update law below.

$$\omega|_p \models q \;=\; \frac{\omega \models p \,\&\, q}{\omega \models p} \;=\; \frac{(\omega|_q \models p) \cdot (\omega \models q)}{\omega \models p} \qquad\qquad \omega|_p|_q \;=\; \omega|_{p \& q}, \tag{1}$$

where the conjunction $p \,\&\, q$ is pointwise multiplication: $(p \,\&\, q)(x) = p(x) \cdot q(x)$. For more details,see *e.g.* [22, 19, 17].

Each channel $c\colon X \to\!\!\!\!\bullet\, Y$ gives rise to a predicate transformation function $\mathrm{Pred}(Y) \to \mathrm{Pred}(X)$, acting in the opposite direction. For a predicate $q\colon Y \to [0,1]$ one gets $c \mathbin{=\!\!\ll} q\colon X \to [0,1]$ via:

$$\big(c \mathbin{=\!\!\ll} q\big)(x) \;=\; \sum_{y \in Y} c(x)(y) \cdot q(y).$$

When a channel is identified with a conditional probability table in a Bayesian network, as is done by [22], this predicate transformation corresponds to propagation of evidence along the channel, as path, see [28, §4.3.1].

State and predicate transformation $\gg\!\!=$ and $=\!\!\ll$ are closely related via validity $\models$, since:

$$(c \gg\!\!= \omega) \models q \;=\; \omega \models (c \mathbin{=\!\!\ll} q). \tag{2}$$

3.2 Pearl's updating, increasing validity

The idea of an update $\omega|_p$ is that the evidence p is incorporated into the state ω. Hence it is to be expected that p is 'more true' in $\omega|_p$ than in ω. That is the content of the next 'update rightness' result, mentioned also in [18], but with a different proof.

Theorem 1 (Update rightness)**.** *For a distribution ω and a predicate p on the same set, if the validity $\omega \models p$ is non-zero, one has:*

$$\omega|_p \models p \;\geq\; \omega \models p.$$

Proof. We show that the difference is non-negative:

$$\begin{aligned}
\big(\omega|_p \models p\big) - \big(\omega \models p\big) &\overset{(1)}{=} \tfrac{1}{\omega\models p} \cdot \Big(\omega \models p \mathbin{\&} p \;-\; (\omega \models p)^2\Big) \\
&= \tfrac{1}{\omega\models p} \cdot \Big(\big(\textstyle\sum_x \omega(x) \cdot p(x)^2\big) \;-\; 2(\omega \models p) \cdot (\omega \models p) \;+\; (\omega \models p)^2\Big) \\
&= \tfrac{1}{\omega\models p} \cdot \Big(\big(\textstyle\sum_x \omega(x) \cdot p(x)^2\big) \;-\; 2\big(\textstyle\sum_x \omega(x) \cdot p(x)\big) \cdot (\omega \models p) \\
&\qquad\qquad + \big(\textstyle\sum_x \omega(x) \cdot (\omega \models p)^2\big)\Big) \\
&= \tfrac{1}{\omega\models p} \cdot \textstyle\sum_x \omega(x) \cdot \Big(p(x)^2 \;-\; 2p(x) \cdot (\omega \models p) \;+\; (\omega \models p)^2\Big) \\
&= \tfrac{1}{\omega\models p} \cdot \textstyle\sum_x \omega(x) \cdot \Big(p(x) - (\omega \models p)\Big)^2 \;\geq\; 0.
\end{aligned}$$

□

We now describe Pearl's update rule, following [19]. The setting is like in predictive coding, as sketched in the introduction, with a prediction $c \gg\!= \sigma$ and a confrontation with external information. In Pearl's setting this information is evidence, in the form of a predicate.

Theorem 2 (Pearl's update)**.** *Let $c \colon X \nrightarrow Y$ be a channel with a (prior) state $\sigma \in \mathscr{D}(X)$ on its domain X. For a predicate q on the codomain Y of the channel we get an increase of validity (rightness):*

$$(c \gg\!= \sigma_P) \models q \;\geq\; (c \gg\!= \sigma) \models q \quad \textit{for the updated/posterior state} \quad \sigma_P \,=\, \sigma|_{c \mathrel{=\!\!\ll} q}.$$

The update mechanism $\sigma \mapsto \sigma_P = \sigma|_{c \mathrel{=\!\!\ll} q}$ is Pearl's update rule.

Proof. By combining Theorem 1 with (2) we get:

$$\Big(c \gg\!= \big(\sigma|_{c \mathrel{=\!\!\ll} q}\big)\Big) \models q \;=\; \sigma|_{c \mathrel{=\!\!\ll} q} \models (c \mathrel{=\!\!\ll} q) \;\geq\; \sigma \models (c \mathrel{=\!\!\ll} q) \;=\; (c \gg\!= \sigma) \models q.$$

□

The update rule of Pearl involves an update with a transformed predicate. It forms the basis of probabilistic reasoning in Bayesian networks. Indeed, the conditional probability tables of such networks are channels and reasoning happens by transforming states and predicates up and down these channels, in combination with updating at appropriate points. This perspective comes from [28] and is elaborated by [22] in channel-based form.

Example 2. We continue in the setting of Example 1 and assume that the pupils have done rather poorly, with no-one scoring above 5, as described by the following evidence/predicate q on the set of grades $Y = \{1, 2, \ldots, 10\}$.

$$q \,=\, \tfrac{1}{10} \cdot \mathbf{1}_1 + \tfrac{3}{10} \cdot \mathbf{1}_2 + \tfrac{3}{10} \cdot \mathbf{1}_3 + \tfrac{2}{10} \cdot \mathbf{1}_4 + \tfrac{1}{10} \cdot \mathbf{1}_5.$$

The validity of this predicate q in the predicted state $c \gg\!= \sigma$ is:

$$c \gg\!= \sigma \models q \;=\; \sigma \models c \mathrel{=\!\!\ll} q \;=\; \tfrac{299}{4000} \,=\, 0.07475.$$

The interested reader may wish to check that the Pearl-update $\sigma_P = \sigma|_{c \mathrel{=\!\!\ll} q}$ and the resulting increased validity of q are:

$$\begin{aligned}
\sigma_P \;&=\; \tfrac{77}{299}|p\rangle + \tfrac{162}{299}|n\rangle + \tfrac{60}{299}|o\rangle \;\approx\; 0.2575|p\rangle + 0.5418|n\rangle + 0.2007|o\rangle \\
c \gg\!= \sigma_P \models q \;&=\; \tfrac{15577}{149500} \;\approx\; 0.1042.
\end{aligned}$$

We thus see an increase of validity, roughly from 0.07 to 0.10.

4 Jeffrey's updating, decreasing wrongness

Before we can describe Jeffrey's update rule we need to introduce some additional background material, about Kullback-Leibler divergence and about the 'dagger' inverse of a channel.

4.1 Divergence and inversion

There are several ways to measure the difference between two distributions on the same set. Here we shall use the so-called Kullback-Leibler divergence, which is quite standard. For $\omega, \rho \in \mathscr{D}(X)$ it is defined as:

$$D_{KL}\big(\omega, \rho\big) \;=\; \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right).$$

Here we use the natural logarithm ln, where sometimes the 2-logarithm is used.

One can show that $D_{KL}(\omega, \rho) \geq 0$ and $D_{KL}(\omega, \rho) = 0$ if and only if $\omega = \rho$. In general one has $D_{KL}(\omega, \rho) \neq D_{KL}(\rho, \omega)$, so that D_{KL} is not a metric distance function. Indeed, it is called divergence and not distance. One can show that state transformation is divergence-decreasing, in the sense that $D_{KL}(c \gg\!= \omega, c \gg\!= \rho) \leq D_{KL}(\omega, \rho)$, but we don't need that property here.

We now look at inversion of a channel $c \colon X \nrightarrow Y$, where we assume a distribution $\sigma \in \mathscr{D}(X)$ on its domain. We turn it into a channel $Y \nrightarrow X$ in the other direction, written as $c^{\dagger}_{\sigma} \colon Y \nrightarrow X$, and defined as:

$$c^{\dagger}_{\sigma}(y)(x) \;=\; \sigma|_{c =\!\ll \mathbf{1}_y}(x) \;=\; \frac{\sigma(x) \cdot (c =\!\ll \mathbf{1}_y)(x)}{\sigma \models c =\!\ll \mathbf{1}_y} \;=\; \frac{\sigma(x) \cdot c(x)(y)}{(c \gg\!= \sigma)(y)}. \tag{3}$$

The latter formulation shows that we need to require that the predicted state $c \gg\!= \sigma$ has full support.

If channel c represents a conditional probability $p(y \mid x)$, then its inversion $c^{\dagger}_{\sigma}$ corresponds to the Bayesian inversion $p(x \mid y)$. Such Bayesian inversions play a basic role in predictive coding, see *e.g.* [11]. The dagger notation is used because this inversion behaves like in so-called dagger categories, see [6, 4, 12] for more information. Such daggers/inversions are also used to capture the reversibility of quantum computations via conjugate transposes, see for further information *e.g.* [1].

5 Jeffrey's updating, decreasing divergence

The main result (Theorem 3) below states that Jeffrey's update rule decreases divergence. The proof is non-trivial and can be found in the appendix.

Recall that in Pearl's updating in Theorem 2 the prediction $c \gg\!= \sigma$ is confronted with external evidence, in the form of a predicate. In Jeffrey's case one uses an external state (distribution) as evidence, instead of a predicate.

Theorem 3. *Let $c \colon X \nrightarrow Y$ be a channel, whose codomain Y is a finite set, with a state $\sigma \in \mathscr{D}(X)$ on its domain, such that the predicted state $c \gg\!= \sigma$ on Y has full support. For an 'evidence' state $\tau \in \mathscr{D}(Y)$ there is a reduction of divergence:*

$$D_{KL}\big(\tau, c \gg\!= \sigma_J\big) \;\leq\; D_{KL}\big(\tau, c \gg\!= \sigma\big) \qquad \textit{for} \qquad \sigma_J \;=\; c^{\dagger}_{\sigma} \gg\!= \tau.$$

The update mechanism $\sigma \mapsto \sigma_J = c^{\dagger}_{\sigma} \gg\!= \tau$ is Jeffrey's update rule, see [19]. □

There is an earlier result describing the effect of Jeffrey's update rule, given for instance by [14, Prop. 3.11.2]. Translated to the current context it describes the divergence between the orginal state σ and its Jeffrey update as infimum:

$$D_{KL}\big(\sigma, \langle f\rangle^{\dagger}_{\sigma} \gg\!= \tau\big) \;=\; \bigwedge\{D_{KL}\big(\sigma,\omega\big) \mid \omega \in \mathscr{D}(X) \text{ with } \langle f\rangle \gg\!= \omega = \tau\}.$$

Recall that $\langle f\rangle \colon X \nrightarrow Y$ is the promotion of an ordinary function $f\colon X \to Y$ to a deterministic channel, with $\langle f\rangle(x) = 1|f(x)\rangle$. Indeed, this earlier result is restricted, since it only works for deterministic channels, and not for channels in general, like Theorem 3. In the deterministic case things are easy: one can update σ to $\sigma' = \langle f\rangle^{\dagger}_{\sigma} \gg\!= \tau$ and get a perfect prediction, since $\langle f\rangle \gg\!= \sigma' = \tau$.

Further, we mention that the update rules of Pearl and Jeffrey are interdefinable, see [3] or [19]. Let $c\colon X \nrightarrow Y$ and $\sigma \in \mathscr{D}(X)$ be given. For an evidence predicate q we can obtain Pearl's update on the left below, in terms of Jeffrey's update on the right:

$$\sigma|_{c \mathrel{=\!\!<} q} \;=\; c^{\dagger}_{\sigma} \gg\!= \Big((c \gg\!= \sigma)\big|_{q}\Big).$$

This uses the predicted state $c \gg\!= \sigma$, updated with the Pearl-evidence q, as Jeffrey-evidence. Similarly, if we have an evidence state $\tau \in \mathscr{D}(Y)$ we have:

$$c^{\dagger}_{\sigma} \gg\!= \tau \;=\; \sigma|_{c \mathrel{=\!\!<} q} \quad \text{for} \quad q(y) \;=\; \frac{\tau(y)}{(c \gg\!= \sigma)(y)}.$$

One may need to rescale the fraction to obtain a predicate, but such rescaling does not affect updating.

We take another look at our leading example, but now from Jeffrey's perspective.

Example 3. Recall the situation of Example 1, with a teacher predicting the performance of pupils, depending on the teacher's mood. The evidence predicate q from Example 2 can be translated into a state τ on the set G of grades:

$$\tau \;=\; \tfrac{1}{10}|1\rangle + \tfrac{3}{10}|2\rangle + \tfrac{3}{10}|3\rangle + \tfrac{2}{10}|4\rangle + \tfrac{1}{10}|5\rangle.$$

There is an a priori divergence $D_{KL}(\tau, c \gg\!= \sigma) \approx 1.336$. With some effort one can prove that the Jeffrey-update of σ is:

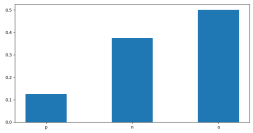
$$\sigma_J \;=\; c^{\dagger}_{\sigma} \gg\!= \tau \;=\; \tfrac{972795}{3913520}|p\rangle + \tfrac{1966737}{3913520}|n\rangle + \tfrac{973988}{3913520}|o\rangle \;\approx\; 0.2486|p\rangle + 0.5025|n\rangle + 0.2489|o\rangle.$$

The divergence has now dropped, from 1.336 to $D_{KL}(\tau, c \gg\!= \sigma_J) \approx 1.087$.

In the end it is interesting to compare the original (prior) mood with its Pearl- and Jeffrey-updates. In Figure 2 the prior mood is reproduced from Figure 1, for easy comparison. The Pearl and Jeffrey updates differ only slightly, to be precise with $D_{KL}(\sigma_P, \sigma_J) \approx 0.007$. They both show that the bad grades evidence deteriorates the teacher's mood. Examples where the Pearl- and Jeffrey-updates differ wildly are given by [19].

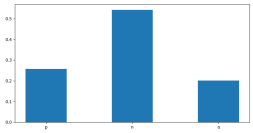
Remark 1. At this stage one could say: well, fair enough, so Pearl's rule increases validity and Jeffrey's rule decreases divergence, but how "exclusive" are these results? Maybe there is also a decrease of divergence in Pearl's updating and an increase of validity in Jeffrey's updating.

Recall that Pearl's rule uses a predicate q as evidence and Jeffrey's rules uses a state τ. If we keep states and predicates apart, as mathematical entities of different types, there is no way to express a divergence-decrease in Pearl's setting or a validity-increase in Jeffrey's setting.



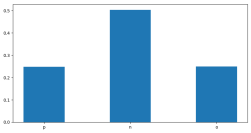
prior mood

$0.125|p\rangle+0.375|n\rangle+0.5|o\rangle$



mood after Pearl-update

$0.2575|p\rangle+0.5418|n\rangle+0.2007|o\rangle$



mood after Jeffrey-update

$0.2486|p\rangle+0.5025|n\rangle+0.2489|o\rangle$

Figure 2: Mood updates from Examples 1 and 3.

Nevertheless, in Examples 2 and 3 we have seen that the evidence q and τ are basically the same. In such a situation we can show that Pearl's update rule need not give a divergence-decrease and Jeffrey's rule need not produce a validity-increase. We give an example which demonstrates both points at the same time.

Take sets $X = \{0,1\}$ and $Y = \{a,b,c\}$ with uniform prior $\sigma = \frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle \in \mathscr{D}(X)$. We use the channel $c\colon X \nrightarrow Y$ given by:

$$c(0) = \tfrac{1}{9}|a\rangle + \tfrac{2}{3}|b\rangle + \tfrac{2}{9}|c\rangle \qquad \text{and} \qquad c(1) = \tfrac{7}{25}|a\rangle + \tfrac{7}{25}|b\rangle + \tfrac{11}{25}|c\rangle.$$

The predicted state is then $c \gg\!= \sigma = \frac{44}{225}|a\rangle + \frac{71}{150}|b\rangle + \frac{149}{450}|c\rangle$. We use as 'equal' evidence predicate and state:

$$q = \tfrac{1}{2}\cdot\mathbf{1}_a + \tfrac{1}{3}\cdot\mathbf{1}_b + \tfrac{1}{6}\cdot\mathbf{1}_c \qquad \text{and} \qquad \tau = \tfrac{1}{2}|a\rangle + \tfrac{1}{3}|b\rangle + \tfrac{1}{6}|c\rangle.$$

We then get the following updates, according to Pearl and Jeffrey, respectively:

$$\begin{aligned} \sigma_P &= \tfrac{425}{839}|0\rangle + \tfrac{414}{839}|1\rangle \\ &\approx 0.5066|0\rangle + 0.4934|1\rangle \end{aligned} \qquad \text{and} \qquad \begin{aligned} \sigma_J &= \tfrac{805675}{1861904}|0\rangle + \tfrac{1056229}{1861904}|1\rangle \\ &\approx 0.4327|0\rangle + 0.5673|1\rangle. \end{aligned}$$

The validities and divergences are summarised in the following tables.

description	formula	value
prior validity	$c \gg\!= \sigma \models q$	0.31074
after Pearl	$c \gg\!= \sigma_P \models q$	0.31079
after Jeffrey	$c \gg\!= \sigma_J \models q$	0.31019

description	formula	value
prior divergence	$D_{KL}(\tau, c \gg\!= \sigma)$	0.238
after Pearl	$D_{KL}(\tau, c \gg\!= \sigma_P)$	0.240
after Jeffrey	$D_{KL}(\tau, c \gg\!= \sigma_J)$	0.221

The differences are small, but relevant. Pearl's updating increases validity, as Theorem 1 dictates, but Jeffrey's updating does not, in this example. Similarly, Jeffrey's updating decreases divergence, in line with Theorem 3, but Pearl's updating does not.

One can ask if a divergence-decrease also happens for other forms of divergence (or distance) between distributions. We do not have an exhaustive answer but we do know that this fails for total variantion distance. If we replace the above channel c by c' with $c'(0) = \frac{1}{10}|a\rangle + \frac{1}{2}|b\rangle + \frac{2}{5}|c\rangle$ and $c'(1) = \frac{11}{100}|a\rangle + \frac{33}{100}|b\rangle + \frac{56}{100}|c\rangle$, and keep everything else as it is, then both Pearl's and Jeffrey's update rule produce an increase of the total variation distance.

6 Application to predictive coding

This section first explains how our main result, Theorem 3, can be seen as strengthening the mathematical basis of predictive coding theory. Then, it illustrates (for the runing example) how the current framework can be extended with selective focus and managed expectations.

6.1 Going beyond free energy for point observations

The concept of free energy plays an important role in predictive coding. We briefly describe how it fits into the current framework, see also *e.g.* [2, 9, 10]. Free energy has its basis in statistical physics, going back to Ludwig Bolzmann in the 19th century, where it is used to describe a thermal equilibrium in gases. In predictive coding the human mind is also seen as striving for an equilibrium by reducing prediction errors. This reduction can be achieved either by internally updating the state or by externally performing an action. Using the notation of Theorem 3, the former involves changing the internal state σ and the latter involves changing the external state τ by taking action. Here we only look at (internal) updating.

In our set-up in Theorem 3 we describe an internal update $\sigma \mapsto \sigma_J$ triggered by confrontation with an external state τ. In the predictive coding framework free energy is described not with respect to an entire distribution $\tau \in \mathscr{D}(Y)$, but with respect to a single, point observation $y \in Y$ only. One often thinks of this y as a sample from some external distribution τ. This single observation y corresponds to a point state $1|y\rangle$ and the resulting Jeffrey update is:

$$c^\dagger_\sigma \gg\!= 1|y\rangle \;=\; c^\dagger_\sigma(y) \overset{(3)}{=} \sum_{x\in X} \frac{\sigma(x)\cdot c(x)(y)}{(c \gg\!= \sigma)(y)}|x\rangle. \tag{4}$$

Calculating this distribution may be computationally demanding, especially in the setting of continuous probability. It is in particular the normalising factor $(c \gg\!= \sigma)(y)$ that one wishes to avoid. Hence the strategy is not to compute (4) but to find a state ω that diverges minimally from (4). One thus looks for ω with minimal divergence:

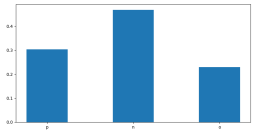
$$\begin{aligned} D_{KL}\big(\omega, c^\dagger_\sigma(y)\big) \;&=\; \sum_{x\in X} \omega(x)\cdot \ln\left(\frac{\omega(x)\cdot(c \gg\!= \sigma)(y)}{\sigma(x)\cdot c(x)(y)}\right) \\ &=\; \sum_{x\in X} \omega(x)\cdot \ln\left(\frac{\omega(x)}{\sigma(x)\cdot c(x)(y)}\right) + \sum_{x\in X} \omega(x)\cdot \ln\Big((c \gg\!= \sigma)(y)\Big) \\ &=\; -\mathscr{F}(\omega) + \ln\Big((c \gg\!= \sigma)(y)\Big), \end{aligned}$$

where $\mathscr{F}$ is the free energy, defined as:

$$\mathscr{F}(\omega) \;=\; \sum_{x\in X} \omega(x)\cdot \ln\left(\frac{\gamma(x,y)}{\omega(x)}\right) \qquad \text{with joint state} \qquad \gamma(x,y) \;=\; \sigma(x)\cdot c(x)(y).$$

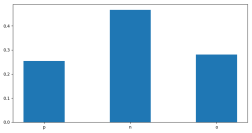
This joint state $\gamma \in \mathscr{D}(X\times Y)$ is the generative model associated with the state and channel σ, c. Thus, by finding ω with maximal free energy $\mathscr{F}(\omega)$ one obtains at the same time a state ω that diverges minimally from the Jeffrey update $c^\dagger_\sigma \gg\!= 1|y\rangle$. This works since the normalisation factor $(c \gg\!= \sigma)(y)$ does not depend on ω and can thus by ignored in the optimalisation.

Our Theorem 3 enriches predictive coding theory and shows how to go beyond point observations $y \in Y$, and use a distribution $\tau \in \mathscr{D}(Y)$ as external evidence. Updating with such τ as evidence reduces divergence — and thus prediction error.



Pearl-update

$0.3033|p\rangle+0.4672|n\rangle+0.2295|o\rangle$



Jeffrey-update

$0.2541|p\rangle+0.4657|n\rangle+0.2802|o\rangle$

Figure 3: Mood updates with focus.

6.2 Incorporating focus and expectation management

We briefly discuss how the situation in this paper, with a channel $c\colon X \nrightarrow Y$ mediating between an internal world X and an external world Y can be used to incorporate aspects of focus and expectation management (preparation). This is done via appropriately placed updates, and can be considered both from Pearl's and from Jeffrey's perspective, each with their own objectives. We illustrate how this can be done mathematically, without any cognitive claims.

We continue Examples 1, 2 and 3 with evidence on the set of marks $Y = \{1,2,\ldots,10\}$, either in the form of a predicate q (Pearl) or a state τ (Jeffrey). Suppose the teacher focuses on the bad grades. This focus may be a conscious decision or instruction, or may happen unconsciously, *e.g.* through some form of bias or tunnel vision. We illustrate how the focus can happen via a subset/event $F \subseteq Y$, say $F = \{1,2,3\}$ containing bad marks. We write $\mathbf{1}_F\colon Y \to [0,1]$ for the associated sharp predicate, with $\mathbf{1}_F(y) = 1$ if $y \in F$ and $\mathbf{1}_F(y) = 0$ if $y \notin F$. We show how to use this focus predicate $\mathbf{1}_F$ for a tunnel vision on the external world, by incorporating it in the following manner.

- Pearl: update internal state σ to $\sigma|_{c \mathrel{=\!\!\ll} (q \,\&\, \mathbf{1}_F)}$.
- Jeffrey: update σ to $c^{\dagger}_{\sigma} \gg\!= (\tau|_{\mathbf{1}_F})$

In Pearl's case the focus predicate $\mathbf{1}_F$ is combined with the evidence q via conjunction. In Jeffrey's case the focus predicate is used to update the external state τ to $\tau|_{\mathbf{1}_F}$ — which itself may be understood as an action in predictive coding theory. In both cases the effect is that the evidence is masked (restricted). The resulting updates of the mood state σ are described in Figure 3. With respect to the updates without focus in Figure 2 there is more pessimism, as a result of the focus on the bad marks. This pessimistic shift is greater in Pearl's update.

In predictive coding theory, and more generally in cognition theory, the notion of attention plays an important role, see *e.g.* [31, 5]. It refers to the processing of the prediction error. Above we have deliberately used the informal term focus, to avoid confusion.

Instead of masking the external evidence via a predicate on the outside world Y the teacher can also prepare for the outcome by updating his/her internal state σ before adapting to the external evidence. This can be seen as a form of managing one's own expectations. Consider for instance the predicate r on the set $X = \{p,n,o\}$ of mood options given by:

$$r = \tfrac{7}{10} \cdot \mathbf{1}_p + \tfrac{1}{2} \cdot \mathbf{1}_n + \tfrac{3}{10} \cdot \mathbf{1}_o.$$

Clearly, it favours pessimism. We can incorporate this predicate r on the internal side, both in Pearl's and in Jeffrey's approach:

- Pearl: update internal state σ to $\sigma|_r|_{c \mathrel{=\!\!\ll} q} = \sigma|_{r \,\&\, (c \mathrel{=\!\!\ll} q)}$.

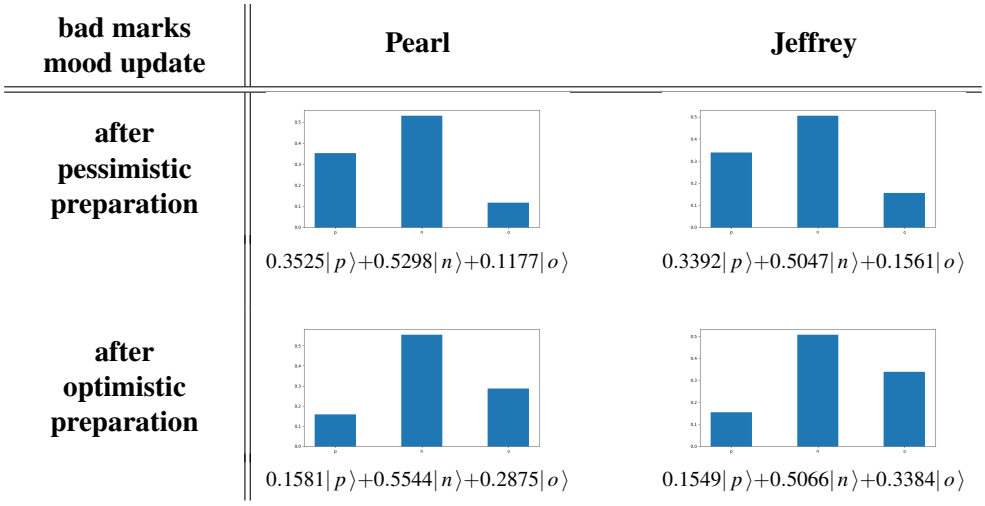


Figure 4: Managing expectations before mood update with bad marks, after pessimistic preparation in the top row (with r) and after optimistic preparation (with the negation $r^{\perp}$) in the bottom row.

- Jeffrey: update σ to $c^{\dagger}_{\sigma|_r} \gg\!= \tau$.

The resulting updated moods are described in the top row of Figure 4. Interestingly, if you try to prepare for bad marks via a 'pessimistic' predicate r, as in the above bullet points, the resulting mood is more pessimistic than without the preparation with r. In order to reduce the negative impact of expected bad marks it works better to prepare positively, so with the negation $r^{\perp}$ instead of with r, in the above two points. This follows common wisdom: bracing for impact works better by cheering up, since the bad news then hits less hard.

Overall we see that although there are considerable mathematical differences between Pearl's and Jeffrey's update mechanism — between learning from what's right and learning from what's wrong — they react in the same directions to changes of focus and preparation. In this example we have described focus and preparation separately, but of course they can be combined.

7 Concluding remarks

This paper is a follow-up to earlier work of [19], where mathematically precise formulations were introduced for Pearl's and Jeffrey's update rules. There, the distinction between the two rules was described only in qualitative terms, namely as 'improvement' (for Pearl) versus 'correction' (for Jeffrey). Here, this qualitative characterisation is turned into a mathematical characterisation: Pearl's rule increases validity, whereas Jeffrey's rule decreases divergence. The proof of the latter fact is the main technical achievement of this paper.

The two update rules of Pearl and Jeffrey have been placed in the setting of predictive coding theory. It remains an open question whether these two update mechanisms can be distinguished empirically in neuroscience.

Acknowledgements

Thanks are due to Harald Woracek and Ana Sokolova for pointing out the relevance of the reference [8] for the proof of Proposition 2 and for helpful subsequent discussions.

References

[1] S. Abramsky and B. Coecke. A categorical semantics of quantum protocols. In K. Engesser, Dov M. Gabbay, and D. Lehmann, editors, *Handbook of Quantum Logic and Quantum Structures: Quantum Logic*, pages 261–323. North-Holland, Elsevier, Computer Science Press, 2009. doi:10.1016/b978-0-444-52869-8.50010-4.

[2] R. Bogacz. A tutorial on the free-energy framework for modelling perception and learning. *Journ. Math. Psychology*, 81:198–211, 2017. doi:10.1016/j.jmp.2015.11.003.

[3] H. Chan and A. Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artif. Intelligence*, 163:67–90, 2005. doi:10.1016/j.artint.2004.09.005.

[4] K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, 29(7):938–971, 2019. doi:10.1017/s0960129518000488.

[5] A. Clark. *Surfing Uncertainty. Prediction, Action, and the Embodied Mind*. Oxford Univ. Press, 2016. doi:10.1093/mind/fzx038.

[6] F. Clerc, F. Dahlqvist, V. Danos, and I. Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, *Foundations of Software Science and Computation Structures*, number 10203 in Lect. Notes Comp. Sci., pages 355–369. Springer, Berlin, 2017. doi:10.1007/978-3-662-54458-7_21.

[7] F. Dietrich, C. List, and R. Bradley. Belief revision generalized: A joint characterization of Bayes' and Jeffrey's rules. *Journ. of Economic Theory*, 162:352–371, 2016. doi:10.1016/j.jet.2015.11.006.

[8] S. Friedland and S. Karlin. Some inequalities for the spectral radius of non-negative matrices and applications. *Duke Math. Journ.*, 42(3):459–490, 1975. doi:10.1215/s0012-7094-75-04244-1.

[9] K. Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009. doi:10.1016/j.tics.2009.04.005.

[10] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi:10.1038/nrn2787.

[11] K. Friston and S. Kiebel. Predictive coding under the free-energy principle. *Phil. Trans. of the Royal Society B: Biological sciences*, 364(1521):1211–1221, 2009. doi:10.1098/rstb.2008.0300.

[12] T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020. doi:10.1016/J.AIM.2020.107239.

[13] I. Gelfand. Normierte ringe. *Sbornik Mathematics*, 9(51):3–24, 1941.

[14] J. Halpern. *Reasoning about Uncertainty*. MIT Press, Cambridge, MA, 2003. doi:10.7551/mitpress/10951.001.0001.

[15] J. Hohwy. *The Predictive Mind*. Oxford Univ. Press, 2013. doi:10.1093/acprof:oso/9780199682737.001.0001.

[16] B. Jacobs. New directions in categorical logic, for classical, probabilistic and quantum logic. *Logical Methods in Comp. Sci.*, 11(3), 2015. doi:10.2168/lmcs-11(3:24)2015.

[17] B. Jacobs. From probability monads to commutative effectuses. *Journ. of Logical and Algebraic Methods in Programming*, 94:200–237, 2018. doi:10.1016/j.jlamp.2016.11.006.

[18] B. Jacobs. Learning along a channel: the Expectation part of Expectation-Maximisation. In B. König, editor, *Math. Found. of Programming Semantics*, number 347 in Elect. Notes in Theor. Comp. Sci., pages 143–160. Elsevier, Amsterdam, 2019. doi:10.1016/j.entcs.2019.09.008.

[19] B. Jacobs. The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. *Journ. of Artif. Intelligence Research*, 65:783–806, 2019. doi:10.1613/jair.1.11349.

[20] B. Jacobs. Multinomial and hypergeometric distributions in Markov categories. In A. Sokolova, editor, *Math. Found. of Programming Semantics*, 2021.

[21] B. Jacobs. Multisets and distributions, in drawing and learning. In A. Palmigiano and M. Sadrzadeh, editors, *Samson Abramsky on Logic and Structure in Computer Science and Beyond*. Springer, 2021, to appear.

[22] B. Jacobs and F. Zanasi. The logical essentials of Bayesian reasoning. In G. Barthe, J.-P. Katoen, and A. Silva, editors, *Foundations of Probabilistic Programming*, pages 295–331. Cambridge Univ. Press, 2021. doi:10.1017/9781108770750.010.

[23] R. Jeffrey. *The Logic of Decision*. The Univ. of Chicago Press, 2nd rev. edition, 1983.

[24] S. Karlin. *Mathematical Methods and Theory in Games, Programming, and Economics. Vol I: Martrix Games, Programming, and Mathematical Economics*. The Java Series. Addison-Wesley, 1959.

[25] P. Lax. *Linear Algebra and Its Applications*. John Wiley & Sons, 2nd edition, 2007.

[26] H. Minc. *Nonnegative Matrices*. John Wiley & Sons, 1998.

[27] A. Mrad, V. Delcroix, S. Piechowiak, P. Leicester, and M. Abid. An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. *Applied Intelligence*, 23(4):802–824, 2015.

[28] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Graduate Texts in Mathematics 118. Morgan Kaufmann, 1988. doi:10.1016/C2009-0-27609-4.

[29] J. Pearl. Jeffrey's rule, passage of experience, and neo-Bayesianism. In Jr. H. Kyburg, editor, *Knowledge Representation and Defeasible Reasoning*, pages 245–265. Kluwer Acad. Publishers, 1990. doi:10.1007/978-94-009-0553-5_10.

[30] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999. doi:10.1038/4580.

[31] E. Schröger, A. Marzecová, and I. SanMiguel. Attention and prediction in human audition: a lesson from cognitive psychophysiology. *Eur. Journ. Neuroscience*, 41:641–664, 2015. doi:10.1111/ejn.12816.

A Appendix

This appendix contains a proof of the main result of this paper (Theorem 3). The proof is extracted from [8] and is specialised here to a conditional expectations matrix. The original proof is formulated more generally. We use some basic facts from linear algebra, esp. about non-negative matrices, see [26] for background information. The proof also uses Gelfand's spectral radius formula [13]. This is a "mathematical bazooka", with a non-trivial proof, using linear analysis. It is an open question if there is an easier proof for Theorem 3 in this paper.

We recall that for a square matrix A the *spectral radius* $\rho(A)$ is the maximum of the absolute values of its eigenvalues:

$$\rho(A) = \max\Big\{ |\lambda| \;\Big|\; \lambda \text{ is an eigenvalue of } A \Big\}.$$

We shall make use the following result. The first point is known as Gelfand's formula, originally from [13]. The proof is non-trivial and is skipped here; for details see *e.g.* [25, Appendix 10]. For convenience we include short (standard) proofs of the other two points.

Theorem 4. *Let A be a (finite) square matrix, and $\|-\|$ be a matrix norm.*

1. *The spectral radius satisfies:*

$$\rho(A) = \lim_{n\to\infty} \big\|A^n\big\|^{1/n}.$$

2. *Here we shall use the* 1*-norm* $\|A\|_1 = \max_j \sum_i |A_{ij}|$. *It yields that* $\rho(A) = 1$ *for each stochastic matrix* A.
3. *Let square matrix* A *now be non-negative, that is, satisfy* $A_{ij} \geq 0$, *and let* x *be a positive vector, so each* $x_i > 0$. *If* $Ax \leq r \cdot x$ *with* $r > 0$, *then* $\rho(A) \leq r$.

Proof. As mentioned, we skip the proof of the Gelfand's formula. If A is stochastic, one gets:

$$\|A\|_1 \;=\; \max_j \textstyle\sum_i |A_{ij}| \;=\; \max_j \textstyle\sum_i A_{ij} \;=\; \max_j 1 \;=\; 1.$$

Stochastic matrices are closed under matrix multiplication, so $\|A^n\|_1 = 1$ for each n. Hence $\rho(A) = 1$ via Gelfand's formula.

We next show how the third point can be obtained from the first one, as in [24, Cor. 8.2.2]. By assumption, each entry x_i in the (finite) vector x is positive. Let's write x_- for the least one and x_+ for the greatest one. Then $0 < x_- \leq x_i \leq x_+$ for each i. For each n we have:

$$\begin{aligned} \left\|A^n\right\|_1 \cdot x_- \;&=\; \max_j \textstyle\sum_i \left|\left(A^n\right)_{ij}\right| \cdot x_- \\ &\leq\; \max_j \textstyle\sum_i \left(A^n\right)_{ij} \cdot x_i \\ &=\; \max_j \left(A^n x\right)_j \\ &\leq\; \max_j \left(r^n \cdot x\right)_j \\ &\leq\; r^n \cdot x_+. \end{aligned}$$

Hence:

$$\left\|A^n\right\|_1^{1/n} \;\leq\; \left(r^n \cdot \frac{x_+}{x_-}\right)^{1/n} \;=\; r \cdot \left(\frac{x_+}{x_-}\right)^{1/n}.$$

Thus, by Gelfand's formula, in the first point,

$$\begin{aligned} \rho(A) \;&=\; \lim_{n\to\infty} \left\|A^n\right\|^{1/n} \\ &\leq\; \lim_{n\to\infty} r \cdot \left(\frac{x_+}{x_-}\right)^{1/n} \;=\; r \cdot \lim_{n\to\infty} \left(\frac{x_+}{x_-}\right)^{1/n} \;=\; r \cdot 1 \;=\; r. \end{aligned}$$

□

For the remainder of this appendix the setting is as follows. Let $\omega \in \mathscr{D}(X)$ be a fixed state, with an n-test $p_1, \ldots, p_n \in \mathrm{Pred}(X)$ so that $\sum_i p_i = \mathbf{1}$, by definition (of test). We shall assume that the validities $v_i = \omega \models p_i$ are non-zero. Notice that $\sum_i v_i = 1$. We organise these validities in a vector v and in diagonal matrix V:

$$v \;=\; \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \;=\; \begin{pmatrix} \omega \models p_1 \\ \vdots \\ \omega \models p_n \end{pmatrix} \qquad\qquad V \;=\; \begin{pmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_n \end{pmatrix}.$$

In addition, we use two $n \times n$ (real, non-negative) matrices B and C given by:

$$\begin{aligned} B_{ij} \;&=\; \frac{\omega \models p_i \,\&\, p_j}{(\omega \models p_i) \cdot (\omega \models p_j)} \\ C_{ij} \;&=\; \omega|_{p_j} \models p_i \;\overset{(1)}{=}\; \frac{\omega \models p_i \,\&\, p_j}{\omega \models p_j} \;=\; (\omega \models p_i) \cdot B_{ij}. \end{aligned}$$

The next series of facts is extracted from [8].

Lemma 1. *The above matrices B and C satisfy the following properties.*

1. *The matrix B is non-negative and symmetric, and satisfies $B\,v = \mathbf{1}$. Moreover, B is positive definite, so that its eigenvalues are positive reals.*
2. *As a result, the inverse B^{-1} and square root $B^{1/2}$ exist — and $B^{-1/2}$ too.*
3. *The matrix C of conditional expectations is stochastic and thus its spectral radius $\rho(C)$ equals 1, by Theorem 4 (2). Moreover, C satisfies $C\,v = v$ and $C = V \cdot B$.*
4. *For an $n \times n$ real matrix D, $\rho(DC) = \rho\big(B^{1/2}DVB^{1/2}\big)$.*
5. *Assume now that D is a diagonal matrix with numbers $d_1, \ldots, d_n \geq 0$ on its diagonal. Then:*

$$\textstyle\sum_i d_i \cdot v_i \;\leq\; \rho(DC).$$

Proof. 1. Clearly, $B_{ij} = B_{ji}$. Further,

$$\big(B\,v\big)_i = \sum_j \frac{\omega \models p_i \,\&\, p_j}{(\omega \models p_i)\cdot(\omega \models p_j)} \cdot (\omega \models p_j) = \frac{\omega \models p_i \,\&\, (\sum_j p_j)}{\omega \models p_i} = \frac{\omega \models p_i \,\&\, \mathbf{1}}{\omega \models p_i} = \frac{\omega \models p_i}{\omega \models p_i} = 1.$$

The matrix B is positive definite since for a non-zero vector $z = (z_i)$ of reals:

$$\begin{aligned} z^T B z &= \sum_{i,j} z_i \cdot \frac{\omega \models p_i \,\&\, p_j}{(\omega \models p_i)\cdot(\omega \models p_j)} \cdot z_j \\ &= \omega \models \Big(\textstyle\sum_i \frac{z_i}{v_i} \cdot p_i\Big) \,\&\, \Big(\sum_j \frac{z_j}{v_j} \cdot p_j\Big) \\ &= \omega \models q \,\&\, q \qquad\qquad \text{for } q = \textstyle\sum_i \frac{z_i}{v_i} \cdot p_i \\ &> 0. \end{aligned}$$

We have a strict inequality $>$ here since $q \geq \frac{z_1}{v_1} \cdot p_1$ and $\omega \models p_1 > 0$, by assumption; in fact this holds for each p_i. Thus:

$$\omega \models q \,\&\, q \;\geq\; \frac{z_1^2}{v_1^2} \cdot (\omega \models p_1 \,\&\, p_1) \;\geq\; \frac{z_1^2}{v_1^2} \cdot (\omega \models p_1)^2 \;>\; 0.$$

The last inequality follows from Theorem 1 and Bayes' rule (1):

$$\omega \models p \;\leq\; \omega|_p \models p \;=\; \frac{\omega \models p \,\&\, p}{\omega \models p} \qquad \text{so} \qquad \big(\omega \models p\big)^2 \;\leq\; \omega \models p \,\&\, p.$$

2. The square root $B^{1/2}$ and inverse B^{-1} are obtained in the standard way via spectral decomposition $B = Q\Lambda Q^T$ where Λ is the diagonal matrix of eigenvalues $\lambda_i > 0$ and Q is an orthogonal matrix (so $Q^T = Q^{-1}$). Then: $B^{1/2} = Q\Lambda^{1/2}Q^T$ where $\Lambda^{1/2}$ has entries $\lambda_i^{1/2}$. Similarly, $B^{-1} = Q\Lambda^{-1}Q^T$, and $B^{-1/2} = Q\Lambda^{-1/2}Q^T$.

3. It is easy to see that all C's columns add up to one:

$$\textstyle\sum_i C_{ij} \;=\; \sum_i \omega|_{p_j} \models p_i \;=\; \omega|_{p_j} \models \sum_i p_i \;=\; \omega|_{p_j} \models \mathbf{1} \;=\; 1.$$

This makes C stochastic, so that $\rho(C) = 1$. Next:

$$\begin{aligned}\big(Cv\big)_i &= \textstyle\sum_j (\omega|_{p_j} \models p_i) \cdot (\omega \models p_j) \\ &= \textstyle\sum_j \omega \models p_i \,\&\, p_j \qquad \text{by Bayes' rule (1)} \\ &= \omega \models p_i \,\&\, (\textstyle\sum_j p_j) \;=\; \omega \models p_i \,\&\, \mathbf{1} \;=\; \omega \models p_i \;=\; v_i.\end{aligned}$$

Further, $\big(VB\big)_{ij} = v_i \cdot B_{ij} = C_{ij}$.

4. We show that DC and $B^{1/2}DVB^{1/2}$ have the same eigenvalues, which gives $\rho(DC) = \rho(B^{1/2}DVB^{1/2})$. First, let $DCz = \lambda z$. Take $z' = B^{1/2}z$ one gets:

$$B^{1/2}DVB^{1/2}z' \;=\; B^{1/2}DVB^{1/2}B^{1/2}z \;=\; B^{1/2}DCz \;=\; B^{1/2}\lambda z \;=\; \lambda B^{1/2}z \;=\; \lambda z'.$$

In the other direction, let $B^{1/2}DVB^{1/2}w = \lambda w$. Now take $w' = B^{-1/2}w$ so that:

$$DCw' \;=\; B^{-1/2}B^{1/2}DVBB^{-1/2}w \;=\; B^{-1/2}\big(B^{1/2}DVB^{1/2}\big)w \;=\; B^{-1/2}\lambda w \;=\; \lambda w'.$$

5. We use the standard fact that for non-zero vectors z one has:

$$\frac{|\,(Az,z)\,|}{(z,z)} \;\leq\; \rho(A),$$

where $(-,-)$ is inner product. In particular,

$$\frac{|\,(B^{1/2}DVB^{1/2}z,z)\,|}{(z,z)} \;\leq\; \rho\big(B^{1/2}DVB^{1/2}\big) \;=\; \rho(DC).$$

We instantiate with $z = B^{1/2}v$ and use that $VBv = Cv = v$ and $Bv = \mathbf{1}$ in:

$$\begin{aligned}\rho(DC) \;&\geq\; \frac{|\,(B^{1/2}DVB^{1/2}B^{1/2}v,B^{1/2}v)\,|}{(B^{1/2}v,B^{1/2}v)} \;=\; \frac{|\,(DVBv,B^{1/2}B^{1/2}v)\,|}{(v,B^{1/2}B^{1/2}v)} \\ &=\; \frac{|\,(Dv,Bv)\,|}{(v,Bv)} \;=\; \frac{|\,(Dv,\mathbf{1})\,|}{(v,\mathbf{1})} \;=\; \frac{|\,\sum_i d_i \cdot v_i\,|}{\sum_i v_i} \;=\; \textstyle\sum_i d_i \cdot v_i. \qquad \square\end{aligned}$$

Proposition 2. *Let $\omega \in \mathscr{D}(X)$ be a state with predicates $p_1,\ldots,p_n \in \mathrm{Pred}(X)$ forming a 'test', so that $p_1 + \cdots + p_n$ equal the constant 1 predicate $\mathbf{1}$. We assume $\omega \models p_i \neq 0$, for each i. For all numbers $r_1,\ldots,r_n \in (0,1]$ with $\sum_i r_i = 1$, one has:*

$$\sum\nolimits_i \frac{r_i \cdot (\omega \models p_i)}{\sum_j r_j \cdot (\omega|_{p_j} \models p_i)} \;\leq\; 1. \tag{5}$$

Proof. Consider a vector r of non-zero numbers $r_i \in (0,1]$ with $\sum_i r_i = 1$. We form a diagonal matrix D with non-zero diagonal entries $d_1,\ldots,d_n$ with:

$$d_i \;=\; \frac{r_i}{\big(Cr\big)_i} \;=\; \frac{r_i}{\sum_j C_{ij} \cdot r_j}.$$

A crucial observation is that r is an eigenvector of the matrix DC, with eigenvalue 1, since:

$$\big(DCr\big)_i = \sum_j \big(DC\big)_{ij} \cdot r_j = \sum_j d_i \cdot C_{ij} \cdot r_j = \frac{r_i}{\big(Cr\big)_i} \cdot \big(\sum_j C_{ij} \cdot r_j\big) = r_i.$$

Theorem 4 (3) now yields $\rho(DC) = 1$.

By Lemma 1 (5) we get the required inequality in Proposition 2:

$$\sum_i \frac{r_i \cdot v_i}{\big(Cr\big)_i} = \sum_i d_i \cdot v_i \leq \rho(DC) = 1. \qquad \square$$

Finally we come to the proof of the main result.

Proof. [Of Theorem 3] We reason as follows:

$$\begin{aligned}
&D_{KL}\Big(\tau, c \gg\!= \big(c^\dagger_\sigma \gg\!= \tau\big)\Big) \;-\; D_{KL}\Big(\tau, c \gg\!= \sigma\Big) \\
&= \sum_y \tau(y) \cdot \log\left(\frac{\tau(y)}{\big(c \gg\!= (c^\dagger_\sigma \gg\!= \tau)\big)(y)}\right) - \sum_y \tau(y) \cdot \log\left(\frac{\tau(y)}{\big(c \gg\!= \sigma\big)(y)}\right) \\
&= \sum_y \tau(y) \cdot \log\left(\frac{\tau(y)}{\big(c \gg\!= (c^\dagger_\sigma \gg\!= \tau)\big)(y)} \cdot \frac{\big(c \gg\!= \sigma\big)(y)}{\tau(y)}\right) \\
&\leq \log\left(\sum_y \frac{\tau(y) \cdot \big(c \gg\!= \sigma\big)(y)}{\big(c \gg\!= (c^\dagger_\sigma \gg\!= \tau)\big)(y)}\right) \\
&\leq \log\big(1\big) \;=\; 0.
\end{aligned}$$

The first inequality is an instance of Jensen's inequality. The second one follows from Proposition 2, via:

$$\begin{aligned}
\sum_y \frac{\tau(y) \cdot \big(c \gg\!= \sigma\big)(y)}{\big(c \gg\!= (c^\dagger_\sigma \gg\!= \tau)\big)(y)} &= \sum_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_x (c^\dagger_\sigma \gg\!= \tau)(x) \cdot c(x)(y)} \\
&= \sum_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_{x,z} \tau(z) \cdot c^\dagger_\sigma(z)(x) \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_y)(x)} \\
&\overset{(3)}{=} \sum_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_{x,z} \tau(z) \cdot \frac{\sigma(x) \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_z)(x)}{\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_z} \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_y)(x)} \\
&= \sum_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_z \tau(z) \cdot \frac{\sigma \models (c \mathrel{=\!\!\ll} \mathbf{1}_z) \,\&\, (c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_z}} \\
&\overset{(1)}{=} \sum_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_z \tau(z) \cdot (\sigma|_{c \mathrel{=\!\!\ll} \mathbf{1}_z} \models c \mathrel{=\!\!\ll} \mathbf{1}_y)} \\
&\leq 1, \qquad \text{by Proposition 2.}
\end{aligned}$$

In the last line we apply Proposition 2 with test $p_i \coloneqq c \mathrel{=\!\!\ll} \mathbf{1}_{y_i}$, where $Y = \{y_1, \ldots, y_n\}$. The point predicates $\mathbf{1}_{y_i}$ form a test on Y. Predicate transformation preserves tests. As assume in Theorem 3, $c \gg\!= \sigma$ has full support, so that $\sigma \models p_i = \sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_{y_i} = (c \gg\!= \sigma)(y_i)$ is non-zero for each i. $\square$