Mathematical Structures in Computer Science (2020), **30**, pp. 44–61
doi:10.1017/S0960129519000082

CAMBRIDGE UNIVERSITY PRESS

PAPER

A channel-based perspective on conjugate priors[†]

B. Jacobs*

Institute for Computing and Information Sciences, Radboud University, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
*Email: bart@cs.ru.nl

(Received 1 July 2017; revised 5 March 2019; accepted 14 April 2019)

Abstract

A desired closure property in Bayesian probability is that an updated posterior distribution be in the same class of distributions – say Gaussians – as the prior distribution. When the updating takes place via a statistical model, one calls the class of prior distributions the 'conjugate priors' of the model. This paper gives (1) an abstract formulation of this notion of conjugate prior, using channels, in a graphical language, (2) a simple abstract proof that such conjugate priors yield Bayesian inversions and (3) an extension to multiple updates. The theory is illustrated with several standard examples.

1. Introduction

The main result of this paper, Theorem 6.3, is mathematically trivial. But it is not entirely trivial to see that this result is trivial. The effort and contribution of this paper lie in setting up a framework – using the abstract language of channels, Kleisli maps and string diagrams for probability theory – to define the notion of conjugate prior in such a way that there is a trivial proof of the main statement, saying that conjugate priors yield Bayesian inversions. This is indeed what conjugate priors are meant to be.

Conjugate priors form a fundamental topic in Bayesian theory. They are commonly described via a closure property of a class of prior distributions, namely as being closed under certain Bayesian updates. Conjugate priors are especially useful because they do not only involve a closure *property*, but also a particular *structure*, namely an explicit function that performs an analytical computation of posterior distributions via updates of the parameters. This greatly simplifies Bayesian analysis. For instance, the Beta distribution is conjugate prior to the Bernoulli (or 'flip') distribution, and also to the binomial distribution: updating a Beta(α, β) prior via a Bernoulli/binomial statistical model yields a new Beta(α', β') prior, with adapted parameters α', β' that can be computed explicitly from α, β and the observation at hand. Despite this importance, the descriptions in the literature of what it means to be a conjugate prior are remarkably informal. One does find several lists of classes of distributions, for instance at Wikipedia,[1] together with formulas about how to recompute parameters. The topic has a long and rich history in statistics (see e.g. Bishop (2006)), with much emphasis on exponential families (Diaconis and Ylvisaker 1979), but a precise, general definition is hard to find.

We briefly review some common approaches, without any pretension to be complete: the definition in Alpaydin (2010) is rather short, based on an example, and just says: 'We see that the

[†]The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement nr. 320571.

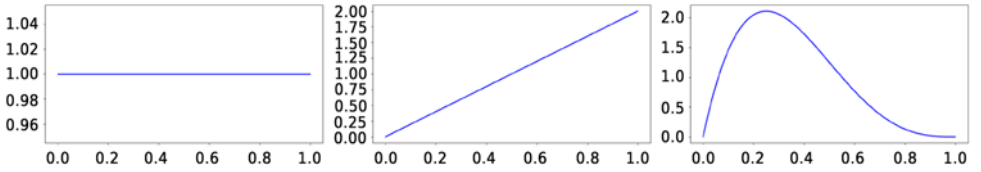


Figure 1. Uniform prior, and two posterior probability density functions on [0, 1], after observing head, and after observing head–tail–tail–tail. These functions correspond respectively to Beta(1, 1), Beta(2, 1), and Beta(2, 4). These three plots can also be obtained via suitable Bayesian updates (inversions).

posterior has the same form as the prior and we call such a prior a *conjugate prior*'. Also Russell and Norvig (2003) mention the term 'conjugate prior' only in relation to an example. There is a separate section in Bishop (2006) about conjugate priors, but no precise definition. Instead, there is the informal description '... the posterior distribution has the same functional form as the prior'. The most precise definition (known to the author) is in Bernardo and Smith (2000, Section 5.2), where the conjugate family with respect to a statistical model, assuming a 'sufficient statistic', is described. It comes close to our channel-based description, since it explicitly mentions the conjugate family as a conditional probability distribution with (recomputed) parameters. The approach is rather concrete however, and the high level of mathematical abstraction that we seek here is missing in Bernardo and Smith (2000).

This paper presents a novel systematic perspective for precisely defining what conjugate priorship means. It uses the notion of 'channel' as starting point. The basis of this approach lies in category theory, especially effectus theory (Cho et al. 2015; Jacobs 2015). However, we try to make this paper accessible to non-category theorists, by using the term 'channel' instead of morphism in a Kleisli category of a suitable monad. Moreover, a graphical language is used for channels, which hopefully makes the approach more intuitive. Thus, the emphasis of the paper is on *what it means* to have conjugate priors. It does not offer new perspectives on how to find/obtain them.

The paper is organised as follows. It starts in Section 2 with a high-level description of the main ideas, without going into technical details. Preparatory definitions are provided in Sections 3 and 4, dealing with channels in probabilistic computation, with a diagrammatic language for channels, and with Bayesian inversion. Then, Section 5 contains the novel channel-based definition of conjugate priorship; it also illustrates how several standard examples fit in this new setting. Section 6 establishes the (expected) close relationship between conjugate priors and Bayesian inversions (Cho and Jacobs, 2019; Clerc et al., 2017). Section 7 illustrates that multiple updates are handled basically in the same way as single updates, and also how the notion of 'sufficient statistic' fits in.

2. Main Ideas

This section gives an informal description of the main ideas underlying this paper. It starts with a standard example, and then proceeds with a step-by-step introduction to the essentials of the perspective of this paper.

A well-known example in Bayesian reasoning is inferring the (unknown) bias of a coin from a sequence of consecutive head/tail observations. The bias is a number $r \in [0, 1]$ in the unit interval, giving the 'Bernoulli' or 'flip' probability r for head, and $1 - r$ for tail. Initially we assume a uniform distribution for r, as described by the constant probability density function (pdf) on the left in Figure 1. After observing one head, this pdf changes to the second picture. After observing head–tail–tail–tail we get the third pdf. These pictures are obtained by Bayesian inversion, see Section 4.

It is a well-known fact that all the resulting distributions are instances of the Beta(α, β) family of distributions, for different parameters α, β. After each observation, one can recompute the entire updated distribution, via Bayesian inversion. But in fact there is a much more efficient way

to obtain the revised distribution, namely by computing the new parameter values: increment α by one, for head, and increment β by one for tail, see Example 5.3 for details. The family of distributions $\mathrm{Beta}(\alpha, \beta)$, indexed by parameters α, β, is thus suitably closed under updates with Bernoulli. It is the essence of the statement that Beta is conjugate prior to Bernoulli. This will be made precise later on.

Let $X = (X, \Sigma)$ be a measurable space, where $\Sigma \subseteq \mathcal{P}(X)$ is a σ-algebra of measurable subsets. We shall write $\mathcal{G}(X)$ for the set of probability distributions on X. Elements $\omega \in \mathcal{G}(X)$ are thus countably additive functions $\omega\colon \Sigma \to [0, 1]$ with $\omega(X) = 1$.

Idea 1: A *family* of distributions on X, indexed by a measurable space P of parameters, is a (measurable) function $P \to \mathcal{G}(X)$. Categorically, such a function is a *Kleisli* map for $\mathcal{G}$, considered as monad on the category of measurable spaces (see Section 3). These Kleisli maps are also called *channels* and will be written simply as arrows $P \to X$, or diagrammatically as boxes  where we imagine that information is flowing upwards.

The study of families of distributions goes back a long way, e.g. as 'experiments' (Blackwell, 1951).

Along these lines we shall describe the family of Beta distributions as a channel with $P = \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and $X = [0, 1]$, namely as function:

$$\mathbb{R}_{>0} \times \mathbb{R}_{>0} \xrightarrow{\text{Beta}} \mathcal{G}([0, 1]) \tag{1}$$

For $(\alpha, \beta) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ there is the probability distribution $\mathrm{Beta}(\alpha, \beta) \in \mathcal{G}([0, 1])$ determined by its value on a measurable subset $M \subseteq [0, 1]$, which is obtained via integration:

$$\mathrm{Beta}(\alpha, \beta)(M) = \int_M \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\, \mathrm{d}x, \tag{2}$$

where $B(\alpha, \beta) = \int_{[0,1]} x^{\alpha-1}(1 - x)^{\beta-1}\, \mathrm{d}x$ is a normalisation constant.

A conjugate prior relationship involves a family of distributions $P \to \mathcal{G}(X)$ which is closed wrt. updates based on observations (or: data) from a separate domain O. Each 'parameter' element $x \in X$ gives rise to a separate distribution on O. This is what is usually called a *statistical* or *parametric* model. We shall also describe it as a channel.

Idea 2: The observations for a family $P \to \mathcal{G}(X)$ arise via another 'Kleisli' map $X \to \mathcal{G}(O)$ representing the statistical model. Conjugate priorship will be defined for two such composable channels $P \to X \to O$, where O is the space of observations.

In the above coin example, the space O of observations is the two-element set $2 = \{0, 1\}$ where 0 is for tail and 1 for head. The Bernoulli channel is written as $\mathrm{Flip}\colon [0, 1] \to \mathcal{G}(2)$. A probability $r \in [0, 1]$ determines a Bernoulli/flip/coin probability distribution $\mathrm{Flip}(r) \in \mathcal{G}(2)$ on 2, formally sending the subset $\{1\}$ to r and $\{0\}$ to $1 - r$.

Idea 3: A channel $c\colon P \to X$ is a conjugate prior to a channel $d\colon X \to O$ if there is a *parameter translation function* $h\colon P \times O \to P$ satisfying a suitable equation.

The idea is that $c(p)$ is a prior, for $p \in P$, which gets updated via the statistical model (channel) d, in the light of observation $y \in O$. The revised, updated distribution is $c(h(p, y))$. The model d is usually written as a conditional probability $d(y \mid \theta)$.

In the coin example we have $h\colon \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times 2 \to \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ given by $h(\alpha, \beta, 1) = (\alpha + 1, \beta)$ and $h(\alpha, \beta, 0) = (\alpha, \beta + 1)$, see Example 5.3 below for more information.

What has been left unexplained is the 'suitable' equation that the parameter translation function $h\colon P \times O \to P$ should satisfy. It is not entirely trivial, because it is an equation between channels in what is called the Kleisli category $\mathcal{K}\ell(\mathcal{G})$ of the Giry monad $\mathcal{G}$. At this stage we need to move to a more categorical description. The equation, which will appear in Definition 5.1, bears similarities with the notion of Bayesian inversion, which will be introduced in Section 4.

3. Channels and conditional probabilities

This section will describe conditional probabilities as arrows and will show how to compose them. Thereby we are entering the world of category theory. We aim to suppress the underlying categorical machinery and make this work accessible to readers without such background. For those with categorical background knowledge: we will be working in the Kleisli categories of the distribution monad $\mathcal{D}$ for discrete probability, and of the Giry monad $\mathcal{G}$ for continuous probability, see e.g. Giry (1982), Panangaden (2009) and Jacobs (2018). Discrete distributions may be seen as a special case of continuous distributions, via a suitable inclusion map $\mathcal{D} \to \mathcal{G}$. Hence one could give one account, using $\mathcal{G}$ only. However, in computer science, unlike for instance in statistics, discrete distributions are so often used that they merit separate treatment.

We thus start with discrete probability. We write a (finite, discrete) distribution on a set X as a formal convex sum $r_1|x_1\rangle + \cdots + r_n|x_n\rangle$ of elements $x_i \in X$ and probabilities $r_i \in [0,1]$ with $\sum_i r_i = 1$. The 'ket' notation $|-\rangle$ is syntactic sugar, used to distinguish elements of x from their occurrence $|x\rangle$ in such formal convex sums.[2] A distribution as above can be identified with a 'probability mass' function $\omega\colon X \to [0,1]$ which is r_i on x_i and 0 elsewhere. We often implicitly identify distributions with such functions. We shall write $\mathcal{D}(X)$ for the set of distributions on X.

We shall focus on functions of the form $c\colon X \to \mathcal{D}(Y)$. They give for each element $x \in X$ a distribution $c(x)$ on Y. Hence such functions form an X-indexed collection $\big(c(x)\big)_{x\in X}$ of distributions $c(x)$ on Y. They can be understood as *conditional* probabilities $P(y \mid x) = r$, if $c(x)$ is of the form $\cdots r|y\rangle \cdots$, with weight $r = c(x)(y) \in [0,1]$ for $y \in Y$. Thus, by construction, $\sum_y P(y \mid x) = 1$, for each $x \in X$. Moreover, if the sets X and Y are finite, we can describe $c\colon X \to \mathcal{D}(Y)$ as a stochastic matrix, with entries $P(y \mid x)$, adding up to one – per row or column, depending on the chosen orientation of the matrix.

We shall often write functions $X \to \mathcal{D}(Y)$ simply as arrows $X \to Y$, call them 'channels', and write them as 'boxes' in diagrams. This arrow notation is justified, because there is a natural way to compose channels, as we shall see shortly. But first we describe *state transformation*, also called *prediction*. Given a channel $c\colon X \to \mathcal{D}(Y)$ and a state $\omega \in \mathcal{D}(X)$, we can form a new state, written as $c \gg \omega$, on Y. It is defined as:

$$c \gg \omega \coloneqq \sum_y \Big(\sum_x \omega(x) \cdot c(x)(y) \Big) \big| y \big\rangle. \tag{3}$$

The outer sum $\sum_y$ is a formal convex sum, whereas the inner sum $\sum_x$ is an actual sum in the unit interval $[0,1]$. Using state transformation $\gg$ it is easy to define composition of channels: given functions $c\colon X \to \mathcal{D}(Y)$ and $d\colon Y \to \mathcal{D}(Z)$, we use the ordinary composition symbol $\circ$ to form a composite channel $d \circ c\colon X \to \mathcal{D}(Z)$, where:

$$(d \circ c)(x) \coloneqq d \gg c(x) = \sum_{z\in Z} \Big(\sum_y c(x)(y) \cdot d(y)(z) \Big) \big| z \big\rangle. \tag{4}$$

Essentially, this is matrix composition for stochastic matrices. Channel composition $\circ$ is associative and also has a neutral element, namely the identity channel $\eta\colon X \to X$ given by the 'Dirac' function $\eta(x) = 1|x\rangle$. It is not hard to see that $(d \circ c) \gg \omega = d \gg (c \gg \omega)$, see e.g. Jacobs and Zanasi (2019) for more details.

We turn to channels in continuous probability. As already mentioned in Section 2, we write $\mathcal{G}(X)$ for the set of probability distributions $\omega\colon \Sigma_X \to [0,1]$, where $X = (X, \Sigma_X)$ is a measurable space. These probability distributions are (also) called states. The set $\mathcal{G}(X)$ carries a σ-algebra itself, but that does not play an important role here. Each element $x \in X$ yields a probability measure $\eta(x) \in \mathcal{G}(X)$, with $\eta(x)(M) = \mathbf{1}_M(x)$, which is 1 if $x \in M$ and 0 otherwise. This map $\mathbf{1}_M\colon X \to [0,1]$ is called the indicator function for the subset $M \in \Sigma_X$.

For a state/measure $\omega \in \mathcal{G}(X)$ and a measurable function $f\colon X \to \mathbb{R}_{\geq 0}$ we write $\int f \,\mathrm{d}\omega$ for the Lebesgue integral, if it exists. We follow the notation of Jacobs (2013) and refer there for details, or alternatively, to Panangaden (2009). We recall that an integral $\int_M f \,\mathrm{d}\omega$ over a measurable subset $M \subseteq X$ of the domain of f is defined as $\int \mathbf{1}_M \cdot f \,\mathrm{d}\omega$, and that $\int \mathbf{1}_M \,\mathrm{d}\omega = \omega(M)$. Moreover, $\int f \,\mathrm{d}\eta(x) = f(x)$.

For a measurable function $g\colon X \to Y$ between measurable spaces X, Y there is the 'push forward' function $\mathcal{G}(g)\colon \mathcal{G}(X) \to \mathcal{G}(Y)$, given by $\mathcal{G}(g)(\omega)(N) = \omega\big(g^{-1}(N)\big)$. It satisfies:

$$\int f \,\mathrm{d}\mathcal{G}(g)(\omega) = \int f \circ g \,\mathrm{d}\omega. \tag{5}$$

Often, the measurable space X is a subset $X \subseteq \mathbb{R}$ of the real numbers and a probability distribution ω on X is given by a probability density function (pdf), that is, by a measurable function $f\colon X \to \mathbb{R}_{\geq 0}$ with $\int_X f(x)\,\mathrm{d}x = 1$. Such a pdf f gives rise to a state $\omega \in \mathcal{G}(X)$, namely:

$$\omega(M) = \int_M f(x)\,\mathrm{d}x. \tag{6}$$

We then write $\omega = \int f$. In such cases where the measure is not explicitly mentioned, one should assume that it is the Lebesgue measure on $\mathbb{R}$.

In this continuous context a channel is a measurable function $c\colon X \to \mathcal{G}(Y)$, for measurable spaces X, Y. Like in the discrete case, it gives an X-indexed collection $\big(c(x)\big)_{x\in X}$ of probability distributions on Y. The channel c can transform a state $\omega \in \mathcal{G}(X)$ on X into a state $c \gg \omega \in \mathcal{G}(Y)$ on Y, given on a measurable subset $N \subseteq Y$ as:

$$\big(c \gg \omega\big)(N) = \int c(-)(N) \,\mathrm{d}\omega. \tag{7}$$

For another channel $d\colon Y \to \mathcal{G}(Z)$ there is a composite channel $d \circ c\colon X \to \mathcal{G}(Z)$, via integration:

$$\big(d \circ c\big)(x)(K) \coloneqq \big(d \gg c(x)\big)(K) = \int d(-)(K) \,\mathrm{d}c(x) \tag{8}$$

In many situations a channel $c\colon X \to \mathcal{G}(Y)$ is given by an indexed probability density function (pdf) $u\colon X \times Y \to \mathbb{R}_{\geq 0}$, with $\int u(x,y)\,\mathrm{d}y = 1$ for each $x \in X$. The associated channel c is:

$$c(x)(N) = \int_N u(x,y)\,\mathrm{d}y. \tag{9}$$

In that case we simply write $c = \int u$ and call c a pdf-channel. We have already seen such a description of the Beta distribution as a pdf-channel in (2).

(In these pdf-channels $X \to Y$, we use a collection of pdf's $u(x,-)$ which are all dominated by the Lebesgue measure. This domination happens via the relationship $\ll$ of absolute continuity, using the Radon–Nikodym Theorem, see e.g. Panangaden (2009).)

Various additional computation rules for integrals are given in the Appendix.

4. Bayesian inversion in string diagrams

In this paper, we make superficial use of string diagrams to graphically represent sequential and parallel composition of channels, mainly in order to provide an intuitive visual overview. We refer to Selinger (2011) for mathematical details and mention here only the essentials.

A channel $X \to Y$, for instance of the sort discussed in the previous section, can be written as a box  with information flowing upwards, from the wire labelled with X to the wire labelled with Y. Composition of channels, as in (4) or (8), simply involves connecting wires (of the same type). The identity channel is just a wire. We use a triangle notation  for a state on X. It is special case of a channel, namely of the form $1 \to X$ with trivial singleton domain 1.

In the present (probabilistic) setting, we allow copying of wires, written diagrammatically as . We briefly describe such copy channels for discrete and continuous probability:

$$\begin{array}{ccc} X \xrightarrow{\quad} \mathcal{D}(X \times X) & & X \xrightarrow{\quad} \mathcal{G}(X \times X) \\ x \longmapsto 1|x, x\rangle & & x \longmapsto \big(M \times N \mapsto \mathbf{1}_{M \cap N}(x)\big) \end{array}$$

After such a copy we can use parallel channels. We briefly describe how this works, first in the discrete case. For channels $c\colon X \to \mathcal{D}(Y)$ and $d\colon A \to \mathcal{D}(B)$, we have a channel $c \otimes d\colon X \times A \to \mathcal{D}(Y \times B)$ given by:

$$(c \otimes d)(x, a) = \sum_{y,b} c(x)(y) \cdot d(a)(b)\big|y, b\big\rangle.$$

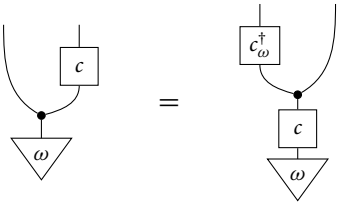
Similarly, in the continuous case, for channels $c\colon X \to \mathcal{G}(Y)$ and $d\colon A \to \mathcal{G}(B)$ we get $c \otimes d\colon X \times A \to \mathcal{G}(Y \times B)$ given by:

$$(c \otimes d)(x, a)(M \times N) = c(x)(M) \cdot d(a)(N).$$

Recall that the product σ-algebra on $Y \times B$ is generated by measurable rectangles of the form $M \times N$, for $M \in \Sigma_Y$ and $N \in \Sigma_B$. Hence measures are uniquely determined by their actions on such rectangles.

We can now formulate what Bayesian inversion is. The definition is couched in purely diagrammatic language, but is applied only to probabilistic interpretations in this paper.

Definition 4.1. *The* Bayesian inversion *of a channel $c\colon X \to Y$ with respect to a state ω of type X, if it exists, is a channel in the opposite direction, written as $c^\dagger_\omega\colon Y \to X$, such that the following equation holds.*



(10)

The dagger notation $c^\dagger_\omega$ is copied from Clerc et al. (2017), see also Cho and Jacobs (2019). There the state ω is left implicit, via a restriction to a certain comma category of kernels. In that setting the operation $(-)^\dagger$ is functorial and forms a dagger category (see e.g. Abramsky and Coecke (2009), Selinger (2007) for definitions). In particular, it preserves composition and identities of channels. Equation (10) can also be written as: $\langle \mathrm{id}, c\rangle \gg\!= \omega = \langle c^\dagger_\omega, \mathrm{id}\rangle \gg\!= (c \gg\!= \omega)$. Alternatively, in the discrete case, with variables explicit, it says: $c(x)(y) \cdot \omega(x) = c^\dagger_\omega(y)(x) \cdot (c \gg\!= \omega)(y)$. This comes close to the 'adjointness' formulations that are typical for daggers.

It is relatively easy to define Bayesian inversion in discrete probability theory: for a channel $c\colon X \to \mathcal{D}(Y)$ and a state/distribution $\omega \in \mathcal{D}(X)$ one can define a channel $c^\dagger_\omega\colon Y \to \mathcal{D}(X)$ as:

$$c^\dagger_\omega(y)(x) = \frac{\omega(x) \cdot c(x)(y)}{(c \gg \omega)(y)} = \frac{\omega(x) \cdot c(x)(y)}{\sum_z \omega(z) \cdot c(z)(y)}, \tag{11}$$

assuming that the denominator is non-zero. This corresponds to the familiar formula $P(B \mid A) = P(A,B)/P(A)$ for conditional probability. The state $c^\dagger_\omega(y)$ can alternatively be defined via updating the state ω with the point predicate $\{y\}$, transformed via c into a predicate $c \ll \mathbf{1}_{\{y\}}$ on X, see Cho and Jacobs (2019) for details.

The situation is much more difficult in continuous probability theory, since Bayesian inversions may not exist (Ackerman et al. 2011; Stoyanov 2014) or may be determined only up to measure zero. But when restricted to e.g. standard Borel spaces, as in Clerc et al. (2017), existence is ensured, see also Culbertson and Sturtz (2014) and Faden (1985). Another common solution is to assume that we have a pdf-channel: there is a map $u\colon X \times Y \to \mathbb{R}_{\geq 0}$ that defines a channel $c\colon X \to \mathcal{G}(Y)$, like in (9), as $c(x)(N) = \int_N u(x,y)\,\mathrm{d}y$. Then, for a distribution $\omega \in \mathcal{G}(X)$ we can take as Bayesian inversion:

$$\begin{aligned} c^\dagger_\omega(y)(M) &= \frac{\int_M u(-,y)\,\mathrm{d}\omega}{\int_X u(-,y)\,\mathrm{d}\omega} \\ &= \frac{\int_M f(x) \cdot u(x,y)\,\mathrm{d}x}{\int_X f(x) \cdot u(x,y)\,\mathrm{d}x} \qquad \text{when } \omega = \int f(x)\,\mathrm{d}x. \end{aligned} \tag{12}$$

We prove that this definition satisfies the inversion Equation (10), using the calculation rules from the Appendix.

$$\begin{aligned} \big(\langle c^\dagger_\omega, \mathrm{id}\rangle \gg (c \gg \omega)\big)(M \times N) &\overset{(7)}{=} \int \langle c^\dagger_\omega, \mathrm{id}\rangle(-)(M \times N)\,\mathrm{d}(c \gg \omega) \\ &\overset{(A.2,A.3)}{=} \int \left(\int f(x) \cdot u(x,y)\,\mathrm{d}x \right) \cdot \langle c^\dagger_\omega, \mathrm{id}\rangle(y)(M \times N)\,\mathrm{d}y \\ &\overset{(A.6)}{=} \int \left(\int f(x) \cdot u(x,y)\,\mathrm{d}x \right) \cdot c^\dagger_\omega(y)(M) \cdot \mathbf{1}_N(y)\,\mathrm{d}y \\ &\overset{(12)}{=} \int_N \left(\int f(x) \cdot u(x,y)\,\mathrm{d}x \right) \cdot \frac{\int_M f(x) \cdot u(x,y)\,\mathrm{d}x}{\int f(x) \cdot u(x,y)\,\mathrm{d}x}\,\mathrm{d}y \\ &= \int_N \int_M f(x) \cdot u(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &\overset{(A.7)}{=} (\langle \mathrm{id}, c\rangle \gg \omega)(M \times N). \end{aligned}$$

5. Conjugate priors

We now come to the core of this paper. As described in the introduction, the informal definition says that a class of distributions is conjugate prior to a statistical model if the associated posteriors are *in the same class* of distributions. The posteriors can be computed via Bayesian inversion (12) of the statistical model.

This definition of 'conjugate prior' is a bit vague, since it loosely talks about 'classes of distributions', without further specification. As described in 'Idea 1' in Section 2, we interpret 'class of states on X' as a channel $P \to X$, where P is the type of parameters of the class.

We have already seen this channel-based description for the class Beta distributions, in (1), as channel $\mathrm{Beta}\colon \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to [0,1]$. This works more generally, for instance for Gaussian

(normal) distributions $\mathrm{Norm}(\mu, \sigma)$, where μ is the mean parameter and σ is the standard deviation parameter, giving a channel of the form:

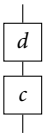
$$\mathbb{R} \times \mathbb{R}_{>0} \xrightarrow{\text{Norm}} \mathcal{G}(\mathbb{R}) \tag{13}$$

It is determined by its value on a measurable subset $M \subseteq \mathbb{R}$ as the standard integral:

$$\mathrm{Norm}(\mu, \sigma)(M) = \int_M \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, \mathrm{d}x \tag{14}$$

Given a channel $c \colon P \to X$, we shall look at states $c(p)$, for parameters $p \in P$, as priors. The statistical model, for which these $c(p)$'s will be described as conjugate priors, goes from X to some other object O of 'observations'. Thus our starting point is a statistical model consisting of a pair of (composable) channels of the form:

$$P \xrightarrow{c} X \xrightarrow{d} O \qquad \text{or, as diagram,} \qquad \tag{15}$$

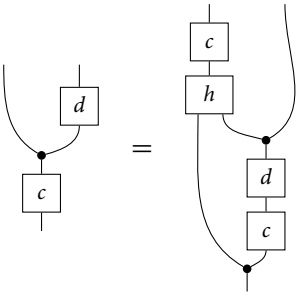


Such a pair of composable channels may be seen as a two-stage hierarchical Bayesian model, in which we standardly use $O =$ observables, $X =$ parameters of the statistical model, and $P =$ parameters over the parameters, i.e. hyperparameters, see e.g. Bernardo and Smith (2000). There, esp. in Definition 5.6 of conjugate priorship, one can also distinguish two channels, written as $p(\theta \mid \tau)$ and $p(x \mid \theta)$, corresponding respectively to our channels c and d. The τ form the hyperparameters.

In this setting we come to our main definition that formulates the notion of conjugate prior in an abstract manner, avoiding classes of distributions. It contains the crucial equation that was missing in the informal description in Section 2.

All our examples of (conjugate prior) channels are maps in the Kleisli category of the Giry monad, but the formulation applies more generally. In fact, abstraction purifies the situation and shows the essentials. The definition below speaks of 'deterministic' channels, between brackets. This part will be explained later on, in the beginning of Section 6. It can be ignored for now.

Definition 5.1. *In the situation* (15) *we call channel c a* conjugate prior *to channel d if there is a (deterministic) channel $h \colon P \times O \to P$ for which the following equation holds:*



(16)

Equivalently, in equational form: $\langle \mathrm{id}, d \rangle \circ c = ((c \circ h) \otimes \mathrm{id}) \circ \langle \mathrm{id}, Y \circ d \circ c \rangle$.

The idea is that the map $h \colon P \times O \to P$ translates parameters, with an observation from O as additional argument. Informally, one gets a posterior state $c(h(p, y))$ from the prior state $c(p)$, given the observation $y \in O$. The power of this 'analytic' approach is that it involves simple recomputation of parameters, instead of more complicated updating of entire states. This will be illustrated in several standard examples below.

The above Equation (16) is formulated in an abstract manner – which is its main strength. We will derive an alternative formulation of Equation (16) for pdf-channels. It greatly simplifies the calculations in examples.

Lemma 5.2. *Consider composable channels $P \xrightarrow{c} X \xrightarrow{d} O$, as in (15), for the Giry monad $\mathcal{G}$, where $c\colon P \to \mathcal{G}(X)$ and $d\colon X \to \mathcal{G}(O)$ are given by pdf's $u\colon P \times X \to \mathbb{R}_{\geq 0}$ and $v\colon X \times O \to \mathbb{R}_{\geq 0}$, as pdf-channels $c = \int u$ and $d = \int v$. Let c be conjugate prior to d, via a measurable function $h\colon P \times O \to P$.*

Equation (16) then amounts to, for an element $p \in P$ and for measurable subsets $M \subseteq X$ and $N \subseteq O$,

$$\begin{aligned} &\int_N \int_M u(p,x) \cdot v(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_N \left(\int u(p,x) \cdot v(x,y)\,\mathrm{d}x \right) \cdot \left(\int_M u(h(p,y),x)\,\mathrm{d}x \right) \mathrm{d}y. \end{aligned} \tag{17}$$

In order to prove this equation, it suffices to prove that the two functions under the outer integral $\int_N$ are equal, that is, it suffices to prove for each $y \in O$,

$$\int_M u(p,x) \cdot v(x,y)\,\mathrm{d}x = \left(\int u(p,x) \cdot v(x,y)\,\mathrm{d}x \right) \cdot \left(\int_M u(h(p,y),x)\,\mathrm{d}x \right). \tag{18}$$

This formulation will be used in the examples below.

Proof. We extensively use the equations for integration from Section 3 and from the Appendix, in order to prove (17). The left-hand side of Equation (16) gives the left-hand side of (17):

$$\big(\langle \mathrm{id}, d\rangle \circ c\big)(p)(M \times N) \overset{(8)}{=} \big(\langle \mathrm{id}, d\rangle \gg c(p)\big)(M \times N) \overset{(A.7)}{=} \int_N \int_M u(p,x) \cdot v(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Unravelling the right-hand side of (16) is a bit more work:

$$\begin{aligned} &\big((c \circ h) \otimes \mathrm{id}) \circ \langle \mathrm{id}, Y \circ d \circ c\rangle\big)(p)(M \times N) \\ &\overset{(8)}{=} \int (c \circ h) \otimes \mathrm{id})(-)(M \times N)\,\mathrm{d}\langle \mathrm{id}, Y \circ d \circ c\rangle(p) \\ &\overset{(A.6)}{=} \int ((c \circ h) \otimes \mathrm{id})(-)(M \times N)\,\mathrm{d}\big(\eta(p) \otimes (Y \circ d \circ c)(p)\big) \\ &\overset{(A.1)}{=} \int \int ((c \circ h) \otimes \mathrm{id})(-,-)(M \times N)\,\mathrm{d}\eta(p)\,\mathrm{d}(Y \circ d \circ c)(p)\big) \\ &= \int ((c \circ h) \otimes \mathrm{id})(p,-)(M \times N)\,\mathrm{d}\mathcal{G}(Y)(d \gg c(p)) \\ &\overset{(5)}{=} \int ((c \circ h) \otimes \mathrm{id})(p, Y(-))(M \times N)\,\mathrm{d}(d \gg c(p)) \\ &\overset{(A.2,A.4)}{=} \int \left(\int u(p,x) \cdot v(x,y)\,\mathrm{d}x \right) \cdot ((c \circ h) \otimes \mathrm{id})(p,y,y)(M \times N)\,\mathrm{d}y \\ &= \int \left(\int u(p,x) \cdot v(x,y)\,\mathrm{d}x \right) \cdot c(h(p,y))(M) \cdot \mathbf{1}_N(y)\,\mathrm{d}y \\ &= \int_N \left(\int u(p,x) \cdot v(x,y)\,\mathrm{d}x \right) \cdot \left(\int_M u(h(p,y),x)\,\mathrm{d}x \right) \mathrm{d}y. \end{aligned}$$

By combining this outcome with the earlier one, we get the desired equation (17). □

One can reorganise Equation (18) as a normalisation fraction:

$$\int_M u(h(p,y),x)\,\mathrm{d}x = \frac{\int_M u(p,x)\cdot v(x,y)\,\mathrm{d}x}{\int u(p,x)\cdot v(x,y)\,\mathrm{d}x}. \tag{19}$$

It now strongly resembles Equation (12) for Bayesian inversion. This connection will be established more generally in Theorem 6.3. Essentially, the above normalisation fraction (19) occurs in Bernardo and Smith (2000, Defnition 5.6).

We are now ready to review some standard examples. The first one describes the structure underlying the coin example in Section 2.

Example 5.3. It is well known that the beta distributions are conjugate prior to the Bernoulli 'flip' likelihood function. We shall reformulate this fact following the pattern of Definition 5.1, with two composable channels, as in (15), namely:

$$\mathbb{N}_{>0}\times\mathbb{N}_{>0} \xrightarrow{\text{Beta}} [0,1] \xrightarrow{\text{Flip}} 2 \qquad \text{where } 2=\{0,1\}.$$

The Beta channel is as in (1), but now restricted to the non-negative natural numbers $\mathbb{N}_{>0}$. We recall that the normalisation constant $B(\alpha,\beta)$ is $\int_{[0,1]} x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x$.

The Flip channel sends a probability $r\in[0,1]$ to the Bernoulli(r) distribution, which can also be written as a discrete distribution $\mathrm{Flip}(r) = r|1\rangle + (1-r)|0\rangle$. More formally, as a Kleisli map $[0,1]\to\mathcal{G}(2)$ it is, for a subset $N\subseteq 2$,

$$\mathrm{Flip}(r)(N) = \int_N r^i\cdot(1-r)^{1-i}\,\mathrm{d}i = \sum_{i\in N} r^i\cdot(1-r)^{1-i} = \begin{cases} 0 & \text{if } N=\emptyset \\ r & \text{if } N=\{1\} \\ 1-r & \text{if } N=\{0\} \\ 1 & \text{if } N=\{0,1\}. \end{cases}$$

The i in $\mathrm{d}i$ refers here to the counting measure.

In order to show that Beta is a conjugate prior of Flip, we have to produce a parameter translation function $h\colon \mathbb{N}_{>0}\times\mathbb{N}_{>0}\times 2\to\mathbb{N}_{>0}\times\mathbb{N}_{>0}$. It is defined by distinguishing the elements in $2=\{0,1\}$

$$h(\alpha,\beta,1) = (\alpha+1,\beta) \qquad \text{and} \qquad h(\alpha,\beta,0) = (\alpha,\beta+1). \tag{20}$$

Thus, in one formula, $h(\alpha,\beta,i) = (\alpha+i,\beta+(1-i))$.

We prove Equation (18) for $c = \mathrm{Beta} = \int u$ and $d = \mathrm{Flip} = \int v$. We start from its right-hand side, for an arbitrary $i\in 2$,

$$\begin{aligned}
&\left(\int u(\alpha,\beta,x)\cdot v(x,i)\,\mathrm{d}x\right)\cdot\left(\int_M u(h(\alpha,\beta,i),x)\,\mathrm{d}x\right) \\
&= \left(\int \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\cdot x^i\cdot(1-x)^{1-i}\,\mathrm{d}x\right)\cdot\left(\int_M \frac{x^{\alpha+i-1}(1-x)^{\beta+(1-i)-1}}{B(\alpha+i,\beta+(1-i))}\,\mathrm{d}x\right) \\
&= \left(\frac{\int x^{\alpha+i-1}(1-x)^{\beta+(1-i)-1}\,\mathrm{d}x}{B(\alpha,\beta)}\right)\cdot\left(\int_M \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha+i,\beta+(1-i))}\cdot x^i\cdot(1-x)^{1-i}\,\mathrm{d}x\right) \\
&= \left(\frac{B(\alpha+i,\beta+(1-i))}{B(\alpha,\beta)}\right)\cdot\left(\int_M \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha+i,\beta+(1-i))}\cdot x^i\cdot(1-x)^{1-i}\,\mathrm{d}x\right) \\
&= \int_M \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\cdot x^i\cdot(1-x)^{1-i}\,\mathrm{d}x \\
&= \int_M u(\alpha,\beta,x)\cdot v(x,i)\,\mathrm{d}x.
\end{aligned}$$

The latter expression is the left-hand side of (18). We see that the essence of the verification of the conjugate prior equation is the shifting of functions and normalisation factors. This is a general pattern.

In the remainder of this section, we review how some of the standard examples fit in the current setting, without each time doing the entire 'big' calculation to verify Equation (18).

Example 5.4. In a similar way, one verifies that the Beta channel is a conjugate prior to the binomial channel. For the latter we fix a natural number $n > 0$, and consider the two channels:

$$\mathbb{N}_{>0} \times \mathbb{N}_{>0} \xrightarrow{\text{Beta}} [0,1] \xrightarrow{\text{Binom}_n} \{0, 1, \ldots, n\}$$

The binomial channel Binom_n is defined for $r \in [0,1]$ and $M \subseteq \{0, 1, \ldots, n\}$ as:

$$\text{Binom}_n(r)(M) = \int_M \binom{n}{i} \cdot r^i \cdot (1-r)^{n-i}\, \mathrm{d}i = \sum_{i \in M} \binom{n}{i} \cdot r^i \cdot (1-r)^{n-i}.$$

The conjugate prior property requires in this situation a parameter translation function $h\colon \mathbb{N}_{>0} \times \mathbb{N}_{>0} \times \{0, 1, \ldots, n\} \to \mathbb{N}_{>0} \times \mathbb{N}_{>0}$, given by: $h(\alpha, \beta, i) = (\alpha + i, \beta + n - i)$.

Here is another well-known conjugate prior relationship, namely between Dirichlet and 'multinomial' distributions. The latter are simply called discrete distributions in the present context.

Example 5.5. Here we shall identify a number $n \in \mathbb{N}$ with the n-element set $\{0, 1, \ldots, n-1\}$. We then write $\mathcal{D}_*(n)$ for the set of n-tuples $(x_0, \ldots, x_{n-1}) \in (\mathbb{R}_{>0})^n$ with $\sum_i x_i = 1$.

For a fixed $n > 0$, let $O = \{y_0, \ldots, y_{n-1}\}$ be a set of 'observations'. We consider the following two channels.

$$(\mathbb{N}_{>0})^n \xrightarrow{\text{Dir}_n} \mathcal{D}_*(n) \xrightarrow{\text{Mult}} O$$

The multinomial channel is defined as $\text{Mult}(x_0, \ldots, x_{n-1}) = x_0|y_0\rangle + \cdots + x_{n-1}|y_{n-1}\rangle$. The Dirichlet channel Dir_n is more complicated: for an n-tuple $\vec{\alpha} = (\alpha_0, \ldots, \alpha_{n-1})$ it is given via pdf's d_n, in:

$$\text{Dir}_n(\vec{\alpha}) = \int d_n(\vec{\alpha}) \qquad \text{where} \qquad d_n(\vec{\alpha})(x_0, \ldots, x_{n-1}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \cdot \prod_i x_i^{\alpha_i - 1},$$

for $(x_0, \ldots, x_{n-1}) \in \mathcal{D}_*(n)$. The operation Γ is the 'Gamma' function, which is defined on natural numbers $k > 1$ as $\Gamma(k) = (k-1)!$.

The parameter translation function $h\colon (\mathbb{N}_{>0})^n \times O \to (\mathbb{N}_{>0})^n$ is:

$$h(\alpha_0, \ldots, \alpha_{n-1}, y) = (\alpha_0, \ldots, \alpha_i + 1, \ldots, \alpha_{n-1}) \quad \text{if } y = y_i.$$

Example 5.6. The γ distribution is conjugate prior to the Poisson distribution, in:

$$\mathbb{R}_{>0} \times \mathbb{R}_{>0} \xrightarrow{\text{Gam}} \mathbb{R}_{>0} \xrightarrow{\text{Pois}} \mathbb{N}$$

where:

$$\text{Gam}(\alpha, \beta)(M) = \int_M \frac{\beta^\alpha \cdot x^{\alpha-1} \cdot e^{-\beta x}}{\Gamma(\alpha)}\, \mathrm{d}x \quad \text{and} \quad \text{Pois}(x) = \sum_{k=0}^{k=\infty} e^{-x} \cdot \frac{x^k}{k!} |k\rangle.$$

The parameter translation function $h\colon \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{N} \to \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ is: $h(\alpha, \beta, k) = (\alpha + k, \beta + 1)$.

We include one more example, illustrating that normal channels are conjugate priors to themselves. This fact is also well known.

Example 5.7. Consider the following two normal channels.

$$\mathbb{R} \times \mathbb{R}_{>0} \xrightarrow{\text{Norm}} \mathbb{R} \xrightarrow{\text{Norm}(-,\nu)} \mathbb{R}_{>0}$$

The channel Norm is described in (13); it is used twice here, the second time with a fixed standard deviation ν, for 'noise'. This second channel is typically used for observation, like in Kalman filtering, for which a fixed noise level can be assumed. In this situation, the first normal channel Norm is a conjugate prior to the second channel $\text{Norm}(-, \nu)$, via the parameter translation function $h\colon \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R} \times \mathbb{R}_{>0}$ given by:

$$h(\mu, \sigma, y) = \left(\frac{\mu \cdot \nu^2 + y \cdot \sigma^2}{\nu^2 + \sigma^2}, \frac{\nu \cdot \sigma}{\sqrt{\nu^2 + \sigma^2}} \right)$$

6. Conjugate priors form bayesian inversions

This section connects the main two notions of this paper, by showing that conjugate priors give rise to Bayesian inversion. The argument is a very simple example of diagrammatic reasoning. Before we come to it, we have to clarify an issue that was left open earlier, regarding 'deterministic' channels, see Definition 5.1.

Definition 6.1. *A channel c is called* deterministic *if it commutes with copiers, that is, if it satisfies the equation on the left below.*

As a special case, a state ω is called deterministic if it satisfies the equation on the right, above.

The state description is a special case of the channel description since a state on X is a channel $1 \to X$ and copying on the trivial (final) object 1 does nothing, up to isomorphism.

Few channels (or states) are deterministic. In deterministic and continuous computation, the ordinary functions $f\colon X \to Y$ are deterministic, when considered as a channel $\eta \circ f$. We check this explicitly for point states, since this is what we need later on.

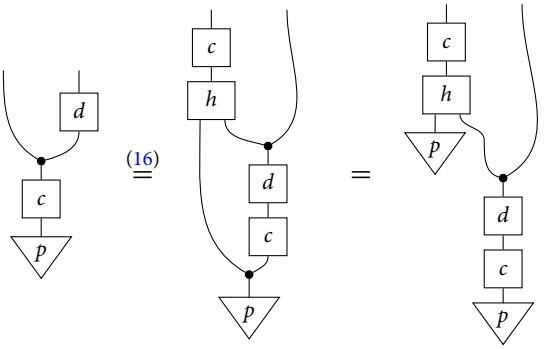
Example 6.2. Let x be an element of a measurable space X. The associated point state $\eta(x) \in \mathcal{G}(X)$ is deterministic, where $\eta(x)(M) = \mathbf{1}_M(x)$. We check the equation on the right in Definition 6.1:

$$\begin{aligned} \big(\curlyvee \circ \eta(x)\big)(M \times N) &= \eta(x, x)(M \times N) = \mathbf{1}_{M \times N}(x, x) = \mathbf{1}_M(x) \cdot \mathbf{1}_N(x) \\ &= \eta(x)(M) \cdot \eta(x)(N) = \big(\eta(x) \otimes \eta(x)\big)(M \times N). \end{aligned}$$

We now come to the main result.

Theorem 6.3. *Let $P \xrightarrow{c} X \xrightarrow{d} O$ be channels, where c is conjugate prior to d, say via $h\colon P \times O \to P$. Then for each deterministic (copyable) state p, the map $c \circ h(p, -)\colon O \to X$ is a Bayesian inversion of d, wrt. the transformed state $c \gg p$.*

Proof. We have to prove Equation (10), for channel d and state $c \gg p$, with the channel $c \circ h(p, -)$ playing the role of Bayesian inversion $d^{\dagger}_{c \gg p}$. This is easiest to see graphically, using that the state p is deterministic and thus commutes with copiers $\curlyvee$, see the equation on the right in Definition 6.1.



This is it. □

When we specialise to Giry-channels, we get an 'if-and-only-if' statement, since there we can reason elementwise.

Corollary 6.4. *Let $P \xrightarrow{c} X \xrightarrow{d} O$ be two channels in $\mathcal{K}\ell(\mathcal{G})$, and let $h\colon P \times O \to P$ be a measurable function. The following two points are equivalent:*

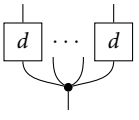
(i) c is a conjugate prior to d, via h;
(ii) $c(h(p, -))\colon O \to \mathcal{G}(X)$ is a Bayesian inverse for channel d with state $c(p)$, i.e. is $d^{\dagger}_{c(p)}$, for each parameter $p \in P$. □

7. Multiple updates

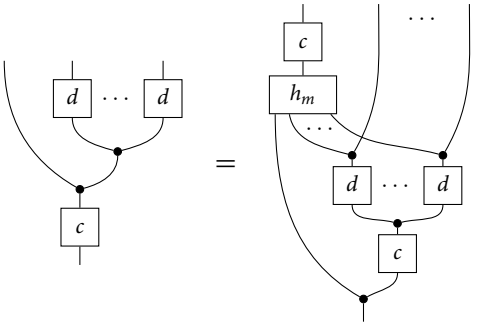
So far we have dealt with the situation where there is a single observation $y \in O$ that leads to an update of a prior distribution. In this final section we briefly look at how to handle multiple observations $y_1, \ldots, y_m \in O$. It will lead to the notion of *sufficient statistic*.

Multiple updates fit in our set-up $P \xrightarrow{c} X \xrightarrow{d} O$ by using a tuple-channel as output:

$$P \xrightarrow{\ c\ } X \xrightarrow{\ \langle d,\ldots,d\rangle\ } O^m \qquad \text{where} \qquad \langle d, \ldots, d\rangle \quad =$$



Following Definition 5.1 the channel c is conjugate prior to the m-tuple $\langle d, \ldots, d\rangle$ if there is a (deterministic) parameter translation function $h_m\colon P \times O^m \to P$ with:



(21)

This construction involves in essence a Bayesian inversion, as in Section 6.

We shall illustrate this for the Beta–Flip relationship from Example 5.3. Suppose we have multiple head/tail observations $y_1, \ldots, y_m \in 2 = \{0, 1\}$ which we wish to incorporate into a prior

distribution Beta(α, β). We use $\sum_i y_i$ as the number of 1's among the observations y_i and $\sum_i (1 - y_i)$ as the number of 0's, and then define as (multiple) parameter update function:

$$\mathbb{N}_{>0} \times \mathbb{N}_{>0} \times 2^m \xrightarrow{h_m} \mathbb{N}_{>0} \times \mathbb{N}_{>0} \quad \text{via} \quad h_m(\alpha, \beta, \vec{y}) = (\alpha + \textstyle\sum_i y_i, \beta + \sum_i (1 - y_i)) \tag{22}$$

Equation (21) then holds: using a reformulation in the style of (18), and with pdf's u, v, as in Example 5.3, it suffices to prove:

$$\left(\int u(\alpha, \beta, x) \cdot v(x, y_1) \cdot \ldots \cdot v(x, y_m) \, \mathrm{d}x \right) \cdot \left(\int_M u(h_m(\alpha, \beta, \vec{y}), x) \, \mathrm{d}x \right)$$
$$= \int_M u(\alpha, \beta, x) \cdot v(x, y_1) \cdot \ldots \cdot v(x, y_m) \, \mathrm{d}x.$$

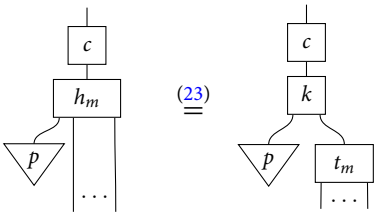
We see that for this multiple updating, we do not need the individual observations, $y_1, \ldots, y_m \in 2$, but only the sums $\sum_i y_i$ and $\sum_i (1 - y_i) = m - \sum_i y_i$. In many such cases the situation can be simplified via factorisation, using the notion of sufficient statistic, see e.g. Koopman (1936) and Bishop (2006). We adapt it to the current setting.

Definition 7.1. *A* sufficient statistic *for a collection of parameter update functions* $\big(h_m \colon P \times O^m \to P\big)_{m \in \mathbb{N}}$ *is a collection of functions* $\big(t_m \colon O^m \to Z\big)_{m \in \mathbb{N}}$ *with* $k \colon P \times Z \to P$ *such that the following triangle commutes for each* $m \in \mathbb{N}$.

$$\begin{array}{ccc} P \times O^m & \xrightarrow{h_m} & P \\ {\scriptstyle \mathrm{id} \times t_m}\downarrow & \nearrow_{k} & \\ P \times Z & & \end{array} \tag{23}$$

Crucially, the set Z *does not depend on* m *and* t_m *does not act on the parameters* P. *Moreover, in examples,* Z *is often simpler than these* O^m.

Let's return to the situation in (21) where a channel c is conjugate prior to an m-tuple $\langle d, \ldots, d \rangle \colon X \to O^m$. In line with (the proof of) Theorem 6.3, for an element $p \in P$ we obtain a Bayesian inversion $O^m \to X$ of the tuple $\langle d, \ldots, d \rangle$ wrt. state $c \gg p$ on the left:



In the presence of a sufficient statistic, this Bayesian inversion takes the form as on the right.

The factorisation in the above triangle (23) may satisfy an obvious universal property once we add the requirement that the functions t_m satisfy a certain invariance property. The latter property does not seem to be uniform across examples, as will be illustrated below.

Example 7.2. We mention sufficient statistics for some of the examples from Section 5.

(1) As discussed in the beginning of this section, for Beta–Flip in Example 5.3 we have $P = \mathbb{N}_{>0} \times \mathbb{N}_{>0}$ and $O = 2 = \{0, 1\}$. The sufficient statistic has $Z = \mathbb{N} \times \mathbb{N}$ with:

$$2^m \xrightarrow{t_m} \mathbb{N} \times \mathbb{N} \qquad \mathbb{N}_{>0} \times \mathbb{N}_{>0} \times \mathbb{N} \times \mathbb{N} \xrightarrow{k} \mathbb{N}_{>0} \times \mathbb{N}_{>0}$$
$$\vec{y} \longmapsto (\textstyle\sum_i y_i, \sum_i (1 - y_i)) \qquad (\alpha, \beta, n_1, n_0) \longmapsto (\alpha + n_1, \beta + n_0).$$

Clearly, $k \circ (\mathrm{id} \times t_m) = h_m$, for the functions h_m in (22).

We notice that the functions t_m are 'invariant under permutation' in the sense that:

$$t_m(y_1, \ldots, y_i, \ldots, y_j, \ldots, y_m) = t_m(y_1, \ldots, y_j, \ldots, y_i, \ldots, y_m),$$

for each $1 \leq i, j \leq m$. This invariance allows us to state that the above (t_m) and k satisfy the following universal property: for each collection of functions $s_m \colon O^m \to W$ invariant under permutation, together with $\ell \colon P \times W \to P$ such that $\ell \circ (\mathrm{id} \times s_m) = h_m$, there is a unique $f \colon Z \to W$ with $f \circ t_m = s_m$, for each $m \in \mathbb{N}$, and $\ell \circ (\mathrm{id} \times f) = k$.
Indeed, we can define $f \colon \mathbb{N} \times \mathbb{N} \to W$ as:

$$f(n_1, n_0) = s_{n_1+n_0}\big(\underbrace{1, \ldots, 1}_{n_1 \text{ times}}, \underbrace{0, \ldots, 0}_{n_0 \text{ times}}\big).$$

Since these s functions are invariant under permutation, the precise order of the 1's and 0's does not matter. Then:

$$\begin{aligned}
\big(f \circ t_m\big)(y_1, \ldots, y_m) &= f\big(\textstyle\sum_i y_i, \sum_i 1 - y_i\big) \\
&= s_m\big(1, \ldots, 1, 0, \ldots, 0\big) = s_m(y_1, \ldots, y_m) \\
\big(\ell \circ (\mathrm{id} \times f)\big)(\alpha, \beta, n_1, n_0) &= \ell\big(\alpha, \beta, s_{n_1+n_0}(1, \ldots, 1, 0, \ldots, 0)\big) \\
&= h_{n_1+n_0}\big(\alpha, \beta, 1, \ldots, 1, 0, \ldots, 0)\big) \\
&\overset{(22)}{=} (\alpha + n_1, \beta + n_0) = k(\alpha, \beta, n_1, n_0).
\end{aligned}$$

We further have to check uniqueness: if $g \colon \mathbb{N} \times \mathbb{N} \to W$ also satisfies $g \circ t_m = s_m$ and $\ell \circ (\mathrm{id} \times g) = k$, then $f = g$ since:

$$f(n_1, n_0) = s_{n_1+n_0}(1, \ldots, 1, 0, \ldots, 0) = g\big(t_m(1, \ldots, 1, 0, \ldots, 0)\big) = g(n_1, n_0).$$

(2) For the Gam–Pois conjugate priorship in Example 5.6, we have $P = \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and $O = \mathbb{N}$ as spaces of parameters and observations, and $h_m \colon P \times O^m \to P$ is defined by $h_m(\alpha, \beta, \vec{y}) = (\alpha + \sum_i y_i, \beta + m)$. The sufficient statistic has the same $Z = \mathbb{N} \times \mathbb{N}$ and $k(\alpha, \beta, z, n) = (\alpha + z, \beta + n)$, much as in the previous point, but with:

$$\mathbb{N}^m \xrightarrow{t_m} \mathbb{N} \times \mathbb{N} \qquad \text{given by} \qquad t_m\big(\vec{y}\big) = \big(\textstyle\sum_i y_i, m\big).$$

The invariance property in this case for these t_m is:

$$t_m(y_1, \ldots, y_i + z, \ldots, y_j, \ldots, y_m) = t_m(y_1, \ldots, y_i, \ldots, y_j + z, \ldots, y_m),$$

With this invariance, we can prove a universal property of the factorisation (t_m) and k.

(3) For the Norm–Norm conjugate priorship from Example 5.7, we have $P = \mathbb{R} \times \mathbb{R}_{>0}$ and $O = \mathbb{R}_{>0}$. We can take $Z = \mathbb{R}_{>0} \times \mathbb{N}$ with:

$$\begin{array}{ccc}
(\mathbb{R}_{>0})^m \xrightarrow{t_m} \mathbb{R}_{>0} \times \mathbb{N} & \qquad & \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{N} \xrightarrow{k} \mathbb{R} \times \mathbb{R}_{>0} \\
\vec{y} \longmapsto \big(\textstyle\sum_i y_i, m\big) & & (\mu, \sigma, s, n) \longmapsto \Big(\frac{\mu\nu^2 + s\sigma^2}{\nu^2 + n\sigma^2}, \frac{\nu\sigma}{\sqrt{\nu^2 + n\sigma^2}}\Big)
\end{array}$$

The proof that this works is laborious but essentially straightforward. Universality is obtained if we use the same invariance property as in the previous point.

8. Conclusions

This paper contains a novel view on conjugate priors, using the concept of channel in a systematic manner. It introduces a precise definition for conjugate priorship, using a pair of composable channels $P \to X \to O$ and a parameter translation function $P \times O \to P$, satisfying a non-trivial

equation, see Definition 5.1. It has been shown that this equation holds for several standard conjugate prior examples. There are many more examples that have not been checked here. One can be confident that the same equation holds for those unchecked examples too, since it has been shown here that conjugate priors amount to Bayesian inversions. This inversion property is the essential characteristic for conjugate priors. It has been extended to multiple updates, using the concept of sufficient statistic.

Acknowledgements. Thanks are due to the two anonymous reviewers, one of whose comments significantly improved Section 7.

Notes

1 See https://en.wikipedia.org/wiki/Conjugate_prior or online lists, such as https://www.johndcook.com/CompendiumOfConjugatePriors.pdf, consulted on Sept. 10, 2018

2 Sometimes these distributions $\sum_i r_i|x_i\rangle$ are called 'multinomial' or 'categorical'; the latter terminology is confusing in the present context.

3 In Jacobs (2013), integration $\int f\,\mathrm{d}\omega$ is defined only for $[0,1]$-valued functions f, but that does not matter for the relevant equations, except that integrals may not exist for $\mathbb{R}_{\geq 0}$-valued functions (or have value ∞). These integrals are determined by their valued $\int \mathbf{1}_M\,\mathrm{d}\omega = \omega(M)$ on indicator functions $\mathbf{1}_M$ for measurable subsets, via continuous and linear extensions, see also Jacobs and Westerbaan (2015).

References

Abramsky, S. and Coecke, B. (2009). A categorical semantics of quantum protocols. In: Engesser, K., Gabbay, D. M. and Lehmann, D. (eds.) *Handbook of Quantum Logic and Quantum Structures: Quantum Logic*, North-Holland, Elsevier, Computer Science Press, 261–323.

Ackerman, N., Freer, C. and Roy, D. (2011). Noncomputable conditional distributions. In: Dawar, A. and Grädel, E. (eds.) *Logic in Computer Science*, IEEE, Computer Science Press.

Alpaydin, E. (2010). *Introduction to Machine Learning*, 2nd edn., Cambridge, MA, MIT Press.

Bernardo, J. and Smith, A. (2000). *Bayesian Theory*. Chichester, John Wiley & Sons.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Chichester, Springer.

Blackwell, D. (1951). Comparison of experiments. In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Springer/British Computer Society, 93–102.

Cho, K. and Jacobs, B. (2019). Disintegration and Bayesian inversion, both abstractly and concretely. In: *Mathematical Structures in Computer Science*. See also arxiv.org/abs/1709.00322.

Cho, K., Jacobs, B., Westerbaan, A. and Westerbaan, B. (2015). An introduction to effectus theory. See arxiv.org/abs/1512.05813.

Clerc, F., Dahlqvist, F., Danos, V. and Garnier, I. (2017). Pointless learning. In: Esparza, J. and Murawski, A. (eds.) *Foundations of Software Science and Computation Structures*, Lecture Notes in Computer Science, vol. 10203, Berlin, Springer, 355–369.

Culbertson, J. and Sturtz, K. (2014). A categorical foundation for Bayesian probability. *Applied Categorical Structures* **22**(4) 647–662.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics* **7**(2) 269–281.

Faden, A. (1985). The existence of regular conditional probabilities: Necessary and sufficient conditions. *The Annals of Probability* **13**(1) 288–298.

Giry, M. (1982). A categorical approach to probability theory. In: Banaschewski, B. (ed.) *Categorical Aspects of Topology and Analysis*, Lecture Notes in Computer Science, vol. 915, Berlin, Springer, 68–85.

Jacobs, B. (2013). Measurable spaces and their effect logic. In: *Logic in Computer Science*, IEEE, Computer Science Press.

Jacobs, B. (2015). New directions in categorical logic, for classical, probabilistic and quantum logic. *Logical Methods in Computer Science* **11**(3). See https://lmcs.episciences.org/1600.

Jacobs, B. (2018). From probability monads to commutative effectuses. *Journal of Logical and Algebraic Methods in Programming* **94** 200–237.

Jacobs, B. and Westerbaan, A. (2015). An effect-theoretic account of Lebesgue integration. In: Ghica, D. (ed.) *Mathematical Foundations of Programming Semantics*, Electronic Notes in Theoretical Computer Science, vol. 319, Amsterdam, Elsevier, 239–253.

Jacobs, B. and Zanasi, F. (2019). The logical essentials of Bayesian reasoning. In: Barthe, G. and Katoen, J.-P. and Silva, A. *Probabilistic Programming*, Cambridge, Cambridge University Press. See arxiv.org/abs/1804.01193.

Koopman, B. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society* **39** 399–409.
Panangaden, P. (2009). *Labelled Markov Processes*. London, Imperial College Press.
Russell, S. and Norvig, P. (2003). *Artificial Intelligence. A Modern Approach*. Englewood Cliffs, NJ, Prentice Hall.
Selinger, P. (2007). Dagger compact closed categories and completely positive maps (extended abstract). In: Selinger, P. (ed.) *Proceedings of the 3rd International Workshop on Quantum Programming Languages (QPL 2005)*, Electronic Notes in Theoretical Computer Science, vol. 170, Amsterdam, Elsevier, 139–163. doi: http://dx.doi.org/10.1016/j.entcs.2006.12.018.
Selinger, P. (2011). A survey of graphical languages for monoidal categories. In: Coecke, B. (ed.) *New Structures in Physics*, Lecture Notes in Physics, vol. 813, Berlin, Springer, 289–355.
Stoyanov, J. (2014). *Counterexamples in Probability*, 2nd rev. edn. Chichester, Wiley.

Appendix A. Calculation laws for Giry–Kleisli Maps with pdf's

We assume that for a probability distribution (state) $\omega \in \mathcal{G}(X)$ and a measurable function $f\colon X \to \mathbb{R}_{\geq 0}$, the integral $\int f \,\mathrm{d}\omega \in [0, \infty]$ can be defined as a limit of integrals over simple functions that approximate f. We shall follow the description of Jacobs (2013), to which we refer for details.[3] This integration satisfies the Fubini property, which can be formulated, for states $\omega \in \mathcal{G}(X)$, $\rho \in \mathcal{G}(Y)$ and measurable function $h\colon X \times Y \to \mathbb{R}_{\geq 0}$, as:

$$\int h \,\mathrm{d}(\omega \otimes \rho) = \int \int h \,\mathrm{d}\omega \,\mathrm{d}\rho. \tag{A.1}$$

The product state $\omega \otimes \rho \in \mathcal{G}(X \times Y)$ is defined by $(\omega \otimes \rho)(M \times N) = \omega(M) \cdot \rho(N)$.

A.1 States via pdf's

For a subset $X \subseteq \mathbb{R}$, a measurable function $f\colon X \to \mathbb{R}_{\geq 0}$ is called a probability density function (pdf) for a state $\omega \in \mathcal{G}(X)$ if $\omega(M) = \int_M f(x)\,\mathrm{d}x$ for each measurable subset $M \subseteq X$. In that case we simply write $\omega = \int f(x)\,\mathrm{d}x$, or even $\omega = \int f$. If ω is given by such a pdf f, integration with state ω can be described as:

$$\int g \,\mathrm{d}\omega = \int f(x) \cdot g(x)\,\mathrm{d}x. \tag{A.2}$$

A.2 Channels via pdf's

Let channel $c\colon X \to \mathcal{G}(Y)$ be given as $c = \int u$ by pdf $u\colon X \times Y \to \mathbb{R}_{\geq 0}$ as $c(x)(N) = \int_N u(x,y)\,\mathrm{d}y$, for each $x \in X$ and measurable $N \subseteq Y$, like in (9). If $\omega = \int f$ is a state on X, then state transformation $c \gg \omega \in \mathcal{G}(Y)$ is given by:

$$\begin{aligned}(c \gg \omega)(N) &\overset{(7)}{=} \int c(-)(N) \,\mathrm{d}\omega \overset{(A.2)}{=} \int f(x) \cdot c(x)(N)\,\mathrm{d}x \\ &= \int_N \int f(x) \cdot u(x,y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned} \tag{A.3}$$

Hence the pdf of the transformed state $c \gg \omega$ is $y \mapsto \int f(x) \cdot u(x,y)\,\mathrm{d}x$.

Given a channel $d\colon Y \to \mathcal{G}(Z)$, say with $d = \int v$, then sequential channel composition $d \circ c$ is given, for $x \in X$ and $K \subseteq Z$, by:

$$\begin{aligned}(d \circ c)(x)(K) &\overset{(8)}{=} \int d(-)(K) \,\mathrm{d}c(x) \overset{(A.2)}{=} \int u(x,y) \cdot d(y)(K)\,\mathrm{d}y \\ &= \int_K \int u(x,y) \cdot v(y,z)\,\mathrm{d}y\,\mathrm{d}z\end{aligned} \tag{A.4}$$

We see that the pdf of the channel $d \circ c$ is $(x,z) \mapsto \int u(x,y) \cdot v(y,z)\,\mathrm{d}y$.

For a channel $e = \int w \colon A \to \mathcal{G}(B)$, we get a parallel composition channel $c \otimes e \colon X \times A \to \mathcal{G}(Y \times B)$ given by:

$$\begin{aligned}(c \otimes e)(x,a)(M \times N) &= c(x)(M) \otimes e(a)(N) \\ &= \left(\int_M u(x,y)\,\mathrm{d}y\right) \cdot \left(\int_N w(a,b)\,\mathrm{d}b\right) \\ &= \int_{M\times N} u(x,y) \cdot w(a,b)\,\mathrm{d}(y,b).\end{aligned} \tag{A.5}$$

Hence the pdf of the channel $c \otimes d$ is $(x,a,y,b) \mapsto u(x,y) \cdot w(a,b)$.

A.3 Graph channels and pdf's

For a channel $c \colon X \to \mathcal{G}(Y)$, we can form 'graph' channels $\langle \mathrm{id}, c\rangle = (\mathrm{id} \otimes c) \circ \curlyvee \colon X \to \mathcal{G}(X \times Y)$ and $\langle c, \mathrm{id}\rangle = (c \otimes \mathrm{id}) \circ \curlyvee \colon X \to \mathcal{G}(Y \times X)$. For $x \in X$ we have:

$$\langle \mathrm{id}, c\rangle(x) = \eta(x) \otimes c(x) \qquad \text{and} \qquad \langle c, \mathrm{id}\rangle(x) = c(x) \otimes \eta(x). \tag{A.6}$$

If $c = \int u$ and $\omega = \int f$ is a state on X, then:

$$\begin{aligned}(\langle \mathrm{id}, c\rangle \gg \omega)(M \times N) &\overset{(A.2)}{=} \int f(x) \cdot \langle \mathrm{id}, c\rangle(x)(M \times N)\,\mathrm{d}x \\ &\overset{(A.6)}{=} \int f(x) \cdot \eta(x)(M) \cdot c(x)(N)\,\mathrm{d}x \\ &= \int_N \int_M f(x) \cdot u(x,y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned} \tag{A.7}$$

We also consider the situation where $d \colon X \times Y \to \mathcal{G}(Z)$ is of the form $d = \int v$, with composition:

$$\begin{aligned}\big(d \circ \langle \mathrm{id}, c\rangle\big)(x)(K) &\overset{(A.6)}{=} \int d(-)(K)\,\mathrm{d}(\eta(x) \otimes c(x)) \\ &\overset{(A.1)}{=} \int d(-)(K)\,\mathrm{d}\eta(x)\,\mathrm{d}c(x) \\ &= \int d(x,-)(K)\,\mathrm{d}c(x) \\ &\overset{(A.2)}{=} \int u(x,y) \cdot d(x,y)(K)\,\mathrm{d}y \\ &= \int_K \int u(x,y) \cdot v(x,y,z)\,\mathrm{d}y\,\mathrm{d}z.\end{aligned} \tag{A.8}$$

Hence the pdf of the channel $d \circ \langle \mathrm{id}, c\rangle$ is $(x,z) \mapsto \int u(x,y) \cdot v(x,y,z)\,\mathrm{d}y$.

Cite this article: Jacobs B (2020). A channel-based perspective on conjugate priors. *Mathematical Structures in Computer Science* **30**, 44–61. https://doi.org/10.1017/S0960129519000082