A Channel-Based Perspective on Conjugate Priors

Simons Institute, Berkeley

Bart Jacobs, Radboud University Nijmegen
bart@cs.ru.nl

Dec 12, 2017

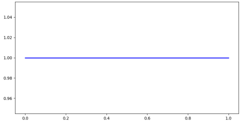
Where we are, so far

Standard example: Finding the bias of a coin

- Suppose we have a coin with unkown bias, and perform a number of tests, giving certain head/tail outcomes
  - We like to learn what the bias is
- This bias is an (unkown) number $r \in [0, 1]$.
  - with discrete coin distribution $\mathrm{Flip}(r) = r| H \rangle + (1 - r)| T \rangle$
  - aim: learn a continuous probability distribution on $[0, 1]$ for r
  - and possibly also a resulting expected value
- Standard procedure:
  - start from a uniform probability distribution on $[0, 1]$
  - update this distribution for each observation
  - (these updates are different for head and for tail)

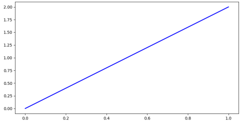
Finding the bias of a coin, II

The probability density functions (pdf's) of the resulting distributions are:



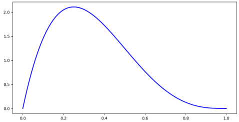
Initially

$\mathrm{Beta}(1,1)$



After H = head

$\mathrm{Beta}(2,1)$



After H-T-T-T

$\mathrm{Beta}(2,4)$

- As is well-known, one does not have to re-compute the distributions each time
- It suffices to *re-compute the parameters* α, β in $\mathrm{Beta}(\alpha, \beta)$
- One says: Beta is conjugate prior to Flip/Bernouilli

What does *Conjugate priorship* mean, precisely??

- The literature is remarkably informal on this topic:
  - sometimes exaplained by example, like for $\mathrm{Beta}/\mathrm{Flip}$ above
  - e.g. in Russell-Norvig's *Artifical Intelligence* book
- Alternatively, informal descriptions are given:
  - Alpaydin'10: "We see that the posterior has the same form as the prior and we call such a prior a conjugate prior"
  - Bishop'06: "...the posterior distribution has the same functional form as the prior."
- Most precise/technical description in Bernardo & Smith'00.

> Our aim: give a mathemtically precise account of conjugate priorship

The account relies on channels and their inversion (via disintegration)

General picture

$$
\begin{tikzcd}
\boxed{\text{(hyper) parameters}} \arrow[r, "\mathrm{Beta}"] & \boxed{\text{probability space}} \arrow[r, "\mathrm{Flip}"] & \boxed{\text{Observations}}
\end{tikzcd}
$$

$$\alpha, \beta > 0 \qquad\qquad [0,1] \qquad\qquad \{H, T\}$$

- Now, Beta and Flip are both channels — technically, Kleisli maps
- Conjugate priorship involves parameter translation function:

$$
\begin{array}{rcl}
\text{Parameters} \times \text{Observations} & \longrightarrow & \text{Parameters} \\
(\alpha, \beta, H) & \longmapsto & (\alpha + 1, \beta) \\
(\alpha, \beta, T) & \longmapsto & (\alpha, \beta + 1)
\end{array}
$$

Question: which equations should hold? **Answer** involves “inversion”

Where we are, so far

Disintegration

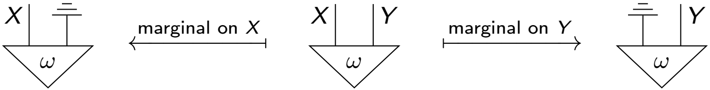
- Informally, disintegration involves turning a joint probability into a conditional probability
  - going from $P(A, B)$ to $P(A \mid B)$
  - i.e. turning a joint state on $A \times B$ into a channel $B \to A$
- It is fundamental for turning a (big) joint probability distribution into a Bayesian network
  - the graph's edges are channels (Kleisli maps)
  - useful perspective, following Brendan Fong, Fabio Zanasi, BJ
- Here it will be presented graphically (Kenta Cho & BJ)
  - main models: Kleisli categories of distribution $\mathcal{D}$ / Giry $\mathcal{G}$ monad
  - copying is allowed in a probabilistic (non-quantum) setting
  - inversion is then (best) explained as special case of disintegration
- Existence of disintegration is a separate topic — ignored here
  - easy for $\mathcal{D}$, non-trivial for $\mathcal{G}$

Graphical language: channels as boxes (flow is upwards)

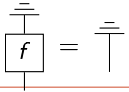
For symmetric monoidal categories with discarding (tensor unit is final):

- Sequential and parallel composition:

- States $1 \to X$ are triangles that can be marginalised via discarding ⏚

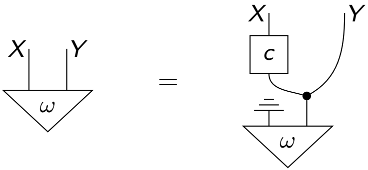


- By finality, channels are causal (or unital):



Disintegration: extraction of channel

- Assume a joint state (distribution) ω on X, Y as depicted below
- A disintegration of ω is a channel $c\colon Y \to X$ such that:



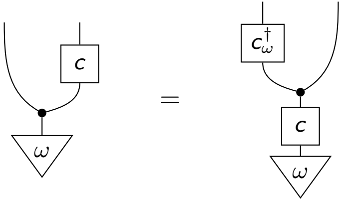
- Equationally, $\omega(x, y) = \omega(x \mid y) \cdot \omega(y)$
- Disintegration is a fundamental concept, esp. in conditional probability theory
  - e.g. to define conditional independence abstractly

Bayesian inversion via disintegration

Bayesian inversion (Clerc et al 2017) turns a state and a channel into an inverted channel, written $c_\omega^\dagger$ in:

$$\begin{tikzcd} 1 \arrow[r, "\omega"] & X \arrow[r, "c"] & Y \arrow[l, dotted, bend left, "c_\omega^\dagger"] \end{tikzcd}$$

Graphically, disintegration is applied to the diagram on the left, giving the defining equation for Bayesian inversion $c_\omega^\dagger$ in:



Where we are, so far

Main ideas behind abstract description

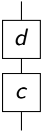
- The informal descriptions of 'conjugate priorship' speak about classes of distributions which are suitably closed
- Such a class will form a channel $c\colon P \to X$
  - P is the object of parameters
  - informally, for each $p \in P$ we have distribution $c(p)$ on X
  - recall, we think of Kleisli maps $P \to \mathcal{D}(X)$ or $P \to \mathcal{G}(X)$
- Observations happen via another channel $d\colon X \to O$
  - O is the object of observations
  - now we can look at inversion of d for each state $c(p)$ on X
  - we will seek a function $h\colon P \times O \to P$ to do so

Definition of conjugate prior

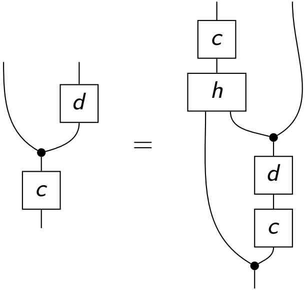
Setting: a pair of composable channels:

$$P \xrightarrow{c} X \xrightarrow{d} O$$

or, as diagram,



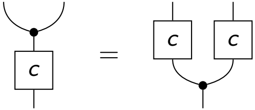
Definition: Channel c a conjugate prior of d if there is a (deterministic) channel $h\colon P \times O \to P$ such that:

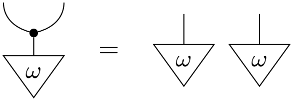


Intuition:
$c \circ h$ forms
inversion of d

Intermezzo on ‘deterministic’ channels

- In general, channels do *not* commute with copying
- If it does commute, then the channel is called deterministic as in:





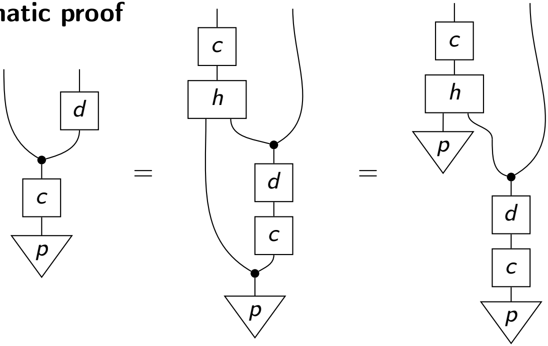
- For a state ω this amounts to the equation on the right
  - the state is then also called copyable
- Measurable functions form deterministic channels in $\mathcal{K}\ell(\mathcal{G})$
  - point (Dirac) states are deterministic/copyable

Conjugate priors involve Bayesian inversion

Theorem

Given $P \xrightarrow{c} X \xrightarrow{d} O$, where c is conjugate prior to d via $h\colon P \times O \to P$. Then for each copyable state p, the map $c \circ h(p, -)\colon O \to X$ is a Bayesian inversion of d.

Diagrammatic proof



Down-to-earth: what does this mean in practice?

- Suppose channels $c\colon P \to X$ and $d\colon X \to O$ are given by likelihoods
  - $c = \int u$ and $d = \int v$, for $u\colon P \times X \to \mathbb{R}_{\geq 0}$, $v\colon X \times O \to \mathbb{R}_{\geq 0}$
  - thus $c(p)(M) = \int_M u(p,x)dx$ and d(x)(N) $= \int_N v(x,y)dy$
- If c is conjugate prior to d, then the defining equation amounts to:

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

- This is essentially Defn. 5.6 of Bernardo & Smith, *Bayesian Theory*, 2000
- This equation holds in the well-known examples of conjugate priorship (that I checked)
  - e.g. $\mathrm{Beta} - \mathrm{Flip}$, or $\mathrm{Beta} - \mathrm{Binom}$, or $\mathrm{Norm} - \mathrm{Norm}$

Where we are, so far

Concluding remarks

- Conjugate priorship is an frequently used fundamental concept
  - that is often introduced only informally — e.g. via examples
- A definition is given here, that is both precise and abstract
  - formulated via channels, with a non-trivial definining equation
- The (expected) relationship with Bayesian inversion holds via a simple (diagrammatic) proof
- Details in arXiv:1709.00322
- This (hopefully) demonstrates that categorical/graphical abstraction can indeed contribute to probability theory
  - but my colleagues in machine learning could not read the report ☹; there is still work to do
  - maybe you can!