Pearl's and Jeffrey's Update as Modes of Learning in Probabilistic Programming

Bart Jacobs, **Dario Stein**

Radboud University Nijmegen

MFPS 2023
Bloomington, Indiana – 23 June 2023

Statistics Example

Covid tests have a known sensitivity/specificity.

- A patient takes three tests: two are positive, one negative
- What's the probability they have covid?
- How does this relate to taking a single test with an uncertain outcome: 66% positive, 33% negative?
- How to interpret uncertain evidence in the first place?

Overview

Recap: Modes of learning from uncertain evidence [Jacobs'21]

- Pearl's update
- Jeffrey's update

New Insights:

- Sampling interpretation: Probabilistic Programming & Nested Normalization
- Learning from datasets: Mixture Modelling & Variational Inference

Pearl's & Jeffrey's update

We need finite distributions

$$\mathcal{D}(X) = \left\{ \sum_{i=1}^{n} p_i |x_i\rangle \, : \, p_i \in [0,1], \sum_i p_i = 1 \right\}$$

Basic setting for learning

1. beliefs $\omega \in \mathcal{D}(X)$
2. prediction channel $c : X \to \mathcal{D}(Y)$
3. uncertain evidence $\tau \in \mathcal{D}(Y)$ $\leftarrow$ e.g. noisy measurement

Question: How to update ω given τ?

Calculus of distributions and predicates

Allowing possibly unnormalized distributions: $[0, \infty)^{X \times Y}$

1. states $\omega : I \to X$
2. predicates $p : X \to I, q : Y \to I$
   - every state gives rise to a predicate $\widehat{\omega}(x) = \omega(x)$.
3. pushforward/pullback: for $c : X \to Y$

$$c_*\omega = c \circ \omega : I \to Y, \qquad c^*q = q \circ c : X \to I$$

4. Validity pairing

$$(\omega \models p) = \sum_x \omega(x) \cdot p(x) \quad (\omega \models \widehat{\tau}) = \langle \omega, \tau \rangle \quad \leftarrow L^2\text{-inner product}$$

5. conditioning (Bayes' rule)

$$\omega|_p = \frac{\omega \cdot p}{\omega \models p}, \quad \text{i.e. } \omega|_p(x) = \frac{\omega(x)p(x)}{\sum_x \omega(x)p(x)}$$

Pearl's & Jeffrey's update

Recall Bayesian inversion $c^{\dagger}_{\omega} : Y \to D(X)$ given by $c^{\dagger}_{\omega}(y) = \omega|_{c^{*}1_{y}}$.

Def: Pearl's update

$$\omega_{\text{Pearl}} = \omega|_{c^{*}\widehat{\tau}}$$

Def: Jeffrey's update

$$\omega_{\text{Jeffreys}} = c^{\dagger}_{\omega} \circ \tau$$

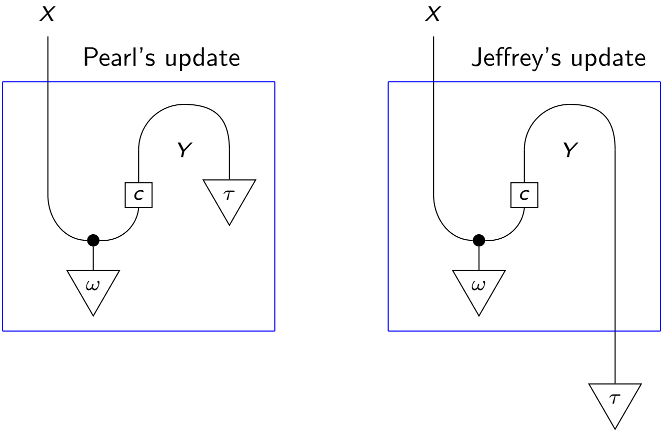
Sharp evidence = Bayesian inversion

For sharp evidence $\tau = |y_0\rangle$, both updates equal Bayesian inversion

$$\omega_{\text{Pearl}} = \omega_{\text{Jeffrey}} = c^{\dagger}_{\omega}(y_0)$$

Pearl's & Jeffrey's update

Key difference: When to normalize?



In probabilistic programs

Pearl's

```
y = τ()
x = ω()
condition(c(x) == y)
return x
```

Jeffrey's using Nested Normalization

```
y = τ()
return normalize(
  x = ω()
  condition(c(x) == y)
  return x
)
```

Rejection samplers

```
# Pearl's update
while True:
  x = ω()
  y = τ() # new target in every iteration
  if c(x) == y:
        yield x
```

```
# Jeffreys's update
while True:
        x = ω()
        if c(x) == y:
                y = τ() # new target after accept
                yield x
```

Pearl's & Jeffrey's update

Pearl's update

1. Symmetric: belief is revised $\leftrightarrow$ evidence is distrusted
2. Repeated updates commute
3. Learns nothing if $\tau =$ uniform.

Jeffrey's update

1. Asymmetric: Belief changes, evidence does not
   - take a random sample $y \sim \tau$ of the evidence, treat it as certain
2. Repeated updates do not commute
3. Learns nothing if $\tau = c\omega$

Learning from what's right and wrong

From [Jacobs'21]:

Pearl's update [easy]

Pearl's update **increases validity** of the model under the evidence

$$\langle \tau, c \circ \omega \rangle \leq \langle \tau, c \circ \omega_{\text{Pearl}} \rangle$$

Jeffrey's update [surprisingly tricky!]

Jeffrey's update **reduces divergence** of evidence and prediction

$$D(\tau \,\|\, c \circ \omega) \geq D(\tau \,\|\, c \circ \omega_{\text{Jeffreys}})$$

where

$$D(\tau \,\|\, \sigma) = \sum_x \tau(x) \log \left(\frac{\tau(x)}{\sigma(x)} \right)$$

Datasets

Statistical datasets naturally organize as multisets (unordered lists)

$$\mathcal{M}[n](X) = \left\{ \sum_i n_i | x_i \rangle : n_i \in \mathbb{N}, \sum_i n_i = n \right\}$$

We have a natural transformations

$$acc : X^n \to \mathcal{M}[n](X), (x_1, \ldots, x_n) \mapsto |x_1\rangle + \ldots + |x_n\rangle$$
$$flrn : \mathcal{M}[n](X) \to \mathcal{D}(X), \varphi \mapsto \frac{\varphi}{N}$$
$$mn[n] : \mathcal{D}(X) \to \mathcal{D}(\mathcal{M}[n](X)), \omega \mapsto \mathcal{D}(acc)(\omega^{\otimes n})$$

Multiset lifting

$\mathcal{M}[n]$ extends to a functor $\mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$.

Pearl's & Jeffrey's update

Back to the Covid example: [all numbers purely hypothetical]

- $X = \{true, false\}$ $\leftarrow$ covid or not
- $Y = \{pos, neg\}$ $\leftarrow$ test result
- $\omega = 0.05|\,true\,\rangle + 0.95|\,false\,\rangle$ $\leftarrow$ base rate
- $c : X \to \mathcal{D}(Y)$, 10% false negatives, 5% false positives
- $\psi = 2|\,pos\,\rangle + 1|\,neg\,\rangle \in \mathcal{M}[3](Y)$ $\leftarrow$ observations

Possible inferences

- 3 x Bayes rule $\Rightarrow$ 64%
- 1 x Pearl's update with $\tau = \mathit{flrn}(\psi)$ $\Rightarrow$ 9%
- 1 x Jeffreys's update with $\tau = \mathit{flrn}(\psi)$ $\Rightarrow$ 33%

How to interpret the results? What are the underlying generative models?

Generative models & Likelihoods

Pearl style mixture model

We use the same latent value x for all datapoints (single patient)

$$x \sim \omega, \quad y_i \sim c(x) \text{ iid.}$$

That is we take the mixture

$$\Phi_{\text{Pearl}} = \sum_x \omega(x) \cdot mn[n](c(x))$$

Jeffrey style multinomial model

All datapoints $\{y_i\}$ are independently sampled (population of patients)

$$x_i \sim \omega \text{ iid.}, \quad y_i \sim c(x_i)$$

hence

$$\Phi_{\text{Jeffrey}} = mn[n](c \circ \omega)$$

Increasing likelihoods via updates

Let a dataset $\psi = acc(y_1, \ldots, y_n) \in \mathcal{M}[n](X)$ be given.

Pearl's likelihood

Repeated application of Bayes's rule

$$\begin{aligned} \omega \mapsto \omega\big|_{mn[n](c)^* 1_\psi} \\ = \omega\big|_{c^* 1_{y_1}}\big|_{c^* 1_{y_2}} \cdots \big|_{c^* 1_{y_n}} \end{aligned}$$

increase the likelihood $\Phi_{\text{Pearl}}(\psi)$.

Jeffrey's likelihood

The likelihood of ψ under the multinomial model is inversely related to the divergence

$$D(flrn(\psi) \,\|\, c\omega)$$

Jeffrey's update $\omega \mapsto c^\dagger_\omega \circ flrn(\psi)$ increases the multinomial likelihood $\Phi_{\text{Jeffrey}}(\psi)$.

Variational Inference for Multinomial models

New Perspective

Jeffrey's update is variational approximation to Bayesian inversion on multisets, under an independence assumption.

Let a population be modelled by $\Phi = mn[n](\omega)$ for some $\omega \in \mathcal{D}(X)$. Each member performs a test c, and we observe outcomes $\psi \in \mathcal{M}[n](Y)$ sharply. The Bayesian inverse is

$$\Phi' = \mathcal{M}[n](c)^{\dagger}_{\Phi}(\psi) \in \mathcal{D}(\mathcal{M}[n](X))$$

- Φ' is no longer a multinomial distribution!
- The best approximation is given by Jeffrey's update

Theorem

$$\arg\min_{\omega'} D(mn[n](\omega') \,\|\, \Phi') = c^{\dagger}_{\omega} \circ \mathit{flrn}(\psi)$$

Summary

We saw

- Difference between Jeffrey's and Pearl's update is subtle
  - Rejection samplers differ in 1 line
- Different modelling assumptions:
  - Pearl's: repeated information about a single individual
  - Jeffrey's: population level / independence assumptions
  - both updates increase a model likelihood
- Jeffrey's update $\leftrightarrow$ nested normalization in PPL
- Variational principle

Outlook: Further Connections with active inference / predictive coding

- Free-energy principle
- Operational differences (sampling, particle filters) $\leftarrow$ Jeffrey needs only a single sample of τ

Thank you!