Bayesian models of learning

$P(H | D)$
Bayes’ Theorem

Bayes’ theorem is most commonly used to estimate the state of a hidden, causal variable $H$ based on the measured state of an observable variable $D$:

$$p(H | D) = \frac{p(D | H)p(H)}{p(D)}$$

- **Likelihood**
- **Prior**
- **Posterior**
- **Evidence**
Bayesian models

- Bayesian models tell you how you should integrate your prior beliefs with new evidence.
  - \( P(H \mid D) \propto P(D \mid H) \times P(H) \)
    - \( P(H) \) = your prior beliefs
    - \( P(D \mid H) \) = how likely the data is under a hypothesis (or theory)

**NOTE:** \( \propto \) means “proportional to”
Bayesian models

- Bayesian models tell you how you should integrate your prior beliefs with new evidence.
  - $P(H | D) \propto P(D | H) \times P(H)$
    - $P(H)$ = your prior beliefs
    - $P(D | H)$ = how likely the data is under a hypothesis (or theory)
  - “Just” compute $P(H | D)$ for all possible hypotheses (H).

*NOTE: $\propto$ means “proportional to”*
Bayes’ rule in odds form

\[
\frac{P(H_1 | D)}{P(H_2 | D)} = \frac{P(D | H_1)}{P(D | H_2)} \times \frac{P(H_1)}{P(H_2)}
\]

D: data

H_i: hypothesis/model

P(H_i | D): posterior probability H_i generated the data

P(D | H_i): likelihood of data under model H_i

P(H_i): prior probability H_i generated the data
Model selection

- You have a coin that either comes up heads 90% of the time or comes up heads 50% of the time (you don’t know!)
- How do you figure it out?
Model selection

- You have a coin that *either* comes up heads 90% of the time *or* comes up heads 50% of the time (you don’t know!)
- How do you figure it out?
  - Flip it!
Model selection

Suppose you flip the coin 10 times and each time it comes up heads.

- What is the probability that it’s the fair coin?
- The probability that it’s the 90%-heads coin?
Model selection

- Suppose you flip the coin 10 times and each time it comes up heads.
  - What is the probability that it’s the fair coin?
  - The probability that it’s the 90%-heads coin?

- Compute the odds:
  \[
  \frac{P(p=90\% \mid 10 \text{ heads})}{P(p=50\% \mid 10 \text{ heads})} = \frac{P(10 \text{ heads} \mid p=90\%)}{P(10 \text{ heads} \mid p=50\%)} \times \frac{P(p=90\%)}{P(p=50\%)}
  \]

  **POSTERIOR ODDS**  **LIKELIHOOD RATIO**  **PRIOR RATIO**
Model selection

\[ P(p=90\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=90\%) \times P(p=90\%) \]

\[ P(p=50\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=50\%) \times P(p=50\%) \]
Model selection

\[ P(p=90\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=90\%) \times P(p=90\%) \]

\[ P(p=50\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=50\%) \times P(p=50\%) \]

\[ \frac{P(p=90\%)}{P(p=50\%)} = ? \]
Model selection

\[ P(p=90\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=90\%) \times P(p=90\%) \]

\[ P(p=50\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=50\%) \times P(p=50\%) \]

\[ \frac{P(p=90\%)}{P(p=50\%)} = 1 \]
Model selection

\[
P(p=90\% \mid 10 \text{ heads}) = \frac{P(10 \text{ heads} \mid p=90\%) \times P(p=90\%)}{P(p=50\% \mid 10 \text{ heads}) = \frac{P(10 \text{ heads} \mid p=50\%) \times P(p=50\%)}{\quad}}
\]

\[
\frac{P(p=90\%)}{P(p=50\%)} = 1 \quad \frac{P(10 \text{ heads} \mid p=90\%)}{P(10 \text{ heads} \mid p=50\%)} = ?
\]
Model selection

\[
\frac{P(p=90\% \mid 10 \text{ heads})}{P(p=50\% \mid 10 \text{ heads})} = \frac{P(10 \text{ heads} \mid p=90\%) \times P(p=90\%)}{P(10 \text{ heads} \mid p=50\%) \times P(p=50\%)}
\]

\[
\frac{P(p=90\%)}{P(p=50\%)} = 1
\]

\[
\frac{P(10 \text{ heads} \mid p=90\%)}{P(10 \text{ heads} \mid p=50\%)} = \frac{0.9^{10}}{0.5^{10}}
\]
Model selection

\[
P(\text{p=90\% } | \text{ 10 heads}) = P(\text{10 heads } | \text{ p=90\%}) \times P(\text{p=90\%})
\]

\[
P(\text{p=50\% } | \text{ 10 heads}) = P(\text{10 heads } | \text{ p=50\%}) \times P(\text{p=50\%})
\]

\[
\frac{P(\text{p=90\% } | \text{ 10 heads})}{P(\text{p=50\% } | \text{ 10 heads})} = \frac{0.9^{10}}{0.5^{10}} \times 1
\]

90\%-H coin is 357 times more likely than the 50\%-H coin.
Model selection

\[ P(p=90\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=90\%) \times P(p=90\%) \]
\[ P(p=50\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=50\%) \times P(p=50\%) \]

\[ \frac{P(p=90\% \mid 10 \text{ heads})}{P(p=50\% \mid 10 \text{ heads})} = \frac{357}{1} \]

90\%-H coin is 357 times more likely than the 50\%-H coin.
Odds to probability

90%-H coin is 357 times more likely than the 50%-H coin.

⇒ $P(p=90\% \mid 10\, \text{heads}) = \frac{357}{358}$

⇒ $P(p=50\% \mid 10\, \text{heads}) = \frac{1}{358}$
Model selection

\[ P(p=90\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=90\%) \times P(p=90\%) \]
\[ P(p=50\% \mid 10 \text{ heads}) = P(10 \text{ heads} \mid p=50\%) \times P(p=50\%) \]

\[ \frac{P(p=90\% \mid 10 \text{ heads})}{P(p=50\% \mid 10 \text{ heads})} = \frac{357}{0.001} \]

90\%-H coin is 0.357 times more likely than the 50\%-H coin.
Comparing infinitely many hypotheses

- Assume data (coin flips) are generated from a model:

- What is the value of $p$?
  - Each possible value of $p$ is a hypothesis
  - Inference performed over infinite values of $p$
Posterior distribution over \( p \):

\[
\text{Expected value of } p \approx \frac{\#\text{heads}}{\#\text{heads} + \#\text{tails}}
\]
Comparing infinitely many hypotheses

- Flip a coin 10 times and see 5 heads, 5 tails.
  - P(H) on next flip? 50%. Same as expected value.
  - “Future will be like the past.”
Comparing infinitely many hypotheses

- Flip a coin 10 times and see 5 heads, 5 tails.
  - $P(H)$ on next flip? 50%. Same as expected value.
  - Why? $50\% = \frac{5}{5+5} = \frac{5}{10}$.
  - “Future will be like the past.”

- Suppose we had seen 4 heads and 6 tails.
  - $P(H)$ on next flip? Closer to 50% than to 40%.
Origin of prior knowledge

• Tempting answer: prior experience
• Suppose you have previously seen 2000 coin flips: 1000 heads, 1000 tails

• By assuming all coins (and flips) are alike, these observations of other coins are as good as observations of the present coin
Limitations

- Suppose you flip a new coin 25 times and get all heads. *Something funny is going on...*
- But having seen 1000 heads and 1000 tails, $P(H)$ on next flip = $\frac{1025}{1025+1000} = 50.6\%$.

*But no way it’s a fair coin!*
Hierarchical Priors

- Higher-order hypothesis: is this coin fair or unfair?
Constraints guide learning

- Children need to learn concepts at multiple levels of abstraction.
- E.g.
  - Word/category learning:
    - Level 1: Balls are round, teacups have handles, ...
    - Level 2: Things in the same category tend to have the same shape
The image shows three open cardboard boxes, each with a different set of objects. The first box contains orange objects, the second contains blue objects, and the third is labeled "Mystery Box" with a question mark. The objects in the boxes are arranged in a manner that suggests a sense of mystery or surprise.
Constraints guide learning

- \( P(\Box = \bullet \mid D) \)?
Constraints guide learning

- $P(\bigcirc = \bigcirc | D)$?
- $U$: balls tend to be uniform in color
- $P(\bigcirc = \bigcirc | U \text{ is true}; D)$?
Can be formalized as inference:

Level 3: Constraints
Level 2: Category means
Level 1: Data

Kemp, Perfors & Tenenbaum (2008);
The acquisition of inductive constraints;
*Developmental Science*
Can be formalized as inference:

\[ p(\theta^i|y) = \int_{\alpha, \beta} p(\theta^i|\alpha, \beta, y)p(\alpha, \beta|y) d\alpha d\beta. \]

Kemp, Perfors & Tenenbaum (2008): The acquisition of inductive constraints; Developmental Science
Constraints guide learning

- $H$: balls tend to be uniform in color
- $P(\square = \bigcirc | H \text{ is true})$?
- **Key point:** Learning abstract constraints allows people to make reliable inferences and learn more quickly.
So what?

- This provides a framework for studying human cognition.
- Assume humans rationally integrate evidence with their prior beliefs.
So what?

- This provides a framework for studying human cognition.
- Assume humans rationally integrate evidence with their prior beliefs. Can ask:
  - What evidence are people using to update their prior beliefs?
  - Formalizes "nature versus nurture": how much prior information do newborns have about _____?
  - Can also ask: how do people deviate from ideal?
Subjective randomness

- Consider two possible sequences of coin flips:
  - HHHHHHHHHHHHHHHHHHHHHH
  - HHTHTTTHTTHHHHTHHTTTHTTHH
- They are *equally probable from a fair coin*, but one looks a lot more random than the other. Why?
Subjective randomness ratings

- Falk & Konold (1997)
- DP model
- Finite state model

Subjective randomness vs. Number of alternations
<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>function <code>complexSequence()</code></td>
<td>A longer program is required to produce sequences of high complexity.</td>
</tr>
<tr>
<td>print(H)</td>
<td></td>
</tr>
<tr>
<td>print(T)</td>
<td></td>
</tr>
<tr>
<td>print(T)</td>
<td></td>
</tr>
<tr>
<td>print(H)</td>
<td></td>
</tr>
<tr>
<td>print(H)</td>
<td></td>
</tr>
<tr>
<td>print(H)</td>
<td></td>
</tr>
<tr>
<td>print(T)</td>
<td></td>
</tr>
<tr>
<td>function <code>notSoComplexSequence()</code></td>
<td>A shorter program can produce sequences of low complexity.</td>
</tr>
<tr>
<td>for i = 1..8:</td>
<td></td>
</tr>
<tr>
<td>print(H)</td>
<td></td>
</tr>
</tbody>
</table>

Griffiths et al. (2018); Subjective randomness as statistical inference; Cognitive Psychology
Kolmogorov complexity

- \( K(x) = \) shortest program that generates \( x \)
  - High \( K(x) \): HHTHTTTHTTHHHHHTHHTTTHTHH
  - Low \( K(x) \): HHHHHHHHHHHHHHHHHHHHHHH
Kolmogorov complexity

- $K(x) =$ shortest program that generates $x$
  - High $K(x)$: HHTHTTHTHHHHTHTTTHTHH
  - Low $K(x)$: HHHHHHHHHHHHHHHHHHHHHHHHHHHH
- Longer programs (high $K(x)$) are less likely a priori
  - “Bayesian Occam’s razor”
A longer program is required to produce sequences of high complexity.

```plaintext
function complexSequence():
    print(H)
    print(T)
    print(T)
    print(H)
    print(H)
    print(H)
    print(T)
    print(T)
```

A shorter program can produce sequences of low complexity.

```plaintext
function notSoComplexSequence():
    for i = 1..8:
        print(H)
```
Subjective randomness as inference

- **Posterior odds**: \( p(\text{random} \mid x) / p(\text{regular} \mid x) \)

- **Likelihood**: 
  - \( p(x \mid \text{random}) = 2^{-\text{LENGTH}(X)} \)
  - \( p(x \mid \text{regular}) = 2^{-\text{KOLMOGOROV\_CPLX}(X)} \)

- **Prior**: 
  - \( p(\text{random}) = 0.5 \)
  - \( p(\text{regular}) = 0.5 \)