Bayesian Belief Revision and Language Acquisition

Thomas Bayes (1734) offered a way to interpret probabilities as hypotheses (beliefs).

It turned out to be a theorem in Probability Theory. It relates conditional probabilities to prior probabilities (via joint probabilities).
\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \quad P(B \mid A) = \frac{P(A \cap B)}{P(A)}
\]

\[
P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)
\]

\[
P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad \text{(Bayes’ Rule)}
\]
If beliefs are considered statistical hypotheses $H$, we can rephrase Bayes’ theorem as belief revision:

$$P(H_0 | X) = \frac{P(X|H_0)P(H_0)}{P(X)}$$

$H_0$: hypothesis

$P(H_0 | X)$: posterior probability

$P(H_0)$: prior probability

$P(X | H_0)$: likelihood

$P(X)$: marginal probability of $X$ (summed over all mutually exclusive hypotheses)
Cookie bowl

There are two bowls full of cookies.

Bowl #1 has 10 chocolate chip and 30 plain cookies.

Bowl #2 has 20 of each.

Fred picks a bowl at random, and then picks a cookie at random.

We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies.

The cookie turns out to be a plain one. How probable is it that Fred picked it out of bowl #1?
\( H_1 \) : bowl #1 \hspace{1cm} \( H_2 \) : bowl #2

Bowls are identical: \( P(H_1) = P(H_2) \) and \( P(H_1) + (PH_2) = 1 \)

\( P(C_p) \): marginal probability of plain cookies.

\[
P(H_1 \mid C_p) = \frac{P(C_p \mid H_1)P(H_1)}{P(C_p)} = \frac{31}{\frac{13}{24} + \frac{11}{22}} = 0.6
\]

Before observing \( C_p \), the probability of picking bowl #1 is \( P(H_1) = 0.5 \)

After observing \( C_p \), it is \( P(H_1 \mid C_p) = 0.6 \)
Example 2: Bayesian justice (not recommended—banned in UK)

$G$: the event that the defendant is guilty.

$E$: the event that the defendant’s DNA matches DNA found at the crime scene.

$P(E | G)$: the probability of seeing event $E$ assuming that the defendant is guilty. (Usually this would be taken to be unity.)

$P(G | E)$: the probability that the defendant is guilty assuming the DNA match event $E$

$P(G)$: the juror’s personal estimate of the probability that the defendant is guilty, based on the evidence other than the DNA match.

\[
P(G | E) = \frac{P(E|G)P(G)}{P(E)}\
\]
Something must bias priors and likelihoods,

if word learning is not random,

speedy at early ages, settling down later.

One conjecture: It’s the Universal Grammar.
Bayesian Word Learning (Tenenbaum and Xu, 2002)

Three factors in learning nouns (Markman 1989):

**Taxonomic assumption:** new words refer to taxonomic classes.

**Mutual exclusivity:** learner assumes there is one word that applies to each object.

**Basic level map:** new words map to basic levels in a taxonomy (these levels maximize category utility).
These are complementary features, all tackling the problem of induction from one aspect.

As hard-coded constraints, they would allow learning a class of words, but make the rest of language unlearnable.

Bayesian approach formulates them as probabilistic constraints.
Experiment

Phase 1: **word learning:** subjects are given examples of words in a novel language and asked to choose the other instances that each word applies to (from a large set of test objects)

Phase 2: **similarity judgment:** judge the similarity of pairs of objects used in Phase 1.

Procedure: Subjects were told they were helping a puppet who speaks a different language to pick out the objects he needs. (e.g., pick out other **blicks** in the set, when a dog is shown on screen)

The results are submitted to a clustering algorithm to reconstruct the taxonomic hypothesis space.
Basic-level bias in generalizing from a single example.

All or nothing behaviour in generalizing from 3 examples (most specific generalization)
Bayesian Model

hypothesis space: concepts $H$ probabilistically related to data $X$

$X$: $n$ observed examples of a novel word $C$.

Each $h \in H$ is a pointer to some subset of objects in the world that is a candidate extension for $C$.

Priors: pre-existing knowledge (e.g. taxonomic, basic-level map)

Likelihoods: statistical patterns in the data.

$P(h \mid X) \approx p(X \mid h) p(h)$
Priors and likelihoods are based on concept cluster: a node $h$ measures the dissimilarity of objects within $h$.

More distinctive clusters are a priori more likely to have distinguishing names.

$$p(h) \approx \text{height}(\text{parent}(h)) - \text{height}(h)$$

Smaller hypotheses assign greater likelihood than large hypotheses

$$p(X \mid h) \approx \left(\frac{1}{\text{height}(h) + \varepsilon}\right)^n$$
Dalmatian as first example of “blick” assigns $14.08/3.50 \approx 4$ in favor of dalmatians to dogs

Dalmatians as three examples assign $(14.08/3.50)^3 \approx 65$ in favor of dalmatians to dogs

**Generalization:** The learner must use $p(h \mid X)$ to decide how to generalize the word $C$ to unlabeled objects:

$$p(y \in C \mid X) = \sum_{h \in H} p(y \in C \mid h) p(h \mid X)$$

If $y \in h$, then $p(y \in C \mid H) = 1$
The model captures the qualitative features of the data:

- similarity-like gradient in generalizing from a single example
- all-or-none behaviour at the most specific level with 3 examples

Only Bayesian models capture the transition from single examples to multiple examples in this task.
Pilot data on children (4 year olds)

Same task (their superordinate generalizations were less consistent)
Children’s response seems to be Bayesian *without* the basic-level bias (do not favor basic-level matches in sets of 3)

This seems to be consistent with language acquisition research that basic-level matching occurs later as children have more linguistic experience.
Probabilistic inference and UG

Something must bias priors and likelihoods,

if word learning is not random,

speedy at early ages, settling down later.

One conjecture: It’s the Universal Grammar.
A Phrase Structure Grammar

(1) \[ S \rightarrow NP \ VP \]
\[ VP \rightarrow V_{iv} \]
\[ VP \rightarrow V_{tv} NP \]
\[ V_{iv} \rightarrow slept \quad e \mapsto t \]
\[ V_{tv} \rightarrow hit \quad e \mapsto (e \mapsto t) \]
A fully lexicalized grammar (CCG)

\[
(2) \quad (S\backslash NP) =: \text{slept} \quad e \mapsto t \\
(S\backslash NP)/NP =: \text{hit} \quad e \mapsto (e \mapsto t)
\]