Probabilistic Context-Free Grammars and Bayesian Inference

Bonan Zhao

1 Overview

Probabilistic Context-Free Grammars (PCFGs) provide a principled way to define probability distributions over structured objects such as logical expressions, programs, or natural language sentences. In cognitive science and machine learning, PCFGs are often used to model a structured hypothesis space, where each expression generated by the PCFG serves as a candidate hypothesis.

This handout introduces:

1. The definition of a PCFG.
2. How PCFGs induce prior probabilities over hypotheses.
3. Bayesian updating using observed data.
4. A causal-program interpretation in which each expression corresponds to a causal hypothesis.
5. Posterior predictive inference obtained by marginalizing over hypotheses.

2 Context-Free Grammars

A context-free grammar (CFG) consists of:

$$G = (N, \Sigma, R, S)$$

where

- N is a set of nonterminal symbols,
- Σ is a set of terminal symbols,
- R is a set of production rules,
- $S \in N$ is the start symbol.

Activity 1.

$$S \to \texttt{(+ E E)}$$

$$E \to x$$

$$E \to 1$$

$$E \to 2$$

What does this CFG generate?

3 Probabilistic Context-Free Grammars

As the name suggests, A PCFG augments each production rule in a CFG with a probability distribution: Suppose a nonterminal A has productions

$$A \to \beta_1, \ldots, \beta_k.$$

Each rule receives a probability

$$P(A \to \beta_i)$$

such that

$$\sum_{i=1}^{k} P(A \to \beta_i) = 1.$$

Activity 2.

$$E \to x \qquad (\quad)$$

$$E \to 1 \qquad (\quad)$$

$$E \to 2 \qquad (\quad)$$

Fill the ( ) with numbers such that they define a probability distribution over productions of E (assume E only has these three production rules).

4 Generation Probabilities

Starting from the start symbol S, a PCFG terminates when there are no nonterminal symbols anymore. We call this resulting expression a complete derivation. In Bayesian terms, a complete derivation corresponds to a hypothesis h.

Assuming production rule choices are independent (i.e, context-free), the probability of a hypothesis h is the product of the probabilities of all the production rules used during generation: Let $r_1, r_2, \ldots, r_m$ be the production rules used to generate hypothesis h, then

$$P(h) = \prod_{j=1}^{m} P(r_j).$$

Activity 3. *Calculate the probability of generating $(+\ x\ 1)$ with the PCFG defined in Activities 1-2.*

Activity 4. *Circle the correct description:*

Shorter or simpler derivations often receive higher / lower *probability, because they can be generated using* more / fewer *rule applications.*

Nb. If multiple derivation trees yield the same expression h, then

$$p(h) = \sum_{t:\text{yield}(t)=h} p(t).$$

In many simple grammars, we design the grammar to be approximately unambiguous, so we often write

$$p(h) \approx p(t_h).$$

5 Recursion in PCFGs

A key expressive feature of PCFGs is *recursion*, where a nonterminal symbol can eventually expand into a structure that includes itself. Recursion allows grammars to generate unboundedly complex hypotheses from a finite set of rules.

Formally, a grammar is recursive if there exists a nonterminal A such that:

$$A \Rightarrow^{+} \alpha A \beta$$

It means that starting from A, through one or more derivation steps ($\Rightarrow^{+}$), you can produce a string that contains A again—possibly with some stuff (α) before it and some stuff (β) after it.

Activity 5. *Add a production rule in our example PCFG and make it recursive.*

$$E \rightarrow$$

Activity 6. *Optional: Write a few derivations under this new PCFG and calculate the probability of generating each derivation.*

Activity 7. *Optional: Design a PCFG for the Magic Egg task.*

6 Likelihood Functions

Following the Magic Egg example, where a stands for the causal agent, r the recipient, and r' the result object, a complete causal interaction is recorded in data $d = (a_i, r_i, r'_i)$. For a deterministic hypothesis h, interpreted as an executable function

$$f_h : (a, r) \mapsto r',$$

a simple likelihood assigns probability one to exact consistency and zero otherwise:

$$P(d \mid h) = \mathbf{1}\left[f_h(a_i, r_i) = r'_i\right],$$

where $\mathbf{1}[\cdot]$ is an indicator function.

This formulation encodes a strict notion of correctness: a hypothesis is either fully consistent with the data or ruled out entirely.

Alternatively, considering stochasticity in the observation process, Goodman et al., 2008 introduced a "soft" interpretation of hypothesis evaluation: Instead of requiring exact equality, we assign higher probability to hypotheses that are closer to the observed outcomes:

$$P(d \mid h) = \frac{1}{Z} \exp\left(-\beta \, \mathrm{dist}\big(f_h(a_i, r_i), r'_i\big)\right),$$

where:

- $\mathrm{dist}(\cdot, \cdot)$ is a loss or discrepancy function (e.g., Hamming distance or squared error),

- $\beta \geq 0$ controls how strongly deviations are penalized,
- Z is a normalization constant ensuring a valid probability distribution.

In the limit $\beta \to \infty$, the soft likelihood reduces to the deterministic case, recovering the hard constraint formulation.

This soft likelihood allows partially correct or approximately consistent hypotheses to retain non-zero posterior probability, supporting graded inference over hypotheses.

7 Bayesian Inference

In the Magic Egg task, a hypothesis h is an executable function

$$f_h(a, r).$$

Executing the program produces a prediction

$$r' = f_h(a, r).$$

The hypothesis therefore encodes a causal theory of the environment:

$$(a, r) \longmapsto r'.$$

Activity 8. *Fill in the missing information:*

In the single-shot scenario, let h denote a hypothesis and d denote an observed data (causal interaction (a_i, r_i, r'_i)), Bayes' rule gives the posterior probability of a hypothesis:

$$\begin{aligned} P(h \mid d) &= \frac{P(d \mid h)P(h)}{P(d)} \\ &= \frac{\qquad\qquad P(d \mid h)P(h)}{\qquad\qquad\qquad\qquad}. \end{aligned}$$

Usually shortened as

$$P(h \mid d) \propto P(d \mid h)P(h).$$

8 Marginalized Posterior Prediction

In the generalization task, given a partial observation of a new agent and a new recipient

$$(a, r),$$

to make predictions on the result, we generally do not commit to a single hypothesis, but average over the posterior distribution.

Activity 9. *Optional: Why averaging over the posterior distribution is better than committing to a single hypothesis?*

Activity 10. *The posterior predictive distribution is*

$$P(r' \mid a, r, D) =$$