

Pearl
How statistical learning can play well with
Universal Grammar (1.2.1)
Theory and predictions for the
development of morphology and syntax: A
Universal Grammar + statistics approach
(2.2)

Melina Noruz

“The Statistics Part”



“The statistics part refers to **statistical learning**...At its core, statistical learning is about **counting things** (this is the “statistical” part), and **updating hypotheses on the basis of those counts** (this is the “learning” part, sometimes also called inference. Counting things is a domain-general ability, because we can count lots of different things, both linguistic and non-linguistic”

“...a child has to know what to count. UG can identify what to count, because UG defines the hypothesis space...the statistical learning mechanism itself doesn’t seem to change – once the child knows the units over which inference is operating, counts of the relevant units are collected and inference can operate”





Table 1: Common inference mechanisms in statistical learning that are used by UG+stats proposals for different morphology and syntax phenomena: basic syntactic categories (**syn cat**), basic word order (**word order**), inflectional morphology (**infl mor**), showing a temporary lack of inflection (**no infl**), movement (**mvmt**), and constraints on utterance form and interpretation (**constr**).

| | syn cat | word order | infl mor | no infl | mvmt | constr |
|------------------------------|----------------|-------------------|-----------------|----------------|-------------|---------------|
| Basic counts & probabilities | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Bayesian inference | ✓ | | ✓ | | ✓ | ✓ |
| Reinforcement learning | | ✓ | | ✓ | ✓ | |
| Tolerance & Sufficiency | | | ✓ | | | |

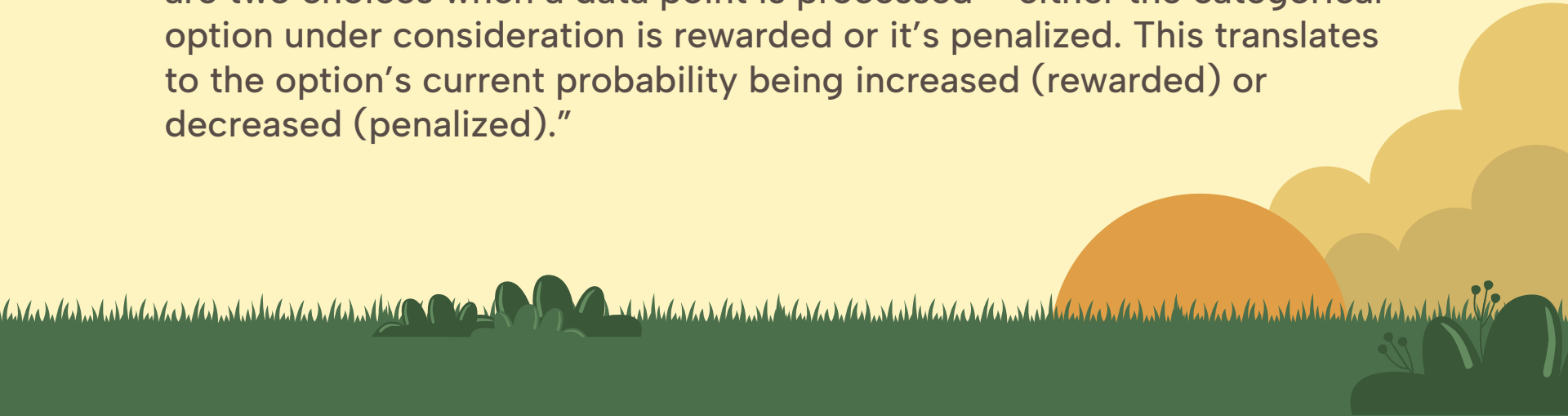


Reinforcement Learning (Operant Conditioning)



“**Reinforcement learning** is a principled way to update the probability of a categorical option which is in competition with other categorical options.”

“A common implementation...is the **linear reward–penalty scheme**...there are two choices when a data point is processed – either the categorical option under consideration is rewarded or it’s penalized. This translates to the option’s current probability being increased (rewarded) or decreased (penalized).”



Tolerance & Sufficiency Principles



“The **Tolerance and Sufficiency Principles** together describe a particular **inference mechanism**, and this mechanism operates over specific kinds of counts that have already been collected...these principles together provide a formal approach for when a child would choose to adopt a “rule”...to account for a set of items.”

“The learning innovation of these principles is that they’re designed for situations where there are exceptions to a potential rule... these two principles help the child infer whether the rule is robust enough to bother with, despite the exceptions...a rule should be bothered with if it speeds up average retrieval time for any item ”

“The **Tolerance Principle** determines how many exceptions a rule can “tolerate” in the data before it’s not worthwhile for the child to have that rule at all; the **Sufficiency Principle** uses that tolerance threshold to determine how many rule-abiding items are “sufficient” in the data to justify having the rule.”



Bayesian Inference



“Bayesian inference operates over probabilities, and involves both **prior assumptions about the probability of different hypotheses** and **an estimation of how well a given hypothesis fits the data.**”



Bayesian Inference



“A **Bayesian model** assumes the learner...has some space of hypotheses H , each of which represents a possible explanation for how the data D in the relevant part of the child’s input were generated...Given D , the modeled child’s goal is to determine the probability of each possible hypothesis $h \in H$, written as $P(h|D)$, which is called the **posterior** for that hypothesis. This is calculated via **Bayes’ Theorem**.”

$$(1) \quad P(h|D) = \frac{P(D|h)*P(h)}{P(D)} = \frac{P(D|h)*P(h)}{\sum_{h' \in H} P(D|h')*P(h')} \propto P(D|h) * P(h)$$

“In the numerator, $P(D|h)$ represents the **likelihood** of the data D given hypothesis h , and describes how compatible that hypothesis is with the data...hypotheses with a good fit to the data have a higher **likelihood**. $P(h)$ represents the **prior** probability of the hypothesis...this corresponds to how plausible the hypothesis is, irrespective of any data.”



Bayesian Inference

$$(1) \quad P(h|D) = \frac{P(D|h)*P(h)}{P(D)} = \frac{P(D|h)*P(h)}{\sum_{h' \in H} P(D|h')*P(h')} \propto P(D|h) * P(h)$$

“A hypothesis’s **prior** is something that could be specified by UG – but all that matters is that the **prior** is specified beforehand somehow...The **likelihood** and **prior** make up the numerator of the posterior calculation, while the denominator consists of the normalizing factor $P(D)$, which is the probability of the data under any hypothesis. Mathematically, this is the summation of the **likelihood * prior** for all possible hypotheses in H , and ensures that all the hypothesis **posteriors** sum to 1.

Notably, because we often only care about how one hypothesis compares to another, calculating $P(D)$ can be skipped over and the numerator alone used (hence, the \propto in (1)).”

