Consistent unsupervised estimators for anchored PCFGs

Alexander Clark
Department of Philosophy
King's College London
alexsclark@gmail.com

Nathanaël Fijalkow
CNRS, LaBRI, Bordeaux, and
The Alan Turing Institute of data science, London
nathanael.fijalkow@labri.fr

1 Introduction

Learning probabilistic context-free grammars just from a sample of strings from the grammars is a classic problem going back to Horning (1969). This abstract, based on the full paper in Clark and Fijalkow (2020), presents an approach for strongly learning a linguistically interesting subclass of probabilistic context free grammars from strings in the realizable case. Unpacking this, we assume that we have some PCFG that we are interested in learning and that we have access only to a sample of strings generated by the PCFG – *i.e.* sampled from the distribution defined by the context free grammar. Crucially we do not observe the derivation trees – the hierarchical latent structure. Strong learning means that we want the learned grammar to define the same distribution over derivation trees – *i.e.* the labeled trees — as the original grammar and not just the same distribution over strings.

The motivation for this work is to get some theoretical insight into first language acquisition, and particularly into the information sources necessary for the acquisition of syntactic structure. The standard view is that children learn the syntactic structure of their first languages not by purely syntactic means, but rather by using information about the range of available interpretations, derived from the situational context of the sentences they hear and inferences about the intentions and goals of the speaker. This work strongly suggests that the surface strings alone contain enough information for the gross constituent structure to be acquired, without the necessity for such external information sources.

2 Anchored PCFGs

The first and most important condition is that each nonterminal has a single terminal that is only derived from that nonterminal: if so then we say that the terminal is an *anchor* for that nonterminal. This is in essence an exemplar based approach.

Given a probability distribution over strings, for each string w we can define a distribution over contexts, where the probability of a context l, r is proportional to the probability of lwr. We can of course measure distances of various types between these distributions. A long standing idea in structuralist linguistics was to use these distances to infer syntactic information. With the anchoring assumption the core idea is given terminals (words) a, b, c that are anchors for nonterminals A, B, C respectively, we should be able to infer something about the possible production $A \to BC$, and its parameter by comparing the distributions of a and bc; and given some terminal d, comparing the distributions of d and a should tell us something about the production $A \to d$. It turns out that the appropriate measure to use is a Renyi divergence defined as:

$$R_\infty(P\|Q) = \log \sup_x \frac{P(x)}{Q(x)} \tag{1}$$

We will write $\rho(v \to w)$ for this divergence between the context distributions of v and w.

3 Bottom up WCFGs

A weighted CFG is a CFG where each production is associated with a parameter given by a function θ. One important technical detail is to reparameterize the probabilistic grammar, where the parameters correspond to conditional probabilities of the right hand side, given the left, in a top down generative process as a bottom-up process. If a WCFG is in bottom-up form then the parameters satisfy:

$$\theta(A \to BC) = \frac{\mathbb{E}(A \to BC)}{\mathbb{E}(B)\mathbb{E}(C)} \tag{2}$$
$$\theta(A \to a) = \mathbb{E}(A \to a).$$

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 471-473.
Held on-line February 14-19, 2021

This parameterization defines, apart from a few uninteresting edge cases, the same family of distributions as a PCFG, and we can efficiently convert between the two parameterizations.

4 Parameter equalities

The paper shows that with two additional assumptions, detailed below, the following equations hold between the log parameters of each production, in the bottom-up format.

$$\log \theta(A \to BC) = \log \frac{\mathbb{E}(bc)}{\mathbb{E}(b)\mathbb{E}(c)} - \rho(a \to bc) \tag{3}$$

Here the first term on the right hand side is very close to the pointwise mutual information between b and c and depends only on the right hand side of the production. The term $\rho(a \to bc)$ is the Renyi divergence between the distributions of a and bc; this takes a value between 0 and ∞, and penalises productions where distributions of the left hand side and right hand side are far apart.

$$\log \theta(A \to d) = \log \mathbb{E}(d) - \rho(a \to d) \tag{4}$$

The equation for a simple lexical rule is even simpler: the first term is just the log of a quantity that is approximately the lexical frequency of the item involved, and the second term is again the divergence between left hand side and right hand side.

The two additional conditions are also fairly natural, and both informally bound the degree of ambiguity.

- The first is called *Strict Upward Monotonicity* (SUM): a grammar satisfies this condition if adding any new production (in Chomsky normal form) will strictly increase the set of strings generated.

- The second is *Local unambiguity*. This requires that for any production there are some sentences that can only be generated by using that production "in the same place". In other words for a production $A \to \alpha$ there must be strings l, u, r such that every derivation of the string lur contains a tree of the form $A \stackrel{*}{\Rightarrow} u$.

Given the first two conditions the context distributions of terminals that are anchored will lie at the vertices of a simplex in high dimensional space, and those of ambiguous words will lie in the interior. A simple algorithm derived from Stratos et al. (2016) allows us to recover these. Using some naive estimators for the divergences and the expectations in Equations 3 and 4 we can then directly give estimates of the parameters of the productions from a large sample of strings.

5 Discussion

The main result is that there is a simple computationally efficient algorithm that consistently learns all PCFGs that satisfy these three conditions, only seeing the strings that are sampled from the distribution defined by the PCFG. This algorithm uses naive estimators that would be extremely slow to converge; but Clark and Fijalkow (2020) also give some computational experiments with synthetic data that show that even when none of the conditions are satisfied, a variant of this algorithm performs well with reasonably sized samples. However this still relies heavily on the assumption that the samples come from a distribution generated by a PCFG.

While there are many empirically evaluated heuristic algorithms in the literature for unsupervised learning of PCFGs, this is more or less the first algorithm with any nontrivial theoretical guarantees. The three conditions give successively stronger conditions:

- If the grammar is anchored then the algorithm will converge to some grammar that generates the correct set of strings.

- If it is anchored and SUM, then it will converge to the correct CFG, but with possibly incorrect parameters

- If it is satisfies all three conditions it converges to the correct PCFG.

The resulting grammars satisfy a *minimax* property: they are not the smallest grammars but rather the largest grammars with the minimal number of nonterminals. The anchoring condition is clearly too strong in its naive form: a corpus study shows that in a corpus of child directed speech in English (Pearl and Sprouse, 2012) clausal categories are not represented by single word, and fine grained distinctions between lexical categories need a more refined approach, but this is perhaps a limitation of the CFG formalism per se rather than the learning algorithm.

References

Alexander Clark and Nathanaël Fijalkow. 2020. Consistent unsupervised estimators for anchored pcfgs. *Transactions of the Association for Computational Linguistics*, 8:409–422.

James Jay Horning. 1969. *A study of grammatical inference*. Ph.D. thesis, Computer Science Department, Stanford University.

L. Pearl and J. Sprouse. 2012. Computational models of acquisition for islands. In J. Sprouse and N. Hornstein, editors, *Experimental syntax and island effects*. Cambridge University Press, Cambridge, UK.

Karl Stratos, Michael Collins, and Daniel Hsu. 2016. Unsupervised part-of-speech tagging with anchor hidden markov models. *Transactions of the Association for Computational Linguistics*, 4:245–257.