Stochastic Relaxation and the Innatist Approach to Language Acquisition

Proponents of innatist theories often emphasize the intuition that learners need a highly constrained hypothesis space. This by no means implies that such a space is small. For example, in the Principle and Parameters (P&P) framework, a set of N binary parameters implies a set of 2^N binary vectors. In Optimality Theory (OT), a set of N constraints lead to $N!$ total ordering of constraints. The exponentially large hypothesis space makes researchers go beyond brute-force enumeration, and propose a number of learning strategies based on local searches. For P&P, the most well-known local search strategy is *trigger-based learning* (TLA) (Gibson and Wexler, 1994), which updates the hypothesis by flipping a single parameter. Another example of a local search strategy is the *constraint demotion* (CD) algorithm of (Tesar and Smolensky, 2000) in the OT literature. Both TLA and CD have been criticized as not robust with regard to noise and incapable of modeling gradience. In addition, TLA suffers from its sensitivity to the structure of the hypothesis space (Berwick and Niyogi, 1996), but this problem is smaller for CD, since it has access to a much larger hypothesis space; the set of stratified grammars, which admits a natural partial ordering.

Stochastic relaxation is a strategy that replaces the discrete target hypothesis space (H_d) of the learner with a continuous one (H_c), where continuous parameters in H_c can be related to probability distributions. In the present work, we argue that stochastic relaxation provides a unifying perspective between two separate lines of inquiry, which improve on CD and TLA, respectively. In the Stochastic OT model (Boersma and Hayes, 2001) improving on CD, the original space H_d is the set of constraint permutations, with cardinality $N!$, while in the new hypothesis space, each of the $N-1$ constraints is parameterized by a ranking value, which is the mean of a normal distribution of selection points. This continuous space H_c can be represented as $\mathbf{R}^{N-1}$. Improving on TLA, the variational P&P framework of (Yang, 2000) proposes replacing each binary parameter with a binomial probability, thus effectively replacing the discrete hypothesis space $H_d = \{0,1\}^N$ with a continuous space $H_c = [0,1]^N$. For both Stochastic OT and variational P& P, each hypothesis in the continuous space H_c determines a probability distribution in the discrete space H_d, and this probability distribution is responsible for noise/gradience observed in the data (see Figure 1). This type of hierarchical architecture makes it possible to learn probabilistic models based on innatist theories from empirical data.

One advantage of this unifying perspective is that the Bayesian inference algorithm in (Lin, 2005), originally proposed for learning Stochatic OT models, can be generalized and applied to any hierarchical model like Figure 1. In particular, a major problem in the parameter-setting literature – the ambiguity of evidence, can be overcome since the Bayesian inference algorithm does not rely on heuristics for local searches. A general outline of such a learner is given in Table 1.

This discussion also sheds light on the debate between formal and statistical approaches to language acquisition. The recent work in context-free grammar induction (Klein, 2005) has left the impression that adopting a statistical approach implies abandoning innatist theoretical frameworks altogether. By showing that (i) both P&P and OT can be enriched with statistics and (ii) algorithms based on the perspective of stochastic relaxation provide solutions to the learning problems within these theories, we argue that statistics v.s. UG is a false dichotomy, and that statistics is better seen as a methodology for fitting linguistic models.

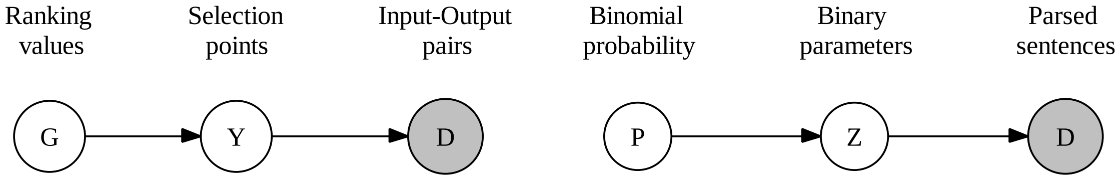

Figure 1: Stochastic OT and variational P&P recaptured as hierarchical models

```
• Let $t \leftarrow 0$ and set the initial hypothesis in the continuous space $h^{(0)} \in H_c$
• Iterate until convergence:
    – Draw a set of grammars from the discrete space $H_d$
        ∗ Draw evidence according to its frequency in the observed data
        ∗ From the distribution governed by $h^{(t)}$, draw a discrete hypothesis from $H_d$
          that is consistent with the evidence
    – Update the $h^{(t+1)}$ by drawing from the appropriate posterior distribution in $H_c$,
      based on the grammars obtained above.
    – Let $t \leftarrow t + 1$.
```

Table 1: Generalized learning algorithm for hierarchical linguistic models

References

Berwick, Robert C. and Partha Niyogi. 1996. Learning from triggers. *Linguistic Inquiry*, 27:605–622.

Boersma, Paul and Bruce P. Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32:45–86.

Gibson, Edward and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry*, 25:407–454.

Klein, Dan. 2005. *The Unsupervised Learning of Natural Language Structure*. Ph.D. thesis, Stanford University.

Lin, Ying. 2005. Learning stochastic OT grammars: A Bayesian approach using data augmentation and Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 346–353, Ann Arbor, Michigan.

Tesar, Bruce and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, Massachusetts: MIT Press.

Yang, Charles. 2000. *Knowledge and learning in natural language*. Ph.D. thesis, Massachusetts Institute of Technology.