

Learning Hidden Structure with a Log-Linear Model of Grammar

Log-linear grammar is a probabilistic extension of Optimality Theory (OT; Prince and Smolensky 1993), or more directly, of Harmonic Grammar (HG; see overviews in Smolensky and Legendre 2006, Pater 2009). Also known as Maximum Entropy grammar, it was originally proposed for syntax by Johnson (2002), and subsequently applied to phonology by Goldwater and Johnson (2003), Wilson (2006), Jäger (2007), and Hayes, Zuraw, Siptar and Londe (2008), amongst others. Log-linear models have a longer history in statistics and in NLP, and their current popularity in generative linguistics largely stems from the availability of provably convergent learning algorithms, which sets them apart from other stochastic versions of OT.

The literature on log-linear grammar sometimes refers to these convergence guarantees without mentioning an important caveat: that they hold only if the learner has access to the full structure of the learning data (on this caveat for the OT Constraint Demotion Algorithms, see Tesar and Smolensky 2000, and for log-linear learning in NLP, see Riezler 2000). Eisenstat (2008) provides a general model for the learning of hidden structure in the log-linear framework, and shows that it succeeds on a toy case of learning of phonological underlying representations (URs). It remains unknown the extent to which language learning problems create local maxima that can trap such a learner, and the extent to which these local maxima can be avoided by applying existing unsupervised learning techniques from connectionist and statistical learning.

We first show that local maxima do arise for log-linear learners, as for most other OT and HG learners, in even the simplest of scenarios involving hidden linguistic structure, including Eisenstat's UR learning case. We further demonstrate that these traps can be eliminated by making use of regularization, which penalizes the objective function as weights diverge from zero. In addition to preventing overfitting, regularization can smooth a function by removing some local optima (Chen and Rosenfeld 1999).

Regularization is not a simple panacea for hidden structure problems, since in changing the shape of the learning space, it can also steer the learner away from the global maximum. We therefore also present tests of log-linear learners on two more complex learning cases. In the first, we found that Tesar's (2006) UR learning problem can be solved using log-linear learning with constraints on URs (Boersma 2001, Apoussidou 2008). This problem can also be solved in OT using Apoussidou's application of Tesar and Smolensky's Robust Interpretive Parsing with Boersma's (1998) GLA, but the OT learners' success is heavily dependent on initial conditions.

The second, more challenging, test involved the learning of stress systems with hidden prosodic structure, using a benchmark set of learning problems developed by Tesar and Smolensky (2000). The problem set consists of 124 languages that can be generated by 12 metrical constraints. The learner is given the stress patterns, and must infer the correct prosodic structure. The goal for log-linear learning was to maximize the probability of the observed data, subject to a Gaussian regularization prior with unit variance. A batch gradient-based optimization algorithm found constraint weights that assigned highest probability to the observed stress pattern in 88% of the cases, averaging over languages and word types. This improves significantly on Tesar and Smolensky's OT results, and is equal to Boersma and Pater's (2008) results using Robust Interpretive Parsing and a stochastic ("noisy") version of HG.

These initial results are especially promising in that the availability of a mathematically well-understood, consistent probabilistic model of the learning of hidden structure opens the door to a range of further approaches to these problems. In ongoing research, we are experimenting with deterministic annealing (Smith and Eisner 2006), which gradually reduces a constraint on the entropy of the model, as well as with on-line learning (Stochastic Gradient Ascent). This line of research has the potential to yield models of language learning that can cope with hidden structure, and that at the same time deal with language variation, and display realistic learning paths (cf. Tesar 2000, 2006).