An Overview of Stochastic Methods in Optimality-Theoretic Approaches

Mafuyu Kitahara*

Key words: Gradual Learning Algorithm, Harmonic Grammar, Maximum Entropy, convergence, on-line, probability distribution

1. Introduction

The purpose of this paper is to give an overview of stochastic methods in Optimality Theory (OT; Prince and Smolensky 1993/2004) and related approaches, which has been developed and spread quite rapidly in these 10 years.

Stochastic methods in linguistics is not uncommon at all. Phoneticians, psycholinguists, and sociolinguists, in particular, cannot ever conduct their research without a reasonable background in statistics. However, the core part of theoretical linguistics had been rigorously symbolic and categorical. There had been little room for quantitative and probabilistic account in phonology, syntax, and semantics. Introducing stochastic methods in OT may still look somewhat out of place for a number of generative linguists. Thus, I believe it is worth reviewing the basic premises, scaffolds, and insights of stochastic methods in OT.

As a basis of the following discussion, let me introduce a “ladder of abstraction” diagram. The idea is originally given in Pierrehumbert (2003), but schematization is my own. In Fig. 1, each tier consists of an abstraction of probability distribution of index values in one step below. Tier I consists of phonetic parameters whose values largely vary depending on the context. Tier II picks up the statistical summary of tier I to form its representations (which roughly corresponds to traditional phonemes). Tier III is based on probabilistic timing relations of units in tier II to form a word-sized units. Word frequencies and neighborhoods come in at tier III as well. Tier IV is a set of templates or constraints of possible word forms (i.e., phonotactics). They are all dependent on the probabilistic distribution of existing words. Finally, tier V captures paradigmatic relations among morphemes, which cannot simply be defined by templatic operations. This tier, again, is based on probabilistic distribution of relations between morphemes.

According to Pierrehumbert, each tier is not considered as an independent module, but as a part of a massively connected network. Lines on the right end of the figure are intended to represent the inter-tier connections.

The reason for presenting this figure is to give a map on which we can locate the area corresponding to each of the analysis reviewed in this paper. In addition, as with Pierrehumbert, I would like to emphasize that probability distributions are everywhere in the models of speech production and perception.

Another, simpler diagram in Fig. 2 shows an example of realistic data we often see in linguistic analyses. If we are interested in a gradual change over time, such as

* Professor, School of Law, Waseda University (早稲田大学法学部教授)
L1 or L2 acquisition, language contact, and historical change, one aspect of such change can be represented as an S-shaped curve on a plain with time on the x-axis as in Fig. 2(a).

If we are interested in synchronic variations of some kind, we can take stimulus number on the x-axis instead. An S-shaped curve typically appears in a categorical perception, such as a VOT continuum and perception of voiced/voiceless plosives. Furthermore, intermediate well-formedness can be thought of as the relatively steep part in the middle. When stimuli with systematic phonetic/phonological deviations are rated by subjects, the ratings will be distributed over such a curve. Synchronic variations can also be ratios. A free variation between two linguistic forms can be depicted as bar-plots such as the one in Fig. 2(b).

We have seen so far that both the model (Fig. 1) and the data (Fig. 2) include non-categorical, quantitative,
and probabilistic distributions. Now we turn our attention to grammars and learning algorithms.

The two grammars we will see in this paper are OT and its sibling, Harmonic Grammar (HG; Legendre, Sorace, and Smolensky 2006). Both grammars, as the theoretical framework, have the potential to cover all the tiers in Fig. 1. Most studies reviewed in this paper focus on specific phenomena and run simulations on concrete examples. The choice of the grammar and a particular linguistic and/or probabilistic aspect of the data characterize each study.

A phenomenon characterized as a function of time, such as the one in Fig. 2(a) requires a method that enables the change of grammar across time. That is, we need something outside of the grammar to make that change. One of the striking advantages of OT compared with the preceding generative framework is that it comes with an explicit learning algorithm.

Constraint Demotion (CD; Tesar and Smolensky 1998) is the first of such algorithms developed for classic OT. CD, as its name stands for, can manipulate the discrete ranking of constraints. Gradual Learning Algorithm (GLA; Boersma 1998) introduced stochastic methods to deal with continuous ranking values while adapting the core procedures of CD.

Maximum Entropy method (ME) has also been frequently mentioned in the recent (stochastic) OT literature. It is an independently developed modeling method outside of the OT community, and quite popular in the field of machine learning and computational linguistics due to its sophisticated mathematical properties. ME model has a capability of a grammar component in a broad sense: it can handle a series of weights for constraints to evaluate a set of candidates. Learning algorithms for ME are abundant: they are gradient ascent (GA), conjugate gradient ascent (CGA), and simulated annealing (SA), among others.

The organization of this paper is as follows. The next section reviews the basic mechanism of classic OT along with a comparison to HG. Section 3 gives the details of how stochastic methods are introduced in the grammar combined with a particular learning algorithm. The section also covers the key points of simulation studies in the light of probabilistic nature of the model and the data as presented above. The final section summarizes the paper and gives miscellaneous information for further research.

2. The Basics of Classic OT and HG

It has been almost two decades since the advent of OT. We have seen a flourishing development of a good many different versions of the theory so far. However, let us concentrate on a comparison between the classic OT proposed in Prince and Smolensky (1993/2004) and HG in Legendre et al. (2006) in this section.

The constraint ranking in classic OT is strict: a candidate with many violations of lower-ranked constraints can still beat other candidates with only one violation of a higher ranked constraint, as shown in (1). The strict ranking only asserts the order of constraints, not their distance. For example, we cannot know whether the distance between C\textsubscript{1} and C\textsubscript{2} is greater than that between C\textsubscript{2} and C\textsubscript{3} in (1).

\[(1)\]  

\begin{array}{|c|c|c|c|}
\hline
i & C\textsubscript{1} & C\textsubscript{2} & C\textsubscript{3} \\
\hline
1 & \star & \star & \\
\hline
2 & \star & \star & \\
\hline
\end{array}

On the contrary, constraints in HG are not just ranked but weighted: a higher ranking corresponds to a larger...
weight value. Thus, the distance between constraints can be simply defined as the difference of weights. As a consequence, multiple violations of a lower ranked constraint can override a single violation of a higher ranked constraint if the number of violations exceeds the difference of the weights of the two constraints. Following traditions in HG-tableau, the number of violations is shown as a negative integer and the weight for each constraint, which is the Harmony of that candidate. In this example, candidate \(o_1\) has a violation of \(C_1\), the highest ranked constraint, can still survive due to \(o_1\)’s multiple violations of lower ranked \(C_2\) and \(C_3\). This overriding situation is called “gang-up effect” in the literature.

(2) HG

<table>
<thead>
<tr>
<th>(i)</th>
<th>(C_1) (w=3)</th>
<th>(C_2) (w=2)</th>
<th>(C_3) (w=1)</th>
<th>(H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(o_1)</td>
<td>(-1)</td>
<td>(-2)</td>
<td>(-1\times2-2\times1=-4)</td>
<td></td>
</tr>
<tr>
<td>(o_2)</td>
<td>(-1)</td>
<td>(-3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The main difference between OT and HG is that the former has a strict domination while the latter does not, which allows the latter to have a gang-up effect. It is in fact difficult to express the strict domination in a series of weights. For example, if we put \(w=30\) as the weight for \(C_1\) in (2), the 31st violation of \(o_1\) on \(C_1\) can still make the other candidate to survive. In other words, the distance between any two constraints must be greater than \(n\) for a \(n\)-times violated lower ranked constraint not to override a higher constraint.

3. Combination of grammars, stochastic methods, and learning algorithms

Boersma and Hayes (2001) is reviewed first in this section. In their model, the ranking of constraints are probabilistically defined, which is sometimes called Stochastic OT (StOT).

Following their step-by-step presentation of StOT, let us first introduce arbitrary numerical values for the ranking in Fig. 3(a) and put all constraints on a scale as in (b). If we allow constraints to have a range (or band) of ranking values as in (c), it is possible to have a flexible ranking in which constraints with overlapping ranges can be swapped. If we further hypothesize that random noise is added to the range to form a normal distribution as in (d), the swapping can now be incorporated into the grammar with probabilities.

It is noteworthy that the random noise is added only at the evaluation, but not in the constraint formulation. In other words, probability distributions of ranking values are not a parameter for each constraint. Thus, the ranking between \(C_2\) and \(C_3\) are, until the evaluation, not determined and can only be construed as probabilities.

Since the noise is not a property of each constraint, the width of the distribution is shared by all the constraints. Hence, constraints are different only on their center values and ranked on the scale as in Fig. 3(d). Boersma and Hayes (2001) carefully eliminated the possibility of specifying different distributions for each constraint: it is too unrestricted as the model of a grammar.

If we allow different distributions for each constraint as in (e), constraints may have a radically wider distributions than other constraints. Here, the constraint \(C_2\) can be ranked anywhere in the hierarchy, which is not a desirable situation for a language learner.

The learning algorithm coupled with StOT is GLA proposed in Boersma (1998). It is a specially developed method for OT/HG learning. The difference between the target ranking and the current ranking is first detected and it gradually makes the difference smaller by modifying the ranking values. The mobility of constraints are set by a parameter called “plasticity”, the amount of change of ranking values in one step. The change can take place only at the constraints that are detected to differ from the target ranking.

The StOT-GLA combination is applied for the metathesis and reduplication in Ilokano and genitive plurals in Finnish. Data from these two languages both show a massive amount of free variation, which has been a challenge for classic OT analyses.

Ilokano data in Table 1 show that some surface representations are expected to appear in the equal ratio as a result of free variation but others are not due to fatal

<table>
<thead>
<tr>
<th>Underlying</th>
<th>Surface</th>
<th>Target</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>/taw.en/</td>
<td>taw.en</td>
<td>50%</td>
<td>52.2%</td>
</tr>
<tr>
<td>ta?wen</td>
<td>50%</td>
<td>47.9%</td>
<td></td>
</tr>
<tr>
<td>ta?wen</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ta?zen</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>/HRED-bwaja/</td>
<td>bu?bwa.ja</td>
<td>33.33%</td>
<td>36.7%</td>
</tr>
<tr>
<td>bw?bwa.ja</td>
<td>33.33%</td>
<td>31.2%</td>
<td></td>
</tr>
<tr>
<td>bw?bwa.ja</td>
<td>33.33%</td>
<td>32.1%</td>
<td></td>
</tr>
<tr>
<td>bw?bwa.ja</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
violations (e.g., \( \text{MAX} \) violation for deleting a glottal stop as in /ta.wen/). The data includes both phonotactic constraints and paradigmatic relations among morphemes and thus corresponds to tiers IV and V in Fig. 1. The simulation results closely trace the ratio of free variation as can be seen in the third and the fourth column of the table.

Boersma and Hayes (2001) also attack the problem of intermediate well-formedness of clear and dark /l/ in English. This phenomenon is in the area of allophonic realization corresponding to tier II in Fig. 1. However, subjects rated allophones only through word forms. Thus, tier III must also be related to this phenomenon. In the analysis, Boersma and Hayes first computed the expected ratio of occurrences for each word forms based on the ratings in well-formedness judgments. Then they conducted a learning simulation whose target is the computed ratio. The simulated ratio of word forms are then reverse-computed from the learning results, which closely corresponds to the real data.

Overall, their analyses are simulating the actual distribution of linguistic data. That is, the data in their study are those in Fig. 2 which are explained by stochastic methods both in the grammar and the learning algorithm. The probabilistic nature of tiers in Fig. 1 is not directly accounted for, though it may be working in the background.

Keller and Asudeh (2002) point out several problems in Boersma and Hayes (2001)'s approach. There are two aspects in the problems: one is related to the fre-
quency in the corpus for learning, and the other is the issue of convergence.

Frequency of occurrence is crucial for the intermediate well-formedness issue just reviewed above. A tacit assumption there is that there is a monotonic relationship between the frequency and well-formedness: the more you hear, the more it is well-formed. Keller and Asudeh (2002) criticize that this is a dangerous move to destroy competence-performance distinction. In my view, the well-formedness data in Boersma and Hayes (2001)’s analysis is just a type of word familiarity. Grammatical well-formedness is a controversial notion covering extremely wide range of studies, not just phonology, but morphology and syntax. Somewhere in the range, there might be a considerable overlap between well-formedness and familiarity, but they are not the same. Familiarity is more strongly influenced by frequency, but grammaticality is less so. Competence-performance distinction is still maintained if we interpret their analysis in this way.

The convergence problem is seriously taken up in the subsequent studies. Three studies are summarized in the below: Goldwater and Johnson (2003), Jäger (2003), and Boersma and Pater (2008).

Goldwater and Johnson (2003) compared ME with OT-GLA and found that one of the advantages of ME to OT-GLA is that ME is proved to converge. They claim that, all else being equal, ME is the better model because of its mathematical sophistication. They re-run all the simulations in Boersma and Hayes (2001) in ME method to find quite similar results.

Jäger (2003) pursues an ME method to cope with the convergence problem, too. He then raises an issue of on-line/off-line learning. On-line learning takes one datum at a time, which is supposed to happen in human language acquisition. Off-line learning is, on the contrary, not taking data one by one. Pre-processing of the entire corpus is required before the actual learning, which is quite unlikely for human to perform. Gradient ascent coupled with ME in Goldwater and Johnson (2003) is an off-line algorithm, which is severely criticized not only by Jäger (2003) but also by Boersma and Pater (2008). Jäger proposed that an on-line version of GA called Stochastic Gradient Ascent (SGA) can work together with ME.

Boersma and Pater (2008) deals mainly with HG combined with GLA. Their concern with OT-GLA is also that it is not guaranteed to converge. Thus, they first concentrate on mathematical proofs of convergence for HG-GLA. Simulations in various conditions support their argument that HG-GLA succeeds to converge while OT-GLA cannot. Their HG also incorporates stochastic methods to handle variation. As in StOT, random noise at evaluation is introduced for constraint weights to fluctuate. The GLA runs in a similar fashion as in StOT. The only difference is that the change of weight is sensitive to the distance between the result of each evaluation and the target grammar. Inheriting from OT-GLA, learning process in HG-GLA is on-line as well.

### 4. Concluding remarks

In the previous section, we have seen several combinations of grammar and learning algorithm, which are summarized in Table 2.

However, the tiers in Fig. 1 are not fully covered in the models mentioned in this table. In particular, tier I, phonetic parameters are not covered in Boersma and Hayes (2001) and the subsequent models. Development and testing of the models seem to be focusing on the issues of learnability and cognitive plausibility.

In fact, probabilistic aspects of phonetic tiers have been implemented not in the learning module but directly in some constraints. Kirchner (1998), for example, proposed that Lazy is an n-ary constraint based on physical definition of effort. There are numerous approaches to quantify some of the probabilistic distributions in the phonetic data in formulating constraints.

There is yet another venue for incorporating probability distribution in the model: the candidate set. Simulated-Annealing OT (SA-OT: Biró 2006) is the one that seeks this option to deal with optionality and gradient properties. SA-OT is proposed as a real-time simulator.
of linguistic performance (not competence).

Candidates, constraints, grammar, and algorithm: there are four areas we can choose whether to incorporate probabilistic mechanisms. Fig. 1 and 2 show that the data are full of probabilistic properties. The choice from the four areas will be determined first by the nature of the data and second by the viewpoint over the higher-level issues, such as “what is grammaticality”.

In conclusion, the current research paradigm of stochastic methods in OT is almost ready to flourish. The formal aspect of the paradigm, such as the issues around convergence and on-line/off-line learning are frequently discussed, but the empirical aspect needs wider and more thorough inspection of, for example, longitudinal data of L1 and L2 acquisition, sociolinguistic variation, and language change. Combination of computational models with empirical data is a rich and still uncultivated area of research, as pointed out in Boersma and Levelt (2003). In the field of sociolinguistics, Sano (2009) has shown a combination of the StOT/ME and large-scale corpora to analyze variations in the voice system in Japanese verbs. Jäger (2004) gives an insightful analysis of applying stochastic methods in OT to language change, using his own-developed “evo-IOT” software. Conveninetly, most simulation tools are available on-line for usual personal computers, such as StOT/HG/GLA[1], ME[2], SA[3], and evo10T[4]. Now, it is your turn to run a simulation to open up a new dimension of research.

**URL list**

[4] evo10T http://www2.sfs.uni-tuebingen.de/jaeger/evo10T/

**References**


(Received Jul. 10, 2009 Accepted Jul. 14, 2009)