Decision Theoretic Models of Optimality

Betsy McCall
University of Pittsburgh

Abstract. This paper examines variability in Optimality Theoretic models by considering their mathematical representations. To this end, four variations on Optimality Theory are modeled as simple Decision Theoretic utility functions that are then analyzed and compared. These versions include a strict version of OT, a version of OT that permits obligatory constraint tying, a version that permits multiple violations of individual constraints, and a stochastic model. The mathematical models help to highlight any of the theoretical difficulties in each version, as well as the power of a simple stochastic model. This paper will consider the implications that such models have for linguistic theory and for future research with respect to Universal Grammar, language acquisition, natural language processing and the dynamics of language change.

1. Introduction

Optimality Theory was first introduced to the linguistics community in 1993 in Prince and Smolensky's seminal work "Optimality Theory: Constraint Interaction in Generative Grammar". In very simple terms, Optimality Theory describes a series of ranked and interacting constraints that represent two opposing forces in language: faithfulness to some underlying representation, and well-formedness. According to the principles of Universal Grammar, all these constraints are spelled out and, while they can be reranked to accommodate acquiring a particular language, cannot be added to. This implies that there is a fixed number of N constraints.

Since the introduction of Optimality Theory, the theoretical details have been expanded by a number of people. In this paper we will not primarily be considering the different types of constraints, but the way in which constraints are violated and ranked.

Decision Theory is a science and mathematics dedicated to understanding decision-making under uncertainty. Uncertainty is present in all levels of a speaker's language understanding—in learning; in comprehension (when dealing with ambiguity resolution, for instance); and in production. By this reckoning, understanding language models through Decision Theory is a necessary approach, as Decision Theory helps us determine which strategies are reasonable when all factors affecting a situation are not known. Decision Theory allows us to convert our knowledge of the world, usually gained through statistical knowledge, into a utility function which helps us analyze future decisions based on previously acquired information.

In this paper, we will examine four different versions of Optimality Theoretic models in Decision Theoretic terms. The goal is to examine theoretical strengths and weaknesses of the different versions of Optimality Theory in order to determine which models need to be reexamined or discarded, and the nature of future research into the nature of remaining models.

2. Models

Decision Theoretic models are mathematical formulae that relate the utility of an outcome, whether it's desired or undesired and to what degree, with the expectation or probability of that outcome. For our purposes, Optimality Theory itself has taken care of this with the constraint

ranking. Rather than introducing a complex statistical utility function, we will adopt the notion of the constraint ranking, which already incorporates the notions of expectation of success or failure, and transform this into a mathematical equation that captures the violation of constraints and the relative weights of constraints. The utility functions discussed in this section are mathematically simple, yet telling.

Each of the models described are made of up two principle features: the constraints themselves and the constraint ranking. Each of the constraints, according to the theory of Universal Grammar must be listed in each speaker's grammar at birth; therefore, there cannot be an infinite number of constraints, but some finite number N of them. Each of our utility functions will be based on a summation of successfully satisfied constraints, as well as the value of success for that constraint. Because there are a finite number of constraints, we need not concern ourselves here with notions of mathematical convergence. Each of the constraints will be indicated by a variable. In the case where we do not allow multiple violations of constraints, the constraints will be given by I_{0j}. This notation indicates that each constraint is marked by an indicator variable, taking on the values of zero or one to indicate failure or success respectively, and numbered with the j-subscript as one of the N total constraints in the grammar. A constraint ranking will permute the constraints and associate them with a ranking according to their utility in a given language. The ranking itself will be given by a variable a_i that will indicate the value of a constraint associated with it being satisfied. If the constraint is satisfied the coefficient will add that much utility to the overall value of the function; if the constraint is not satisfied, the coefficient will be multiplied by a zero and no additional utility will be contributed. Each of the models considered below will rely on some variation of these basic ingredients.

2.1. Strict Optimality Theory

The first of the Optimality Theoretic models we will consider is a strict version of Optimality Theory at its bare bones. This version of Optimality Theory is similar to that ascribed to by John McCarthy (2002) and others. The components of this version of Optimality Theory are quite restrictive. First, although it may not be clear to an outsider how constraints are ranked when two constraints do not appear to interact, the speaker must, in fact, rank them, permitting minimal variations within a constraint ranking to produce identical grammars. Second, constraints may not accept multiple violations. Constraints are naturally only satisfied or unsatisfied—thus requiring the use of the indicator variable. Thirdly, constraint rankings, once fixed at the conclusion of language acquisition, cannot be modified and constraint rankings are impermeable, not admitting to probabilistic variation. The utility function for this strict OT is given in (1).

$$\mathrm{U}(x) = \sum_{i=1}^{n} a_i I_{0_j} \tag{1}$$

This equation says simply that the utility function U, operating on some element of language x, an input for instance that the grammar is analyzing for speech production, has a utility in the language equivalent to the sum of the values of the satisfied constraints. The winning candidate will be the candidate with the highest utility.

We can also take a more literal interpretation of Optimality Theory. Typically, in OT tableaux, constraint violations are marked rather than constraint successes. We can instead consider a loss function, given in (2), where constraints that are satisfied receive a loss value of zero, and constraints that are violated receive a loss value of one times the α_i value for its place in the constraint ranking. An input to the function that receives the lowest value of L is the

winning candidate. It can be shown that maximizing utility and minimizing loss are equivalent results (Berger, 1985), so that for the remainder of the paper I will primarily only be considering optimizing the relevant utility functions, although I will comment further if the correspondence between loss and utility is not obvious.

$$\mathrm{L}(x) = \sum_{i=1}^{n} \alpha_i I_{0_j} \tag{2}$$

In (1) and (2), there is no specification of the values of a_i. In order to achieve the kind of constraint ranking that is described in Optimality Theory, a further specification of the values of a_i needs to be added here. So that a single constraint cannot have a lesser utility than the sum of lower ranked constraints, each coefficient in the ranking must satisfy the equation in (3).

$$\sum_{k=1}^{i-1} a_k \leq a_i \tag{3}$$

So consider, if the lowest constraint in the ranking is equivalent to a value of one, the next highest ranked constraint must be a little higher, say, (1+ε), where epsilon is some small amount greater than zero. The next ranked constraint must be at least this sum, and so forth. If we continue with this scheme, then if there are N constraints, the utility value of the highest ranked constraint is 2^{N-1}, and the total possible utility would be approximately 2^N. This relationship between the highest and lowest ranked constraints would be true, regardless of the scaling factor used. Since it is unlikely that for a given constraint, all constraints of lower utility will be satisfied—the higher the constraint is ranked, the less likely this becomes—we can simplify the equation in (3) so that there is just an equal sign.

Constraint interaction may also occur in strict Optimality Theory in a limited fashion through constraint conjunction. The utility model described here can be made to naturally incorporate constraint conjunction. Constraint conjunction represents a logical AND between two independent constraints. These can be derived from constraint interaction in our model by permitting multiplication of the two constraint variables that are conjoined. Both must achieve a value of one for the multiplication to be nonzero. Constraint conjunction has logical consequences for the grammar. Even if we permit only two constraints being conjoined at once, if all the possible conjunctions must be listed in Universal Grammar and not acquired during the learning process, we increase the maximal number of constraints by N(N-1); i.e. the maximal utility of the grammar is now two raised to the N^2 power. If we were also to admit of language specific constraints, and expanding OT to other parts of the grammar, N becomes large very quickly and N^2 lager still. This relates directly to the problem of the infinities. Though not technically, infinite, the size of appears to be capable of growing nearly without bound.

2.2. Other constraint impermeable models

Linguists champion this kind of strict model of Optimality Theory because it is theoretically simple. Just as we can see from the mathematical representation, it requires only two relationships between the grammar and the value of an element: the ranking itself, and the relationship between the constraints and the ranking. The simpler a model is, the easier it should be to acquire and encode in UG. The drawback to the model remains in the question of whether or not it can capture all of the features of known languages and language acquisition. Thus, other models have arisen. In this section we will consider two possible variations on Optimality Theory that preserve the notions of constraint impermeability.

2.2.1. Tied constraints

A model of Optimality Theory that satisfies the second and third features of strict OT as described in §2.1, but which permits constraint tying is described here. Versions of Optimality Theory that incorporate constraint tying do so for two possible reasons. The first of these reasons was initially proposed as a possible account of producing variation within Optimality Theoretic grammars, particularly with effects such as emergence of the unmarked and context effects. The second possibility is that tied constraints can produce the effects of a logical OR within the grammar. The general utility function is given in (4). We call it U_t for 'tied' to distinguish it from the function for strict OT, although the equation is identical. The changes come in the way we define the coefficients that figure into the constraint ranking, given in (5).

$$U_t(x) = \sum_{i=1}^{n} a_i I_{0_j} \tag{4}$$

$$a_i = \begin{cases} a_k & k = i-1 \text{ and constraint tied} \\ \sum_{k=1}^{i-1} a_k & \text{otherwise} \end{cases} \tag{5}$$

In order to achieve constraint tying, the possibility for two successive constraint weights being equal must be allowed. Equation (5) says that for most constraints, we define successive constraints as we would for strict Optimality Theory, as equal to (or greater than) the sum of all lower ranked constraints. However, this definition of the a_i's leaves open the possibility that a constraint may be tied in utility to the one immediately preceding it in the ranking. This formulation only tells us that a constraint, as it is added to the ranking, may be ranked equal to the previous one in the ranking. This particular model does not specify any limit on the number of constraints that may be ranked equally. To prevent this from happening, we would require another constraint, perhaps that $a_i \neq a_{i-2}$. Without this additional constraint, this clearly can be a way to reduce the maximum utility (numerical size) of the grammar by not requiring non-interacting constraints to be ranked with respect to each other, particularly for very highly constraints that are never violated in the working language, or for very low ranked constraints that are never satisfied, to be ranked equally and contribute less to the ratio between the highest ranked constraint and the lowest. Reducing the unused portions of the grammar should result in simpler computation of winning candidates by placing more emphasis on constraints that are actually decisive.

2.2.2. Multiple violations

A model of Optimality Theory that satisfies the first and third requirements described in §2.1 for strict OT, but that allows multiple violations of constraints is described in this section. Multiple violations of a constraint, or gradient effects, arise typically in certain well-formedness constraints such as those governing right- or left-headedness. If a constraint receives a violation for each syllable, for instance, as it moves into a word, it may be recorded in an analysis as receiving multiple violations if it moves beyond the first syllable. Distinguishing the accent placement, for instance between the first syllable versus the second or later syllables then, can be easily obtained from single violations, but distinguishing between second and third syllable is often achieved through allowing multiple violations. John McCarthy (2002) specifically rejects such gradient effects, but since the process is common in existing models of a wide range of phenomena, we describe it here.

Achieving multiple violations cannot be achieved through the use of an indicator variable. Rather, another variable, here labeled, z_j, is an ordinal variable. For constraints that can achieve only success or failure, nothing has changed except the label, since not all constraints need to be gradient. However, for constraints that achieve multiple violations, values of two, three, four, or whatever whole number is needed can be achieved. Our utility function now looks like (6).

$$U_{mv}(x) = \sum_{i=1}^{n} a_i z_j \tag{6}$$

Because we are no longer considering a simple indicator variable, we once again need to reconsider our coefficient ranking. In order to keep the strict ranking approach of previous models, we need to adjust our a_i values to accommodate multiple violations of a constraint. To guarantee that higher ranked constraints will always have a higher utility value than constraints that can have multiple violations, we need to consider the maximum utility value of the constraint in question given complete success. Our indicator variables allowed for a zero value if the constraint failed to be satisfied and a value of one if it succeeds. Now, since there are different degrees of failure, there must also be different degrees of success. Negative numbers are not allowed, so one way around this is to determine the maximum number of violations that are permitted that are still useful in the grammar. If an accent, for example, appears only on the last three syllables of a word, for instance, then three violations guarantee failure. There is no need for a fourth degree. This maximum number of violations achieves a zero value, and complete success, or no violations, receives this ordinal value in the constraint ranking. The maximum value will be something learned in language acquisition. The equation for this scheme is given in (7) and (8). We choose (7) if we wish to consider the maximum total violations (regardless of where the usefulness of such violations ends) which depends entirely upon observation, and (8) if m_j represents the maximal decisive violations associated with each constraint, something that would require a deeper understanding of the grammar. This value may indeed be one (the minimum value), and we return to strict OT if this were true for all constraints.

$$\sum_{k=1}^{i-1} a_k \max(z_j) \le a_i \tag{7}$$

$$\sum_{k=1}^{i-1} a_k m_j \le a_i \tag{8}$$

One of the weaknesses of such a model is that it increases the size and complexity of the grammar. The value of utilities for all successive constraints must be ranked higher to maintain the constraint ranking. The same effect might conceivably be achieved by splitting up the constraints, just as we do for place feature faithfulness and as would be done in a statistical analysis of an ordinal variable, into pieces labeled with indicator variables and ranking these successively, one after another (Kleinbaum, et al., 1998). It also forces us to establish an additional relationship between the constraints to ensure that the constraint with three violations is not ranked above the one with two violations. This approach, of course, increases our value for N. Constraint conjunction also represents a problem for constraints with multiple violations. Would conjoined constraints reduce to I_0 or maintain the gradience of the bare constraint.

It is certainly conceivable that variations on Optimality Theory exist that incorporate features of both tied constrains and multiple violations of constraints. Combining features of both constraint tying and multiple violations would not change our general utility function much, as we've seen, but would change dramatically the way in which we define our coefficients, particularly for tied, gradient constraints. I leave these variations to the imagination of the reader.

2.3. Stochastic Optimality Theory

Stochastic OT was introduced as yet another method of handling variation in a synchronic grammar. Constraint tying was proposed originally as a way of achieving variation, but in the end, this technique only permits lower ranked constraints to be the deciding factor, leading to variation which is ultimately contextual. Stochastic OT permits variation which is truly random. The mathematical model of a stochastic model of Optimality Theory is given in (9).

$$U_s(x) = \sum_{i=1}^{n} (a_i + b_i Y_j) I_{0_j} \tag{9}$$

The model given in (9) contains the usual features of strict OT, indicator variables for each constraint, and a coefficient a_i for the constraint ranking. The second term b_iY_j of the coefficient is the stochastic portion of the grammar, which is irrelevant if the constraint itself is not satisfied. Each Y_j represents a random variable associated with each constraint. Each Y_j takes on the value of one with probability p_j and zero with probability $(1- p_j)$ When the random variable Y_j achieves a value of one, then the value of the coefficient b_i adds to the value of the utility function. (We assume here that the random variable is evaluated once for every input, and not once for each candidate individually.) A model for the strict version of OT can be achieved when all the p_j's are very close to or equal to zero, as this would leave only the bare constraint ranking. However, when we change the value of some of the p_j's, constraint permeability appears.

If the magnitude of the coefficient is free, the degree of permeability depends upon the magnitude of the coefficient of the random variable in relation to the value of the constraint itself. Values of b_i significantly smaller (or larger) than the corresponding a_i permit contextual variation with random variability, as a combination of smaller ranked constraints may combine to produce a utility greater than the single constraint alone. Values of b_i that are equal to the corresponding a_i will cause variation with the constraint ranked immediately above it. When we combine this with a p_j value equal to one, we regain the tied constraints model. The ability to recover several other models here is a strong plus for this model. This is straightforward for indicator variable constraints, but becomes more complicated for constraints that permit multiple violations, and I will not address those complications here.

In order to achieve maximum learnability, we need to gain maximum control of the theory; we would like to reduce the variation in the model to only what is needed to account for behaviour. Ideally, allowing the value of b_i to depend directly on the corresponding a_i, and $b_i+a_i \approx a_{i+1}$, so that constraint permeability is possible in only one direction, and the values of the b_i's do not need to be acquired separately. This would permit constraint stochastic effects only with two successively ranked constraints. However, this restriction leaves open certain theoretical questions. When we consider small segments of a grammar in analyzing a particular behaviour of interest, it is not difficult to get two constraints that are varying with each other to be ranked together. The question that remains, however, is will these constraints remain

consecutively ranked when the full grammar is considered? Until complete Optimality Theoretic grammars are developed, and analyzed, that are meant to account for an entire language, complete with variation, what restrictions can be placed on the stochastic portion of this model remains to be seen.

3. Implications of the models

These mathematical models of Optimality Theory have implications for linguistic theory. Some of these implications have already been addressed above, but in this section, I would like to highlight these and others relating to some specific theoretical issues.

3.1. Universal Grammar

Universal Grammar is a central feature of modern linguistic theory. These models have a lot to say about what UG would have to contain with respect to Optimality Theory. We saw in (3), given in §2.1, how our constraint ranking must be accounted for in our utility function. Given that multiple violations and tied constraints are not a feature of this version of OT, the values of the constraint ranking for our utility function, and the utility function itself can be listed in UG. A speaker would have to acquire the permutation of constraints so that the coefficients can be associated with the correct utility values. The coefficients themselves, however, may be contained in UG since, given a fixed number of constraints, under this model the value of each coefficient would be invariant across languages. This would also be true of the stochastic model given here if we assume that the b_i's are dependent upon the a_i's and that p_j=0 is the default for all constraints initially.

On the other hand, as we've seen, if we assume that all constraints (and their binary conjunctions) are listed in UG, we have a very large grammar which to work from. This is powerful, but unwieldy. Models of UG that permit constraint learning can help to minimize the size of a grammar significantly. Humans are known to have difficulty managing small and large numbers simultaneously, so reducing a grammar to its minimal parts could be advantageous.

3.2. Language acquisition

These models address several features relevant to language acquisition. Assuming that UG conforms to the strict version of OT described in §2.1, language acquisition would be at its simplest of the four models. A speaker would have only to acquire the constraint ranking that maximized the utility function. Other models present more difficulty for language acquisition. That in itself should not be interpreted to mean that they are wrong as each has its own benefits.

The tied constraints model has the benefit of reducing the final grammar almost as much as acquiring constraints reduces it. However, if a tied constraints model is accurate, then the values of the coefficients in the model must be acquired as well. Because of the possibility of constraint tying, the coefficients are no longer regular.

Multiple constraint rankings likewise have additional features that need to be acquired, such as the maximal number of violations. This occurs regardless of whether the speaker is merely tallying, or actually calculating the number that is useful. This increases the numerical size of the grammar but reduces the number of variables that need to be manipulated. As we've seen, trading off features of UG and additional complexity in acquisition may lead to models of grammar that are ultimately easier to manipulate once learned.

The stochastic model presents the greatest challenge for learning. I assume here that the p_j values for the probability of a constraint varying begin with a value of zero. Before the

variation can be considered the constraint ranking must be established. If we assume that irregularities are established after regular behaviours, then it is clear that once the constraint ranking is established, the p_j values can be adjusted where needed to account for nuances. I assume for the moment that the probabilities would be adjusted via Bayesian principles, and if they are established only after the constraint ranking, it is reasonable to predict that this portion of the grammar may be adjustable over time, even while the constraint ranking itself remains fixed.

An alternative approach to the stochastic model is that the stochastic portion is the source of probabilistic behaviour, and that these probabilities diverge from zero very early, only to have the constraint ranking imposed upon a purely probabilistic model at a later date. More research into language acquisition will have to be done to determine which of these is a more accurate model of learning behaviour. Without the constraint ranking, however, the grammar is no longer Optimality Theoretic.

3.3. Multiple constraint rankings

As we mentioned in the discussion of the model of strict OT, multiple constraint rankings are possible for a given language. The theory tells us that constraints must be ranked, but that constraints that don't interact in a given language may be ranked in one order in one speaker's grammar, but ranked in a slightly different order in another's. A model that permits tied constraints, as described in this paper, does not require non-interacting constraints to be tied. Such a requirement would help reduce the size of the grammar and reduce or eliminate differences in the grammars across speakers of a single language. More than these minor variations, however, it may be also be possible to produce identical linguistic outputs but appealing to very different constraint interactions (McCall, 2002). These models do not make any predictions about how this might occur or how the utility values may differ. However, it should be possible to test in each case how accurate the predictions of each model are by conducting experimental studies in the field, and modeling the behaviour of each model to determine which values for p_j work best, and which models match the study's behaviour most closely. If it can be shown by these or other means that multiple rankings exist, the notion of language change through constraint reranking becomes, at the least, more complex than currently envisioned.

3.4. Linearity and nonlinearity in OT

The mathematical models described here also suggest another feature of Optimality Theory, which is a strong linear quality. While there is some allowance for constraint conjunction, the variables for the conjoined constraints are also zero or one. Gradience effects, while linear in individual constraints are the first suggestion of possible nonlinearity in Optimality Theory when we begin to consider conjoining them. However, nonlinearities are concealed in OT in the guise of output-output faithfulness constraints and sympathy theory. Sympathy theory, in particular, is language specific, and amounts to a clever way of masking constraint interaction. As we have seen from the discussion of gradience constraints that gradience, as difficult as it is for Optimality Theory, comes with certain advantages, one of these being to reduce the overall number of possible constraints. Likewise, by permitting more complicated interactions among constraints, further reductions may be possible, at the cost of additional complexity in the model.

3.5.Dynamics

Optimality Theory postulates two functions, EVAL and GEN. Most of this paper has been dedicated to discussing the EVAL function. However, the analysis of the EVAL function may bear directly upon an analysis of the GEN function in OT. GEN is the function which generates candidates for EVAL to evaluate, and it is usually seen as generating an infinite number of candidates which EVAL considers in parallel. However, a human brain cannot, in fact, evaluate an infinite number of candidates simultaneously. This is another problem that has been referred to as the problem of the infinities. A mathematical model of EVAL predicts that there will be some minimal utility value that can be a winning candidate. By allowing the two functions to interact, we can make GEN more efficient, and more difficult to modify once the grammar has been established. By preventing GEN from providing candidates to EVAL that have no chance of succeeding, such a model may provide another explanation for why second language acquisition is so difficult, since GEN may not be capable of even supplying winning candidates.

The stochastic model also has something to say about language dynamics over a speaker's lifetime, and for language change. If the value of p_j is adjusted in a Bayesian fashion over the life of the speaker, changes in the linguistic environment can be learned and the language of the speaker adjusted, even while the constraint ranking for that speaker remains fixed. Within a limited domain, new speakers might perceive the constraint ranking as already adjusted—even when variation still exists. We need only have some $p_j>0.5$ to cause a change in the constraint ranking, since in language acquisition we assume that p_j would be adjusted upwards from zero.

4. Conclusions and future research

One can see that the mathematical models of Optimality Theory described here show in detail some of the theoretical consequences that variations on a basic theme can have. A strict model of OT has benefits that arise from its simplicity, but it forces grammars under the assumption of UG to be extremely large in relation to other models. The stochastic version of Optimality Theory shows the greatest promise for maintaining behavioural features of other models, and still being capable of adding new features to tackle linguistic variation across speakers, within a speaker's grammar over time and through the process of language change. Such mathematical models in general provide a concrete means of constraining aspects of the theory and using OT in other fields of language modeling such as natural language processing and producing simulations of studies to better determine whether the model proposed can actually produce the observed behaviours. Furthermore, the models help us see best where theoretical tools such as sympathy theory and other features introduce nonlinearities into a model that is otherwise very linear. This allows us to begin asking questions about these features if they do not point in a new direction for linguistic theory beyond OT.

References

Antilla, A 1995 Deriving Variation from Grammar: a study of Finnish genitives. ROA-63

Archangeli D & Langendoen DT eds. 1997 Optimality Theory: an Overview (Blackwell: Malden, MA)

Berger JO 1985 Statistical Decision Theory and Bayesian Analysis, 2nd ed (Springer: New York)

Boersma P & Escudero P 2002 Optimality-Theoretic modeling of microvariation in phonological perception and production

Bonilha G 2002 Conjoined Constraints and Phonological Acquisition ROA-533-0802
Chernoff H & Moses LE 1959 Elementary Decision Theory (Dover Publications: New York)
Goldsmith JA 1995 The Handbook of Phonological Theory (Blackwell: Malden, MA)
Hammond M 2000 The Logic of Optimality Theory ROA-390-0400
Kleinbaum DG, Kupper LL, Muller KE & Nizam A 1998 Applied Regression Analysis and Other Multivariable Methods (Duxbury Press: New York)
McCall B 2002 Solving Palatalization in Japanese Mimetics, ms.
McCarthy J 2002 Against Gradience ROA-510-0302
Prince A & Smolenksy P 1993 Optimality Theory: Constraint Interaction in Generative Grammar ROA-537-0802