

Chapter 3

Formal Approaches to SA-OT

The goal of this chapter is to underpin the *Simulated Annealing Optimality Theory Algorithm* (Fig. 2.8) in general, and the definition (2.17) of the *transition probability* $P(w \rightarrow w' | T)$, that is, the *Rules of moving* (page 63) in particular. We demonstrate that this definition follows directly from the *Strict Domination Hypothesis*, which constitutes the basis of OT. Therefore, we introduce several formal representations of Optimality Theory. Each of them is first proved to realise the *Strict Domination Hypothesis*, and then to lead to the same definition of temperature and to the same transition probabilities $P(w \rightarrow w' | T)$ —independently of each other. Before introducing different representations (real numbers in section 3.2, followed by polynomials in section 3.3, and finally ordinal numbers in section 3.4), however, we have to formally define what a representation is in Optimality Theory (section 3.1). Fig. 3.1 on page 85 might serve to the reader as a road map to the present chapter.¹

As the formal models developed here result in the same algorithm, latter chapters, as well as further implementations of SA-OT, can certainly be understood without the mathematically demanding details of the present chapter.

3.1 Towards a formal definition of OT

The *violation profile* as introduced by Prince and Smolensky (2004) (Prince and Smolensky, 1993) is a list of violation marks—or, rather, a set of tokens of violation marks. The *Harmony function* is the mapping that assigns a violation profile to a candidate. Nonetheless, Prince and Smolensky’s “list of violation marks” is a less convenient construction. Therefore, here we (re-)introduce the *vector representation* of a *violation profile*, a straightforward translation (and generalisation) of Prince and Smolensky’s idea. We repeat the idea already presented in section 2.2.3, and elaborate more on this approach. Then, we consider it as the standard for introducing two further representations: polynomials and ordinal numbers.

¹A summary of the present chapter has been published as Bíró (2005b).

As our starting point, we are given GEN, a mapping from the set of possible underlying representations to the set of possible candidates. Let $GEN(UR)$ denote the *set of candidates* corresponding to a specific underlying representation UR . The set \mathcal{PC} of *all possible candidates* is the union of $GEN(UR)$ for all possible UR .

3.1.1 Constraints

Let $C_i(w)$ be the number of times candidate w violates constraint C_i . In general, we shall call $C_i(w)$ the *level of violation* incurred by candidate w with respect to constraint C_i , and in specific models we may speak of the *number of violation marks* assigned by the constraint.

Indeed, C_i most often takes non-negative integer values in linguistic practice, conform to the original idea of a “list of violation marks” in Prince and Smolensky (1993). Yet, here we generalise the concept:

Definition 3.1.1. *Constraint C_i is a function on the set \mathcal{PC} of all possible candidates, such that for each possible UR : the set $\{C_i(w) \mid w \in GEN(UR)\}$ (the image of the candidate set corresponding to UR) is a totally ordered set with some ordering relation $\mathcal{R}_{i,UR}$, and any of its subsets has a lower bound contained by the subset.²*

Notice that different constraints may have different ranges. Moreover, the same constraint with different underlying representations could also have very different ranges in theory. Moreover, the requirement of a lower bound is only important to ensure that an optimal form always exists. The set of non-negative integer values with the simple *greater than* relation used in practice for C_i clearly satisfies all our requirements.

Although it has been already mentioned, it may be useful to repeat here:

Definition 3.1.2. *Let S be a set, and $>$ a binary relation on S .³ Then, the pair $(S, >)$ is a totally (fully) ordered set, iff:*

1. *The law of trichotomy: for all $x, y \in S$ exactly one of the following three statements holds: 1. $x > y$, 2. or $y > x$, 3. or $x = y$.*
2. *Transitivity: for all $x, y, z \in S$, if $x > y$ and $y > z$ then $x > z$.*

3.1.2 Hierarchies

First, let us introduce the notion of *isomorphism*.⁴

Definition 3.1.3. *The totally ordered sets $(A, <)$ and (B, \prec) are ORDER ISOMORPHIC, iff there is a bijection⁵ f from A to B such that for all $a_1, a_2 \in A$, $a_1 < a_2$ iff $f(a_1) \prec f(a_2)$.*

²In brief, the set $\{C_i(w) \mid w \in UR\}$ is well-ordered with the relation $\mathcal{R}_{i,UR}$.

³That is, $<$ is a subset of $S \times S$.

⁴Cf. e.g. Eric W. Weisstein: “Bijection”, from MathWorld—A Wolfram Web Resource, <http://mathworld.wolfram.com/Bijection.html>; Holz et al. (1999, p. 11).

⁵A bijection f is “a transformation which is one-to-one and onto”. That is, its domain covers A and its inverse f^{-1} is also a function ($f(a_1) = f(a_2)$ iff $a_1 = a_2$) whose domain covers the entire set B .

In other words, the *isomorphism* f translates the order $<$ on set A into the order $<$ on set B .

The subsequent key concept in OT is a *constraint hierarchy*:

Definition 3.1.4. A CONSTRAINT HIERARCHY \mathcal{H} , is a finite set of totally ordered constraints $\{C_N, C_{N-1}, \dots, C_1, C_0\}$ with an ordering relation \gg .

Any two totally ordered sets with finite k elements (for any nonnegative integer k) are order isomorphic.⁶ Consequently, the above hierarchy \mathcal{H} is order isomorphic to the ordered set $(N, N - 1, \dots, 0)$, so it can be easily represented as a vector:

$$\mathcal{H} = (C_N, C_{N-1}, \dots, C_1, C_0) \quad (3.1)$$

In turn, the VECTOR REPRESENTATION of the *Harmony function* of a candidate w with respect to a constraint hierarchy $\mathcal{H} = (C_N, C_{N-1}, \dots, C_1, C_0)$ is defined as the vector

$$H_{\mathcal{H}}(w) = (C_N(w), C_{N-1}(w), \dots, C_1(w), C_0(w)) \quad (3.2)$$

In subsection 2.2.3 (equation (2.10)), we already saw that this vector representation is but a shorthand for a row in a traditional tableau.

Most frequently the hierarchy will be constant, so the subscript \mathcal{H} may be left out. Notice that the subscripts of the constraints are written in a decreasing order: this minor inconvenience at this point will help us later in keeping our notations simple.

The way Prince and Smolensky's original "list of violation marks" can be translated into this vector representation is straightforward: first, if the list of violation marks incurred by candidate w contains n tokens of violation marks $*C_i$, then let $C_i(w) = n$; second, the ranking of the constraints can simply be mapped onto a vector using the order isomorphism. Therefore, this representation of a violation profile will serve as the formalisation of standard Optimality Theory.

3.1.3 An order on violation profile-like vectors

Our goal is to formulate the central idea of Optimality Theory in (3.3) that says that the surface representation is the candidate that maximises the Harmony function. Therefore, our next step is to define an order between two *violation profile-like vectors*. First we introduce

Definition 3.1.5. A VIOLATION PROFILE-LIKE VECTOR with respect to underlying form UR is an element of the following Cartesian product:

$$Range_{UR}(C_N) \times Range_{UR}(C_{N-1}) \times \dots \times Range_{UR}(C_0)$$

Here, $Range_{UR}(C_i) = \{C_i(w) \mid w \in GEN(UR)\}$. We shall, however, omit the reference to UR for the sake of simplicity.

Now, we define an order \succ on two, violation profile-like vectors:

⁶Stated for instance in the MathWorld of Eric W. Weisstein. "Ordinal Number" at <http://mathworld.wolfram.com/OrdinalNumber.html>.

Definition 3.1.6. Let $A = (a_N, a_{N-1}, \dots, a_0)$ and $B = (b_N, b_{N-1}, \dots, b_0)$ be two violation profile-like vectors (with respect to the same UR). Then, A is MORE HARMONIC THAN B , $A \succ B$ if and only if there is an integer $k \in \{N, N-1, \dots, 1, 0\}$ such that

1. $a_k < b_k$
2. and for all $i \in \{N, N-1, \dots, 1, 0\}$: if $i > k$ then $a_i = b_i$

Moreover, $A = B$ iff for all $k \in \{N, N-1, \dots, 1, 0\}$, $a_k = b_k$.

The set $\{N, N-1, \dots, 1, 0\}$, which is going to recur very frequently, could have been replaced by an arbitrary finite set of indices \mathcal{I} with some order $>$. This more general and shorter notation is however equivalent to the more transparent notation we use, since, as noted, any two totally ordered finite sets of equal cardinality are order isomorphic.

We may call C_k the *critical*⁷ or *fatal constraint*: this is the constraint that determines the relative harmony of two violation profile-like vectors. This definition of the binary relation \succ may be also called *lexicographic ordering* (Eisner, 2000b).

It may be confusing that the *more harmonic than* relation has an opposite direction to the *bigger than* relation on the number of violation marks: *more harmonic* ($A \succ B$) corresponds to *fewer marks* ($a_k < b_k$). In the following sections, the rule is that the Harmony function is to be optimised or maximised, whereas the energy function (cost function, the violation marks) minimised.

In what follows, we demonstrate that relation \succ is a total order on any set of violation profile-like vectors: that is, both trichotomy and transitivity hold.

Theorem 3.1.7. TRANSITIVITY: Suppose that $A = (a_N, a_{N-1}, \dots, a_0)$, $B = (b_N, b_{N-1}, \dots, b_0)$ and $C = (c_N, c_{N-1}, \dots, c_0)$ are violation-like vectors (with respect to the same UR). If $A \succ B$ and $B \succ C$, then also $A \succ C$ holds.

Proof. Suppose that $A \succ B$ with C_k being the crucial (fatal) constraint ($a_k < b_k$; for all $i > k$: $a_i = b_i$). Furthermore, suppose $B \succ C$ with C_l as crucial constraint ($b_l < c_l$; for all $i > l$: $b_i = c_i$). Now, we have to demonstrate that $A \succ C$. Let us distinguish between three cases: $l > k$, $l = k$ and $l < k$. If $l > k$, then, by the definition of \succ , $a_l = b_l < c_l$, and for all $i > l > k$: $a_i = b_i = c_i$. Thus, $A \succ C$, and the crucial constraint is C_l . Secondly, if $l = k$, then $a_l = a_k < b_k = b_l < c_l$, and for all $i > l = k$: $a_i = b_i = c_i$. Again, $A \succ C$, with the crucial constraint being $C_l = C_k$. Finally, whenever $l < k$, $a_k < b_k = c_k$, and $a_i = b_i = c_i$ for all $i > k > l$. In this case, $A \succ C$ because the crucial constraint is C_k . \square

Theorem 3.1.8. TRICHOTOMY: Suppose that $A = (a_N, a_{N-1}, \dots, a_0)$ and $B = (b_N, b_{N-1}, \dots, b_0)$ are violation-like vectors (with respect to the same UR). Then, exactly one of the following three relations hold:

- $A \succ B$
- $B \succ A$

⁷This notion of critical constraint should not be confused with the *critical cut-off point* in Coetzee (2004)'s proposal (cf. section 1.3.2).

- $A = B$

Proof. Suppose that $A \neq B$. By the last part of definition 3.1.6, two vectors are not equal if at least one of their components is different. Take the set $S = \{i \in \{N, N-1, \dots, 1, 0\} \mid a_i \neq b_i\}$. Observe that S is a finite set, thus it has a maximum k , and therefore contains it: $k = \max(S) \in S$. We are demonstrating now that C_k is the crucial constraint. Because k is the maximum of S , it is true that for all $i \in [N, \dots, 1, 0]$: if $i > k$ then i is not in S , so $a_i = b_i$. As for the first requirement in the definition of \succ : because $k \in S$, either $a_k > b_k$ or $a_k < b_k$. (Note that here becomes important that the range of the constraints are also fully ranked sets.) In the first case $B \succ A$, and in the second case $A \succ B$. \square

In sum, we have shown that any set of violation profile-like vectors are totally ordered with respect to the relation \succ . Now, we demonstrate that any set of violation-like vectors has a minimum, and also contains it:

Theorem 3.1.9. THE MAXIMUM-THEOREM ON VIOLATION PROFILE-LIKE VECTORS: *Let S be a non-empty set of violation-like vectors (with respect to the same UR). Then, there is exactly one violation-like vector $A_0 = \max(S)$ such that: 1. $A_0 \in S$; and 2. for all $A \in S$, if $A_0 \neq A$ then $A_0 \succ A$.*

Proof. We shall find A_0 the same way as a linguist finds the best candidate in a tableau.

Let $S_{N+1} = S$. Further, for all $i \in \{N, N-1, \dots, 0\}$: suppose that $m_i = \min\{w_i \mid W \in S_{i+1}\}$ and $S_i = \{W \in S_{i+1} \mid w_i = m_i\}$, where we use the abbreviation $W = (w_N, w_{N-1}, \dots, w_0)$. In other words, m_i is the lowest violation level for constraint C_i attested among the elements of S_{i+1} ; whereas S_i is the subset of S_{i+1} containing the elements which have exactly violation level m_i for constraint C_i . Observe that the definition of m_i makes crucially reference to a property in Definition 3.1.1 of a constraint: a subset of $\text{Range}(C_i)$ (here $\{w_i \mid W \in S_{i+1}\}$) always has a lower bound. Not only does it have a lower bound, but the subset also contains its bound. Consequently, there is at least one $W \in S_{i+1}$ such that $w_i = m_i$, which is why S_i is not empty. In brief, S_i is the set of violation profile-like vectors that have “survived” the filtering effect of constraint C_i .

Now, we show that S_0 has exactly one element. First, we have just seen that S_0 is not empty, similarly to all S_i s. Second, suppose both $A \in S_0$ and $B \in S_0$. Then $A \in S_0 \subseteq S_1 \subseteq \dots \subseteq S_i$, that is, $a_i = m_i$, for all $i \in \{N, N-1, \dots, 0\}$. Similarly, $b_i = m_i$, by definition of S_i . Consequently, all components of A and B are equal, that is, by definition 3.1.6, $A = B$.

Last, we show that the only element A of S_0 is the minimum predicted by the theorem. Clearly, $A \in S_0 \subseteq S_{N+1} = S$. Moreover, take any $B \in S$ that is different from A (hence, not a member of S_0). Let k be such that B is an element of S_{k+1} , but not an element of S_k . Such a k exists, because $S_0 \subseteq S_1 \subseteq \dots \subseteq S_{N+1} = S$, and B is element of S , but not of S_0 . Now, as $A \in S_0 \subseteq S_k$, $a_k = m_k < b_k$, by definition of m_k . Yet, for all $i > k$, both A and B are in S_i : in other words, $a_i = m_i = b_i$. Therefore, we have demonstrated, by definition 3.1.6, that $A \succ B$. In sum, the only element A of S_0 is a maximal element of S .

Finally, S cannot have two different maximal elements. Suppose that both A_1 and A_2 were maximal elements. Because A_1 is a maximal element, and A_2 is

a different element of S , then $A_1 \succ A_2$. Similarly, $A_2 \succ A_1$ should hold, which contradicts the law of trichotomy (theorem 3.1.8). \square

3.1.4 Comparing candidates

The following definition follows closely the proposal of Prince and Smolensky, also called *strict domination*:

Definition 3.1.10. *For a given hierarchy $\mathcal{H} = (C_N, C_{N-1}, \dots, C_1, C_0)$ and candidates w_1 and w_2 , w_1 is MORE HARMONIC THAN w_2 , or $w_1 \succ_{\mathcal{H}} w_2$, if and only if there is an integer $k \in \{N, N-1, \dots, 0\}$ such that*

1. $C_k(w_1) < C_k(w_2)$
2. and for all $i \in \{N, N-1, \dots, 1, 0\}$: if $i > k$ then $C_i(w_1) = C_i(w_2)$

Moreover, two candidates w_1 and w_2 are EQUIVALENT, $w_1 \simeq w_2$ if and only if for all $i \in \{N, N-1, \dots, 1, 0\}$: $C_i(w_1) = C_i(w_2)$.

The reference to the hierarchy \mathcal{H} may be omitted whenever obvious. The expression *strict domination* refers to a very important property of this definition: if a candidate meets its Waterloo at a given constraint, it can never come back to the battle field. Even by satisfying all lower ranked constraints, behaving with respect to them much better than all surviving candidates, it is definitely defeated.

Observe the following properties:

Corollary 3.1.11. *The relation \simeq is an equivalence relation on the set of candidates. That is, if w_1, w_2 and w_3 are candidates, then*

1. $w_1 \simeq w_1$ (*reflexivity*)
2. $w_1 \simeq w_2$ iff $w_2 \simeq w_1$ (*symmetry*)
3. $w_1 \simeq w_2$ and $w_2 \simeq w_3$ the $w_1 \simeq w_3$ (*transitivity*)

Furthermore, for a given hierarchy \mathcal{H} , if $w_1 \simeq w_2$ and $w_2 \succ_{\mathcal{H}} w_3$ the $w_1 \succ_{\mathcal{H}} w_3$.

By comparing the definition 3.1.10 of the \succ and \simeq relations on candidates to the definition 3.1.6 of \succ and $=$ on violation profile-like vectors, we immediately see by equation (3.2) that

Corollary 3.1.12. **THE EQUIVALENCE OF STRICT DOMINATION AND VIOLATION PROFILES:** *The vector-representation of the violation profiles realises Prince and Smolensky's definition of strict domination:*

1. $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ if and only if $w_1 \succ_{\mathcal{H}} w_2$;
2. $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$ if $w_1 = w_2$;

Moreover, $w_1 \simeq w_2$ if $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$.

In other words, the function $H_{\mathcal{H}}$ is a *homomorphism* (Holz et al., 1999, p. 10-11) with respect to relations $\succ_{\mathcal{H}}$ on the set of candidates and \succ on the set of violation profile-like vectors.

In the following two subsections, we shall demonstrate that the alternative representations to be proposed are also equivalent to the violation profiles, hence, to strict domination.

Now, relation \succ is almost a total ordering on the candidate set corresponding to a given underlying representation UR . Transitivity holds, as a consequence of the transitivity on the set of violation profile-like vectors. Yet, the Law of Trichotomy only holds in a weaker modified version: whenever neither $w_1 \succ w_2$ nor $w_2 \succ w_1$, then w_1 and w_2 are *equivalent* ($w_1 \simeq w_2$), that is, they incur the same violation level by all constraints ($H(w_1) = H(w_2)$). To prove it, one has to apply the law of trichotomy on violation profile-like vectors to the vectors $H(w_1)$ and $H(w_2)$, and use corollary 3.1.12.

Similarly, the consequence of the *Maximum-theorem on violation profile-like vectors* is the following:

Theorem 3.1.13. THE MAXIMUM-THEOREM ON CANDIDATES: *Let S be a set of candidates (corresponding to the same UR). Then, for a given hierarchy of constraints \mathcal{H} , S has a unique subset $S_0 = \min_{\mathcal{H}}(S) \subseteq S$ such that*

1. *if $w_1 \in S_0$ and $w_2 \in S_0$, then $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$;*
2. *if $w_1 \in S_0$ and $w_3 \in S \setminus S_0$, then $w_1 \succ_{\mathcal{H}} w_3$.*

Proof. To prove this statement, one has to apply the *Maximum-theorem on violation profile-like vectors* to the set $\{H_{\mathcal{H}}(w) \mid w \in S\}$. This is a set of violation profile-like vectors, and has exactly one maximal element A_0 . Now, the maximum subset S_0 of S is formed by the elements $w \in S$ such that $H_{\mathcal{H}}(w) = A_0 = \min\{H_{\mathcal{H}}(w) \mid w \in S\}$. In other words: $S_0 = \operatorname{argmin}_{w \in S}(H_{\mathcal{H}}(w))$.

Set S_0 is not empty, because the maximal element $A_0 \in \{H_{\mathcal{H}}(w) \mid w \in S\}$, that is, for at least one $w \in S$, $H_{\mathcal{H}}(w) = A_0$. If both w_1 and $w_2 \in S_0$, then $H_{\mathcal{H}}(w_1) = A_0 = H_{\mathcal{H}}(w_2)$. Finally, if $w_1 \in S_0$ and $w_3 \in S \setminus S_0$: $H_{\mathcal{H}}(w_3) \in \{H_{\mathcal{H}}(w) \mid w \in S\}$, but $H_{\mathcal{H}}(w_3) \neq A_0$ (otherwise, $w_3 \in S_0$), thus $H_{\mathcal{H}}(w_1) = A_0 \succ_{\mathcal{H}} H_{\mathcal{H}}(w_3)$. Then, by corollary 3.1.12, $w_1 \succ_{\mathcal{H}} w_3$.

Finally, we show that the maximum subset S_0 is unique. Namely, suppose that two such maximum subsets, S_0 and S'_0 exist at the same time. If the two subsets are different, then there exist an element $w \in S$ such that either $w \in S'_0$ and $w \notin S_0$, or $w \notin S'_0$ and $w \in S_0$. As the two cases are symmetrical, let us take the former case. Then, $w \in S \setminus S_0$, hence $w_1 \succ_{\mathcal{H}} w$ for any $w_1 \in S_0$. As S'_0 is also a maximum subset, all its elements $w_2 \in S'_0$ are equivalent to w ($H_{\mathcal{H}}(w) = H_{\mathcal{H}}(w_2)$). Consequently, for any $w_1 \in S_0$ and $w_2 \in S'_0$, it is true that $w_1 \succ_{\mathcal{H}} w_2$, and therefore S_0 and S'_0 are disjoint sets (by the Law of Trichotomy). Furthermore, S'_0 cannot be a maximum subset, for its elements are less harmonic than some elements of $w \in S \setminus S'_0$, namely, the elements of S_0 , which fact would contradict the Law of Trichotomy. \square

3.1.5 The definition of Optimality Theory

Finally, we are able to formulate what Optimality Theory is about, and see the soundness of this formulation.

Remember that GEN is a function that maps each underlying representation (UR) to a set of candidates. The central idea of Optimality Theory is that the surface representation is the optimal (maximal) candidate of the candidate set with respect to an ordering defined by the given hierarchy:

$$\begin{aligned} SR &= \max_{\mathcal{H}}(GEN(UR)) = \\ &= \operatorname{argmax}_{w \in GEN(UR)} H_{\mathcal{H}}(w) \end{aligned} \quad (3.3)$$

The first line of this definition makes sense because of theorem 3.1.13, and is equal to the second line by corollary 3.1.12.

In words: the surface representation(s) maximise(s) the Harmony function. Whenever more candidates $w \in GEN(UR)$ maximise $H(\cdot)$, all of these candidates are predicted to appear as a grammatical form on the surface (Prince and Smolensky (2004) p. 82): this is the approach mentioned in section 1.3.1.

Sometimes, the candidates include information not present in the overt linguistic form, such as parsing brackets. Nevertheless, the surface form can be readily arrived at from the winning candidate by a simple function (e.g. by erasing these brackets). Note that because the inverse of this transformation is not always a function, and more candidates can correspond to the same form, learning algorithms face extra difficulties. Tesar and Smolensky (2000) propose using *Robust Interpretive Parsing* in order to decide which candidate corresponding to the overt learning data form to employ in the learning algorithm.

3.1.6 Realisations of the Harmony function

Subsection 2.2.3 showed how Prince and Smolensky (2004)'s concept of a set of violation mark tokens can be translated into the *vector representation* of the Harmony function. There, we referred crucially to Prince and Smolensky's *Cancellation/Domination Lemma*. The present subsection has demonstrated formally that this representation makes sense, and that the formulation of an OT grammar as equation (3.3) is well-founded.

In the following subsections, we introduce two new representations of the Harmony function. We do that in order to carry out again the agenda of introducing SA-OT:

- Firstly, we represent the violation profiles in an appropriate way.
- Secondly, we define the *difference* of two violation profiles.
- Thirdly, we define *temperature* in a similar format.
- Fourthly, we define the exponential of their quotient.
- Lastly, we introduce the SA-OT algorithm.

Before launching this program, however, let us define what an appropriate representation of a violation profile is. Corollary 3.1.12 has already stated the (almost) equivalence of the ranking \succ on the candidate set and the ranking \succ on the set of vector representation of the candidates' violation profiles. Thus, the vector representation of a violation profile is a typical example of *order isomorphism* as introduced in definition 3.1.3.

To be more exact, based on corollary 3.1.11, we introduce the *set of violation profiles*, which is the set of equivalence classes on the set of candidates with respect to equivalence relation \simeq . Again by corollary 3.1.11, but by its second part this time, a total order $\succ_{\mathcal{H}}$ on the violation profiles may be introduced: an equivalence class is *more harmonic* than another equivalence class, if some element of the first class is more harmonic than some element of the second class (by definition 3.1.10). Now, the set of violation profiles (equivalence classes on the set of candidates) with order $\succ_{\mathcal{H}}$ is *order isomorphic* to the set $\{H_{\mathcal{H}}(w) | w \text{ is a candidate}\}$ with the order \succ on violation profile-like vectors (def. 3.1.6).

The new representations of the Harmony function must be isomorphic to the set of the violation profiles, too. Only this can ensure that the new representations will yield the grammar defined by equation 3.3. In other words:

Definition 3.1.14. A REALISATION of the Harmony function $H(w)$ is a mapping $E(w) : \mathcal{PC} \rightarrow X$ (from the set of all possible candidates to some set X), such that:

- a total ordering relation \prec and an equivalence relation $=$ is defined on the set X ;
- for all candidates w_1 and w_2 : $H(w_1) \succ H(w_2)$ iff $E(w_1) \prec E(w_2)$;
- for all candidates w_1 and w_2 : $H(w_1) = H(w_2)$ iff $E(w_1) = E(w_2)$.

Observe that the new representation E is compared to the vector representation H , which can be done because isomorphy is a transitive relation between ordered sets. We know that the set of violation profiles is order isomorphic to the set of vector representations; hence, a new representation is isomorphic to the set of violation profiles if and only if it is isomorphic to their vector representation.

Besides being clearer, an additional advantage of using the vector representation as the starting point—as opposed to a set of violation mark tokens—is that definition 3.1.1 allows for more flexibility concerning the range of the constraints.⁸

Note that the ordering will be reversed: while the Harmony function H is to be maximised, the new representations—seen as energy or cost function—will be minimised.⁹ The advantages of reversing the \succ relation are manifold. It is simpler to derive formally the new representations from the constraints seen as non-negative valued functions in a way that results in this reversed relation. Intuitively, the minimisation approach parallels better the idea of minimising the violation marks—that is, the punishment symbols. Moreover, simulated annealing is traditionally formulated for minimising the cost function. Observe that subsection 2.2.3 already defined the difference of two violation profiles so that it is positive if less violation marks is subtracted from more violation marks:

⁸Most often in practice, violation levels are non-negative integers. Nonetheless, equation (4.8) introduces a constraint whose possible range is $\mathbb{N}_0 + z \cdot \mathbb{N}_0$ with $z \in \mathbb{R}$ —a different set, which still meets the requirements of definition 3.1.1. Note that the polynomial approach will allow for this more general type of constraints, whereas the ordinal number approach requires this set be mapped by an isomorphism to the set of non-negative integers.

⁹Adding a minus sign would be possible in the case of the polynomial representation, but not possible for the ordinal numbers. Furthermore, maximising negative numbers is probably less intuitive than minimising positive numbers.

that is, the goal was there also to minimise the violation marks. Similarly, even definition 3.1.6 would be simpler (the usual definition of *lexicographic ordering*) if we reversed the \succ sign. Yet, there the motivation was to follow the OT concept of Harmony maximisation.

Fig. 3.1 summarises the different levels of representations.

Definition 3.1.14 requires us to add new items on our agenda. In turn, the introduction of the two new representations will follow this agenda:

1. Introduce the representation $E(w)$.
2. Define the relations \prec and $=$ on the range of E .
3. Prove that for all candidates w_1 and w_2 : $H(w_1) \succeq H(w_2)$ if and only if $E(w_1) \preceq E(w_2)$.¹⁰
4. Define the *difference* of two violation profiles.
5. Define *temperature* in a similar format.
6. Define the exponential of their quotient.
7. Introduce the SA-OT algorithm.

3.2 Violation profiles as real numbers

In what follows, we introduce two further representations of a violation profile. The goal of both approaches is to interpret Eq. (2.2) in the context of Optimality Theory, and thereby to implement simulated annealing.

Surprisingly or not, both approaches will result in the SA-OT algorithm already presented in Figure 2.8.

As an introduction, we try to realise violation profiles as real numbers. If it worked, SA-OT could be implemented as a real-valued optimisation problem. As it does not, we will have to proceed to the realisations using polynomials and ordinal numbers.¹¹ Yet, it is educative to understand why $H_{\mathcal{H}}(w)$ cannot

¹⁰We shall also demonstrate the law of trichotomy on the range of E , and therefore the two latter points of definition 3.1.14 can be summarised as above.

¹¹Gerhard Jäger pointed out that it is possible to define order preserving mappings from violation profiles into the real numbers. Their applicability to SA-OT should be tested in the future, even if these functions do not preserve the magnitude of differences between violation vectors, as defined in the previous chapter.

Jäger's proposal is based on the fact that the function $f(x) = 1 - (x+2)^{-1}$ is order preserving on non-negative reals and maps all positive numbers into the interval $(0, 1)$. Therefore, he proposes the following recursive definition, supposing that the violation levels $C_i(w)$ are non-negative integers:

$$\begin{aligned} g_0(w) &= C_0(w) \\ g_{i+1}(w) &= C_{i+1}(w) + f(g_i(w)) \\ E(w) &= g_n(w) \end{aligned}$$

A similar solution is $E(w) = \sum_{i=0}^n f_i(C_i(w))$, where $f_i(x) = 2^{2^i} - \frac{2^{2^i}}{x+1}$. Observe that $f_i(x)$ grows monotonously, $f_i(0) = 0$, $f_i(1) = 2^{2^i-1}$ and $\lim_{x \rightarrow \infty} f_i(x) = 2^{2^i}$, whence it is easy to show that constraints ranked lower than C_k can never accumulate a sum larger than the weight of a single violation of constraint C_k , for $\sum_{i=0}^{k-1} 2^{2^i} = \frac{4^k - 1}{4 - 1} > 2^{2^k - 1}$ if $k > 0$.

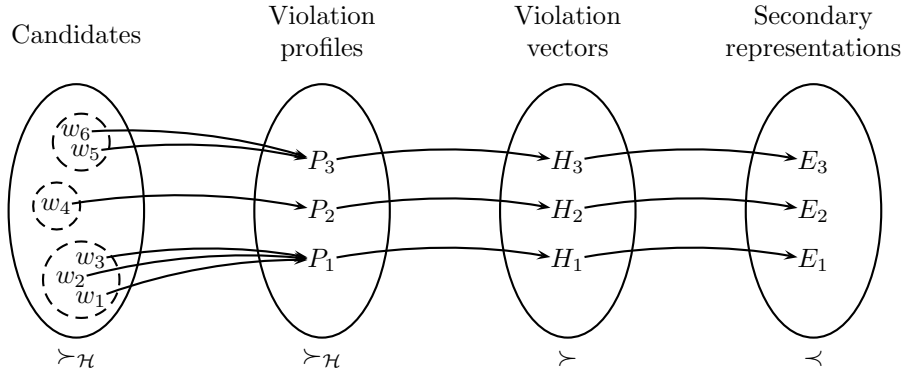


Figure 3.1: **Different levels of representation:** A candidate w incurs a certain number of violation marks from each of the constraints. Its violation profile $P(w)$ can be seen as a set of tokens of these marks. Given a certain hierarchy \mathcal{H} , the violation profile $P(w)$ corresponds to a vector $H_{\mathcal{H}}(w)$, introduced by equation (3.2). Finally, $H_{\mathcal{H}}(w)$ will be translated to different secondary representations $E_{\mathcal{H}}(w)$.

Definition 3.1.10 has introduced the order $\succ_{\mathcal{H}}$ and the equivalence relation \simeq on the candidate set. The relation \simeq defines equivalence classes on the candidate set (the dashed circles on the figure), which are then identified with the violation profiles: all elements of such an equivalence class have the same profile. Relation \simeq depends only on the definition of the constraints, which is universal, similarly to the candidate set. Thus, the set of violation profiles is also universal. The language dependent hierarchy \mathcal{H} determines the ranking $\succ_{\mathcal{H}}$ both on the candidate set and on the set of violation profiles.

This distribution of tasks is reversed in the right half of the figure. Different hierarchies map the same profile to different vectors. Therefore, different languages involve different subsets of the set of all possible vectors (that is, \mathbb{N}_0^{N+1} if the violation levels of the $N + 1$ constraints are the non-negative integers). The total order relation \succ used is however universal, as introduced in definition 3.1.6.

The secondary representations to be introduced will behave similarly. Each vector H_i will be mapped by an isomorphism onto some E_i . The order \prec on the E_i s is universal, but the hierarchy \mathcal{H} determines the specific value $E_{\mathcal{H}}(w)$ associated with the candidate w .

be realised as a real number, and this train of thought will lead us in a natural way to the subsequent proposals.

In the present section, as well as in section 3.3 (but not in 3.4), we could suppose that the violation levels are non-negative real numbers ($C_i(w) \in \mathbb{R}_0^+$ for any w and i) from the point of view of the definitions. Some of the theorems will, however, require that they are non-negative integers—which requirement is met by most applications in practice.

As mentioned in section 3.1, a crucial feature of Optimality Theory is *strict domination*: if a candidate is suboptimal for a higher ranked constraint, it can never win, even if it satisfies the lower ranked constraints best. Losing a battle means definitely being out of the game. Prince and Smolensky (2004) present on page 236 why a harmonic function $H(w)$ satisfying strict domination cannot be realised with a real-valued function.

Suppose first that there exists an upper bound $q - 1 > 0$ on the violation level a candidate can incur: for all $i \in \{N, \dots, 1, 0\}$ and for all $w \in \text{GEN}(UR)$, $0 \leq C_i(w) \leq q - 1$. (Note that this is exactly the condition required by the finite state approach of Frank and Satta (1998).) In such a case, the following real-valued Energy function $E(w)$ realises the Harmony function $H(w)$ perfectly:¹²

$$E(w) = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + \dots + C_1(w) \cdot q + C_0(w) \quad (3.4)$$

Following definition 3.1.14, we mean by $E(w)$ *realising* $H(w)$ that for all w_1 and w_2 , $E(w_1) \leq E(w_2)$ if and only if $H(w_1) \succeq H(w_2)$. In other words, optimising the Harmony function is equivalent to minimising the Energy function.

Indeed, equation (3.4) assigns candidate w a number $E(w)$ in a number system of base q whose digits are the violation levels. Informally speaking, this observation already proves that $E(w)$ defined accordingly realises strict domination.

Formally, we demonstrate this fact in two steps.

Lemma 3.2.1. *Given a hierarchy \mathcal{H} ($C_N \gg \dots \gg C_1 \gg C_0$), suppose that some $q \in \mathbb{R}$ exists such that for all constraints C_i and for all candidates w , the inequality $0 \leq C_i(w) \leq q - 1$ holds. Moreover, let $C_i(w) \in \mathbb{N}_0$, while $E_{\mathcal{H}}(w) = \sum_{i=0}^N C_i(w) \cdot q^i$. Then, with the Harmony function $H_{\mathcal{H}}(w)$, as defined in (3.2), for any two candidates w_1 and w_2 : if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$, then*

$$E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2) .$$

Proof. Following the definition 3.1.6, if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$, then there exists a $k \in \{N, \dots, 0\}$ such that

$$E_{\mathcal{H}}(w_1) - E_{\mathcal{H}}(w_2) = \sum_{i=0}^k (C_i(w_1) - C_i(w_2)) \cdot q^i \quad (3.5)$$

¹²On page 61 we mentioned that the domains (the first component of the temperature, the indices of the constraints) are not necessarily consecutive integers, but can be arbitrary real numbers. Hence, a more general formulation of the real valued representation of a violation profile could be thus:

$$E(w) = \sum_{i \in \mathcal{I}} C_i(w) \cdot q^i$$

where \mathcal{I} is a finite set of real valued indices. We could but we shall not use this notation.

Moreover, $C_k(w_1) - C_k(w_2) < 0$. As the violation levels are integers,

$$C_k(w_1) - C_k(w_2) \leq -1 \quad (3.6)$$

Recall the sum of a geometric series:

$$\sum_{i=0}^{k-1} q^i = \frac{q^k - 1}{q - 1} \quad (3.7)$$

Therefore, and because 0 and $q - 1$ are lower and upper bounds on the number of violation marks ($C_i(w_1) - C_i(w_2) \leq q - 1$):

$$\sum_{i=0}^{k-1} (C_i(w_1) - C_i(w_2)) \cdot q^i \leq (q - 1) \frac{q^k - 1}{q - 1} \quad (3.8)$$

Summarising, from (3.5), (3.6) and (3.8), we obtain:

$$E_{\mathcal{H}}(w_1) - E_{\mathcal{H}}(w_2) \leq -q^k + q^k - 1 < 0 \quad (3.9)$$

That is, $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$. \square

Next, we can prove that the Harmony function can be realised with real numbers under some specific conditions:

Theorem 3.2.2. *Given a hierarchy \mathcal{H} ($C_N \gg \dots \gg C_1 \gg C_0$), suppose that some $q \in \mathbb{R}$ exists such that for all constraints C_i and for all candidates w , the inequality $0 \leq C_i(w) \leq q - 1$ holds. Moreover, $C_i(w) \in \mathbb{N}_0$. Then, the energy function*

$$E_{\mathcal{H}}(w) = \sum_{i=0}^N C_i(w) \cdot q^i$$

realises the Harmony function $H_{\mathcal{H}}(w)$, as defined in (3.2). That is, for any two candidates w_1 and w_2 ,

- $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$ iff $E_{\mathcal{H}}(w_1) = E_{\mathcal{H}}(w_2)$
- $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ iff $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$

Proof. This theorem comprises four statements. We have already demonstrated in lemma 3.2.1 that if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ then $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$.

Furthermore, if $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$ then $E_{\mathcal{H}}(w_1) = E_{\mathcal{H}}(w_2)$. Namely, due to the second part of definition 3.1.6, each coefficient in the definition of $E_{\mathcal{H}}(w_1)$ and of $E_{\mathcal{H}}(w_2)$ are equal.

Suppose now that $E_{\mathcal{H}}(w_1) = E_{\mathcal{H}}(w_2)$. By the law of trichotomy (theorem 3.1.8), either $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ or $H_{\mathcal{H}}(w_2) \succ H_{\mathcal{H}}(w_1)$ or $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$. We have already demonstrated that the first two possibilities would involve $E_{\mathcal{H}}(w_1) \neq E_{\mathcal{H}}(w_2)$, which leaves us the only possibility of $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$.

Similarly, suppose now that $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$. Because the law of trichotomy also applies on the set of real numbers with the usual $>$ relation, this supposition would be contradicted if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ did not hold, but one of the two other possibilities in theorem 3.1.8. \square

3.3 Violation profiles as polynomials

However, nothing in the general theory of Optimality Theory guarantees that such an upper bound $q - 1$ exists.¹³ The behaviour of an energy function (3.4) with some q only approximates the behaviour of the Harmony function.¹⁴

Then, why not consider the behaviour of this polynomial as q goes to infinity?¹⁵ We propose to see the violation profiles as a polynomials of $q \in \mathcal{R}^+$ ($q > 0$):¹⁶

$$E(w)[q] = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + \dots + C_1(w) \cdot q + C_0(w) \quad (3.10)$$

This equation defines the *polynomial representation* of a violation profile: each candidate, or each violation profile is realised as a real-valued polynomial of q . The energy or the Eval-function $E(w)$ is not any more a real number, but a function mapping \mathbb{R} to \mathbb{R} . It is $E(w)[q]$, but not $E(w)$ which is in \mathbb{R} .

This proposal is opposed to seeing a violation profile as a real number, as a vector, or as some other construct. Namely, equations (2.10) and (3.2) introduced the *vector representation* of a violation profile. Equation (3.4), for a constant q , attempted to introduce a *real valued representation* (different qs would correspond to different representations); even though we have just seen that this approach would not work in the general case. The next section presents how to realise a profile as an ordinal ("infinite") number (cf. equation (3.19)). All these representations correspond to different ways of defining the rightmost set ($\{E(w) \mid w \in GEN(UR)\}$) on Fig. 3.1.

3.3.1 Comparing polynomials

This new representation now requires us to introduce the relations \prec and $=$ on the range of the representation E . As the $=$ relation is simply the identity relation, introducing the order \prec is always the less trivial task. So far, we used the lexicographic order on the vector representations, and the everyday "less than" relation on the real valued representation. How shall we deal now with the polynomial representations?

Obviously, $E(w)[q]$ goes to infinity as q grows without bound:

$$\lim_{q \rightarrow +\infty} E(w)[q] = +\infty$$

Therefore, observing directly the limit of $E(w)[q]$ will not work, whatever we would like to do with the violation profiles. (First, we will aim at defining the \prec relation in order to prove the soundness of our approach. And then, we

¹³Notice that this problem arises only if the candidate set corresponding to a certain input is infinite. Otherwise the real valued representation would work, even if different inputs required different qs . In unidirectional Optimality Theory, the candidate sets of different inputs may overlap, but do not interact with each other.

¹⁴For cases when any monotonically decreasing series of weights can be used, see Prince (2002).

¹⁵The idea of using polynomial arithmetics originates from Balázs Szendrői.

¹⁶The more general formulation mentioned in footnote 12 looks as:

$$E(w)[q] = \sum_{i \in \mathcal{I}} C_i(w) \cdot q^i$$

shall interpret Eq. (2.2) for SA-OT.) The trick will always be *first* to perform an operation, or *first* to check the behaviour of the energy function, and only *subsequently* bring q to the infinity. In using *continuous* operations, it makes sense to change the order of the operation and of the limit to infinity.

First, how shall we compare two violation profiles seen as polynomials? The following definition—comparing the limits—is useless: $E(w_1) \prec E(w_2)$ if and only if

$$\lim_{q \rightarrow +\infty} E(w_1)[q] < \lim_{q \rightarrow +\infty} E(w_2)[q]$$

We may, however, consider the limit of the comparisons, instead of the comparison of the limits. The following definition works consequently perfectly, that is, it realises the harmony function:

Definition 3.3.1. $E(w_1) \prec E(w_2)$ if and only if either

$$\lim_{q \rightarrow +\infty} (E(w_2)[q] - E(w_1)[q]) > 0$$

or

$$\lim_{q \rightarrow +\infty} (E(w_2)[q] - E(w_1)[q]) = +\infty$$

Furthermore, $E(w_1) = E(w_2)$ if and only if $E(w_1)[q] = E(w_2)[q]$ for all $q \in \mathbb{R}^+$.

By using the definition of the limits and the properties of polynomials,¹⁷ we may reformulate the first part of this definition thus:

Corollary 3.3.2. $E(w_1) \prec E(w_2)$ if and only if there exists a $q_0 \in \mathbb{R}$ such that for all $q \in \mathbb{R}^+$: if $q > q_0$ then $E(w_2)[q] - E(w_1)[q] > 0$.

In other words, for any two candidates one can choose a q that is high enough so that we can simply compare the “energies” as real values. The problem with the real valued representation was that no single q exists that would always work perfectly. But the polynomial representation allows for choosing different qs for any two candidates w_1 and w_2 , and hence we have circumvented the problem.

Indeed, energy-polynomials with this definition realise the Harmony function: $E(w_1) \prec E(w_2)$ if and only if $H(w_1) \prec H(w_2)$. Similarly, $H(w_1) = H(w_2)$ if and only if $E(w_1) = E(w_2)$ (that is, $E(w_1)[q] = E(w_2)[q]$ for all $q \in \mathbb{R}^+$). We are going to prove this *equivalence* of the Harmony function to the energy polynomials in three steps.

First we demonstrate the *law of trichotomy* on the set $\{E(w) | w \in \text{GEN}(UR)\}$ with respect to the relation \prec . Namely:

Theorem 3.3.3. LAW OF TRICHOTOMY FOR THE ENERGY POLYNOMIALS: for all candidates w_1 and $w_2 \in \text{GEN}(UR)$, exactly one of the following three statements hold: either $E(w_1) \prec E(w_2)$, or $E(w_2) \prec E(w_1)$, or $E(w_1) = E(w_2)$.

Proof. First, recall first that polynomials are continuous functions, and that a basic property of continuous functions is that they map an interval onto an interval. In other words, if the continuous function $f(x)$ is defined on the interval $[a, b]$, and X is in the interval $[f(a), f(b)]$ (or $[f(b), f(a)]$, depending on whether

¹⁷Namely, the fact that if a real valued polynomial $P(x)$ is not constant, then it converges to infinity: $\lim_{x \rightarrow +\infty} P(x) = \pm\infty$.

$f(a) \leq f(b)$ or $f(b) \leq f(a)$ holds), then there exists an $x \in [a, b]$ such that $f(x) = X$.

If $E(w_1) = E(w_2)$, then by definition, $\lim_{q \rightarrow +\infty} (E(w_2)[q] - E(w_1)[q]) = 0$, so neither $E(w_1) \prec E(w_2)$ nor $E(w_2) \prec E(w_1)$.

Suppose now that $E(w_1) \neq E(w_2)$, thus we have to demonstrate that exactly one of the first two statements applies. In this case, $P[q] := E(w_1)[q] - E(w_2)[q]$ is a polynomial, which is not constant zero, and whose order is maximally N (the order of $E(w_1)[q]$ and $E(w_2)[q]$). Such a function may have maximally N different roots, that is $q_{(i)}$ values rendering it zero: $P[q_{(i)}] = 0$.

If no such real valued root exists, $P[q]$ is either positive or negative for all qs . Otherwise, $P[q_1] > 0$ and $P[q_2] < 0$ for some q_1 and q_2 would force $P[q]$ to have a root, due to the property of the continuous functions mentioned at the beginning of this proof. In turn, any $q_0 \in \mathbb{R}$ can be chosen to show by corollary 3.3.2 that $E(w_1) \prec E(w_2)$ if $P[q]$ is always negative, and that $E(w_2) \prec E(w_1)$ if $P[q]$ is always positive.

If, on the other hand, $P[q]$ does have at least one root, let q_0 be the greatest root. Now, for all $q > q_0$ the value of $P[q]$ has the same sign (always positive, or always negative): if there existed a $q_1 > q_0$ such that $P[q_1] > 0$ and another $q_2 > q_0$ such that $P[q_2] < 0$, then $P[q]$ would have a root greater than q_0 , between q_1 and q_2 , again because $P[q]$ is a continuous function. Consequently, either $P[q] = E(w_1)[q] - E(w_2)[q] > 0$ for all $q > q_0$, proving that $E(w_2) \prec E(w_1)$; or $P[q] = E(w_1)[q] - E(w_2)[q] < 0$ for all $q > q_0$, and then $E(w_1) \prec E(w_2)$, by corollary 3.3.2. \square

In the next step, we demonstrate that

Lemma 3.3.4. *If $H(w_1) \succ H(w_2)$, then $E(w_1) \prec E(w_2)$.*

Proof. If $H(w_1) \succ H(w_2)$, then, by definition, there exists an integer $k \in [N, N - 1, \dots, 1, 0]$ such that

1. $C_k(w_2) - C_k(w_1) > 0$, and
2. for all $i \in [N, N - 1, \dots, 1, 0]$: if $i > k$, then $C_i(w_2) - C_i(w_1) = 0$.

If $k = 0$ then for all q

$$E(w_2)[q] - E(w_1)[q] = \sum_{i=0}^N [C_i(w_2) - C_i(w_1)]q^i = C_k(w_2) - C_k(w_1) > 0$$

Therefore $E(w_1) \prec E(w_2)$, and any $q_0 \in \mathbb{R}$ may be chosen.

In the case, however, when $k > 0$, let us define c such that for all $i < k$: $c > C_i(w_1), C_i(w_2) \geq 0$. Such a c exists because a finite number of violation levels always have a finite upper bound. First note that for all $i < k$:

$$c > C_i(w_2) - C_i(w_1) > -c \tag{3.11}$$

Second, remember the sum of a geometric series ($q \neq 1$):

$$\sum_{i=0}^{k-1} q^i = \frac{q^k - 1}{q - 1} \tag{3.12}$$

Now let $q_0 = \max(\frac{2c}{C_k(w_2) - C_k(w_1)}, 2)$. For all $q > q_0$, then

$$\begin{aligned} E(w_2)[q] - E(w_1)[q] &= \sum_{i=0}^N [C_i(w_2) - C_i(w_1)]q^i = \\ &= [C_k(w_2) - C_k(w_1)]q^k + \sum_{i=0}^{k-1} [C_i(w_2) - C_i(w_1)]q^i \end{aligned} \quad (3.13)$$

due to the definition of the fatal constraint C_k . Because $q > q_0 \geq \frac{2c}{C_k(w_2) - C_k(w_1)}$, in the first component of the sum we can use $C_k(w_2) - C_k(w_1) > 2c/q$. For the second component, we may use equations (3.11) and (3.12). In turn, we obtain:

$$\begin{aligned} E(w_2)[q] - E(w_1)[q] &> \frac{2c}{q}q^k - c\frac{q^k - 1}{q - 1} = \\ &= \frac{c}{q - 1}[q^k - 2q^{k-1} + 1] > 0 \end{aligned} \quad (3.14)$$

because $q > q_0 \geq 2$.

In sum, either $k = 0$ or $k > 0$, we have shown that there exists a q_0 such that for all $q > q_0$: $E(w_2)[q] - E(w_1)[q] > 0$. Therefore, $E(w_1) \prec E(w_2)$. \square

Observe that the present proof did not require $C_i(w)$ be an integer, unlike the proof of the corresponding lemma for the real-number representation (Lemma 3.2.1). The reason of this difference is that now, if $C_k(w_2) - C_k(w_1) < 1$, we could simply increase q_0 . Similarly, Lemma 3.2.1 (and hence, Theorem 3.2.2) could be generalised, if a positive lower bound existed for the difference of different violation levels of the constraints. Nevertheless, the real-number representation requires a universal upper bound on the violation levels and a global lower bound on the difference of the violation levels in order to specify some q , the base of the exponential weight system. The advantage of the polynomial approach is that q is handled in a flexible way, and thus a different q_0 (or any $q > q_0$) can be used for any pair of candidates. A pair of candidates has a finite number of violation levels, which guarantees the existence of the required upper and lower bounds.

Third, we can formulate and prove that energy polynomials realise Harmony function, if using definitions 3.1.6 and 3.3.1:

Theorem 3.3.5. ENERGY POLYNOMIALS REALISE THE HARMONY FUNCTION:
 $E(w_1) = E(w_2)$ if and only if $H(w_1) = H(w_2)$;
 $E(w_1) \prec E(w_2)$ if and only if $H(w_1) \succ H(w_2)$.

Proof. This statement includes four substatements. First, if $H(w_1) = H(w_2)$, then, by definition, $C_i(w_1) = C_i(w_2)$ for all $i \in [N, N-1, \dots, 1, 0]$. Consequently, for all $q \in \mathcal{R}^+$, $E(w_1)[q] = E(w_2)[q]$.

Second, if $H(w_1) \succ H(w_2)$, then $E(w_1) \prec E(w_2)$, as demonstrated by the previous lemma.

Third, if $E(w_1) = E(w_2)$, then $H(w_1) = H(w_2)$. This is true, because either $H(w_1) = H(w_2)$, or $H(w_1) \succ H(w_2)$, or $H(w_2) \succ H(w_1)$, due to the law of trichotomy on vectors (theorem 3.1.8). Using an indirect proof, suppose $H(w_1) \succ H(w_2)$. As just shown, $E(w_1) \prec E(w_2)$ would follow, which would

contradict the law of trichotomy for the energy polynomials (theorem 3.3.3). Similarly, $H(w_2) \succ H(w_1)$ is also impossible, leaving us the only possibility $H(w_1) = H(w_2)$.¹⁸

Fourth, if $E(w_1) \prec E(w_2)$, then $H(w_1) \succ H(w_2)$. This can be demonstrated similarly to the third case, by referring to the laws of trichotomy for the Harmony function and for the energy polynomials. Namely, $H(w_2) \succeq H(w_1)$ would require $E(w_1) \preceq E(w_2)$ by the statements of the present theorem already demonstrated, which would contradict the law of trichotomy on polynomials (Theorem 3.3.3). \square

3.3.2 Simulated annealing with polynomials

So far, we have seen that energy polynomials can be used to model the behaviour of the Harmony function of Optimality Theory: their definition is sound and they realise the Harmony function. The trick was first to compare two candidates, and only then take the $q \rightarrow \infty$ limit.

Can we use energy polynomials to formulate simulated annealing for Optimality Theory? The recurring problem has been how to define the transition probability $P(w \rightarrow w')$ from candidate w to a worse candidate w' in a form that is reminiscent of the traditional expression $e^{(E(w')-E(w))/T}$. How would we define the transition probabilities in the polynomial representation of violation profiles?

Using the polynomial representation of two violation profiles, translating the expression $E(w') - E(w)$ is straightforward. It is simply another polynomial of q , namely $P[q] = E(w')[q] - E(w)[q]$. The difference of two real valued function is given for free by elementary school arithmetic. Observe that if C_k is the fatal constraint, the highest ranked constraint—the constraint with the highest index—that assigns different violation levels to w and w' , then the dominant component in $P[q]$ is q^k .

Temperature in OT simulated annealing, as explained in section 2.2.3, should have a structure similar to that of the difference of two violation profiles. If presently the difference $E(w') - E(w)$ is a polynomial of $q \in \mathbb{R}^+$, so must the temperature $T = \langle K, t \rangle$ be, as well:

$$T[q] = \langle K, t \rangle [q] = t \cdot q^K \quad (3.15)$$

The attentive reader will notice that this formulation allows K to be a real number, not only an integer, if $q > 0$.¹⁹ Nonetheless, we shall not really exploit this opportunity, besides the fact that we theoretically allow any real K values in the outer cycle of the SA-OT algorithm (Fig. 2.8).

¹⁸An alternative proof of this third substatement would employ the fact that a constant zero polynomial—such as $P[q] := E(w_1)[q] - E(w_2)[q]$ in the case $E(w_1) = E(w_2)$ —must have but zero coefficients. Thus, $C_i(w_1) - C_i(w_2) = 0$ for all $i \in [N, N-1, \dots, 1, 0]$, yielding $H(w_1) = H(w_2)$ by Definition 3.1.6.

¹⁹The polynomials proposed to represent violation profiles have been defined on the domain of positive real numbers ($q \in \mathbb{R}^+$). Although this restriction might have appeared to be unnecessary, now we can see its advantage. Besides, this restriction also allows generalising the polynomial representation to the case if the indices of the constraints are real numbers. Furthermore, q was originally the base of an exponential weight system in (3.4), which makes sense only if $q > 1$. In any case, as only the $q \rightarrow +\infty$ limit will be of interest, we can always remove a lower subset of q 's domain.

The last step is to formulate the probability of moving from candidate w to a neighbour candidate w' . If $w' \succeq w$, the probability is 1. Otherwise, we shall repeat the trick: *first* perform the operation, and only *afterwards* take the $q \rightarrow \infty$ limit.

Thus, the probability of moving from a candidate w to a worse candidate w' shall be defined as:

$$P(w \rightarrow w') = \lim_{q \rightarrow +\infty} e^{-\frac{E(w')[q] - E(w)[q]}{T[q]}} \quad (3.16)$$

Analysing the defining equation (3.16), one can quickly check that this definition yields the RULES OF MOVING on page 63, which we have been hoping for:

- If w' better than w : move $w \rightarrow w'$!
- If w' loses due to constraint $C_k > K$: don't move ($P = 0$)!
- If w' loses due to constraint $C_k < K$: move ($P = 1$)!
- If w' loses due to the constraint $C_k = K$: move with transition probability $P(w \rightarrow w') = e^{-(C_k(w') - C_k(w))/t}$.

This is so because the mathematical operations involved are continuous. Further, as q grows very large, the dominant component in $E(w')[q] - E(w)[q]$ will be the highest non-zero component, which is $(C_k(w') - C_k(w))q^k$ where C_k is the fatal constraint when comparing these two candidates:

$$\begin{aligned} P(w \rightarrow w') &= \lim_{q \rightarrow +\infty} e^{-\frac{E(w')[q] - E(w)[q]}{T[q]}} = \\ &= \lim_{q \rightarrow +\infty} e^{-\frac{(C_k(w') - C_k(w))q^k}{tq^K}} \\ &= \left[\lim_{q \rightarrow +\infty} e^{(-q^{k-K})} \right]^{\frac{C_k(w') - C_k(w)}{t}} \\ &= \begin{cases} 0 & \text{if } k > K \\ e^{-\frac{C_k(w') - C_k(w)}{t}} & \text{if } k = K \\ 1 & \text{if } k < K \end{cases} \quad (3.17) \end{aligned}$$

For a visualisation, recall Fig. 2.6. The expression $e^{(-q^\alpha)}$ is equal to e^{-1} if $\alpha = 0$. If however $\alpha < 0$, then it converges to 1 with $q \rightarrow +\infty$, similarly to the function $e^{-1/x} = e^{(-x^{-1})}$ on Fig. 2.6. In the third case, that is when $\alpha > 0$, the expression e^{-q^α} converges to 0, because this case corresponds to the $x \rightarrow +0$ limit of the function $e^{-1/x}$ (replace x with $q^{-\alpha}$).

In (3.15), we could have used a more complex expression as the definition of $T[q]$, but the form $t \cdot q^K$ will be good enough. As we take the $q \rightarrow +\infty$ limit, where only the highest component of a polynomial plays a role, adding lower components would not influence the behaviour of the system. Temperature could also have been defined not as a polynomial, but as a different function of q . Nevertheless, if $T[q]$ did not converge like some polynomial ($T[q] = \mathcal{O}(q^K)$), it would not turn useful in the equation 3.16 defining the transition probability, for the latter would always be 0 (if $T[q] = o(q^K)$) or 1 (if $T[q]/(q^K) \rightarrow \infty$). This

is the reason why we required $T[q]$ to be a polynomial of the form appearing in (3.15). In sum, the polynomial approach also advocates temperature to be a pair $T = \langle K, t \rangle$, similarly to the approach proposed in section 2.2.3, on page 58.

3.4 Violation profiles as ordinal numbers

In the present section, an alternative way is presented to introduce *Optimality Theory Simulated Annealing*. Instead of considering real-valued polynomials $E(w)[q]$ in the limit $q \rightarrow +\infty$, we immediately take *infinite weights* for q .

As demonstrated, no finite weights can reproduce, in the general case, the *strict constraint ranking* postulated by Optimality Theory. A series of exponential weights, such as in

$$E(w) := \sum_{i=0}^N C_i(w)q^i \quad (3.18)$$

realises the constraint hierarchy $C_N \gg \dots \gg C_1 \gg C_0$ only if each constraint can assign at most $q - 1$ violation marks to any candidate.

However, the number of violation marks assigned by most constraints used in linguistics does not have any upper bound theoretically. Even if one argues that performance usually limits the length of words and sentences that can be uttered, still, linguistic models can require generating never winning candidates of an unbounded length. It would be nice consequently to allow unbounded weights in (3.18). In other words, to let equation (3.18) introduce a value $E(w)$ in a number system of infinite base.

Axiomatic Set Theory proposes a solution to carry out this idea in a mathematically sound framework. When the possible levels of violation formed the well ordered set $\{0, 1, 2, \dots, q - 1\}$ —which is the definition of the integer q (Holz et al., 1999, p. 19).—, we used q as the base of an exponential weight system. In the case of unbounded violations, the possible levels of violation most often form the ordered set $\{0, 1, 2, \dots\}$. This well ordered set is called ω , the first limit ordinal (Suppes, 1972; Holz et al., 1999). In other words, ω is the upper limit of the set of the natural numbers \mathcal{N} .

Arithmetic can be defined on ordinal numbers, including comparison, as well as addition and multiplication (Holz et al., 1999).²⁰ These latter operations are associative, but not commutative. Therefore, we can redefine the E function as:²¹

²⁰See also references under: Eric W. Weisstein: *Ordinal Number*, From MathWorld—A Wolfram Web Resource, <http://mathworld.wolfram.com/OrdinalNumber.html>.

²¹Footnotes 12 and 16 proposed a more general formulation, which would translate now as:

$$E_{\mathcal{I}}(w) = \sum_{i \in \mathcal{I}} \omega^i \cdot C_i(w)$$

with the important caveat that the elements of the finite set of indices \mathcal{I} are ordinal numbers (practically: non-negative integers). Further, as ordinal addition is not commutative, we have to specify that the elements of \mathcal{I} are read in a decreasing order. This formulation naturally allows us not using certain numbers as indices; whereas in (3.19) one has to stipulate $C_j = 0$ in the case we would like to associate no “real” constraint with the index j .

$$\begin{aligned}
E_{\mathcal{H}}(w) &= \omega^N C_N(w) + \dots + \omega C_1(w) + C_0(w) \\
&= \sum_{i=N}^0 \omega^i C_i(w)
\end{aligned} \tag{3.19}$$

This expression introduces the *ordinal number representation* of a violation profile for hierarchy \mathcal{H} . We will nonetheless dismiss the index \mathcal{H} , as long as we work with a constant constraint ranking. Observe that unlike in the polynomial representation, the violation levels must be ordinals, such as non-negative integers, in order to 3.19 be meaningful.

Because ω is the upper limit of the natural numbers, $\omega^i n < \omega^{i+1}$ for any finite n . This property will guarantee that if candidate w_1 is less harmonic than candidate w_2 then $E(w_1) > E(w_2)$. In other words, ordinal arithmetic furnishes us with the relation $<$ and $=$ for free in the ordinal number representation of a violation profile.

3.4.1 Ordinal numbers can realise violation profiles

We heavily rely on results demonstrated by Holz et al. (1999), while we are proving trichotomy and representation:

Lemma 3.4.1. TRICHOTOMY ON ORDINAL NUMBERS: *Let σ and τ be ordinal numbers. Then exactly one of the following three statements hold: 1. $\sigma < \tau$; 2. $\sigma = \tau$; 3. $\tau < \sigma$.*

Proof. Lemma 1.2.3 in Holz et al. (1999, p. 16) demonstrates that for any two ordinal numbers at least one of the three statements holds. Lemma 1.2.1.c states that $\sigma \not< \sigma$; hence statements 1 and 2, as well as statements 2 and 3 cannot simultaneously hold. Similarly, by the latter lemma and by the transitivity of the $<$ relation, statements 1 and 3 cannot hold in the same time. \square

Lemma 3.4.2. *Let w_1 and w_2 be two candidates, and let $E(w_1)$ and $E(w_2)$ be the ordinal number representation of their violation profile with respect to some hierarchy \mathcal{H} . If $H(w_1) \succ_{\mathcal{H}} H(w_2)$, then $E(w_1) < E(w_2)$.*

Proof. As $H(w_1) \succ_{\mathcal{H}} H(w_2)$, and violation levels are integers, there is a constraint C_k such that

- [1] $C_k(w_1) + 1 \leq C_k(w_2)$, and
- [2] $C_i(w_1) = C_i(w_2)$ for all $i > k$.

Lemma 1.4.3 in Holz et al. (1999, p. 33) contains among others the following properties, if α , β and γ are cardinal numbers:

- [3] if $0 < \alpha$ and $\beta < \gamma$, then $\alpha \cdot \beta < \alpha \cdot \gamma$
- [4] if $\beta < \gamma$, then $\alpha + \beta < \alpha + \gamma$
- [5] $\alpha^{\beta+\gamma} = \alpha^\beta \cdot \alpha^\gamma$
- [6] if $1 < \alpha$ and $\beta < \gamma$, then $\alpha^\beta < \alpha^\gamma$

$$[7] \quad \alpha \cdot (\beta + \gamma) = \alpha \cdot \beta + \alpha \cdot \gamma$$

$$[8] \quad \alpha \cdot 1 = \alpha = 1 \cdot \alpha$$

From properties [1] and [3] it follows that $\omega^k(C_k(w_1) + 1) \leq \omega^k C_k(w_2)$. Therefore, and due to [2], [4], [7] and [8],

$$\sum_{i=N}^k \omega^i C_i(w_1) + \omega^k \leq \sum_{i=N}^k \omega^i C_i(w_2) \quad (3.20)$$

Furthermore, from $0 \leq C_i(w_1) < \omega$, by [3] and [5], follows that $\omega^j \cdot C_i(w_1) < \omega^j \cdot \omega = \omega^{j+1}$ for any i and j . Hence, due to [6], $\omega^j \cdot C_i(w_1) < \omega^k$ for any $j < k$ (that is, $j + 1 \leq k$).

From Lemma 1.4.7.b of Holz et al. (1999, p. 37) follows that for all j and $\alpha < \omega^j$, $\alpha + \omega^j = \omega^j$. Now, if both $\alpha < \omega^j$ and $\beta < \omega^j$, property [4] ensures that $\alpha + \beta < \alpha + \omega^j = \omega^j$. That is, the sum of any two ordinals smaller than ω^j is smaller than ω^j . Using this observation recursively in the case of $j = k$, we obtain:

$$\sum_{i=k-1}^0 \omega^i C_i(w_1) < \omega^k \quad (3.21)$$

From (3.21) and (3.20), by using repeatedly [4]:

$$\begin{aligned} E(w_1) &= \sum_{i=N}^k \omega^i C_i(w_1) + \sum_{i=k-1}^0 \omega^i C_i(w_1) < \\ &< \sum_{i=N}^k \omega^i C_i(w_1) + \omega^k \leq \\ &\leq \sum_{i=N}^k \omega^i C_i(w_2) \leq \\ &\leq \sum_{i=N}^0 \omega^i C_i(w_2) = E(w_2) \end{aligned} \quad (3.22)$$

□

Now, we can formally prove that the representation of a violation profile using ordinal numbers is isomorphic to the vector representation:

Theorem 3.4.3. ORDINAL NUMBERS REALISE VIOLATION PROFILES: *Let w_1 and w_2 be two candidates, and let $E(w_1)$ and $E(w_2)$ be the ordinal number representation of their violation profile with respect to some hierarchy \mathcal{H} . Then,*

- $E(w_1) = E(w_2)$ if and only if $H(w_1) = H(w_2)$;
- $E(w_1) < E(w_2)$ if and only if $H(w_1) \succ_{\mathcal{H}} H(w_2)$.

Proof. This theorem contains four substatements. If $H(w_1) = H(w_2)$, then by definition, $C_i(w_1) = C_i(w_2)$ for all i , and therefore $E(w_1) = E(w_2)$ follows directly.

If $H(w_1) \succ_{\mathcal{H}} H(w_2)$, then we have just demonstrated in the previous lemma that $E(w_1) < E(w_2)$.

If $E(w_1) = E(w_2)$, then Theorem 1.4.6 of Holz et al. (1999, p. 36) (Cantor Normal Form for the base ω) ensures that $C_i(w_1) = C_i(w_2)$ for all i s. Namely, from the theorem follows that if

$$\omega^N \cdot a_N + \omega^{N-1} \cdot a_{N-1} + \dots \omega^0 \cdot a_0 = \omega^N \cdot b_N + \omega^{N-1} \cdot b_{N-1} + \dots \omega^0 \cdot b_0$$

then $a_i = b_i$ for all i s. Consequently, $H(w_1) = H(w_2)$.

Last, if $E(w_1) < E(w_2)$, then $H(w_1) \succ H(w_2)$. Suppose this does not hold. Then $H(w_2) \succeq H(w_1)$ should be true, because of the trichotomy on the set of violation profile-like vectors (Theorem 3.1.8). This, however would involve $E(w_2) \leq E(w_1)$ due to the previously proven parts of the present theorem, which in turn would contradict the trichotomy on the class of ordinal numbers (Lemma 3.4.1). (A similar proof is also possible for the third substatement of the present theorem, if you would like to avoid the Cantor Normal Forms; cf. the relevant part of the proof of theorem 3.3.5.) \square

3.4.2 SA-OT with ordinal numbers

The next step towards SA-OT is the definition of the difference of two $E(w)$ values, which will pave the way for the introduction of temperature, necessary to interpret the expression $e^{-(E(w')-E(w))/T}$ in the context of Optimality Theory.

On the class ON of all ordinal numbers, subtraction is not defined as it is defined on the set \mathbb{Z} of the integers, or on the set \mathbb{R} of the real numbers. The class ON of all ordinal numbers can be seen as a generalisation of the natural numbers (non-negative integers), and observe that the difference $a - b$ of two natural numbers a and b is defined on the set \mathbb{N} only if $a \geq b$.

(Holz et al., 1999, p. 34) proves the following

Lemma 3.4.4. SUBTRACTION LEMMA *If $\alpha \leq \beta$ are ordinal numbers, then there is a unique ordinal γ such that $\alpha + \gamma = \beta$.*

Based on this lemma, we introduce the notation $\Delta(a, b)$ for ordinals $a \geq b$, to denote the unique ordinal x that satisfies $a = b + x$. As addition is not commutative on the class of ordinals ON, $a = \Delta(a, b) + b$ does not follow (and usually does not hold) from $a = b + \Delta(a, b)$. The notation $a - b$ and the term ‘‘subtraction’’ will be avoided in order to remind us this caveat, as well as the fact that $\Delta(a, b)$ is defined only if $a \geq b$.

Violation profiles are represented with a subset of ON, namely, with ordinals of the form $a = \sum_{i=N}^0 \omega^i a_i$, where $a_i \in \mathbb{N}_0$ (a_i is a non-negative integer). Thus, the elements of the set $\sum_{i=N}^0 \omega^i \mathbb{N}_0$ will be referred to as *violation profile-like ordinal numbers*.

The following proposition sheds light on how ordinal numbers represent violation profiles:

Proposition 3.4.5. *Given violation profile-like ordinals $a = \sum_{i=N}^0 \omega^i a_i$ and $b = \sum_{i=N}^0 \omega^i b_i$, such that $a > b$,*

$$\Delta(a, b) = \sum_{i=N}^0 \omega^i \delta_i$$

where for all $0 \leq i \leq N$

$$\delta_i = \begin{cases} a_i - b_i & \text{if } a_j = b_j \forall j. (j > i \wedge j \leq N) \\ a_i & \text{otherwise} \end{cases}$$

Proof. As the Subtraction lemma 3.4.4 proves uniqueness, it is satisfactory to show that $a = b + \sum_{i=N}^0 \omega^i \delta_i$.

Recall that ordinal addition is associative (Lemma 1.4.3.a.(v) in Holz et al. (1999, p. 33)), as well as that $\omega^i a + \omega^j b = \omega^j b$ if $i < j$.²²

Let k be the lowest index to which $\forall j > k : a_j = b_j$ holds (in the case of violation profiles, this is the index of the fatal constraint). Such a k exists, because the set $\{0, \dots, N\}$ is finite, hence well-ordered: each set, for instance the set $\{i \in \{0, \dots, N\} \mid \forall j \in \{0, \dots, N\} : (j > i) \Rightarrow (a_j = b_j)\}$, has a least element. Then,

$$\begin{aligned} b + \Delta(a, b) &= \sum_{i=N}^0 \omega^i b_i + \sum_{i=N}^0 \omega^i \delta_i = \\ &= \sum_{i=N}^k \omega^i b_i + \left(\sum_{i=k-1}^0 \omega^i b_i + \omega^k \delta_k \right) + \sum_{i=k-1}^0 \omega^i \delta_i = \\ &= \sum_{i=N}^{k+1} \omega^i a_i + \omega^k b_k + \omega^k (a_k - b_k) + \sum_{i=k-1}^0 \omega^i a_i = a \quad (3.23) \end{aligned}$$

□

In the case of violation profile-like ordinal numbers, the co-efficient δ_k of the highest non-zero term in $\Delta(a, b)$ is the difference of the respective terms in a and b . In OT, this co-efficient will reflect the difference of violation marks (i.e. the uncanceled marks) of the constraint C_k where the fatal violation takes place when comparing these two candidates. All the lower terms $\omega^i \delta_i$ are equal to the respective terms in a .

By neglecting the lower terms, which are negligible compared to the highest one, we can define another difference-like function, which better reflects what is relevant for OT. In addition, its use saves us from some unnecessary calculation.

Definition 3.4.6. Given $a = \sum_{i=N}^0 \omega^i a_i$ and $b = \sum_{i=N}^0 \omega^i b_i$, where $a > b$, let be $\Delta'(a, b) = \sum_{i=N}^0 \omega^i \delta'_i$ such that

$$\delta'_i = \begin{cases} a_i - b_i & \text{if } a_j = b_j \forall j. (j > i \wedge j \leq N) \\ 0 & \text{otherwise} \end{cases}$$

This function returns the difference of violations of the constraint where the fatal violation takes place when we compare the two candidates. It is still somehow a sort of difference, because $b + \Delta'(a, b)$ differs from a only in lower terms than what is relevant when comparing the two violation sets.

First, SA-OT will be introduced by using some intuitive conventions, as a short cut, and then we argue for using this conventions.

²²By Lemma 1.4.3.c.(iii) in Holz et al. (1999, p. 33), $\omega^i < \omega^j$. Furthermore, due to Lemma 1.4.7.b in Holz et al. (1999, p. 37), $\alpha + \omega^j = \omega^j$ for all $\alpha < \omega^j$.

Thus, I propose the following notations, reflecting the idea that ω is a form of “infinity”:

$$e^{-\frac{\omega^i a}{\omega^j b}} := e^{-\omega^{i-j} \frac{a}{b}} := \begin{cases} 1 & \text{if } i < j \\ e^{-\frac{a}{b}} & \text{if } i = j \\ 0 & \text{if } i > j \end{cases} \quad (3.24)$$

$$e^{-\frac{x+y}{z}} := e^{-\frac{x}{z}} e^{-\frac{y}{z}} \quad (3.25)$$

where a, b, i and j are positive natural numbers, while x, y and z are ordinal numbers.

Employing these notational conventions, we can directly introduce the transition probabilities required by simulated annealing:

$$\begin{aligned} &\text{If } E(w) \geq E(w') \text{ then } P(w \rightarrow w' | T) = 1, \text{ otherwise} \\ &P(w \rightarrow w' | T) := e^{-\frac{\Delta(E(w'), E(w))}{T}} = e^{-\frac{\Delta'(E(w'), E(w))}{T}} \end{aligned} \quad (3.26)$$

Therefore, temperature T is also an ordinal number of the form:

$$T = \langle K_T, t \rangle = t\omega^{K_T} \quad (3.27)$$

One can simply check that both notions of difference, $\Delta(E(w'), E(w))$ and $\Delta'(E(w'), E(w))$, define the same probability. Using the second notion is somewhat farther from the traditional idea in SA (it is not exactly the difference of the energy levels), but it is closer to the philosophy of OT (ignore the constraints below the fatal constraint), and it is simpler to calculate.

By representing the Harmony function as an ordinal-valued energy function, we could formulate equation 3.26, which has a form that is fully analogous to the traditional transition probability equation used in real-valued simulated annealing:

$$P(w \rightarrow w' | T) = e^{-\frac{\Delta E}{T}} = e^{-\frac{E(w') - E(w)}{T}} \quad (3.28)$$

The interpretation of equation 3.26, in turn, leads to the same rules determining transition probabilities (the *Rules of moving* on page 63) that we have formulated earlier, in section 2.2.3. Namely, if temperature is $T = \langle K_T, t \rangle$, then:

- If w' is better than w ($w' \succ w$, that is, $C_k(w') < C_k(w)$), then move from w to w' .
- If w' loses due to the critical constraint $C_k > K_T$: don't move!
- If w' loses due to the critical constraint $C_k < K_T$: move!
- If w' loses due to the critical constraint $C_k = K_T$: move with probability $P(w \rightarrow w') = e^{-d/t}$, where $d = C_k(w') - C_k(w)$.

3.4.3 Arguing more for the definition of $e^{-d/t}$

Conventions (3.24) and (3.25), on the one hand, “make sense” because ω is but a mathematically sound way of saying “infinite”, and these proposals lead directly to a formulation of SA-OT in the ordinal representation of the violation profiles. On the other hand, they might nonetheless seem to the reader as *ad hoc*, and therefore spoil the mathematically precise underpinning of SA-OT. In the remaining pages of the present section we argue for this short-cut.

First, we quote another lemma from Holz et al. (1999, p. 34). Not only does the *Subtraction Lemma* hold on the class of ordinals, but also the *Division Lemma* and the *Logarithm Lemma*. The former states the following:

Lemma 3.4.7. DIVISION LEMMA: *Let a and b be ordinals. If $b \neq 0$, then there are unique ordinals q and m satisfying $a = b \cdot q + m$ and $m < b$.*

For ordinals a and $b \neq 0$, let $q(a, b)$ and $r(a, b)$ therefore denote the unique ordinals such that $a = b \cdot q(a, b) + r(a, b)$ and $r(a, b) < b$. $q(a, b)$ will be referred to as the *quotient*, and $r(a, b)$ as the *remainder* of a and b .

Our goal is to translate the expression $e^{-\frac{E(w')-E(w)}{t}}$ into ordinal arithmetic. The quotient $\frac{E(w')-E(w)}{t}$ can be easily rewritten as $q(\Delta(E(w'), E(w)), T)$ if $E(w') \geq E(w)$ —that is, exactly in the case we actually need this expression for the transition probabilities (if $w \succeq w'$). But we are still not able to interpret the expression $e^{-(E(w')-E(w))/t}$, because of the negative sign and because e is not an integer. Not surprisingly, for the value of this expression, a real number between 0 and 1, is unquestionably beyond the scope of ordinal arithmetic.

However, the following two observations can help us overcome this difficulty.

First, observe that the expression $e^{-d/t}$ can be replaced by the expression $a^{-d/t}$ for any real number $a \neq 1$, by simply rescaling temperature (or the violation levels), because²³

$$e^{-\frac{d}{T}} = a^{-\frac{d}{T \ln a}} \quad (3.29)$$

The concept of *rescaling* originates from physics. One can measure a quantity using different scales—e.g., metres, kilometres, feet, yards, lightyears, etc. for distance—and the difference is but a constant multiplicative factor. Now, $T' = T \ln a$ will replace the earlier T , and then the form of the equations can be kept unchanged.

In turn, ordinal exponentiation can be used by replacing e with an arbitrarily chosen *integer* base $a > 1$ —for instance $a = 2$ or $a = 3$ in order to remain close to the original exponentiation of base $e \approx 2.71$.

Second, we can also get rid of the $-$ sign in the exponent. A transition probability $p = e^{-d/t}$ means that first we generate a random number r in the interval $]0, 1[$ with an equal distribution, and then we move the random walker iff $r < p = e^{-d/t}$. The transition probability is the *measure* of the set of the r values that result in moving—that is, of the r values that satisfy this inequality. Now, the negative sign can be removed by rewriting this inequality. We can say therefore that we move iff $r^{-1} > e^{d/t}$; that is, if and only if for all $\alpha > 0$

$$r^{-\alpha} > e^{\frac{d \cdot \alpha}{T}} \quad (3.30)$$

²³Recall that $\log_a b = \frac{\log_c b}{\log_c a}$, that is, $\log_a e = \frac{1}{\ln a}$, if $a \neq 1$.

If P is a probability measure on $\{r|0 < r < 1\}$ (specifically, we use an equal distribution: $P(\{r|a < r < b\}) = b - a$), then

$$P(w \rightarrow w'|T) = P(\{r|\forall \alpha > 0 : r^{-\alpha} > e^{\frac{d-\alpha}{T}}\}) \quad (3.31)$$

Introducing the arbitrary multiplier *alpha* will help us in the case d is not dividable by T in integer arithmetic.

In order to be able to compare formally a real number, such as $r^{-\alpha}$, with an ordinal derived from the representation of the violation profiles, we introduce the following

Definition 3.4.8. Let $\mathbb{R}^\infty := \mathbb{R} \cup \{\infty\}$, the enlargement of the set of the real numbers with $+\infty$.²⁴ Let the relation $>'$ be the enlargement of the usual order $> \subset \mathbb{R} \times \mathbb{R}$ on \mathbb{R}^∞ :

$$>' := \left\{ (a, b) \in \mathbb{R} \times \mathbb{R} \mid a > b \right\} \cup \left\{ (\infty, a) \mid a \in \mathbb{R} \right\}$$

If O is a set of ordinal numbers, then the function $R : O \rightarrow \mathbb{R}^\infty$ is defined as

$$R[a] := \begin{cases} a & \text{if } a < \omega \\ \infty & \text{if } a \geq \omega \end{cases}$$

One can simply demonstrate that the relation $>'$ is indeed a total order on \mathbb{R}^∞ . The symbol $>$ is usually used both on the set \mathbb{R} and on the class of ordinals, whereas we rather use $>'$ on \mathbb{R}^∞ in order to avoid confusion.

The definition of the function R makes use of the fact that the integers are defined as ordinals less than ω , and then they are injected into \mathbb{R} . The class of all ordinals is not a set, yet we can for instance take the set $O = \omega^\omega$ for our purposes, which contains all violation profile-like ordinals.

After all these remarks and definitions, we can reformulate within ordinal arithmetic how to decide in SA-OT whether to move from candidate w to candidate w' , if $w \succ w'$, that is, if $E(w) < E(w')$.

The straightforward solution would be to generate a real number r between 0 and 1 with equal distribution, and then move if and only if

$$\frac{1}{r} >' R \left[2^q \left(\Delta(E(w'), E(w)), T \right) \right] \quad (3.32)$$

The problem with this proposal is that the division is performed in a coarse way, similarly to division in integer arithmetic. Suppose for instance that $T = \omega^k \cdot 3$. Then, no distinction is made between $\Delta(E(w'), E(w))$ being $\omega^k \cdot 5$ or $\omega^k \cdot 3$, for in both cases $q \left(\Delta(E(w'), E(w)), T \right) = 1$ and only the remainders are different. Even though the empirical predictions of such a model might be worth investigating, this is probably not what we want. Namely, this model could not make the difference between a step that increases the violation level of constraint C_k by 3 or by 5.

The problem is that we cannot make use of the remainder of the division. How do you get a higher precision if you are forced to use exclusively integer division? You multiply the numerator by 10 or by 100, and then you consider

²⁴Compare to the addition of the *point at infinity* to each line in projective geometry.

the last digits of the quotient as being beyond the decimal point. Applying this trick has been the purpose of introducing the arbitrary positive multiplier α in equation (3.30).

Therefore, we rather propose moving from w to w' if and only if

$$\forall \alpha \in \mathbb{N}^+ : r^{-\alpha} >' R \left[2^q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) \right] \quad (3.33)$$

What follows from this definition? Let $T = \omega^K \cdot t$, and $\Delta'(E(w'), E(w)) = \omega^k \cdot d$. In words, the fatal constraint is C_k , and $d = C_k(w') - C_k(w) > 0$. Then, we consider the following cases:

1. Suppose $K > k$ (informally, $TT \gg \Delta(E(w'), E(w))$, cf. definition 2.2.5). Then, for all $\alpha \in \mathbb{N}^+$, $q\left(\Delta(E(w'), E(w)) \cdot \alpha, T\right) = 0$, because the divider is always larger than the numerator ($\omega^j \cdot \alpha < \omega^{j+1}$). It follows that for all $r < 1$, rule (3.33) prescribes to move to w' . That is, the measure of the set of the r values resulting in move—the transition probability—is $P(w \rightarrow w'|T) = 1$.
2. Suppose now $K = k$ (informally, $TT \approx \Delta(E(w'), E(w))$, cf. definition 2.2.5). Then for all $\alpha \in \mathbb{N}^+$,

$$R \left[q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) \right] = \left[\frac{d\alpha}{t} \right] \leq \frac{d\alpha}{t} \quad (3.34)$$

where $\left[\frac{d\alpha}{t} \right]$ denotes the integer part of $\frac{d\alpha}{t}$.

Hence, (3.33) holds if and only if $r < 2^{-d/t}$. Namely, if $r < 2^{-d/t}$, then for all $\alpha \in \mathbb{N}^+$,

$$r^{-\alpha} > 2^{\frac{d\alpha}{t}} \geq 2^{\left[\frac{d\alpha}{t} \right]} = R \left[2^q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) \right] \quad (3.35)$$

Further, if $r \geq 2^{-d/t}$, then choose $\alpha = ct$ (for any positive integer c) to show that the condition for moving is not satisfied anymore:

$$r^{-\alpha} \leq 2^{d\alpha/t} = 2^q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) \quad (3.36)$$

As r is chosen with an equal distribution, the measure of the set of the r values causing the system to move is thus $2^{-d/t}$. By rescaling temperature, we obtain the usual rule for this case: “move with probability $P(w \rightarrow w'|T) = e^{-d/t}$!”

3. Suppose finally $K < k$ (informally, $TT \gg \Delta(E(w'), E(w))$). Then

$$q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) \geq \omega^{k-K} \cdot (d\alpha) \geq \omega \quad (3.37)$$

Hence, the exponentiation in (3.33) returns an infinite ordinal ($2^\omega = \omega$) in the present case. As $r^{-\alpha} \not>' \infty$ by the definition of $>'$, the consequence is that no r ever results in moving. Thus, the measure of the set of r values causing the system to move is zero: $P(w \rightarrow w'|T) = 0$.

Summarising, we have again derived the *Rules of moving* from page 63.

3.5 Summary of the formal approaches

When we introduced the idea of applying simulated annealing to Optimality Theory, many different options could have been followed. Indeed, we saw in section 2.3.2 that the proposed solution does not always work. Due to the Strict Domination Hypothesis, some models are always stuck in local optima, and the algorithm’s precision—the likelihood of finding the global optimum—does not converge to 1 as the number of iterations grows infinite. There, we speculated about possible ways to solve this problem, without much success.

I will therefore argue that these failures are inevitable in SA-OT, and actually we can make use of them in building linguistic models. Yet, before making these statements, I have to convince the reader that the proposed SA-OT is indeed the most appropriate implementation of simulated annealing for Optimality Theory. This has exactly been the goal of the present chapter. Already section 2.2.3 contained a train of thought that introduced SA-OT, whereas the present chapter formally showed how the Strict Domination Hypothesis leads directly to the same *Rules of moving*—twice, at that.

In sum, we may conclude that the transition probabilities driving OT-SA are well-founded: we have seen several ways in which they may be derived from the basic ideas of Optimality Theory. The bottom line was in both cases the same *Rules of moving*.

One may ask what the polynomial approach and the ordinal number approach can contribute to each other. The answer is manifold. Firstly, the two approaches are based on very different mathematical concepts, and yet, they led to the same algorithm. Secondly, the mathematical beauty of a model is a very subjective feature, hence, different readers may prefer one or the other approach. For instance, one may not like the way calculating the limit is proposed in (3.16), or not be convinced of the necessity of introducing an arbitrary α in (3.33). Additionally, the beauty of transfinite arithmetic (analysed as conceptual blending by Núñez, 2005) may arguably be an additional subjective value of the cardinal approach in the eyes of some readers. Indeed, the contradicting reviews I received to my article (Bíró, 2005b) demonstrated that different people are more convinced by one or the other approach.

Formal arguments can also be made why to introduce both approaches in this dissertation. In most linguistic models, violation profiles are non-negative integers (represented as a certain number of stars in a tableau). These are exactly the values allowed the $C_i(w)$ s to take by the representation of violation profiles as ordinal numbers in (3.19). If the range of $C_i(w)$ is some other well-ordered set (such as the set of the consonants ordered according to sonority in the Berber example of Prince and Smolensky, 2004), an isomorphy could be applied to map this set onto some ordinal numbers. Indeed, definition 3.1.1 (page 76) requires the set $\{C_i(w) \mid w \in UR\}$ be a well-ordered set in order to make definition 3.3 introducing the main idea of OT (page 82) well-founded. This observation invites the generalisation to allow constraints that can take any ordinal numbers as values. Such an approach would naturally prohibit ganging up effects for most phenomena (if $C_i(w) < \omega$), and allow them in some special cases by stipulating $C_i(w) \geq \omega$.²⁵

²⁵Similarly to the way we propose here to generalise the range of the constraints from \mathbb{N}_0 to larger sets of ordinal numbers, further research might also enlarge the set of constraints. Indeed, the main restriction of the constraint set is that it must be a totally ordered set such

Nonetheless, the polynomial representation in (3.10) furnishes us with more flexibility, as the only requirement is that $C_i(w)$ take real values—even if the formal definition of OT (3.3) was based on the definition 3.1.1 of a constraint requiring its range to be well-ordered. And indeed, the SA-OT algorithm in Fig. 2.8 follows the idea that the violation levels are real numbers.

Even though both approaches have been introduced in order to faithfully realise the Strict Domination Hypothesis, both of them point towards a possibility of representing situations that do not satisfy this hypothesis. In the polynomial approach, one may decide not to perform the $q \rightarrow \infty$ limit, but to replace it by stipulating a high value for q , as an approximation of the $q \rightarrow \infty$ limit. Then, even if most constraints follow the Strict Domination Hypothesis, some special cases can display cumulativity effects. In the ordinal approach, one can argue for using non-finite violation levels ($C_i(w) \geq \omega$) if forced to account for cumulativity phenomena.

The advantage of both approaches—especially of the ordinal approach—over the traditional real valued realisation is that the Strict Domination Hypothesis can be saved as categorically true for the cases where it really applies; and towards the cases where it does not, the border is sharp.

that each subset have a unique *maximal* element. Take for instance Prince and Smolenksy’s “bag of violation marks” approach and their *Cancellation Lemma*: after cancelling the shared violations, the task is to identify the *unique* highest ranked constraint that still has a violation in one of the bags.

By reversing the direction of the ranking relation among the constraints, we could therefore propose simply a constraint set $\{C_i \mid i \in \mathcal{I}\}$ where \mathcal{I} can be any well-ordered set (hence, even larger than ω). The indices i of the constraints would then be ordinal numbers. The problem is that the *inverse* of the indices would be needed in equation (3.19) (page 95) due to the reversion of the constraint ranking relation, which operation is not defined among ordinal numbers. Nevertheless, this observed “duality” of the range of the constraints and the constraint set is probably worth analysing further, and might have consequences for the relationship of generation and learning in OT in general (cf. Turkel, 1994).