Calibration of constraint promotion does not help with learning variation in stochastic OT

Giorgio Magri and Benjamin Storme

Abstract

The *Calibrated error-driven ranking algorithm* (CEDRA; Magri 2012) is shown to fail on two test cases of phonologically conditioned variation from Boersma and Hayes (2001). The failure of CEDRA raises a serious unsolved challenge for learnability research in stochastic OT, because CEDRA itself was proposed to repair a learnability problem (Pater 2008) encountered by the original GLA. This result is supported by both simulation results and a detailed analysis whereby a few constraints and a few candidates at the time are recursively "peeled off" until we are left with a "core" small enough that the behavior of the learner is easy to interpret.

Keywords: phonological variation; Stochastic Optimality Theory; Gradual Learning Algorithm; Error-Driven Constraint Demotion.

1 Introduction

This article offers a method of diagnosis applicable to proposed learning algorithms for stochastic Optimality Theory (OT; Prince and Smolensky 2004; Boersma 1997, 1998). The core intuition (which stems from Tesar and Smolensky's 1998 seminal learnability analysis) is that we can tackle a complex test case with many constraints and many candidates because the structure of (deterministic or stochastic) OT allows us to recursively "peel off" a few constraints and a few candidates at the time, until we are left with a "core" which is small enough to immediately reveal the behavior of the learner.

Using this method, we scrutinize four different ranking algorithms. They differ along two dimensions: the amount of constraint promotion performed (none, small, or large); and which loser-preferring constraints are demoted (all of them or just the undominated ones). We apply these four algorithms to two test cases from Boersma and Hayes (2001; henceforth BH): Ilokano segmental phonology and Finnish genitive plural allomorphy. We find that of the four algorithms, only the original *Gradual Learning Algorithm* (GLA; Boersma 1997, 1998) as employed by BH can learn the relevant patterns. Among the rival algorithms that fail is the *Calibrated error-driven ranking algorithm* (CEDRA) proposed by Magri (2012). The failure of CEDRA raises a serious unsolved challenge for learnability research in stochastic OT, because CEDRA itself was proposed to repair a learnability problem (Pater 2008) encountered by the original GLA. Thus at present, there is no algorithm for stochastic OT that works in all cases.

The paper is organized as follows. Section 2 briefly reviews the various implementations of OT stochastic error-driven learning compared. Section 3 illustrates our technique for the analysis of stochastic error-driven learners on the Ilokano metathesis test case. The analysis leads to a *proof* that the GLA always succeeds on this test case, without any need to run any simulations. Section 4 looks at the Finnish test case and shows that the glitch in the GLA's performance in BH's simulations is due to a failure of the grammatical analysis, not to a shortcoming of the learner. The proposed analyses straightforwardly explain why the GLA's variants considered, and in particular the CEDRA, fail on these test cases. Section 5 concludes by discussing the implications for the theory of error-driven learning in stochastic OT. The presentation is kept informal with details relegated to an appendix available online.

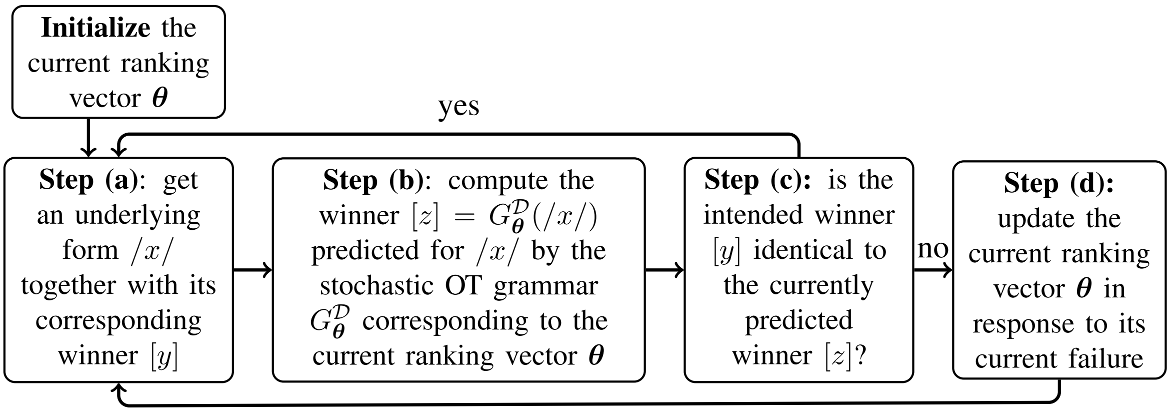
2 OT stochastic error-driven learning

Boersma (1997, 1998) introduces the following *stochastic* variant of OT. We are given a set of candidate pairs of underlying and surface forms together with a constraint set consisting of n constraints $C_1, \ldots, C_n$. Each constraint C_k is assigned a *ranking value* θ_k. These ranking values are collected together into a *ranking vector* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. Let $\epsilon_1, \ldots, \epsilon_n$ be n numbers sampled independently from each other according to the same underlying continuous distribution $\mathcal{D}$. These numbers are collected together into a stochastic vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. The sum $\theta_k + \epsilon_k$ of each ranking value θ_k with the corresponding stochastic value ϵ_k is called a *stochastic ranking value*. These sums are collected together into the *stochastic ranking vector* $\boldsymbol{\theta} + \boldsymbol{\epsilon} = (\theta_1 + \epsilon_1, \ldots, \theta_n + \epsilon_n)$. Since the numbers ϵ_h, ϵ_k are sampled according to a distribution $\mathcal{D}$ which is continuous, the probability that two stochastic ranking values $\theta_h + \epsilon_h$ and $\theta_k + \epsilon_k$ are identical is equal to zero. The stochastic ranking vector $\boldsymbol{\theta} + \boldsymbol{\epsilon}$ thus represents the unique constraint ranking which ranks a constraint C_h above another constraint C_k if and only if the stochastic ranking value $\theta_h + \epsilon_h$ of the former is larger than the stochastic ranking value $\theta_k + \epsilon_k$ of the latter. The *stochastic OT grammar* $G^{\mathcal{D}}_{\boldsymbol{\theta}}$ corresponding to a ranking vector $\boldsymbol{\theta}$ and a continuous distribution $\mathcal{D}$ (given a candidate set and a constraint set) is the function from underlying to surface forms defined as follows: whenever it is called on an underlying form $/x/$, it samples the components of $\boldsymbol{\epsilon}$ independently from each other according to the distribution $\mathcal{D}$ and it returns the surface form $[y]$ such that (the unique constraint ranking represented by) the stochastic ranking vector $\boldsymbol{\theta} + \boldsymbol{\epsilon}$ prefers the candidate pair $(/x/, [y])$ to any other candidate $(/x/, [z])$ which pairs that underlying form $/x/$ with a different surface form $[z]$.

Boersma assumes $\mathcal{D}$ to be a gaussian distribution with zero mean and small variance. Since the tails of the gaussian distribution decrease exponentially fast, the stochastic value ϵ_k is bounded between $-\Delta$ and $+\Delta$ with high probability (which of course depends on the threshold Δ). Thus, whenever the distance between two ranking values θ_h and θ_k is large (namely, larger than 2Δ), the condition $\theta_h > \theta_k$ is equivalent to the condition $\theta_h + \epsilon_h > \theta_k + \epsilon_k$ with high probability. In other words, the original ranking vector $\boldsymbol{\theta}$ and the corresponding stochastic ranking vector $\boldsymbol{\theta} + \boldsymbol{\epsilon}$ agree on the relative ranking of the two constraints C_h and C_k. From the analytical perspective adopted in this paper, it is nonetheless convenient to make each stochastic value ϵ_k deterministically bounded between $-\Delta$ and $+\Delta$, rather than bounded with high probability.[1] The analyses developed in this paper extend with high probability to the gaussian case.

Error-driven learning within this framework takes the form of the *stochastic error-driven ranking algorithm* (EDRA) in (1). This learner maintains a current stochastic OT grammar, represented through a current ranking vector $\boldsymbol{\theta}$. Following BH, these current ranking values are initialized all to the same value, say 100 for concreteness. These initial ranking values are then updated by looping though the four steps (1a)-(1d).

(1)



At step (1a), the algorithm receives a piece of data consisting of an underlying form $/x/$ together with the corresponding surface realization $[y]$ according to the target grammar. At step (1b), the algorithm computes the candidate $[z]$ predicted to be the winner for the underlying form $/x/$ by the stochastic OT grammar $G^{\mathcal{D}}_{\boldsymbol{\theta}}$ corresponding to the current ranking vector $\boldsymbol{\theta}$ (and a certain distribution $\mathcal{D}$ used to sample the stochastic values). If the predicted winner $[z]$ coincides with the intended winner $[y]$, the current ranking vector has performed impeccably on the current piece of data. The EDRA thus has nothing to learn from the current piece of data, loops back to step (1a), and waits for more data. If instead the predicted winner $[z]$ differs from the intended winner $[y]$, the current ranking vector is updated at step (1d).

The learner focuses on the comparison between the intended winner $[y]$ and the incorrectly predicted winner $[z]$. The latter is usually referred to as the current *loser* form. The failure of the current stochastic ranking vector suggests that the constraints which prefer the current loser $[z]$ over the intended winner $[y]$ (namely, the constraints which assign less violations to the former than to the latter) are currently ranked too high while the constraints which prefer the intended winner $[y]$ over the current loser $[z]$ are ranked too low. The re-ranking rule used by the EDRA at step (1d) tries to remedy to these shortcomings: it promotes winner-preferring constraints by a certain promotion amount p; and it demotes loser-preferring constraints by a certain demotion amount d. What matters for the behavior of the algorithm is the ratio between p and d (not their individual size). The demotion amount d can thus be set equal to 1 for concreteness.[2]

Different re-ranking rules differ with respect to two choice-points. The first choice point concerns the promotion component of the re-ranking rule: how should we choose the promotion amount p? Various options have been explored in the literature, summarized in the second column of table 1. The gradual EDCD re-ranking rule (*Error-Driven Constraint Demotion*; Tesar and Smolensky 1998)[3] assumes a null promotion amount $p = 0$. The GLA re-ranking rule (*Gradual Learning Algorithm*; Boersma 1997, 1998) assumes that the promotion amount equals the demotion amount, namely $p = d = 1$. Finally, the CEDRA (*Calibrated EDRA*; Magri 2012) assumes a promotion amount calibrated on the number w of winner-preferring constraints promoted through the identity $p = \frac{1}{w+1}$, whereby the promotion amount p ends up being always smaller than the demotion amount $d = 1$.

The second choice point concerns the demotion component: which loser-preferring constraints are demoted? (All proposals in the literature agree on treating all the winner-preferring constraints on a par). Various options have been explored in the literature, summarized in the third column of table 1. The GLA demotes all loser-preferring constraints. EDCD and the CEDRA instead only demote those loser-preferring constraints which need to be demoted, namely those that are currently *undominated*, in the sense that their stochastic ranking values are not smaller than the stochastic ranking value of some winner-preferring constraint. Finally, the minGLA only demotes the unique undominated loser-preferring constraint which is currently ranked highest, namely which has the largest stochastic ranking value.[4]

3 The Ilokano metathesis test case

This section analyzes the performance of the stochastic EDRA on BH's Ilokano metathesis test case (based on data and analysis from Hayes and Abad 1989).

3.1 Description of the test case

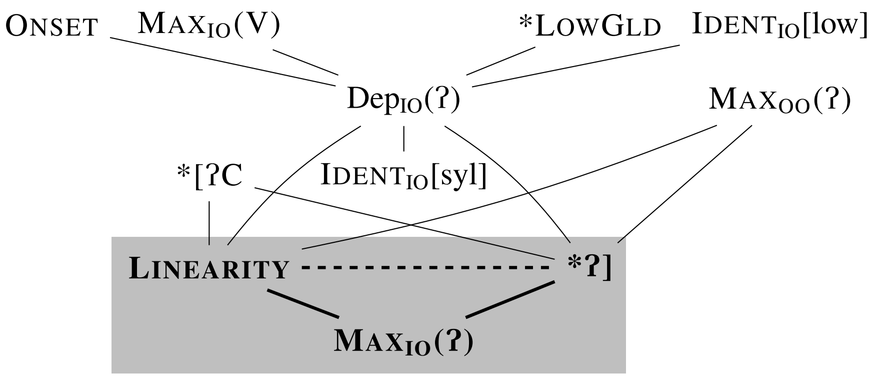
The Ilokano metathesis test case is summarized in (2) and (3). It features four underlying forms listed in (2a), which are assumed to have the same frequency. The corresponding candidates are listed in (2b) together with their probabilities. The three underlying forms /paʔlak/, /ʔajo-en/, and /basa-en/ admit a unique optimal candidate surface form (with probability 1). The underlying form /taʔo-en/ instead displays variation between the two surface forms [taʔ.wen] and [taw.ʔen], which BH assume to be equally frequent.

(2)

a.	/paʔlak/		/ʔajo-en/		/basa-en/		/taʔo-en/	
b.	[paʔ.lak]	0	[ʔa.jo.en]	0	[ba.sa.en]	0	[ta.ʔo.en]	0
	[pa.lak]	**1**	[ʔaj.wen]	**1**	[bas.a̯en]	0	[taʔ.wen]	**.5**
	[pa.ʔlak]	0	[ʔa.jo.ʔen]	0	[bas.wen]	0	[taw.ʔen]	**.5**
	[pal.ʔak]	0	[ʔa.jen]	0	[ba.sa.ʔen]	**1**	[ta.wen]	0
					[ba.sen]	0	[ta.ʔwen]	0
							[ta.ʔo.ʔen]	0
							[ta.ʔen]	0

BH assume the constraints listed in (3) and show that the data are properly accounted for by the ranking conditions displayed. The solid lines connect constraints whose ranking values need to be separated by a large distance, so as not to be swappable by the stochastic component. The dotted line instead indicates that the ranking values of the constraints LINEARITY and *ʔ] must be equal in order to model the free variation displayed by /taʔo-en/.

(3)



3.2 Simulation results

We have run the stochastic EDRA on the Ilokano metathesis test case (2)-(3) with the four re-ranking rules in table 1.[5] Table 2 reports the final ranking vector learned for each of the four re-ranking rules.[6] The quality of these final ranking vectors is evaluated in table 3. The first two columns list all pairs of an underlying form and a corresponding candidate together with the actual probability of the corresponding mapping. The four remaining columns provide the frequency of each mapping predicted by the four final ranking vectors listed in table 2.[7]

We see in table 3 that all four algorithms manage to learn the stochastic behavior of the underlying form /taʔo-en/. Furthermore, all four algorithms manage to learn the deterministic behavior of the underlying forms /ʔajo-en/ and /basa-en/. The critical test case is the underlying form /paʔlak/: both the GLA and the minGLA succeed, while the CEDRA comes short and EDCD fails. What makes /paʔlak/ hard to learn? How do the GLA and the minGLA actually manage to succeed? Why is it that EDCD and the CEDRA instead fail?

3.3 Restating the test case in ERC notation

To start, we describe the Ilokano metathesis test case (2)-(3) in ERC (*Elementary Ranking Condition*; Prince 2002) notation, as in table 4. We list in the left most column all possible triplets ($/x/$, $[y]$, ~~$[z]$~~) of an underlying form $/x/$, *a* corresponding winner $[y]$ (namely any candidate for that underlying form which has a non-null probability), and any other candidate $[z]$ different from that winner, which therefore counts as a loser in the current comparison. For instance, the first triplet (/paʔlak/, [pa.lak], ~~[pa.ʔlak]~~) consists of the underlying form /paʔlak/, its winner candidate [pa.lak], and one of its loser candidates, in this case ~~[pa.ʔlak]~~. We adopt the convention of striking out the loser in each triplet, in order to distinguish it from the winner. The remaining triplets in the first block are obtained by considering all possible loser candidates for that underlying form /paʔlak/. The next two blocks corresponding to the underlying forms /ʔajo-en/ and /basa-en/ are constructed analogously. Finally, the underlying form /taʔo-en/ comes with two winners [taw.ʔen] and [taʔ.wen]. It thus yields two blocks of triplets, each corresponding to this underlying form, one of the two winners and all remaining loser candidates.

Each underlying/winner/loser form triplet ($/x/$, $[y]$, ~~$[z]$~~) sorts the constraints into winner-preferring (i.e. those which assign less violations to the winner $[y]$ than to the loser $[z]$), loser-preferring (i.e. those which assign less violations to the loser $[z]$ than to the winner $[y]$), and even. We write a W (or an L) when the constraint corresponding to the column considered is winner-preferring (loser-preferring) relative to the triplet corresponding to the row considered (while the entries corresponding to even constraints are left empty for readability). To illustrate, the entry corresponding to the first ERC (/paʔlak/, [pa.lak], ~~[pa.ʔlak]~~) and the markedness constraint *[ʔC is a W because that constraint is winner-preferring, as it is violated by the loser ~~[pa.ʔlak]~~ but not by the winner [pa.lak]. The ERC matrix thus obtained summarizes the actions available to the EDRA: each update is triggered by one of these ERCs and it can be described as promoting the constraints with a W and demoting (some of) the constraints with an L.

3.4 First round of simplifications

The six left-most columns in table 4 only have W's but no L's. The corresponding constraints are therefore never demoted, namely they never drop below their initial ranking value. We focus on the triplets that have a W corresponding to one of these six constraints. The following fact 1 guarantees that these triplets can trigger only "few" updates in any run of the stochastic EDRA. The appendix available online provides a more explicit formulation of this fact (with an explicit bound on the number of updates that these ERCs can trigger).

Fact 1 *Consider an arbitrary run of the stochastic EDRA on the Ilokano metathesis test case (2)-(3) with any of the four re-ranking rules listed in table 1. Assume that the stochastic values are sampled between $-\Delta$ and $+\Delta$. Each of the triplets in table 4 which has a* W *corresponding to one of the six left-most constraints can trigger only a "small" number of updates.*

Fact 1 is empirically confirmed by a closer look at the simulations. The first column of table 5 lists all triplets. The columns headed by "#U" ("number of updates") report the total number of updates triggered by that triplet in one of our simulations. For instance, the entry 2 corresponding to the first triplet and the GLA re-ranking rule says that that triplet has triggered only 2 updates in a run of the GLA. For each simulation, we have also maintained an incremental counter of the number of updates: it starts at zero and it is increased by 1 whenever the algorithm performs an update. The columns headed "LU" ("last update") report the value of the counter when the corresponding triplet has triggered its last update in the simulation considered. For instance, the entry 26 corresponding to the first triplet and the GLA re-ranking rule says that that triplet has triggered the 26th update and has not triggered any additional updates afterwards. Finally, we have also maintained an incremental counter of the number of iterations: it starts at zero and it is increased by 1 at every iteration, no matter whether an update is performed or not at that iteration. The columns headed "LI" ("last iteration") report the value of the counter when the corresponding triplet has triggered its last update in the simulation considered. For instance, the entry 43 corresponding to the first triplet and the GLA re-ranking rule says that that triplet has triggered an update at the 43rd iteration and has not triggered any additional updates afterwards. Overall, table 5 thus provides information on the number of updates triggered by each triplet and on how late into the run each triplet has remained active. The triplets that have a W corresponding to one of the six left-most constraints in table 4 have been highlighted in dark gray in table 5. They are shown to trigger only a small number of updates and only at the very beginning of the run, as indeed stated by fact 1.

Fact 1 holds trivially in the case of EDCD, which performs constraint demotion but no constraint promotion. For concreteness, let's focus on the first underlying/winner/loser form triplet (/paʔlak/, [pa.lak], ~~[pa.ʔlak]~~). Its corresponding ERC has a W corresponding to the constraint *[ʔC which is indeed among the six left-most constraints with no L's. Thus, *[ʔC is never demoted during the run and always sits at its initial ranking value, as represented in (4).

(4)

separation of 2Δ { constraint *[ʔC sits at its initial ranking value

constraint Max$_{IO}$(ʔ) sits somewhere in this region

After this triplet has triggered 2Δ updates, its loser-preferring constraint Max$_{IO}$(ʔ) has been demoted at least 2Δ times. As this constraint is never promoted (because EDCD performs no constraint promotion), the ranking value of Max$_{IO}$(ʔ) has decreased by at least[8] 2Δ (possibly more, in case Max$_{IO}$(ʔ) has been demoted by some other triplets as well). The winner-preferring constraint *[ʔC is thus ranked above the loser-preferring constraint Max$_{IO}$(ʔ) with a separation of at least 2Δ. If the stochastic values are sampled between $-\Delta$ and $+\Delta$, they will never be able to swap the two constraints. This triplet will therefore never be able to trigger any further update, as stated by fact 1. Appendix A formalizes this intuitive reasoning. The proof of fact 1 for the GLA, the minGLA, and the CEDRA is based on the same intuition as for

EDCD, but it is more involved and therefore relegated to appendix B. The additional difficulty is that, after the triplet considered has triggered 2Δ updates in a run of these three algorithms, we cannot straightforwardly conclude that the ranking value of its loser-preferring constraint MAX$_{IO}$(ʔ) has decreased by at least 2Δ because that constraint could also have been promoted by some other triplets in the meantime. In the case of EDRAs which also perform constraint promotion, the connection between the number of updates triggered by an ERC and the ranking values of its loser-preferrers is harder to establish and requires a more subtle analysis.

Let's take stock. The six left-most constraints in table 4 are never loser-preferring and are therefore never demoted. We focus on the ERCs that have a W corresponding to one of these six constraints. Fact 1 guarantees that the current ranking vector after a few updates always satisfies these ERCs. The learning dynamics is therefore governed in the long run by the remaining ERCs, which are collected for convenience in the simplified table 6.

3.5 Second round of simplifications

We can now repeat the same reasoning. Constraint DEP$_{IO}$(ʔ) in table 6 is winner-preferring but never loser-preferring, and it is therefore never demoted. Fact 2 thus guarantees that those triplets in table 6 that have a W corresponding to this constraint DEP$_{IO}$(ʔ) can only trigger "few" updates. Indeed, these triplets are shown in table 5 (where they have been highlighted in light gray) to trigger only "few" updates and only towards the beginning of the run. A more precise formulation of fact 2 is provided in the appendix, together with a proof.

Fact 2 *Consider an arbitrary run of the stochastic EDRA on the Ilokano metathesis test case (2)-(3) with any of the four re-ranking rules listed in table 1. Assume that the stochastic values are sampled in between $-\Delta$ and $+\Delta$. Each of the three triplets in table 6 which has a* W *corresponding to the constraint* DEP$_{IO}$(ʔ) *can trigger only a "small" number of updates.*

Fact 2 guarantees that the current ranking vector after a few updates always satisfies the ERCs corresponding to those triplets in table 6 which have a W corresponding to the constraint DEP$_{IO}$(ʔ) and an L corresponding to the constraint IDENT$_{IO}$(syllabic). The learning dynamics is therefore governed in the long run by the remaining triplets, which are collected for convenience in the further simplified table 7. The kernel of the Ilokano metathesis test case thus consists of learning the ranking conditions highlighted in the lightgray box in (3): the two constraints LINEARITY and *ʔ] must be assigned the same ranking value while the constraint MAX_{IO}(ʔ) must slide underneath them, settling on a smaller ranking value, so that it cannot be swapped easily with the other two constraints.

3.6 Why the GLA and the minGLA succeed

We focus on the kernel of the Ilokano metathesis test case described in table 7. The dynamics of the ranking values in a run of the GLA in this case is plotted in table 8a. The horizontal axis plots the number of iterations and the vertical axis plots the ranking values of the three active constraints. The two constraints LINEARITY and *ʔ] quickly raise to their final ranking values (namely 115.88 and 116.12, respectively) and then just keep oscillating without moving away from that position.[9] The constraint MAX_{IO}(ʔ) quickly drops to its final ranking value (namely 68.0) well separated underneath the other two constraints, and then stays put there.

To gain some intuition into this ranking dynamics, suppose the GLA is trained only on the underlying form /paʔlak/. The two corresponding ERCs 1 and 2 in table 7 promote their winner-preferring constraints LINEARITY and *ʔ] and both demote their shared loser-preferring constraint MAX_{IO}(ʔ). These ERCs trigger updates until the two winner-preferring constraints are separated from the loser-preferring constraint by a distance large enough that they cannot be swapped. As soon as that ranking configuration is achieved after a few updates, the GLA stops performing any further updates and learning effectively ceases, as plotted in table 8b.

The ranking dynamics in table 8b corresponding to a run of the GLA on only the underlying form /paʔlak/ and the dynamics in table 8a corresponding to both underlying forms /paʔlak/ and /taʔo-en/ have the same shape (ignoring the oscillations in the upper branch). The reason is easy to grasp. Suppose that the GLA is trained only on the underlying form /taʔo-en/. The two corresponding triplets (/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~) and (/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~) differ because the winner and loser forms are swapped. Thus, their two corresponding ERCs 3 and 4 in table 7 are one the opposite of the other: one has a W where the other has an L. Crucially, the GLA promotes and demotes winner- and loser-preferring constraints by exactly the same amount. The updates by these two ERCs 3 and 4 thus cancel each other out: one of them displaces the two constraints LINEARITY and *ʔ] up or down and the other ERC shifts them back to their original position. If the two constraints start with the same initial ranking value, ERCs 3 and 4 do not displace them but just keep them oscillating, as in table 8c.

As all ERCs in table 7 have a single L, the re-ranking rules of the GLA and minGLA coincide in this case and the preceding considerations extend to the minGLA, leading to the following fact 3. A more precise formulation is provided in appendix C, together with a proof.

Fact 3 *Consider an arbitrary run of the stochastic EDRA on the Ilokano metathesis test case (2)-(3) with the GLA or the minGLA re-ranking rules (described in table 1). Assume that the stochastic values are sampled in between $-\Delta$ and $+\Delta$. The two ERCs 1 and 2 corresponding to the underlying form /paʔlak/ can trigger only a "small" number of updates.*

Facts 1, 2, and 3 together guarantee that all ERCs of the Ilokano metathesis test case listed in the original table 4 can trigger only a few updates but the two ERCs corresponding to (/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~) and (/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~). In other words, from a certain moment on in any run of the GLA and the minGLA, only these two triplets (/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~) and (/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~) will trigger updates. Since they differ because the winner and loser forms are swapped, their corresponding ERCs are one the opposite of the other. Since the GLA and the minGLA promote and demote by exactly the same amount, these two ERCs thus effectively maintain their two active constraints LINEARITY and *ʔ] close to each other as required to model variation. We conclude that facts 1, 2, and 3 *prove* that the GLA and the minGLA always succeed on the Ilokano metathesis test case.

3.7 Why EDCD and the CEDRA fail

We focus again on the kernel of the Ilokano metathesis test case described in table 7. The dynamics of the ranking values in a run of EDCD in this case is plotted in table 8a. The ranking values of the three active constraints LINEARITY, *ʔ], and MAX$_{\text{IO}}$(ʔ) start out all together and are never separated (the final ranking values are -1884.38, -1884.64, and -1886.98, respectively). EDCD has thus failed to learn that the constraint MAX$_{\text{IO}}$(ʔ) needs to be ranked at a safe distance underneath both constraints LINEARITY and *ʔ]. As shown by the ERC matrix in table 7, the latter ranking condition is needed to account for the deterministic mapping of /paʔlak/ to [pa.lak], thus explaining the failure of EDCD on this mapping diagnosed in table 3.

What explains the difference in behavior between EDCD and the GLA? To start, suppose that EDCD is trained on the underlying form /paʔlak/ only. Since EDCD performs no constraint promotion, the two corresponding ERCs 1 and 2 do not re-rank the two winner-preferring constraints LINEARITY and *ʔ]. But they both demote the loser-preferring constraint MAX_{IO}(ʔ). After a few updates, this loser-preferring constraint has dropped at a safe distance. EDCD thus stops performing any updates and learning effectively ceases, as plotted in table 8b. EDCD's dynamics is analogous (in scale and overall shape) to the GLA's dynamics in table 8b: when trained on /paʔlak/ only, the two algorithms behave roughly in the same way.

The crucial difference shows up when the GLA and EDCD are trained on the underlying form /taʔo-en/ which displays variation. As noted above, the two corresponding ERCs 3 and 4 are the opposite of each other: LINEARITY and *ʔ] are winner-preferring in one of the two ERCs and loser-preferring in the other. Crucially, EDCD performs constraint demotion but no constraint promotion. When trained on these two ERCs 3 and 4, EDCD thus forces LINEARITY and *ʔ] into a free fall, as shown by the ranking dynamics plotted in table 8c.[10] This dynamics is completely different from the GLA's dynamics in table 8c: LINEARITY and *ʔ] keep oscillating around themselves because the GLA's promotions compensate for the demotions.

As ERCs 3 and 4 force LINEARITY and *ʔ] into a free fall in the case of EDCD, ERCs 1 and 2 need to keep triggering updates in order to try to slide the constraint MAX_{IO}(ʔ) underneath them, without ever managing to achieve the needed separation. Indeed, table 5 for EDCD shows that ERCs 1 and 2 (corresponding to the top white rows in the table) trigger many updates (roughly 600) and remain active until the end of the simulation (close to the last 21,000th iteration). In the case of GLA, ERCs 3 and 4 instead do not displace LINEARITY and *ʔ] and ERCs 1 and 2 thus easily slide the constraint MAX_{IO}(ʔ) underneath them. Indeed, table 5 for the GLA shows that ERCs 1 and 2 trigger few updates (roughly 20) and only towards the beginning of the simulation (they become inactive after the 7,000th iteration).

The case of the CEDRA is analogous to the case of EDCD. The two learners differ because the latter performs *no* constraint promotion while the former performs *little* promotion: it crucially promotes less than it demotes. For instance, ERC 3 promotes LINEARITY by 0.5 (that is $\frac{1}{w+1}$, where $w = 1$ is the number of winner-preferring constraints relative to ERC 3) and ERC 4 demotes it by 1.[11] Overall, the ranking value of LINEARITY has thus decreased by 0.5 , while it would have decreased by 1 in the case of EDCD. In the case of the CEDRA, we thus expect the same free fall of the three constraints LINEARITY, *ʔ], and MAX_{IO}(ʔ) as in the case of EDCD, only slower. That is indeed what happens: the shape of the ranking dynamics reported in table 8 for the CEDRA is identical to the shape of the dynamics reported in table 8 for EDCD, only with a smaller scale because of a promotion amount equal to 0.5 rather than just 0.

This analysis shows that the failure of EDCD and the CEDRA does not depend on the choice of the initial ranking values. Indeed, set the initial ranking values of EDCD or the CEDRA equal to the final ranking values learned by the GLA, as reported in the first column of table 2. These ranking values already account for all the attested frequencies. A learner starting from those ranking values thus effectively has nothing to learn. Yet, EDCD and the CEDRA fail also with such a favorable choice of the initial ranking values: the underlying form /taʔo-en/ which displays variation keeps triggering updates for ever and thus forces the constraints into a free fall because nothing cancels out the demotions it triggers.

4 The Finnish genitive plurals test case

This section analyzes the performance of the stochastic EDRA on BH's Finnish genitive plurals test case (based on data and analysis from Anttila 1997b,a).

4.1 Description of the test case

This test case consists of twenty-two underlying forms paired up with two candidates each, corresponding to the two genitive plural suffixes /-jen/ and /-iden/. Fourteen of those underlying forms select their genitive form deterministically. They therefore give rise to a unique underlying/winner/loser form triplet each, listed in rows a-n of table 9. The remaining eight underlying forms display variation in the choice of the genitive plural suffix. They therefore give rise to two underlying/winner/loser form triplets each, listed in rows o-v of table 9. The second and third column of the table provide the frequency of each underlying form and the frequency of the two variants (conditioned on the underlying form). BH propose an analysis of this phonological pattern based on ten constraints described in ERC notation in the rest of table 9. The two constraints WEIGHT-TO-STRESS (WTS; no unstressed heavy syllables) and *LAPSE (no consecutive unstressed syllables) are familiar from the OT literature on stress (Prince 1983). The two constraints *H.H and *L.L prohibit consecutive heavy and consecutive light syllables. Finally the three constraints *Í, *Ó, and *Á and the three constraints *Ĭ, *Ŏ, and *Ă penalize surface stressed and unstressed syllables with underlying high/mid/low vowels.

4.2 Simulation results

BH report that the final ranking vector learned by the GLA in their simulations closely matches the attested frequencies of the suffixes /-iden/ and /-jen/ but for the three underlying forms /ministeri/ (the actual probabilities are .143 and .857 while the probabilities predicted by the ranking vector learned by the GLA are .3051 and .6949), /aleksanteri/ (the actual probabilities are .118 and .882 while the predicted probabilities are .3049 and .6951), and /naapuri/ (the actual probabilities are .369 and .631 while the predicted probabilities are .3049 and .6951). The other three stochastic error-driven learners (the minGLA, EDCD and the CEDRA) instead massively fail on this test case (detailed simulation results are provided below in subsection 4.7). How does the GLA manage to largely succeed? why does it fall on the those three problematic underlying forms? why do other implementations of stochastic error-driven learning (especially the minGLA, which is so similar to the GLA) instead fail?

4.3 Simplifying the candidate set: first round

The ERC matrix in table 9 can be substantially simplified into the reduced ERC matrix in table 10.[12] A first round of simplifications involve underlying forms that display a categorical behavior. The six underlying forms /sosialisti/, /margariini/, /edustusto/, /italiaanno/, /luonnehdinta/, and /evankelista/ in rows a-f of table 10 yield ERCs which have no loser-preferring constraints. They can thus be dropped because they impose no ranking conditions and never trigger any update. The two underlying forms /kala/ and /lasi/ in rows g-h yield identical ERCs. One of them (say /lasi/ for concreteness) can thus be dropped and its probability mass added to the mass of the other one. Thus, the probability mass reported in table 10 for /kala/ (namely .1755) is the sum of the two masses reported in table 9 for /kala/ and /lasi/ (which are .08775 each). Also the two underlying forms /luettelo/ and /televisio/ in rows i-j yield identical ERCs. One of them (say /televisio/ for concreteness) can thus be dropped and its probability mass added to the mass of the other one. Finally, the three underlying forms /kamera/, /ajattelija/, and /taiteilija/ in rows k-m also yield three identical ERCs. Two of them (say /ajattelija/ and /taiteilija/ for concreteness) can be dropped and their mass added to the mass of the remaining one.

4.4 Simplifying the candidate set: second round

A second round of simplifications of BH's Finnish genitive test case involve underlying forms that display variation. The two underlying forms /korjaamo/ and /koordinaatisto/ in rows o-p of table 9 each yield variation between the same two ERCs. The probabilities of these two ERCs conditioned on the two underlying forms are almost identical. One of the two underlying forms (say /koordinaatisto/ for concreteness) can thus be dropped and its probability mass added to the mass of the other one. Also the three underlying forms /aleksanteri/, /ministeri/, and /naapuri/ in rows q-s each yield variation between the same two ERCs. The probabilities of these two ERCs conditioned on the first two underlying forms /aleksanteri/ and /ministeri/ are almost identical, as reported in the third column. Yet, the probabilities of these two ERCs conditioned on the last underlying form /naapuri/ are very different. BH's constraint set is therefore not sufficiently fine-grained to capture the difference between the two patterns of variation corresponding to /naapuri/ on the one hand and to /aleksanteri/ and /ministeri/ on the other hand. The failure of the GLA on these three underlying forms reported above in subsection 4.2 is thus due not to the learning procedure but to an insufficiency of the constraint set.

The incorrect frequencies predicted by the GLA for these three underlying forms can be explained as follows. The frequencies of the three underlying forms /aleksanteri/, /ministeri/, and /naapuri/ are $q^{\text{aleksanteri}} = .0029$, $q^{\text{ministeri}} = .0479$ and $q^{\text{naapuri}} = .1023$, respectively. As they do not add up to 1, we normalize them as $Q^{\text{aleksanteri}} = .0194$, $Q^{\text{ministeri}} = .3127$ and $Q^{\text{naapuri}} = .6678$.[13] Let $p_1^{\text{aleksanteri}} = .88$ and $p_2^{\text{aleksanteri}} = .12$ be the probabilities of the two shared ERCs conditioned on the underlying form /aleksanteri/; let $p_1^{\text{ministeri}} = .86$ and $p_2^{\text{ministeri}} = .14$ be the probabilities of those two shared ERCs conditioned on the underlying form /ministeri/; finally, let $p_1^{\text{naapuri}} = .63$ and $p_2^{\text{naapuri}} = .37$ be the probabilities of those two shared ERCs conditioned on the underlying form /naapuri/. We define the probabilities P_1, P_2 of these two shared ERCs as in (5), namely as the average of the three original distributions for those two ERCs, weighted by the (normalized) frequencies of the three corresponding underlying forms.

(5) $P_1 = Q^{\text{aleksanteri}} p_1^{\text{aleksanteri}} + Q^{\text{ministeri}} p_1^{\text{ministeri}} + Q^{\text{naapuri}} p_1^{\text{naapuri}} = 0.7067$

$P_2 = Q^{\text{aleksanteri}} p_2^{\text{aleksanteri}} + Q^{\text{ministeri}} p_2^{\text{ministeri}} + Q^{\text{naapuri}} p_2^{\text{naapuri}} = 0.2932$

The probabilities P_1 and P_2 just computed are indeed the probabilities learned by the GLA in BH's simulations for the three underlying forms /aleksanteri/, /ministeri/, and /naapuri/ as reported above in subsection 4.2. These are also the probabilities assumed in table 10.

4.5 Simplifying the candidate set: third round

A third round of simplifications of BH's Finnish genitive test case involve interactions between underlying forms that display a categorical behavior and underlying forms that display variation. The underlying form /poliisi/ in table 9t yields variation between the triplet (/poliisi/, [pó.lii.sèi.den], ~~[pó.lii.si.en]~~) which has a very small conditional probability (namely .02) and the triplet (/poliisi/, [pó.lii.si.en], ~~[pó.lii.sèi.den]~~) which instead has a very large probability (namely .98) and furthermore yields the same ERC as the categorical triplet (/avantgardisti/, [á.vant.gàr.dis.ti.en], ~~[á.vant.gàr.dis.tèi.den]~~). We can thus drop the underlying form /poliisi/ and add its probability mass to the categorical underlying form /avantgardisti/. Analogous considerations hold for the underlying form /hetero/ in table 9u. The corresponding triplet (/hetero/, [hé.te.ròi.den], ~~[hé.te.ro.jen]~~) has very large conditional probability (namely .99) and yields the same ERC as the categorical triplet (/luettelo/, [lú.et.te.lòi.den], ~~[lú.et.te.lo.jen]~~). Once again, we can thus drop the underlying form /hetero/ and add its probability mass to the categorical underlying form /luettelo/. In the end, BH's original Finnish genitive test case can be substantially simplified as in table 10.[14]

4.6 Simplifying the constraint set

The constraint set used by BH and listed in tables 9 and 10 is a subset of Anttila's (1997) original constraint set. The restriction to this subset is motivated by BH through the observation that "we found that we could derive the corpus frequencies accurately using only a subset of [Anttila's] constraints" (p. 68). It turns out that the constraint set can be substantially further simplified: the four constraints *LAPSE, *Ó, *Á, and *Ĭ can be dropped without affecting the ability of stochastic OT to model the corpus frequencies accurately.[15] The underlying/winner/loser form triplet (/avantgardisti/, [á.vant.gàr.dis.ti.en], ~~[á.vant.gàr.dis.tèi.den]~~)

can then also be dropped, because it has no loser-preferring constraints once *LAPSE and *Ĭ are dropped.[16] In conclusion, the ERC matrix in table 10 is further simplified to table 11.

The ranking conditions (6) required by the Finnish test case are easily determined from table 11. The underlying form /koorjamo/ requires *H.H and *Ŏ to stay close. Hence, the underlying form /luttelo/ requires *L.L to be ranked above *H.H. Analogously, the underlying form /maailma/ requires *H.H and *Ă to stay close and the underlying form /kamera/ thus requires *L.L to be ranked above *H.H. The latter two blocks are completely analogous, but for the fact that it is *Ŏ which is active in one case and *Ă that is active in the other case. The underlying form /kala/ requires WTS to be ranked well above *L.L. As *L.L is deterministically ranked above *H.H, the underlying form /naapuri/ requires Í and *L.L to stay close.

(6)
```
WTS
 |
*L.L ------ *Í
 |
*H.H ----- *Ŏ ----- *Ă
```

4.7 Explaining the simulation results

We have run the stochastic EDRA on the simplified Finnish test case described in table 11 with the four re-ranking rules in table 1.[17] Table 12 plots the ranking dynamics for each re-ranking rule. Table 13 provides the final ranking values and table 14 compares the frequencies of each underlying/surface form mapping predict by those ranking values to the actual frequencies.

The GLA succeeds for the by now familiar reason. Since it promotes and demotes by exactly the same amount, the three underlying forms /korjaamo/, /maailma/, and /naapuri/ which display variation keep their active constraints oscillating but do not displace them. The three categorical underlying forms /kala/, /luttelo/, and /kamera/ thus have the time to space the constraints apart as required by the categorical backbone of the target ranking (6). The ranking dynamics in table 12a indeed shows the six constraints nicely organizing into the required three strata. And the final ranking vector matches the attested frequencies, as shown in table 14.

EDCD and the CEDRA fail on the Finnish genitive test case for the by now familiar reason. Constraint WTS is never demoted because it is never loser-preferring in table 11. The other constraints are instead all forced into a free fall. This is due to the fact that the three underlying forms /korjaamo/, /maailma/, and /naapuri/ which display variation keep triggering updates forever and the resulting demotions are not balanced by corresponding promotions, because EDCD and the CEDRA perform no or only little constraint promotion. The constraint *Ă drops more slowly than the other four constraints, as shown by the ranking dynamics for EDCD and the CEDRA in table 12c-d. The reason is that *Ă is only demoted by the triplet (/maailma/, [máa.il.mo.jen], ~~[máa.il.mòi.den]~~) whose underlying form is extremely infrequent. Table 14 indeed shows that this underlying form is effectively treated as categorical by the final ranking vectors learned by EDCD and the CEDRA.

Contrary to the GLA, the minGLA fails on the Finnish test case: table 14 shows that the final ranking vector learned by the minGLA fails at matching the attested frequencies.[18] The minGLA differs from the GLA only because the latter demotes all loser-preferring constraints while the former only demotes the one with the largest stochastic ranking value. This means that the GLA and the minGLA come apart only in response to ERCs with two or more L's. In the specific case of the simplified Finnish test case, the behavior of the minGLA and the GLA can thus come apart only on the ERC corresponding to (/naapuri/, [náa.pu.rèi.den], ~~[náa.pu.ri.en]~~), which is the only one in table 11 with two L's.

Since it displays variation, the underlying form /naapuri/ will keep triggering updates for ever, as the current stochastic ranking vector will never be able to satisfy both contradictory ERCs corresponding to the two triplets (/naapuri/, [náa.pu.rèi.den], ~~[náa.pu.ri.en]~~) and (/naapuri/, [náa.pu.ri.en], ~~[náa.pu.rèi.den]~~). The latter triplet will always promote both *H.H and *Í. The former triplet instead will always demote only one of the two, namely the one with the largest stochastic ranking value. As a result, the combined effect of the two ERCs corresponding to /naapuri/ is that the ranking value of *H.H keeps increasing indefinitely over time. And the minGLA thus does not manage to slide *L.L above *H.H. The ranking dynamics plotted in table 12b indeed shows that all constraints raise and the learner fails at spacing them apart.[19]

This analysis entails that the failure of the minGLA does not depend on the choice of the initial ranking values. Indeed, set the initial ranking values of the minGLA equal to the final ranking values learned by the GLA, as reported in table 13. These ranking values already account for all the attested frequencies. But the minGLA fails also with such a favorable choice of the initial ranking values: the underlying form /naapuri/ which displays variation keeps triggering updates for ever and thus forces the constraints into a free raise because its two corresponding ERCs overall promote more than they demote.

5 Conclusions

The language acquisition and learnability literature (at least since Wexler and Culicover 1980) has endorsed error-driven learning because it imposes no memory requirements (training data are not stored but discarded after each update), because it straightforwardly models acquisition gradualness (child acquisition paths can be modeled through the predicted sequences of grammars), and because of its noise robustness (a single piece of corrupted data only has a tiny effect on the final grammar).[20] A core issue of the theory of error-driven learning within OT concerns the specifics of the re-ranking rule used to revise the current ranking whenever it makes an error on the current piece of data. There are two crucial choice-points in the theory. The first concerns which loser-preferring constraints are demoted: all or just a subset of them (say, just the undominated ones or just the top ranked one)? The second choice point concerns the promotion amount: should it be equal to the demotion amount or smaller than that (and thus possibly null)? Some combinations of these options pursued in the literature are summarized in table 1.

The issue of the proper definition of the re-ranking rule has been tackled from two perspectives. One perspective focuses on categorical phonology within deterministic OT and asks the following question: which re-ranking rule ensures that the deterministic EDRA converges to a constraint ranking which captures the categorical pattern? In this categorical context, *convergence* means that the EDRA eventually stops making errors. Another perspective focuses on phonological variation within stochastic OT and asks the following question: which re-ranking rule ensures that the stochastic EDRA converges to a ranking vector which captures the pattern of variation? In this stochastic context, *convergence* means that, although the EDRA keeps making updates for ever (because of the competition between variants), the ranking values eventually stop shifting away and instead oscillate around a mid point which remains fixed.

Unfortunately, these two perspectives lead to contradicting conclusions. From the categorical perspective, it is better for the promotion amount to be smaller rather than equal to the demotion amount; and it is better to demote only the loser-preferring constraints which are undominated rather than all the loser-preferring constraints (Pater 2008; Magri 2012). In other words, from this perspective, EDCD and the CEDRA outperform the GLA. But from the perspective of variation, it is better for the promotion amount to be exactly equal to the demotion amount; and it is better to demote all the loser-preferring constraints rather than just the undominated ones or the top-ranked one. In other words, from this perspective, the GLA outperforms EDCD and the CEDRA.

This paper has explained in detail why the latter conclusion holds. Patterns of variation lead to pairs of contradicting ERCs. These contradicting ERCs keep triggering updates for ever, no matter which re-ranking rule is chosen. These updates thus need to cancel each other out, so that they do not indefinitely displace the constraints. This canceling out requires demotions and promotions to exactly balance each other. Thus, the promotion amount cannot be smaller than the demotion amount (otherwise, these ERCs do not promote enough, forcing their active constraints into a free fall). And no loser-preferring constraint can forgo demotion (otherwise, these ERCs do not demote enough, forcing their active constraints into a free raise). Unfortunately, we have to leave unresolved for the time being this tension between the two opposing conclusions reached from the the two perspectives of categorical phonology and variation in stochastic OT.

Notes

[0]The research reported in this paper has been supported by a grant from the Agence National de la Recherche (project title: ‘The mathematics of segmental phonotactics’) and by a grant from the MIT France Seed Fund Award (project title: ‘Phonological Typology and Learnability’). We would like to thank an anonymous *LI* reviewer for very useful comments.

[1]This can be achieved for instance by clipping the original gaussian outside of $[-\Delta, +\Delta]$.

[2] BH assume that the demotion and promotion amounts d and p are multiplied by a *step-size* or *plasticity* which decreases with time, making the updates smaller towards the end of the run. We consider plasticity only in the simulations (see footnotes 5 and 17), while in the analyses we ignore it (namely, assume it is equal to 1) for simplicity.

[3]The re-ranking rule originally proposed by Tesar and Smolensky can be easily restated in

terms of ranking vectors, yielding (apart from a glitch discussed in Boersma 2009) the non-gradual counterpart of the algorithm that we refer to here as *gradual* EDCD; see Magri (2012, section 3) for details.

[4]In the case of EDCD, the CEDRA, and the minGLA, the same stochastic values $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ are used twice: once at step (b), to compute the current prediction $[z]$; and then again at step (d), to determine which loser-preferring constraints are demoted.

[5] As in BH, each simulation consists of three learning stages of 7,000 iterations each; plasticity (see footnote 2) is equal to 2.0, 0.2, and 0.02 in the three stages; the distribution $\mathcal{D}$ is a gaussian with zero mean and variance equal to 10.0, 2.0, and 2.0 in the three stages.

[6]The final ranking vector reported here for the GLA is almost identical to the one reported by BH. They note that multiple runs of the GLA yield almost identical final ranking vectors. The same holds for the other three re-ranking rules according to our simulations.

[7]As in BH, the probabilities in table 3 are obtained by sampling 100,000 times the stochastic OT grammar with a gaussian distribution with zero mean and variance 2.0.

[8]Recall from footnote 2 that we assume plasticity to be equal to 1 for simplicity.

[9]The amplitude of the oscillations decreases with time, due to the fact that plasticity decreases after each stage of 7,000 iterations (see footnote 5).

[10]The slope of the fall decreases with time, due to the fact that plasticity decreases after each stage of 7,000 iterations (see footnote 5).

[11]Both amounts are actually multiplied by the plasticity corresponding to the current iteration.

[12]Goldwater and Johnson (2003) also report that they have been able to simplify the Finnish test case. Yet, they do not provide specifics on what their simplified test case looks like and therefore we cannot build on their observation.

[13]To illustrate, $Q^{\text{aleksanteri}}$ is computed as $Q^{\text{aleksanteri}} = \dfrac{q^{\text{aleksanteri}}}{q^{\text{aleksanteri}} + q^{\text{ministeri}} + q^{\text{naapuri}}} = .0194$.

[14]The probabilities of the underlying forms in the second column of table 10 need to be normalized to add up to 1, compensating for the mass initially allocated to the six underlying forms in rows a-f of table 10, that have been dropped because they have no L's.

[15]The three constraints *Ó, *Á, and *Ĭ that we drop head columns in the original table 9 which are the opposite (namely, have an L instead of a W and vice versa) of the columns headed by

*Ŏ, *Ă, and *Í, respectively. Furthermore, the fourth constraint *LAPSE that we drop heads a column in table 9 which is the opposite of the column headed by *H.H apart for eight ERCs: the six ERCs a-f, which have no L's and can thus be ignored; and the two ERCs corresponding to the categorical underlying forms /kala/ and /lasi/ which trigger only few updates because they have a W corresponding to the constraint WTS which has no L's and is therefore never demoted. Indeed, the GLA in BH's simulations (see BH's (29) on p. 70) assigns to these four constraints the smallest ranking values, whereby they play no role in predicting the attested frequencies.

[16]Indeed, this triplet triggers no updates in our simulations on the original test case in table 9.

[17] As in BH, each simulation consists of five learning stages; the first four stages consist of 22,000 iterations each while the fifth stage consists of 300,000 iterations; plasticity is equal to 2.0, 2.0, 0.2, 0.02, and 0.002 in the five stages; the distribution $\mathcal{D}$ is a gaussian with zero mean and variance equal to 10.0 in the first stage and to 2.0 in the following four stages.

[18]Another variant of the GLA which fails on the Finnish test case is the stochastic EDRA which (like the GLA) promotes all winner-preferring constraints by 1 but (like EDCD and the CEDRA) only demotes the loser-preferring constraints which are undominated relative to the current stochastic ranking vector (rather than just the highest ranked one, like the minGLA).

[19]The ranking values effectively keep increasing for ever, although the rate of increase in the last learning stage is very small because the plasticity is very small (0.002) in the last stage.

[20]Non-error-driven learning requires special *ad hoc* provisions in order to achieve these advantages, such as an iterated implementation through multiple batches of data to model gradualness, a stochastic implementation to achieve robustness, a developing proto-lexicon to get around the lack of a stored fully-blown lexicon in the early acquisition stages; see Gibson and Wexler (1994, p. 410) for discussion.

References

Anttila, Arto. 1997a. Deriving variation from grammar: A study of Finnish genitives. In *Variation, change and phonological theory*, ed. Frans Hinskens, Roeland van Hout, and Leo Wetzels, 35–68. Amsterdam: John Benjamins. Rutgers Optimality Archive ROA-63, http://ruccs.rutgers.edu/roa.html.

Anttila, Arto. 1997b. Variation in finnish phonology and morphology. Doctoral Dissertation, Stanford University.

Boersma, Paul. 1997. How we learn variation, optionality and probability. In *Proceedings of the Institute of Phonetic Sciences (IFA) 21*, ed. Rob van Son, 43–58. University of Amsterdam: Institute of Phonetic Sciences.

Boersma, Paul. 1998. Functional phonology. Doctoral Dissertation, University of Amsterdam, The Netherlands. The Hague: Holland Academic Graphics.

Boersma, Paul. 2009. Some correct error-driven versions of the constraint demotion algorithm. *Linguistic Inquiry* 40:667–686.

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

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

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–120. Stockholm University.

Hayes, Bruce, and Mary Abad. 1989. Reduplication and syllabification in ilokano. *Lingua* 77:331–374.

Magri, Giorgio. 2012. Convergence of error-driven ranking algorithms. *Phonology* 29:213–269.

Pater, Joe. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39:334–345.

Prince, Alan. 1983. Relating to the grid. *Linguistic Inquiry* 14:19–100.

Prince, Alan. 2002. Entailed ranking arguments. URL `http://roa.rutgers.edu/files/500-0202/500-0202-PRINCE-0-1.PDF`, manuscript (Rutgers University). Available from the Rutgers Optimality Archive as ROA 500.

Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. Oxford: Blackwell. URL `http://roa.rutgers.edu`, original version, Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, April 1993. Available from the Rutgers Optimality Archive as ROA 537.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Wexler, Kenneth, and Peter W. Culicover. 1980. *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Giorgio Magri
`magrigrg@gmail.com`
SFL UMR 7023 (UPL, CNRS, University of Paris 8)
61 Rue Pouchet, 75017 Paris, France

Benjamin Storme
`benjastorme@hotmail.com`
SFL UMR 7023 (UPL, CNRS, University of Paris 8)
61 Rue Pouchet, 75017 Paris, France

re-ranking rule	promotion amount	which constraints are demoted
(Gradual) Error-Driven Constraint Demotion (EDCD; Tesar and Smolensky 1998)	$p = 0$	only the undominated loser-preferrers
Gradual Learning Algorithm (GLA; Boersma 1997, 1998)	$p = 1$	all loser-preferrers
Minimal Gradual Learning Algorithm (minGLA; Boersma 1997, 1998)	$p = 1$	only the highest loser-preferrer
Calibrated Error-Driven Ranking Algorithm (CEDRA; Magri 2012)	$p = \frac{1}{w+1}$	only the undominated loser-preferrers

Table 1: Four re-ranking rules for the stochastic EDRA.

GLA		minGLA		EDCD		CEDRA	
IDENT$_{IO}$[low]	142.0	IDENT$_{IO}$[low]	154.0	ONSET	100.0	ONSET	113.0
MAX$_{IO}$(V)	140.0	*LOWGLIDE	152.0	MAX$_{OO}$(ʔ)	100.0	IDENT$_{IO}$[low]	111.3
ONSET	138.0	ONSET	152.0	MAX$_{IO}$(V)	100.0	MAX$_{IO}$(V)	111.0
*LOWGLIDE	138.0	MAX$_{IO}$(V)	150.0	IDENT$_{IO}$[low]	100.0	*LOWGLIDE	108.7
*[ʔC	114.0	MAX$_{OO}$(ʔ)	120.0	*[ʔC	100.0	*[ʔC	103.0
MAX$_{OO}$(ʔ)	110.0	*[ʔC	120.0	*LOWGLIDE	100.0	MAX$_{OO}$(ʔ)	100.0
DEP$_{IO}$(ʔ)	98.0	DEP$_{IO}$(ʔ)	106.0	DEP$_{IO}$(ʔ)	50.0	DEP$_{IO}$(ʔ)	64.0
LINEARITY	67.0	LINEARITY	81.14	IDENT$_{IO}$[syl]	10.0	IDENT$_{IO}$[syl]	22.0
*ʔ]	67.0	*ʔ]	80.86	*ʔ]	-897.9	*ʔ]	-304.3
MAX$_{IO}$(ʔ)	24.0	IDENT$_{IO}$[syl]	72.0	LINEARITY	-898.0	LINEARITY	-304.6
IDENT$_{IO}$[syl]	24.0	MAX$_{IO}$(ʔ)	38.0	MAX$_{IO}$(ʔ)	-900.8	MAX$_{IO}$(ʔ)	-309.1

Table 2: Ranking values learned by the stochastic EDRA with the four re-ranking rules listed in table 1 on the Ilokano metathesis test case described in (2)-(3).

	actual	GLA	minGLA	EDCD	CEDRA
(/paʔlak/, [pa.lak])	1.0	1.0	1.0	**0.75**	**0.91**
(/paʔlak/, [paʔ.lak])	0.0	0.0	0.0	**0.12**	**0.04**
(/paʔlak/, [pal.ʔak])	0.0	0.0	0.0	**0.12**	**0.05**
(/paʔlak/, [pa.ʔlak])	0.0	0.0	0.0	0.0	0.0
(/ʔajo-en/, [ʔaj.wen])	1.0	1.0	1.0	1.0	1.0
(/ʔajo-en/, [ʔa.jen])	0.0	0.0	0.0	0.0	0.0
(/ʔajo-en/, [ʔa.jo.ʔen])	0.0	0.0	0.0	0.0	0.0
(/ʔajo-en/, [ʔa.jo.en])	0.0	0.0	0.0	0.0	0.0
(/basa-en/, [ba.sa.ʔen])	1.0	1.0	1.0	1.0	1.0
(/basa-en/, [bas.a̯en])	0.0	0.0	0.0	0.0	0.0
(/basa-en/, [ba.sen])	0.0	0.0	0.0	0.0	0.0
(/basa-en/, [ba.sa.en])	0.0	0.0	0.0	0.0	0.0
(/basa-en/, [bas.wen])	0.0	0.0	0.0	0.0	0.0
(/taʔo-en/, [taʔ.wen])	0.5	0.49	0.54	0.49	0.45
(/taʔo-en/, [taw.ʔen])	0.5	0.50	0.46	0.51	0.54
(/taʔo-en/, [ta.ʔo.en])	0.0	0.0	0.0	0.0	0.0
(/taʔo-en/, [ta.ʔen])	0.0	0.0	0.0	0.0	0.0
(/taʔo-en/, [ta.wen])	0.0	0.0	0.0	0.0	0.0
(/taʔo-en/, [ta.ʔwen])	0.0	0.0	0.0	0.0	0.0
(/taʔo-en/, [ta.ʔo.ʔen])	0.0	0.0	0.0	0.0	0.0

Table 3: Probabilities of each underlying/surface form mapping predicted by the four final ranking vectors reported in table 2.

(/x/, [y], ~~[z]~~)	ONSET	MAX$_{IO}$(V)	*LOWGLIDE	IDENT$_{IO}$[low]	*[ʔC	MAX$_{OO}$(ʔ)	DEP$_{IO}$(ʔ)	IDENT$_{IO}$[syl]	LINEARITY	*ʔ]	MAX$_{IO}$(ʔ)
(/paʔlak/, [pa.lak], ~~[pa.ʔlak]~~)					W						L
(/paʔlak/, [pa.lak], ~~[pal.ʔak]~~)									W		L
(/paʔlak/, [pa.lak], ~~[paʔ.lak]~~)										W	L
(/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jen]~~)		W						L			
(/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jo.en]~~)	W							L			
(/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jo.ʔen]~~)							W	L			
(/basa-en/, [ba.sa.ʔen], ~~[bas.a̯en]~~)			W				L	W			
(/basa-en/, [ba.sa.ʔen], ~~[bas.wen]~~)				W			L	W			
(/basa-en/, [ba.sa.ʔen], ~~[ba.sen]~~)		W					L				
(/basa-en/, [ba.sa.ʔen], ~~[ba.sa.en]~~)	W						L				
(/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~)									L	W	
(/taʔo-en/, [taw.ʔen], ~~[ta.wen]~~)						W			L		W
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔen]~~)		W						L	L		
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔo.en]~~)	W							L	L		
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔo.ʔen]~~)							W	L	L		
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔwen]~~)					W				L		
(/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~)									W	L	
(/taʔo-en/, [taʔ.wen], ~~[ta.wen]~~)						W				L	W
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔen]~~)		W						L		L	
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔo.en]~~)	W							L		L	
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔo.ʔen]~~)							W	L		L	
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔwen]~~)					W					L	

Table 4: ERC description of the Ilokano metathesis test case (2)-(3)

(/x/, [y], ~~[z]~~)	GLA #U	GLA LU	GLA LI	minGLA #U	minGLA LU	minGLA LI	EDCD #U	EDCD LU	EDCD LI	CEDRA #U	CEDRA LU	CEDRA LI
(/paʔlak/, [pa.lak], ~~[pa.ʔlak]~~)	2	26	43	5	121	294	0	0	0	4	15	20
(/paʔlak/, [pa.lak], ~~[pal.ʔak]~~)	19	822	5,398	17	1,021	6,380	630	3,832	20,994	223	3,096	20,965
(/paʔlak/, [pa.lak], ~~[paʔ.lak]~~)	21	899	6,059	19	621	3,413	589	3,829	20,988	205	3,092	20,938
(/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jen]~~)	1	4	6	1	2	2	0	0	0	1	8	10
(/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jo.en]~~)	2	5	9	4	55	104	3	15	20	3	46	81
(/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jo.ʔen]~~)	10	236	975	32	1,080	6,940	20	1,357	6,997	20	723	4,143
(/basa-en/, [ba.sa.ʔen], ~~[bas.a̯en]~~)	18	584	3,611	26	1,081	6,946	3	891	4,494	13	407	2,054
(/basa-en/, [ba.sa.ʔen], ~~[bas.wen]~~)	20	875	5,894	27	920	5,632	5	200	735	16	944	5,626
(/basa-en/, [ba.sa.ʔen], ~~[ba.sen]~~)	16	824	5,410	20	978	6,027	8	656	3,288	7	235	951
(/basa-en/, [ba.sa.ʔen], ~~[ba.sa.en]~~)	15	997	6,881	19	1,069	6,890	8	1,110	5,687	9	911	5,436
(/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~)	1,275	2,730	20,968	1,285	2,839	20,995	1,253	3,826	20,967	1,269	3,095	20,951
(/taʔo-en/, [taw.ʔen], ~~[ta.wen]~~)	2	50	101	6	829	4,907	0	0	0	2	34	61
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔen]~~)	1	13	19	3	57	114	1	5	6	1	29	51
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔo.en]~~)	1	17	27	1	53	99	2	52	109	0	0	0
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔo.ʔen]~~)	24	782	5,091	35	1,082	6,947	21	1,339	6,904	16	686	3,879
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔwen]~~)	1	159	546	2	1,084	6,957	2	28	52	0	0	0
(/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~)	1,262	2,733	21,000	1,300	2,840	21,000	1,274	3,830	20,989	1,288	3,100	20,996
(/taʔo-en/, [taʔ.wen], ~~[ta.wen]~~)	3	78	207	4	946	5,821	0	0	0	0	0	0
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔen]~~)	1	42	79	1	26	50	2	14	19	2	7	8
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔo.en]~~)	2	51	102	2	82	182	2	45	90	0	0	0
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔo.ʔen]~~)	32	965	6,609	28	806	4,756	8	575	2,773	21	936	5,586
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔwen]~~)	5	559	3,388	3	271	1,015	1	42	80	0	0	0

Table 5: Contribution of each underlying/winner/loser form triplet $(/x/, [y], \not{[z]})$ in table 4 to the learning dynamics: number of updates triggered by that triplet (#U); number of updates overall performed before that triplet has triggered its last update (LU); iteration at which that triplet has triggered its last update (LI).

(/x/, [y], ~~[z]~~)	ONSET	MAX$_{IO}$(V)	*LOWGLIDE	IDENT$_{IO}$[low]	*[ʔC	MAX$_{OO}$(ʔ)	DEP$_{IO}$(ʔ)	IDENT$_{IO}$[syl]	LINEARITY	*ʔ]	MAX$_{IO}$(ʔ)
(/paʔlak/, [pa.lak], ~~[pal.ʔak]~~)									W		L
(/paʔlak/, [pa.lak], ~~[paʔ.lak]~~)										W	L
(/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jo.ʔen]~~)							W	L			
(/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~)									L	W	
(/taʔo-en/, [taw.ʔen], ~~[ta.ʔo.ʔen]~~)							W	L	L		
(/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~)									W	L	
(/taʔo-en/, [taʔ.wen], ~~[ta.ʔo.ʔen]~~)							W	L		L	

Table 6: ERC description of the Ilokano metathesis test case after the first round of simplifications.

(/x/, [y], ~~[z]~~)	ONSET	MAX$_{IO}$(V)	*LOWGLIDE	IDENT$_{IO}$[low]	*[ʔC	MAX$_{OO}$(ʔ)	DEP$_{IO}$(ʔ)	IDENT$_{IO}$[syl]	LINEARITY	*ʔ]	MAX$_{IO}$(ʔ)
ERC 1 = (/paʔlak/, [pa.lak], ~~[pal.ʔak]~~)									W		L
ERC 2 = (/paʔlak/, [pa.lak], ~~[paʔ.lak]~~)										W	L
ERC 3 = (/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~)									W	L	
ERC 4 = (/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~)									L	W	

Table 7: ERC description of the kernel of the Ilokano metathesis test case after the second round of simplifications.

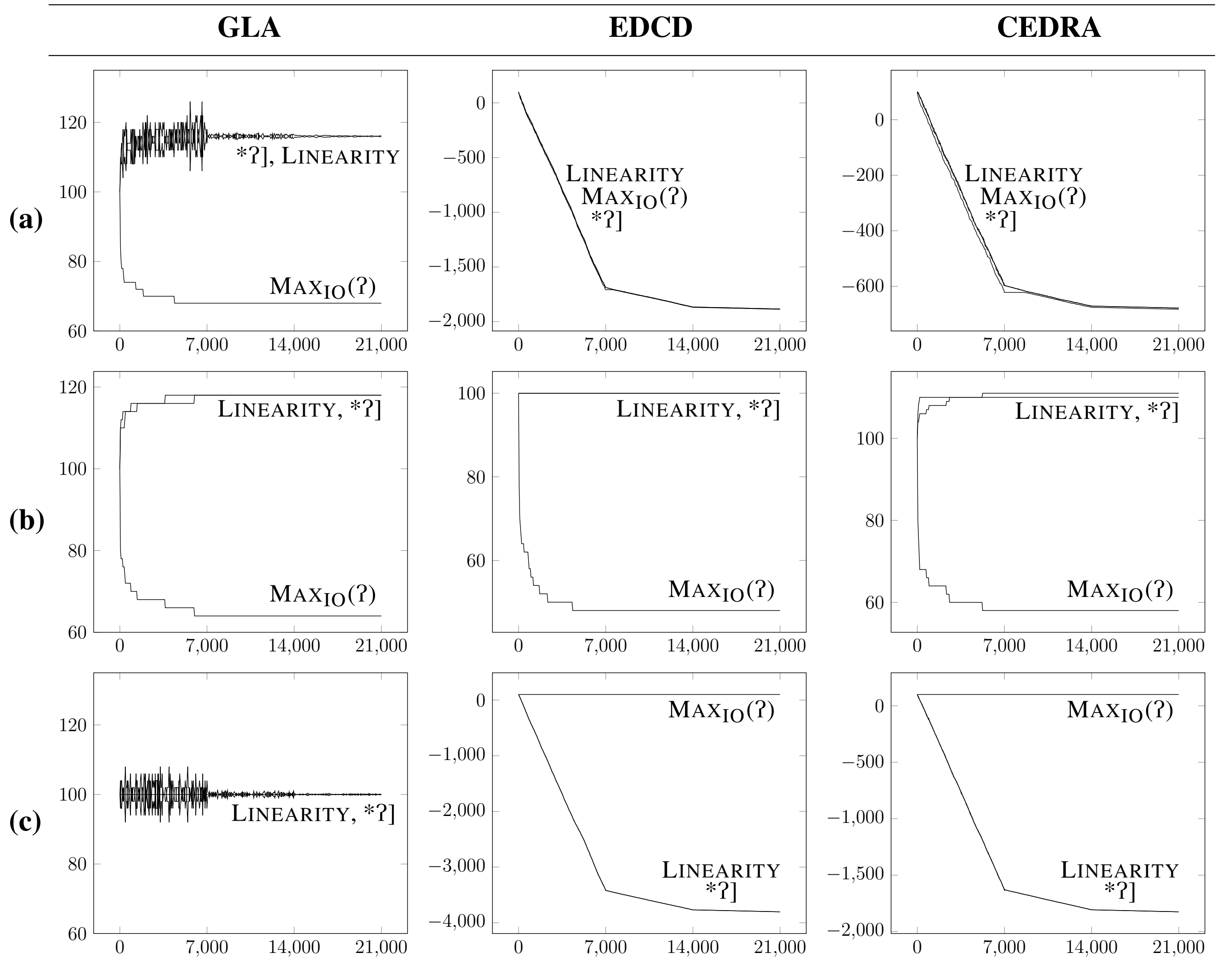


Table 8: Ranking dynamics of the three constraints LINEARITY, *ʔ], and MAX$_{IO}$(ʔ) when the GLA, EDCD, and the CEDRA are trained on: **(a)** both underlying forms /paʔlak/ and /taʔo-en/ in table 7; **(b)** only the underlying form /paʔlak/; **(c)** only the underlying form /taʔo-en/.

	undrlying/winner/loser form triplet (/x/, [y], ~~[z]~~)	p(x)	p(y|x)	WTS	*Í	*L.L	*Ă	*H.H	*Ŏ	*LAPSE	*Ó	*Á	*Ĭ
a.	(/sosialisti/, [só.si.a.lìs.ti.en], ~~[só.si.a.lìs.tei.den]~~)	.01737		W				W					
b.	(/margariini/, [már.ga.rìi.ni.en], ~~[már.ga.rìi.nei.den]~~)	.1292		W				W					
c.	(/edustusto/, [é.dus.tùs.to.jen], ~~[é.dus.tùs.toi.den]~~)	.01474		W				W					
d.	(/italiaanno/, [í.ta.li.àa.no.jen], ~~[í.ta.li.àa.noi.den]~~)	.0001		W				W					
e.	(/luonnehdinta/, [lúon.neh.dìn.to.jen], ~~[lúon.neh.dìn.toi.den]~~)	.0001		W				W					
f.	(/evankelista/, [é.van.ke.lìs.to.jen], ~~[é.van.ke.lìs.toi.den]~~)	.0003		W				W					
g.	(/kala/, [ká.lo.jen], ~~[ká.loi.den]~~)	.0877		W		L		W					
h.	(/lasi/, [lá.si.en], ~~[lá.sei.den]~~)	.0877		W		L		W					
i.	(/luettelo/, [lú.et.te.lòi.den], ~~[lú.et.te.lo.jen]~~)	.0044				W		L	W	W	L		
j.	(/televisio/, [té.le.vi.si.òi.den], ~~[té.le.vi.si.o.jen]~~)	.0072				W		L	W	W	L		
k.	(/kamera/, [ká.me.ròi.den], ~~[ká.me.ro.jen]~~)	.1264				W	W	L		W		L	
l.	(/ajattelija/, [á.jat.te.li.jòi.den], ~~[á.jat.te.li.jo.jen]~~)	.0177				W	W	L		W		L	
m.	(/taiteilija/, [tái.tei.li.jòi.den], ~~[tái.tei.li.jo.jen]~~)	.0484				W	W	L		W		L	
n.	(/avantgardisti/, [á.vant.gàr.dis.ti.en], ~~[á.vant.gàr.dis.tèi.den]~~)	.0003			W			W		L			L
o.	(/korjaamo/, [kór.jaa.mo.jen], ~~[kór.jaa.mòi.den]~~)	.0748	.82					W	L	L	W		
	(/korjaamo/, [kór.jaa.mòi.den], ~~[kór.jaa.mo.jen]~~)		.18					L	W	W	L		
p.	(/koordinaatisto/, [kóor.di.nàa.tis.to.jen], ~~[kóor.di.nàa.tis.tòi.den]~~)	.0017	.8					W	L	L	W		
	(/koordinaatisto/, [kóor.di.nàa.tis.tòi.den], ~~[kóor.di.nàa.tis.to.jen]~~)		.2					L	W	W	L		
q.	(/aleksanteri/, [á.lek.sàn.te.ri.en], ~~[á.lek.sàn.te.rèi.den]~~)	.0029	.88		W	L		W		L			L
	(/aleksanteri/, [á.lek.sàn.te.rèi.den], ~~[á.lek.sàn.te.ri.en]~~)		.12		L	W		L		W			W
r.	(/ministeri/, [mí.nis.te.ri.en], ~~[mí.nis.te.rèi.den]~~)	.0479	.86		W	L		W		L			L
	(/ministeri/, [mí.nis.te.rèi.den], ~~[mí.nis.te.ri.en]~~)		.14		L	W		L		W			W
s.	(/naapuri/, [náa.pu.rèi.den], ~~[náa.pu.ri.en]~~)	.1023	.37		L	W		L		W			W
	(/naapuri/, [náa.pu.ri.en], ~~[náa.pu.rèi.den]~~)		.63		W	L		W		L			L
t.	(/poliisi/, [pó.lii.sèi.den], ~~[pó.lii.si.en]~~)	.1437	.02		L			L		W			W
	(/poliisi/, [pó.lii.si.en], ~~[pó.lii.sèi.den]~~)		.98		W			W		L			L
u.	(/hetero/, [hé.te.ròi.den], ~~[hé.te.ro.jen]~~)	.0686	.99			W		L	W	W	L		
	(/hetero/, [hé.te.ro.jen], ~~[hé.te.ròi.den]~~)		.01			L		W	L	L	W		
v.	(/maailma/, [máa.il.mo.jen], ~~[máa.il.mòi.den]~~)	.0159	.5				L	W		L		W	
	(/maailma/, [máa.il.mòi.den], ~~[máa.il.mo.jen]~~)		.5				W	L		W		L	

Table 9: ERC description of BH's Finnish genitive plurals test case.

(/x/, [y], ~~[z]~~)	p(x)	p(y|x)	WTS	*Í	*L.L	*Ă	*H.H	*Ŏ	*LAPSE	*Ó	*Á	*Ĭ
(/kala/, [ká.lo.jen], ~~[ká.loi.den]~~)	.1755		W		L		W					
(/luettelo/, [lú.et.te.lòi.den], ~~[lú.et.te.lo.jen]~~)	.0802				W		L	W	W	L		
(/kamera/, [ká.me.ròi.den], ~~[ká.me.ro.jen]~~)	.1925				W	W	L		W		L	
(/avantgardisti/, [á.vant.gàr.dis.ti.en], ~~[á.vant.gàr.dis.tèi.den]~~)	.1440			W			W		L			L
(/korjaamo/, [kór.jaa.mo.jen], ~~[kór.jaa.mòi.den]~~)	.0765	0.82					W	L	L	W		
(/korjaamo/, [kór.jaa.mòi.den], ~~[kór.jaa.mo.jen]~~)		.18					L	W	W	L		
(/naapuri/, [náa.pu.rèi.den], ~~[náa.pu.ri.en]~~)	.1532	.29		L	W		L		W			W
(/naapuri/, [náa.pu.ri.en], ~~[náa.pu.rèi.den]~~)		.71		W	L		W		L			L
(/maailma/, [máa.il.mo.jen], ~~[máa.il.mòi.den]~~)	.0159	.5				L	W		L		W	
(/maailma/, [máa.il.mòi.den], ~~[máa.il.mo.jen]~~)		.5				W	L		W		L	

Table 10: ERC description of BH's Finnish genitive plurals test case after the simplifications of the candidate set described in subsection 4.3, 4.4, and 4.5.

(/x/, [y], ~~[z]~~)	p(x)	p(y|x)	WTS	*L.L	*H.H	*Í	*Ŏ	*Ă
(/kala/, [ká.lo.jen], ~~[ká.loi.den]~~)	.25		W	L	W			
(/luettelo/, [lú.et.te.lòi.den], ~~[lú.et.te.lo.jen]~~)	.12			W	L		W	
(/kamera/, [ká.me.ròi.den], ~~[ká.me.ro.jen]~~)	.28			W	L			W
(/korjaamo/, [kór.jaa.mo.jen], ~~[kór.jaa.mòi.den]~~)	.11	.82			W		L	
(/korjaamo/, [kór.jaa.mòi.den], ~~[kór.jaa.mo.jen]~~)		.18			L		W	
(/maailma/, [máa.il.mo.jen], ~~[máa.il.mòi.den]~~)	.02	.5			W			L
(/maailma/, [máa.il.mòi.den], ~~[máa.il.mo.jen]~~)		.5			L			W
(/naapuri/, [náa.pu.rèi.den], ~~[náa.pu.ri.en]~~)	.22	.29		W	L	L		
(/naapuri/, [náa.pu.ri.en], ~~[náa.pu.rèi.den]~~)		.71		L	W	W		

Table 11: ERC description of BH's Finnish genitive test case after the simplifications of the constraint set described in subsection 4.6.

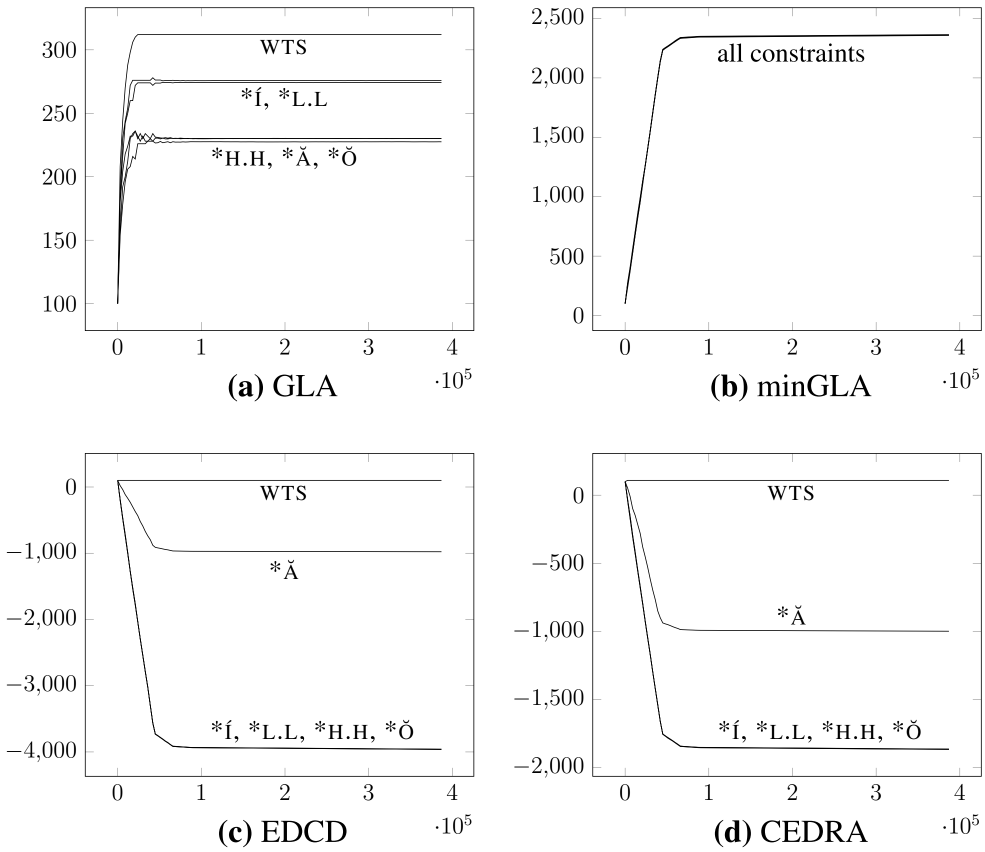


Table 12: Ranking dynamics predicted by **(a)** the GLA, **(b)** the minGLA, **(c)** EDCD, and **(d)** the CEDRA trained on the simplified Finnish genitive plurals test case described in table 11.

GLA		minGLA		EDCD		CEDRA	
WTS	314.0	WTS	2,364.474	WTS	100.0	WTS	108.667
*Í	277.734	*Ă	2,361.834	*Ă	-977.468	*Ă	-998.819
*L.L	276.266	*Í	2,361.512	*Í	-3958.608	*Í	-1863.969
*Ă	231.522	*L.L	2,360.848	*L.L	-3960.004	*L.L	-1864.652
*H.H	231.458	*H.H	2,359.406	*H.H	-3962.116	*H.H	-1866.018
*Ŏ	228.754	*Ŏ	2,357.548	*Ŏ	-3962.286	*Ŏ	-1866.616

Table 13: Ranking values learned by the stochastic EDRA with the four re-ranking rules listed in table 1 on the simplified Finnish Genitive test case described in table 11.

	actual	GLA	minGLA	EDCD	CEDRA
(/kala/, [ká.lo.jen])	1	1.0	0.906	1.0	1.0
(/kala/, [ká.loi.den])	0	0.0	0.094	0.0	0.0
(/luettelo/, [lú.et.te.lòi.den])	1	1.0	**0.722**	**0.821**	**0.746**
(/luettelo/, [lú.et.te.lo.jen])	0	0.0	**0.278**	**0.179**	**0.254**
(/kamera/, [ká.me.ròi.den])	1	1.0	**0.884**	1.0	1.0
(/kamera/, [ká.me.ro.jen])	0	0.0	**0.116**	0.0	0.0
(/korjaamo/, [kór.jaa.mòi.den])	0.18	0.178	0.255	**0.474**	**0.417**
(/korjaamo/, [kór.jaa.mo.jen])	0.82	0.822	0.745	**0.526**	**0.583**
(/naapuri/, [náa.pu.rèi.den])	0.29	0.278	0.354	0.287	0.345
(/naapuri/, [náa.pu.ri.en])	0.71	0.722	0.646	0.713	0.655
(/maailma/, [máa.il.mòi.den])	0.51	0.503	**0.803**	**1.0**	**1.0**
(/maailma/, [máa.il.mo.jen])	0.5	0.497	**0.197**	**0.0**	**0.0**

Table 14: Probabilities of each underlying/surface form mapping predicted by the four final ranking vectors reported in table 13.

Calibration of constraint promotion does not help with learning variation in stochastic OT:

online supplementary materials

Giorgio Magri and Benjamin Storme

September 18, 2018

The following appendices provide precise formulations and proofs of facts 1, 2, and 3 informally stated in section 3 of the paper, concerning the behavior of the stochastic EDRA on the Ilokano metathesis test case. This test case has been described in ERC notation in table 4, repeated below as table 15, properly reordered. We have numbered the eleven constraints as $C_1, C_2, \ldots, C_{11}$. We have also reordered the twenty-two ERCs and numbered them as ERC_1, ERC_2, ..., ERC_{22}, in such a way that the ERCs 1 -15 in the top block are those with a W corresponding to one of the constraints $C_1, \ldots, C_6$ and the three ERCs 16-18 in the second block are those with a W corresponding to constraints C_7.

Throughout this appendix, we assume for simplicity that plasticity (see footnote 2 in the paper) is a constant η through the run (while BH assume plasticity to decrease over learning stages). Furthermore, we assume the ranking values are all initially set equal to zero (while BH assume they are initially set equal to 100). Finally, we assume that the stochastic values $\epsilon_1, \epsilon_2, \ldots$ are sampled between $-\Delta$ and $+\Delta$ (both edges excluded). The statements and the proofs extend *with high probability* to the case where the stochastic values $\epsilon_1, \epsilon_2, \ldots$ are sampled according to a gaussian distribution with zero mean, as in that case they are bounded between $-\Delta$ and $+\Delta$ with high probability. The ranking value of a constraint C_h entertained by the learner at the tth iteration of an arbitrary run is denoted by θ_h^t.

A Facts 1 and 2 for EDCD

A.1 The following lemma lower-bounds the ranking values of constraints $C_1, \ldots, C_6, C_7$ in an arbitrary run of the stochastic EDRA with any of the re-ranking rules listed in table 1.

Lemma 1 *Consider a run of the stochastic EDRA with any of the four re-ranking rules listed in table 1 on the Ilokano metathesis test case described by the ERC matrix in table 15. The ranking values $\theta_1^t, \ldots, \theta_6^t, \theta_7^t$ of the constraints $C_1, \ldots, C_6, C_7$ satisfy*

(7) a. $\theta_1^t \geq 0, \ldots, \theta_6^t \geq 0$

b. $\theta_7^t > -2\Delta - \eta$

at any time t.

Proof. The inequalities (7a) follow straightforwardly from the fact that $C_1, \ldots, C_6$ start out with an initial ranking value equal to 0 and are never demoted because they are never loser-preferring in table 15. The inequality (7b) trivially holds at the initial time $t = 0$, because $\theta_7^{t=0} = 0$ (all ranking values are initialized to zero). It thus suffices to prove that, if the inequality (7b) holds at time t, it also holds at the subsequent time $t + 1$. That is trivially the case if the update between times t and $t + 1$ does not demote C_7 (because in that case the ranking value of C_7 does not decrease between times t and $t + 1$). Thus, assume that C_7 is demoted by the EDRA

	Onset $= C_1$	MaxIO(V) $= C_2$	*LowGlide $= C_3$	IdentIO[low] $= C_4$	*[ʔC $= C_5$	MaxOO(ʔ) $= C_6$	DepIO(ʔ) $= C_7$	IdentIO[syl] $= C_8$	Linearity $= C_9$	*ʔ] $= C_{10}$	MaxIO(ʔ) $= C_{11}$
$\mathbf{ERC_1}$ = (/paʔlak/, [pa.lak], ~~[pa.ʔlak]~~)					W						L
$\mathbf{ERC_2}$ = (/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jen]~~)		W						L			
$\mathbf{ERC_3}$ = (/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jo.en]~~)	W							L			
$\mathbf{ERC_4}$ = (/basa-en/, [ba.sa.ʔen], ~~[bas.a̯en]~~)			W				L	W			
$\mathbf{ERC_5}$ = (/basa-en/, [ba.sa.ʔen], ~~[bas.wen]~~)				W			L	W			
$\mathbf{ERC_6}$ = (/basa-en/, [ba.sa.ʔen], ~~[ba.sen]~~)		W					L				
$\mathbf{ERC_7}$ = (/basa-en/, [ba.sa.ʔen], ~~[ba.sa.en]~~)	W						L				
$\mathbf{ERC_8}$ = (/taʔo-en/, [taw.ʔen], ~~[ta.wen]~~)						W			L		W
$\mathbf{ERC_9}$ = (/taʔo-en/, [taw.ʔen], ~~[ta.ʔen]~~)		W						L	L		
$\mathbf{ERC_{10}}$ = (/taʔo-en/, [taw.ʔen], ~~[ta.ʔo.en]~~)	W							L	L		
$\mathbf{ERC_{11}}$ = (/taʔo-en/, [taw.ʔen], ~~[ta.ʔwen]~~)					W				L		
$\mathbf{ERC_{12}}$ = (/taʔo-en/, [taʔ.wen], ~~[ta.wen]~~)						W				L	W
$\mathbf{ERC_{13}}$ = (/taʔo-en/, [taʔ.wen], ~~[ta.ʔen]~~)		W						L		L	
$\mathbf{ERC_{14}}$ = (/taʔo-en/, [taʔ.wen], ~~[ta.ʔo.en]~~)	W							L		L	
$\mathbf{ERC_{15}}$ = (/taʔo-en/, [taʔ.wen], ~~[ta.ʔwen]~~)					W					L	
$\mathbf{ERC_{16}}$ = (/ʔajo-en/, [ʔaj.wen], ~~[ʔa.jo.ʔen]~~)							W	L			
$\mathbf{ERC_{17}}$ = (/taʔo-en/, [taw.ʔen], ~~[ta.ʔo.ʔen]~~)							W	L	L		
$\mathbf{ERC_{18}}$ = (/taʔo-en/, [taʔ.wen], ~~[ta.ʔo.ʔen]~~)							W	L		L	
$\mathbf{ERC_{19}}$ = (/paʔlak/, [pa.lak], ~~[pal.ʔak]~~)									W		L
$\mathbf{ERC_{20}}$ = (/paʔlak/, [pa.lak], ~~[paʔ.lak]~~)										W	L
$\mathbf{ERC_{21}}$ = (/taʔo-en/, [taw.ʔen], ~~[taʔ.wen]~~)									L	W	
$\mathbf{ERC_{22}}$ = (/taʔo-en/, [taʔ.wen], ~~[taw.ʔen]~~)									W	L	

Table 15: ERC description of BH's Ilokano metathesis test case (repeated from table 4 in the paper, slightly reordered).

between times t and $t+1$. Every ERC in table 15 which has an L corresponding to C_7, also has a W corresponding to one of $C_1, \ldots, C_6$. Let C_k be the constraint among $C_1, \ldots, C_6$ which is winner preferring relative to the ERC which triggers that update between times t and $t+1$. Since C_7 is demoted by that update, it must be undominated relative to the stochastic ranking vector $\boldsymbol{\theta}^t + \boldsymbol{\epsilon}$ entertained at time t. Since C_k is winner-preferring, this means in particular that the stochastic ranking value of C_7 is not smaller than that of C_k at time t:

(8) $\theta^t_7 + \epsilon_7 \geq \theta^t_k + \epsilon_k$

Since $\theta^t_k \geq 0$ by (7a) and $-\Delta < \epsilon_h, \epsilon_k < +\Delta$, the inequality (8) entails that $\theta^t_7 > -2\Delta$. Since C_7 is demoted by η between times t and $t+1$, it follows that $\theta^{t+1}_7 > -2\Delta - \eta$, thus establishing the inequality (7b) at time $t+1$. □

A.2 The following lemma focuses on EDCD and connects the number N^t of updates triggered by an ERC up to time t and the ranking value θ^t_h of any constraints C_h which is loser-preferring relative to that ERC. The inequality (9) is obvious for an ERC with a unique loser-preferring constraint, because that loser-preferrer must be demoted by η every time that ERC triggers an update. For ERCs with multiple loser-preferring constraints, some more care is needed to establish the inequality (9), because not all loser-preferring constraints are necessarily demoted with each update (only the loser-preferrers which are currently undominated are indeed demoted).

Lemma 2 *Consider a run of the stochastic EDCD on an arbitrary ERC matrix (not necessarily consistent with any ranking). For any ERC in that matrix, let N^t be the number of updates it has triggered in that run up to time t. The following inequality*

(9) $\theta_h^t \leq -\eta\, N^t$

holds at any time t and for any constraint C_h which has an L *in the ERC considered.*

Proof. The inequality (9) trivially holds at the initial time $t = 0$, because $\theta_h^{t=0} = 0$ (all ranking values are initialized to zero) and $N^{t=0} = 0$ (no ERC has triggered any update yet). It thus suffices to prove that, if the inequality (9) holds at time t, it also holds at the subsequent time $t+1$. That is trivially the case if the update between times t and $t+1$ has not been triggered by the ERC considered, because the left hand side of (9) does not increase between times t and $t+1$ (namely, $\theta_h^{t+1} \leq \theta_h^t$ due to the lack of constraint promotion) and the right hand side does not change (namely $N^{t+1} = N^t$, because it is not the ERC considered which triggers the update between times t and $t+1$).

Suppose next that the update between times t and $t+1$ has been triggered by the ERC considered. By adding $-\eta$ at both sides of the inequality (9) at time t and using the obvious identity $N^{t+1} = N^t + 1$, we obtain:

(10) $\theta_h^t - \eta \leq -\eta\, N^{t+1}$

We now distinguish two cases. To start, consider the case where C_h is a loser-preferrer which is demoted between times t and $t+1$. Thus $\theta_h^{t+1} = \theta_h^t - \eta$. Hence, (10) becomes $\theta_h^{t+1} \leq -\eta\, N^{t+1}$, thus establishing the inequality (9) at time $t+1$. Next, consider the case where C_h is a loser-preferrer which is not demoted between times t and $t+1$. Consider another constraint $C_{\widehat{h}}$ which is loser preferring relative to the ERC considered and is indeed demoted between times t and $t+1$ (such a constraint $C_{\widehat{h}}$ exists because the ERC considered has indeed triggered an update between times t and $t+1$). The chain of inequalities in (11) then establishes the inequality (9) at time $t+1$.

(11) $$\theta_h^{t+1} \overset{(a)}{=} \theta_h^t \overset{(b)}{\leq} \theta_{\widehat{h}}^t - \eta \overset{(c)}{=} \theta_{\widehat{h}}^{t+1} \overset{(d)}{\leq} \eta\, N^{t+1}$$

In step (11a), we have used the fact that C_h is not demoted between times t and $t+1$, whereby $\theta_h^{t+1} = \theta_h^t$. Since $C_{\widehat{h}}$ is instead demoted between times t and $t+1$, it ought to be $\theta_h^t < \theta_{\widehat{h}}^t$. Since all ranking values start out at zero and are demoted by η, they are spaced out by multiples of η. Thus, $\theta_h^t < \theta_{\widehat{h}}^t$ entails $\theta_h^t \leq \theta_{\widehat{h}}^t - \eta$, yielding step (11b). In step (11c), we have used the fact that $C_{\widehat{h}}$ is demoted between times t and $t+1$, whereby $\theta_{\widehat{h}}^{t+1} = \theta_{\widehat{h}}^t - \eta$. By the first case considered, the inequality (9) holds at time $t+1$ for the demoted constraint $C_{\widehat{h}}$, yielding step (11d). □

A.3 We can now state and prove the following explicit formulations of facts 1 and 2 for EDCD, informally anticipated in subsections 3.4 and 3.5 of the paper. The proof straightforwardly formalizes the intuitive reasoning depicted in (4) in the paper.

Fact 1 (case of EDCD). *Consider a run of stochastic EDCD on the Ilokano metathesis test case described by the ERC matrix in table 15. The number of updates triggered by any of the ERCs 1 through 15 (in the top block of the matrix) cannot be larger than $\frac{2\Delta}{\eta}$.*

Proof. Suppose by contradiction that one of the ERCs 1 through 15 has triggered $\frac{2\Delta}{\eta}$ updates up to time t and yet triggers one more update between times t and $t+1$. Let C_k be the constraint among $C_1, \ldots, C_6$ which is winner-preferring relative to that ERC. There has got to

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
ERC_8						W			L		W
ERC_{12}						W				L	W
ERC_{19}									W		L
ERC_{20}										W	L
ERC_{21}									L	W	
ERC_{22}									W	L	

Table 16: Subset of ERCs in table 15 which have a W corresponding to one the three constraints C_9, C_{10}, or C_{11}.

exist a constraint C_h which is loser-preferring relative to the ERC considered and furthermore undominated relative to the stochastic ranking vector $\boldsymbol{\theta}^t + \boldsymbol{\epsilon}$ entertained at time t. Since C_k is winner-preferring, the stochastic ranking value of C_h cannot be smaller than the stochastic ranking value of C_k at time t:

(12) $\theta_h^t + \epsilon_h \geq \theta_k^t + \epsilon_k$

Since $-\Delta < \epsilon_h, \epsilon_k < +\Delta$, since $\theta_k^t \geq 0$ (by (7a) of lemma 1), and since $\theta_h^t \leq -2\Delta$ (by lemma 2 together with the hypothesis that the ERC considered has already triggered $\frac{2\Delta}{\eta}$ updates up to time t), the inequality (12) is a contradiction. □

Fact 2 (case of EDCD). *Consider a run of stochastic EDCD on the Ilokano metathesis test case described by the ERC matrix in table 15. The number of updates triggered by any of the ERCs 16 through 18 (in the second block of the matrix) cannot be larger than* $\frac{4\Delta}{\eta} + 1$.

Proof. The proof is completely analogous to the proceeding proof of fact 1, only using (7b) instead of (7a). Suppose by contradiction that one of the ERCs 16 through 18 has triggered $\frac{4\Delta}{\eta} + 1$ updates up to time t and yet triggers one more update between times t and $t+1$. There has got to exist a constraint C_h which is loser-preferring relative to that ERC and furthermore undominated relative to the stochastic ranking vector $\boldsymbol{\theta}^t + \boldsymbol{\epsilon}$ entertained at time t. Since C_7 is winner-preferring relative to the three ERCs 16 through 18, the stochastic ranking value of C_h cannot be smaller than the stochastic ranking value of C_7 at time t:

(13) $\theta_h^t + \epsilon_h \geq \theta_7^t + \epsilon_7$

Since $-\Delta < \epsilon_h, \epsilon_k < +\Delta$, since $\theta_7^t \geq -2\Delta - \eta$ (by (7b) of lemma 1), and since $\theta_h^t \leq -4\Delta - \eta$ (by lemma 2 together with the hypothesis that the ERC considered has already triggered $\frac{4\Delta}{\eta}+1$ updates up to time t), the inequality (13) is a contradiction. □

B Facts 1 and 2 for the GLA, the minGLA, and the CEDRA

B.1 The ERCs in table 15 which have a W corresponding to one the three constraints C_9, C_{10}, or C_{11} and can thus contribute to their promotion are singled out in table 16. These ERCs share the crucial property that they promote one of the three constraints C_9, C_{10}, or C_{11}, they demote another one of those three constraints and they do not re-rank the remaining one of the three constraints. The following lemma uses this fact to deduce an upper bound on the ranking values of the three constraints C_9, C_{10}, or C_{11}.

Lemma 3 *Consider a run of the stochastic EDRA with any of the four re-ranking rules described in table 1 on the Ilokano metathesis test case summarized by the ERC matrix in table 15. The ranking values* θ_9^t, θ_{10}^t, *and* θ_{11}^t *of the three constraints* C_9, C_{10}, *and* C_{11} *satisfy*

(14) $\theta_9^t, \theta_{10}^t, \theta_{11}^t \leq 2\Delta + 2\eta$

at any time t.

Proof. In order to carry out the proof by induction on time t, we strengthen the claim (14) as in (15). Not only each of the three ranking values individually is never larger than $2\Delta + 2\eta$, as stated in (14) and repeated in (15a). But also the sum of any two of those ranking values is never larger than $2\Delta + 2\eta$, as stated in (15b). And the sum of the three ranking values is never larger than zero, as stated in (15c).

(15) a. $\theta_9^t,\ \theta_{10}^t,\ \theta_{11}^t \leq 2\Delta + 2\eta$

b. $\theta_9^t + \theta_{10}^t,\ \theta_9^t + \theta_{11}^t,\ \theta_{10}^t + \theta_{11}^t \leq 2\Delta + 2\eta$

c. $\theta_9^t + \theta_{10}^t + \theta_{11}^t \leq 0$

The inequalities (15) trivially hold at the initial time $t = 0$, as the ranking values are all initialized to zero. It thus suffices to prove that, if the inequalities hold at time t, they also hold at the subsequent time $t + 1$.

Inequalities (15a) For concreteness, we focus on C_9 (analogous considerations hold for C_{10} and C_{11}). If C_9 is not promoted between times t and $t + 1$, the inductive hypothesis that the corresponding inequality (15a) holds at time t immediately entails that it also holds at time $t + 1$. Thus, we assume that C_9 is promoted between times t and $t + 1$, as stated in (16a). This means in turn that the ERC which triggers that update is one of those singled out in table 16. Since these ERCs have no L's outside of C_9, C_{10} or C_{11}, the loser-preferring constraint which triggers the update and undergoes demotion is either C_{10} or C_{11}. For concreteness, we assume it is C_{10}, as stated in (16b).

(16) a. C_9 is promoted between times t and $t + 1$;

b. C_{10} is undominated relative to the stochastic ranking vector at time t.

We assume by contradiction that the ranking value of C_9 after the update described in (16) fails to satisfy the corresponding inequality (15a), as stated in (17).

(17) $\theta_9^{t+1} > 2\Delta + 2\eta$

Since the ranking value of C_9 is large at time $t+1$ by (17), it is also large at time t, as shown in (18). Step (18a) uses the fact (16a) that C_9 is promoted between times t and $t+1$, so that its ranking value at time $t+1$ coincides with its ranking value at time t increased by the promotion amount p times the plasticity η. Step (18b) uses the fact that the promotion amount p is not larger than 1 according to any of the re-ranking rules considered in table 1. Step (18c) uses the contradictory assumption (17).

(18) $\theta_9^t \overset{(a)}{=} \theta_9^{t+1} - p\eta \overset{(b)}{\geq} \theta_9^{t+1} - \eta \overset{(c)}{>} 2\Delta + \eta$

Since the ranking value of the winner-preferrer C_9 at time t is large by (18), the ranking value of the loser-preferrer C_{10} at time t is also large, as shown in (19). Step (19a) uses the fact (16b) that C_{10} is undominated relative to the stochastic ranking vector at time t. That means that C_{10} cannot be ranked underneath the winner-preferrer C_9 by more than 2Δ, because their separation would otherwise be too large for the additive stochastic values $\epsilon_1, \epsilon_2, \ldots$ to swap the two constraints. Step (19b) uses the inequality $\theta_9^t > 2\Delta + \eta$ obtained in (18).

(19) $\theta_{10}^t \overset{(a)}{\geq} \theta_9^t - 2\Delta \overset{(b)}{>} \eta$

From (18) and (19), we conclude that $\theta_9^t + \theta_{10}^t > 2\Delta + 2\eta$, contradicting the inductive hypothesis that (15b) holds at time t.

Inequalities (15b) For concreteness, we focus on the sum of the ranking values of C_9 and C_{10} (analogous considerations hold for the other two pairs of constraints). If neither C_9 nor C_{10} is promoted in between times t and $t+1$, the sum of their ranking values does not increase. And the inductive hypothesis that the corresponding inequality (15b) holds at time t immediately entails that it also holds at time $t+1$. We thus assume that either C_9 or C_{10} is promoted; for concreteness, we assume it is C_9, as stated in (20a). This means in turn that the update between times t and $t+1$ is triggered by one of the ERCs singled out in table 16. If C_{10} is demoted in that update, the promotion of C_9 and the demotion of C_{10} cancel out and again the sum of their ranking values cannot have increased between times t and $t+1$. Thus, we assume that C_{10} is not demoted. Since no ERC in table 16 has two W's in the last three columns, C_{10} is not promoted either, namely it is not re-ranked, as stated in (20b). Finally, since the ERCs in table 16 have no L's outside of C_9, C_{10}, and C_{11}, the loser-preferring constraint which triggers the update and undergoes demotion must be C_{11}, as stated in (20c).

(20) a. C_9 is promoted between times t and $t+1$;
b. C_{10} is not re-ranked between times t and $t+1$;
c. C_{11} is undominated relative to the stochastic ranking vector used at time t.

We assume by contradiction that the sum of the ranking values of C_9 and C_{10} after the update described in (20) fails to satisfy the corresponding inequality (15b), as stated in (21).

(21) $\theta_9^{t+1} + \theta_{10}^{t+1} > 2\Delta + 2\eta$

Since the sum of the ranking values of C_9 and C_{10} at time $t+1$ is large by (21) and since the ranking value of C_{10} at time t is small by (15a), the ranking value of C_9 at time t must be large, as shown in (22). Step (22a) uses the fact (20a) that C_9 is promoted between times t and $t+1$, so that its ranking value θ_9^{t+1} at time $t+1$ coincides with its ranking value θ_9^t at the preceding time t increased by at most η. Step (22b) uses the contradictory assumption (21) to lower bound the ranking value θ_9^{t+1}. Step (22c) uses the fact (20b) that C_{10} is not re-ranked between times t and $t+1$, so that its ranking values θ_{10}^t and θ_{10}^{t+1} at times t and $t+1$ coincide. Step (22d) uses the inductive hypothesis that the inequalities (15a) hold at time t to upper bound θ_{10}^t.

(22) $$\theta_9^t \overset{(a)}{\geq} \theta_9^{t+1} - \eta \overset{(b)}{>} (2\Delta + 2\eta - \theta_{10}^{t+1}) - \eta \overset{(c)}{=} 2\Delta + \eta - \theta_{10}^t \overset{(d)}{\geq} 2\Delta + \eta - (2\Delta + 2\eta) = -\eta$$

Since the ranking value of the winner-preferring C_9 at time t is large by (22), the ranking value of the loser-preferring constraint C_{11} at time t must be large as well, as shown in (23). Step (23a) uses the fact (20c) that C_{11} is undominated relative to the stochastic ranking vector at time t. That means that C_{11} cannot be ranked underneath the winner-preferring C_9 by more than 2Δ. Step (23b) uses the inequality $\theta_9^t > -\eta$ obtained in (22).

(23) $$\theta_{11}^t \overset{(a)}{\geq} \theta_9^t - 2\Delta \overset{(b)}{>} -\eta - 2\Delta$$

The sum of the ranking values at time t of C_9, C_{10}, and C_{11} can now be bounded as in (24). Step (24a) uses the hypothesis (20a) that C_9 is promoted by at most η between times t and $t+1$. Step (24b) uses the hypothesis (20b) that C_{10} is not re-ranked between times t and $t+1$, so that $\theta_{10}^t = \theta_{10}^{t+1}$. Step (24c) uses the contradictory assumption (21). Step (24d) uses the lower bound (23) on θ_{11}^t.

(24) $$\theta_9^t + \theta_{10}^t + \theta_{11}^t \overset{(a)}{\geq} \theta_9^{t+1} - \eta + \theta_{10}^t + \theta_{11}^t \overset{(b)}{=} \theta_9^{t+1} - \eta + \theta_{10}^{t+1} + \theta_{11}^t \overset{(c)}{>} 2\Delta + \eta + \theta_{11}^t \overset{(d)}{\geq} 0$$

The chain of inequalities in (24) says that the sum of the ranking values of C_9, C_{10}, and C_{11} at time t is larger than 0, contradicting the inductive hypothesis that (15c) holds at time t.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
ERC_4			W				L	W			
ERC_5				W			L	W			
ERC_{16}							W	L			
ERC_{17}							W	L	L		
ERC_{18}							W	L		L	

Table 17: Subset of ERCs in table 15 which have a W corresponding to constraints C_7 or C_8.

Inequality (15c) It suffices to show that the sum of the three ranking values of C_9, C_{10}, and C_{11} does not increase between times t and $t+1$. If the ERC which triggers the update between times t and $t+1$ is not one of those in table 16 which have a W corresponding to C_9, C_{10}, or C_{11}, the sum of their three ranking values does not increase because none of them is promoted. Thus, we assume that the update between times t and $t+1$ is triggered by one of the ERCs in table 16. Since any of these ERCs has exactly one W and exactly one L corresponding to C_9, C_{10}, or C_{11}, the sum of their ranking values is decreased by η and increased by at most η, and thus does not overall increase. □

B.2 The ERCs in table 15 which have a W corresponding to constraints C_7 or C_8 and can thus contribute to their promotion are singled out in table 17. These ERCs share the crucial property that they promote one of the two constraints C_7 or C_8 and demote the other. The following lemma uses this fact to deduce an upper bound on the ranking values of the two constraints. The proof is analogous to the proof of the preceding lemma and therefore only briefly sketched.

Lemma 4 *Consider a run of the stochastic EDRA with any of the four re-ranking rules described in table 1 on the Ilokano metathesis test case summarized by the ERC matrix in table 15. The ranking values θ_7^t and θ_8^t of the two constraints C_7 and C_8 satisfy*

(25) $\theta_7^t, \theta_8^t \leq 4\Delta + 3\eta$

at any time t.

Proof. By reasoning as in the proof of the preceding lemma, we immediately conclude that the sum of the two ranking values θ_7^t and θ_8^t never becomes larger than zero, as stated in (26).

(26) $\theta_7^t + \theta_8^t \leq 0$

The inequalities (25) trivially hold at the initial time $t = 0$, as the ranking values are all initialized to zero. It thus suffices to prove that, if the inequalities (25) hold at time t, they also hold at the subsequent time $t+1$. For concreteness, we focus on C_7 (analogous considerations hold for C_8). If C_7 is not promoted between times t and $t+1$, the claim is trivial. Thus, we assume that C_7 is promoted between times t and $t+1$, as stated in (27a). This means that the ERC which triggers that update is one of those singled out in table 17. One of C_8, C_9 or C_{10} must therefore be the loser-preferrer which triggers the update and undergoes demotion, as stated in (16b).

(27) a. C_7 is promoted between times t and $t+1$;

b. C_8, C_9 or C_{10} is undominated relative to the stochastic ranking vector at time t.

We assume by contradiction that the ranking value of C_7 after the update described in (27) fails to satisfy the corresponding inequality (25), as stated in (28).

(28) $\theta_7^{t+1} > 4\Delta + 3\eta$

By reasoning as above in (18), this contradictory assumption entails that:

(29) $\theta_7^t > 4\Delta + 2\eta$

Let C_h be the constraint among C_8, C_9 or C_{10} which is undominated at time t by (27b). By reasoning as above in (19), (29) entails that:

(30) $\theta_h^t \geq \theta_7^t - 2\Delta > 4\Delta + 2\eta - 2\Delta = 2\Delta + 2\eta$

By (14), the inequality (30) ensures that C_h cannot be neither C_9 nor C_{10}. Hence, it must be C_8. In this case (29) and (30) yield $\theta_7^t + \theta_8^t > 6\Delta + 4\eta$, which contradicts (26). □

B.3 Constraints $C_1, \dots, C_6$ in table 15 are never loser-preferring. The following lemma uses this fact to deduce an upper bound on their ranking values. The proof is analogous to the proof of the two preceding lemmas and therefore only briefly sketched.

Lemma 5 *Consider a run of the stochastic EDRA with any of the four re-ranking rules described in table 1 on the Ilokano metathesis test case summarized by the ERC matrix in table 15. The ranking values $\theta_1^t, \dots, \theta_6^t$ of the constraints $C_1, \dots, C_6$ satisfy*

(31) $\theta_1^t, \dots, \theta_6^t \leq 6\Delta + 4\eta$

at any time t.

Proof. Since the inequalities (31) trivially hold at the initial time $t = 0$, it thus suffices to prove that, if they hold at time t, they also hold at the subsequent time $t + 1$. For concreteness, we focus on C_1 (analogous considerations hold for $C_2, \dots, C_6$). If C_1 is not promoted between times t and $t + 1$, the claim is trivial. Thus, we assume that C_1 is promoted between times t and $t+1$, as stated in (32a). Since $C_1, \dots, C_6$ are never loser-preferring, one of the other constraints $C_7, \dots, C_{11}$ must be the loser-preferrer which triggers the update and undergoes demotion, as stated in (32b).

(32) a. C_1 is promoted between times t and $t + 1$;

b. one among $C_7, \dots, C_{11}$ is undominated relative to the stochastic ranking vector at time t.

We assume by contradiction that the ranking value of C_1 after the update described in (32) fails to satisfy the corresponding inequality (31), as stated in (33).

(33) $\theta_1^{t+1} > 6\Delta + 4\eta$

By reasoning as above in (18), this contradictory assumption entails that:

(34) $\theta_1^t > 6\Delta + 3\eta$

Let C_h be the constraint which is undominated at time t. By reasoning as above in (19), (34) entails that:

(35) $\theta_h^t \geq \theta_1^t - 2\Delta > 4\Delta + 3\eta$

By (14), the inequality (35) ensures that C_h cannot be C_9 or C_{10}, or C_{11}. By (25), it ensures that C_h cannot be C_7 or C_8. This conclusion contradicts (32b). □

B.4 Using lemma 5, we can now state and prove the following explicit formulations of facts 1 and 2 for the GLA, the minGLA, and the CEDRA, informally anticipated in subsections 3.4 and 3.5.

Fact 1 (case of the GLA, minGLA, and CEDRA). *Consider a run of the stochastic GLA, minGLA, or CEDRA on the Ilokano metathesis test case summarized by the ERC matrix in table 15. The number of updates triggered by any of the ERCs 1 through 15 (in the top block of the matrix) cannot be larger than $\frac{6\Delta}{\eta} + 4$ in the case of the GLA and the minGLA and not larger than $3\left(\frac{6\Delta}{\eta} + 4\right)$ in the case of the CEDRA.*

Proof. Consider the case of the GLA or the minGLA. Suppose by contradiction that there exists an ERC 1 through 15 which has triggered more than $\frac{6\Delta}{\eta} + 4$ updates. Let C_k be the constraint among $C_1, \ldots, C_6$ which is winner-preferring relative to that ERC. Since C_k starts out with a null initial ranking value, since it is never demoted (because never loser-preferring), and since it is promoted by η with each update by the ERC considered, its ranking value must have grown larger than $6\Delta + 4\eta$. This conclusion contradicts lemma 5.

Analogous considerations hold for the case of the CEDRA. Suppose by contradiction that there exists an ERC 1 through 15 which has triggered more than $3\left(\frac{6\Delta}{\eta} + 4\right)$ updates. Let w be the number of winner-preferring constraints in that ERC. Hence, $w \leq 2$, because no ERC 1 through 15 has more than two winner-preferring constraints. Let C_k be the constraint among $C_1, \ldots, C_6$ which is winner preferring relative to the ERC considered. Since C_k starts out with a null initial ranking value, since it is never demoted (because never loser-preferring), and since it is promoted by $\frac{\eta}{w+1} \geq \frac{\eta}{3}$ with each update by the ERC considered, its ranking value must have grown larger than $6\Delta + 4\eta$. This conclusion again contradicts lemma 5. □

Fact 2 (case of the GLA, minGLA, and CEDRA). *Consider a run of the stochastic GLA, minGLA, or CEDRA on the Ilokano metathesis test case summarized by the ERC matrix in table 15. The number of updates triggered by any of the ERCs 16 through 18 (in the second block of the matrix) cannot be larger than* $\frac{28\Delta}{\eta} + 19$ *in the case of the GLA and the minGLA and not larger than* $2\left(\frac{28\Delta}{\eta} + 19\right)$ *in the case of the CEDRA.*

Proof. Consider the case of the GLA or the minGLA. Suppose by contradiction that there exists an ERC 16 through 18 which has triggered more than $\frac{28\Delta}{\eta} + 19$ updates by some time t. Each of these updates promotes C_7 by η, yielding the term (36a) in the lower bound on the ranking value θ_7^t of C_7. On the other hand, only the four ERCs 4 through 7 can demote C_7 by η. And each of those ERCs can trigger at most $\frac{6\Delta}{\eta} + 4$ updates by the preceding fact 1, yielding the term (36b).

$$(36)\quad \theta_7^t > \underbrace{\left(\frac{28\Delta}{\eta} + 19\right)\eta}_{(a)} - \underbrace{4\left(\frac{6\Delta}{\eta} + 4\right)\eta}_{(b)}$$

The inequality (36) thus obtained contradicts the upper bound $\theta_7^t \leq 4\Delta + 3\eta$ on the ranking value of C_7 provided by lemma 4.

Analogous considerations hold for the case of the CEDRA. The only difference is that each of the three ERCs 16 through 18 promotes C_7 not by η but by $\eta/2$. In the case of the CEDRA, the inequality (36) is thus replaced by the following:

$$(37)\quad \theta_7^t > 2\left(\frac{28\Delta}{\eta} + 19\right)\frac{\eta}{2} - 4\left(\frac{6\Delta}{\eta} + 4\right)\eta$$

which again contradicts lemma 4. □

C Fact 3 for the GLA and the minGLA

C.1 The ERCs in table 15 which have an L corresponding to constraints C_9 or C_{10} and can thus contribute to their demotion are singled out in table 18.

Lemma 6 *Consider a run of the stochastic GLA or minGLA on the Ilokano metathesis test case summarized by the ERC matrix in table 15. The ranking values* θ_9^t *and* θ_{10}^t *of the constraints* C_9 *and* C_{10} *satisfy*

$$(38)\quad \theta_9^t, \theta_{10}^t \geq -53\Delta - 36\eta$$

at any time t.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
ERC_8						W			L		W
ERC_9		W						L	L		
ERC_{10}	W							L	L		
ERC_{11}					W				L		
ERC_{12}						W				L	W
ERC_{13}		W						L		L	
ERC_{14}	W							L		L	
ERC_{15}					W					L	
ERC_{17}							W	L	L		
ERC_{18}							W	L		L	
ERC_{21}									L	W	
ERC_{22}									W	L	

Table 18: Subset of ERCs in table 15 which have an L corresponding to constraints C_9 or C_{10}.

Proof. To start, we note that the sum of the two ranking values $\theta_9^t, \theta_{10}^t$ can be lower-bounded as in (39). In fact, each of the eight ERCs 8 through 15 in table 18 demotes either C_9 or C_{10} by η and can trigger at most $\frac{6\Delta}{\eta} + 4$ updates by fact 1, yielding the term (39a). Furthermore, both ERCs 17 and 18 in table 18 demote either C_9 or C_{10} by η and can trigger at most $\frac{28\Delta}{\eta} + 19$ updates by fact 2, yielding the term (39b). Finally, ERCs 21 and 22 in table 18 promote one of the two constraints C_9 or C_{10} by η and demote the other by η as well, thereby contributing nothing to the sum of the two ranking values.

(39) $$\theta_9^t + \theta_{10}^t \geq -\underbrace{8\left(\frac{6\Delta}{\eta} + 4\right)\eta}_{(a)} - \underbrace{2\left(\frac{28\Delta}{\eta} + 19\right)\eta}_{(b)} = -104\Delta - 70\eta$$

We can turn to the inequalities (38). These inequalities trivially hold at time $t = 0$. It thus suffices to prove that, if they hold at time t, they also hold at the subsequent time $t+1$. For concreteness, we focus on C_9 (analogous considerations hold for C_{10}). If C_9 is not demoted between times t and $t+1$, the claim is trivial. Thus assume by contradiction that C_9 is demoted between times t and $t+1$ and that its ranking value after that update fails to satisfy the corresponding inequality (38), namely $\theta_9^{t+1} < -53\Delta - 36\eta$. Since the ranking value of C_9 is decreased by η between times t and $t+1$ the latter means that:

(40) $$\theta_9^t < -53\Delta - 35\eta$$

Let C_k be a winner-preferring constraint relative to the ERC in table 18 which has triggered the update between times t and $t+1$. Since C_9 is undominated relative to the ranking vector at time t, the stochastic ranking value of C_9 must be larger than the stochastic ranking value of C_k at time t:

(41) $$\theta_9^t + \epsilon_9 \geq \theta_k^t + \epsilon_k$$

Since $-\Delta < \epsilon_9, \epsilon_k < +\Delta$, (40) and (41) entail that:

(42) $$\theta_k^t < -51\Delta - 35\eta$$

By lemma 1, condition (42) says that C_k cannot be one of the constraints $C_1, \ldots, C_6, C_7$. We thus conclude from table 18 that C_k must be C_{10}. The two inequalities (40) and (42) on the ranking values of C_9 and C_{10} then contradict the inequality (39) on their sum. □

C.2 We can now state and prove the following explicit formulation of fact 3, informally anticipated in subsection 3.6.

Fact 3. *Consider a run of the stochastic GLA or minGLA on the Ilokano metathesis test case described by the ERC matrix in table 15. The number of updates triggered by ERCs 19 and 20 (in the third block of the matrix) cannot be larger than* $\frac{67\Delta}{\eta} + 45$.

Proof. Let's focus on ERC 19 (analogous considerations hold for ERC 20). Suppose by contradiction that ERC 19 has triggered $\frac{67\Delta}{\eta} + 35$ updates up to time t and yet triggers one more update between times t and $t+1$. Since C_{11} is the only loser-preferrer relative to ERC 19, C_{11} must be undominated relative to the stochastic ranking vector $\boldsymbol{\theta}^t + \boldsymbol{\epsilon}$ entertained at time t. Since C_9 is winner-preferring relative to ERC 19, the stochastic ranking value of C_{11} cannot be smaller than the stochastic ranking value of C_9 at time t:

(43) $\theta^t_{11} + \epsilon_{11} \geq \theta^t_9 + \epsilon_9$

The ranking value of C_{11} at time t can be upper bounded as in (44). In fact, C_{11} is promoted only by the two ERCs 8 and 12 in table 15 and each of them can trigger at most $\frac{6\Delta}{\eta} + 4$ updates by fact 1, yielding the term (44a). Furthermore, C_{11} has been demoted by η for each of the $\frac{67\Delta}{\eta} + 35$ updates triggered by ERC 19, yielding the term (44b).

(44) $$\theta^t_{11} \leq \underbrace{2\left(\frac{6\Delta}{\eta} + 4\right)\eta}_{(a)} - \underbrace{\left(\frac{67\Delta}{\eta} + 45\right)\eta}_{(b)}$$

Since $-\Delta < \epsilon_h, \epsilon_k < +\Delta$, since $\theta^t_9 \geq -53\Delta - 27\eta$ (by lemma 6), the two inequalities (43) and (44) contradict each other. □