HG HAS NO COMPUTATIONAL ADVANTAGES OVER OT:
TOWARDS A NEW TOOLKIT FOR COMPUTATIONAL OT

ONLINE SUPPLEMENTARY MATERIALS

GIORGIO MAGRI

Appendix . . .

Appendix A: Proof of Lemma 1

What is currently known in the literature concerning the relationship between the Ranking and the Weighting problem corresponding to a finite set of data is that a solution to the former provides a solution to the latter, as stated in Lemma 1 repeated below. In this subsection, I review the classical proof of this result offered in Prince and Smolensky (2004) and Keller (2000, 2005).

Lemma 1. If a finite set \( D \) of underlying/winner/loser form triplets is OT-compatible, then it is also HG-compatible. More precisely, let \( \gg \) be a ranking OT-compatible with \( D \). Without loss of generality, assume that it is (101a), with \( C_n \) ranked at the top, \( C_{n-1} \) below it and so on, until the bottom ranked \( C_1 \).

\[
\begin{align*}
\text{(101)} \quad & a. \quad C_n & b. \quad \theta_n = (\Delta + 1)^n \\
& C_{n-1} & \theta_{n-1} = (\Delta + 1)^{n-1} \\
& \vdots & \vdots \\
& C_1 & \theta_1 = (\Delta + 1)
\end{align*}
\]

Then, the weight vector \( \theta = (\theta_1, \ldots, \theta_n) \) defined in (101b) is HG-compatible with \( D \), where \( \Delta \) is the largest constraint difference (ignoring sign) over all constraints and all data triplets in the data set \( D \).

Proof. Let \( A = A(D) \) be the set of EWCs corresponding to the set of data triplets \( D \), as defined in (11); let \( A = A(D) \) be the corresponding set of ERCs, as defined in (15). Consider an arbitrary EWC \( \pi = [\pi_1, \ldots, \pi_n] \) in \( A \). Let \( a = [a_1, \ldots, a_n] \) be the corresponding ERC in \( A \). Since the ranking (101a) satisfies the OT-compatibility condition (16) with this ERC \( a \), then there is some \( k \in \{n, n-1, \ldots, 1\} \) such that the top ranked constraints \( C_n, C_{n-1}, \ldots, C_k+1 \) are even for the ERC \( a \) and constraint \( C_k \) is winner-preferrer, as stated in (102a). Thus, the corresponding vector \( \bar{a} \) of constraint differences satisfies (102b).

\[
\begin{align*}
\text{(102)} \quad & a. \quad a_n = a_{n-1} = \ldots = a_{k+1} = e, \quad a_k = w. \\
& b. \quad \bar{a}_n = \bar{a}_{n-1} = \ldots = \bar{a}_{k+1} = 0, \quad \bar{a}_k > 0.
\end{align*}
\]

To simplify notation, let \( B = \Delta + 1 \). The chain of inequalities in (103) shows that the weight vector \( \theta = (\theta_1, \ldots, \theta_n) \) defined in (101b) does indeed satisfy the HG-compatibility condition \( \sum_{i=1}^{n} \theta_i \bar{a}_i > 0 \) in (12).
\[ \sum_{i=1}^{n} \theta_i a_i \overset{(a)}{=} k - 1 \sum_{i=1}^{k-1} \theta_i a_i + \theta_k a_k + \sum_{i=k+1}^{n} \theta_i a_i \]

\[ \overset{(b)}{=} \sum_{i=1}^{k-1} \theta_i a_i + \theta_k a_k \]

\[ \overset{(c)}{=} \sum_{i=1}^{k-1} \theta_i a_i + \theta_k \]

\[ \overset{(d)}{=} - \sum_{i=1}^{k-1} \theta_i (B - 1) + \theta_k \]

\[ \overset{(e)}{=} - \sum_{i=1}^{k-1} B^i (B - 1) + B^k \]

\[ = - \sum_{i=2}^{k-1} B^i + \sum_{i=1}^{k-1} B^i + B^k \]

\[ = -B^k + B + B^k \]

\[ > 0 \]

In (103), I have reasoned as follows: in step (a), I have split up the set \( \{1, \ldots, n\} \) that the index \( i \) runs over into the three subsets \( \{1, \ldots, k - 1\} \), \( \{k\} \) and \( \{k + 1, \ldots, n\} \); in step (b), I have used the fact that \( \pi_0 = \pi_{k-1} = \cdots = \pi_{k+1} = 0 \) by (102b); in step (c), I have lower-bounded by replacing \( \pi_k \) with 1, since the fact that \( \pi_k > 0 \) by (102b) entails that \( \pi_k \geq 1 \), as constraint differences are integers; in step (d), I have lower-bounded by replacing \( \pi_1, \ldots, \pi_{k-1} \) with \(- (B - 1)\), since the absolute value of the entries in the EWC is upper bounded by \( \Delta = B - 1 \); in step (e), I have used the definition (101b) of the weight vector \( \theta = (\theta_1, \ldots, \theta_n) \), restated in terms of \( B = \Delta + 1 \); the remaining steps are simple algebraic manipulations. \( \square \)

**Appendix B: Proof of Lemma 2**

In this Subsection, I present a proof of the main Lemma 2, repeated below. The proof is actually just a straightforward generalization of the reasoning illustrated in Subsection 3.1 with the three examples (26)-(28). It rests on the trivial observation (104). Suppose we have a certain number of weights, say three. The sum of these three weights can always be upper bounded by taking three times the largest among them.

\[ \begin{align*}
\text{Lemma 2.} \ & \text{Given a set } A \text{ of ERCs, consider the corresponding set } \overline{A} \text{ of EWCs derived from } A \text{ as in (29), repeated in (105). Here, } w \text{ is the total number of } w \text{'s in the ERC } a. \\
\text{(105)} \ & a = [a_1, \ldots, a_n] \longrightarrow \overline{a} = [\overline{a}_1, \ldots, \overline{a}_n] \text{ such that } \overline{a}_k \doteq \begin{cases} 
-1 & \text{if } a_k = 1, \\
0 & \text{if } a_k = e, \\
1/w & \text{if } a_k = w
\end{cases}
\end{align*} \]

If \( \overline{A} \) is HG-compatible, then \( A \) is OT-compatible. Furthermore, if \( \theta \) is a solution of the instance WP(\( \overline{A} \)) of the Weighting problem, then any ranking derived from \( \theta \) according to (30) is a solution of the instance RP(\( A \)) of the Ranking problem. \( \blacksquare \)
Proof. Let me show that, if a weight vector $\theta$ is HG-compatible with the EWC $\pi$ in (105), then any of its derived rankings is OT-compatible with the original ERC $a$. Let $W$ and $L$ be the sets of constraints that have a $w$ and an $l$ in the ERC $a$, respectively. Let $w$ be cardinality of the set $W$, namely the total number of $w$’s in $a$. The following chain of inequalities (106) holds for every loser-prefering constraint $k \in L$. Here, I have reasoned as follows: in step (106a), I have used the hypothesis that the weight vector $\theta$ is HG-compatible with $a$ and thus satisfies condition (12); in step (106b), I have split up the set $\{1, \ldots, n\}$ that $h$ runs over into the three sets $W$, $L$ and their complement; in step (106c), I have noted that $a_h = 1/w$ for every $h \in W$, that $a_h = -1$ for every $h \in L$ and that $a_h = 0$ for every $h \not\in W \cup L$, by the definition (105) of the derived EWC $\pi$; in step (106d), I have used (104) to upper bound the sum $\sum_{h \in W} \theta_h a_h$ with its biggest term $\max_{h \in W} \theta_h$ multiplied by the total number $w$ of terms; in step (106e), I have used the hypothesis that all the components of $\theta$ are nonnegative and thus the sum $\sum_{h \in L} \theta_h a_h$ is lower-bounded by one of its terms $\theta_k$ with $k \in L$.

\begin{align}
(106) & \quad 0 \ \overset{(a)}{<} \ \sum_{h=1}^{n} \theta_h a_h \\
& \overset{(b)}{=} \sum_{h \in W} \theta_h a_h + \sum_{h \in L} \theta_h a_h + \sum_{h \not\in W \cup L} \theta_h a_h \\
& \overset{(c)}{=} \frac{1}{w} \sum_{h \in W} \theta_h - \sum_{h \in L} \theta_h \\
& \overset{(d)}{\leq} \frac{1}{w} w \max_{h \in W} \theta_h - \sum_{h \in L} \theta_h \\
& \overset{(e)}{\leq} \max_{h \in W} \theta_h - \theta_k
\end{align}

By reordering the inequality obtained in (106), I conclude that the following strict inequality (107) holds for any loser-prefering constraint $k \in L$. This inequality says that the weight vector $\theta$ has the following property: the largest weight among winner-preferers is strictly larger than the weight of any loser-preferer.

\begin{align}
(107) & \quad \max_{h \in W} \theta_h > \theta_k
\end{align}

Consider a ranking $\gg$ derived from this weight vector $\theta$ according to (30). This means that $\gg$ ranks a constraint above another constraint whenever the weight of the former is larger than the weight of the latter. The inequality (107) thus guarantees that the derived ranking $\gg$ ranks a winner-preferer above any loser-preferer, and is thus OT-compatible with the original EWC $a$. □

Appendix C: Dropping the restriction against negative weights

So far, I have stucked to the restriction (4) that HG weights are nonnegative. In this Subsection, I want to discuss what happens to Lemma 2 when this non-negativity restriction (4) is dropped. This digression is useful for cases where we want to adapt to OT algorithms for HG that do not return weights that are necessarily nonnegative, such as the Perceptron algorithm considered in Section 4. Thus, consider the \textit{unrestricted} Weighting problem (108), which is the Weighting problem (24) without the non-negativity restriction on the weights. I will denote by WP$_{\text{unr}}(A)$ the instance of problem (108) corresponding to a set of EWCs $A$, or equivalently the set of its solutions.

\begin{align}
(108) & \quad \text{Given:} \quad \text{a finite set } A \text{ of EWCs;} \\
& \quad \text{Return:} \quad \perp, \text{ if the data } A \text{ are not HG-compatible; otherwise, a weight vector } \theta \text{ (with no restriction on the sign of the weights) that is HG-compatible with } A.
\end{align}
Lemma 2 establishes an equivalence between the Ranking problem (23) and the Weighting problem (24). Unfortunately, this equivalence does not extend to the unrestricted variant (108) of the latter. Here is a trivial counterexample. Consider the ERC \( \mathbf{a} \) in (109), together with the corresponding EWC \( \mathbf{\pi} \) derived according to (29)-(105), namely by replacing each \( l \) with \(-1\) and the \( w \) with 1. Consider the weight vector that assigns the weight \( \theta_1 = -4 \) to constraint \( C_1 \) and the weight \( \theta_2 = \theta_3 = -3 \) to both constraints \( C_2 \) and \( C_3 \). This weight vector is HG-compatible with the EWC \( \mathbf{\pi} \) (as \( 1 \cdot (-4) + (-1) \cdot (-3) + (-1) \cdot (-3) > 0 \)). Yet, it admits the derived ranking \( C_2 \gg C_3 \gg C_1 \) which is not OT-compatible with the ERC \( \mathbf{a} \) we started from.

\[
\begin{align*}
\mathbf{a} &= \begin{bmatrix} c_1 & c_2 & c_3 \\ w & l & l \end{bmatrix} \quad \mathbf{\pi} &= \begin{bmatrix} c_1 & c_2 & c_3 \\ +1 & -1 & -1 \end{bmatrix}
\end{align*}
\]

Yet, the assumption (4) that the weights \( \theta_1, \ldots, \theta_n \) be non-negative was used only once in the proof of Lemma 2, namely in step (106e), in order to lower-bound the sum \( \sum_{h \in L} \theta_h \) of the weights of loser-preferrers with one of those weights \( \theta_k \) with \( k \in L \). If the assumption (4) that the weights be non-negative is dropped, then this lower-bound (106e) does not hold any more (the sum of negative numbers is smaller than each of the negative numbers). Yet the bound trivially holds, even without the non-negativity restriction, provided that the ERC \( \mathbf{a} \) has a unique loser-prefering constraint \( C_k \), as in this case \( L = \{k\} \) and thus \( \sum_{h \in L} \theta_h = \theta_k \). In conclusion, the equivalence between the Ranking and the Weighting problem established by Lemma 2 does extend to the unrestricted variant (108) of the Weighting problem, provided that there is a unique \( l \) per ERC, as stated in the following variant of Lemma 2.

**Lemma 1.** Given a set \( \mathbf{A} \) of ERCs with at most one \( l \) each, consider the corresponding set \( \mathbf{\bar{A}} \) of EWCs derived from \( \mathbf{A} \) as in (105). If \( \mathbf{\bar{A}} \) is HG-compatible, then \( \mathbf{\bar{A}} \) is OT-compatible. More precisely, if a weight vector \( \mathbf{\theta} \) (with possibly negative weights) solves the instance \( \text{WP}_{\text{unr}}(\mathbf{\bar{A}}) \) of the unrestricted Weighting problem (108), then any ranking derived from \( \mathbf{\theta} \) according to (30) solves the instance \( \text{RP}(\mathbf{A}) \) of the Ranking problem (23).

As already noted in Subsection 4.3, the assumption that each ERC in the set \( \mathbf{A} \) has a unique \( l \) is not restrictive. In fact, suppose that \( \mathbf{A} \) contains an ERC that has, say, two \( l \)'s. Let \( \mathbf{A}' \) be the set of ERCs obtained by replacing ERC with two identical ERCs, but for the fact that each of them retains only one of the two original \( l \)'s, while the other is replaced by \( e \), as illustrated in (110). Obviously, a ranking \( \gg \) is OT-compatible with \( \mathbf{A} \) iff it is OT-compatible with \( \mathbf{A}' \). In other words, \( \mathbf{A} \) and \( \mathbf{A}' \) yield the same instance of the Ranking problem (23), namely \( \text{RP}(\mathbf{A}) = \text{RP}(\mathbf{A}') \).

\[
\begin{align*}
\mathbf{A} &= \begin{bmatrix} e & w & \ldots & l & l \\ \ldots \end{bmatrix} \quad \mathbf{A}' &= \begin{bmatrix} e & w & \ldots & l & e \\ e & w & \ldots & e & l \end{bmatrix}
\end{align*}
\]

In conclusion, I can always assume without loss of generality that a given instance of the Ranking problem consists of ERCs with at most one \( l \) each, because if that is not the case, then I can pre-process them into ERCs that indeed have a unique \( l \) each.

**Appendix D: Proof of Lemma 3**

In this section, I prove Lemma 3, which ensures that a run of the revised GLA can be mimicked by a run of the Perceptron. To start in subsection D.1, I consider the restricted case of input ERCs that have a unique \( l \) each, where the core idea is not obscured by technical details. In subsection D.2, I then extend the reasoning to the general case where input ERCs have an arbitrary number of \( l \)'s.
Appendix D.1: Restricted case of a single loser-preferrer per input ERC. The key ingredient of the reasoning is again the mapping (29)=(105) from an ERC \( a \) into a derived EWC \( \overline{a} \), repeated once more in (111): each \( e \) is mapped to 0; each \( t \) is mapped to \(-1\); and each \( w \) is mapped to \( 1/w \), where \( w \) is the total number of \( w \)'s in the ERC considered.

\[
(111) \quad a = [a_1, \ldots, a_n] \rightarrow \overline{a} = [\overline{a}_1, \ldots, \overline{a}_n] \quad \text{where} \quad \overline{a}_k = \begin{cases} 
-1 & \text{if } a_k = t \\
0 & \text{if } a_k = e \\
1/w & \text{if } a_k = w 
\end{cases}
\]

Lemma 3 for the restricted case of input ERCs with a unique \( t \) each can be stated more explicitly as follows.

**Lemma 3 (restricted case).** Consider a run (112) of the revised GLA (45) on an input set \( A \) of ERCs. Assume without loss of generality that an update is triggered at every time (i.e. \( \theta_t \neq \theta_{t+1} \) for every \( t \)). The initial weight vector is \( \theta_1 \); the algorithm is then fed a certain ERC from \( A \), call it \( a_1 \); and it updates the current weight vector to \( \theta_2 \); and so on. Let \( \theta_1, \theta_2, \ldots, \theta_t, \ldots \) be the sequence of weight vectors entertained by the revised GLA in this run.

\[
(112) \quad \theta_1 \rightarrow a_1 \theta_2 \rightarrow a_2 \theta_3 \rightarrow a_3 \ldots \rightarrow \theta_t \rightarrow a_t \theta_{t+1} \rightarrow \ldots
\]

Consider the corresponding run (113) of the Perceptron, whereby the Perceptron is fed at every time the EWC \( \overline{a}_t \) corresponding through (111) to the ERC \( a_t \) fed to the revised GLA at that same time in the run (112). Let \( \overline{\theta}_1, \overline{\theta}_2, \ldots, \overline{\theta}_t, \ldots \) be the sequence of weight vectors entertained by the Perceptron in this run.

\[
(113) \quad \overline{\theta}_1 \rightarrow \overline{a}_1 \overline{\theta}_2 \rightarrow \overline{a}_2 \overline{\theta}_3 \rightarrow \overline{a}_3 \ldots \rightarrow \overline{\theta}_t \rightarrow \overline{a}_t \overline{\theta}_{t+1} \rightarrow \ldots
\]

Assume that each input ERC in \( A \) fed to the revised GLA in the run (112) has a unique \( t \). Assume furthermore that the run (112) of the revised GLA and the corresponding run (113) of the Perceptron start from the same initial weight vector, namely \( \theta_{1} = \overline{\theta}_{1} \). Then, the two runs always entertain the same current weight vector, namely \( \theta_{t} = \overline{\theta}_{t} \) at any time \( t \), as stated in (114).

\[
(114) \quad \text{If } \theta_{t} = \overline{\theta}_{t}, \text{ then } \theta_{t} = \overline{\theta}_{t} \text{ for every time } t.
\]

Finally, OT-compatibility of the set \( A \) of input ERCs fed to the revised GLA in the run (112) entails HG-compatibility of the set \( \overline{A} \) of input EWCs fed to the Perceptron in the corresponding run (113).

**Proof.** The proof of (114) is by induction on time \( t \). The hypothesis that \( \theta_{t} = \overline{\theta}_{t} \) provides the basis of induction. As the inductive hypothesis, assume that \( \theta_{t} = \overline{\theta}_{t} \). As the inductive step, let me show that \( \theta_{t+1} = \overline{\theta}_{t+1} \). The ERC \( a_t \) fed to the revised GLA at time \( t \) in the run (112) triggers an update of the current weight vector \( \theta_t \), as we have discarded ERCs that trigger no update. This means that the current weight vector \( \theta_t \) admits derived rankings that are not OT-compatible with the current ERC \( a_t \). Because of the assumption that all input ERCs have a unique \( t \), Lemma 1 applies. This Lemma ensures that HG-compatibility between a weight vector and a derived EWC entails OT-compatibility between all the derived rankings of that weight vector and the original ERC. As the weight vector \( \theta_t \) admits derived ranking not OT-compatible with the ERC \( a_t \), this Lemma ensures by contraposition that this weight vector \( \theta_t \) is not HG-compatible with the derived EWC \( \overline{a}_t \). As the two weight vectors \( \theta_t \) and \( \overline{\theta}_t \) coincide by inductive hypothesis, I conclude that the current weight vector \( \overline{\theta}_t \) in the Perceptron run (113) is not HG-compatible with the current EWC \( \overline{a}_t \). Thus, an update is performed at time \( t \) both in the revised GLA run (112) and in the Perceptron run (113). Update according to the
revised GLA update rule (45) based on the current ERC $a_t$ denotes the loser-preferrer by 1 and promotes each winner-preferrer by $1/w$, where $w$ is the total number of w's in the ERC $a_t$. Update according to the Perceptron update rule (48) based on the EWC $\mathbf{\bar{a}}_t$ derived from $a_t$ adds $-1$ to the component corresponding to the loser-preferrer and adds $1/w$ to the components corresponding to the winner-preferrers. The two updates are thus identical. In conclusion, as the two current weight vectors $\mathbf{\theta}_t$ and $\mathbf{\bar{\theta}}_t$ entertained by the revised GLA in the two runs are identical (by the inductive hypothesis), as they are both updated, and as they are updated in exactly the same way, then the two subsequent weight vectors $\mathbf{\theta}_{t+1}$ and $\mathbf{\bar{\theta}}_{t+1}$ in the two runs are identical too. Finally, HG-compatibility of $\mathbf{\bar{X}}$ follows from OT-compatibility of $\mathbf{A}$ through Lemma 1. □

Appendix D.2: General case with an arbitrary number of loser-preferrers. So far, I have assumed that the ERCs fed to the revised GLA have only one $l$ each. Let me make explicit why I had to make this restrictive assumption. As noted in footnote 9, the Perceptron does not ensure that the current weights are nonnegative. Even if the algorithm is initialized with very large positive initial weights, there is no guarantee that the weights will stay positive until convergence, as the total number of updates depends on the initial weights. Since I cannot guarantee that the weights be nonnegative, then I cannot use Lemma 2 in order to prove the crucial fact that OT-incompatibility between the current ERC and the current weight vector in the run (112) of the revised GLA entails HG-incompatibility between the derived EWC and the current weight vector in the run (113) of the Perceptron. Thus, I restricted myself to input ERCs with a unique $l$ in order to apply Lemma 1 instead, that does not require the weights to be non-negative.

In order to get around this difficulty, the mapping (111) used above needs to be replaced with the mapping (115). Given an ERC $a$ and a weight vector $\mathbf{\theta} = (\theta_1, \ldots, \theta_n)$, an undominated loser-preferrer is a constraint $C_k$ that has an $l$ in the ERC $a$ and whose weight $\theta_k$ is not smaller than the weight $\theta_l$ of some constraint $C_h$ that has a $w$ in $a$. The mapping (115) that takes an ERC $a$ and a weight vector $\mathbf{\theta}$ and returns the derived EWC $\mathbf{\bar{a}}$ defined as follows: each $l$ corresponding to an undominated loser-preferrer is mapped to $-1$; each $w$ is mapped to $\ell/w$, where $\ell$ is the number of undominated loser-preferrers and $w$ is the number of winner-preferrers; all other entries are set equal to 0.

$$ (a, \mathbf{\theta}) \mapsto \mathbf{\bar{a}} = [\pi_1, \ldots, \pi_n] \text{ where } \pi_k = \begin{cases} -1 & \text{if } C_k \text{ is an undom. loser-preferrer} \\ \ell/w & \text{if } C_k \text{ is a winner-preferrer} \\ 0 & \text{otherwise} \end{cases} $$

The crucial difference between the two mappings (111) and (115) is as follows: according to the former, the derived EWC $\mathbf{\bar{a}}$ only depends on an ERC $a$; according to the latter, it depends instead also on a weight vector $\mathbf{\theta}$, as the weight vector is used to determine the undominated loser-preferrers. The additional power that comes with this new mapping (115) allows me to drop the restriction to input ERCs with a unique $l$, as stated in the following formulation of Lemma 3. The proof is identical to the proof provided above for the restricted case where input ERCs have a unique $l$ each. The only difference is that Lemma 2 below is used instead of Lemma 1.

**Lemma 3 (general case).** Consider a run (116) of the revised GLA (45) on an input set $A$ of ERCs. Assume without loss of generality that an update is triggered at every time (i.e. $\mathbf{\theta}_t \neq \mathbf{\theta}_{t+1}$ for every $t$). Let $\mathbf{\theta}_1, \mathbf{\theta}_2, \ldots, \mathbf{\theta}_t, \ldots$ be the sequence of weight vectors entertained by the revised GLA in this run.

$$ (116) \quad \mathbf{\theta}_1 \xrightarrow{a_1} \mathbf{\theta}_2 \xrightarrow{a_2} \mathbf{\theta}_3 \xrightarrow{a_3} \ldots \xrightarrow{a_t} \mathbf{\theta}_t \xrightarrow{a_{t+1}} \mathbf{\theta}_{t+1} \xrightarrow{\ldots} $$

The notation $\mathbf{\pi}$ is thus a slight abuse for the mapping (115), as it reveals only the dependence of the derived EWC on the ERC, but hides the dependence on the weight vector. Nonetheless, I stick to this notation for consistency with the rest of the paper.
Consider the corresponding run (117) of the Perceptron, whereby the Perceptron is fed at every time \( t \) the EWC \( \overline{a}_t \) corresponding through (115) to the ERC \( a_t \) fed to the revised GLA at that same time in the run (116).

\[
\begin{align*}
\overline{a}_1 \xrightarrow{\pi_1} \overline{a}_2 \xrightarrow{\pi_2} \overline{a}_3 \xrightarrow{\pi_3} \ldots \xrightarrow{\pi_t} \overline{a}_t \xrightarrow{\pi_{t+1}} \ldots
\end{align*}
\]

Assume that the run (116) of the revised GLA and the derived run (117) of the Perceptron start from the same initial weight vector, namely \( \theta_1 = \overline{\theta}_1 \). Then, the two runs always entertain the same current weight vector, namely \( \theta_t = \overline{\theta}_t \) for any time \( t \), as stated in (118).

(118) If \( \theta_1 = \overline{\theta}_1 \), then \( \theta_t = \overline{\theta}_t \) for every time \( t \).

Finally, the set \( A \) of input EWCs fed to the Perceptron in (117) is always finite; it is furthermore HG-compatible, if the set \( A \) of input ERCs fed to the revised GLA in the run (116) is OT-compatibility. ■

I conclude this Subsection with a proof of Lemma 2, which is actually just a slight variant of the proof of Lemma 2 presented above in Appendix ??.

**Lemma 2.** Given an ERC \( a \) and a weight vector \( \theta \), consider the EWC \( \overline{a} = (\overline{a}_1, \ldots, \overline{a}_n) \) derived from them according to (115). If the ERC \( a \) is not OT-compatible with some ranking derived from the weight vector \( \theta \), then the corresponding derived EWC \( \overline{a} \) is not HG-compatible with \( \theta \) either.

**Proof.** Let \( W \) be the set of (indices corresponding to) constraints that have a \( w \) in the ERC \( a \), as in (119a); let \( L \) be the set of (indices corresponding to) constraints that have an \( l \) in the ERC \( a \) and are furthermore undominated relative to the weight vector \( \theta \), in the sense that their weight vector is not strictly smaller than the weight vector of some winner-preferrer, as in (119b). The hypothesis that the ERC \( a \) is not OT-compatible with some ranking derived from the weight vector \( \theta \) ensures that \( L \) is not empty.

\[
\begin{align*}
W &= \{ k \mid a_k = w \} & \text{(a)} \\
L &= \left\{ k \mid a_k = l, \; \theta_k \geq \max_{h \in W} \theta_h \right\} & \text{(b)}
\end{align*}
\]

The chain of inequalities (120) shows that the EWC \( \overline{a} \) derived from \( a \) and \( \theta \) according to (115) is not HG-compatible with the weight vector \( \theta \). This chain of inequalities is analogous to the chain of inequalities (106) used in Appendix ?? to prove Lemma 2.

\[
\begin{align*}
\sum_{h=1}^n \theta_h \overline{a}_h &= \sum_{h \in W} \theta_h \overline{a}_h + \sum_{h \in L} \theta_h \overline{a}_h + \sum_{h \notin W \cup L} \theta_h \overline{a}_h \\
&\overset{(b)}{=} \ell \frac{w}{w} \sum_{h \in W} \theta_h - \sum_{h \in L} \theta_h \\
&\overset{(c)}{\leq} \ell \frac{w}{w} \max_{h \in W} \theta_h - \sum_{h \in L} \theta_h \\
&\overset{(d)}{\leq} \ell \max_{h \in W} \theta_h - \ell \min_{h \in L} \theta_h \\
&= \ell \left( \max_{h \in W} \theta_h - \min_{h \in L} \theta_h \right) \\
&\overset{(e)}{\leq} 0
\end{align*}
\]

Here, I have reasoned as follows: in step (a), I have split the set \{1, \ldots, n\} that the index \( h \) runs over into the the set of winner-preferrers \( W \) in (119a), the set of undominated loser-preferrers \( L \) in (119b) and their complement; in step (b), I have used the definition (115) of the components \( \overline{a}_1, \ldots, \overline{a}_n \) of the derived EWC \( \overline{a} \); in step (c), I have used once
more (104) to upper bound the sum $\sum_{h \in W} \theta_h$ with its biggest term $\max_{h \in W} \theta_h$ multiplied by the number $w$ of terms; in step (d), I have reasoned analogously to lower bound the sum $\sum_{h \in L} \theta_h$ with its smallest term $\min_{h \in L} \theta_h$ multiplied by the number $\ell$ of terms; in step (e), I have used the hypothesis that the weight vector $\theta$ admits a refinement which is not OT-compatible with the ERC $a$, which means in turn that there is a loser-preferer whose weight is at least as large as the largest weight among winner-preferers, so that the quantity (*) is not positive.

\[ \square \]

**APPENDIX E: PROOF OF THE PERCEPTRON CONVERGENCE THEOREM**

In this Subsection, I recall for completeness the proof of the classical Perceptron convergence Theorem 1; see Block (1962), Novikoff (1962), and Rosenblatt (1958, 1962), as well as Cristianini and Shawe-Taylor (2000, Theorem 2.3) and Cesa-Bianchi and Lugosi (2006, Chp. 12) for a more recent presentation and a broader perspective. Throughout this Subsection, I use standard notations from Linear Algebra: $\langle \cdot, \cdot \rangle$ is the *Euclidean scalar product*, defined by $\langle v, w \rangle = \sum_{i=1}^{n} v_i w_i$ for any pair of vectors $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$; $\| \cdot \|$ is the *Euclidean norm*, defined by $\|v\|^2 = \langle v, v \rangle = \sum_{i=1}^{n} v_i^2$; they are connected by the Cauchy-Schwartz inequality $\langle v, w \rangle \leq \|v\| \|w\|$. The HG-compatibility condition (12) between a weight vector $\theta$ and an EWC $\overline{a}$ can be expressed in terms of the Euclidean scalar product as $\langle \theta, \overline{a} \rangle > 0$.

**Theorem 1.** The HG online algorithm (47) with the HG Perceptron update rule (48) converges, provided that the finite input set of EWCs is HG-compatible.

**Proof.** Without loss of generality, assume that the initial weight vector is the null vector. Let $\theta^{t+1}$ be the weight vector obtained at time $t$ by updating the current weight vector $\theta^t$ in response to the current EWC $\overline{a}$ according to the Perceptron update rule (48). As current EWCs that do not trigger an update can be discarded, I can assume without loss of generality that an update is performed at each time $t$. The proof has three parts. The first part of the proof estimates the norm of the current weight vector entertained by the algorithm. To start, note that the norm of the updated weight vector $\theta^{t+1}$ can be bound as in (121) in terms of the norm of the current weight vector $\theta^t$. Here, I have reasoned as follows: in step (a), I have used the definition (48) of the Perceptron HG update rule; in step (b), I have used the identity $\|v + w\|^2 = \|v\|^2 + \|w\|^2 + 2 \langle v, w \rangle$; in step (c), I have upper bounded by dropping the term $\langle \theta^t, \overline{a} \rangle$, which is negative by the hypothesis that the current weight vector $\theta^t$ is not HG-compatible with the current EWC $\overline{a}$; in step (d), I have upper bounded the norm of the current EWC $\overline{a}$ with the largest norm $R$ over all input EWCs (note that $R$ is finite, as there are only a finite number of input EWCs).

\begin{align}
\|\theta^{t+1}\|^2 & \leq \langle \theta^t + \overline{a} \rangle^2 \\
& \stackrel{(a)}{=} \|\theta^t + \overline{a}\|^2 \\
& \stackrel{(b)}{=} \|\theta^t\|^2 + \|\overline{a}\|^2 + 2 \langle \theta^t, \overline{a} \rangle \\
& \stackrel{(c)}{\leq} \|\theta^t\|^2 + \|\overline{a}\|^2 \\
& \stackrel{(d)}{\leq} \|\theta^t\|^2 + R^2 \quad \text{where } R = \max \{ \|\overline{a}\| : \overline{a} \text{ is an input EWC} \}
\end{align}

Since the inequality (121) holds at every time $t$ and since the initial weight vector has null norm, then we obtain the inequality (122), which concludes the first part of the proof, showing that the norm of the current weight vector grows with time $t$ slower than $\sqrt{t}$.

\begin{align}
\|\theta^t\|^2 & \leq tR^2
\end{align}

Consider now a weight vector $\theta$ that is HG-compatible with the input set of EWCs, that exists because of the hypothesis that the latter is HG-compatible. The second part of the proof estimates the scalar product between the latter weight vector and the current weight vector entertained by the algorithm. To start, note that the scalar product between the
weight vector $\theta$ and the updated weight vector $\theta^{t+1}$ can be bound as in (123) in terms of the scalar product between that same weight vector $\theta$ and the current weight vector $\theta^t$. Here, I have reasoned as follows: in step (a), I have used the definition (48) of the Perceptron HG update rule; in step (b), I have used the linearity of the scalar product; in step (c), I have lower bounded the quantity $\langle \theta, \pi \rangle / ||\theta||$ with the smallest such quantity over all input EWCs, called $\mu(\theta)$ (note that this quantity is finite, as there are only a finite number of input EWCs).

$$(123) \quad \frac{\langle \theta, \theta^{t+1} \rangle}{||\theta||} \overset{(a)}{=} \frac{\langle \theta, \theta^t + \pi \rangle}{||\theta||} \overset{(b)}{=} \frac{\langle \theta, \theta^t \rangle}{||\theta||} + \frac{\langle \theta, \pi \rangle}{||\theta||} \overset{(c)}{=} \frac{\langle \theta, \theta^t \rangle}{||\theta||} + \mu(\theta)$$

Since the inequality (123) holds at every time $t$ and since the initial weight vector is null, then we obtain the inequality (124), which concludes the second part of the proof, showing that the scalar product between the weight vector $\theta$ and the current weight vector $\theta^t$ grows faster than $t$.

$$(124) \quad \frac{\langle \theta, \theta^t \rangle}{||\theta||} \geq t\mu.$$  

The third part of the proof connects the two inequalities (122) and (124) as in (125). Here, I have reasoned as follows: in step (a), I have used (124); in step (b), I have used the Cauchy-Schwartz inequality; in step (c), I have used (122).

$$$(125) \quad (t\mu)^2 \overset{(a)}{=} \left( \frac{\langle \theta, \theta^t \rangle}{||\theta||} \right)^2 \overset{(b)}{=} \frac{||\theta^t||^2}{||\theta||^2} \overset{(c)}{=} ||\theta^t||^2 \leq tR^2$$

The chain of inequalities (125) entails in particular that $t \leq R^2/\mu^2$, namely that the number of updates $t$ is finite. \hfill \square

The proof just presented actually shows that the total number $T$ of updates performed by the Perceptron can be bound as in (126), in terms of the two quantities $R$ and $\mu(\theta)$ defined in (121) and (123).

$$$(126) \quad T \leq \left( \frac{R}{\mu(\theta)} \right)^2$$

The bound can be optimized by considering the weight vector $\hat{\theta}$ that maximizes $\mu(\theta)$. The corresponding quantity $\mu = \mu(\hat{\theta})$ is called the margin of the set of input EWCs.

**Appendix F: Proof of Corollaries 1 and 2**

In this section, I present a proof of Corollaries 1 and 2, that bound the worst-case number of updates performed by the revised GLA (45) on diagonal sets of input ERCs as well as on. As these ERCs have a unique $l$ per row, consider again the restricted formulation of Lemma 3 in terms of the mapping (111) considered at the beginning of Subsection ??.. That Lemma entails in particular that bounds on the number of updates performed by the revised GLA on a set of input ERCs (with a unique $l$ each) are straightforwardly obtained by bounding the number of updates performed by the Perceptron on the set of input EWCs.
derived through (111). The proof presented below thus just expresses in terms of the number $n$ of constraints the bound (126) provided at the end of Appendix ?? for the number of updates performed by the Perceptron. The standard notations form Linear Algebra used in Appendix ?? are carried over to this Subsection.

**Corollary 1.** The revised GLA (45) cannot perform more than $n(n^2 - 1)/6$ updates on the diagonal set of input ERCs corresponding to $n$ constraints (starting from the null initial weight vector).

**Proof.** Let me denote by $\bar{A}_n$ the set of EWCs derived through the mapping (111) from the diagonal set $A_n$ of ERCs corresponding to $n$ constraints, as illustrated in (54b). As noted above, the worst case number of updates $T_{\text{GLA}}(n)$ performed by the revised GLA on the diagonal set $A_n$ of input ERCs is at most as large as the worst case number of updates $T_{\text{Per}}(n)$ performed by the Perceptron on the set $\bar{A}_n$ of derived EWCs, as stated in (127a). As noted at the end of Appendix ??, the latter worst-case number $T_{\text{Per}}(n)$ of updates can in turn be bound as in (127b), in terms of the radius $R$ and the margin $\mu$ of $A_n$.

\begin{equation}
T_{\text{GLA}}(n) \leq T_{\text{Per}}(n) \leq \frac{R^2}{\mu^2}
\end{equation}

where $R = R(\bar{A}_n) = \text{maximum norm of the EWCs in } \bar{A}_n$

$\mu = \mu(\bar{A}_n) = \text{margin of the set of EWCs } \bar{A}_n$

The squared norm of any EWC in $\bar{A}_n$ is 2. In order to conclude the proof, I thus only need to lower bound the squared inverse $1/\mu^2$ of the margin. Vapnik (1998, Theorem 10.2) ensures that $1/\mu^2$ coincides with the (unique) solution of the quadratic optimization problem (128) in the decision variable $\theta \in \mathbb{R}^n$.

\begin{equation}
\begin{aligned}
m\text{inimize: } & \|\theta\|^2 \\
\text{subject to: } & \langle \theta, \pi \rangle \geq 1 \text{ for every ERC } \pi \text{ in the set } \bar{A}_n
\end{aligned}
\end{equation}

An EWC $\pi$ in $\bar{A}_n$ is called a support vector iff the condition $\langle \theta, \pi \rangle \geq 1$ in the definition of the feasible set in (128) holds tight at optimality, namely $\langle \theta^*, \pi \rangle = 1$, where $\theta^*$ is the unique solution of the optimization problem (128). In the case of diagonal sets of ERCs, all EWCs are support vectors, so that the optimization problem (128) is equivalent to (129).²

\begin{equation}
\begin{aligned}
m\text{inimize: } & \|\theta\|^2 \\
\text{subject to: } & \langle \theta, \pi \rangle = 1 \text{ for every EWC } \pi \text{ in the set } \bar{A}_n
\end{aligned}
\end{equation}

To illustrate, consider for instance the case $n = 5$, repeated in (130) together with the corresponding derived EWCs. For convenience, I have numbered the constraints from right to left and the rows from bottom to top.

\begin{equation}
\begin{bmatrix}
C_5 & C_4 & C_3 & C_2 & C_1
\end{bmatrix}
\begin{bmatrix}
W & L
W & L
W & L
W & L
\end{bmatrix}
\implies
\begin{bmatrix}
\theta_5 & \theta_4 & \theta_3 & \theta_2 & \theta_1
\end{bmatrix}
\begin{bmatrix}
+1 & -1 & +1 & -1 & +1 & -1
\end{bmatrix}
\end{equation}

If we fix, say, the weight $\theta_1$ corresponding to constraint $C_1$, then there exists a unique weight vector that satisfies the condition $\langle \theta, \pi \rangle = 1$ for every derived EWC $\pi$, namely the

²Let me explain why all rows are support vectors. Let $\bar{A} \setminus \pi$ be the HG-comparative tableau obtained from $\bar{A}$ by suppressing one of its row $\pi$. It is well known that a row $\pi$ is not a support vector iff it can be dropped without affecting HG-compatibility, namely the two tableaux $\bar{A}$ and $\bar{A} \setminus \pi$ are HG-compatible with exactly the same weight vectors. There is no single row of the HG-comparative tableau derived from a diagonal OT-comparative tableau that can be dropped without affecting HG-compatibility.
weight vector \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n) \) constructed as in (131): we use the value of \( \hat{\theta}_1 \) and EWC 1 in order to conclude that \( \hat{\theta}_2 \) must exceed \( \hat{\theta}_1 \) by 1; then, we use this value of \( \hat{\theta}_2 \) and EWC 2 in order to conclude that \( \hat{\theta}_3 \) must exceed \( \hat{\theta}_1 \) by 2; and so on.

(131) \[
\begin{align*}
\text{row } 1 & \Rightarrow \hat{\theta}_2 - \hat{\theta}_1 = 1 \Rightarrow \hat{\theta}_2 = \hat{\theta}_1 + 1 \\
\text{row } 2 & \Rightarrow \hat{\theta}_3 - \hat{\theta}_2 = 1 \Rightarrow \hat{\theta}_3 = \hat{\theta}_2 + 1 = \hat{\theta}_1 + 2 \\
\text{row } 3 & \Rightarrow \hat{\theta}_4 - \hat{\theta}_3 = 1 \Rightarrow \hat{\theta}_4 = \hat{\theta}_3 + 1 = \hat{\theta}_1 + 3 \\
\text{row } 4 & \Rightarrow \hat{\theta}_5 - \hat{\theta}_4 = 1 \Rightarrow \hat{\theta}_5 = \hat{\theta}_4 + 1 = \hat{\theta}_1 + 4
\end{align*}
\]

In the general case, there is a unique weight vector \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n) \) such that \( \hat{\theta}_1 \) is equal to a fixed value and furthermore \( \langle \hat{\theta}, \pi \rangle = 1 \) for every EWC \( \pi \) in \( \mathcal{A}_n \), namely (132).

(132) \[
\hat{\theta}_k = \hat{\theta}_1 + (k - 1) \quad k = 1, 2, \ldots, n
\]

Thus, in order to solve the optimization problem (129) it is sufficient to solve the optimization problem (133a) in the scalar decision variable \( \theta_1 \), as the weight vector that solves (129) can then be reconstructed thorough (132) from the value of \( \theta_1 \) that solves (133a). As the objective function of the optimization problem (133a) is (strictly) convex, the unique solution can be determined by setting its derivative to zero, which yields the solution (133b).

(133) a. minimize: \[
\sum_{k=1}^{n} \left( \theta_1 + k - 1 \right)^2
\]
subject to: \( \theta_1 \in \mathbb{R} \)

b. \( \hat{\theta}_1 = -\frac{1}{2} (n - 1) \)

In conclusion, the squared inverse of the margin \( \mu \) can be computed as in (134). Here, I have reasoned as follows: in step (a), I have used the fact that \( 1/\mu^2 \) coincides with the squared norm of the unique solution \( \hat{\theta} \) of the optimization problem (128), or equivalently (129); in step (b), I have used the fact that the latter vector can be described as in (132); in step (c), I have used the fact that the value \( \hat{\theta}_1 \) is provided by (133b); the remaining identities are simple algebraic manipulations.

(134) \[
\frac{1}{\mu^2} = |\hat{\theta}|^2 = \sum_{k=1}^{n} \left( \hat{\theta}_1 + k - 1 \right)^2
= \sum_{k=1}^{n} \left( -\frac{1}{2} (n - 1) + k - 1 \right)^2
= \sum_{k=1}^{n} \left( \frac{1}{4} (n - 1)^2 + (k - 1)^2 - (n - 1)(k - 1) \right)
= \frac{1}{4} n(n - 1)^2 + \sum_{k=1}^{n-1} k^2 - (n - 1) \sum_{k=1}^{n-1} k
= \frac{1}{4} n(n - 1)^2 + \frac{1}{6} n(n - 1)(2n - 1) - \frac{1}{2} n(n - 1)^2
= \frac{1}{12} n(n^2 - 1)
\]

The claim thus follows from the upper bound \( R^2/\mu^2 \) provided in (127), together with the identities \( R^2 = 2 \) and \( 1/\mu^2 = n(n^2 - 1)/12 \). \( \square \)

**Corollary 2** The worst-case number of updates of the revised GLA (45) run on Pater’s comparative tableau of order \( n \) starting from the null initial vector grows with \( n \) at most at the order of \( n^5 \). \( \blacksquare \)
Proof. Let me denote by \( \overline{A}_n \) the HG-comparative tableau derived through the mapping (??) from Pater’s OT-comparative tableau \( A_n \) of order \( n \). To illustrate, I provide in (135) the derived HG-comparative tableaux corresponding to the three Pater’s tableaux in (55).

\[
\begin{align*}
(135) \quad \overline{A}_4 &= \begin{bmatrix}
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
1 & -1 & 1 & -1 & 1 \\
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
\end{bmatrix} \quad \overline{A}_5 &= \begin{bmatrix}
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
1 & -1 & 1 & -1 & 1 \\
\end{bmatrix} \quad \overline{A}_6 &= \begin{bmatrix}
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
\frac{1}{2} & -1 & \frac{1}{2} & -1 & \frac{1}{2} \\
1 & -1 & 1 & -1 & 1 \\
\end{bmatrix}
\end{align*}
\]

Again as in the preceding proof, the worst case number of updates required by the revised GLA on Pater’s comparative tableau \( A_n \) can be bound by \( R^2/\mu^2 \). The maximum norm \( R \) of the rows of \( \overline{A}_n \) is a constant, namely it does not depend on \( n \). Again as in the preceding proof, the squared inverse \( 1/\mu^2 \) of the margin coincides with the solution of the optimization problem (129), repeated in (136).

\[
(136) \quad \text{minimize: } \|\theta\|^2 \\
\text{subject to: } \langle \theta, \overline{a} \rangle = 1 \quad \text{for every row } \overline{a} \text{ of Pater’s derived tableau } \overline{A}_n
\]

Suppose that the weight \( \theta_1 \) is fixed to zero. Then the condition that \( \langle \theta, \overline{a} \rangle = 1 \) for every row \( \overline{a} \) of the derived Pater HG-comparative tableau univocally determines the weight vector \( \theta \). To illustrate, consider Pater’s comparative tableau of order \( n = 5 \), repeated in (137) together with the corresponding derived Pater HG-comparative tableau. For convenience, I have numbered the constraints from right to left and the rows from bottom to top.

\[
(137) \quad \begin{align*}
\text{row 4} & \quad \begin{bmatrix} C_5 & C_1 & C_2 & C_1 & C_1 \end{bmatrix} \\
\text{row 3} & \quad \begin{bmatrix} W & L & W & W & W \end{bmatrix} \\
\text{row 2} & \quad \begin{bmatrix} W & L & W & W & W \end{bmatrix} \\
\text{row 1} & \quad \begin{bmatrix} W & L & W & W & W \end{bmatrix} \\
\end{align*} \quad \implies \begin{align*}
\text{row 4} & \quad \begin{bmatrix} \theta_5 & \theta_4 & \theta_3 & \theta_2 & \theta_1 \end{bmatrix} = \begin{bmatrix} +1 & -2 & +1 & 0 & 0 \end{bmatrix} \\
\text{row 3} & \quad \begin{bmatrix} \theta_5 & \theta_4 & \theta_3 & \theta_2 & \theta_1 \end{bmatrix} = \begin{bmatrix} +1 & +1 & 0 & 0 & 0 \end{bmatrix} \\
\text{row 2} & \quad \begin{bmatrix} \theta_5 & \theta_4 & \theta_3 & \theta_2 & \theta_1 \end{bmatrix} = \begin{bmatrix} +1 & +1 & 0 & 0 & 0 \end{bmatrix} \\
\text{row 1} & \quad \begin{bmatrix} \theta_5 & \theta_4 & \theta_3 & \theta_2 & \theta_1 \end{bmatrix} = \begin{bmatrix} +1 & +1 & 0 & 0 & 0 \end{bmatrix}
\end{align*}
\]

There exists a unique weight vector \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_5) \) such that \( \hat{\theta}_1 = 0 \) and furthermore \( \langle \hat{\theta}, \overline{a} \rangle = 1 \) for every row \( \overline{a} \) of the derived Pater HG-comparative tableau, namely the vector constructed as in (138): to start, we set \( \hat{\theta}_1 = 0 \); then, we use this value of \( \hat{\theta}_1 \) and row 1 in order to conclude that \( \hat{\theta}_2 \) must be equal to 1; then, we use these values of \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) and row 2 in order to conclude that \( \hat{\theta}_3 \) must be equal to 3; and so on.

\[
(138) \quad \begin{align*}
\hat{\theta}_1 &= 0 \\
\hat{\theta}_2 &= 1 \\
\hat{\theta}_3 &= 2\hat{\theta}_2 - \hat{\theta}_1 = 3 \\
\hat{\theta}_4 &= 2\hat{\theta}_3 + \hat{\theta}_2 = 6 \\
\hat{\theta}_5 &= 2\hat{\theta}_4 + \hat{\theta}_3 = 10
\end{align*}
\]

In the general case, there is a unique weight vector \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n) \) such that \( \hat{\theta}_1 = 0 \) and furthermore \( \langle \hat{\theta}, \overline{a} \rangle = 1 \) for every row \( \overline{a} \) of the derived Pater HG-comparative tableau, namely the weight vector defined by the recursion in (139a). It is trivial to prove by induction on \( k \) that the recursion (139a) can be made explicit as in (139b).

\[
(139) \quad \begin{align*}
a. \quad \hat{\theta}_1 &= 0 \\
\hat{\theta}_2 &= 1 \\
\hat{\theta}_k &= 1 + 2\hat{\theta}_{k-1} - \hat{\theta}_{k-2}, \quad \text{for } k = 3, \ldots, n \\
b. \quad \hat{\theta}_k &= \frac{k^2 - k}{2}, \quad \text{for } k = 3, \ldots, n
\end{align*}
\]

The squared inverse of the margin \( \mu \) can thus be bounded as in (140), since \( 1/\mu^2 \) coincides with the solution of the optimization problem (136), which is upper-bounded by the
HG HAS NO COMPUTATIONAL ADVANTAGES OVER OT

The squared norm of the weight vector defined in (139b). The leading term of the expression obtained in (140) is \( \sum_{k=1}^{n} k^4 \) and thus the claim follows from the well-known identity 
\[
\sum_{k=1}^{n} k^4 = \frac{1}{5} n^5 + \frac{1}{2} n^4 + \frac{1}{3} n^3 - \frac{1}{30} n.
\]

(140) \[ \frac{1}{\mu^2} \leq ||\hat{\theta}||^2 = \sum_{k=1}^{n} \left( k^2 - \frac{k}{2} \right)^2 \]

It is easy to verify that the replacement of the solution of the optimization problem (136) with the squared norm of the feasible weight vector (139) does not affect the bound asymptotically. \( \square \)

References