The Hegselmann-Krause Dynamics for the Continuous-Agent Model and a Regular Opinion Function do not always lead to Consensus

Edvin Wedin and Peter Hegarty

Abstract—We present an example of a regular opinion function which, as it evolves in accordance with the discrete-time Hegselmann-Krause bounded confidence dynamics, always retains opinions which are separated by more than two. This confirms a conjecture of Blondel, Hendrickx and Tsitsiklis.

Index Terms—Hegselmann-Krause model, continuous agent model, regular function.

I. INTRODUCTION

There is a rapidly expanding vista for the application of mathematics to multi-agent systems, with applications ranging from engineering to the life and social sciences. One major theme of this effort is emergence, the name given to the idea that patterns in the collective behaviour of large groups of interacting agents may be explicable even if each individual is assumed to obey only rules which are both simple and local, the latter meaning that each agent is only influenced by its close neighbours, in some appropriate metric.

The field of opinion dynamics is concerned with how human agents modify their opinions on social issues as a result of the influence of others. This paper is a contribution to the study of a particularly elegant and well-known mathematical model, the bounded confidence model of Hegselmann and Krause [7], or simply the HK-model for brevity. In the simplest formulation of the model, we have a finite number, say $N$, of agents, indexed by the integers $1, 2, \ldots, N$. The opinion of agent $i$ is represented by a real number $x(i)$, where the convention is that $x(i) \leq x(j)$ whenever $i \leq j$. The dynamics are as follows: There is a fixed parameter $r > 0$ such that, after each unit of time, every individual replaces their current opinion by the average of those which currently lie within distance $r$ of themselves. This is summarised by the formula

$$x_{t+1}(i) = \frac{1}{|N_t(i)|} \sum_{j \in N_t(i)} x_t(j)$$

(1.1)

where $N_t(i) = \{ j : |x_t(j) - x_t(i)| \leq r \}$. As the dynamics is obviously unaffected by rescaling all opinions and the confidence bound $r$ by a common factor, we can assume without loss of generality that $r = 1$.

Note that the HK-model seems to implicitly assume that each agent is aware of the opinions of all other agents, even if he chooses to ignore most of them when modifying his view. In one sense, this is a matter of interpretation. For example, a conservatively inclined Swedish citizen may switch the channel whenever a member of the Left party is giving an interview, or may keep watching but shake his head and mutter under his breath. In other words, the agent adopts strategies which both filter out unwelcome opinions and prevent him from being aware of them in the first place. On the other hand, the HK-model clearly assumes that an agent is aware of all opinions within his current confidence range. There are no restrictions imposed by, for example, geography, which prevent certain agents from sharing opinions a priori. In other words, agents do not follow local rules in the sense described above, though this is the case when the HK-model is reinterpreted in terms of multi-agent rendezvous [2].

Other important features of the model are that it is fully deterministic and that all agents act simultaneously. Hence the model differs in important respects from other famous models of opinion dynamics such as classical voter models [11] or the Deffuant-Weisbuch model [5].

The update rule (1.1) is certainly simple to formulate, though the simplicity is deceptive. Associated to a given configuration $(x(1), \ldots, x(N))$ of opinions is a receptivity graph $G$, whose nodes are the $N$ agents and where an edge is placed between agents $i$ and $j$ whenever $|x(i) - x(j)| \leq 1$. In this case, agents $i$ and $j$ are said to be neighbours, alternatively that they see or interact with one another. The transition in the configuration from time $t$ to time $t+1$ is determined by this graph at time $t$. However, it is clear from (1.1) that the dynamics will affect the graph, which in turn affects the dynamics. This feedback is the basic reason why many beautiful conjectures about the HK-model remain unresolved, as we shall now explain.

We begin with the necessary notation and terminology. The state space for a system of $N$ agents obeying the HK-dynamics is the set of non-decreasing functions $x : \{1, 2, \ldots, N\} \rightarrow \mathbb{R}$, equivalently, the set of vectors $(x(1), \ldots, x(N)) \in \mathbb{R}^N$ such that $x(i) \leq x(j)$ whenever $i \leq j$. An equilibrium state is one such that $|x(i) - x(j)| > 1$ whenever $x(i) \neq x(j)$. Clearly, once an equilibrium state is reached, then the opinion of every agent will be frozen for all future time. It is also easy to see that the converse holds: if $x_{t+1}(i) = x_t(i)$ for all $i$, then $x_t$ must be an equilibrium state. Any set of agents sharing a common opinion are referred to as a cluster. By a slight abuse of terminology, the term “cluster” may refer either to the set of agents with a certain opinion or the real number representing that opinion. Hence, a HK-system is in equilibrium if and only if no two clusters are within unit distance of each other. The simplest kind of equilibrium state is a consensus, in which there is only one cluster. Given a cluster $c \in \mathbb{R}$, its weight $w(c)$ is the number of agents sharing opinion $c$. A stable equilibrium is one in which, for any two clusters $a$ and $b$,

$$|b - a| \geq 1 + \frac{\min\{w(a), w(b)\}}{\max\{w(a), w(b)\}}$$

(1.2)

This last notion was introduced in [2], which is the paper that directly inspired the present work. The word “stability” refers to the fact that, if we extend the model to allow non-integer weights and add an agent of sufficiently small weight to an equilibrium configuration satisfying (1.2), then when the system is allowed to evolve again the
new equilibrium will not differ much from the old one, no matter the opinion of the perturbing agent - see [2] for precise statements and proofs.

The two fundamental facts about the HK-model are the following: (A) Any initial state will evolve to equilibrium within a finite time. (B) Even if the receptivity graph is initially connected, the subsequent equilibrium state need not be a consensus.

Fact (A) seems to have been rediscovered several times over and there are a number of different proofs in the literature. Indeed, the same fact has been proven for a wide class of models of which HK is just one particularly simple example, see [8]. Some of the known proofs of (A) give effective bounds for the time taken to reach equilibrium, as a function of the number \( N \) of agents only. The best-known bound is \( O(N^3) \), which was proven independently in [1] and [10]. It had been speculated that equilibrium is always reached within \( O(N) \) steps, and that the worst-case scenario is given by the initial state \( E_N = (1, 2, \ldots, N) \). This is false, however. Recent work of the authors [8], [12] shows that equilibrium state is almost surely stable under the weaker assumption that the initial opinions are chosen independently from any continuous field to determine the best-possible general upper bound.

Regarding (B), it is easy to see that consensus may not be achieved if the initial distribution of opinions is very uneven. For example, suppose we have 100 agents and the initial state is:

\[
x_0(i) = \begin{cases}
-1, & 1 \leq i \leq 98, \\
0, & i = 99, \\
1, & 100.
\end{cases}
\]

At \( t = 1 \), the opinion of agent 99 will be pulled very close to \(-1\), while agent 100 will only modify his opinion to \( x_1(100) = 1/2 \). Thus, agent 100 will now be isolated from everyone else and will form a cluster by himself in the equilibrium configuration. What is more interesting is that consensus may not emerge even when there is no such unevenness in the initial configuration. The simplest example is the initial state \( E_6 \). A direct computation shows that the resulting equilibrium consists of clusters at \( 4014 \) and \( 3728 \), each of weight three. At this point it seems natural to ask what a equilibrium, as a function of the number \( N \) of agents only. Indeed, the uncountably many agents are indexed by numbers in the closed interval \([0, 1]\) and the state space consists of non-decreasing, bounded Lebesgue measurable functions \( x : [0, 1] \to \mathbb{R} \). The analogue of (1.1) is

\[
x_{t+1}(\alpha) = \frac{1}{|\mathcal{N}_t(\alpha)|} \int_{\mathcal{N}_t(\alpha)} x_t(\beta) \, d\beta,
\]

where \( \mathcal{N}_t(\alpha) = \{ \beta : |x_t(\beta) - x_t(\alpha)| \leq 1 \} \) and \( |\cdot| \) denotes the length of an interval. Note that, if \( x_t \) is non-decreasing then \( \mathcal{N}_t(\alpha) \) is indeed always an interval, justifying this notation. It is also clear that if \( x_t \) is non-decreasing then so is \( x_{t+1} \), so our choice of state-space also makes sense. An equilibrium state in this setting is a function attaining only finitely many values, such that the difference between any two such values exceeds one whenever both are attained on sets of positive measure. Stable equilibrium can be defined as in (1.2), where now \( \omega(c) = |x^{-1}(c)| \), and the inequality is required to hold only when both clusters have positive weight. In particular, consensus means a constant function, whereas any equilibrium state which is not a consensus is represented by a discontinuous function. Note that, even if \( x_t \) is continuous then \( x_{t+1} \) may not be, if \( x_t \) is constant on an interval of positive measure. For example, if

\[
x_0(\alpha) = \begin{cases}
0, & 0 \leq \alpha \leq 1/2, \\
4\alpha - 2, & 1/2 < \alpha \leq 1,
\end{cases}
\]

then one may check that

\[
x_1(\alpha) = \begin{cases}
1/6, & 0 \leq \alpha \leq 1/2, \\
2(\alpha - 1/4)^2, & 1/2 < \alpha \leq 3/4, \\
2(\alpha - 4/3), & 3/4 < \alpha \leq 1.
\end{cases}
\]

There is now a discontinuity at \( \alpha = 3/4 \), since \( \lim_{\alpha \to 3/4^+} x_1(\alpha) = 1 > 1/2 = x_1(3/4) \). This is caused by the fact that, if \( \alpha \leq 3/4 \) then \( x_0(\alpha) \leq 1 \) and so \( \mathcal{N}_0(\alpha) \) reaches all the way down to zero, whereas if \( \alpha > 3/4 \) then \( \mathcal{N}_0(\alpha) \subset [1/2, 1] \).

It is reasonable to restrict attention to initial states \( x_0 \) which are injective. Indeed, \( x_0^{-1} \) should correspond to the cdf in Conjecture 1.2 above and Conjecture 1 of [3]. We shall assume henceforth that the initial state is regular, by which we mean that it is almost everywhere \( C^1 \), with strictly positive lower and upper bounds on its derivative where it exists. This is a slight strengthening of the notion of regularity as defined in [3].

In contrast to the discrete case, it is not clear which initial states will reach equilibrium in finite time. In [2], it is conjectured that a regular initial state \( x_0 \) will converge almost everywhere to a stable equilibrium, that is: there is a stable equilibrium \( x_\infty \) such that, for each \( \varepsilon > 0 \) there is a \( T_\varepsilon > 0 \) such that \( \mu(\{ \alpha : |x_\infty(\alpha) - x_\infty(\alpha)| > \varepsilon \}) < \varepsilon \) for all \( t > T_\varepsilon \), where \( \mu \) denotes Lebesgue measure. They prove a weaker statement in [3], but this fundamental conjecture about the continuous agent model remains open.

It is also proven in Lemma 4 of [2] (see also Proposition 3 of [3]) that if \( x_0 \) is regular and if

\[
x_t(1) - x_t(0) \geq 2 \quad \text{for all} \quad t
\]

then \( x_t \) will also be regular for all \( t \). Hence in such a situation \( x_t \) will not reach equilibrium in finite time, nor will it converge to a consensus. This brings us to perhaps the most curious aspect of the
Theorem 1.3. There exists a regular function $x_0 : [0, 1] \to \mathbb{R}$ such that, if the sequence $(x_t)_{t \in \mathbb{N}}$ is defined according to (1.4), then $x_t(1) - x_t(0) > 2$ for all $t$.

Section III contains a proof of this result and Section IV contains a discussion of some open problems.

II. PROOF OF MAIN THEOREM

The opinion function to be described below will converge pointwise to a non-regular stable state with 3 clusters of positive weight, and the construction can be extended to allow convergence to (at least) any odd number of such clusters.

Since scaling in the agent space $I$ does not affect the dynamics, we will loosen the definition of $I = [0, 1]$ and let $I$ be a longer interval. This is done to facilitate some computations at the end of the proof. To construct our initial state $x_0$, we first partition the set of agents into successive closed intervals $A$, $B$, $C$, $D$ and $E$, each intersecting the next in exactly one point. We choose a small positive $\varepsilon$ and let these intervals have the lengths $|A| = |E| = 1$, $|B| = |D| = \varepsilon^4$, and $|C| = \varepsilon^2$, so that the endpoints of the intervals lie symmetrically around the centre of $C$, which we denote by $c = 1 + \varepsilon^4 + \varepsilon^2$. The proof will go through for any sufficiently small $\varepsilon$, but we will fix $\varepsilon = \frac{1}{100}$ which will certainly be small enough. In an analogous manner, we partition the opinion space into closed intervals $A$, $B$, $C$, $D$, and $E$. We will consistently use Roman capitals for intervals in the agent space and script capitals for intervals in the opinion space. We take $|A| = |C| = |E| = \varepsilon$, and $|B| = |D| = d = \frac{3}{2}$. The choice of $d$ is somewhat arbitrary, but depends on the choice of $\varepsilon$ and must always lie in the open interval $[1, 2]$. To have some co-ordinates to work with, we place the origin at the lower endpoints of the intervals $A$ and $A$, and we thus have $I = [0, 2\varepsilon] = [0, 2 + \varepsilon^2 + 2\varepsilon^3] = [0, 2.000100002]$ and opinions ranging from 0 to $2d + 3\varepsilon$ is 3.03.

We now define $x_0$ to be linear on each of the subsets $A-E$, in such a way that $x_0(A) = A$, $x_0(B) = B$, $x_0(C) = C$, $x_0(D) = D$ and $x_0(E) = E$. This will force $x_0$ to stay within the “boxes” illustrated in Fig. 2. With this definition, the derivatives of $x_0$ on $A$ and $B$ will be $x_0(A) = 0 = \frac{\varepsilon^4}{\varepsilon^2}$, respectively, $x_0(E) = 0$, which will force $x_0$ to stay within the “boxes” illustrated in Fig. 3. The function $x_0$ is anti-symmetrical about $c$, in particular $x_0(A) = x_0(B)$ and $x_0(D) = x_0(B)$. It is clear that $x_0(x) = x_0(x)$ for all time steps $t$, and that the anti-symmetry remains.

We define $A_t = x_t^{-1}([0, 2\varepsilon])$ and $B_t = x_t^{-1}([2\varepsilon, \varepsilon + d])$ to be the sets of agents with opinions in $[0, 2\varepsilon]$ and $[2\varepsilon, \varepsilon + d]$, respectively, at time $t$, see Fig. 3. Note that $A$ is a proper subset of $A_0$. The reason for using $2\varepsilon$ instead of just $\varepsilon$ will become clear in the proof of Lemma 2.4 below. We also let $A_t = x_t(A)$. Again, Roman capitals denote sets of agents while script capitals are reserved for sets of opinions. We will let $A_t$ denote the average opinion on $A_t$ at time $t$.

We also define sequences $(s_t)_{t \geq 2}$ and $(s_t)_{t \geq 0}$: For $t = 0$ we will use the previously defined numbers $s_0 = \varepsilon$ and $s_0 = \frac{d}{\varepsilon^4}$, and then recursively set $s_{t+1} = \frac{2d}{s_t}$ and $s_{t+1} = \frac{2d}{s_t^2}$.

Using the above definitions, we can note that when $t = 0$ the following properties hold:

I: $A_t \subseteq A$. 
II: $B \supseteq B_t$. 
III: $x_t |_A \leq x_t |_E$. 
IV: $x_t |_B |_E \geq s_t$. 
V: $\varepsilon - A_t \geq 4\varepsilon^2$.

Indeed, II-IV hold at $t = 0$ for any $\varepsilon > 0$, and it is easy to check that, for $d = \frac{3}{2}$,

$$A_0 \leq \frac{\varepsilon}{2} + \varepsilon^6$$

and hence $\varepsilon = \frac{1}{100}$ is enough for V to hold as well.

We will prove by strong induction that properties II-V hold at all time steps. Note that Theorem 1.3 will follow immediately, since if property I holds for all $t \leq T$ then, in particular, $x_t(2\varepsilon) - x_t(0) \geq 2$ for all $t \leq T$ and hence, by Proposition 3 of [3], the functions $x_t$ remain regular up to time $T$.

Lemma 2.1. Assume properties II-V hold at all $t \leq T$. Then

i) $B_{t+1} \subseteq x_t^{-1}(A_{t+1} + 1)$ where $A_{t+1} = \{x_t + 1 : x_t \in A_t\}$. 
ii) $A_{t+1} \subseteq A_t$. 
iii) $A_{t+1} \supseteq A_T$. 
iv) $B_{t+1} \supseteq B_T$.

Proof: Let

$$\beta_T = x_T^{-1}(\min A_T + 1), \quad \gamma_T = x_T^{-1}(\max A_T + 1).$$

denote the two extreme agents in $x_T^{-1}(A_T + 1)$. We will show that

$$x_{T+1}(\beta_T) \leq \varepsilon, \quad x_{T+1}(\gamma_T) \geq d + \varepsilon$$

which together with monotonicity is easily checked to imply all four parts of the lemma.

By definition of $\gamma_T$, the leftmost agent he can see is the rightmost agent of $A$. Property II implies that $x_T(\gamma_T) \in [1, 1 + \varepsilon]$, and hence the rightmost agent he can see has an opinion between 2 and $2 + \varepsilon$ at time $T$. Because of property II and symmetry and since $d = \frac{3}{2}$ and $\varepsilon < \frac{1}{2}$, no agent in $C$ will have an opinion above 2, so $\gamma_T$ cannot see any agents in $C$. By property II and symmetry the agents in $E$ all have opinions larger than $2d + 2\varepsilon$ at time $T$, and hence $\gamma_T$ cannot see any agents in $E$. Thus $B \cup C \subseteq N_T(\gamma_T) \subseteq B \cup C \cup D$. The integral of the function $\tilde{x}_T(\alpha)$ is smaller than or equal to that of $x_T$ over all intervals containing $C$, since the average opinion on $C$ will always be $\frac{3}{2} + d$ by symmetry. We thus have that

$$x_{T+1}(\gamma_T) \geq \frac{1}{|N_T(\gamma_T)|} \int_{N_T(\gamma_T)} \tilde{x}_T(\alpha)d\alpha \geq \frac{|C|}{|B| + |C| + |D|} \geq \frac{|C|}{|C|} \geq 1$$

The reason for using $2\varepsilon$ instead of just $\varepsilon$ will become clear in the proof of Lemma 2.4 below. We also let $A_t = x_t(A)$. Again, Roman capitals denote sets of agents while script capitals are reserved for sets of opinions. We will let $\bar{A}_t$ denote the average opinion on $A_t$ at time $t$.

We also define sequences $(s_t)_{t \geq 2}$ and $(s_t)_{t \geq 0}$: For $t = 0$ we will use the previously defined numbers $s_0 = \varepsilon$ and $s_0 = \frac{d}{\varepsilon^4}$, and then recursively set $s_{t+1} = \frac{2d}{s_t}$ and $s_{t+1} = \frac{2d}{s_t^2}$.

Using the above definitions, we can note that when $t = 0$ the following properties hold:

I: $A_t \subseteq A$. 
II: $B \supseteq B_t$. 
III: $x_t |_A \leq x_t |_E$. 
IV: $x_t |_B |_E \geq s_t$. 
V: $\varepsilon - A_t \geq 4\varepsilon^2$.

Indeed, II-IV hold at $t = 0$ for any $\varepsilon > 0$, and it is easy to check that, for $d = \frac{3}{2}$,
Fig. 2. Some time invariant subspaces of $I$ and the opinion space for a piecewise linear initial opinion function along with some of the end points. Note that this figure is not to scale.

Fig. 3. Overview of the subsets of the agent space and the opinion space at time $t$. 
using \( \varepsilon = \frac{1}{100} \) and \( d = \frac{1}{2} \).

Next, consider the agent \( \beta_T \). By definition this agent can see every agent in \( A \). Property \([\ddagger]\) implies again that \( x_T(\beta_T) \in [1, 1 + \varepsilon] \), and hence the same argument as above implies that \( \beta_T \) can also see all agents in \( B \) and \( C \), but no agents in \( E \). The integral of the function

\[
\hat{x}_T(\alpha) = \begin{cases} \bar{A}_T, & \text{if } \alpha \in A_T, \\ 2\varepsilon + 2d, & \text{else} \end{cases}
\]

over \( \mathcal{N}_T(\beta_T) \) is thus greater than or equal to that of \( x_T \) over the same set. Since property \([\ddagger]\) implies \( A_T \supseteq A \) we can thereby use \( \hat{x}_T \) to get the following bound:

\[
x_{T+1}(\beta_T) \leq \frac{1}{|\mathcal{N}_T(\beta_T)|} \int_{A \cup \{\beta_T\} \cup D} \hat{x}_T(\alpha)d\alpha \leq \frac{|A_T|\bar{A}_T + (2\varepsilon + 2d)(B \cup C \cup D)}{|A_T|} \leq \bar{A}_T + (2\varepsilon + 2d)(\varepsilon^2 + 2\varepsilon^4) \leq \bar{A}_T + 4\varepsilon^2 \leq \varepsilon
\]

where the last inequality is true since we assume property \([\ddagger]\). We have now established the inequalities in (2.2) so the proof of the lemma is complete.

**Lemma 2.2.** Assume properties \([\ddagger]\) and \([\ddagger]\) hold at all \( t \leq T \). Then the increase in the mean opinion from \( A_T \) at time \( T \) to \( A_{T+1} \) at time \( T+1 \) is at most linear in \( |B_T| \). More precisely, \( A_{T+1} - A_T \leq 4|B_T| \).

**Proof:** Lemma 2.1 tells us that \( A_{T+1} \supseteq A_T \), and this allows us to write \( A_{T+1} = A_T \cup (A_{T+1} \setminus A_T) \), a disjoint union of two sets.

As for the first of these sets, recall that at time \( T \) agents in \( A_T \) have opinions in \([0, 2\varepsilon] \), and agents in \( B_T \) have opinions in \([2\varepsilon, \varepsilon + d] \). Hence with \( \varepsilon = \frac{1}{100} \) all agents in \( A_T \) can see one another, together with some agents in \( B_T \) whose opinions are all at most \( 1 + 2\varepsilon \). Thus the average of \( x_{T+1} \) over \( A_T \) at time \( T+1 \) will be

\[
\bar{A}_T \leq (A_T)_{T+1} \leq \frac{|A_T|\bar{A}_T + |B_T|(1 + 2\varepsilon)}{|A_T| + |B_T|} \leq \bar{A}_T + \frac{|B_T|(1 + 2\varepsilon)}{|A_T|} \leq \bar{A}_T + 2|B_T|,
\]

where the last inequality uses that \( A \subseteq A_T \), which follows from property \([\ddagger]\).

The average of \( x_{T+1} \) over \( A_{T+1} \setminus A_T \) at time \( T+1 \) is certainly at most \( 2\varepsilon \), by definition of the set \( A_{T+1} \). It also follows immediately from Lemma 2.1(iv) and monotonicity that \( A_{T+1} \setminus A_T \subseteq B_T \), and we thereby get the total average

\[
\bar{A}_{T+1} \leq (A_{T+1})_{T+1} + \frac{|B_T|(2\varepsilon)}{|A|} \leq \bar{A}_T + 2|B_T| + |B_T|(2\varepsilon) \leq \bar{A}_T + 4|B_T|.
\]

**Lemma 2.3.** Assume properties \([\ddagger]\) and \([\ddagger]\) hold at all \( t \leq T \). Then \( x_{T+1} \big|_A \leq c_{T+1} \).

**Lemma 2.4.** Assume properties \([\ddagger]\) and \([\ddagger]\) hold at all \( t \leq T \). Then \( x_{T+1} \big|_{B_{T+1}} \geq s_{T+1} \) and \( |B_{T+1}| \leq \frac{d}{\alpha^2} \).

In the proofs of Lemmas 2.3 and 2.4 the following additional lemma will be used:

**Lemma 2.5.** Let \( x_t \) be a regular opinion function on \( I \) such that \( x_t(2c) - x_t(0) > 2 \) and define the functions \( u_t, \; v_t, \; w_t : I \rightarrow I \) as follows:

\[
u_t(\alpha) = \begin{cases} 2c, & \text{if } x_t(\alpha) \geq x_t(2c) - 1 \\
(x_t(\alpha) + 1), & \text{otherwise} \end{cases} \\
v_t(\alpha) = v_t(\alpha) - u_t(\alpha)
\]

In words \( u_t(\alpha) \) and \( v_t(\alpha) \) are the leftmost and rightmost agents, respectively, that interact with agent \( \alpha \) at time \( t \), and \( w(\alpha) \) is the length of the set of neighbours of \( \alpha \) at time \( t \).

Then the updated function \( x_{t+1} \) is regular with derivative, where it exists, given by

\[
x_{t+1}(\alpha) = \frac{1}{w_t(\alpha)} \left[ u_t'(\alpha) \cdot (1 + x_t(\alpha) - x_t(\alpha)) + v_t'(\alpha) \cdot (1 + x_t(\alpha) - x_t(\alpha)) \right].
\]

**Proof:** That \( x_{t+1} \) is regular was proven in Proposition 3 of [3].

Assuming \( x_{t+1}, u_t, v_t \) and \( w_t \) are differentiable at \( \alpha \), we can compute as follows:

We first use the definition in (1.3) along with the product rule for derivatives to get

\[
x_t'(\alpha) = \frac{1}{w_t(\alpha)} \left[ u_t'(\alpha) \cdot (1 + x_t(\alpha) - x_t(\alpha)) + v_t'(\alpha) \cdot (1 + x_t(\alpha) - x_t(\alpha)) \right].
\]

The first term simplifies to

\[
x_t'(\alpha) = \frac{-w_t'(\alpha)}{w_t(\alpha)} \int_{u_t(\alpha)}^{v_t(\alpha)} x_t(\beta)d\beta + \frac{1}{w_t(\alpha)} \frac{d}{d\alpha} \left( \int_{u_t(\alpha)}^{v_t(\alpha)} x_t(\beta)d\beta \right).
\]

But by definition of the functions \( u_t \) and \( v_t \), we have

\[
x_t(u_t(\alpha)) = \begin{cases} x_t(\alpha) + 1, & \text{if } x_t(\alpha) \leq x_t(2c) - 1 \\
2c, & \text{otherwise} \end{cases}
\]

\[
x_t(v_t(\alpha)) = \begin{cases} x_t(\alpha) - 1, & \text{if } x_t(\alpha) \geq x_t(0) + 1 \\
0, & \text{otherwise} \end{cases}
\]

We would like to substitute the values \( x_t(v_t(\alpha)) = x_t(\alpha) + 1 \) and \( x_t(u_t(\alpha)) = x_t(\alpha) - 1 \) into (2.8). The former doesn’t hold when \( x_t(\alpha) > x_t(2c) - 1 \), but in this range \( v_t(\alpha) = 2c \) so \( v_t'(\alpha) = 0 \), so the substitution can be made in any case. A similar reasoning applies to the latter substitution. Hence the right-hand side of (2.8) simplifies to

\[
v_t'(\alpha) \cdot (x_t(\alpha) + 1) - u_t'(\alpha) \cdot (x_t(\alpha) - 1).
\]

Substituting (2.7) and (2.9) into (2.6) leads after a little computation to (2.5).

**Proof of Lemma 2.3** By property \([\ddagger]\) for agents \( \alpha \) in \( A \) we have that \( w_T(\alpha) = 0 \), so \( w_T(\alpha) = 0 \). Using Lemma 2.5 this gives

\[
x_{t+1}(\alpha) = 1 + x_T(\alpha) - x_{T+1}(\alpha) \cdot w_T(\alpha)
\]

for all \( \alpha \in A \). We also know from property \([\ddagger]\) that, at time \( T \), all agents in \( A \) can see each other, so \( w_T(\alpha) = 1 \).

To get a bound on \( v_T'(\alpha) \), we use the definition of \( v_t \), the chain rule, and the formula for the derivative of an inverse function:

\[
v_T'(\alpha) = \frac{d}{d\alpha} x_T^{-1}(x_T(\alpha) + 1) = \frac{1}{x_T'(x_T^{-1}(x_T(\alpha) + 1))} = \frac{x_T'(\alpha)}{x_T'(x_T(\alpha) + 1))}.
\]
To bound this, first note that $\alpha \in A$ implies $x'_t(\alpha) \leq e_T$ by property III. Second, note that since we assume $A_T \subseteq A$, it follows that $x_T(v_T(\alpha)) \in [1, 1 + \epsilon]$. In particular, $v_T(\alpha) \in B_T$, and thus $x'_T(v_T(\alpha)) \geq s_T$ by property [V]. Putting this together results in the bound

$$v'_T(\alpha) \leq e_T s_T. \quad (2.11)$$

Finally, we observe that $1 + x_T(\alpha) - xT+1(\alpha) \leq 2$ holds trivially, and we can now insert this and (2.11) into (2.10) to obtain

$$x'_{T+1}(\alpha) \leq 2 \frac{e_T s_T}{e_T} = e_{T+1}$$

as desired.

**Proof of Lemma 2.4.** First observe that since both the terms within brackets in (2.3) are positive, only one of them is needed to construct a lower bound for the derivative:

$$x'_{T+1}(\alpha) \geq \frac{1}{w_T(\alpha)} u'_T(\alpha)[1 + xT+1(\alpha) - x_T(\alpha)]. \quad (2.12)$$

We know from assuming property [I] and symmetry that no agent in $B_T$ can see as far as $E$, so $w_T(\alpha) \leq |A \cup B \cup C \cup D| \leq 2|A| = 2$. Lemma 2.1(i) and the definition of $B_T$ together assure us that $(1 + x_T(\alpha) - xT+1(\alpha)) \geq \epsilon$: All the agents in $A_{T+1}$ must have had opinions in $A_T+1$ at time $T$, according to Lemma 2.1(i). This is the motivation for using 2$e$ in the definitions of $A_T$ and $B_T$. It also lets us use $e_T$ and $s_T$ in a way similar to what was done in the proof of Lemma 2.3 to get that $u'_T(\alpha) = \frac{w_T(\alpha)}{w_T(\alpha)} \geq \frac{2e_T}{e_T}$. Applying these inequalities to (2.12) gives the result.

The upper bound on the size of $B_{T+1}$ simply comes from multiplying the inverse of the bound on the derivative with the height of $B_{T+1}$, which we know is constantly $d - e < d$ by construction.

**Proof of Theorem 1.3.** We would like properties I–V to hold for all time steps, for then we would be done.

By Lemmas 2.1, 2.3 and 2.4, if properties [IV] hold for all $t \leq T$, then property [V] will still hold at time $T+1$. To complete the induction it remains to show that [V] still holds at time $T+1$.

Lemma 2.3 allows us to bound each of the increments $A_{t+1} - A_t$, yielding

$$A_{T+1} = A_0 + \sum_{t=0}^{T} (A_{t+1} - A_t) \leq$$

$$\leq A_0 + 4T \left| B_t \right| \leq \tilde{A}_0 + 4e^4 + 4T \left( \frac{\bar{s} t}{s_t} \right), \quad (2.13)$$

where the last inequality follows from Lemma 2.3 plus the fact that $B_0 \subseteq B$ and $|B| = \epsilon^3$. We have $s_t = \frac{d}{4e^3 s_t}$ and it can easily be checked that the sequence $(s_t)$ is decreasing. Hence, for $t \geq 1$,

$$s_{t+1} = \frac{s_t}{2e^t} \geq \frac{s_1}{2e} \geq \frac{d}{4e^3 s_1}$$

and, hence, by iteration

$$s_t \geq (\frac{d}{4e^3})^{t-1} s_1$$

Substituting this into (2.13) and using (2.1) we get

$$A_{T+1} \leq \frac{\epsilon}{2} + \epsilon^6 + 4e^4 + 4d \sum_{t=1}^{T} \left( \frac{d}{4e^3} \right)^{t-1} \leq$$

$$\leq \frac{\epsilon}{2} + \epsilon^6 + 4e^4 + \frac{8e^4}{1 - \left( \frac{d}{4e^3} \right)} \leq \epsilon - 4e^2$$

where the last inequality holds for $\epsilon = \frac{1}{10}$ and $d = \frac{3}{2}$. This completes the proof.

### III. Discussion

When thinking about how to construct an example to prove Theorem 1.3 we first considered a "single-S" shape, without the narrow plateau in the middle, but with the height of the narrow strip connecting the two tails still being above two. We could not prove that the updates of such an initial state would also satisfy (1.4), though we suspect this is the case. In fact, what we think happens when the function is updated is that a narrow plateau will form in the middle, thus yielding the "double-S" shape of the function in Section 1 as an intermediate step in the evolution.

In any case, there should be even simpler examples of regular functions which satisfy (1.4). Indeed, Conjecture 1.1 suggests the following corresponding hypothesis for the continuous agent model:

**Conjecture 3.1.** Let $x_0 : [0, 1] \rightarrow \mathbb{R}$ be given by $x_0(\alpha) = L_0$. Then there is a critical value $L_0^*$ such that the updates $x_t$ satisfy (1.4) whenever $L > L_0^*$, whereas $x_0$ will evolve to consensus when $L < L_0$. Moreover, $L_0^* < L_0$, the critical value in Conjecture 1.1.

In fact, we also conjecture there will not be evolution to consensus at the critical value $L_0^*$. Intuitively, the reason for this is as follows. For as long as the updates $x_t$ are continuous and non-constant, the ranges $x_t(1) - x_t(0)$ will be strictly decreasing with $t$. The "2$e$ conjecture" suggests that, given a linear $x_0$, there will eventually be consensus if and only if the range of opinions shrinks to strictly below two at some point. Hence, at $L = L_0^*$, we should converge almost everywhere to an equilibrium consisting of two clusters of equal measure and separated by exactly two.

This leads in turn to another obvious remaining question, namely whether it is possible for a regular initial state to fail to satisfy (1.4) and yet never reach consensus.

### ACKNOWLEDGMENT

We thank the referees for their very careful reading of the original manuscript, which led to substantial improvements in the revised version.

### REFERENCES


