Frustration, glassy behavior and dynamical annealing in societies of Neural Networks.

Felippe Alves and Nestor Caticha

Instituto de Física,
Universidade de São Paulo,
CEP 05315-970, São Paulo, SP, Brazil

(Dated: August 17, 2021)
Abstract

We study maximum entropy mechanisms of information exchange between agents modeled by neural networks and the macroscopic states of a society of such agents in a few situations. Mathematical quantification of surprise, distrust of other agents and confidence about its opinion emerge as essential ingredients in the entropy based learning dynamics. Learning is shown to be driven by surprises, i.e. the receptor agent is confronted with the concurring opinion of a distrusted agent or with a trusted agent’s disagreeing opinion. Attribution of blame for the surprise derives from measures of distrust of the receiver towards the emitter agent and the receiver’s confidence about its own opinion. The dynamics proceeds by changes of mainly one or the other: the receptor opinion about the issue or the distrust about the emitter. A society with $N$ agents exchanging binary opinions about a set of issues show rich behavior which depend on the complexity of the agenda. For small sets the society reaches a steady state polarized into antagonistic factions, where balanced norms such as “the friend of an enemy is an enemy” are strictly satisfied. For larger sets of issues, societies can persist for a long time in spin-glass like states. There are two types of frustration: ideological and affective, with dynamical annealing properties depending on the complexity of the set of questions under discussion, leading to the lack of sharply defined parties for long transients.

I. INTRODUCTION

The analysis of opinion polarization in societies is central in several areas, see [1] for a review of the literature. Agent based modeling has has been used to build interacting systems that show collective properties of antagonistic bands. Some models [2] include the multivariate representation of the cultural state of an agent and the information exchanges with neighbors on a given geographical setting. Alternatively, agents can be Ising like variables, i.e. represented by the two states of “for” or “against” opinions on a given single issue [3] [4] which, with different combinations of ferromagnetic and anti-ferromagnetic interactions yields interesting collective behaviors. Even with these simplifications of the agents, several empirical successes can be obtained in complex situations. Agents may have continuous or vector opinions [5] or have simple Bayesian learning [6].

*flipgm@gmail.com
†ncaticha@usp.br
We follow a different modeling path, using agents that are simple neural networks (NN) with psychologically inspired traits and rich learning dynamics, as recommend by psychologists [7]. Our agents exchange information in a way that saturates bounds of efficiency in different student-teacher scenarios. Information theory and Statistical Mechanics guide the development of a framework, the entropic dynamics for NNs algorithm (EDNNA) [8], for general architectures. The resulting learning algorithms reveal aggregate variables identifiable with correlates in humans, such as cognitive dissonance, surprises, performance, the distrust of an individual and confidence on an opinion, that characterize the interactions of a pair of agents. Macroscopically, frustration, social balance, its emergence or breakdown and different types of affinities, e.g. affective or ideological [9] can drive polarization and lead to different types of spin-glass like states.

Adaptive changes through learning is the basic assumption of this work. Therefore, we choose the NN of simplest architecture, the single layer perceptron, as a model for each agent. See [10] for early work on interacting perceptrons. Linearly separable models have been shown to be useful in describing human performance in several cases [11]. The resulting societies will be shown to be already sufficiently complex that at least, this step is justified as necessary in the path towards richer agent models.

There are several learning algorithms for a perceptron, but we can imagine a process of evolution, where the selection pressure comes from the needs of better generalization ability. This has been studied in several papers analytically [12], [13], [14], [15], [16], [17], [8] and using evolutionary programming [18],[19]. Different simpler versions have already been applied to societies [20],[21], [22] and [23]. Data from decisions in 3 judges panels of US appellate courts has been studied [24] with these models. Other societies of NN have been studied by [25] and [26].

While selection of psychological traits, or analogously optimization of learning algorithms can be performed under certain conditions, such as in small groups, people and agents are called upon to engage in social interactions in groups far larger than those of the original environment. Optimality depends on the context and as the agents here studied, we have not evolved to deal within the modern contexts of large groups [27].

The performance of a NN learning a classification task from examples, will suffer from noise affecting the learning set. Introducing the possibility of noise among the assumptions used to optimize a learning rule enhances the machine’s performance. A measure of how
much a receiver should distrust the emitter’s signal is equivalent to a noise level estimate and  
entropic inference determines the rule for dynamical update of distrust. Earlier attempts  
to introduce adaptive noise estimation for a single agent with this architecture include [28],  
[29]. Communication can also be disrupted by parsing errors which would introduce noise  
to the mathematical representation of an assertion under discussion.

The society evolves under a discrete time dynamics. At a given time, two agents are se-  
lected to interact and randomly assigned the emitter and receiver roles. A question is chosen  
from the prominent questions of policy which occupy the public attention. The emitter sends  
its opinion. The receiver changes its state using the learning algorithm. There are several  
conceivable ways to choose the pair of agents. Borrowing language from biology, [23] use the  
term *allopatric* group formation when the probability of choosing an emitter depends on the  
receiver’s distrust. An effective communication barrier is created by repeated disagreement  
and this pair of agents cease to interact, grouping into factions holding opposing opinions.  
Alternatively, *sympatric* describes the process where agents keep interacting despite holding  
opposing views on the set of issues. We show that this dynamics of distrust permits the  
possibility of anti-learning, i.e learning the reverse of the emitter’s opinion.

In section II, we present the technical aspects of learning by entropic dynamics in general  
settings. This theory is applied in section II A to agents that have two sectors, for ideolog-  
cal/opinion and affective/distrust parameters, respectively, the weights of the NN and the  
noise level estimate of the channel. In section III we show that learning occurs with high  
intensity when there is a surprise: the receiver agrees with a distrusted emitter or disagrees  
with a trusted one. A change occurs in the receptor to decrease this dissonant interaction.  
The receiver’s changes resemble the learning algorithms for a tree committee machine [13]  
and a parity machine [30]. These Bayesian algorithms are poorly approximated by the least  
anaction introduced by [31] for the parity machine and [32] by hand. There is an interaction  
between the distrust and the weight sectors. The largest change occurs in the sector in  
which the agent is the least confident. Learning proceeds by predominantly changing one  
sector for the surprise such that the dissonance decreases.

In this large $N$ society model, section IV there is no absolute truth since there is no  
objective external fact checking mechanism. As Lincoln [33] already knew “The process is  
this: Three, four or half a dozen questions are prominent at a given time, the party selects  
its candidate, and he takes his position on each of these questions.” If only a few assertions
are under discussion i.e. the prominent issues, the system typically evolves into two separate parties. Membership can be attributed in two ways, by distrust or opinions. Distrust induced polarization is analogous to affective polarization discussed in political science and opinion polarization to ideological polarization on the basis of beliefs about policies. Scholarship \[34\] [35] about their theoretical characterization and experimental measurement is rapidly accumulating.

Frustration, \[36\] [37], also comes in two flavors. One similar to the structural balance theory \[38\] in social psychology is affective frustration. Agents $I, J$ and $M$ are in a balanced relation if the distrusts of $I$ towards $J$, from $J$ towards $M$ and from $M$ towards $I$ contains an even number of antagonistic relations. The unbalanced case of an odd number of antagonistic relations is frustrated and describes triplets of agents where for example “the friend of a friend is not a friend.” The second type, ideological frustration, describes unbalanced ideological alignment. As the number of issues under discussion becomes larger the society resembles more a spin-glass, where the number of frustrated triples is macroscopically large for persistently longer times. An interesting question in current affairs is to establish whether ideological polarization leads or follows affective polarization. We find that for agents with the full EDNNA discussing a simple agenda, affective polarization is established faster and drives ideological polarization. As the set of issues gets larger this is reversed and ideological polarization drives affective polarization. For a simpler learning algorithm, this reversal doesn’t occur.

II. ENTROPIC LEARNING DYNAMICS

In online learning, an example input-output pair $(x_t, \sigma_e)$ drives the dynamics one time step. In Bayesian learning \[14\], the likelihood $L$ is a probability that incorporates the information about the architecture of the networks and pushes the prior to a posterior distribution of the NN parameters. We choose to describe, at time $t$, the prior by a member of a MaxEnt parametric family of densities such that the knowledge about the parameters $u$ of the network, conditional on the data prior to $t + 1$, is represented by $Q_t$

$$Q_t(u|D_t) = \frac{1}{\zeta} e^{-\sum_a \lambda_{a,t} f_a(u)}, \quad (1)$$

5
where the set of generator functions \( \{ f_a(u) \} \) defines the manifold; \( D_t = \{ \lambda_a \} \) is the set of Lagrange multipliers that enforce a set of constraints \( F_a^t = E_Q(f_a) \) and \( \zeta \) ensures normalization. Bayes update is given by

\[
P_{t+1} := P(u|x_t, \sigma_e, D_t) = \frac{Q_t(u|D_t)L_t(\sigma_e|x_t, u, D_t)}{Z(\sigma_e|x_t, D_t)},
\]

where the evidence for the model is the probability \( Z_t := Z(\sigma_e|x_t, D_t) = \int L_t Q_t du \).

The likelihood doesn’t depend on \( D_t \), only on \( (\sigma_e, x_t) \) and the state of the network \( u \). In general the Bayes posterior will not be in the manifold, i.e. not conjugated. Nevertheless, the expected values of the Maximum Entropy generator functions of the family, under the Bayes posterior, point uniquely to a new distribution \( Q_{t+1} \) in the manifold. This MaxEnt posterior is the new representation of knowledge for the network and the prior for the next step, obtained by maximizing

\[
S[Q_{t+1}||Q_t] = -\int Q_{t+1} \log \frac{Q_{t+1}}{Q_t} du \nonumber \\
- \Delta \lambda_a \left( E_{Q_{t+1}}(f_a) - E_{P_{t+1}}(f_a) \right),
\]

subject to the constraints that its expected values \( E_{Q_{t+1}}(f_a) \) are equal to the Bayes posterior expected values \( E_{P_{t+1}}(f_a) \). The Lagrange multipliers are denoted by \( \Delta \lambda_a \) since they are the increments of the \( \{ \lambda_a \} \) of the prior \( Q_t \). It follows that

\[
F_a^{t+1} - F_a^t = -\frac{\partial \log Z_t}{\partial \lambda_a},
\]

which makes no mention of the Bayes posterior. These equations hold for any well behaved family. Since this learning dynamics is deduced from entropy maximization it is called Entropic dynamics. Learning occurs along the gradient of the log evidence. An element of the MaxEnt manifold can be represented equivalently by different sets of coordinates, the constraints \( F_a \) or the Lagrange multipliers \( \lambda_a \) or any Routhian mixture of the Legendre transforms. Convenience dictates which representation should be used. We are interested here in the case of linear and quadratic generators: \( f_0 = 1, f_i = u_i \) and \( f_{ij} = u_i u_j, \) the Gaussian family, which we write in the conventional way \( Q_{t+1} \propto \exp\left[ -\frac{1}{2}(u - \hat{u}_t)^T \Sigma_t^{-1} (u - \hat{u}_t) \right] \). The entropic dynamics update equations [4] driven by the arrival of the \( t^{th} \) example describe the changes in the parameters of \( Q \), its mean \( \hat{u}_t \) and covariance matrix \( \Sigma_t \):

\[
\hat{u}_{t+1} = \hat{u}_t + \Sigma_t \nabla_{\hat{u}_t} \log Z_t,
\]

\[
\Sigma_{t+1} = \Sigma_t + \Sigma_t (\nabla_{\hat{u}_t} \nabla_{\hat{u}_t}^T \log Z_t) \Sigma_t.
\]
These were first obtained by Opper [14] using a different argument, based on minimizing the expected value of a log loss function, which turns out to be equivalent to the (less intuitive) maximization of $S[P||Q_{t+1}]$, the entropy of the Bayes posterior relative to a “prior” $Q_{t+1}$, which is the entropic posterior.

A. Ideological and affective sectors

Equations 5 and 6 are general, they implement learning for any architecture and therefore to move on, we have to make some structural hypothesis. We split the degrees of freedom $u$ into two subspaces, the ideological sector $w \in \mathbb{IR}^K$ and the affective sector $z \in \mathbb{IR}$. Given the internal state of an agent, $w$, the forced decision of being for or against an issue $x$ is given by $\sigma = \text{sign}(w \cdot x)$.

The affective sector models the distrust that the receiver agent has for the emitter agent. It could be higher dimensional, but here we restrict to one dimension, because we model it as a noisy channel, which has been studied [28] [39] for perceptrons learning from examples in the student-teacher scenario. Information about the distrust $z$ of the receiver towards the emitter is encoded in a Gaussian with mean $\mu_{e|r}$ and variance $V_{e|r}$.

Although the receiver distrusts the emitter $\sigma_e$ is useful information about the putative “true label” $\sigma_T$. Another source of noise may act on the issues themselves, the issue $x_t$ considered by the receiver is not $y_t$, considered by the emitter, a “parsing error”. We take the components of $x_t$ and $y_t$ to be Gaussian correlated variables, $\mathbb{E}x_{t,i} = \mathbb{E}y_{t,i} = 0, \mathbb{E}x_{t,i}y_{t,i} = v_p$. To obtain the likelihood we marginalize the joint distribution of labels and the issue $y_t$:

$$L(\sigma_e|x_t, u) = \sum_{\sigma_T} \int P(\sigma_e|\sigma_T, y_t, x_t, u)P(\sigma_T|y_t, x_t, u)\frac{P(y_t|x_t, u)dy_t}{P(y_t|x_t, u)}.$$  

Furthermore, $\sigma_T = \text{sign}(y_t \cdot w_e)$ for an unknown $w_e$ but it is reasonable to suppose that $\sigma_e$ depends on $y_t$ and $w_e$ only through $\sigma_T$, so

$$P(\sigma_e|\sigma_T, y_t, x_t, u, w_e) = P(\sigma_e|\sigma_T, z)$$

$$= (1 - \varepsilon(z))\delta_{\sigma_e, \sigma_T} + \varepsilon(z)\delta_{\sigma_e, -\sigma_T},$$  

7
where \( \varepsilon(z) : \mathbb{R} \to [0, 1] \) is a function of \( z \). This random variable, limited to the unit interval, is the probability of the emitter conveying \( \sigma_e \neq \sigma_T \) to the receiver. While tempting, the choice of a Beta distribution for \( \varepsilon \) results in inconvenient technical difficulties. With \( \varepsilon(z) = \Phi(z) \), the cumulative distribution of a standard Gaussian and \( z \) a Gaussian with mean \( \mu_e | r \) and standard deviation \( V_e | r \), which are a representation of the distrust of the receiver and its uncertainty about it, with regard to the emitter agent, we maintain the Gaussian simplicity of equations \([5]\) and \([6]\).

Choosing a Gaussian \( G(y_t | x_t, v_p(w)) \) for \( P(y_t, | x_t, u) \) and

\[
P(\sigma_T | y_t, w, x_t, u) = P(\sigma_T | y_t, w) = \Theta(\sigma_T y_t \cdot w),
\]

where \( \Theta \) is the step function, the likelihood becomes

\[
L(\sigma_e | x, w) = \varepsilon(z) \Phi(-w \cdot x \sigma_e) + (1 - \varepsilon(z)) \Phi(w \cdot x \sigma_e)
\]

with the choice \( v_p = ||w||^{-2} \) accounting for the effect of inexperience on issue parsing.

These modeling steps allows the partition of the covariance \( \Sigma \) to remain block diagonal if it starts block diagonal, where the blocks are the ideological covariance \( C_t \) and the affective variance \( V_e | r \). The evidence

\[
Z(\sigma_e | x_t, D_t) = \mathbb{E}_{Q_t}(L(\sigma_e | u, x_t))
\]

\[
= \sum_{\sigma_T} \mathbb{E}_z(P(\sigma_e | \sigma_T, x_t, z))\mathbb{E}_w(\Theta(\sigma_T w \cdot x_t))
\]

with the expectation taken with respect to the two Gaussian sectors at time \( t \). At this point the information concerning the receiver about the past is \( D_t = \{ \hat{w}_t, C_t; \mu_e | r, V_e | r \} \). To reveal interesting symmetries it is useful to introduce the scaled stabilities

\[
h_w = \frac{\hat{w} \cdot x \sigma_e}{\gamma_C}, \quad h_\mu = \frac{\mu_e | r}{\gamma_V}
\]

where

\[
\gamma_C = \sqrt{1 + \hat{x} \cdot C_t \hat{x}}, \quad \gamma_V = \sqrt{1 + V_e | r}
\]
are the *internal response uncertainties*. The evidence depends on the agents state only through the \( h_w, h_\mu \) variables: \( Z = Z(h_w, h_m) \). The Gaussian integrals lead to

\[
Z = \Phi(h_w) + \Phi(h_\mu) - 2\Phi(h_w)\Phi(h_\mu)
\]

where \( \Phi(s) = \int_{-\infty}^{s} du g(u) \) is the cumulative of the standard Gaussian \( g(u) = (2\pi)^{-1/2}e^{-u^2/2} \).

The learning algorithm, obtained from 14, 5 and 6 is

\[
\hat{w}_{t+1} = \hat{w}_t + \frac{1}{\gamma_C} F_w C_t \sigma_e
\]

\[
C_{t+1} = C_t + \frac{1}{\gamma_C^2} F_C C_t \sigma_x C_t
\]

\[
\mu_{e|r}(t+1) = \mu_{e|r}(t) + \frac{1}{\gamma_V} F_m V_{e|r}(t)
\]

\[
V_{e|r}(t+1) = V_{e|r}(t) + \frac{1}{\gamma_V^2} F_V V_{e|r}(t)^2
\]

The first two describe the evolution of the ideological sector and the last two of the affective sector. The coupling between the two sets is due to the four \( F \) functions which we now describe. Note the appearance of the Hebbian-like term \( x\sigma_e \) in equation 15 which is the effective learning algorithm of the perceptron with a tensorial and adaptive learning rate modulated by \( F_w \). The rest of the dynamics is the adaptive estimation of the noisy channel, equations 17 and 18. \( F_w \) and \( F_\mu \) are the *Modulation* functions, since they set the scale of the changes of the parameters of the network for a given example. We also call \( F_C \) and \( F_V \) modulation functions, which set the scale of changes of the annealing schedule of the learning algorithm. These modulation functions \( F_w, F_C, F_m \) and \( F_V \) are given by

\[
F_w(h_w, h_\mu) = \frac{\partial \log Z}{\partial h_w} = (1 - 2\Phi(h_\mu)) \frac{g(h_w)}{Z}
\]

\[
F_C(h_w, h_\mu) = \frac{\partial^2 \log Z}{\partial h_w^2} = -F_w(F_w + h_w)
\]

\[
F_\mu(h_w, h_\mu) = \frac{\partial \log Z}{\partial h_\mu} = (1 - 2\Phi(h_w)) \frac{g(h_\mu)}{Z}
\]

\[
F_V(h_w, h_\mu) = \frac{\partial^2 \log Z}{\partial h_\mu^2} = -F_\mu(F_\mu + h_\mu)
\]

An important feature of equations 19 and 21 are the prefactors \( 1 - 2\Phi \) which, when negative, allow to learn the opposite of the arriving information and dynamically lead the society into a group of ferromagnetic or anti-ferromagnetic interactions. A similar reversal
FIG. 1. Modulation functions as functions of disagreement/agreement ($h_w$) and trust/distrust ($h_{\mu_e|r}$). They are significantly different from zero in the regions of surprises. Top row: $F_\mu$ for the distrust, left and $F_w$ for the weights, right. The floor shows a contour graph of the evidence $Z$. The dark regions are where surprises of agreeing with a distrusted or disagreeing with a trusted agent occur. Note the complementarity or blame attribution of the two modulation functions. Bottom: Modulation functions for the covariance modulation functions $F_V$ (left, distrust/affinity) and $F_C$ (right, opinion sector). Slightly negative values show the increase in certainty about the distrust and opinion sectors. Large positive values are triggered by surprises and show regions where uncertainty increases. Uncertainties grow along the diagonal $h_w = -h_{\mu}$.

of negative norm-support mechanism was studied in humans in [40]. The fields $h_w, h_{\mu}$ set the scale of the receiver’s opinion on the issue, and its distrust on the emitter, respectively.

From equations [19] to [22] it is easy to verify the symmetry between the two sectors, when
FIG. 2. A cut of the functions in figures 1 for constant $h_w$. Left: agents agree on an issue, Right: disagree. Top: $F_\mu$, continuous black line and $F_w$, dashed blue line. Notice the crossover at $h_\mu = h_w$, when the blame for the surprise changes sector. Bottom: $F_V$ shows the increase in uncertainty in the crossover region.

appropriately scaled

\[ F_w(h_w, h_\mu) = F_\mu(h_\mu, h_w), \]
\[ F_C(h_w, h_\mu) = F_V(h_\mu, h_w), \]

which have important consequences on the dynamics.

III. SURPRISE DRIVEN LEARNING

When two agents interact, the receiver undergoes meaningful changes only when there is something that could be interpreted as “cognitive dissonance” or surprise, see figure 1. The quadrants $h_w, h_\mu > 0$ are the regions where the modulation functions are significantly different from zero. Agreeing on an issue with a foe occurs for $h_w > 0$ and $h_\mu > 0$. When both are negative, the agent disagrees with a trusted agent. But the modulation functions are not uniform on the surprise region. If $|h_w| > |h_\mu|$, then the affective sector undergoes
a large change and the ideological sector remains rather unchanged and vice versa for the reverse inequality. The absolute values $|h_w|$ and $|h_\mu|$ are a measure of how sure the receiver is of its opinions about the issue ($|h_w|$) or about the emitter ($|h_\mu|$). The sector with the smaller confidence suffers a large change, that tends to eliminate the dissonance. $h_w$ and $h_\mu$ are fields re-scaled by measures of certainty $\gamma_C$ and $\gamma_V$. Since they are a measure of the width of the prior distribution $Q_t$, they represent a measure of how seriously an agent should take into account its estimates of weights and distrust and are obviously related to Bayesian credible intervals on the sectors. These are also dynamically updated in a complex but intuitive way, shown in figure 1. Figure 2 shows a cut of $F_\mu$ and $F_w$ (top) and $F_V$ (bottom) for constant $h_w$. Agents disagree $h_w = -6$ on the right, and agree $h_w = 6$ on the left. This case illustrates the following features of the modulation functions:

- Region of corroboration, $h_\mu < 0$: All the modulation functions are effectively zero since there is no surprise (receiver agrees with trusted agent).

- For $0 \lesssim h_\mu \lesssim 5.5 < h_w$ blame the surprise on the distrust toward the emitter. $F_w$ remains zero since the opinion sector is not blamed for the surprise. Here $F_V$ is negative since the receiver is effectively learning something about the emitter’s deserved distrust and thus slightly decreases its uncertainty $V$ ($F_V \lesssim 0$).

- Transition surprise: $h_\mu \sim h_w$. There is a surprise and no clear blame attribution since both sectors are similarly confident. $F_\mu$ grows quickly back to zero as $h_\mu$ goes across the transition region and there is a big increase in the uncertainty about the the distrust of the emitter.

- Surprise ($h_\mu > 6.5 > h_w$), the receiver blames the opinion sector $h_w$. Once $h_\mu$ is significantly larger than $h_w$, the blame for the surprise falls on the opinion sector and nothing happens on the affective sector. Note that $F_w$ is negative: the receiver learns the opposite of what the emitter is saying.

The agents interaction is driven by the evidence towards minimizing the surprises. The role of the response uncertainties $\gamma_C$ and $\gamma_V$ (equation 13) is to set the proper scale of the effective responses, according to how certain the agent is about the assigned value for that behavior. Higher values of $\gamma_C$ or $\gamma_V$ throw the agent’s effective response closer to zero, and therefore closer to the surprising zone, so agents who are very sure of their assessments about
either their opinion or the distrust they attribute to others are more stubborn. Certainty leads to immunization against surprises. The uncertainties also create an asymmetry on the modulation functions regarding novel versus corroborative information, as the less certain an agent is, the smaller is the difference between agreeing or disagreeing with a trusted emitter. For a given fixed distrust, as an agent becomes more convinced, i.e. smaller norm of the covariance of the Bayesian posterior, surprises become more effective for novel rather than corroborative information. A possible link between the width of the posterior, its role in weighting the importance of surprises and possible political orientations of agents in a small political agenda context is discussed in [21], [22].

Equipped with the full description of EDNNA for perceptron agents, we proceed with the study of societies of EDNNA agents learning through communication.

IV. A SOCIETY OF $N$ AGENTS

Consider a society of $N$ agents, exchanging $\pm 1$ opinions about $P$ issues $x \in \mathbb{R}^K$, independently chosen with $|x| = 1$. Different levels of frustration, polarization and dynamical annealing can result depending on the parameters and the initial conditions. In order to control the effect of the initial conditions we need to parametrize them in an appropriate manner.

We start with agents, with weight vectors $w_i \in \mathbb{R}^K$, chosen uniformly at random independently from anything else, on the sphere of radius $|w| = \sqrt{1+c}$. For each agent the initial co-variance matrix is the same $C = cI$. Each agent $i$ has an initial set of distrusts $\mu_j$ towards agents $j \neq i$ drawn from a normal distribution $\mu_j \sim G(\mu_0, \sqrt{1+V_0})$. These choices permit comparing the typical changes $\langle|\Delta w|\rangle$ e $\langle|\Delta \mu|\rangle$, as initially $\langle h_w \rangle \approx \langle h_\mu \rangle$. If all agents trust every other agent, $\mu_0 \ll 0$, the system typically evolves toward consensus, with the overlaps $\rho_{ij} = \frac{|w_i \cdot w_j|}{|w_i||w_j|} \lesssim 1$, so we will not deal with this trivial case.

To characterize the state of the ideological and affective sectors, [34] [35], we introduce order parameters that describe the balance of trust [41] [38] and balance of opinions in triads of agents. Receiver $i$, which distrusts emitter $j$ by a scaled distrust $h_\mu$, communicates with an information channel with flipping probability $\varepsilon_{ji} = \Phi(h_\mu)$. Define $v_{ji} = 1 - 2\varepsilon_{ji}$, which is positive/negative for trust/distrust. For any three agents, following [41], we say that the trust relation is balanced if $b_{ijk} := v_{ji}v_{kj}v_{ij} > 0$. This is as frustration in spin glasses but
FIG. 3. Evolution of the overlap and distrust matrices for a society with $N = 20$ agents with $K = 5$ ideological space dimensions, full covariance matrix with initial norms $C_0 = 1.0$ and $V_0 = 1.0$ for different number of issues. As the number of issues increase, the polarization goes from affective to ideological. $\varphi = N(N + K - 1) = 480$ is the number of (effectual) degrees of freedom.

for directed graphs since the odd permutation $b_{ikj}$ may be of opposite sign. The affective balance is the population average

$$B_A = \frac{1}{2N_T} \sum_{(ijk)} (b_{ijk} + b_{ikj}),$$

characterizes the state of affective frustration of a society. The sum runs over the set of $N_T = N$ choose 3 triplets. In addition to frustration associated to breakdown of affective balance, a second type of frustration can be defined. The opinion alignment is measured by the symmetric overlap $\rho_{ij}$. We can measure opinion balance by introducing for every triplet of agents the ideological balance and characterize the state by the average over the
FIG. 4. The mean value of the balance distribution describes the frustration based on ideology/opinion $B_I$ (left) and on affinity/distrust $B_A$ (right); for small (upper) or large (lower) dimension of issues compared to the size of the population, as a function of the number of issues under discussion, for different simulation times $\tau = \frac{t}{\varphi} = \frac{t}{N(N+K-1)}$. The agents evolve with a full covariance matrix $C$.

The population

$$B_I = \frac{1}{N_T} \sum_{\langle ijk \rangle} \rho_{ij} \rho_{jk} \rho_{ki}. \quad (26)$$

A ferromagnetic state or a staggered polarized society will have a $B_I$ and $B_A$ close to one; values for a highly frustrated society will be close to zero. These quantities may anneal to 1, but if the time scale is very large it points to a society that perseveres in a spin glass like state, without the need to have an infinite separation of time scales.

For agents using the tensorial learning adaptive EDNNA algorithm, with a full covariance $C$, a pattern emerges as the number $P$ of issues is varied, see figures 3, 4 and 5. For a few issues the society polarizes rapidly into two factions in the affective sector and then more slowly the affective polarization drives the ideological polarization. However as the agenda grows in complexity the ideological polarization sets in before and then drives the affinity.
FIG. 5. Time flow for $B_A$ versus $B_I$ for different sizes of the agenda. Same data as in figure 4. Full covariance case. For simple agendas affinity polarization sets in before ideological polarization. This is reversed when larger sets of issues are under discussion.

Affinity polarization. This is seen in the dark lines below the diagonal in figure 5. Times to anneal into the polarized state grow and the society lingers in a spin-glass state, with small $B_A$, thus large frustration. The time to achieve a balanced society increases as $P$ increases. Times $\tau$ measures the number of learning interactions per degrees of freedom of the system ($\varphi = N(N+K-1)$) so that we can compare simulations with different parameters. This reversal disappears when the agents use a simpler algorithm, with the covariance $C = cI$, see figures 6, 7. For these simpler agents, affinity polarization always seem to occur first and has slow behavior annealing form the glassy state to the ideological polarized state.
V. DISCUSSION AND CONCLUSIONS

Evolution under certain conditions leads to learning algorithms adapted to that environment. Our agents use algorithms designed to be efficient in a small group scenario, more specifically where there is a rule to be learned and followed by members of this group. This gives rise to the opinion sector, with a mechanism to emit opinions about issues. The interaction with other agents and the possibility of noisy communications or concealed cheating demands that the optimized algorithm incorporates a defense mechanism that appears in the assignment of a level of distrust to other members of the group. The entropic dynamics for NN architectures (EDNNA) analysis provides a general method to obtain such optimized learning algorithm. It could also result of an evolutionary process, similar to that shown in [18]. Once a two agent interaction is defined by the exchange and learning of issues with these rich algorithms, we construct a society. The reader should not be induced to think that we have obtained the Bayesian algorithm. There is no such thing. There are Bayesian algorithms conditioned on some informational context. Failure to understand this has lead
FIG. 7. Time flow of $B_A$ versus $B_I$ for different sizes of the agenda. Same data as in figure 6. Simple covariance case $C = cI$. For these simpler agents glassy behavior persists longer and affinity polarization sets in before ideological polarization.

To claims that humans’ lack of rationality can’t be modeled by Bayesian methods [12], but the problem in modeling humans is not that information theory does not apply but that the informational structure is poorly determined in the model. A simpler covariance (multiple of the identity) model, such as in figure 7 leads to algorithms just as Bayesian, but under a different set of constraints.

The society of $N$ agents has, as expected, a great variety of possible behaviors. The interesting result we report is that if the agenda is simple opposing parties can form around polarized affinities which proceed and push on a slower time scale to shared ideologies. However as it grows in complexity (larger $P$) the time to anneal down frustration grows and unbalanced societies have an effective spin-glass regime in finite times that might be
much larger than the life of an agent, and thus seems to evolve in the presence of quenched disorder. Now, the ideological sector anneals first, driving the slower affinity polarization. The reduction of times to polarize that accompanies the reduction of the agenda is in accordance with examples discussed in [43] where rapid partisan polarization is not accompanied by rapid changes in positions on economic policy issues. But this doesn’t happen for the simpler scale covariance agents. Affinity polarization sets in first and ideological polarization comes later at a much slower scale.

Agents with information processing capabilities permit a general framework useful in different situations. Questionnaires are a central method to gather information in the humanities and these NN agents can include such scenarios [21]. The different sectors are natural from an information processing formulation, lead to different frustrations and polarizations which are compatible with those observed in political science and permit sorting conditions when affective polarization leads ideological polarization [35]. We have used mistrust as a portmanteau for several associated concepts [44] [45], since their nuances are beyond our capacity. The NN we use are certainly simple and limited, despite the rich learning algorithms they use. The EDNNA formalism applies to any architecture, however it becomes intractable fast with architecture complexity. This of course is just an obvious reminder that care should be taken in building metaphors that map simple ideas in societies of NN to human societies.

**Acknowledgment:** We thank A. Caticha, O. Kinouchi, R. Vicente, M. Copelli, JP Neirotti for discussions. FA received financial support from a Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) PhD fellowship. This work received partial support from CNAIPS-NAP USP.

---

[1] L. Mastroeni, P. Vellucci, and M. Naldi, Agent-based models for opinion formation: A bibliographic survey, IEEE Access 7, 58836 (2019).

[2] R. Axelrod, The Dissemination of Culture: A Model with Local Convergence and Global Polarization, J. Conflict Resolut. 41, 203 (1997).

[3] S. Galam, Sociophysics: a review of galam models, International Journal of Modern Physics C 19, 409 (2008).
[4] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, Reviews of modern physics 81, 591 (2009).

[5] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Mixing beliefs among interacting agents, Adv. Complex Syst. 3, 87 (2000).

[6] A. C. R. Martins, Bayesian updating rules in continuous opinion dynamics models, Journal of Statistical Mechanics: Theory and Experiment 2009, P02017 (2009).

[7] W. Jager, Enhancing the realism of simulation (eros): On implementing and developing psychological theory in social simulation, Journal of Artificial Societies and Social Simulation 20, 14 (2017).

[8] N. Caticha, Entropic dynamics in neural networks, the renormalization group and the hamilton-jacobi-bellman equation, Entropy 22 (2020).

[9] J. Jost, C. Federico, and J. Napier, Political ideology: its structure, functions, and elective affinities., Annual review of psychology 60, 307 (2009).

[10] R. Metzler, W. Kinzel, and I. Kanter, Interacting neural networks, Physical Review E 62, 2555 (2000).

[11] R. A. Rescorla and A. R. Wagner, A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, in Classical Conditioning II: Current Research and Theory, edited by P. W. Black AH (New York: Appleton Century Crofts, 1972) pp. 64–99.

[12] O. Kinouchi and N. Caticha, Optimal generalization in perceptrons, Journal of Physics A 25, 6243 (1992).

[13] M. Copelli and N. Caticha, On-line learning in the committee machine, Journal of Physycs A 28, 1615 (1995).

[14] M. Opper, On-line versus off-line learning from random examples: General results, Physical Review Letters 77, 4671 (1996).

[15] N. Caticha and O. Kinouchi, Time ordering in the evolution of information processing and modulation systems, Philosophical Magazine B 77, 1565 (1998).

[16] M. Opper and O. Winther, A bayesian approach to on-line learning, in On-line learning in neural networks, Publications of the Newton Institute, edited by D. Saad (Cambridge University Press, United Kingdom, 1999) pp. 363–378, copyright of Cambridge University Press Available on Google Books.
[17] S. Solla and O. Winther, Optimal perceptron learning: an online bayesian approach, in *On-Line Learning in Neural Networks. Combridge* (University Press, 1998).

[18] J. P. Neirotti and N. Caticha, Dynamics of the evolution of learning algorithms by selection, Phys. Rev. E 67, 041912 (2003).

[19] N. Caticha and J. P. Neirotti, The evolution of learning systems: to bayes or not to be, AIP Conference Proceedings 872 (2006).

[20] R. Vicente, A. C. R. Martins, and N. Caticha, Opinion dynamics of learning agents: does seeking consensus lead to disagreement?, Journal of Statistical Mechanics: Theory and Experiment 2009, P03015 (2009).

[21] N. Caticha and R. Vicente, Agent-based social psychology: from neurocognitive processes to social data, Advances in Complex Systems 14, 711 (2011).

[22] N. Caticha, J. Cesar, and R. Vicente, For whom will the Bayesian agents vote?, Frontiers in Physics 3, 10.3389/fphy.2015.00025 (2015).

[23] F. Alves and N. Caticha, Sympatric multiculturalism in opinion models, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering 35*, Vol. 1757 (AIP Publishing LLC, 2016) p. 060005.

[24] N. Caticha and F. Alves, Trust, law and ideology in a neural network agent model of the US-Appellate Courts, ESANN 2019 - Proceedings 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2019).

[25] J. P. Neirotti, Anisotropic opinion dynamics, Physical Review E 94, 012309 (2016).

[26] J. P. Neirotti, Consensus formation times in anisotropic societies, Phys. Rev. E 95, 062305 (2017).

[27] R. Dunbar, Neocortex size as a constraint on group size in primates, Journal of Human Evolution 20, 469 (1992).

[28] M. Biehl, P. Riegler, and M. Stechert, Learning from noisy data: an exactly solvable model, Physical Review E 52, R4624 (1995).

[29] M. Copelli, R. Eichhorn, O. Kinouchi, M. Biehl, R. Simonetti, P. Riegler, and N. Caticha, Noise robustness in multilayer neural networks, Europhysics Letters 37, 427 (1997).

[30] R. Simonetti and N. Caticha, On-line learning in parity machines, J. Phys. A: Math. Gen. 29, 4859–4867 (1996).

[31] Y. Kabashima, Perfect loss of generalization due to noise in k=2 parity machines, J. Phys. A:
Math. Ge 27, 1917 (1994).

[32] G. J. Mitchison and R. M. Durbin, Bounds on the learning capacity of some multi-layer networks, Biol. Cybern 60, 345 (1989).

[33] A. Lincoln, Speech in U. S. House of Representatives on the Presidential Question, Collected works Vol 1. The Abraham Lincoln Association, Springfield, Illinois, edited by e. R. P. Basler (New Brunswick, N.J: Rutgers University Press, 1953).

[34] S. Klar, Y. Krupnikov, and J. B. Ryan, Affective Polarization or Partisan Disdain?: Untangling a Dislike for the Opposing Party from a Dislike of Partisanship, Public Opinion Quarterly 82, 379 (2018).

[35] S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood, The origins and consequences of affective polarization in the united states, Annual Review of Political Science 22, 129 (2019).

[36] P. W. Anderson, Amorphous Magnetism II, edited by R. Levy and R. Hasegawa (Plenum Press, New York, 1977) pp. 1 – 26.

[37] G. Toulouse, Theory of the frustration effect in spin glasses : I, Communications on Physics 2, 115 (1977).

[38] D. Cartwright and F. Harari, Structural balance: a generalization of heider’s theory, The Psychological Review 63 (1956).

[39] O. Kinouchi and N. Caticha, Lower bounds for generalization with drifting rules, Journal of Physics A 26, 6161 (1993).

[40] K. Keizer, S. Lindenberg, and L. Steg, The reversal effect of prohibition signs, Group Processes & Intergroup Relations 14, 681 (2011).

[41] F. Heider, The Psychology of Interpersonal Relations (J. Wiley and Sons, New York, 1958).

[42] E. Frederick and D. Danks, Confirmation in the cognitive sciences: The problematic case of bayesian models, Minds and Machines 21, 389 (2011).

[43] M. P. Fiorina and S. J. Abrams, Political polarization in the american public, Annual Review of Political Science 11, 563 (2008).

[44] T. Yamagishi, Trust: The Evolutionary Game of Mind and Society (Springer, 2011).

[45] E. Shockley, T. Neal, L. PytlíkZillig, and B. B. (Eds.), Interdisciplinary Perspectives on Trust Towards Theoretical and Methodological Integration (Springer, 2016).