Continuous time limits of the utterance selection model

Jérôme Michaud*

School of Physics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom

(Received 27 June 2016; revised manuscript received 18 November 2016; published 14 February 2017)

In this paper we derive alternative continuous time limits of the utterance selection model (USM) for language change [G. J. Baxter *et al.*, Phys. Rev. E **73**, 046118 (2006)]. This is motivated by the fact that the Fokker-Planck continuous time limit derived in the original version of the USM is only valid for a small range of parameters. We investigate the consequences of relaxing these constraints on parameters. Using the normal approximation of the multinomial approximation, we derive a continuous time limit of the USM in the form of a weak-noise stochastic differential equation. We argue that this weak noise, not captured by the Kramers-Moyal expansion, cannot be neglected. We then propose a coarse-graining procedure, which takes the form of a stochastic version of the heterogeneous mean field approximation. This approximation groups the behavior of nodes of the same degree, reducing the complexity of the problem. With the help of this approximation, we study in detail two simple families of networks: the regular networks and the star-shaped networks. The analysis reveals and quantifies a finite-size effect of the dynamics. If we increase the size of the network by keeping all the other parameters constant, we transition from a state where conventions emerge to a state where no convention emerges. Furthermore, we show that the degree of a node acts as a time scale. For heterogeneous networks such as star-shaped networks, the time scale difference can become very large, leading to a noisier behavior of highly connected nodes.

DOI: 10.1103/PhysRevE.95.022308

I. INTRODUCTION

In the study of complex systems, one important challenge is to deduce the macroscopic behavior of a system from the microscopic dynamics. This problem is at the center of statistical mechanics. In this paper we are interested in the (stochastic) agent-based class of complex systems. In (stochastic) agent-based models, agents interact following some rules (subject to noise) and we would like to characterize the average behavior of the complete population. One possibility to obtain a characterization of the average behavior is by obtaining a mean field approximation. What is meant by a mean field approximation varies between authors. The original idea is to characterize the dynamics of a complex system by choosing a representative agent and approximating the effect of the rest of the population as a mean field (see, for example, [1]). This approach is well adapted to well-mixed populations, but in the case of heterogeneous populations, for example, when the social structure is a complex network, this approach usually fails to describe the dynamics. To tackle this problem, the heterogeneous mean field (HMF) approximation has been proposed. In this approximation, the dynamics of agents in a network is approximated by taking one representative agent for each degree class. For more details on the HMF approximations and other approximations of the dynamics on a complex network, the reader is referred to [2]. For application of the HMF approximation for different models, the interested reader is referred to [3–6]. These two mean field approaches are based on the average influence of the different groups considered and provide a deterministic approximation of the dynamics. They share the property of averaging out the details of the underlying structure of the interactions.

In this paper we present an alternative to the usual mean field approaches by keeping some stochasticity in the HMF approximation. In the HMF approximation, one uses degree-block variables to estimate the dynamics. This is only one of many possible choices to introduce a heterogeneity in the mean field approach. Alternatively, one can group the agents by community or by any relevant criteria instead of by degree. An HMF approximation can then be obtained by using block variables, where the blocks depend on the grouping criteria. This procedure does not imply a deterministic approximation and some stochasticity can be conserved in the coarse-grained approximation; we will refer to this approximation as the stochastic HMF (SHMF) approximation.

As an example, we apply the SHMF approximation procedure to the problem of language evolution. Language is a defining property of humanity and is at the center of human interactions. The study of language dynamics is very important to better understand the formation of human cultures. In particular, the dynamics of language contacts and the formation of new dialects, pidgins, or creoles can shed light on the mechanisms underlying the formation and evolution of sociocultural groups (see, for example, [7]). Language is a complex adaptive system [8,9] and can be described at many different scales [10,11]. It seems that the different scales of language evolution should be accounted for in a better way than it has previously been done. In fact, at the interaction scale languages are highly variable, whereas at the population scale languages are relatively stable and change on a slow time scale. In order to better understand the link between these two time scales a coarsening procedure such as the SHMF approximation is needed.

Here we focus on the specific instance of the utterance selection model (USM) for language change [12] and derive an SHMF approximation of it. The USM is a stochastic agent-based model describing the evolution of a population interacting by stochastically producing utterances and learning

*v1jmicha@staffmail.ed.ac.uk

from them. Although there exists a wide range of models of language evolution (see [13]), we find the USM particularly appealing in that it can describe the process of language both at the time scale of individual interactions and at the time scale of the population. The USM has been applied to evaluate the theory of Trudgill of the emergence of New Zealand English [14]. Under appropriate assumptions, this model is analytically tractable and a wide range of results is available. The main results on the dynamics of the USM have been obtained in [12,15,16] and we review them below.

In [12], continuous time limits at the interaction level have been obtained using the Kramers-Moyal (KM) expansion and provide an analytical tool to study the marginal distribution of a representative agent in a population. However, in order to obtain this continuous time limit, one has to restrict the parameter space to simplify the mathematics. In order to fully characterize the behavior of the model, this restriction on parameters has to be overcome.

In [15], modifications of the USM are investigated in order to characterize under what circumstances language change trajectories follow a so-called S curve. Linguistic corpora studies [17,18] have shown that language change trajectories typically follow S curves.

Finally, in [16] the scaling law for the time needed to achieve consensus is obtained and numerically validated. This paper is one of the few considering parameter values outside the range in which the results of [12] are valid. In this paper we extend and improve previously known results by obtaining a continuous time limit of the USM at the interaction time scale, which does not suffer from any parameter restriction. We also obtain a coarse-grained SHMF approximation of the USM, clarifying the conditions under which a consensus can be achieved in this model.

The remainder of this paper is organized as follows. In Sec. II we discuss the coarse-graining problem and clarify our strategy to obtain a SHMF approximation of the USM. In Sec. III we recall the definition of the USM and some known results. In Sec. IV we derive a weak-noise stochastic differential equation (SDE) generalizing the continuous time limit obtained in [12] and compare the numerical efficiency of different numerical algorithms. This shows that for two agents, the system mainly behaves in a deterministic manner for short times, but the stochastic effects become relevant in long time scales. In Sec. V we derive the SHMF approximation of the USM and apply it to regular and star-shaped networks to validate it. This allows us to obtain a mean field characterization of the noise-driven phase transition separating the conditions under which a consensus can or cannot be formed at the population level. The analysis reveals a finite-size effect, justifying the fact that in a small population it is easier to create conventions. In Sec. VI we summarize the main results of this research and discuss possible future research directions. This paper is complemented by three appendixes. In Appendix A we rederive the continuous time limit obtained in [12] for completeness. In Appendix B the USM is linked with the Wright-Fisher (WF) process and technical details about the SDE are provided. Finally, in Appendix C details on numerical methods used to numerically integrate the SHMF approximation equations are provided.

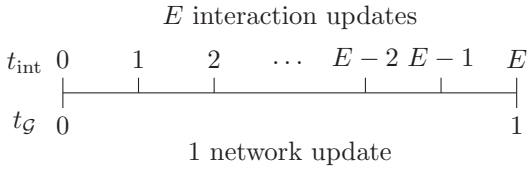

FIG. 1. Illustration of the two time scales of the problem. Here t_{int} represents the time of one interaction and $t_{\mathcal{G}}$ represents the time of one network update.

II. TIME SCALE SEPARATION AND COARSE-GRAINING PROBLEM

In this section we discuss the time scale separation problem inherent in every agent-based model and set out the approach taken in this paper. In order to simplify the discussion, we consider the case in which agents are associated with vertices $\mathcal{V}$ of a static network $\mathcal{G}$. Assuming pairwise interactions, such a system possesses two natural time scales: an interaction time scale t_{int} and a network time scale $t_{\mathcal{G}}$. Imagine that a clock, associated with t_{int}, ticks at every new interaction (assuming sequential updates) and that another clock, associated with $t_{\mathcal{G}}$, ticks when all the edges of the graph have been updated. Then, on average, the interaction clock ticks E times between two ticks of the network clock, where E is the number of edges of the graph. This situation is illustrated in Fig. 1.

If the number of edges E is large, the time scale separation between t_{int} and $t_{\mathcal{G}}$ increases. In fact, the relationship between these two time scales is

$$t_{\text{int}} \approx \frac{1}{E} t_{\mathcal{G}}. \tag{1}$$

In the limit when $E \to \infty$ the dynamics at the interaction level can be considered as continuous, since $t_{\text{int}} \to 0$. This stimulates the need to develop continuous time limits of the dynamics at the agent level in order to derive a population level approximation of the dynamics. If the network is finite, we expect some finite-size effects to occur, modifying the dynamics.

As we have mentioned, we aim to obtain a coarse-grained approximation of the dynamics of an agent-based model (ABM) and we intend for this approximation to be continuous in the network time $t_{\mathcal{G}}$ for large enough network. There are therefore two problems that need to be solved: the coarse-graining problem and the continuous time limit problem.

In Fig. 2 we provide an illustration of this problem. We want to derive a continuous in time ($t_{\mathcal{G}}$) population-based model (PBM) starting from a discrete in time ABM. One can first coarsen the problem and then obtain a continuous time limit or do the opposite. In this paper we will do both in a single step. As mentioned in Fig. 2, for the USM the only approximation that has been studied is a continuous in time approximation at the agent level in the form of a KM expansion leading to a Fokker-Planck (FP) equation (see [12] and Appendix A for details). This approximation suffers from parameter restrictions and cannot be easily coarse grained. It is not clear how one can in general approximate the other arrows for the USM. In this paper we provide both an alternative to the KM expansion, obtaining a continuous time limit at the agent level without parameters restrictions, and a methodology to

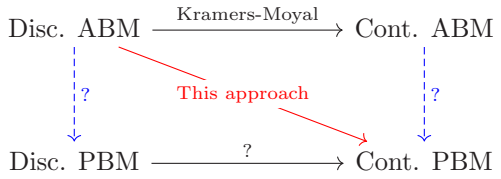

FIG. 2. Illustration of the coarsening problem by a continuous time approximation. The discrete models are on the left. Agent-based models are on top; their evolution depends on t_{int} and on the bottom the population models evolve according to $t_{\mathcal{G}}$. For the USM, it is known how to obtain a continuous time limit at the agent level using the Kramers-Moyal expansion. The other arrows are not clear. In this paper we will take the diagonal approach.

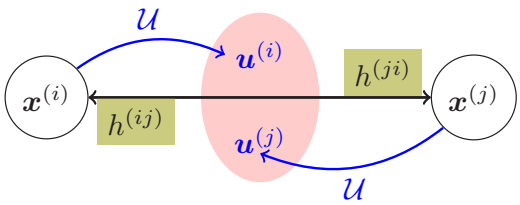

FIG. 3. Structure of the USM interaction. On the (ij) edge, the agents use their internal beliefs $\boldsymbol{x}^{(i)}$ to produce an utterance $\boldsymbol{u}^{(i)}$ through the process $\mathcal{U}$, which depends on the matrix M. The utterances are then used to update the internal beliefs depending on a weighting parameter $h^{(ij)}$.

derive a continuous in time population-based approximation in the form of a SHMF approximation.

III. UTTERANCE SELECTION MODEL

We now recall the definition of the USM. The USM [12] is a stochastic ABM of language evolution based on an evolutionary theory of language change due to Croft [19]. This model is not limited to the cultural evolution of languages but can be interpreted as a general model of cultural evolution.

In the USM, N agents are represented as nodes of a static network $\mathcal{G} = (\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ the set of edges along which the agents interact. We assume this network to be undirected and weighted by a probability distribution $G^{(ij)}$ representing the probability that agent $i \in \mathcal{V}$ interacts with agent $j \in \mathcal{V}$. In order to model the cultural evolution of a trait, the USM assumes that a particular trait can be instantiated in V equivalent variants. The state of an agent is characterized by a probability distribution $\boldsymbol{x}$ over the possible V variants of the cultural trait, which can be interpreted as the agent's belief of the frequency with which the variants should be used. In other words, $\boldsymbol{x}$ models the agent's *idiolect* and cannot be accessed by other agents. Since $\boldsymbol{x}$ is a discrete probability distribution it belongs to

$$\mathbb{P}_V := \left\{ \boldsymbol{x} \in [0,1]^V \,|\, \sum_{v=1}^{V} x_v = 1 \right\}. \tag{2}$$

In order to communicate, an agent produces an utterance $\boldsymbol{u} \in \mathbb{P}_V$ from a production process $\mathcal{U}$ ($\boldsymbol{u} := \mathcal{U}\boldsymbol{x}$), which takes the form of an empirical distribution of a biased sample of length L of the agent's belief distribution or idiolect. The length of the utterance L controls the amount of variability in the speech, since when L is large, the utterances are long and the induced noise small. The biasing process models production errors and/or innovation and is encoded through a stochastic matrix M. The updating (or learning) rule is formed by the weighted average of a process of self-monitoring S (weighted by $1 - h^{(ij)}$) and a process of accommodation A (weighted by $h^{(ij)}$). $h^{(ij)}$ is called the attention parameter. The process of self-monitoring aims at reducing the difference between $\boldsymbol{x}^{(i)}$ and $\boldsymbol{u}^{(i)}$ of an agent i and the accommodation process aims at reducing the difference between $\boldsymbol{x}^{(i)}$ and the utterance $\boldsymbol{u}^{(j)}$ of a neighboring agent j. The model is completed by a small parameter λ modeling the rate of learning.

An interaction time step of the USM can be divided into three substeps: social interaction, utterance production, and retention. A simulation run of the USM iterates such an interaction time step ET times, where E is the number of edges of the network and T is the final time of the simulation in $t_{\mathcal{G}}$ units. The three substeps of an interaction time step are defined below.

(i) Social interaction. The social interaction is simply modeled by choosing a pair of speakers i, j with the prescribed probability $G^{(ij)}$. In this paper we only consider the case where $G^{(ij)} = \frac{1}{E}$, that is, the uniform distribution. Furthermore, in order to be closer to the discussion about time scales of Sec. II, instead of randomly sampling the edges, we randomly order them and go through them in sequence in such a way that when a network update is complete, all the edges have been updated exactly once.

(ii) Utterance production. The production phase is illustrated in Fig. 3 by the $\mathcal{U}$ operator. It occurs at a specified time t_{int}. The two chosen agents generate an utterance $\boldsymbol{u}^{(i)}$.

The sampling process is done by using a multinomial sampling $\mathcal{M}$ and the biasing process is done through the introduction of a mutation matrix M, which is column stochastic. Note that the ordering of the sampling and the biasing processes matters. We therefore have the two possible definitions of the utterance empirical frequency vector $\boldsymbol{u}^{(i)}$:

$$\boldsymbol{u}_{bs} \sim \frac{1}{L}\mathcal{M}(L, M\boldsymbol{x}), \tag{3a}$$

$$\boldsymbol{u}_{sb} \sim \frac{1}{L}M\mathcal{M}(L, \boldsymbol{x}). \tag{3b}$$

In [12], the rule (3a) was chosen to model the utterance process. We argue in Appendix B that the other choice (3b) is more natural and leads to a well-posed SDE, whereas the choice (3b) leads to an ill-posed SDE. In [12], the differences between these two choices are lost during the derivation of the continuous time limit. If the specific rule is not specified, we use the notation $\mathcal{U}\boldsymbol{x} = \boldsymbol{u}$ without a subscript.

The different utterances produced during a communication event form an utterance pool on which the retention phase is based.

(iii) Retention. The retention rule, or updating rule, is a rule to compute $\boldsymbol{x}^{(i)}(t+1)$, where t is measured in t_{int} units. This is the short-time-scale updating rule. An agent i then revises

state $\boldsymbol{x}^{(i)}$ using

$$\begin{aligned}\delta\boldsymbol{x}^{(i)}(t) = \lambda[(1-h^{(ij)})S(\boldsymbol{x}^{(i)}(t),\boldsymbol{u}^{(i)}(t)) \\ + h^{(ij)}A(\boldsymbol{x}^{(i)}(t),\boldsymbol{u}^{(j)}(t))],\end{aligned} \tag{4}$$

where $\delta\boldsymbol{x}^{(i)}(t) = \boldsymbol{x}^{(i)}(t+1) - \boldsymbol{x}^{(i)}(t)$. In this description, the utterance vectors $\boldsymbol{u}$ are stochastic vectors. We define the self-monitoring process S and the accommodation process A as

$$\begin{aligned}S(\boldsymbol{x}^{(i)}(t),\boldsymbol{u}^{(i)}(t)) &:= \boldsymbol{u}^{(i)}(t) - \boldsymbol{x}^{(i)}(t), \\ A(\boldsymbol{x}^{(i)}(t),\boldsymbol{u}^{(j)}(t)) &:= \boldsymbol{u}^{(j)}(t) - \boldsymbol{x}^{(i)}(t).\end{aligned} \tag{5}$$

These two processes are driven by probability matching, since they compare the empirical frequencies $\boldsymbol{u}$ with their belief probability distribution $\boldsymbol{x}$ until they match. The term *probability matching* is widely used by evolutionary linguists to express situations in which speakers adapt their speech distribution to the speech distribution they hear. If the probability distributions are equal, then those terms vanish. Some variants of the USM use a different definition for the self-monitoring and accommodation processes. In [15], the influence of misperception is investigated. This is outside the scope of this paper, but extending our results to these more general cases is in principle possible. In this paper we restrict the discussion to the original choice of probability matching.

The relative weight of these two functions is given by a parameter $h^{(ij)} \in [0,1]$ and $\lambda > 0$ is a usually small positive parameter. For simplicity, we assume that $h^{(ij)} = h$, that is, the attention parameter does not depend on the identity of agents.

The complete mathematical definition of the discrete USM is then given by Eqs. (3)–(5). The USM contains two sources of randomness: The first is contained in the distribution $G^{(ij)}$, which controls the way in which the edges are updated; the second is contained in the utterance process $\mathcal{U}$, which controls the noisy interaction between agents. In order to characterize the model, one is interested in the statistical behavior, which can be studied through approximations. Continuous time limits deal with the noisy utterance process, while coarse-graining approximations deal with the social noise.

In [12], a FP equation has been obtained as an agent-level continuous time limit of the USM using the KM expansion. In Appendix A we recall this procedure and show that the required scaling assumptions [Eq. (A7)] significantly restrict the application of this approximation. In the next section, we derive an alternative continuous time limit of the USM based on a normal approximation of the multinomial distribution (diffusion approximation), which does not suffer from any parameter restriction and generalizes the result obtained with the KM expansion.

IV. STOCHASTIC DIFFERENTIAL EQUATION CONTINUOUS TIME LIMITS

In this section we develop the first main contribution of this paper, that is, we derive a continuous time limit of the USM that captures the dynamics of the USM over the full range of parameters. The limit is derived at the interaction time scale t_{int} and generalizes the FP equation obtained by the KM expansion.

In the rest of this section, we first obtain approximations of the utterance production process. We then derive the weak-noise SDE continuous time limit of the USM. Finally, we test the different approximations against the discrete USM and against the deterministic limit obtained by the KM expansion with scaling $\lambda \propto \delta t$ on a very simple network and argue that the weak noise should not be neglected.

A. Approximations of the multinomial distribution

The USM utterance production mechanism given in Eq. (3) relies on a multinomial sampling and a biasing procedure. In order to obtain a weak-noise SDE continuous time limit of the USM, we need (i) to approximate the sampling process in a continuous-in-L manner and (ii) to decouple the parameters and the source of noise to relate the noise to a Wiener process. To do so, assume that we want to approximate a random vector

$$\boldsymbol{z} \sim \frac{1}{L}\mathcal{M}(L,\boldsymbol{y}), \tag{6}$$

where L is an integer and $\vec{y} \in \mathbb{P}_V$, by a vector $\boldsymbol{w}$. First note that the expectation value and covariance matrix of $\boldsymbol{z}$ are given by

$$\mathbb{E}(\boldsymbol{z}) = \boldsymbol{y}, \tag{7a}$$

$$\text{cov}(\boldsymbol{z},\boldsymbol{z}) = \frac{1}{L}[\text{diag}(\boldsymbol{y}) - \boldsymbol{y}\boldsymbol{y}^T]. \tag{7b}$$

A possible continuous-in-L analog to the multinomial distribution is given by the Dirichlet distribution $\mathcal{D}$ of parameter $L\boldsymbol{y}$ and we can approximate $\boldsymbol{z}$ by

$$\boldsymbol{w}^{\mathcal{D}}(\boldsymbol{y}) \sim \mathcal{D}(L\boldsymbol{y}). \tag{8}$$

This approximation is continuous in L but does not decouple the parameter and the noise source. The good property of this approximation is that $\boldsymbol{w}$ is a discrete probability distribution.

In order to decouple the source of noise from the parameter $\boldsymbol{y}$, one can use the normal approximation of the multinomial distribution. This leads to an approximation

$$\begin{aligned}\boldsymbol{w}^{\mathcal{N}}(\boldsymbol{y}) &\sim \mathbb{E}(\boldsymbol{z}) + [\text{cov}(\boldsymbol{z},\boldsymbol{z})]^{1/2}\mathcal{N}(0,\boldsymbol{I}) \\ &\sim \boldsymbol{y} + \frac{1}{\sqrt{L}}D(\boldsymbol{y})\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),\end{aligned} \tag{9}$$

where the square root has to be taken in the Cholesky sense and the matrix $D(\boldsymbol{y})$ is the square root in the Cholesky sense of $\text{diag}(\boldsymbol{y}) - \boldsymbol{y}\boldsymbol{y}^T$. A definition of a square root in the Cholesky sense is given in Definition 1 in Appendix B and the possible forms of the matrix $D(\boldsymbol{y})$ are given in Appendix B.

The normal approximation given by Eq. (9) both is continuous in L and decouples the source of noise and the parameter $\boldsymbol{y}$. This permits a connection with Wiener processes as will be shown below.

The drawback of the normal approximation is that $\boldsymbol{w}^{\mathcal{N}}$ is not a discrete probability distribution vector in general. This is a consequence of the fact that the normal approximation is unbounded, whereas the multinomial and the Dirichlet distribution are bounded. We also have to note that this approximation is only valid if L is sufficiently large and $\boldsymbol{y}$ is not close to the boundaries of the domain. These assumptions are not always satisfied, but we will assume them anyway.

With these limitations in mind, one can now provide a continuous approximation of the utterance production process and introduce the biasing process through the matrix M. For the Dirichlet approximation and for the normal approximation, one can approximate the continuous-in-L utterance vector as

$$\boldsymbol{u}_{bs}^{\mathcal{T}} = \boldsymbol{w}^{\mathcal{T}}(M\boldsymbol{x}), \quad \mathcal{T} \in \{\mathcal{D},\mathcal{N}\} \tag{10a}$$

$$\boldsymbol{u}_{sb}^{\mathcal{T}} = M\boldsymbol{w}^{\mathcal{T}}(\boldsymbol{x}), \quad \mathcal{T} \in \{\mathcal{D},\mathcal{N}\}, \tag{10b}$$

where $\boldsymbol{x}$ is the belief distribution state vector. The index bs stands for first biasing and then sampling and the index sb for the reverse ordering.

We will show in Sec. IV B that under the normal approximation the order of application of the sampling and the biasing processes matters. This is connected to the fact that the continuous-in-L utterance vector obtained under this approximation does not always represent a discrete probability distribution. For the normal approximation, the continuous-in-L utterance vectors are given by

$$\boldsymbol{u}_{bs}^{\mathcal{N}} = M\boldsymbol{x} + \frac{1}{\sqrt{L}} D(M\boldsymbol{x})\boldsymbol{\xi}, \tag{11a}$$

$$\boldsymbol{u}_{sb}^{\mathcal{N}} = M\boldsymbol{x} + \frac{1}{\sqrt{L}} MD(\boldsymbol{x})\boldsymbol{\xi}, \tag{11b}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\boldsymbol{I})$.

Remark 1. It is mentioned in Appendix A that one usually assumes that the off-diagonal terms of M are small [of $O((\delta t)^{1/2})$ or smaller]. In that case, one can show in general that

$$D(M\boldsymbol{x}) = D(\boldsymbol{x}) + O(\|M - I\|_\infty), \tag{12a}$$

$$MD(\boldsymbol{x}) = D(\boldsymbol{x}) + O(\|M - I\|_\infty). \tag{12b}$$

As a consequence, in the derivation of a continuous time equation, the influence of the matrix M in the noise term can be neglected and the ordering between sampling and biasing no longer matters. Note that using the Dirichlet approximation produces a vector $\boldsymbol{u}^{\mathcal{D}}$ representing a discrete probability distribution under both orderings. This is also true for the discrete USM.

B. Weak-noise SDE limit

We have now collected all the partial results needed to derive continuous time limits of the USM and in particular a weak-noise SDE continuous time limit based upon the normal approximation. The derivation of the continuous time limits is now fairly straightforward; all that needs to be done is to put the continuous-in-L approximations of the utterance vector into Eq. (4), average over the possible interactions of an agent (average over its neighbors), and scale $\lambda = \delta t$ to obtain a continuous time limit.

For the Dirichlet approximation such an approximation is obtained by introducing the random vector $\boldsymbol{u}^{\mathcal{D}}$ defined in Eq. (10a), with either the biasing-sampling order or the reverse order, into (4), summing the contribution of all the neighbors of an agent i, and introducing the scaling $\lambda = dt$. We obtain

$$\dot{\boldsymbol{x}}^{(i)} = \sum_{j\neq i} G^{(ij)}[(1-h)(\boldsymbol{u}^{\mathcal{D},(i)} - \boldsymbol{x}^{(i)}) + h(\boldsymbol{u}^{\mathcal{D},(j)} - \boldsymbol{x}^{(i)})]. \tag{13}$$

This is the first continuous time equation we consider. This is an SDE in the sense that the vectors $\boldsymbol{u}^{\mathcal{D},(i)}$ and $\boldsymbol{u}^{\mathcal{D},(j)}$ are stochastic vectors, but it is not a usual SDE, since the noise is not related to a Wiener process and cannot be analyzed in the framework of SDEs. We use this formulation in the numerical experiments as an accurate continuous time limit of the USM, since the random vector produced always represents a discrete probability distribution.

The derivation of the continuous time limit based on the normal approximation is obtained in a way similar to the Dirichlet approximation. We introduce the normal approximation (11) into Eq. (4), sum the contribution of all the neighbors of an agent i, and introduce the scaling $\lambda = dt$. Letting $dt \to 0$, we obtain the following two equations depending on the ordering choice between the biasing and the sampling processes:

$$d\boldsymbol{x}^{(i)} = \sum_{j\neq i} G^{(ij)}\Bigg[[(1-h)(M-I)\boldsymbol{x}^{(i)} + h(M\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)})]dt + \left(\frac{1-h}{\sqrt{L}} D(M\boldsymbol{x}^{(i)})d\boldsymbol{\xi}_t^{(i)} + \frac{h}{\sqrt{L}} D(M\boldsymbol{x}^{(j)})d\boldsymbol{\xi}_t^{(j)}\right)\Bigg] \tag{14a}$$

or

$$d\boldsymbol{x}^{(i)} = \sum_{j\neq i} G^{(ij)}\Bigg[[(1-h)(M-I)\boldsymbol{x}^{(i)} + h(M\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)})]dt + \left(\frac{1-h}{\sqrt{L}} MD(\boldsymbol{x}^{(i)})d\boldsymbol{\xi}_t^{(i)} + \frac{h}{\sqrt{L}} MD(\boldsymbol{x}^{(j)})d\boldsymbol{\xi}_t^{(j)}\right)\Bigg], \tag{14b}$$

where $d\boldsymbol{\xi}_t = \sqrt{dt}\, d\boldsymbol{W}_t \sim \mathcal{N}(\mathbf{0},dt^2\boldsymbol{I})$. This noise is weaker than a usual Gaussian noise $d\boldsymbol{W}_t$ by a factor $\sqrt{dt}$. We call this limit a weak-noise SDE. This approximation is a diffusion approximation taking into account all the sources of noise in an interaction, that is, the noise originating from the two utterances produced. This is different from the continuous time limit obtained in [12], where only the noise of the speaker is taken into account. Note that a deterministic limit is obtained by neglecting the noise terms in (14a) in which case the solution of the KM expansion when $\lambda = \delta t$ is recovered. This approximation therefore generalizes the KM expansion.

The coefficient of the noise scales as $\sqrt{dt}$ and vanishes in the continuous time limit in agreement with the FP derivation. We argue that the noise term of Eq. (14) should not be neglected for two reasons. First, since the white noise $d\boldsymbol{W}_t$ scales as $\sqrt{dt}$, the noise term scales as dt and can be argued to be of the same order of magnitude as the drift term. Second, the drift term has the property of becoming very small for long times, because $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ converge towards each other and they jointly converge toward a vector $\boldsymbol{b}$ such that $(M-I)\boldsymbol{b} = 0$. As a result, even a weak noise becomes important in the long time as soon as the drift term becomes of order $\sqrt{dt}$. Therefore,

we expect the noise to be important on a long time scale, but negligible on a short time scale. This will be verified with numerical simulations.

We argue that Eq. (14a) is ill posed and that Eq. (14b) is well posed. We recall a problem is said to be well posed if it has a unique solution and if small changes in initial conditions leads to small changes of the solution (stability). From Eqs. (14a) and (14b) it is not straightforward to decide whether or not they are well posed. A detailed discussion of a special case of these equations is treated in Appendix B and explains the origin of the ill-posed nature of Eq. (14a). In the following we will work with the continuous time limit (14b).

If we consider the scaling (A7) instead of scaling only λ, the corresponding SDE reads

$$d\boldsymbol{x}^{(i)} = \sum_{j\neq i} G^{(ij)}\Big[[(\bar{M}-I)\boldsymbol{x}^{(i)} + \bar{h}(\boldsymbol{x}^{(j)}-\boldsymbol{x}^{(i)})]dt + \frac{1}{\sqrt{L}} D(\boldsymbol{x}^{(i)}) d\boldsymbol{W}_t^{(i)}\Big], \tag{15}$$

where $\boldsymbol{W}_t^{(i)}$ is a standard vectorial Wiener process and $d\boldsymbol{W}_t^{(i)}$ is a white noise. Equation (15) is the stochastic counterpart of the FP equation obtained in [12] using the Itô convention. The deterministic limit corresponds to scaling only λ, which is consistent with the FP equation derived by the KM expansion in [12]. In this limit, the noise of agent j becomes irrelevant and can be neglected.

Remark 2. Equation (15) is the same for the two possible orderings of the production, due to Remark 1. In other words, with the scaling used in [12], the two orderings become equivalent.

C. Numerical experiments

We now perform some numerical experiments to validate the continuous time limits derived above. We consider a network of two connected agents 1 and 2 for simplicity. The probability $G^{(12)}$ that they interact is 1. This is the smallest network where interaction is possible. We also consider for simplicity the case of two variants $V = 2$. We compare the weak-noise SDE (14b) with the deterministic limit given by Eq. (14b) in which the noise is neglected. To obtain better insight into the dynamics, we consider the difference between the idiolects of the two agents, that is, we consider the variable $\boldsymbol{z} := \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}$. The deterministic evolution of $\boldsymbol{z}$ is given by

$$\dot{\boldsymbol{z}} = [(1-2h)M - I]\boldsymbol{z} := A\boldsymbol{z}. \tag{16}$$

We choose the formulation of Eq. (14b), since the other ordering of sampling and biasing is shown to be ill posed in Appendix B. The matrix $A := [(1-2h)M - I]$ is negative definite unless $h = 0$ and $M = I$, in which case $A = 0$ and the difference is conserved. The consequence of this equation is that the behavior of the two agents will converge as soon as there are either mutations $M \neq I$ or interactions $h \neq 0$ or both. If $h = 0$ and $M \neq I$, the convergence between the two agents is driven by the self-monitoring process. In fact, if there is mutation, there exists a vector $\boldsymbol{x}$ that minimizes the self-monitoring term and every agent will converge towards this particular idiolect, since the mutation matrix is the same for all agents. If $h \neq 0$ and $M = I$, then it is the interaction process that drives the convergence between the two agents. The greater the parameter h, the faster the convergence.

For this simple case, we compare the behavior of the weak-noise SDE, the Dirichlet approximation, and the deterministic limit. As we discussed in Sec. IV A, the normal approximation does not ensure that the utterance vector $\boldsymbol{u}$ is bounded. This leads to numerical difficulties and various numerical strategies have been proposed. We review them in Appendix C. For the weak-noise SDE, we consider two different implementations: the external limiter (EL) and the backward implicit split step (BISS) implementation (see [20] and Appendix C).

For the parameters used in the simulation, we used short utterances $L = 2$ and a symmetric mutation matrix M defined as

$$M := \begin{bmatrix} 1-q & q \\ q & 1-q \end{bmatrix}, \tag{17}$$

where $q = 0.001$ is a mutation parameter. The initial condition is set to $x_1^{(1)}(0) = 0.2$ and $x_1^{(2)}(0) = 0.6$. This gives an initial difference $z(0) = 0.4$. In the simulation h is varied from 0 to 1 and the statistics is performed on 100 trajectories for each value of h. The results are given in Fig. 4.

In Fig. 4 we display the results for the different algorithm for a short time $T = 1$ and for a long time $T = 100$, where $T := 2^8\delta t$ for continuous time limits and $\lambda = \delta t$ for the USM. Changing the value of λ in the USM therefore changes the time scale of the problem. Results are displayed in Fig. 4. For $T = 1$ we observe that all the different algorithms agree well with the deterministic limit as shown in the first and third

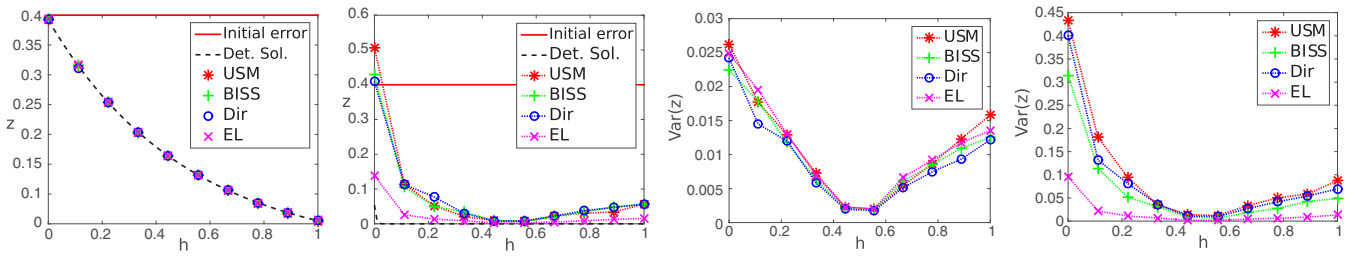

FIG. 4. Comparison of models for the USM and continuous time limits. The red horizontal line represents the initial error $z(0)$. The dashed black line represents the solution of the deterministic limit. We display the average over 100 simulations and the corresponding variances. The continuous time limits have to be compared with the discrete USM displayed in red stars. The first panel shows the averaged z at time $T = 1$, the second panel the averaged z at time $T = 100$, the third panel the variance of z at time $T = 1$, and the fourth panel the variance of z at time $T = 100$.

panels of Fig. 4. For longer times, however, the deterministic limit no longer agrees with the USM and its limits as shown in the second panel of Fig. 4. This is due to the fact that after a long time, the deterministic part of (14b) tends to 0 and the noise starts to contribute significantly to the dynamics. This is a numerical justification that the noise term has to be kept. The target curve in the second panel of Fig. 4 corresponds to the USM discrete solution displayed as red stars. We see that the Dirichlet approximation and the BISS implementation agree well with the discrete USM, but the EL algorithm fails to capture the dynamics. The introduction of a control function that modifies the normal approximation leads to a better approximation. Therefore, we will use the BISS algorithm for other numerical experiments.

Note that the variance of all models vanishes for $h = 0.5$, since in this case the dynamics of the variable z is always deterministic. For small values of h, the coupling is weak between the two agents and, as a consequence, the variance is larger for small values of h than for high values of h. The variance of the BISS algorithm slightly underestimates the variance of the USM and the Dirichlet approximation. This is a feature of this approximation and a consequence of the chosen control function given by Eq. (C6).

These numerical simulations show that the noise term has to be kept to accurately capture the behavior of the discrete USM. Recall that the deterministic limit corresponds to the KM expansion with the scaling $\lambda = \delta t$. The influence of the weak-noise has to be kept and the KM analysis is insufficient to capture this dynamics.

In the next section we discuss the coarse-graining procedure and explain how to obtain a SHMF approximation of the USM.

V. HETEROGENEOUS MEAN FIELD

The main result of this paper is the derivation of a coarse-grained approximation of the USM in the form of a SHMF approximation, which is based on the idea that the behavior of the complete network can be approximated by a smaller network of classes of agents grouped according to a relevant property. The SHMF approximation we present in this paper is based on grouping by degree, similarly to what is done in [6], but another grouping choice can be made. This grouping technique allows a coarse-graining procedure and the time scale of the approximation obtained is $t_{\mathcal{G}}$ instead of t_{int}, that is, we obtain an approximation at the population level, thus realizing the diagonal arrow of Fig. 2.

The main advantage of this approach is to keep the stochasticity of the model, while throwing away much of the network structure. The approximation obtained takes the form of a system of SDEs capturing the behavior of the entire ABM. With this approximation, the influence of the different parameters on the population behavior can be analyzed. In the rest of this section, we discuss the network and the state space reduction induced by a HMF approach, derive the SHMF approximation of the USM, and apply it to simple network topologies. We leave the discussion of complicated topologies for future work and focus in this paper on regular and star-shaped networks. In the case of regular networks, there is a single class of nodes and the SHMF approximation reduces the dynamics to a single SDE. This SDE is of the same form as a Wright-Fisher (WF) diffusion process (see Appendix A) and known results about this process can be applied. We also compare trajectories of the discrete USM with those of the SHMF approximation to qualitatively validate the approximation. Unfortunately, it is not possible to provide a good analysis of pathwise convergence of the SHMF approximation to the USM, since the sources of noise are of different natures. We then discuss the results for a star-shaped network. This example illustrates the robustness of the SHMF approximation for a very heterogeneous network.

A. Graph and state space reduction

We now describe the graph and state space reduction induced by an SHMF approximation. The idea is to group the nodes according the relevant property. This partition of the nodes in classes implies the existence of an equivalence relation, where the elements of the node partition are equivalence classes. In this paper we group the nodes by degree. This grouping is common in HMF approximations (see, for example, [6]). Note that other groupings are possible; one can group all the nodes and obtain a mean field approximation or one can group nodes by communities. In each case, a partition in equivalence classes is implied.

In this paper, we group the nodes by degree κ, that is, we introduce the equivalence relation $\sim_\kappa$ defined as

$$i \sim_\kappa j \quad \text{if} \quad \kappa(i) = \kappa(j),$$

where $\kappa(i)$ is the degree of node i. We set $\kappa(i) = k$ and then denote the corresponding equivalence class by $[k]$. The nodes of the reduced graph are given by the classes $[k]$ and are given a weight N_k representing the number of nodes contributing to the class $[k]$. Links between degree classes $[k]$ and $[k']$ exist whenever there is a link connecting a node of degree k to a node of degree k' in the original network. These directed links are weighted by $p(k'|k)$, which represents the probability that a node of degree k is connected to a node of degree k'. Note that in general $p(k'|k) \neq p(k|k')$ and that self-links are possible, since different nodes of the same degree can be connected together.

Example. The reduced graph of a regular network (a network in which all nodes have the same degree) is a single node with a connection to itself (see the left panel of Fig. 5). The reduced graph of a star-shaped network has two connected nodes, but no connection to itself, since in this topology the node of one class always interacts with nodes of the other class (see the right panel of Fig. 5).

In the SHMF approximation, each degree classes is described by a single belief distribution $\boldsymbol{x}^{(k)} \in \mathbb{P}_V$ defined as

$$\boldsymbol{x}^{(k)} := \frac{1}{N_k} \sum_{i \in [k]} \boldsymbol{x}^{(i)}, \quad \boldsymbol{x}^{(k)} \in \mathbb{P}_V,$$

where N_k is the number of agents of degree k in the network. If there are K classes, the dimension of the state space is $K(V-1)$, since $\mathbb{P}_V$ is of dimension $V-1$ because of the normalization constraint. In the original model, the dimension of the state space is $N(V-1)$. If $K \ll N$, the SHMF approximation significantly reduces the dimension of the state space.

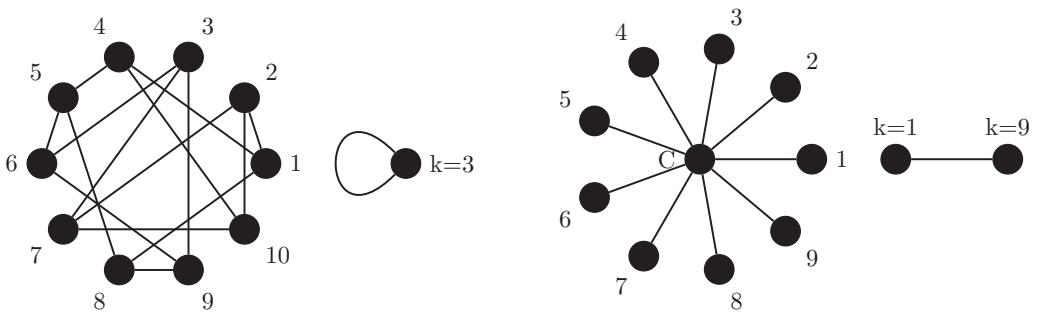

FIG. 5. Illustration of the network reduction for regular networks (left) and star-shaped networks (right). In each panel the left part is the original network and the right part is the reduced network.

B. Derivation of the stochastic heterogeneous mean field approximation

We can now derive the SHMF approximation of the USM. This is where the work done in previous sections, in particular the continuous-in-L normal approximation of Sec. IV A, pays off. The SHMF approximation uses a time unit corresponding to the network time $t_{\mathcal{G}}$. The core idea of the approximation is to consider a class of nodes as a single agent, which corresponds to the vertical arrow between Disc. ABM and Disc. PBM in Fig. 2, and to use the continuous time limit obtained in the previous section to implement the horizontal arrow between Disc. PBM and Cont. PBM in Fig. 2. To do so, we group all the agents belonging to the same degree class and ask them to produce all the utterances they have to utter during a complete network update and consider the results as a single class utterance of length $L_k := \frac{kL}{E} N_k$. At each network time step, all degree classes exchange their class utterance with the other degree classes. A weight $p(k'|k)$ is given to these utterances, proportional to the probability that the two degree classes are connected. Since L_k is usually large, the normal approximation, which fails in the two-agent case, is now justified by the central limit theorem and one can use it to approximate the average utterance by

$$\boldsymbol{u}^{(k)} = M\left(\boldsymbol{x}^{(k)} + \frac{1}{\sqrt{L_k}} D(\boldsymbol{x}^{(k)})\boldsymbol{\xi}^{(k)}\right), \tag{18}$$

where M is the production error, or mutation, matrix and $D(\boldsymbol{x})$ is a Cholesky square root of the covariance matrix of a multinomial distribution (see Appendix B) and $\boldsymbol{\xi}^{(k)} \sim \mathcal{N}(0, \boldsymbol{I})$ is a normally distributed random vector.

We assume that $G^{(ij)} = \frac{1}{E}\delta_{i\leftrightarrow j}$, that is, the probability to pick an edge is uniform. The averaged change of a degree class $[k]$ at the network level is given by

$$\begin{aligned}\delta\boldsymbol{x}^{(k)} &= \lambda\frac{(1-h)k}{E}(\boldsymbol{u}^{(k)} - \boldsymbol{x}^{(k)}) \\ &\quad + \lambda\frac{hk}{E}\sum_{k'} p(k'|k)(\boldsymbol{u}^{(k')} - \boldsymbol{x}^{(k)}),\end{aligned} \tag{19}$$

where $p(k'|k)$ is the probability that a node of degree k is connected to a node of degree k' and E is the number of edges of the network.

Introducing the degree k utterance (18) into Eq. (19) and introducing the scaling $dt = \frac{1}{E}$, which is motivated by the fact that the interaction time is much faster than the network time (see Fig. 1), gives the SDE

$$\begin{aligned}d\boldsymbol{x}^{(k)} &= \lambda\Big[(1-h)k(M\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k)}) \\ &\quad + hk\sum_{k'} p(k'|k)(M\boldsymbol{x}^{(k')} - \boldsymbol{x}^{(k)})\Big]dt \\ &\quad + \lambda\Big[(1-h)\sqrt{\frac{k}{LN_k}}MD(\boldsymbol{x}^{(k)})d\boldsymbol{W}_t^{(k)} \\ &\quad + hk\sum_{k'} p(k'|k)\frac{1}{\sqrt{Lk'N_{k'}}}MD(\boldsymbol{x}^{(k')})d\boldsymbol{W}_t^{(k')}\Big],\end{aligned} \tag{20}$$

where the time is measured in $t_{\mathcal{G}}$ units. Equation (20) is the continuous time SHMF approximation of the USM. The first two terms describe the influence of the self-monitoring and accommodation processes and the last two terms model the corresponding noises. There is one such equation for each degree class $[k]$.

This approximation greatly reduces the number of degrees of freedom whenever $K \ll N$, where K is the number of equivalence classes $[k]$. The number of agents N_k in a class $[k]$ only enters Eq. (20) as a parameter of the noise coefficients. The noises are therefore dependent on the size of the network. For large networks, the contribution of the noise is small and vanishes in the limit of infinite networks. In other words, the global stochastic dynamics of the model is a finite-size effect. The parameter L also controls the amplitude of the noise. The shorter the utterance, the larger the noise. This justifies the interpretation of L as describing the variability of a speaker.

In the SHMF approximation, we are throwing away a great deal of information about the topology of the network, conserving only the different degree classes. If we model the social interaction by randomly ordering the edges and going through them exactly once at each network time, the nodes with a large number of neighbors interact more often than nodes with a small number of neighbors. As a result, we expect the evolution of the different classes of nodes to evolve on a different time scale. In Eq. (20) the time scale difference is encoded in the dependence on k of the dynamics of $\boldsymbol{x}^{(k)}$.

We expect the SHMF approximation to be a good approximation if the number of agents in each degree class is sufficiently large for the normal approximation to hold and if

the nodes forming a class are well connected. Both of these conditions are satisfied for regular networks. A limiting case is given by star-shaped networks, in which there is no direct connection between nodes of degree 1 and where there is a single node of degree $N-1$. In this case, both conditions are violated and we show that the SHMF approximation nevertheless captures well the dynamics of the system.

In the following we apply the SHMF approximation to regular networks and to star-shaped networks. The regular network analysis allows us to study in detail the influence of the different parameters and the star-shaped network illustrates the robustness of the method.

C. Regular networks and the Wright-Fisher SDE

The case of regular networks is particularly interesting, since its SHMF approximation takes the form of a WF diffusion, which has been widely studied, much is known about the behavior of this process and we can apply this knowledge to the study of the SHMF approximation of the USM. The left panel of Fig. 5 illustrates the type of network we are considering, together with the reduced network of degree class on which the SHMF approximation is defined.

For simplicity, we restrict the discussion to the case of two variants $V=2$ and we choose a mutation matrix M of the form (17), with $q=10^{-3}$. The Cholesky square root $D(\boldsymbol{x})$ is given by Eq. (B9b). Under these assumptions, the SHMF approximation of the regular network is given by

$$\begin{aligned} dx_1^{(k)} &= k\lambda\left(x'^{(k)}_1 - x_1^{(k)}\right)dt \\ &\quad + \lambda(1-2q)\sqrt{\frac{k}{LN}}\sqrt{x_1^{(k)}\left(1-x_1^{(k)}\right)}dW_t^{(k)} \\ &= -\gamma\left(x_1^{(k)}-\frac{1}{2}\right)dt + \sigma\sqrt{x_1^{(k)}\left(1-x_1^{(k)}\right)}dW_t^{(k)}, \end{aligned} \tag{21}$$

where $\boldsymbol{x}' = M\boldsymbol{x}$ and $x_2^{(k)} = 1 - x_1^{(k)}$ to conserve probability. We also introduced $\gamma = 2qk\lambda$ and $\sigma = \lambda(1-2q)\sqrt{\frac{k}{LN}}$. The time has to be measured in $t_{\mathcal{G}}$ units.

In order to simplify the discussion, we scale the time variable as $t' := \lambda k t_{\mathcal{G}}$. With this scaling, Eq. (21) can be rewritten as

$$dx_1^{(k)} = -\gamma'\left(x_1^{(k)}-\frac{1}{2}\right)dt' + \sigma'\sqrt{x_1^{(k)}\left(1-x_1^{(k)}\right)}dW_{t'}^{(k)}, \tag{22}$$

where

$$\begin{aligned} \gamma' &= 2q, \\ \sigma' &= (1-2q)\sqrt{\frac{\lambda}{LN}}. \end{aligned} \tag{23}$$

Equation (22) is a WF process, as discussed in Appendix B. This process occurs in many different contexts such as population genetics and economics (see, for example, [21]). The type of noise occurring in Eq. (22) can be found in another model for language change in which an age-structured population is considered (see [22]).

The three relevant parameters controlling the dynamics are λk, q, and $r := \frac{\lambda}{LN}$. The time scale evolution is controlled by the product λk of the learning rate and the degree of the class. This is expected, since λ models the amplitude of change at each time step and since an agent of degree k interacts k times during a single network update. The parameter q models the influence of error production and innovations. If $q=0$, then there is no error and no innovation. In this case, $\lambda' = 0$ and $\sigma' = \sqrt{r}$. In other words, the dynamics is only driven by noise and the boundaries are absorbing. Once the population reaches a consensus, the state of the system no longer changes. The other extreme case is when $q=\frac{1}{2}$. In this case, the multiplication by M in (18) randomizes the output and the noise information is lost. In this case, the noise coefficient σ' vanishes and the dynamics is driven by the drift term and the solution deterministically goes to $x_1^{(k)} = \frac{1}{2}$. The parameter r controls the size of the noise and is proportional to λ and inversely proportional to N and L. When r is large the noise dominates the dynamics and the solution is pushed towards the boundary of the domain; when r is small the drift term dominates the dynamics and the solution is pushed towards the center of the domain. Therefore, we expect a change of the stationary distribution shape between a U-shaped and a bell-shaped distribution by varying r and q.

We can now take advantage of the WF form of the SHMF approximation of the regular network for which the stationary distribution is known and takes the form of a Beta distribution (see, for example, [12]). For long times, the probability $p_*(x)$ that a trajectory reaches a certain value x is given by

$$p_*(x) = 2\frac{\Gamma\left(\frac{\gamma}{\sigma^2}+\frac{1}{2}\right)}{\Gamma\left(\frac{\gamma}{\sigma^2}\right)+\Gamma\left(\frac{1}{2}\right)}[4x(1-x)]^{(\gamma/\sigma^2-1)}. \tag{24}$$

For more than two variants, one can generalize this formula. The resulting Dirichlet distribution can be found in [23]. In our case, the single parameter of this distribution is given by

$$\frac{\gamma}{\sigma^2} = \frac{2q}{(1-2q)^2 r}. \tag{25}$$

We see that this parameter only depends on q and r, as expected. The distribution (24) changes from a bell-shaped distribution for $\sigma^2 > \gamma$ to a U-shaped distribution for $\sigma^2 < \gamma$, with a transition when $\sigma^2 = \gamma$. In the bell-shaped regime, there is no convention emerging and the agents are probability matching and the dynamics is dominated by the deterministic term. In the U-shaped regime, conventions emerge, but are not stable unless $q=0$, in which case the distribution degenerates to the discrete probability mass function weighting only $x=0$ and 1. From Eq. (25) one obtains the critical value for $q_*(r)$ given by

$$q_*(r) = \frac{r}{1+2r+\sqrt{1+4r}}, \tag{26}$$

which behaves as $q_*(r) \propto \frac{r}{2}$, when $r \to 0$.

Figure 6 summarizes the behavior of regular graphs. The exact critical value q_* and its asymptotic behavior are displayed, separating the parameter space into regions of U-shaped and bell-shaped stationary distributions. Since the parameter r is inversely proportional to N, this phase diagram is the signature of a finite-size effect. For an infinite graph, the distribution is always bell shaped and no convention can ever globally emerge.

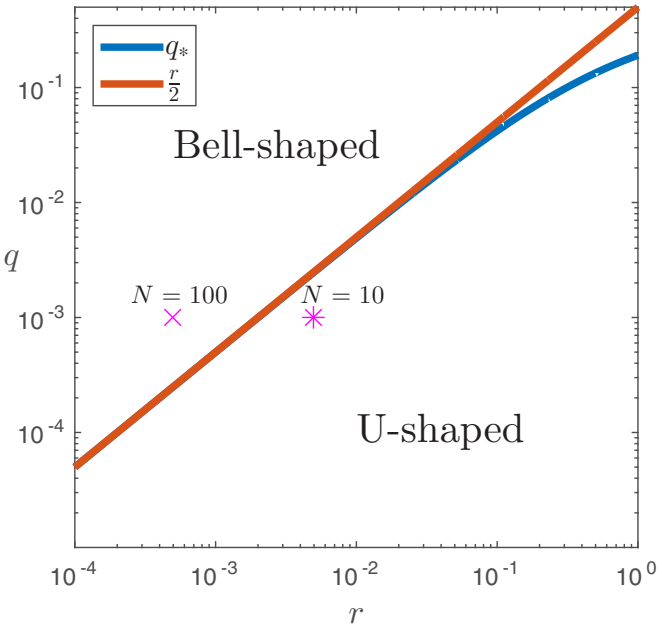

FIG. 6. Illustration of the critical parameter $q_*(r)$ separating the bell-shaped and the U-shaped domain. The red curve is the approximate behavior. For illustration, we display the positions of the two examples considered in this section (regular network of degree 3 for $N = 10$ and 100 agents).

For a fixed parameter q, decreasing the parameter r leads to a phase transition from a U-shaped to a bell-shaped distribution. We recall that r is proportional to λ and inversely proportional to N and L.

The parameter k, representing the degree of the regular graph, only contributes to the λk time scale parameter and therefore has no influence on the shape of the stationary distribution of the averaged system. For the numerical simulations, we choose regular networks of degree $k = 3$. The agents choose between $V = 2$ variants and produce utterances of length $L = 2$ for $T = 10^4$ network updates. For the other parameters, we choose $h = 0.5$ and $q = 0.001$. We then change the number of agents from $N = 10$ to $N = 100$, which corresponds to values of $r = 200$ and 2000, respectively. With these parameters, the critical values of the mutation parameter are $q_* \approx 2.475 \times 10^{-3}$ and 2.498×10^{-4}. These values are plotted in Fig. 6. Since the chosen value of $q < q_*$ for $N = 10$, we expect a U-shaped distribution and since $q > q_*$ for $N = 100$, we expect a bell-shaped distribution.

Results for the trajectories of the discrete USM and for the corresponding SHMF approximation are displayed in Fig. 7. The results of the SHMF approximation are in good qualitative agreement with the results of the discrete USM and can therefore be used to characterize the behavior of the system.

In order to validate the SHMF approximation of the USM, we computed the stationary distribution of the discrete USM and compared the results with the analytical prediction of its SHMF approximation. The results displayed in Fig. 8 are excellent already with relatively few statistics of 1000 trajectories. Since the computation of the stationary distribution of the discrete USM is time consuming and due to the symmetry of the dynamics, we augmented the statistics by considering both

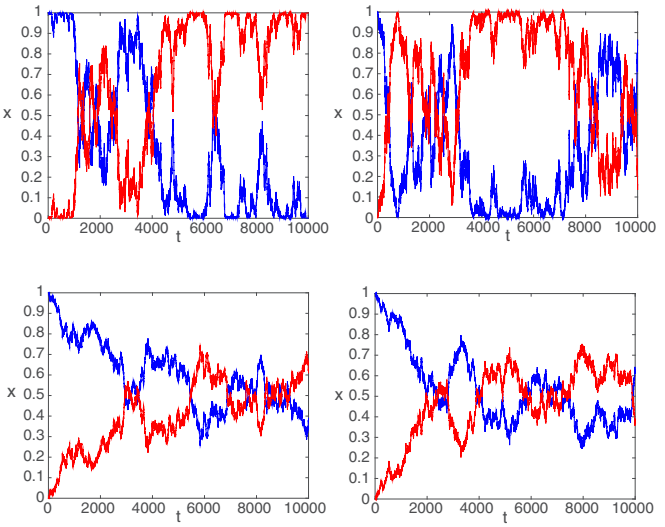

FIG. 7. Comparison between the discrete USM (left column) and the SHMF approximation limit of it (right column). For these simulations, the parameters are $h = 0.5$, $\lambda = 0.1$, $V = 2$, $L = 2$, $T = 10^4$, and $q = 0.001$ and the number of agents is 10 in the top row and 100 in the bottom row. The regular graph is of degree $k = 3$. At the beginning of the simulation, all agents share the convention to use the variant $v = 1$.

x_1 and x_2 at the end of the simulation, since the two variants are equivalent.

For regular networks, the parameter h does not play a role in predicting the population-averaged stationary distribution. This has been verified by performing the simulation for different values of h (not shown). However, in [12] it is shown that h does play a role in the marginal stationary distribution; in other words, h has an influence on the stationary distribution of higher-order moments, rather than on the stationary distribution of the average behavior analyzed here. In order to better understand the effect of h on the population averaged stationary distribution, we now consider the case of star-shaped networks.

D. Star-shaped networks

We now consider the case of a heterogeneous network, namely, the star-shaped network. This kind of network is

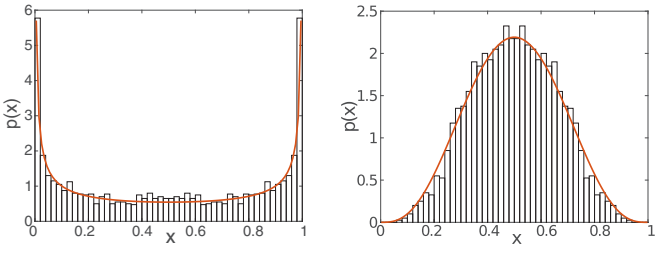

FIG. 8. Comparison between the stationary distribution of the discrete USM on a regular network with the distribution predicted by the SHMF approximation. Shown on the left is the stationary distribution for 10 agents. On the right is the stationary distribution for 100 agents.

characterized by two classes of nodes, a central node of degree $N-1$ and $N-1$ nodes of degree 1 connected to it. The right panel of Fig. 5 illustrates this kind of network, together with the reduced network used in the SHMF approximation.

For this kind of network, the SHMF approximation is expected to fail to capture efficiently the dynamics. This is due to the fact that the normal approximation is not well justified for the central node labeled C in the right panel of Fig. 5. Furthermore, all the degree one nodes interact through the mediation of this poorly approximated node.

In order to simplify the notation, we introduce the quantities

$$\sigma_1 = \lambda(1-2q)\frac{1}{\sqrt{L(N-1)}}, \quad \gamma_1 = 2q\lambda,$$
$$\sigma_N = \lambda(1-2q)\sqrt{\frac{(N-1)}{L}}, \quad \gamma_N = 2q\lambda(N-1)$$

and we have the relations $\gamma_N = (N-1)\gamma_1$ and $\sigma_N = (N-1)\sigma_1$. With this notation, the SHMF formulation of the USM for a star-shaped network of N agents reads

$$\begin{aligned}
dx_1^{(1)} &= \left[\gamma_1(1-h)\left(\frac{1}{2}-x_1^{(1)}\right)+\lambda h\left(x_1'^{(N-1)}-x_1^{(1)}\right)\right]dt \\
&\quad +(1-h)\sigma_1\sqrt{x_1^{(1)}\left(1-x_1^{(1)}\right)}dW_t^{(1)} \\
&\quad +h\sigma_1\sqrt{x_1^{(N-1)}\left(1-x_1^{(N-1)}\right)}dW_t^{(N-1)}, \\
dx_1^{(N-1)} &= \left[(1-h)\gamma_N\left(\frac{1}{2}-x_1^{(N-1)}\right)+\lambda h\left(x_1'^{(1)}-x_1^{(N-1)}\right)\right]dt \\
&\quad +(1-h)\sigma_N\sqrt{x_1^{(N-1)}\left(1-x_1^{(N-1)}\right)}dW_t^{(N-1)} \\
&\quad +h\sigma_N\sqrt{x_1^{(1)}\left(1-x_1^{(1)}\right)}dW_t^{(1)}, \qquad (27)
\end{aligned}$$

where $x_1'^{(i)}$ is the first component of $M\boldsymbol{x}^{(i)}$, $i = 1, N-1$.

For Eq. (27) we do not have an analytical form for the stationary distribution of $x_1^{(1)}$ and $x_1^{(N-1)}$. However, the results obtained for regular networks can be used to gain some insight into this problem. For example, we observe that the noise magnitude is much larger for the central node than for the other nodes. This is a consequence of the time scale difference between the two classes of nodes.

In order to illustrate the behavior of the star-shaped network and, in particular, the influence of the h parameter, we performed simulations of the star-shaped network for parameters similar to those used for the regular network case. We consider $V = 2$ variants that are used to produce utterances of length $L = 2$; the mutation parameter entering the symmetric mutation matrix M is fixed to $q = 10^{-3}$. The learning rate is $\lambda = 0.1$ and the simulation ends after $T = 10^4$ network time steps. For these parameters, we vary the number of agents $N = 10$ or 100 and the parameter $h = 0.9$ and 0.1. In these settings, we compare the behavior of the discrete USM with the behavior of the corresponding SHMF approximation.

The results are displayed in Fig. 9. Figures 9(a)–9(d) show the results for $N = 10$ agents and Figs. 9(e)–9(h) show the results for $N = 100$ agents. Figures 9(a), 9(b), 9(e), and 9(f) correspond to $h = 0.9$ and Figs. 9(c), 9(d), 9(g), and 9(h) correspond to $h = 0.1$. We observe that between $N = 10$ and 100 there is a transition from a U-shaped to a bell-shaped distribution. Since the critical value of q_* is derived for the regular network, this existence of a transition should be fairly robust for different topologies. The exact value of q_* is not known for the star-shaped network case, but such a transition is nevertheless expected.

We now discuss the influence of the h parameter. As expected, the behavior of the central node is noisier than the average of the other nodes. This is a consequence of the time scale difference between the two classes of nodes. If h is reduced, the coupling between the two classes of nodes is weakened and the noise increases. The SHMF approximation

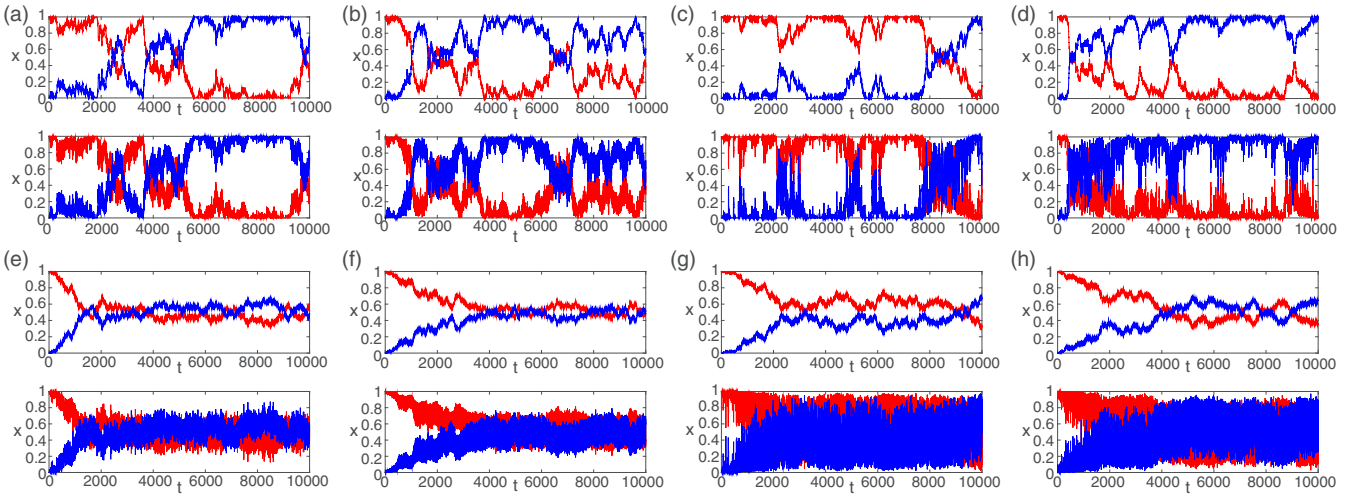

FIG. 9. Comparison of time series of the discrete USM and of the SHMF approximation limit of it for the star-shaped network. For these simulations, the parameters are $\lambda = 0.1$, $V = 2$, $L = 2$, $T = 10^4$, and $q = 0.001$. The top part of each graph displays the behavior of the degree $k = 1$ nodes and the bottom part of each graph displays the behavior of the central node. The value of N is (a)–(d) 10 and (e)–(h) 100. The first and third columns display the results of the USM and the second and fourth columns display the results of the SHMF approximation. In the first two columns $h = 0.9$ and in the last two columns $h = 0.1$. At the beginning of the simulation, all agents share the convention to use the variant $v = 1$.

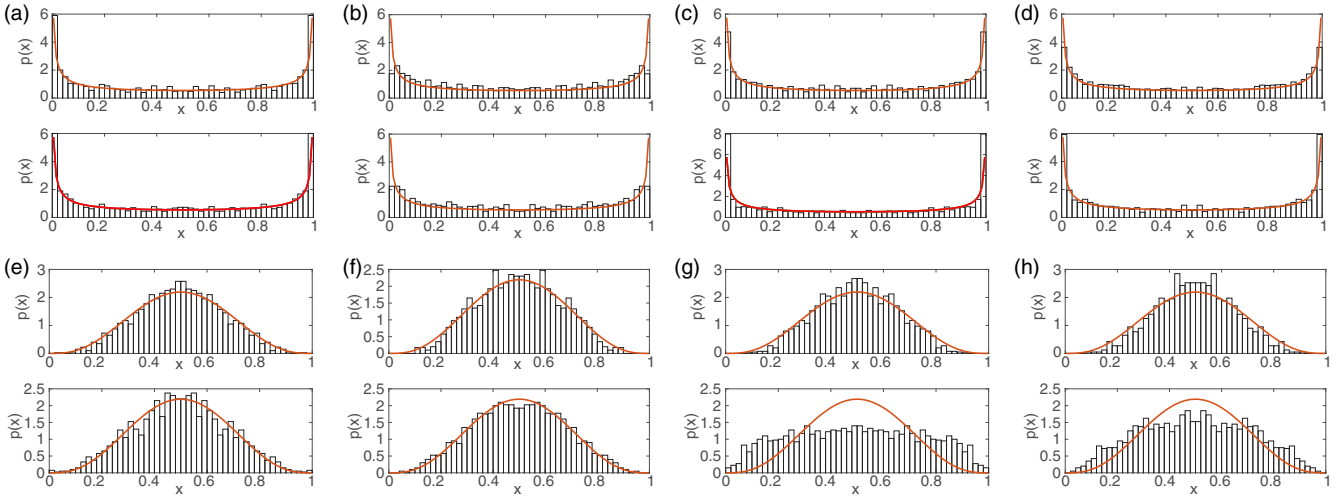

FIG. 10. Comparison between the discrete USM and its SHMF approximation limit for the star-shaped network. For these simulations, the parameters are $\lambda = 0.1$, $V = 2$, $L = 2$, $T = 4 \times 10^3$, and $q = 0.001$. The top part of each graph displays the distribution of the degree $k = 1$ nodes and the bottom part of each graph displays the distribution of the central node. In the top row the value of N is 10 and in the second row the value of N is 100. The first and third columns display the results of the USM and the second and fourth columns display the results of the SHMF approximation. In the first two columns $h = 0.9$ and in the last two columns $h = 0.1$. The red line is the solution of the mean field approximation. It helps to see how the star-shaped network differs from the regular network case. The first two columns and the last two columns have to be compared.

reproduces this behavior and therefore captures the effect of h. However, it seems that the SHMF approximation converges with a slower rate towards the stationary distribution. This could be explained by the fact that in the discrete USM, the edges are updated sequentially, whereas in the SHMF approximation they are updated synchronously. The sequential update might converge faster than the synchronous one, as observed in Fig. 9. For large networks, the difference between sequential and synchronous updates diminishes and the convergence rates of the two approaches become more similar. Even if the convergence rate of the USM and the SHMF approximation might be different, the stationary state should nevertheless be similar for both approaches. In order to verify this prediction, we compare the numerical stationary distribution of the USM and of the SHMF approximation in the same conditions as in Fig. 9, computed at $T = 4000$. We also compare the results with the predicted mean field approximation corresponding to Eq. (21), where the degree k has to be replaced by the averaged degree $\bar{k} = 2 + 2/N$ of star-shaped networks. Since the mean field stationary distribution does not depend on k, we expect it to be a good approximation if the coupling between the two classes of nodes is strong enough.

In Fig. 10 we display the results for the stationary distribution in the same settings as those used in Fig. 9. The sampling is done at the final time $t_G = T$, $T = 4000$. To augment the statistics, we once again considered both x_1 and x_2 in the histograms, which artificially enforces symmetry of the distributions.

For $h = 0.9$, we observe that the distributions of the two degree classes are both in good agreement with the mean field limit. For $N = 10$ [Figs. 10(a) and 10(b)], the SHMF approximation underestimates the behavior at the boundary of the domain. This might be due to a discretization error effect. In fact, we know that the algorithm used converges strongly (trajectorywise), but we do not know at what rate. Since this rate can be arbitrarily slow, the results of the SHMF approximation close to the boundary might not be reliable see Appendix C for details). Apart from this effect, the results of the USM and the SHMF approximation are in good agreement with the mean field limit (red lines in Fig. 10). For $h = 0.9$ and $N = 100$ [Figs. 10(e) and 10(f)], the class of degree $k = 1$ nodes follows, as expected, the mean field limit for both the USM and the SHMF approximation. The behavior of the central node is noisier and the comparison is less straightforward. We observe that for the SHMF approximation the effect of the noise manifests itself by an undersampling of the peak of the distribution. This effect is less clear in the USM case, but the results are quite noisy. We can conclude that the SHMF approximation in this case is slightly better that the mean field, which completely neglects the topology of the graph.

For the weaker coupling $h = 0.1$, the dynamics becomes more interesting. For $N = 10$ [Figs. 10(c) and 10(d)], the agreement between the USM and the SHMF approximation is good and the effect of the stronger noise is mainly seen at the boundaries of the domain, where it is observed that the central nodes spends more time close to the boundary than the average degree $k = 1$ node. The discretization problems might explain the undersampling of the SHMF approximation in the boundary regions. Another explanation can be linked with the network reduction itself. In fact, for star-shaped networks, it is not clear whether the approximation should work at all, since the normal approximation fails for the central node. For $N = 100$ [Figs. 10(g) and 10(h)], the effect of the noise is much clearer. Since the coupling is fairly weak, the time scale difference between the classes of noise leads to a very noisy behavior of the central node. As a result, the distribution of

the central node flattens, while the behavior of the degree 1 nodes remains close to the mean field. We also observe an oversampling effect of the peak for the degree 1 nodes. This can be explained by the fact that in the absence of coupling, the behavior of all the degree 1 agents becomes independent. As a consequence of the central limit theorem, the variance of their average behavior is reduced, explaining the stronger peak observed. The results of the SHMF approximation qualitatively capture the correct behavior and provide a better prediction than the mean field approximation (red line in Fig. 10). In this case, the very strong noise entering the dynamics of the central node leads to greater numerical errors (see Appendix C for details). Another explanation of these differences lies in the difference in the variance of the multinomial distribution and the BISS approximation of it. For instance, in Fig. 4 it is shown that the variance of the BISS approximation is smaller than the variance of the discrete USM. Since the central node is the only node in this class, the hypothesis based on the central limit theorem, needed to justify the normal approximation, no longer holds. This is a possible explanation of the disagreement of the USM and the SHMF approximation results. Since the flattening effect is seen in both simulations, we can nevertheless conclude that the SHMF approximation captures the main characteristics of the dynamics of the star-shaped network and in particular the effect of h better than the mean field approximation (red line).

In this section we have shown that the SHMF approximation is able to capture the dynamics of the USM on different network structures and to reproduce both the trajectories and the stationary distributions of the model. The results of the star-shaped network are less convincing due to numerical problems in the simulation of the SHMF approximation in the presence of strong noise. This is the case, for example, for $N = 100$ and $h = 0.1$. The search for better algorithms to sample these trajectories is left for future work.

VI. CONCLUSION

In this paper we discussed the USM for language change and its continuous time limits. In order to overcome the parameter restrictions of the FP continuous time limit obtained using the KM expansion, we have proposed a continuous time limit based on the normal approximation of the multinomial distribution. For two agents, this approximation leads to a weak-noise SDE generalizing the KM expansion solution. We argued that the weak noise should not be neglected for two reasons: (i) The noise is heuristically of the same order of magnitude as the drift term and (ii) the drift term vanishes in long time simulations. The weak-noise limit also captures the influence of the noise of both utterances, whereas the FP limit of [12] neglects the influence of the noise of the incoming utterance.

Using this continuous time limit, we derived a stochastic version of the HMF approximation and applied it to regular and star-shaped networks. This approximation allows us to study the dynamics of the system at the level of the network instead of at the level of the agents, which is a great improvement on the analysis of agent-based models in that it provides analytical tools to characterize the noise-driven phase transition and therefore opens the door to exciting results, since the grouping procedure can be done using different criteria.

For regular networks, the SHMF formulation turns out to be a Jacobi process described by the WF diffusion SDE. The analysis has shown that the dynamics is controlled by three interdependent parameters λk, q, and $r := \frac{\lambda}{LN}$ and only the last two parameters contribute to the stationary distribution. The h parameter, weighting the self-monitoring and the accommodation process in the USM, does not enter the SHMF approximation. As a result, one can interpret this fact as "prestigious" agents (large h) do not have a particular influence on the dynamics. This is true as long as only the attention parameter is taken into account. If a prestigious agent influences the weighting of its variants, then the effect can be large. This can be modeled, for example, by a preference mechanism (see [24]). For regular networks, we computed the critical value $q_*(r)$ and obtained a phase diagram describing the form of the stationary distribution. Such a distribution is also expected on average for regular graphs, since the SHMF approximation of regular networks can be interpreted as a mean field approximation of any network. Since r is inversely proportional to N, the functional dependence $q_* \propto \frac{\lambda}{2LN}$ is the signature of a finite-size effect. For instance, the stationary distribution of the averaged population transitions from a U-shaped distribution to a bell-shaped distribution when N increases. In the limit $N \to \infty$, the noise term vanishes and the solution exponentially decays to $x = \frac{1}{2}$. This case corresponds to the deterministic limit obtained using the KM expansion and only scaling $\lambda = \delta t$.

For star-shaped networks, a case where the SHMF approximation is expected not to be a very good approximation, the SHMF approximation still provides satisfying results, capturing the time scale difference between the central node and the outer nodes. This effect is not captured by the mean field approximation (which corresponds to applying the results from regular networks to star-shaped networks).

In the context of cultural evolution, the interesting regime is when the stationary distribution is U shaped, which is the signature of the creation of populationwide conventions that can change. In our model, for large populations (for small values of r), we have shown that a convention does not usually emerge (the stationary distribution is bell shaped). This is a signature of what is called by Nettle [25] the threshold problem. This problem states that in large populations, it is really difficult to change an established convention. Nettle proposed a solution by using the social impact theory. In our case, we can obtain populationwide conventions by increasing r or decreasing q (see Fig. 6). In other words, one can explain the emergence of new conventions in a large population if the learning rate λ is sufficiently large or if the variability of speech is sufficiently large, that is, if the utterance length L is small. In both cases, the influence of errors is increased. If q is very small, conventions emerge, but they are stable and cultural change is rare. In fact, if $q = 0$ the boundaries are *exits* according to Feller classification and conventions are absorbing states.

The social impact theory relies on prestigious agents to explain language change. In the USM, the way the influence of a specific agent is encoded through the attention parameter h. Since this parameter does not enter the mean field equation,

our results suggests that an influential agent only has a weak influence on the dynamics. However, if the prestige is associated with the variant used by an influential agent, the conclusion changes and this can have a tremendous influence on the dynamics. In this case, the different variants are no longer equivalent and the learning rule has to be adapted to take this into account. Such a variant weighting can be encoded either in the mutation matrix M if the variant is objectively, or functionally, better or through the introduction of a preference mechanism [24], which allows the agent to adapt one's behavior to the different variants. These modifications have a huge impact on the dynamics of the system and remain to be studied.

In the USM, the influence of the topology can be studied using the SHMF approximation. In this paper we have provided a proof of principle and the complete analysis of the influence of the network remains to be done. In [26], the authors discussed the dynamics of a model of language change in social networks using computer simulations. We believe that our approach can complement and possibly explain the results obtained in [26].

Left for future work is the study of the influence of the topology of the network using the SHMF approximation and the influence of nonconstant $h^{(ij)}$ or asymmetric M, as well as study the influence of different extensions of the USM, such as the presence of preferences for a particular variant and the influence of group membership (different behavior depending on the identity of the interacting agents). The SHMF approximation only characterizes the stationary distribution of the population-averaged behavior. Extending this approach to higher moments will complement the knowledge and provide information on the dispersion around the averaged behavior.

ACKNOWLEDGMENTS

The author was funded by the Swiss National Science Foundation Grant No. P2GEP2_159156. The author would like to thank Richard A. Blythe and Gilles Vilmart for comments and interesting discussions and Edwige Dugas for discussions about the application of this work to linguistics.

APPENDIX A: CONTINUOUS TIME LIMIT OF THE USM USING THE KRAMERS-MOYAL EXPANSION

In this Appendix we provide the derivation of continuous time limits of the USM using the KM expansion [27], similarly to what was done in [12]. This method provides a FP equation for the probability distribution $p(\boldsymbol{x}^{(i)},t;\boldsymbol{X}^{(-i)})$ to find agent i with an idiolect $\boldsymbol{x}^{(i)}$ at time t, knowing the state of the rest of the population $\boldsymbol{X}^{(-i)}$. The exponent $(-i)$ means all agents except agent i. This is notation borrowed from game theory. The time t has to be measured in t_{int} units here.

The KM expansion of a stochastic process $\boldsymbol{x}^{(i)}$ is given by

$$\begin{aligned}\frac{\partial p}{\partial t} &= -\sum_{v=1}^{V-1}\frac{\partial}{\partial x_v^{(i)}}\{\beta_v(\boldsymbol{x}^{(i)})p\}\\ &\quad+\frac{1}{2}\sum_{v=1}^{V-1}\sum_{w=1}^{V-1}\frac{\partial^2}{\partial x_v^{(i)}\partial x_w^{(i)}}\{\beta_{vw}(\boldsymbol{x}^{(i)})p\}\\ &\quad+\cdots,\end{aligned} \tag{A1}$$

where the jump moments are defined as

$$\beta_v(\boldsymbol{x}^{(i)}) = \lim_{\delta t\to 0}\frac{\langle\delta x_v^{(i)}(t)\rangle}{\delta t}, \tag{A2}$$

$$\beta_{vw}(\boldsymbol{x}^{(i)}) = \lim_{\delta t\to 0}\frac{\langle\delta x_v^{(i)}(t)\delta x_w^{(i)}(t)\rangle}{\delta t}. \tag{A3}$$

Here the average is taken over utterance production and over edges of the graph connected to agent i, which are the two sources of randomness in the model.

In order to simplify a bit the presentation, we assume that the off-diagonal terms of the matrix M are of order $(\delta t)^{1/2}$ or smaller. If this is the case, then one can write the condition

$$O(\|M-I\|_\infty) = O((\delta t)^{1/2}). \tag{A4}$$

This assumption was used in [12] and we only do it in this appendix. We also introduce the notation $\boldsymbol{x}' = M\boldsymbol{x}$ for convenience.

Under assumption (A4) one can collect all the terms that depend on the off-diagonal term of M in $O(\|M-I\|_\infty)$. We can then write the first two jump moments as

$$\langle\delta x_v^{(i)}\rangle = \sum_{j\neq i}G^{(ij)}\lambda\left[(1-h)\left(x_v'^{(i)}-x_v^{(i)}\right)+h\left(x_v'^{(j)}-x_v^{(i)}\right)\right] \tag{A5}$$

and

$$\begin{aligned}\langle\delta x_v^{(i)}\delta x_w^{(i)}\rangle = \sum_{j\neq i}G^{(ij)}\lambda^2\Bigg[&\frac{(1-h)^2}{L}x_v^{(i)}\left(\delta_{vw}-x_w^{(i)}\right)\\ &+\frac{h^2}{L}x_v^{(j)}\left(\delta_{vw}-x_w^{(j)}\right)\\ &+h(1-h)\left(x_w^{(j)}-x_w^{(i)}\right)\left(x_v^{(j)}-x_v^{(i)}\right)\\ &+O(\|M-I\|_\infty)\Bigg].\end{aligned} \tag{A6}$$

Equation (A6) has been computed for the definition (3a) of the utterances. The expression for the definition (3b) differs from (A6) and can be computed easily.

In order to obtain a FP equation, scaling assumptions have to be made to ensure that (A5) and (A6) are both of order δt and that higher-order jump moments are of higher order. The scaling chosen in [12] is given by

$$\lambda = (\delta t)^{1/2}, \tag{A7a}$$

$$M_{vw} = \bar{M}_{vw}(\delta t)^{1/2} \quad \text{for } v\neq w, \tag{A7b}$$

$$h = \bar{h}(\delta t)^{1/2}. \tag{A7c}$$

Equation (A7b) is equivalent to assumption (A4). This scaling is the only one compatible with the KM expansion leading to a FP equation with nonvanishing diffusion, given the constraints on the parameters. In particular, if the constraint that L is an integer is relaxed, another scaling would work. It is given by

$$\lambda = \delta t, \tag{A8a}$$

$$L = \bar{L}\delta t. \tag{A8b}$$

The scaling of L means that the number of tokens in an utterance tends to 0. Since $L \geqslant 1$, this is not possible.

The USM scaling is problematic since it requires one to scale the h parameter and the off-diagonal terms of M, limiting this continuous time limit to a small part of the parameter space. The assumption that off-diagonal terms of M are small corresponds to a small probability of innovation and is not really problematic. The restriction on h is much stronger, since it requires the accommodation process to be negligible with respect to the self-monitoring process, which is usually not justified.

The second scaling is not satisfying either since it requires one to scale an integer quantity, namely, L. Therefore, none of these approaches gives a satisfying FP equation.

If one does not want to scale either h or L, the only possible scaling left is to scale $\lambda \propto \delta t$. In this case, the KM expansion is truncated after the first term and there is no diffusion term. In other words, the continuous time limit is deterministic. The KM expansion, therefore, predicts that the behavior of a single agent on the t_{int} time scale is deterministic, unless the attention parameter h and the off-diagonal terms of M are small, in which case we obtain a diffusive dynamics.

APPENDIX B: THE USM AND THE WRIGHT-FISHER PROCESS

In this Appendix we present the WF stochastic process, also called the Jacobi process, and connect it to the USM. We then discuss the different available choices of choosing a noise form in the resulting SDE.

1. Definition of the Wright-Fisher process

The WF models of population genetics [28,29] model the biological transmission of alleles of genes between generations of a population. This model gives rise to a stochastic process described by the SDE

$$d\boldsymbol{x}_t = -\lambda(\boldsymbol{x}_t - \boldsymbol{b}) + c\left[\operatorname{diag}(\boldsymbol{x}_t) - \boldsymbol{x}_t\boldsymbol{x}_t^T\right]^{1/2} d\boldsymbol{W}_t, \tag{B1}$$

where $\lambda > 0$, $\boldsymbol{b} \in \mathbb{P}_V$, c is a positive constant, and the square root of the matrix has to be taken in the Cholesky sense. Finally $d\boldsymbol{W}_t$ is a d-dimensional white noise. Here d is not necessarily equal to the dimension V of $\boldsymbol{x}_t$, since the Cholesky square root is not necessarily a square matrix. Note that one only needs to consider the first $V-1$ components of $\boldsymbol{x}_t$, since the last one can be recovered using the conservation of probability.

Definition 1 (square root in the Cholesky sense). A matrix $D \in \mathbb{R}^{m\times n}$ is said to be a square root in the Cholesky sense of a matrix $A \in \mathbb{R}^{m\times m}$ if

$$DD^T = A.$$

The square root in the Cholesky sense is not uniquely defined and not necessarily a square matrix (see [30] for details).

The WF process (B1) satisfies a sum to unit constraint and a non-negativity constraint. In [31] it is shown that there are only a few stochastic processes that satisfy such a conservation law. The WF process naturally arises from discrete processes when characterized by a multinomial sampling process. This is the case in the original discrete WF model as well as in the USM. For instance, the matrix $\operatorname{diag}(\boldsymbol{x}_t) - \boldsymbol{x}_t\boldsymbol{x}_t^T$ corresponds to the covariance matrix of the normalized multinomial sampling process. The WF stochastic process is sometimes called the Jacobi process by mathematicians and economists [32–35], because the infinitesimal generators of this process, obtained as the eigenfunctions of the backward Kolmogorov equation, are Jacobi polynomials.

We now discuss the form of the matrix $D(\boldsymbol{x})$, which is the square root of the matrix

$$A(\boldsymbol{x}) := \operatorname{diag}(\boldsymbol{x}_t) - \boldsymbol{x}_t\boldsymbol{x}_t^T \tag{B2}$$

in the Cholesky sense. This matrix is needed to complete the formulation of the WF process (B1) and is also used in the normal approximation (9) assumed in the derivation of continuous time limits of the USM. We start with the special case of $V = 2$ and discuss then the general case.

2. Form of D when $V = 2$

In the case $V = 2$, a vector $\boldsymbol{x} \in \mathbb{P}_2$ is such that $x_2 = 1 - x_1$ and the matrix $A(\boldsymbol{x})$ takes the simple form

$$A(\boldsymbol{x}) = x_1(1-x_1)\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{B3}$$

This matrix has many Cholesky square roots. We list three of them here:

$$D_1(\boldsymbol{x}) := \frac{1}{\sqrt{2}}\sqrt{x_1(1-x_1)}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \tag{B4a}$$

$$D_2(\boldsymbol{x}) := \sqrt{x_1(1-x_1)}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{B4b}$$

$$D_3(\boldsymbol{x}) := \begin{bmatrix} (1-x_1)\sqrt{x_1} & -x_1\sqrt{1-x_1} \\ -(1-x_1)\sqrt{x_1} & x_1\sqrt{1-x_1} \end{bmatrix}. \tag{B4c}$$

It is straightforward to check that these matrices are Cholesky square roots of (B3). The matrix $D_1(\boldsymbol{x})$ is also a square root of $A(\boldsymbol{x})$ in the sense that $D_1(\boldsymbol{x})D_1(\boldsymbol{x}) = A(\boldsymbol{x})$. For simplicity, the matrix $D_2(\boldsymbol{x})$ is usually chosen.

3. General case

If $V > 2$, then one can generalize the choices (B4b) and (B4c), but the choice (B4a) is more difficult to generalize. The choice (B4a) corresponds to a Cholesky square root that is also a square root in the sense that $D_1^2 = A$. Finding matrix square roots is not an easy task and therefore this choice is difficult to generalize.

The choice (B4b) takes into account the possible reaction channels and considers one noise for each. For example, if $V = 3$ then there are three mutation channels $x_1 \leftrightarrow x_2$, $x_1 \leftrightarrow x_3$, and $x_2 \leftrightarrow x_3$ and a possible Cholesky square root is given by

$$D(\boldsymbol{x}) := \begin{bmatrix} \sqrt{x_1x_2} & \sqrt{x_1x_3} & 0 \\ -\sqrt{x_2x_1} & 0 & \sqrt{x_2x_3} \\ 0 & -\sqrt{x_3x_1} & -\sqrt{x_3x_2} \end{bmatrix}. \tag{B5}$$

This can be generalized to an arbitrary number of variants. The dimension of this matrix is $V \times \binom{V}{2}$, $\binom{V}{2}$ is the binomial coefficient. This is the formulation used, for example, in [20].

The generalization of Eq. (B4c) is given by

$$\{D(\boldsymbol{x})\}_{vw} := (\delta_{vw} - x_v)\sqrt{x_w}. \tag{B6}$$

This choice is associated with the multivariate Jacobi process (see [33]).

All these choices are equivalent. In the context of SDEs, they correspond to different trajectories of the same Wiener process (see, for example, [30]).

4. The USM and the WF process

We now detail a case in which the USM is related to the WF process. We consider a network of $N = 2$ agents using $V = 2$ variants. Given a mutation matrix M of the form

$$M := \begin{bmatrix} 1-m_2 & m_1 \\ m_2 & 1-m_1 \end{bmatrix}, \tag{B7}$$

from Eq. (14) we obtain

$$d\boldsymbol{x}^{(i)} = G^{(i)}\left[[(M-I)\boldsymbol{x}^{(i)}]dt + \left(\frac{1}{\sqrt{L}}D(M\boldsymbol{x}^{(i)})d\boldsymbol{\xi}^{(i)}\right)\right] \tag{B8a}$$

or

$$d\boldsymbol{x}^{(i)} = G^{(i)}\left[[(M-I)\boldsymbol{x}^{(i)}]dt + \left(\frac{1}{\sqrt{L}}MD(\boldsymbol{x}^{(i)})d\boldsymbol{\xi}^{(i)}\right)\right], \tag{B8b}$$

where the matrix $D(\boldsymbol{x})$ is given by Eq. (B4b). We then have

$$D(M\boldsymbol{x}^{(i)}) = \sqrt{x_1'^{(i)}\left(1-x_1'^{(i)}\right)}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{B9a}$$

$$MD(\boldsymbol{x}^{(i)}) = (1-m_1-m_2)\sqrt{x_1^{(i)}\left(1-x_1^{(i)}\right)}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{B9b}$$

where x_1' is the first component of $\boldsymbol{x}' = M\boldsymbol{x}$.

As stated in Sec. III the components of $\boldsymbol{x}^{(i)} \in \mathbb{P}_2$ are not independent and it is sufficient to only consider the evolution of the first components. We obtain

$$dx_1^{(i)} = -\gamma(x_1^{(i)} - \mu)dt + \sigma_{sb}\sqrt{x_1^{(i)}\left(1-x_1^{(i)}\right)}dW_t^{(i)}, \tag{B10a}$$

$$dx_1^{(i)} = -\gamma(x_1^{(i)} - \mu)dt + \sigma_{bs}\sqrt{x_1'^{(i)}\left(1-x_1'^{(i)}\right)}dW_t^{(i)}, \tag{B10b}$$

where

$$\begin{aligned} \gamma &:= -G^{(i)}(m_1+m_2), \\ \mu &:= \frac{m_2}{m_1+m_2}, \\ \sigma_{sb} &:= \frac{\sqrt{dt}G^{(i)}}{\sqrt{L}}(1-m_1-m_2), \\ \sigma_{bs} &:= \frac{\sqrt{dt}G^{(i)}}{\sqrt{L}}. \end{aligned} \tag{B11}$$

We now discuss the influence of the ordering of sampling and biasing on this weak-noise SDE. Since Eq. (B10) has to satisfy the constraint that $x_1 \in [0,1]$, the SDE has to satisfy a number of properties discussed in [31]. One of these properties is that the noise coefficient has to vanish at the boundaries of the interval, that is, at $x_1 = 1$ and 0. The property is satisfied by Eq. (B10a), but not by Eq. (B10b). One can therefore conclude that Eq. (B10b) is ill posed, since it does not conserve the probability. The well-posed character of Eq. (B10a) then follows from the Yamada-Watanabe theorem [36]. This theorem has to be used because the noise coefficient is not a Lipschitz continuous function. Recall that a Lipschitz continuous function f satisfies

$$\|f(x) - f(y)\|_2 \leqslant C_L|x-y| \,\forall x,y \in \mathcal{D}(f), \tag{B12}$$

where C_L is the Lipschitz constant and $\mathcal{D}(f)$ is the domain of f. This non-Lipschitz aspect of the noise coefficient in Eq. (B10) is at the origin of numerical difficulties (see Appendix C).

Under the normal approximation, the order of the sampling and biasing processes matters. Sampling first and then biasing is the only one that leads to a well-posed SDE. This order is also more natural in a linguistic framework; it corresponds to first sampling for the belief distribution $\boldsymbol{x}$ and then modifying the output as a result of passing through the articulatory-auditory channel. The other ordering corresponds to modifying the belief distribution $\boldsymbol{x}$ and then sampling from the biased distribution $\boldsymbol{x}' = M\boldsymbol{x}$ without error. The origin of errors is more difficult to justify in this case. The most natural ordering is then also the mathematically preferred. Note that in the USM and in the Dirichlet approximation, both orders are possible and the restriction obtained here is intrinsically connected with the normal approximation and its unbounded nature (see Sec. IV A). The discussion about the well-posed character of the equation has been done for two variants. Using the results of [31], one can generalize the results to an arbitrary number of variants and we arrive at the same conclusion that the only ordering leading to a well-posed equation is sampling first and then biasing. Another example in which the USM is linked with the WF model is given by the SHMF approximation for regular graphs given in Sec. V C.

APPENDIX C: NUMERICAL ALGORITHMS

In this Appendix we discuss the possible numerical strategies to solve the WF SDE occurring as the SHMF approximation of the regular network and how to extend the results to the general SHMF equation. We consider the SDE

$$dx_t = -\gamma(x_t - \mu)dt + \sigma\sqrt{x_t(1-x_t)}dW_t, \tag{C1}$$

where x_t is a realization of the stochastic process and dW_t a white noise. Equation (21) is of this form. This equation can be shown to be well posed on $[0,1]$ using a result of Yamada and Watanabe [36]. One difficulty that arises with this kind of SDE is linked with the non-Lipschitz aspect of the multiplicative noise. Most of the usual proofs of convergence rely on a Lipschitz condition (B12).

In order to accurately capture the trajectory of the stochastic process, one needs a strongly convergent numerical method (see [37] for details about the types of convergence). There is a weaker notion of convergence, known as weak convergence, that only requires convergence on average and not trajectory-

wise. Obtaining weakly convergent methods is usually much easier than obtaining strongly convergent methods.

In the rest of this appendix we discuss the performance of different numerical methods for integrating Eq. (C1) and then obtain a numerical method to integrate Eq. (20).

1. Wright-Fisher diffusion

We now discuss the different families of methods that have been used to integrate Eq. (C1).

The first class of methods is the usual algorithms for SDE, such as the Euler-Maruyama (EM) method or the Milstein method. This class of methods fails to capture the correct dynamics of Eq. (C1) due to the non-Lipschitz multiplicative noise and the solution can leave the domain $[0,1]$ of Eq. (C1).

The second class of methods introduces a min-max limiter

$$\Theta(x) = \min[\max(x,0),1] \tag{C2}$$

to project the numerical solution back onto $[0,1]$. The resulting methods are bounded and weakly convergent, but they are not strongly convergent. One can apply this limiter under the square root to get the internal limiter (IL) method (see [38]) or to the complete update to get the external limiter (EL) method.

The third class of methods is based on the fact that Eq. (C1) has an exact solution for particular values of the parameters. Moro and Schurz proposed a splitting method based on this idea (see [39]). The Moro-Schurz (MS) method has parameter restrictions, which limit its applicability.

The fourth and last class of methods uses a control function to keep the solution in the bounded domain. This idea is due Milstein [40] and can be used alone [the balanced implicit method (BIM); see [40]] or in conjunction with a splitting method (the BISS method; see [20]). These methods can be applied without restriction and can be shown to strongly converge. However, the rate at which the method converges is not known.

We now discuss the implementations of the different methods. Let us introduce Δt, a time increment, and ΔW^n, the nth increment of a Wiener process. Then one can obtain the discrete approximation $x^n \approx x(t_n = ndt)$ of the different algorithms.

The EM method is given by

$$x^{n+1} = x^n - \gamma(x^n - \mu)\Delta t + \sigma\sqrt{x^n(1-x^n)}\Delta W^n.$$

This method does not converge at all and leads to unrealistic results.

The IL method is defined as

$$x^{n+1} = x^n - \gamma(x^n - \mu)\Delta t + \sigma\sqrt{\Theta(x^n)[1-\Theta(x^n)]}\Delta W^n.$$

This method is not bounded, but is weakly convergent.

The EL method is defined as

$$x^{n+1} = \Theta[x^n - \gamma(x^n - \mu)\Delta t + \sigma\sqrt{x^n(1-x^n)}\Delta W^n].$$

This method is bounded and weakly convergent, but not strongly convergent.

The MS method is based on the following splitting:

$$dy_1 = \frac{\sigma^2}{2}\left(y_1 - \frac{1}{2}\right)dt + \sigma\sqrt{y_1(1-y_1)}dW_t, \tag{C3a}$$

$$dy_2 = \left[-\gamma(y_2 - \mu) - \frac{\sigma^2}{2}\left(y_2 - \frac{1}{2}\right)\right]dt. \tag{C3b}$$

Equation (C3a) has an exact solution. At each time step, Eq. (C3a) is solved analytically and serves as an initial condition for Eq. (C3b), which is solved using a forward Euler algorithm. This method is only bounded for certain parameters, for which it is both weakly and strongly convergent.

The BIM is defined as

$$x^{n+1} = x^n - \gamma(x^n - \mu)\Delta t + \sigma\sqrt{x^n(1-x^n)}\Delta W^n + D(x^n)(x^n - x^{n+1}),$$

where

$$D(x^n) = d^0(x^n)\Delta t + d^1(x^n)|\Delta W_n| \tag{C4}$$

is a control function. The convergence of this method depends on the choice of d^0 and d^1. For good control functions, this method is both weakly and strongly convergent. The limitation of this method is that it is not always clear how to choose the appropriate control functions.

The BISS method is based on the splitting

$$dy_1 = \sigma\sqrt{y_1(1-y_1)}dW_t, \tag{C5a}$$

$$dy_2 = -\gamma(y_2 - \mu)dt \tag{C5b}$$

and solves Eq. (C5a) using the BIM and Eq. (C5b) using a forward Euler step. In the BIM step, the function $d^0(x^n) \equiv 0$ and

$$d^1(x) = \begin{cases} \sigma\sqrt{\frac{1-\varepsilon}{\varepsilon}} & \text{if } y < \varepsilon \\ \sigma\sqrt{\frac{1-y}{y}} & \text{if } \varepsilon \leqslant y < \frac{1}{2} \\ \sigma\sqrt{\frac{y}{1-y}} & \text{if } \frac{1}{2} < y \leqslant 1-\varepsilon \\ \sigma\sqrt{\frac{1-\varepsilon}{\varepsilon}} & \text{if } y > 1-\varepsilon, \end{cases} \tag{C6}$$

where ε is a small tolerance parameter, defined in [20] as

$$\varepsilon = \min(A\Delta t, B\Delta t, 1 - A\Delta t, 1 - B\Delta t) > 0$$

for Δt small enough and where $A = \gamma\mu$ and $B = \gamma(1-\mu)$. The discretization of Eq. (C5) takes the form

$$y_*^{n+1} = y_1^n + \frac{\sigma\sqrt{y_1^n(1-y_1^n)}\Delta W_n}{1 + d^1(y_1^n)|\Delta W_n|}, \tag{C7a}$$

$$y_1^{n+1} = y_*^{n+1} - \gamma(y_*^{n+1} - \mu)\Delta t. \tag{C7b}$$

This method is bounded and converges weakly and strongly for all parameters if Δt is chosen small enough.

The characteristics of the different methods are summarized in Table I. The two best methods are the BIM and the BISS method. We choose the BISS method because it is easier to adapt to more complex dynamics such as the dynamics of the SHMF approximation. The BIM could also be used; the problem is that for more complex dynamics, a good control function d^0 is difficult to define. Since the convergence rate of

TABLE I. Summary of the properties of the different numerical methods available to solve the WF diffusion equation. We consider whether the method produces a bounded result and is weakly convergent or strongly convergent. In the case of convergence, we specify whether on not there is a restriction on the parameters.

Method	Bounded	Weak convergence	Strong convergence	No restriction
EM	×	×	×	×
IL	×	√	×	×
EL	√	√	×	×
MS	√	√	√	×
BIM	√	√	√	√
BISS	√	√	√	√

the BISS method is unknown, we expect numerical artifacts close to the boundaries of the domain, where the Lipschitz condition is not satisfied.

2. Numerical methods for the SHMF approximation of the USM

In Sec. I of this Appendix we recalled the numerical methods available for solving the WF diffusion equation. For the SHMF approximation of the USM (20), one needs to deal with the noises of all neighboring degree classes. This can be done by a splitting method inspired by the BISS algorithm. We describe it for two variants $V = 2$. The idea is to split the update between the utterance production (which is noisy) and the deterministic learning rule. The continuous time version of the normal approximation (18) is obtained by scaling $\frac{1}{E} = dt$. The first component u_1 is of the form

$$u_1 = a + b\left[x_1 + \sigma_k\sqrt{x_1(1-x_1)}\frac{\Delta W_n}{\Delta t}\right], \tag{C8}$$

where $\sigma_k = (kLN_k)^{-1/2}$, $a = m_1$, and $b = 1 - m_1 - m_2$ for a matrix M defined by Eq. (17).

The idea is to modify Eq. (C8) by introducing the control function d^1 of Eq. (C6), leading to the utterance production

$$u_1^{n+1} = a + b\left[x_1^n + \frac{\sigma_k\sqrt{x_1^n(1-x_1^n)}\frac{\Delta W_n}{\Delta t}}{1 + d^1(x_1)|\Delta W_n|}\right] \tag{C9a}$$

and the learning update given by Eq. (19)

$$x_1^{(k),n+1} = x_1^{(k),n} + \lambda(1-h)k\left(u_1^{(k),n+1} - x_1^{(k),n}\right)\Delta t + \lambda hk\sum_{k'} p(k'|k)\left(u_1^{(k'),n+1} - x_1^{(k),n}\right)\Delta t. \tag{C9b}$$

Equation (C9) is the BISS algorithm for the SHMF approximation of the USM. This approximation ensures that $x_1 \in [0,1]$ for all degree classes. The strong convergence remains to be shown, but since the BISS method is strongly convergent, we have good reason to think that this algorithm conserves this property. For $V > 2$, the same idea can be used. The only difficulty is to find an appropriate control function.

[1] L. P. Kadanoff, J. Stat. Phys. **137**, 777 (2009).
[2] J. P. Gleeson, Phys. Rev. X **3**, 021004 (2013).
[3] V. Sood and S. Redner, Phys. Rev. Lett. **94**, 178701 (2005).
[4] V. Sood, T. Antal, and S. Redner, Phys. Rev. E **77**, 041121 (2008).
[5] C. Castellano, in *Modeling Cooperative Behavior in the Social Sciences*, edited by P. L. Garrido, J. Marro, and M. A. Muñoz, AIP Conf. Proc. No. 779 (AIP, Melville, 2005), p. 114.
[6] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).
[7] F. Tria, V. D. Servedio, S. S. Mufwene, and V. Loreto, PLoS ONE **10**, e0120771 (2015).
[8] C. Beckner, R. A. Blythe, J. Bybee, M. H. Christiansen, W. Croft, N. C. Ellis, J. Holland, J. Ke, D. Larsen-Freeman, and T. Schoenemann, Lang. Learning **59**, 27 (2009).
[9] L. Steels, in *Parallel Problem Solving for Nature—PPSN VI*, edited by M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H.-P. Schwefel, Lecture Notes in Computer Science Vol. 1917 (Springer, Berlin, 2000), pp. 17–26.
[10] J. Michaud, in *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, edited by S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, and T. Verhoef (2016).
[11] M. H. Christiansen and S. Kirby, Trends Cogn. Sci. **7**, 300 (2003).
[12] G. J. Baxter, R. A. Blythe, W. Croft, and A. J. McKane, Phys. Rev. E **73**, 046118 (2006).
[13] P. Grifoni, A. D'Ullizia, and F. Ferri, Artif. Intell. Rev. **45**, 369 (2016).
[14] G. J. Baxter, R. A. Blythe, W. Croft, and A. J. McKane, Lang. Var. Change **21**, 257 (2009).
[15] R. A. Blythe and W. Croft, Language **88**, 269 (2012).
[16] C.-M. Pop and E. Frey, Phys. Rev. E **88**, 022814 (2013).
[17] H. E. Pemberton, Am. Sociol. Rev. **1**, 547 (1936).
[18] F. Ghanbarnejad, M. Gerlach, J. M. Miotto, and E. G. Altmann, J. R. Soc. Interface **11**, 20141044 (2014).
[19] W. Croft, *Explaining Language Change: An Evolutionary Approach* (Pearson, London, 2000).
[20] C. Dangerfield, D. Kay, S. MacNamara, and K. Burrage, BIT Numer. Math. **52**, 283 (2012).
[21] N. Bouleau and C. Chorro, halshs.archives-ouvertes.fr/halshs-01162452/.
[22] W. G. Mitchener, http://mitchenerg.people.cofc.edu/WGM-SMLCSSAPDI-2.pdf.
[23] G. J. Baxter, R. A. Blythe, and A. J. McKane, Math. Biosci. **209**, 124 (2007).
[24] J. Michaud, talk given at the Centre for Language Evolution, http://www.lel.ed.ac.uk/cle/index.php/2016/05/17/24-may-jerome-michaud/.
[25] D. Nettle, Lingua **108**, 95 (1999).
[26] J. Ke, T. Gong, and W. S. Wang, Commun. Comput. Phys. **3**, 935 (2008).
[27] H. Risken, *The Fokker-Planck Equation* (Springer, Berlin, 1989).
[28] S. Wright, Genetics **16**, 97 (1931).
[29] R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).

[30] D. W. Stroock and S. S. Varadhan, *Multidimensional Diffusion Processes* (Springer, Berlin, 2007).
[31] J. Bakosi and J. Ristorcelli, Int. J. Stoch. Anal. **2014**, 603692 (2014).
[32] C. Gouriéroux and P. Valéry, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.196&rep=rep1&type=pdf.
[33] C. Gourieroux and J. Jasiak, J. Econ. **131**, 475 (2006).
[34] A. Kuznetsov, Solvable Markov processes, Ph.D. thesis, University of Toronto, 2004.
[35] S. Karlin and H. E. Taylor, *A Second Course in Stochastic Processes* (Elsevier, Amsterdam, 1981).
[36] T. Yamada and S. Watanabe, J. Math. Kyoto Univ. **11**, 155 (1971).
[37] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Stochastic Modelling and Applied Probability Vol. 23 (Springer, Berlin, 1992).
[38] C. R. Doering, K. V. Sargsyan, and P. Smereka, Phys. Lett. A **344**, 149 (2005).
[39] E. Moro and H. Schurz, SIAM J. Sci. Comput. **29**, 1525 (2007).
[40] G. Milstein, E. Platen, and H. Schurz, SIAM J. Numer. Anal. **35**, 1010 (1998).