Structural Drift: The Population Dynamics of Sequential Learning

James P. Crutchfield
Sean Whalen

SFI WORKING PAPER: 2010-05-011

SANTA FE INSTITUTE

Santa Fe Institute Working Paper 10-05-XXX
arxiv.org:1005.2714 [q-bio.PE]

Structural Drift: The Population Dynamics of Sequential Learning

James P. Crutchfield[1, 2, 3, 4, *] and Sean Whalen[1, 3, †]

[1]*Complexity Sciences Center*
[2]*Physics Department*
[3]*Computer Science Department*
University of California Davis, One Shields Avenue, Davis, CA 95616
[4]*Santa Fe Institute*
1399 Hyde Park Road, Santa Fe, NM 87501

(Dated: May 17, 2010)

We introduce a theory of sequential causal inference in which learners in a chain estimate a structural model from their upstream "teacher" and then pass samples from the model to their downstream "student". It extends the population dynamics of genetic drift, recasting Kimura's selectively neutral theory as a special case of a generalized drift process using structured populations with memory. We examine the diffusion and fixation properties of several drift processes and propose applications to learning, inference, and evolution. We also demonstrate how the organization of drift process space controls fidelity, facilitates innovations, and leads to information loss in sequential learning with and without memory.

PACS numbers: 87.18.-h 87.23.Kg 87.23.Ge 89.70.-a 89.75.Fb
Keywords: neutral evolution, causal inference, genetic drift, allelic entropy, allelic complexity, structural stasis

I. "SEND THREE- AND FOUR-PENCE, WE'RE GOING TO A DANCE"

This phrase was heard, it is claimed, over the radio during WWI instead of the transmitted tactical phrase "Send reinforcements we're going to advance" [1]. As illustrative as it is apocryphal, this garbled yet comprehensible transmission sets the tone for our investigations here. Namely, what happens to knowledge when it is communicated sequentially along a chain, from one individual to the next? What fidelity can one expect? How is information lost? How do innovations occur?

To answer these questions we introduce a theory of sequential causal inference in which learners in a communication chain estimate a structural model from their upstream "teacher" and then, using that model, pass along samples to their downstream "student". This reminds one of the familiar children's game *Telephone*. By way of quickly motivating our sequential learning problem, let's briefly recall how the game works.

To begin, one player invents a phrase and whispers it to another player. This player, believing they have understood the phrase, then repeats it to a third and so on until the last player is reached. The last player announces the phrase, winning the game if it matches the original. Typically it does not, and that's the fun. Amusement and interest in the game derive directly from how the initial phrase evolves in odd and surprising ways.

The game is often used in education to teach the lesson that human communication is fraught with error. The final phrase, though, is not merely accreted error but the product of a series of attempts to parse, make sense, and intelligibly communicate the phrase. The phrase's evolution is a trade off between comprehensibility and accumulated distortion, as well as the source of the game's entertainment. We employ a much more tractable setting to make analytical progress on sequential learning,[1] intentionally selecting a simpler language system and learning paradigm than likely operates with children.

Specifically, we develop our theory of sequential learning as an extension of the evolutionary population dynamics of genetic drift, recasting Kimura's selectively neutral theory [2] as a special case of a generalized drift process of structured populations with memory. Notably, this requires a new and more general information-theoretic notion of fixation. We examine the diffusion and fixation properties of several drift processes, demonstrating that the space of drift processes is highly organized. This organization controls fidelity, facilitates innovations, and leads to information loss in sequential learning and evolutionary processes with and without memory. We close by proposing applications to learning, inference, and evolutionary processes.

*Electronic address: chaos@cse.ucdavis.edu
†Electronic address: whalen@cs.ucdavis.edu

[1] There are alternative, but distinct notions of *sequential learning*. Our usage should not be confused with notions in education and psychology, sometimes also referred to as *analytic* or step-by-step learning [3]. Our notion also differs in motivation from those developed in machine learning, such as with statistical estimation for sequential data [4], though some of the inference methods may be seen as related. Perhaps the notion here is closer to that im-

To get started, we briefly review genetic drift and fixation. This will seem like a distraction, but it is a necessary one since available mathematical results are key. Then we introduce in detail our structured variants of these concepts—defining the *generalized drift process* and introducing a generalized definition of allelic fixation appropriate to it. With the background laid out, we begin to examine the complexity of structural drift behavior. We demonstrate that the diffusion takes place in a space that can be decomposed into a connected network of structured subspaces. We show how to quantify the degree of structure within these subspaces. Building on this decomposition, we explain how and when processes jump between these subspaces—innovating new structural information or forgetting it—thereby controlling the long-time fidelity of the communication chain. We then close by outlining future research and listing several potential applications for structural drift, drawing out consequences for evolutionary processes that learn.

Those familiar with neutral evolution theory are urged to skip to Sec. V, after skimming the next sections to pick up our notation and extensions.

II. FROM GENETIC TO STRUCTURAL DRIFT

Genetic drift refers to the change over time in genotype frequencies in a population due to random sampling. It is a central and well studied phenomenon in population dynamics, genetics, and evolution. A population of genotypes evolves randomly due to drift, but typically changes are neither manifested as new phenotypes nor detected by selection—they are *selectively neutral*. Drift plays an important role in the spontaneous emergence of mutational robustness [6, 7], modern techniques for calibrating molecular evolutionary clocks [8], and non-adaptive (neutral) evolution [9, 10], to mention only a few examples.

Selectively neutral drift is typically modeled as a stochastic process: A random walk in genotype space that tracks finite populations of individuals in terms of their possessing (or not) a variant of a gene. In the simplest models, the random walk occurs in a space that is a function of genotypes in the population. For example, a drift process can be considered to be a random walk of the *fraction* of individuals with a given variant. In the simplest cases there, the model reduces to the dynamics of repeated binomial sampling of a biased coin, in which the empirical estimate of bias becomes the bias in the next round of sampling. In the sense we will use the term, the sampling process is *memoryless*. The biased coin, as the population being sampled, has no memory: past samples are independent of future ones. The current state of the drift process is simply the bias, a number between zero and one that summarizes the state of the population.

The theory of genetic drift predicts a number of measurable properties. For example, one can calculate the expected time until all or no members of a population possess a particular gene variant. These final states are referred to as *fixation* and *deletion*, respectively. Variation due to sampling vanishes once these states are reached and, for all practical purposes, drift stops. From then on, the population is homogeneous. These states are fixed points—in fact, absorbing states—of the drift stochastic process.

The analytical predictions for the time to fixation and time to deletion were developed by Kimura and Ohta [2, 11] in the 1960s and are based on the memoryless models and simplifying assumptions introduced by Wright [12] and Fisher [13] in the early 1930s. The theory has advanced substantially since then to handle more complicated and realistic models and to predict the additional effects due to selection and mutation. One example is the analysis of the drift-like effect of *pseudohitchiking* ("genetic draft") recently given [14].

The following explores what happens when we relax the memoryless assumption. The original random walk model of genetic drift forces the statistical structure at each sampling step to be an independent, identically distributed (IID) stochastic process. This precludes any memory in the sampling. Here, we extend the IID theory to use time-varying probabilistic state machines to describe memoryful population sampling.

In the larger setting of sequential learning, we will show that memoryful sequential sampling exhibits structurally complex, drift-like behavior. We call the resulting phenomenon *structural drift*. Our extension presents a number of new questions regarding the organization of the space of drift processes and how they balance structure and randomness. To examine these questions, we require a more precise description of the original drift theory [15].

III. GENETIC DRIFT

We begin with the definition of an *allele*, which is one of several alternate forms of a gene. The textbook example is given by Mendel's early experiments on heredity [16], in which he observed that the flowers of a pea plant were colored either white or violet, this being determined by the combination of alleles inherited from its parents. A new, *mutant* allele is introduced into a population by the mutation of a *wild-type* allele. A mutant allele can be passed on to an individual's offspring who, in turn, may pass it on to their offspring. Each inheritance occurs with some probability.

Genetic drift, then, is the change of allele frequencies in a population over time: It is the process by which the number of individuals with an allele varies generation after generation. The Fisher-Wright theory [12, 13] models drift as a stochastic evolutionary process with neither selection nor mutation. It assumes random mating

plicated in mimicked behavior which drives financial markets [5].

between individuals and that the population is held at a finite, constant size. Moreover, successive populations do not overlap in time.

Under these assumptions the Fisher-Wright theory reduces drift to a binomial or multinomial sampling process—a more complicated version of familiar random walks such as Gambler's Ruin or Prisoner's Escape [17]. Offspring receive either the wild-type allele A_1 or the mutant allele A_2 of a particular gene $\mathcal{A}$ from a random parent in the previous generation with replacement. A population of N diploid[2] individuals will have $2N$ total copies of these alleles. Given i initial copies of A_2 in the population, an individual has either A_2 with probability $i/2N$ or A_1 with probability $1-i/2N$. The probability that j copies of A_2 exist in the offspring's generation given i copies in the parent's generation is:

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} . \qquad (1)$$

This specifies the Markov transition dynamic of the drift stochastic process over the discrete state space $\{0, 1/N, 2/N, \ldots, N-1/N, 1\}$.

This model of genetic drift is a discrete-time random walk, driven by samples of a biased coin, over the space of biases. The population is a set of coin flips, where the probability of HEADS or TAILS is determined by the coin's current bias. After each generation of flips, the coin's bias is updated to reflect the number of HEADS or TAILS realized in the new generation. The walk's absorbing states—all HEADS or all TAILS—capture the notion of fixation and deletion.

IV. GENETIC FIXATION

Fixation occurs with respect to an allele when all individuals in the population carry that specific allele and none of its variants. Restated, a mutant allele A_2 reaches fixation when all $2N$ alleles in the population are copies of A_2 and, consequently, A_1 has been *deleted* from the population. This halts the random fluctuations in the frequency of A_2, assuming A_1 is not reintroduced.

Let X be a binomially distributed random variable with bias probability p that represents the fraction of copies of A_2 in the population. The expected number of copies of A_2 is $\mathrm{E}[X] = 2Np$. That is, the expected number of copies of A_2 remains constant over time and depends only on its initial probability p and the total number ($2N$) of alleles in the population. However, A_2 eventually reaches fixation or is deleted due to the variance introduced by random finite sampling and the presence of absorbing states.

[2] Though haploid populations can be used, we focus on diploid populations (two alleles per individual) for direct comparison to Kimura's simulations. This gives a sample length of $2N$.

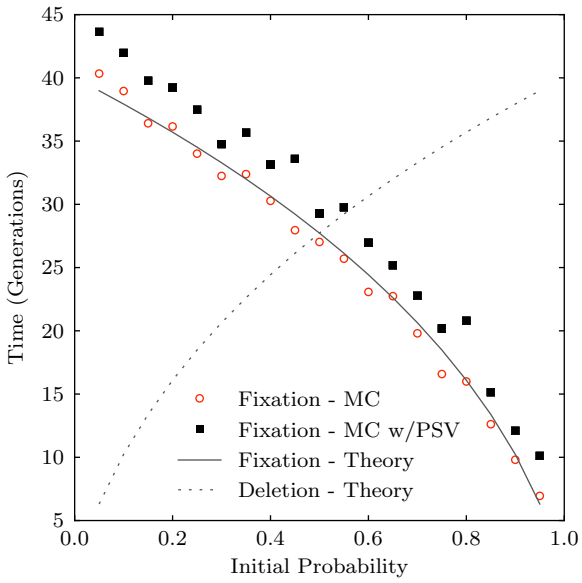

FIG. 1: Number of generations to fixation for a population of $N = 10$ individuals (sample size $2N = 20$), plotted as a function of initial allele frequency p under different sampling regimes: Monte Carlo (MC), Monte Carlo with pseudo-sampling variable (MC w/ PSV), and theoretical prediction (solid line, Theory). The time to deletion is also shown (dashed line, Theory).

Prior to fixation, the mean and variance of the change Δp in allele frequency are:

$$\mathrm{E}[\Delta p] = 0 \text{ and} \qquad (2)$$

$$\mathrm{Var}[\Delta p] = \frac{p\,(1-p)}{2N} , \qquad (3)$$

respectively. On average there is no change in frequency. However, sampling variance causes the process to drift towards the absorbing states at $p = 0$ and $p = 1$. The drift rate is determined by the current generation's allele frequency and the total number of alleles. For the neutrally selective case, the average number of generations until A_1's fixation (t_1) or deletion (t_0) is given by [2]:

$$t_1(p) = -\frac{1}{p}\left[4N_e(1-p)\log(1-p)\right] \text{ and} \qquad (4)$$

$$t_0(p) = -4N_e\left(\frac{p}{1-p}\right)\log p , \qquad (5)$$

where N_e denotes *effective population size*. For simplicity we take $N_e = N$, meaning all individuals in the population are candidates for reproduction. As $p \to 0$, the boundary condition is given by:

$$t_1(0) = 4N_e . \qquad (6)$$

That is, excluding cases of deletion, an initially rare mutant allele spreads to the entire population in $4N_e$ generations.

One important observation that immediately falls out from the theory is that when fixation ($p = 1$) or deletion ($p = 0$) are reached, variation in the population vanishes: $\mathrm{Var}[\Delta p] = 0$. With no variation there is a homogeneous population, and sampling from this population produces the same homogeneous population. In other words, this establishes fixation and deletion as absorbing states of the stochastic sampling process. Once there, drift stops.

Figure 1 illustrates this, showing both the simulated and theoretically predicted number of generations until fixation occurs for $N = 10$, as well as the predicted time to deletion, for reference. Each simulation was performed for a different initial value of p and averaged over 400 realizations. Using the same methodology as [2], we include only those realizations whose allele reaches fixation.

Different kinds of sampling dynamic have been proposed to simulate the behavior of genetic drift. Simulations for two are shown in the figure. The first uses binomial sampling, producing an initial population of $2N$ uniform random numbers between 0 and 1. An initial probability $1 - p$ is assigned to allele A_1 and probability p to allele A_2. The count i of A_2 in the initial population is incremented for each random number less than p. This represents an individual having the allele A_2 instead of A_1. The maximum likelihood estimate of allele frequency in the initial sample is simply the number of A_2 alleles over the sample length: $p = i/2N$. This estimate of p is then used to generate a new population of offspring, after which we re-estimate the value of p. These steps are repeated each generation until fixation at $p = 1$ or deletion at $p = 0$ occurs.

The second sampling dynamic uses a *pseudo-sampling variable* ξ in lieu of direct sampling [18]. The allele frequency p_{n+1} in the next generation is calculated by adding ξ_n to the current frequency p_n:

$$p_{n+1} = p_n + \xi_n \ , \tag{7}$$

$$\xi_n = \sqrt{3\,\sigma_n^2}(2r_n - 1) \ , \tag{8}$$

where σ_n^2 is the current variance given by Eq. (3) and r_n is a uniform random number between 0 and 1. This method avoids the binomial sampling process, sacrificing some accuracy for faster simulations and larger populations. As Fig. 1 shows for fixation, this method (MC w/ PSV, there) overestimates the time to fixation and deletion.

Kimura's theory and simulations predict the time to fixation or deletion of a mutant allele in a finite population by the process of genetic drift. The Fisher-Wright model and Kimura's theory assume a memoryless population in which each offspring inherits allele A_1 or A_2 via an IID binomial sampling process. We now generalize this to memoryful stochastic processes, giving a new definition of fixation and exploring examples of structural drift behavior.

V. SEQUENTIAL LEARNING

How can genetic drift be a memoryful stochastic process? Consider a population of N individuals. Each generation consists of $2N$ alleles and so is represented by a string of $2N$ symbols, e.g. $A_1A_2 \ldots A_1A_1$, where each symbol corresponds to an individual with a particular allele. In the original drift models, a generation of offspring is produced by a memoryless binomial sampling process, selecting an offspring's allele from a parent with replacement. In contrast, the structural drift model produces a generation of individuals in which the sample order is tracked. The population is now a string of alleles, giving the potential for memory and structure in sampling—temporal interdependencies between individuals within a sample or some other aspect of a population's organization. (Later, we return to give several examples of alternative ordered-sampling processes.)

The model class we select to describe memoryful sampling consists of ϵ-machines, since each is a unique, minimal, and optimal representation of a stochastic process [19, 20]. More to the point, ϵ-machines give a systematic representation of all the stochastic processes we consider here. As will become clear, these properties give an important advantage when analyzing structural drift, since they allow one to monitor the amount of structure innovated or lost during drift, as we will show. We next give a brief overview of ϵ-machines and refer the reader to the previous references for details.

ϵ-Machine representations of the finite-memory discrete-valued stochastic processes we consider here form a class of unifilar probabilistic finite-state machine. An ϵ-machine consists of a set of *causal states* $\boldsymbol{\mathcal{S}} = \{0, 1, \ldots, k-1\}$ and a set of transition matrices:

$$\{T_{ij}^{(a)} : a \in \mathcal{A}\} \ , \tag{9}$$

where $\mathcal{A} = \{A_1, \ldots, A_m\}$ is the set of alleles and where the transition probability $T_{ij}^{(a)}$ gives the probability of transitioning from causal state $\mathcal{S}_i$ to causal state $\mathcal{S}_j$ and emitting allele a. Maintaining our connection to diploid theory, we think of an ϵ-machine as a generator of populations or length-$2N$ strings: $\alpha^{2N} = a_1a_2 \ldots a_i \ldots a_{2N}$, $a_i \in \mathcal{A}$. As a model of a sampling process, an ϵ-machine gives the most compact representation of the distribution of strings produced by sampling.

We are now ready to describe *sequential learning*. We begin by selecting an initial population generator M_0—an ϵ-machine. A random walk through an M_0, guided by its transition probabilities, generates a length-$2N$ string $\alpha_0^{2N} = \alpha_1 \ldots \alpha_{2N}$ that represents the first generation of N individuals possessing alleles $a_i \in \mathcal{A}$. We then infer an ϵ-machine M_1 from the population α_0^{2N}. M_1 is then used to produce a new population α_1^{2N}, from which a new ϵ-machine M_2 is estimated. (We describe alternative inference procedures shortly.) This new population has the same allele distribution as the previous, plus

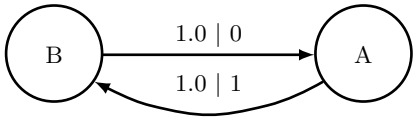

FIG. 2: ϵ-Machine for the Alternating Process, consisting of two causal states $\boldsymbol{\mathcal{S}} = \{A, B\}$ and two transitions. Each transition is labeled $p \mid a$ to indicate the probability $p = T_{ij}^{(a)}$ of taking that transition and emitting allele $a \in \mathcal{A}$. State A generates allele 0 with probability one and transitions to state B, while B generates allele 1 with probability one and transitions to A.

some amount of variance. The cycle of inference and reinference is repeated while allele frequencies drift between generations and until fixation or deletion is reached. At that point, the populations (and so ϵ-machines) cannot vary further. The net result is a stochastically varying time series of ϵ-machines—$M_0, M_1, M_2, \ldots$—that terminates when the populations α_t^N generated stop changing.

Thus, at each step a new representation or model is estimated from the previous step's sample. The inference step highlights that this is learning: a model of the generator is estimated from the given data. The repetition of this step creates a sequential communication chain. Said simply, sequential learning is closely related to genetic drift, except that sample order is tracked and this order is used in estimating the next model.

The procedure is analogous to flipping a biased coin a number of times, estimating the bias from the results, and re-flipping the newly biased coin. Eventually, the coin will be completely biased towards HEADS or TAILS. In our drift model the coin is replaced by an ϵ-machine, which removes the IID constraint and allows for the sampling process to take on structure and memory. Not only do the transition probabilities $T_{ij}^{(a)}$ change, but the structure of the model itself—number of states and transitions—drifts over time.

Before we can explore this dynamic, we first need to examine how an ϵ-machine reaches fixation or deletion.

VI. STRUCTURAL STASIS

Consider the *Alternating Process*—a binary process that alternately generates 0s and 1s. The ϵ-machine for this process, shown in Fig. 2, generates the strings $0101\ldots$ and $1010\ldots$ depending on the start state.

Regardless of the start state, the ϵ-machine is reinferred from any sufficiently long string it generates. In the context of sequential learning, this means the population at each generation is the same. However, if we consider allele A_1 to be represented by symbol 0 and A_2 by symbol 1, neither allele reaches fixation or deletion according to current definitions, which require homogeneous populations. Nonetheless, the Alternating Process prevents any variance between generations and so, despite the population not being all 0s or all 1s, the population does reach an equilibrium: half 0s and half 1s.

For these reasons, one cannot use the original definitions of fixation and deletion. This leads us to introduce *structural stasis* to combine the notions of fixation, deletion, and the inability to vary caused by periodicity. However, we need a method to detect the occurrence of structural stasis in a drift process.

A state machine representing a periodic sampling process enforces the constraint of periodicity via its internal memory. One measure of this memory is the *population diversity* $H(N)$ [21]:

$$H(N) = H[\mathcal{A}_1 \ldots \mathcal{A}_{2N}] \ , \tag{10}$$

$$= -\sum_{a^{2N} \in \mathcal{A}^{2N}} \Pr(a^{2N}) \log_2 \Pr(a^{2N}) \ , \tag{11}$$

where the units are [bits]. (For background on information theory, as used here, the reader is referred to Ref. [22].)

The population diversity of the Alternating Process is $H(N) = 1$ bit at any size $N \geq 1$. This single bit of information corresponds to the sampling process's current phase or state. The population diversity does not change with N, meaning the Alternating Process is always in structural stasis. However, using population diversity as a condition for detecting stasis fails for an arbitrary sampling process.

Said more directly, structural stasis corresponds to a sampling process becoming nonstochastic, since it ceases to introduce variance between generations and so prevents further drift. The condition for stasis could be given as the vanishing (with size N) of the growth rate, $H(N) - H(N-1)$, of population diversity. However, this can be difficult to estimate accurately for finite population sizes.

A related, alternate condition for stasis that avoids these problems uses the entropy rate of the sampling process. We call this *allelic entropy*:

$$h_\mu = \lim_{N \to \infty} \frac{H(N)}{2N} \ , \tag{12}$$

where the units are [bits per allele]. Allelic entropy gives the average information per allele in bits, and structural stasis occurs when $h_\mu = 0$.

This quantity is also difficult to estimate from population samples since it relies on an asymptotic estimate of the population diversity. However, in structural drift we have the ϵ-machine representation of the sampling process. Due to the ϵ-machine's unifilarity, the allelic entropy can be calculated in closed-form over the causal states and their transitions:

$$h_\mu = -\sum_{S \in \mathcal{S}} \Pr(S) \sum_{a \in \mathcal{A}, S' \in \mathcal{S}} T_{SS'}^{(a)} \log_2 T_{SS'}^{(a)} \ . \tag{13}$$

When $h_\mu = 0$, the sampling process has become periodic and lost all randomness generated via its branching transitions. This new criterion simultaneously captures

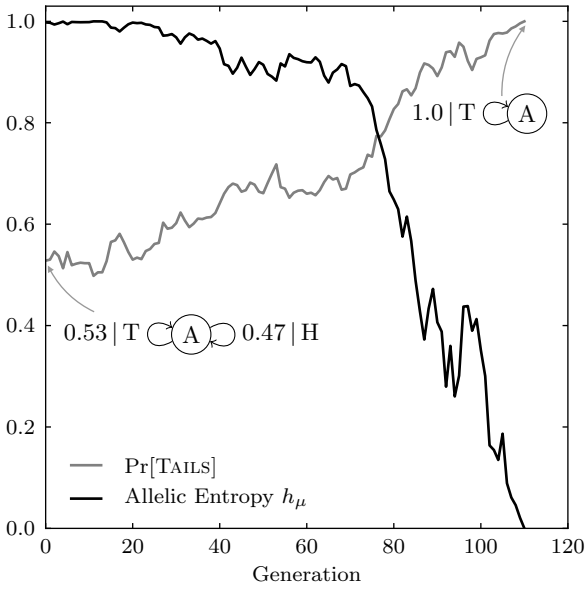

FIG. 3: Drift of allelic entropy h_μ and Pr[Tails] for a single realization of the Biased Coin Process with sample length $2N = 100$ and state splitting reconstruction ($\alpha = 0.01$).

the notions of fixation and deletion, as well as periodicity. An ϵ-machine has zero allelic entropy if any of these conditions occur. More formally, we have the following statement.

Definition. Structural stasis *occurs when the sampling process's allelic entropy vanishes:* $h_\mu = 0$.

Proposition 1. *Structural stasis is a fixed point of finite-memory structural drift.*

Proof. *Finite-memory means that the ϵ-machine representing the population sampling process has a finite number of states. Given this, if $h_\mu = 0$, then the ϵ-machine has no branching in its recurrent states: $T^{(a)}_{ij} = 0$ or 1, where $\mathcal{S}_i$ and $\mathcal{S}_j$ are asymptotically recurrent states. This results in no variation in the inferred ϵ-machine when sampling sufficiently large populations. Lack of variation, in turn, means that $\Delta p = 0$ and so the drift process stops. Since no mutations are allowed, a further consequence is that, if allelic entropy vanishes at time t, then it is zero for all $t' > t$. Thus, structural stasis is an absorbing state of the drift stochastic process.*

VII. EXAMPLES

While more can be said analytically about structural drift, our present purpose is to introduce the main concepts. We will show that structural drift leads to interesting and nontrivial behavior. First, we calibrate the new class of drift processes against the original genetic drift theory.

A. Memoryless Drift

The Biased Coin Process is represented by a single-state ϵ-machine with a self loop for both Heads and Tails symbols. It is an IID sampling process that generates populations with a binomial distribution. Unlike the Alternating Process, the coin's bias p is free to drift during sequential inference. These properties make the Biased Coin Process an ideal candidate for exploring memoryless drift.

Two measures of the structural drift of a single realization of the Biased Coin Process are shown in Fig. 3, with initial $p = \Pr[\text{Tails}] = 0.53$. Structural stasis ($h_\mu = 0$) is reached after 115 generations. Note that the drift of allelic entropy h_μ and $p = \Pr[\text{Tails}]$ are inversely related, with allelic entropy converging quickly to zero as stasis is approached. This reflects the rapid drop in population diversity. The initial Fair Coin ϵ-machine is shown at the left of Fig. 3, and the final completely biased ϵ-machine shown in the lower right. After stasis occurs, all randomness has been eliminated from the transitions at state A, resulting in a single transition that always produces Tails. Anticipating later discussion, we note that during the run one only sees Biased Coin Processes.

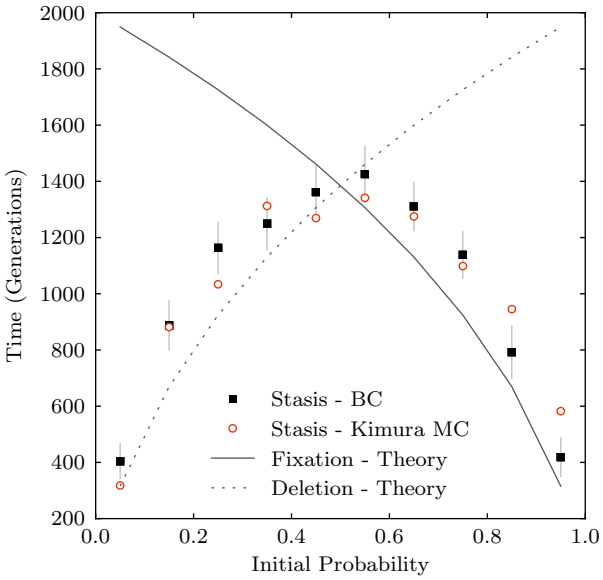

FIG. 4: Average time to stasis as a function of initial Pr[Heads] for the Biased Coin Process: Structural Drift simulation (solid squares) and Monte Carlo simulation of Kimura's equations (hollow circles). Both employ vanishing allelic entropy at stasis. Kimura's predicted times to fixation and deletion are shown for reference. Each estimated time is averaged over 100 drift experiments with sample length $2N = 1000$ and state-splitting reconstruction ($\alpha = 0.01$).

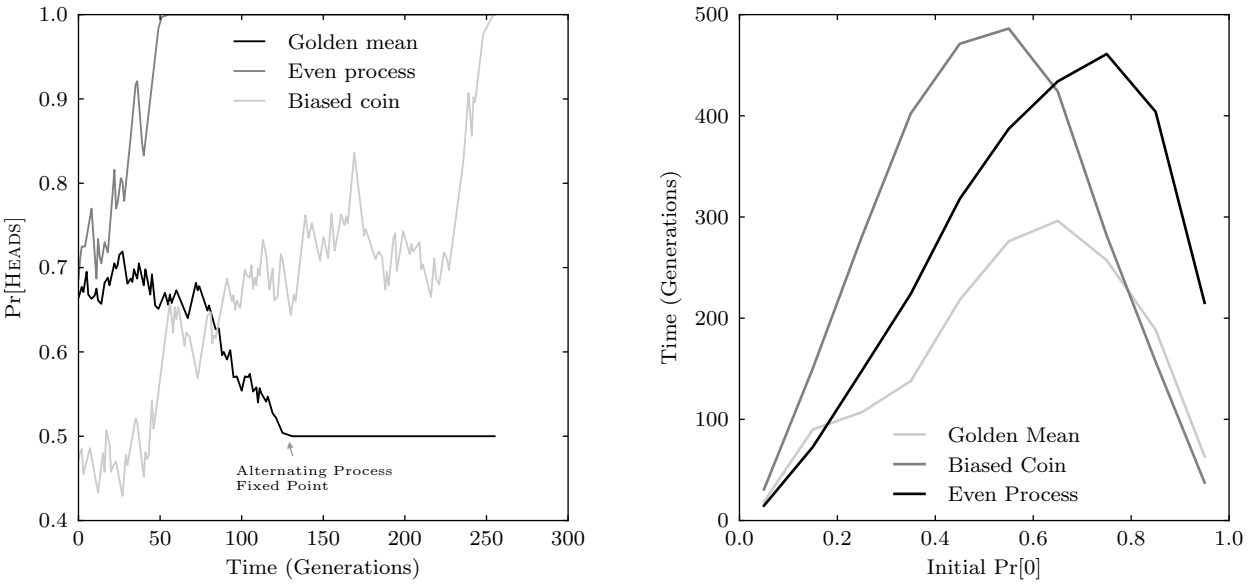

FIG. 5: Comparison of structural drift processes. *Left*: Pr[HEADS] for the Biased Coin, Golden Mean, and Even Processes as a function of generation. The Even and Biased Coin Processes become completely biased coins at stasis, while the Golden Mean becomes the Alternating Process. Note that the definition of structural stasis recognizes the lack of variance in that periodic-process subspace, even though the allele probability is neither 0 nor 1. *Right*: Time to stasis as a function of initial bias parameter for each process. Each estimated time is averaged over 100 drift experiments with sample length $2N = 1000$ and state-splitting reconstruction ($\alpha = 0.01$).

The time to stasis of the Biased Coin Process as a function of initial $p = \Pr[\text{HEADS}]$ is shown in Fig. 4. Also shown is Kimura's previous Monte Carlo drift simulation modified to terminate when either fixation or deletion occurs. This experiment, with a 100 times larger population than Fig. 1, illustrates the definition of structural stasis and allows direct comparison of structural drift with genetic drift in the memoryless case.

Not surprisingly, we can interpret genetic drift as a special case of the structural drift process for the Biased Coin.[3] Both simulations follow Kimura's theoretically predicted curves, combining the lower half of the deletion curve with the upper half of the fixation curve to reflect the initial probability's proximity to the absorbing states. A high or low initial bias leads to a shorter time to stasis as the absorbing states are closer to the initial state. Similarly, a Fair Coin is the furthest from absorption and thus takes the longest average time to reach stasis.

[3] Simulations used both the causal-state splitting [20] and subtree merging [23] ϵ-machine reconstruction algorithms, with approximately equivalent results. Unless otherwise noted, state splitting used a history length of 3, and tree merging used a morph length of 3 and tree depth of 7. Both algorithms used a significance-test value of $\alpha = 0.01$ in the Kolmogorov-Smirnov hypothesis tests for state equivalence. (Other tests, such as χ^2, may be substituted with little change in results.) For the Biased Coin Process, a history length of 1 was used for direct comparison to binomial sampling.

B. Structural Drift

The Biased Coin Process is an IID sampling process with no memory of previous flips, reaching stasis when $\Pr[\text{HEADS}] = 1.0$ or 0.0 and, correspondingly, when $h_\mu(M_t) = 0.0$. We now introduce memory by starting drift with M_0 as the *Golden Mean Process*, which produces binary populations with no consecutive 0s. Its ϵ-machine is shown in Fig. 6.

Like the Alternating Process, the Golden Mean Process has two causal states. However, the transitions from state A have nonzero entropy, allowing their probabilities to drift as new ϵ-machines are inferred from generation to generation. If the $A \to B$ transition parameter p (Fig. 6) drifts towards zero probability and is eventually removed, the Golden Mean reaches stasis by transforming into the Fixed Coin Process identical to that shown at the top right of Fig. 3. Instead, if the same transition drifts towards probability $p = 1$, the $A \to A$ transition is removed. In this case, the Golden Mean Process reaches stasis by transforming into the Alternating Process (Fig. 2).

Naturally, one can start drift from any one of a number of processes. Let's also consider the *Even Process* and then compare drift behaviors. Similar in form to the Golden Mean Process, the Even Process produces populations in which blocks of consecutive 1s must be even in length when bounded by 0s.

Figure 5 (left) compares the drift of Pr[HEADS] for single runs starting with the Biased Coin, Golden Mean,

and Even Processes. One observes that the Biased Coin and Even Processes reach stasis via the Biased Coin fixed point, while the Golden Mean Process reaches stasis via the Alternating Process fixed point.

It should be noted that the memoryful Golden Mean and Even Processes reach stasis markedly faster than the memoryless Biased Coin. While the left panel of Fig. 5 shows only a single realization of each sampling process type, the right panel shows that the large disparity in stasis times holds across all settings of each process's initial bias. This is one of our first general observations about memoryful processes: The structure of memoryful processes substantially impacts the average time to stasis by increasing variance between generations.

VIII. ISOSTRUCTURAL SUBSPACES

To illustrate some of the richness of structural drift and to understand how it affects average time to stasis, we examine the complexity-entropy (CE) diagram [24] of the ϵ-machines produced over several realizations of an arbitrary sampling process. The CE diagram displays how the allelic entropy h_μ of an ϵ-machine varies with its *allelic complexity* C_μ:

$$C_\mu = -\sum_{\sigma \in \mathcal{S}} \Pr(\sigma) \log_2 \Pr(\sigma) \ , \tag{14}$$

where the units are [bits]. The allelic complexity is the Shannon entropy over an ϵ-machine's stationary state distribution $\Pr(\mathcal{S})$. It measures the memory needed to maintain the internal state while producing stochastic outputs. ϵ-Machine minimality guarantees that C_μ is the smallest amount of memory required to do so. Since there is a one-to-one correspondence between processes and their ϵ-machines, a CE diagram is a projection of process space onto the two coordinates (h_μ, C_μ). Used in tandem, these two properties differentiate many types of sampling process, capturing both their intrinsic memory (C_μ) and the diversity (h_μ) of populations they generate.

Let's examine the structure of drift-process space further. The CE diagram for 100 realizations starting with the Golden Mean Process is shown in the left panel of Fig. 7. The M_t reach stasis by transforming into either the Fixed Coin Process or the Alternating Process,

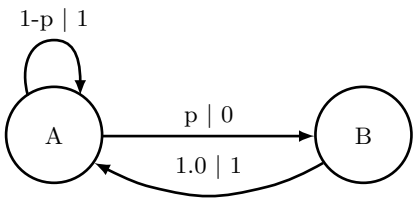

FIG. 6: The ϵ-machine for the Golden Mean Process, which generates a population with no consecutive 0s. In state A the probabilities of generating a 0 or 1 are p and $1-p$, respectively.

depending on how the transition parameter p (Fig. 6) drifts. The Fixed Coin Process, all 1s or all 0s, exists at $(h_\mu, C_\mu) = (0, 0)$ since a process in stasis has no allelic entropy and no memory is required to track a single state. The Golden Mean Process transforms into the Fixed Coin at $p = 0$. The Alternating Process exists at point $(h_\mu, C_\mu) = (0, 1)$ as it also has no allelic entropy, but requires 1 bit of information storage to track the phase of its 2 states. The Golden Mean Process becomes the Alternating Process when $p = 1$. The two stasis points are connected by the isostructural curve $(h_\mu(M(p)), C_\mu(M(p))$, where $M(p)$ is the Golden Mean ϵ-machine of Fig. 6 with $p \in [0, 1]$.

What emerges from examining these overlapping realizations is a broad view of how the structure of M_t drifts in process space. Roughly, they diffuse locally in the parameter space specified by the current, fixed architecture of states and transitions. During this, transition probability estimates vary stochastically due to sampling variance. Since C_μ and h_μ are continuous functions of the transition probabilities, this variance causes the M_t to fall on well defined curves or regions corresponding to a particular process subspace. (See Figs. 4 and 5 in [24] and the theory for these curves and regions there.) We refer to the associated sets of ϵ-machines as *isostructural subspaces*. They are metastable subspaces of sampling processes that are quasi-invariant under the structural drift dynamic. That invariance is broken by jumps between the subspaces in which one or more ϵ-machine parameters diffuse sufficiently that inference is forced to shift ϵ-machine topology—that is, states or transitions are gained or lost.

Such a shift occurs in the lower part of the Golden Mean isostructural curve, where states A and B merge into a single state as transition probability $p \to 0$, corresponding to $(h_\mu, C_\mu) \approx (0.5, 0.5)$. The exact location on the curve where this discontinuity occurs is controlled by the significance level (α), described earlier, at which two causal states are determined to be equivalent by a statistical hypothesis test. Specifically, states A and B are merged closer to point $(h_\mu, C_\mu) = (0, 0)$ along the Golden Mean Process isostructural curve when the hypothesis test requires more evidence.

When causal-state merging occurs, the ϵ-machine leaves the Golden Mean subspace and enters the Biased Coin subspace. In the CE diagram, the latter is the one-dimensional interval $C_\mu = 0$ and $h_\mu \in [0, 1]$. In the new subspace, the time to stasis depends only on the entry value of p. For comparison, the right panel of Fig. 7 shows a CE diagram of 100 realizations starting with the fair Biased Coin Process. The process diffuses along the line $C_\mu = 0$, never innovating a new state or jumping to a new subspace. This demonstrates that movement between subspaces is often not bidirectional—innovations from a previous topology may be lost either temporarily (when the innovation can be restored by returning to the subspace) or permanently. For example, the Golden Mean Process commonly jumps to the Biased Coin, but

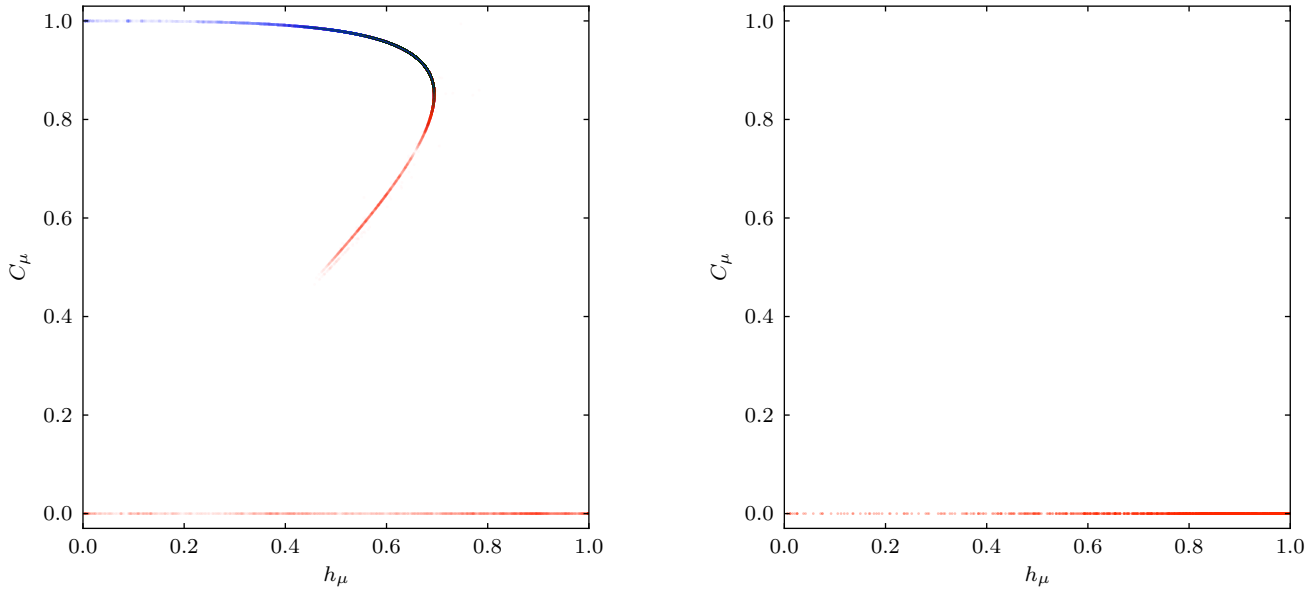

FIG. 7: *Left:* Allelic complexity C_μ versus allelic entropy h_μ for 100 realizations starting with the Golden Mean Process at $p_0 = \frac{1}{2}$ and $(h_\mu, C_\mu) \approx (\frac{2}{3}, 0.918)$, showing time spent in the Alternating and Biased Coin subspaces. *Right:* Drift starting with the Biased Coin Process with initially fair transition probabilities: $p_0 = \frac{1}{2}$ and $(h_\mu, C_\mu) = (1, 0)$. Point density in the higher-h_μ region indicates longer times to stasis compared to drift starting with the Golden Mean Process, which enters the Biased Coin subspace nearer the fixed point $(0, 0)$, jumping around $(h_\mu, C_\mu) \approx (0.5, 0.5)$. Red-colored dots correspond to M_t that go to stasis as the Fixed Coin Process; blue correspond to those which end up in stasis as the Alternating Process. Each of the 100 runs used sample length $2N = 1000$ and state-splitting reconstruction ($\alpha = 0.01$).

the opposite is observed to be improbable.

All realizations eventually find their way to structural stasis on the CE diagram's left boundary at absorbing states $(h_\mu, C_\mu) = (0, \log_2 P)$ with periods P. Nonetheless, it is clear that the Golden Mean process, which starts at $(h_\mu, C_\mu) = (\frac{1}{2}, 1)$, leads the drift to innovate M_ts with substantially more allelic entropy and complexity.

In addition to the CE diagram helping to locate fixed points and quasi-invariant subspaces, the density of points on the isostructural curves gives an alternate view of the time to stasis plots (Fig. 4). Dense regions on the curve correspond to initial p values "furthest away" from the fixed points. ϵ-Machines typically spend longer diffusing on these portions of the curve, resulting in longer stasis times and a higher density of neighboring ϵ-machines.

The difference in densities between the post-jump Biased Coin subspace of the Golden Mean (left) and the initially fair Biased Coin subspace (right) highlights that the majority of time spent drifting in the latter is near high-entropy, initially fair values of p. Golden Mean M_ts jump into the Biased Coin subspace only after a state merging has occurred due to highly biased, low-entropy transition probabilities. This causes the M_t to arrive in the subspace nearer an absorbing state, resulting in a shorter average time to stasis. This is a consequence of the Golden Mean's structured process subspace. However, once the Biased Coin subspace is reached, the time to stasis from that point forward is independent of the time spent in the previous isostructural subspace. It is determined by the effective transition parameters of the M_t at time of entry.

Figure 8 demonstrates how the total stasis times decompose, accounting for the total time as weighted sum of the average stasis times of its pathways. A *pathway* is a set of subspaces visited by all realizations reaching a particular fixed point. The left panel shows time to stasis $T_s(GMP(p_0))$ for starting the Golden Mean Process with initial transition probability $p = p_0$ as the weighted sum of the time spent diffusing in the Alternating and Fixed Coin pathways:

$$T_s(GMP(p_0)) = \sum_{\gamma \in \{FC, AP\}} k_\gamma T_s(\gamma | GMP(p_0)) \ , \quad (15)$$

where $k_\gamma \in [0, 1]$ are the weight coefficients and $T_s(\gamma|\cdot)$ is the time to reach each terminal (stasis) subspace γ, having started in the Golden Mean Process with $p = p_0$.

For low p_0, the transition from state A to state B is unlikely, so zeros are rare and the Alternating Process pathway is taken infrequently. Thus, the total stasis time is initially dominated by the Fixed Coin pathway. As $p_0 \to 0.3$ and above, the Alternating Process pathway becomes more frequent and this stasis time begins to contribute to the total. Around $p_0 = 0.6$ the Fixed

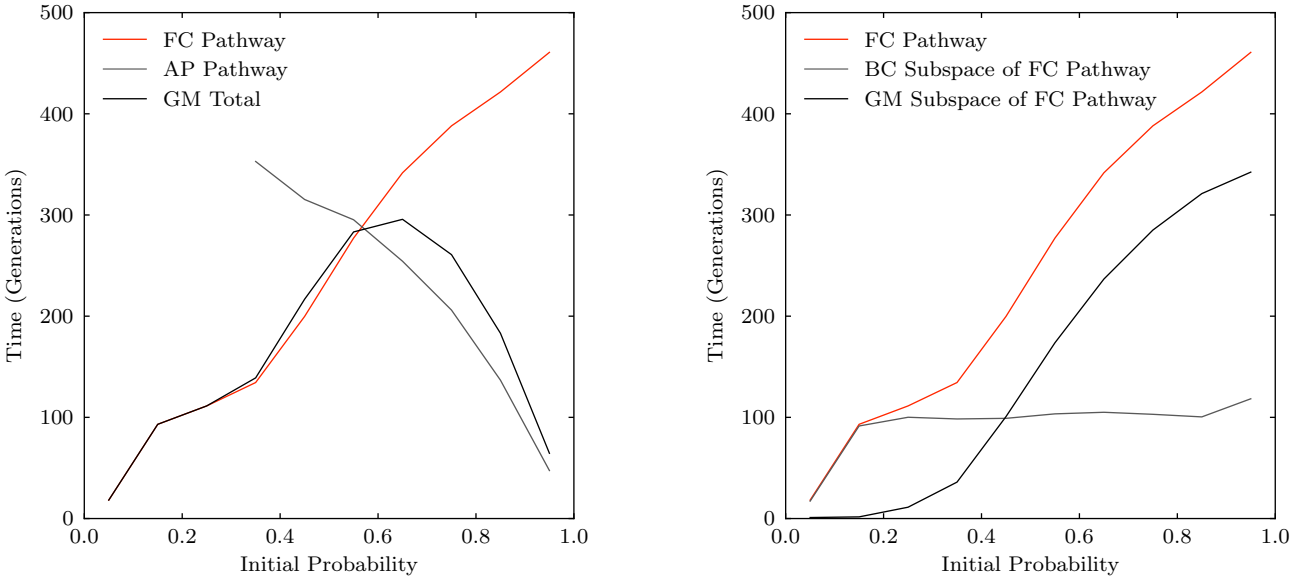

FIG. 8: *Left*: Mean time to stasis for the Golden Mean Process as a function of initial transition probability p. The total time to stasis is the sum of stasis times for the Fixed Coin pathway and the Alternating Process pathway, weighted by their probability of occurrence. For initial p less than ≈ 0.3, the Alternating Process pathway was not observed during simulation due to its rarity. As initial p increases, the Alternating pathway is weighted more heavily while the Fixed Coin pathway occurs less frequently. *Right*: Mean time to stasis for the Fixed Coin pathway as function of initial transition probability p. The total time to stasis is the sum of the pathway's subspace stasis times. The Fixed Coin pathway visits the Golden Mean subspace before jumping to the Biased Coin subspace, on its way to stasis as the Fixed Coin. Each estimated time is averaged over 100 drift experiments with sample length $2N = 1000$ and state-splitting reconstruction ($\alpha = 0.01$).

Coin pathway becomes less likely and the total time becomes dominated by the Alternating Process pathway.

Time to stasis for a particular pathway is simply the sum of the times spent in the subspaces it connects. Figure 8's right panel examines the left panel's Fixed Coin pathway time $T_s(FC|GMP(p_0))$ in more detail. There are two contributions. These include the time diffusing on the portions of the Golden Mean subspace before the subspace jump, as well as the time attributed to the Biased Coin subspace *after* the subspace jump. Note that the Biased Coin subspace stasis time is independent of p_0, because the subspace is entered when the states A and B are merged. This merging occurs at the same p regardless of p_0. There is a dip for values of p_0 that are lower than the state-merging threshold for p, placing the initial subspace bias even closer to its absorbing state.

Thus, the time to reach each stasis from a subspace β consists of the times taken on pathways $c \in \beta$ to structural stasis that can be reached from β. As a result, the total time to stasis starting in β is the sum of each pathway c's stasis time $T_s(c|\beta)$ weighted by the pathway's likelihood $\Pr(c|\beta)$ starting from β:

$$T_s(\beta) = \sum_{c \in \beta} \Pr(c|\beta)\ T_s(c|\beta)\ , \tag{16}$$

where the probabilities and times depend implicitly on the initial process's transition parameter(s).

IX. DISCUSSION

A. Summary

The Fischer-Wright model of genetic drift can be viewed as a random walk of coin biases, a stochastic process that describes generational change in allele frequencies based on a strong statistical assumption: the sampling process is memoryless. Here, we developed a generalized structural drift model that adds memory to the process and examined the consequences of such population sampling memory.

The representation selected for the population sampling mechanism was the class of probabilistic finite-state machines called ϵ-machines. We discussed how a sequential chain of inferring and re-inferring ϵ-machines from the finite data they generate parallels the drift of alleles in a finite population, using otherwise the same assumptions made by the Fischer-Wright model.

We revisited Kimura's early results measuring the time to fixation of drifting alleles and showed that the generalized structural drift process reproduces these well known results, when staying within the memoryless sampling process subspace. Starting with populations outside of that subspace led the sampling processes to exhibit memory effects, including greatly reduced times to stasis, structurally complex transients, structural innovation, and structural decay. We introduced structural sta-

sis to combine the concepts of deletion, fixation, and periodicity for drift processes. Generally, structural stasis occurs when the population's allelic entropy vanishes—a quantity one can calculate in closed form due to the use of the ϵ-machine representation for sampling processes.

Simulations demonstrated how an ϵ-machine diffuses through isostructural process subspaces during sequential learning. The result was a very complex time-to-stasis dependence on the initial probability parameter—much more complicated than Kimura's theory describes. We showed, however, that a process's time to stasis can be decomposed into sums over these independent subspaces. Moreover, the time spent in an isostructural subspace depends on the value of the ϵ-machine probability parameters at the time of entry. This suggests an extension to Kimura's theory for predicting the time to stasis for each isostructural component independently. Much of the phenomenological analysis was facilitated by the global view of drift process space given by the complexity-entropy diagram.

Drift processes with memory generally describe the evolution of structured populations without mutation or selection. Nonetheless, we showed that structure leads to substantially shorter stasis times. This was seen in drifts starting with the Biased Coin and Golden Mean Processes, where the Golden Mean jumps into the Biased Coin subspace close to an absorbing state. This demonstrated that even without selection, population structure and sampling memory matter in evolutionary dynamics. It also demonstrated that memoryless models severely restrict sequential learning, leading to overestimates of the time to stasis.

Finally, we should stress that none of these phenomena occur in the limit of infinite populations or sample size. The variance due to finite sampling drives sequential learning, the diffusion through process space, and the jumps between isostructural subspaces.

B. Applications

Structural drift gives an alternative view of drift processes in population genetics. In light of new kinds of evolutionary behavior, it reframes the original questions about underlying mechanisms and extends their scope to phenomena that exhibit memory in the sampling process. Examples of the latter include environmental toxins [25], changes in predation [26], and socio-political factors [27] where memory lies in the spatial distribution of populations. In addition to these, several applications to areas beyond population genetics proper suggest themselves.

1. Epochal Evolution

An intriguing parallel exists between structural drift and the longstanding question about the origins of *punctuated equilibrium* [28] and the dynamics of *epochal evolution* [29]. The possibility of evolution's intermittent progress—long periods of stasis punctuated by rapid change—dates back to Fisher's demonstration of metastability in drift processes with multiple alleles [13].

Epochal evolution presents an alternative to the view of metastability posed by adaptive landscapes [30]. Within epochal evolutionary theory, equivalence classes of genotype fitness, called *subbasins*, are connected by fitness-changing *portals* to other subbasins. A genotype is free to diffuse within its subbasin via selectively neutral mutations, until an advantageous mutation drives genotypes through a portal to a higher-fitness subbasin. An increasing number of genotypes derive from this founder and diffuse in the new subbasin until another portal to higher fitness is discovered. Thus, the structure of the subbasin-portal architecture dictates the punctuated dynamics of evolution.

Given an adaptive system which learns structure by sampling its past organization, structural drift theory implies is that its evolutionary dynamics are inevitably described by punctuated equilibria. Diffusion in an isostructural subspace corresponds to a period of structured equilibrium and subspace shifts correspond to rapid innovation or loss of organization.

Thus, structural drift establishes a connection between evolutionary innovation and structural change and identifies the conditions for creation or loss of innovation. This suggests that there is a need to bring these two theories together by adding mutation and selection to structural drift.

2. Graph Evolution

The evolutionary dynamics of structured populations have been studied using undirected graphs to represent correlation between individuals. Edge weights w_{ij} between individuals i and j give the probability that i will replace j with its offspring when selected to reproduce.

By studying fixation and selection behavior on different types of graphs, Lieberman et al found, for example, that graph structures can sometimes amplify or suppress the effects of selection, even guaranteeing the fixation of advantageous mutations [31].

Jain and Krishna [32] investigated the evolution of directed graphs and the emergence of self-reinforcing autocatalytic networks of interaction. They identified the attractors in these networks and demonstrated a diverse range of behavior from the creation of structural complexity to its collapse and permanent loss.

Graph evolution is a complementary framework to structural drift. In the latter, graph structure evolves over time with nodes representing equivalence classes of the distribution of selectively neutral alleles. Additionally, unlike ϵ-machines, the multinomial sampling of individuals in graph evolution is a memoryless process. A combined approach would allow one to examine how amplification and suppression are affected by external in-

fluences on the population structure; for example, including how a population might use temporal memory to maintain desirable properties in anticipation of structural shifts in the environment.

3. Molecular Clocks

The notion that evolutionary changes occur at regular time intervals was introduced more than 40 years ago by Zuckerkandl et al [33]. Such regularity, when it holds, allows one to estimate the date of common ancestry for two divergent species. Kimura's theory of neutral evolution lent support to the idea of molecular clocks by stating that most selectively neutral single-nucleotide mutations accumulate at the same rate across species due to fixed error rates in DNA replication [34].

Molecular clocks have been controversial due to uncertainties about the molecular mechanisms on which they rely and due to the discovery of varying mutation rates. Several modifications to the theory were proposed, though none is considered universally satisfactory [35]. Schwartz et al proposed that regular changes in germ cells are "not, in our present understanding of cell biology, tenable" [36] and, instead, they suggested evolutionary changes happen suddenly and without clock-like regularity.

The phenomena of epochal evolution and structural drift offer alternative views of the evolutionary dynamics underlying molecular clocks, modeling periods of stable diffusion punctuated by rapid change. Specifically, the architecture of generalized drift process space could dictate how and where structured populations could have diverged in the past. Indeed, here we demonstrated that structural drift can substantially alter times to stasis. Epochal evolution and structural drift also challenge us, however, to design experiments for testing if their mechanisms operate during evolutionary change. In vitro evolutionary experiments with bacteria [37] seem particularly amenable to this kind of validation.

4. Sequential Learning in Communication Chains

Let's briefly return to our motivating problem of learning in chains of sequentially coupled communication channels. In the drift behaviors explored above, the M_T went to stasis ($h_\mu = 0$) corresponding to periodic formal languages. Clearly, such a long-term condition falls short as a model of human communication chains. In the latter, the resulting messages, though distant from those at the beginning of the chain, are not periodic. To more closely capture drift in the context of sequential language learning, structural drift can be extended to include mutation and selection. Then one can investigate how they prevent permanent stasis and allow preference for intelligible phrases.

However, the current framework does capture the language-centric notion of dynamically changing semantics. The symbols and words in the strings generated have a semantics given by the structure of the ϵ-machine [38]. Briefly, causal states provide dynamic contexts for interpretation of individual symbols and words. Moreover, the allelic complexity is the total amount of semantic content that can be generated by an M_t. In this way, changes in the architecture of the M_t during drift correspond to semantic innovation and loss.

X. FINAL REMARKS

Structural drift is amenable to extension and application to a range of biological problems, as was the original Fisher-Wright drift model. Here, we focused on motivating the model, drawing comparisons to Kimura's theory, and explaining the basic mechanisms underlying the resulting phenomena. We will report elsewhere on more technical aspects including a predictive theory and extensions that include mutation and selection. We close by indicating some of the challenging open technical problems.

ϵ-Machine minimality allowed us to monitor the sampling process's information processing, information storage, and causal architecture during drift. However, it may be more realistic to use nonminimal representations in the drift process as populations and sampling processes need not be minimal. This would alter the organization of sampling process space. Still, one would transform these representations to ϵ-machines so that informational properties can be properly and unambiguously measured.

There are various kinds of memory in structural drift with mutation that one must distinguish. On the one hand, a high mutation rate destroys a population's ability to remember its past. A high mutation rate leads to rapid exploration and discovery of beneficial genotypes, but past innovations become corrupted and are rapidly forgotten. On the other hand, a vanishingly small mutation rate leads to an arbitrarily slow evolution process: There are no innovations worth remembering. Thus, population memory derives from balancing these tendencies. This kind of population memory is not necessarily the same as the sampling memory implicated in basic structural drift. Nonetheless, these types of memory interact and this interaction must eventually be understood.

We demonstrated how structural drift—diffusion and structural innovation and loss—are controlled by the architecture of connected isostructural subspaces. Many questions remain about these subspaces. What is the degree of subspace-jump irreversibility? Can we predict the likelihood of these jumps? What does the phase portrait of a drift process look like? Thus, to better understand structural drift, we need to analyze the high-level organization of generalized drift process space.

Fortunately, ϵ-machines are in one-to-one correspondence with structured processes. Thus, the preced-

ing question reduces to understanding the space of ϵ-machines and how they can be connected by diffusion processes. Is the diffusion within each process subspace predicted by Kimura's theory or some simple variant? We have given preliminary evidence that it does. And so, there are reasons to be optimistic that in face of the original complicatedness of structural drift, a good deal can be predicted analytically. And this, in turn, will lead to quantitative applications.

Acknowledgments

This work was partially supported by the DARPA Physical Intelligence Program.

[1] C.U.M. Smith. Send reinforcements we're going to advance. *Biology and Philosophy*, 3:214–217, 1988.

[2] M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3):763–771, 1969.

[3] R. M. Felder. Learning and teaching styles in engineering education. *Engr. Education*, 78(7):674–681, 1988.

[4] T. G. Dietterich. Machine learning for sequential data: A review. In T. Caelli, editor, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 2396 of *Lecture Notes in Computer Science*, pages 15–30. Springer-Verlag, 2002.

[5] M. Lowry and G. W. Schwert. IPO market cycles: Bubbles or sequential learning? *Journal of Finance*, LXVII(3):1171–1198, 2002.

[6] E. van Nimwegen, J. P. Crutchfield, and M. A. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, 96:9716–9720, 1999.

[7] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA.*, 103:5869–5874, 2006.

[8] A. Raval. Molecular clock on a neutral network. *Phys. Rev. Lett.*, 99:138104, 2007.

[9] J. P. Crutchfield and P. K. Schuster. *Evolutionary Dynamics—Exploring the Interplay of Selection, Neutrality, Accident, and Function*. Santa Fe Institute Series in the Sciences of Complexity. Oxford University Press, 2002.

[10] K. Koelle, S. Cobey, B. Grenfell, and M. Pascual. Epochal evolution shapes the phylodynamics of interpandemic Influenza A (H3N2) in humans. *Science*, 314:1898–1903, 2006.

[11] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, United Kingdom, 1983.

[12] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–126, 1931.

[13] R. A. Fisher. *The genetical theory of natural selection*. Oxford University Press, 1999. Revised reprint of the 1930 original, Edited, with a foreword and notes, by J. H. Bennett.

[14] J. H. Gillespie. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics*, 155:909–919, 2000.

[15] J. H. Gillespie. *Population Genetics: A Concise Guide*. Johns Hopkins University Press, New York, second edition, 2004.

[16] G. Mendel and W. Bateson. *Experiments in plant-hybridisation*. Harvard University Press, 1925.

[17] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, 1968.

[18] M. Kimura. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. *Proc. Natl. Acad. Sci. USA*, 77(1):522–526, 1980.

[19] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[20] C. R. Shalizi, K. L. Shalizi, and J. P. Crutchfield. Pattern discovery in time series, Part I: Theory, algorithm, analysis, and convergence. 2002. Santa Fe Institute Working Paper 02-10-060; arXiv.org/abs/cs.LG/0210025.

[21] E.C. Pielou. The use of information theory in the study of the diversity of biological populations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, number 4, pages 163–177, 1967.

[22] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[23] J. P. Crutchfield and K. Young. Computation at the onset of chaos. In *Complexity, Entropy, and the Physics of Information*, volume VIII, pages 223–269. Addison-Wesley, 1990.

[24] D. P. Feldman, C. S. McTague, and J. P. Crutchfield. The organization of intrinsic computation: Complexity-entropy diagrams and the diversity of natural information processing. *CHAOS*, 18(4):59–73, 2008.

[25] M.H. Medina, J.A. Correa, and C. Barata. Micro-evolution due to pollution: Possible consequences for ecosystem responses to toxic stress. *Chemosphere*, 67(11):2105–2114, 2007.

[26] A. Tremblay, D. Lesbarreres, T. Merritt, C. Wilson, and J. Gunn. Genetic structure and phenotypic plasticity of yellow perch (perca flavescens) populations influenced by habitat, predation, and contamination gradients. *Integrated Environmental Assessment and Management*, 4(2):264–266, 2008.

[27] M. Kayser, O. Lao, K. Anslinger, C. Augustin, G. Bargel, J. Edelmann, S. Elias, M. Heinrich, J. Henke, and L. Henke. Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Human genetics*, 117(5):428–443, 2005.

[28] S. J. Gould and N. Eldredge. Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology*, 3:115–151, 1977.

[29] M. Mitchell, J. P. Crutchfield, and P. T. Hraber. Evolving cellular automata to perform computations: mechanisms and impediments. *Phys. D*, 75(1-3):361–391, 1994.

[30] S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the*

Sixth International Congress on Genetics, pages 355–366, 1932.

[31] E. Lieberman, C. Hauert, and M.A. Nowak. Evolutionary dynamics on graphs. *Nature*, 433:312–316, 2005.

[32] S. Jain and S. Krishna. Graph theory and the evolution of autocatalytic networks. In *Handbook of Graphs and Networks*, pages 355–395. Wiley-VCH Verlag GmbH & Co. KGaA, 2004.

[33] E. Zuckerkandl and L. B. Pauling. Molecular disease, evolution, and genetic heterogeneity. In *Horizons in Biochemistry*, pages 189–225, New York, 1962. Academic Press.

[34] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[35] G. Hermann. Current status of the molecular clock hypothesis. *The American Biology Teacher*, 65(9):661–663, 2003.

[36] J. Schwartz and B. Maresca. Do molecular clocks run at all? A critique of molecular systematics. *Biological Theory*, 1(4):357–371, 2006.

[37] S.F. Elena, V. S. Cooper, and R. E. Lenski. Punctuated evolution caused by selection of rare beneficial mutations. *Science*, 272:1802–1804, 1996.

[38] J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317 – 359, Reading, Massachusetts, 1992. Addison-Wesley.