Evolutionary implementation of Bayesian computations

Dániel Czégel[1,2,3,4], Hamza Giaffar[5], István Zachar[1,3,6], and Eörs Szathmáry[1,2,3]

[1]MTA Centre for Ecological Research, Evolutionary Systems Research Group, Hungarian Academy of Sciences, Tihany, Hungary
[2]Dept. of Plant Systematics, Ecology and Theoretical Biology, Eötvös University, Budapest, Hungary
[3]Parmenides Foundation, Center for the Conceptual Foundations of Science, Pullach/Munich, Germany
[4]Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[6]MTA-ELTE Theoretical Biology and Evolutionary Ecology Research Group, Eötvös University, Budapest, Hungary

Abstract

A wide variety of human and non-human behavior is computationally well accounted for by probabilistic generative models, formalized consistently in a Bayesian framework. Recently, it has been suggested that another family of adaptive systems, namely, those governed by Darwinian evolutionary dynamics, are capable of implementing building blocks of Bayesian computations. These algorithmic similarities rely on the analogous competition dynamics of generative models and of Darwinian replicators to fit possibly high-dimensional and stochastic environments. Identified computational building blocks include Bayesian update over a single variable and replicator dynamics, transition between hidden states and mutation, and Bayesian inference in hierarchical models and multilevel selection. Here we provide a coherent mathematical discussion of these observations in terms of Bayesian graphical models and a step-by-step introduction to their evolutionary interpretation. We also extend existing results by adding two missing components: a correspondence between likelihood optimization and phenotypic adaptation, and between expectation-maximization-like dynamics in mixture models and ecological competition. These correspondences suggest a deeper algorithmic analogy between evolutionary dynamics and statistical learning, pointing towards a unified computational understanding of mechanisms Nature invented to adapt to high-dimensional and uncertain environments.

1 Introduction

The Bayesian framework of performing consistent computations about probabilistic beliefs has become a fundamental language to describe and understand adaptation in the presence of uncertainty [1, 2, 3]. Two factors have recently driven the theoretical development of this machinery: i) the machine learning revolution building on the boom of available data and computational power [4, 5] and ii) the possibility to perform large-scale (human and non-human) cognitive experiments that necessitates a modeling framework accounting for trial-to-trial and inter-individual variability in a probabilistic manner [6] . A Bayesian approach, building on the concept of belief update driven by external evidence, seems to provide a modeling framework that possess just the appropriate amount of richness to make it extremely useful in building adaptive machines by either human designers or by the "blind" processes of inheritance and natural selection [7]. One major advantage of Bayesian models is that they naturally translate between a generative process that predicts data given the current model and the reverse process of inference that updates the model given new data. This allows for online learning, naturally accounting for the optimal level of uncertainty on-the-fly given the amount of data that is available [8].

The fact that many efficient adaptive mechanisms that leverage stochastic information about the external world seemingly converge on the same basic set of Bayesian computations points at the importance of understanding the possible ways of implementing such computations. Within neuroscience, advances have been made in formulating possible implementations of Bayesian computations that are consistent with the constraints given by neural substrates, including locality of both information processing and learning and also the spiking nature of neural computations [9].

In this work, we explore how different flavors of Bayesian computations can be implemented on a fundamentally different substrate: the substrate of Darwinian replicators. The common overarching theme of such implementations is that abundances of types (ribozymes, genes, cells, organisms, etc. in a familiar biological setting) represent probabilities of hypotheses. Analogously to how probabilities of these hypotheses are updated based on their ability of predicting current data, abundances of types are updated based on a similar kind of environmental feedback: fitness in the current environment. This analogy can be formalized by the equivalence of the discrete-time replicator equation and Bayesian update [10, 11]. Bayesian update quantifies how the prior probability $P(h_i)$ of hypothesis h_i is updated to its posterior probability $P(h_i|x)$, given data x and the probability $P(x|h_i)$ of x being generated (or predicted) by h_i called the likelihood of h_i, as

$$P(h_i|x) = P(x|h_i)P(h_i)/\sum_j P(x|h_j)P(h_j). \tag{1}$$

The discrete-time replicator equation has the equivalent algebraic form: the relative abundance of type i, denoted by p_i is updated from time t to $t+1$ according to its fitness $f_i(x)$ in the current environment x as

$$p_i(t+1) = f_i(x)p_i(t)/\sum_j f_j(x)p_j(t). \tag{2}$$

Identifying relative abundances with probabilities and fitness with likelihood shows that these two forms of update are indeed equivalent.

The goal of this paper is to point out how any evolutionary dynamics, natural or engineered, biological or non-biological, fits in a general Bayesian computational framework. Two potential benefits of this approach are i) it might help understanding (macro)evolutionary processes in high-dimensional stochastic environments in normative/computational terms and ii) it can hint at possible ways to combine evolutionary and Bayesian computations, leveraging the advantages of both in a common computational framework [12, 13].

Thanks to recent efforts, many fundamental Darwinian phenomena can now be translated to the language of Bayesian computations, including selection [10, 11], mutation [14] and multilevel evolutionary processes [15]. Here we complement this list with the Bayesian interpretation of phenotypic adaptation and of evolutionary-ecological dynamics.

More importantly, we discuss elementary building blocks of evolutionary and probabilistic models in a coherent narrative, provided by the language of graphical models [3], which, similarly to Feynman diagrams in particle physics, provide an easy-to-manipulate visual representation of relevant aspects of computations, while, just as importantly, hides less relevant ones.

We also believe that one of the most important contributions any theoretical approach can provide is the clear articulation of assumptions behind any argumentation. We therefore attempt to provide exact mathematical formulations wherever possible.

2 Results

In the following, we build up a series of correspondences between simple models of evolutionary dynamics, including that of selection, adaptation, mutation, evolutionary-ecological dynamics and multilevel selection, and Bayesian generative models in a step-by-step manner, highlighting the assumptions we make at each step. Importantly, we also point out what aspects of modeling are not constrained by this theoretical framework.

First and foremost, one of the most unconstrained aspects of this correspondence is the interpretation of environment $x(t)$ at time t. The environment can comprise any number of data points at each time step, and can be governed by any (unknown) generative process. Among others, two possible interpretations are the following: $x(t)$ might correspond to one single data point summarizing the actual environment

without further specification, or, $x(t)$ might correspond to a set of points, whose local density stands for the probability/weight of the environment being at that given state. This includes the interpretation of the density of points corresponding to the density of finite resources, one that is assumed when ecological competition is modeled.

Another cornerstone connection between evolutionary and Bayesian models is the interpretation of likelihood as fitness. The likelihood function $P(x|h_i)$, telling us the probability of x being generated by hypothesis h_i, accounts naturally for an important constraint: adaptation always unfolds under some trade-offs that prohibit unlimited mastery of all possible environments; instead, both generative models and replicator types must give up adeptness in one environment to be more adapted to others. This is formalized in simple mathematical terms as the likelihood function being normalized over possible environments, $\sum_x P(x|h_i) = 1$, corresponding to $\sum_x f_i(x) = 1$ for any hypothesis/type. The evolutionary interpretation of the likelihood function in a simple case is shown in Figure 1.

Given the statistical structure of the environment and the constraint on possible fitness functions described above, replicator types compete to fit the environment in a manner analogous to the competition between hypotheses to predict the statistical structure of the data. Depending on the possible modes and timescales of adaptive mechanisms, minimal dynamical models of evolutionary processes map to different Bayesian models and inference procedures. In particular, two relevant timescales of evolution are i) microevolutionary timescale, where the change in abundance distribution over replicator types is explicitly modeled, ii) macroevolutionary timescale, where only the evolution of phenotypic population average is modeled. These are two complementary building block models of evolutionary dynamics, upon which additional ingredients, such as explicit mutation, ecological competition or multilevel selection, can be added. For simplicity, we refer to the former as selection and the latter as adaptation. Figure 1 illustrates these two elementary dynamics.

Selection, mutation, and adaptation Selection by itself does not account for how variation is generated; instead, it assumes that replicator types are fixed and only their relative abundances $p_i(t)$ that change over time. Indeed, as described above, the abundance of those replicator types that fit the current environment x better increases at a higher rate according to the discrete-time replicator equation; equivalently, the posterior probability of those hypotheses that predict the current data better increase at a higher rate, described by Bayesian update. The corresponding graphical model is depicted in Figure 2. Whatever type/hypothesis had nonzero relative abundance/prior probability initially, only those types/hypotheses compete; no new types emerge. This dynamics is the simplest building block that models selection on a microevolutionary timescale.

Variation, generated by mutations, can be introduced into the selection-only model on two different timescales. The first approach extends the replicator equation by explicitly characterizing the number of mutation events, modeled by mutation rates $\mu_{i\leftarrow j}$ from type j to type i. The evolutionary dynamics is therefore determined by the overall effect of two forces acting on the same timescale: selection and mutation. Mathematically, this simultaneous dynamics can be accounted for by the replicator-mutator equation (cf. eq. (2)),

$$p_i(t+1) = \sum_j \mu_{i\leftarrow j} f_j(x_t) p_j(t) / \sum_k f_k(x_t) p_k(t) \tag{3}$$

where $\mu_{i\leftarrow j}$ determines the mutation probability (in unit time) from type j to type i and therefore is normalized as $\sum_i \mu_{i\leftarrow j} = 1$, and the factor $\sum_k f_k(x_t)p_k(t)$ is the average fitness of the population, responsible for keeping the distribution p_i normalized at all times. Eq. (3) describes how fitness $f_i(x_t)$ in the current environment, together with the mutation probabilities $\mu_{i\leftarrow j}$, update the abundance distribution p_i. Similarly to the simple replicator equation, information about the environment is transferred to the abundance distribution over time; here, however, this information is filtered through the mutation probabilities, including the diagonal elements $\mu_{i\leftarrow i}$ that specify the fidelity of replication. As observed by [14], this model of replication and mutation is equivalent to updating the probability distribution over latent hypotheses h_i in hidden Markov models (HMMs). HMMs extend Bayesian update by introducing probabilistic transitions between hidden states (i.e. hypotheses) h_i, and they infer the joint distribution $\{P(h_i, t=1), P(h_i, t=2), \ldots, P(h_i, t=T)\}$ over the hidden states over time given data history $\{x_1, x_2 \ldots x_T\}$. In order to achieve this, HMMs need two quantities to be pre-determined: hypothesis h_i's likelihood of data, $P(x|h_i)$, just like in the case of

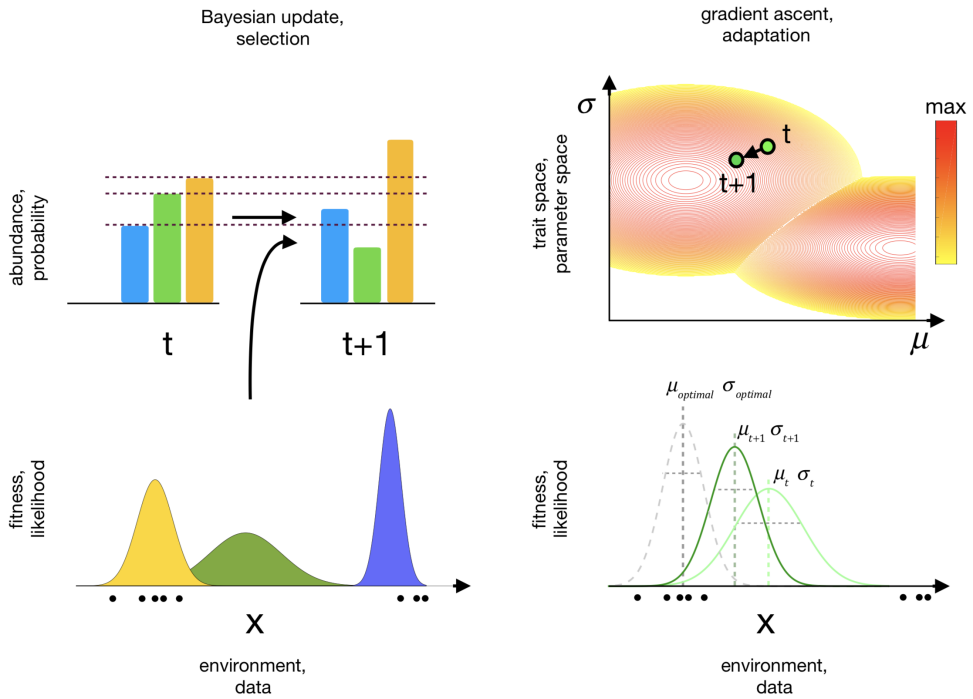

Figure 1: Left: on a microevolutionary timescale, replicator types are assumed to be fixed and their abundance is updated based on their fitness in the current environment. This is mapped to updating the probabilities of hypotheses based on how well they predict current data (i.e., their likelihood). Right: on a macroevolutionary timescale (right), mean phenotype of a population climbs the fitness surface, as a result of selection acting upon generated variation . This is mapped to gradient ascent on the likelihood surface. In both cases, the key ingredient is the identification of fitness of types over current environment with the likelihood of hypotheses over current data, depicted here as Gaussians.

simple Bayesian update, and the transition probabilities from h_j to h_i, denoted by $P(h_i|h_j)$. This joint probability distribution over hypotheses at all times, $\{P(h_i, t=1), P(h_i, t=2), \ldots, P(h_i, t=T)\}$, can be reduced, however, to distributions of interest; if this distribution of interest is the one corresponding to the last timestep, $P(h_i, t)$, dynamically inferring this last distribution given data history is called the *filtering* problem in HMMs, and it is obtained through the dynamics

$$P(h_i, t+1) = \sum_j P(h_i|h_j)P(x_t|h_j)P(h_j, t) / \sum_k P(x_t|h_k)P(h_k, t) \tag{4}$$

that is indeed equivalent to eq. (3), with the identifications depicted in Figure 3. Importantly, HMMs can also be cast as Bayesian graphical models, illustrated also in Figure X. Note that this dynamics computes the predictive distribution at time t, corresponding to taking into account data until time $t-1$ and transitions until time t; if data in the last timestep t is also integrated, eq. (4) needs a slight modification, corresponding to a less frequently used evolutionary counterpart.

The other approach of introducing the effect of mutations, that we refer to as adaptation, corresponds to modeling evolutionary dynamics at a qualitatively longer, macroevolutionary timescale. It does not explicitly account for mutations, instead, it assumes that a cloud of mutants in the (pheno)type space is always available upon which selection can act, irrespectively of what the generative mechanism of this variation is. It also shift the attention from tracking the abundance distribution of all types over time to simply following the population average of some relevant phenotypic traits. As an effect, in simple adaptive scenarios (with the explicit assumptions delineated below), the population average of relevant traits, denoted by $\bar{\theta}$, evolves according to a gradient ascent on the (log) fitness landscape:

$$\Delta\bar{\theta} = \eta \nabla_\theta \ln f_\theta(x)|_{\theta=\bar{\theta}} \tag{5}$$

where $\Delta\bar{\theta} = \bar{\theta}(t+1) - \bar{\theta}(t)$ is the change in the population average trait value (that can be a vector of any dimensionality), $f_\theta(x)$ is the fitness of individuals with trait value θ given environmental state x, and η is the rate of adaptation. Indeed, the Price equation [16], in its simplest form (assuming no intergenerational transmission bias), relates the change in population average $\Delta\bar{\theta}$ and the covariance between fitness and trait value $\mathrm{cov}(\theta, f)$ as $\Delta\bar{\theta} = \bar{f}^{-1}\mathrm{cov}(\theta, f)$, where $\bar{f}$ is the average fitness in the population. Assuming a deterministic fitness function f_θ once the environment x is fixed, an isotropic cloud of mutants in the phenotype space, and that the variance in trait values is small enough to approximate the fitness linearly around the population average by $f_{\bar{\theta}+\Delta\theta} = f_{\bar{\theta}}(x) + \Delta\theta \cdot \nabla_\theta f_\theta(x)|_{\theta=\bar{\theta}}$, the Price equation implies (see Appendix) the dynamics given by eq. (5) with rate of adaptation being equal to the variance of trait values projected to the direction of fitness gradient, $\eta = \mathrm{var}(\theta_{||})$.

This dynamics, assuming the above mentioned normalization constraint on fitness $\sum_x f_\theta(x) = 1$, is equivalent to the gradient optimization of the (log) likelihood function,

$$\Delta\theta = \eta \nabla_\theta \ln P(x|\theta) \tag{6}$$

with learning rate η. The corresponding graphical model is shown in figure 4. Crucially, the connection between the two descriptions is given by the identification of fitness and likelihood again. In this model of macroevolutionary adaptation, new types can and do arise; however, it is only the phenotypic population average that this description accounts for. Furthermore, the rate of adaptation η is given by the phenotypic population variance that i) might change over time, and ii) is not modeled here, in accordance with the dynamical insufficiency of the Price equation [17].

Ecological competition As we shall see, selection and adaptation, as well as the combination of the two, that is necessary to describe evolutionary-ecological interaction of species, form the basis of relating more complex evolutionary scenarios to their counterpart model within learning theories. One additional phenomenon relevant to life across all scales is competition for finite resources. Competition takes place at two timescales: on the ecological timescale, species abundances change according to their current ability to access the resources; on the evolutionary timescale, species themselves (i.e., their phenotypic average) become more adapted to the distribution of environmental resources in niche space. Ecological competition therefore serves as an important and, as we will see, the simplest nontrivial example in which combining

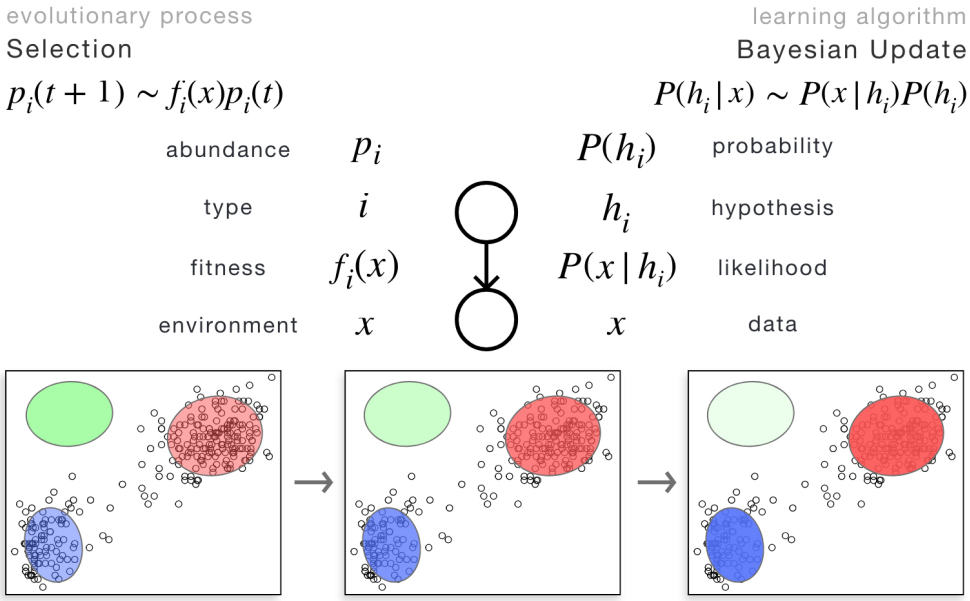

Figure 2: Top: the graphical model corresponding to the replicator equation, describing selection over fixed types favoring types that are more fit in the current environment; and to Bayesian update, describing the change of probabilities of hypotheses, favoring hypotheses that predict current data better (i.e., have higher likelihood). Variables are illustrated by circles; (conditional) dependencies between them are symbolized by arrows. Bottom: Illustration of the process in a two-dimensional environment/data and Gaussian fitness/likelihood, depicted by ellipses. The phenotypes/hypotheses, characterized by the location, size, and orientation of the Gaussians, are fixed; their abundance/probability, signaled by the opacity of the ellipses, change according to how fit they are/how well they predict data.

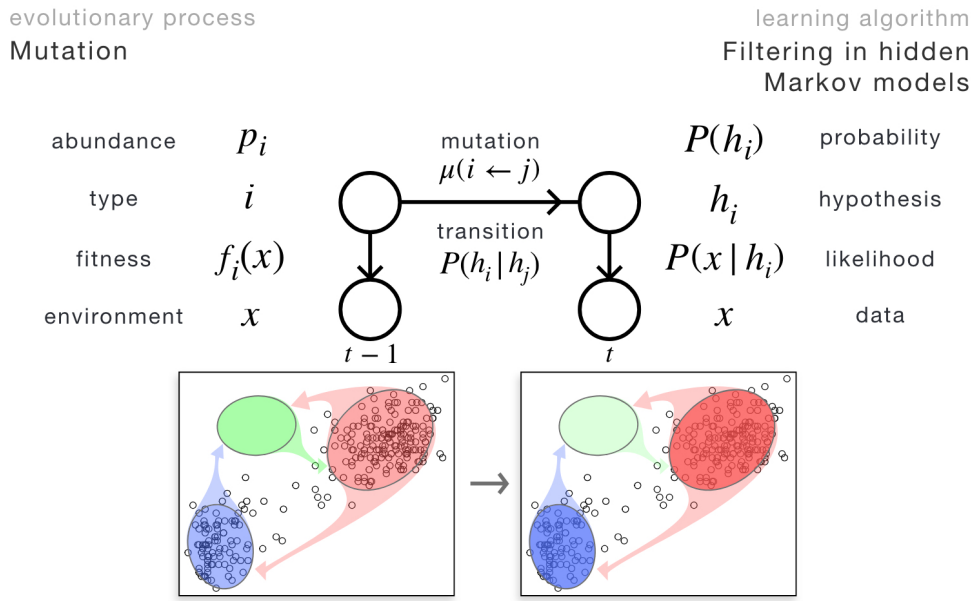

Figure 3: Top: the graphical model corresponding to the replicator-mutator equation, describing the effect of selection and mutation among fixed types; and to hidden Markov models, describing the effect of Bayesian update in light of new data and transitions between fixed hypotheses. Bottom: the phenotypes/hypotheses are fixed, shown here by the fixed location, size, and orientation of the ellipses; their abundance/probability is driven by how fit they are/how well they predict data and by the mutation/transition probabilities between them.

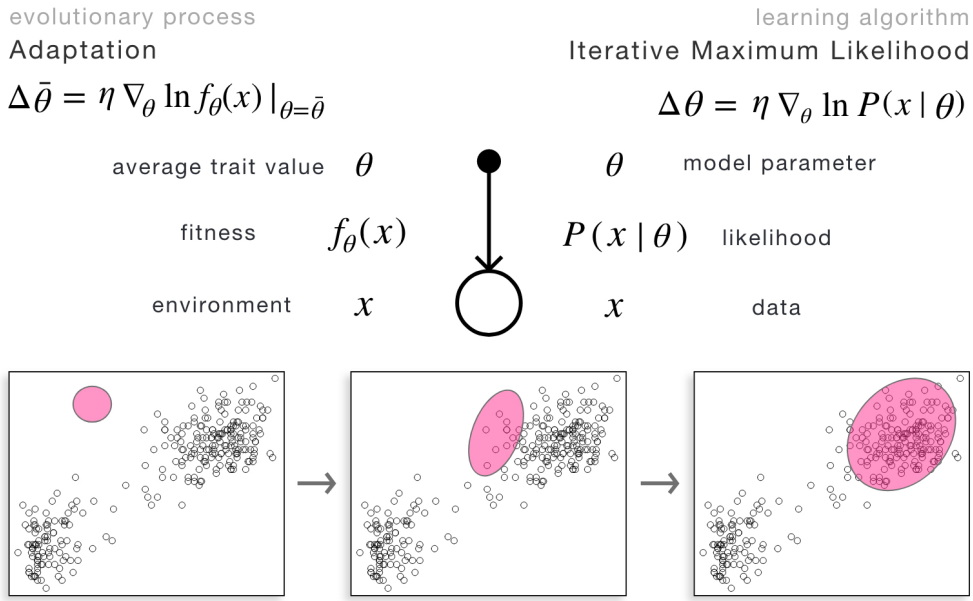

Figure 4: Top: the graphical model corresponding to macroevolutionary adaptation, i.e., the change in mean phenotype of a (quasi)species upwards in the fitness landscape, as a result of selection acting upon generated variation, neither of them is modeled explicitly here. It also describes how the parameters corresponding to a single hypothesis climb the likelihood surface to provide a better and better fit of data. Nodes (such as θ here) represent parameter(s), whereas circles represent variables. Bottom: adaptation of mean phenotype, depicted by the location, size, and orientation of the Gaussian, to the current environment. Equivalently, iterative optimization of the likelihood function to fit data.

dynamics at both timescales, referred to as selection and adaptation, is necessary. We model evolutionary-ecological dynamics as follows. Multiple species compete to access environmental resources; the amount of resources individuals of a species extract from the environment determines the abundance of the species. Mean phenotypic trait values of species and their abundances co-evolve. What makes the unfolding of these two dynamics non-trivial is that they are coupled: species adapt to the resources available to them, however, available resources are determined by the trait values of other species as well as their abundances. Abundances, set by the amount of resources species access, change according to the evolution of trait values of all species.

Mathematically, we model the environment as a set of data points $x_1, \ldots, x_m$, each of them corresponding to one unit of resource in resource space at their corresponding location within the niche space. Species i's ability to access resource x_j, depending on its trait value θ_i, is denoted by $u(\theta_i, x_j)$. We refer to u as utilization. It is these trait values θ that species evolve, and therefore the adaptive trade-off any species face is imposed on their utilization function, $\sum_x u(\theta_i, x) = 1$, instead of on their fitness. Utilization determines how resources are divided: the share each individual of any species with mean phenotype θ receives out of resource j is proportional to its utilization at x_j, $u(\theta, x_j)$. Consequently, species i's share of resource j, γ_{ij}, is proportional to the product of its utilization $u(\theta_i, x_j)$ and its relative abundance π_i

$$\gamma_{ij} = \pi_i u(\theta_i, x_j) / \sum_k \pi_k u(\theta_k, x_j). \tag{7}$$

The relative abundance π_i of species i is, in turn, given by the total amount of resources its members access,

$$\pi_i = \sum_j \gamma_{ij}. \tag{8}$$

Relative abundances π and trait values θ co-evolve: relative abundances are updated given more adaptive trait values, whereas trait values evolve given the updated relative abundances. It is worthwhile to investigate how this model relates to a well-known model of ecological competition: the game-dynamical replicator equation. The continuous time game-dynamical replicator equation, describing ecological competition through abundance-dependent selection reads as $\dot{\pi}_i = \pi_i f_i(\{\pi_k\}) - \bar{f})$, where π_i and f_i is the relative abundance and fitness of type i, respectively, and $\bar{f} = \sum_k \pi_k f_k$ is the average fitness. Importantly, in our model, quasi-equilibrium dynamics is assumed: trait values evolve on a timescale slow enough so that abundances are always in equilibrium given the current trait values. The equilibrium abundance distribution in the game-dynamical replicator equation (if exists), corresponds to all fitnesses being equal to the average fitness, $f_i = \bar{f}$ for all types i. This can be satisfied by interpreting fitnesses as excess utilization values summed over environmental resources:

$$f_i(\{\pi_k\}) = \sum_j u(\theta_i, x_j) / \bar{u}(x_j) \tag{9}$$

where $\bar{u}(x_j) = \sum_k \pi_k u(\theta_k, x_j)$. Indeed, $f_i = \sum_j u(\theta_i, x_j) / \sum_k \pi_k u(\theta_k, x_j) = \pi_i^{-1} \sum_j \pi_i u(\theta_i, x_j) / \sum_k \pi_k u(\theta_k, x_j) = \sum_j \gamma_{ij} / \pi_i = 1$ for all i.

Interestingly, an equivalent formulation of this simple model of evolutionary-ecological dynamics, plays a central role in machine learning and Bayesian statistics: maximum likelihood estimation in mixture models via the expectation-maximization (EM) algorithm. Mixture models fit a weighted sum of distributions $P(x) = \sum_i P(x|\{\theta_i\}, h_i)\pi_i$ to the data such that both the parameters of each component distribution $\{\theta_i\}$ and their relative weights $\{\pi_i\}$ are optimized in parallel improve on the likelihood of the whole model generating the actual data x. Since there is no known general closed-form formula for finding the global maximum of the likelihood function, in practice, mixture models are usually fitted iteratively using a variant of the expectation-maximization algorithm: parameters θ_i are improved given fixed weights π_i, whereas weights π_i are improved given fixed parameters θ_i; these two steps alternate until the parameters converge (assuming the data x does not change). Importantly, both the model framework (mixture models) and the dynamics (EM) are equivalent to the evolutionary-ecological dynamical model described above, with the following translation between the two interpretations: 1. species evolve their trait values θ_i to improve on their utilization function $u(\theta_i, x)$ and therefore extract more resources, corresponding to component distributions $P(x|\{\theta_i\}, h_i)$ improving on their parameter θ_i to increase the overall likelihood of the model; 2. relative

abundances π_i are updated given the amount of resources that is accessible to a species, corresponding to updating the relative weights of the component distribution such that it increases the likelihood of the whole model.

One important message of this equivalence is that even though it is composed of competing agents, evolutionary-ecological dynamics, in this simple model, optimizes a global cost function, given by the likelihood of the counterpart probabilistic model. More generally, as many learning algorithm is based on iteratively optimizing a global cost function, finding such functions that are emergently optimized by evolutionary systems composed of competing agents is an important gateway towards a more unified discussion of evolution and learning. Finding a Lyapunov function (i.e., a scalar-valued function over the space of parameters on which the learning/evolutionary dynamics always follows a downwards path) is not always easy, finding one out of many possible ones that contributes meaningfully to a learning-theory based interpretation of emergent computations is even harder. In evolutionary theory, this line of thought goes back to Fisher's fundamental theorem of natural selection, stating that a scalar function over (phenotypic) parameters, namely, average fitness, always increases if selection is not frequency-dependent [18]. It is perhaps less well known that a Lyapunov function always exists in frequency-dependent scenarios as well, now interpretable in the language of information theory: it is the Kullback-Leibler divergence (or relative entropy) between the current relative abundance distribution over types and the one at equilibrium that always decreases over time [19, 10, 20]. Following a similar path to find learning theoretically relevant Lyapunov/potential/cost functions in case of more complex dynamics of competing structures that are capable of representing patterns in high-dimensional data is a potentially important step towards a more unified theory of adaptation, incorporating both evolutionary systems composed of competing representations of the environment and learning systems that learn to represent data, many cases distributed over myriads of elementary computing units interacting only locally.

Another message the equivalence of this simple model of evolutionary-ecological dynamics and expectation-maximization in mixture models suggest is that it might be possible to engineer meta-level rules in synthetic evolutionary systems such that they learn as a whole to represent complex patterns in high-dimensional data. A further step towards designing creative artificial systems would be to build such an evolutionary system on complex-enough building blocks so that the complexification of representation can follow millions of diverging paths, impossible to list or foresee all of them by any existing technique, therefore rendering them unexpected and possibly creative.

Multilevel selection When the replication of different replicator types are partially but not fully synchronized and groups of replicators inherit information regarding the identity of the group, an effective description of the system is provided by multilevel selection theory (MLS) [21]. MLS decomposes the full effect of selection hierarchically to selection acting between and within groups. Partially synchronized replicators form a necessary intermediate step towards new emerging units of evolution, i.e., transitions in individuality, that is, in turn, a main mechanism responsible for increasing complexity in evolutionary systems. Understanding MLS is therefore of crucial importance regarding any, natural or engineered, open-ended Darwinian system.

MLS acts on a hierarchical population of replicators: types of individuals h_i are grouped into types of collectives z_j. A complete static description of the population is given by the abundance and composition of collectives (for further details, see [15]). The composition of collective z_j, denoted by c_i^j, normalized as $\sum_i c_i^j = 1$, quantifies the relative abundance of individual level replicators within collectives of type z_j. The total abundance of collective z_j, measured as the total abundance of individual-level replicators within collectives of type z_j, is denoted by p^j and normalized as $\sum_j p^j = 1$. These two sets of relative abundances, the compositions of collectives c_i^j and their abundance, p^j, describe the the composition of the hierarchical (two-level) population completely. From these two quantities, a third one, the total abundance of individual-level replicators of type h_i being part of collective z_j, denoted by p_i^j, can be calculated as $p_i^j = c_i^j p^j$. This quantity connects the two population levels in a sense that the abundance distribution of any level can be computed by summing over the other level: at the level of collectives, the abundance distribution, as we have seen, is given by $p^j = \sum_i p_i^j$; at the level of individuals, the abundance of type h_i, p_i, is obtained by adding up the abundances of type h_i being in any collective, $p_i = \sum_j p_i^j = \sum_j c_i^j p^j$.

These quantities above describe the composition of the population at one time instance. How does selection change this hierarchical population over time? According to the discrete-time replicator dynamics, abundances change proportionally to their fitness. Here, however, the fitness of an individual-level type h_i

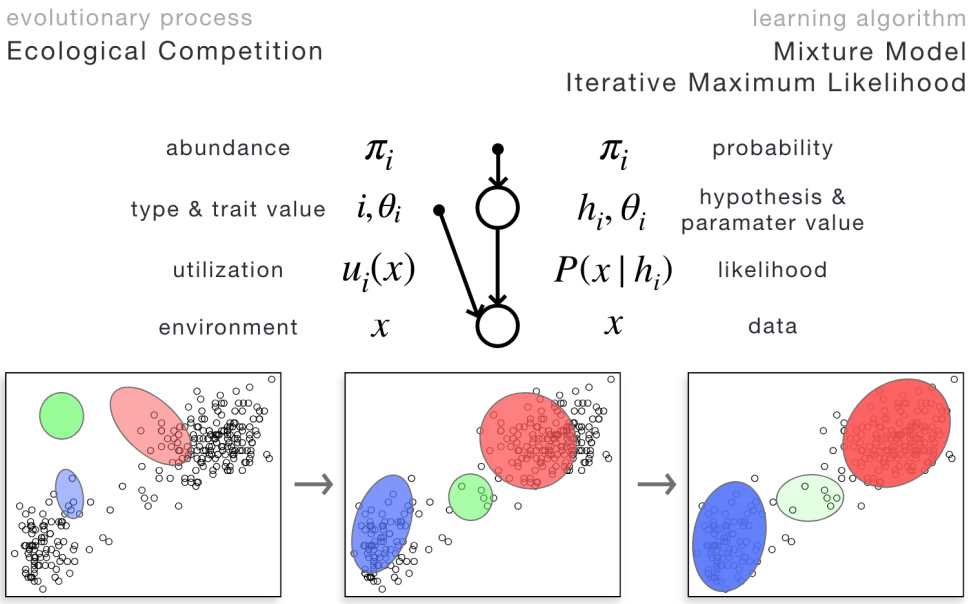

Figure 5: Top: A simple model of evolutionary-ecological dynamics, including intra-species adaptation and inter-species competition for finite resources. Equivalently, this model also corresponds to fitting data with multiple component distributions through optimizing a global likelihood function. This model exemplifies evolutionary systems possessing a global cost (or energy) function that is interpretable in the language of learning algorithms. Bottom: Evolutionary-ecological dynamics of three species, or equivalently, iterative maximum likelihood (e.g., Expectation-Maximization) optimization of a model involving three component distributions. Each datapoint here corresponds to one unit of resources that is shared between species based on their utilization ability. This resource competition drives species to occupy separate niches, or equivalently, component distributions to avoid overlap. Species abundances (opacity) and their mean phenotype (location, size, and orientation of the Gaussian) co-evolve, driven by the amount of resources they access; equivalently, parameters of component distributions and their weights are iteratively optimized such that the global likelihood of the model is improved.

might very well depend on the collective z_j it is part of; we denote this fitness by f_i^j. We also allow this fitness to depend on the environment x. The replicator dynamics, tracking the abundance of individuals of type h_i being part of collectives z_j, then reads as

$$p_i^j(t+1) = f_i^j(x)p_i^j(t)/\bar{f} \tag{10}$$

where the average fitness, $\bar{f}$, is calculated as $\bar{f} = \sum_{i,j} f_i^j p_i^j$. Tracking the abundance distribution at any level is then possible by summing over the other level, as discussed above. This, however, does not mean that the dynamics at the two levels are decoupled, for example, at the level of individuals, $p_i(t+1) = \sum_j f_i^j p_i^j/\bar{f} \neq (\sum_j f_i^j)(\sum_j p_i^j)/\bar{f}$.

Crucially, fitnesses and abundances are assigned to types that connect the two levels, namely, to individuals of type h_i that are part of collective z_j. The replicator equation acts on these quantities, evolving the multilevel population in time. This conceptualization of MLS allows for relating multilevel evolutionary dynamics to hierarchical Bayesian computations over multivariate distributions: these two hierarchical dynamics are structurally equivalent, with the identified quantities listed in Figure 6. Note that this analogy can be extended to arbitrary number of levels, see [15] for details.

An important aspect of multivariate Bayesian models is that it is possible to identify a parametrization-independent backbone of the model in terms of conditional independence relations between variables. This allows for distinguishing between model structure (topology) and parameters, a necessary step towards modeling causality [22]. As discussed in [15], conditional independence relations imposed on the evolutionary implementation of the corresponding Bayesian model have a well-defined intuitive meaning, too: it corresponds to freezing compositions at various levels of the hierarchical population.

Another fundamental feature of hierarchical models is their complexity. Finding the probabilistic model with optimal complexity given data is a non-trivial task that can be approached from many directions. The Bayesian approach is to gauge the model's performance averaged over all possible parameter settings of hidden variables [23]. When comparing two models, possibly having different number of variables and different number of hierarchical levels, one has to favor the one that predict data better on average. Interestingly, this Bayesian agenda translates to a general and simple evolutionary interpretation: if the average fitness of a collection of replicators is higher when they are grouped together in a collective compared to the case when they are "free", a new level emerges. Although this model of evolutionary transition of individuality [24, 25] is rather simplistic in terms of dynamics, it is not in terms of structure, and we believe that this similarity is quite remarkable, having the potential to be refined later on to understand and design complexifying evolutionary systems.

3 Discussion

Bayesian models provide the recipe of performing i) consistent probabilistic computations such that ii) the system extracts the maximum amount of information from the environment given clearly specified model constraints. They therefore form a unique cornerstone of any adaptive dynamics in uncertain environments. Their power lies in their generality: the algorithmic and implementational details constrained by the given physical/informational substrate is not specified. Here we show that it is possible to implement many of the most relevant types of Bayesian computations by simple replicator-based systems. On top of the basic ingredients that include Bayesian update over one hidden variable and iterative optimization of cost surfaces (e.g., maximum likelihood, maximum a posteriori estimation at any level of the Bayesian hierarchical model), we attempt to provide guidance for replicator-based implementation of some fundamental building blocks of the full Bayesian apparatus, such as inference in mixture models, hidden Markov models and hierarchical models. What makes such a translation table especially appealing is the simplicity of such implementations on the evolutionary side: they do not require introducing fundamental conceptual novelties, instead they correspond to well-known dynamics such as models of ecological competition, mutation and multilevel selection. This observation hints at a more unified picture of evolutionary dynamics and probabilistic learning; but it might also provide a ground for transferring more advanced computational techniques from one field to the other. These, we believe, might include leveraging the parallel accumulation of rare subsolutions and of sex, the possibilities of maintained search for (more and more complex) novelties given by an open-ended replicator system, or, from the side of probabilistic computations, finding an emergent global cost function

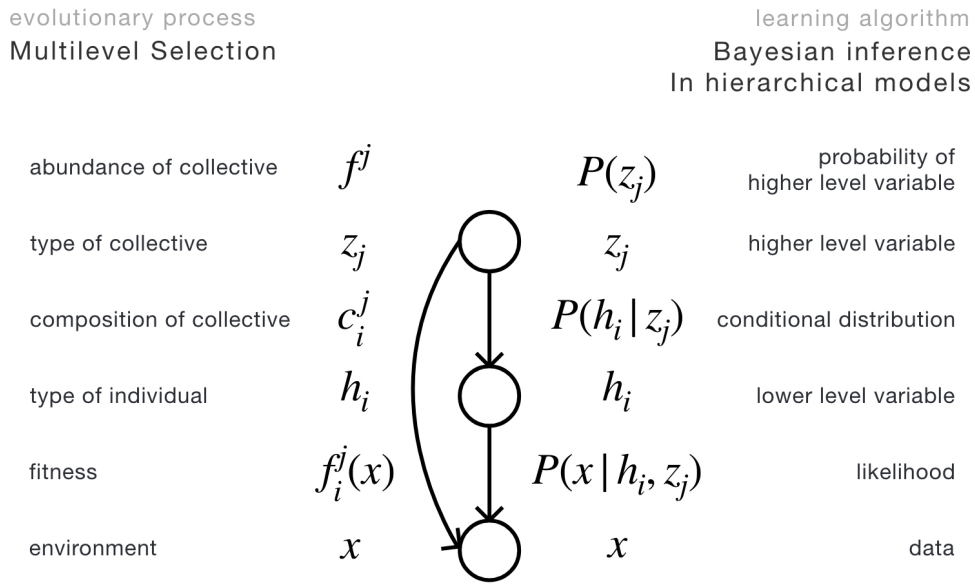

Figure 6: Bayesian inference in hierarchical models is equivalent to a model of multilevel selection. This identification relies on the possibility of representing multivariate probability distributions by hierarchical populations, combined with the equivalence of Bayesian update and discrete replicator equation.

that is improved upon by the dynamics of competing replicating agents, explaining the seemingly paradoxical adaptive potential of evolutionary systems as wholes.

We envisage three aspects of human inquiry that potentially benefit from this correspondence between evolutionary and probabilistic computations.

First, to explain the adaptive potential of any Darwinian dynamics in Nature, both those that we already have plenty of observations and understanding about (Darwinian processes on a genetic basis), but also those that have not yet been fully acknowledged, and the usefulness of replicator-based modeling is questionable at this point, such as memetics [26, 27], Darwinian neurodynamics [28, 29] or quantum Darwinism [30]. Such explanations would mostly be based on either i) finding global cost functions that evolutionary system emergently optimize or ii) relating the computations performed by the system to probabilistic computations that optimally extract information from external data.

Second, many aspects of animal (including human) cognition is efficiently modeled by the framework of Bayesian inference and generative models. This is in line with selective pressures favoring optimal extraction of information about the environment that in turn enables to select action that leads to maximum survival and reproduction probability. The possible implementations of Bayesian computations on neural substrates, however, is not fully explored. Here we hint at one possible implementation, based on copying and evaluation of any neural information. We do not posit that all Bayesian computations are implemented this way, instead, we point at the possibility that the effective information extraction provided by Bayesian computations might be efficiently combined with the open-ended novelty generation evolutionary systems are capable of.

This leads to our third point: the possibility of leveraging the combination of evolutionary and Bayesian computations in designing future AI systems to generate creative yet optimally informed solutions/actions in any probabilistic environment. This direction would, in the very first place, need a more thorough understanding of the relation between evolutionary dynamics and learning theories, possible unified in the same mathematical language.

Appendix

Showing that the rate of adaptation η equals to the phenotypic population variance in the direction of fitness gradient, $(\theta_{||})$. In the following, cov, var, corr and std denotes covariance, variance, correlation and standard deviation, respectively. We assume for simplicity that the variance in trait values

is small enough to approximate the fitness linearly around the population average by $f_{\bar{\theta}+\Delta\theta} = f_{\bar{\theta}}(x) + \Delta\theta \cdot \nabla_\theta f_\theta(x)|_{\theta=\bar{\theta}}$, and we also assume for simplicity that the mutant cloud in the phenotype space is isotropic around the population average. Furthermore, it is also assumed throughout the paper that the mapping $f_\theta(x)$ from trait values θ and environment x to fitness f is deterministic, implying that $\mathrm{corr}(f, \theta) = 1$. Since selection acts in the direction of the (log) fitness gradient $\nabla_\theta \ln f_\theta(x)$, we project the trait vector θ into that direction, denoted by $\theta_{||}$. Applying the one dimensional Price equation to $\theta_{||}$ and using the identity $|\nabla_\theta f_\theta(x)|\mathrm{std}(\theta_{||}) = \mathrm{std}(f)$ and $\bar{f} = f(\bar{\theta})$ validated by the linear approximation of the fitness landscape, we obtain $\Delta\bar{\theta}_{||} = \mathrm{cov}(\theta_{||}, f)/\bar{f} = \mathrm{std}(\theta_{||})\mathrm{std}(f)\mathrm{corr}(\theta_{||}, f)/\bar{f} = \mathrm{std}(\theta_{||})\mathrm{std}(f)/\bar{f} = \mathrm{var}(\theta_{||})|\nabla_\theta f_\theta(x)|_{\theta=\bar{\theta}}|/\bar{f} = \mathrm{var}(\theta_{||})|\nabla_\theta \ln f_\theta(x)|_{\theta=\bar{\theta}}|$.

References

[1] Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006.

[2] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127, 2010.

[3] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[5] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[6] József Fiser. Perceptual learning and representational learning in humans and animals. *Learning & behavior*, 37(2):141–153, 2009.

[7] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

[8] József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130, 2010.

[9] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170, 2013.

[10] Marc Harper. The replicator equation as an inference dynamic. *arXiv preprint arXiv:0911.1763*, 2009.

[11] Cosma Rohilla Shalizi et al. Dynamics of bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1039–1074, 2009.

[12] Richard A Watson and Eörs Szathmáry. How can evolution learn? *Trends in ecology & evolution*, 31(2):147–157, 2016.

[13] Jordan W Suchow, David D Bourgin, and Thomas L Griffiths. Evolution in mind: Evolutionary dynamics, cognitive processes, and bayesian inference. *Trends in cognitive sciences*, 21(7):522–530, 2017.

[14] Ömer Deniz Akyıldız. A probabilistic interpretation of replicator-mutator dynamics. *arXiv preprint arXiv:1712.07879*, 2017.

[15] Dániel Czégel, István Zachar, and Eörs Szathmáry. Major evolutionary transitions as bayesian structure learning. *bioRxiv*, page 359596, 2018.

[16] George R Price et al. Selection and covariance. *Nature*, 227:520–21, 1970.

[17] Steven A Frank. Natural selection. iv. the price equation. *Journal of evolutionary biology*, 25(6):1002–1019, 2012.

[18] R. A. Fisher. The genetical theory of natural selection. 1930.

[19] E Akin et al. The geometry of population genetics. *The geometry of population genetics.*, 1979.

[20] John C Baez and Blake S Pollard. Relative entropy in biological systems. *Entropy*, 18(2):46, 2016.

[21] Samir Okasha. Why won't the group selection controversy go away? *The British journal for the philosophy of science*, 52(1):25–50, 2001.

[22] Judea Pearl. *Causality*. Cambridge university press, 2009.

[23] David JC MacKay. Information theory, inference, and learning algorithms. 2003.

[24] John Maynard Smith and Eors Szathmary. *The major transitions in evolution*. Oxford University Press, 1997.

[25] Stuart A West, Roberta M Fisher, Andy Gardner, and E Toby Kiers. Major evolutionary transitions in individuality. *Proceedings of the National Academy of Sciences*, 112(33):10112–10119, 2015.

[26] Richard Dawkins. The selfish gene. page 224, 1976.

[27] Robert Aunger. Darwinizing culture: The status of memetics as a science. page 242, 2000.

[28] Chrisantha Thomas Fernando, Eörs Szathmáry, and Phil Husbands. Selectionist and evolutionary approaches to brain function: a critical appraisal. *Frontiers in Computational Neuroscience*, 6(24):24, 2012.

[29] András Szilágyi, István Zachar, Anna Fedor, Harold P de Vladar, and Eörs Szathmáry. Breeding novel solutions in the brain: A model of darwinian neurodynamics. *F1000Research*, 5(2416), 2016.

[30] Wojciech Hubert Zurek. Quantum darwinism. *Nature Physics*, 5(3):181, 2009.