axioms

MDPI

Article

On Using Relative Information to Estimate Traits in a Darwinian Evolution Population Dynamics

Eddy Kwessi

Department of Mathematics, Trinity University, 1 Trinity Place, San Antonio, TX 78212, USA; ekwessi@trinity.edu

Abstract: Since its inception, evolution theory has garnered much attention from the scientific community for a good reason: it theorizes how various living organisms came to be and what changes are to be expected in a certain environment. While many models of evolution have been proposed to track changes in species' traits, not much has been said about how to calculate or estimate these traits. In this paper, using information theory, we propose an estimation method for trait parameters in a Darwinian evolution model for species with one or multiple traits. We propose estimating parameters by minimizing the relative information in a Darwinian evolution population model using either a classical gradient ascent or a stochastic gradient ascent. The proposed procedure is shown to be possible in a supervised or unsupervised learning environment, similarly to what occurs with Boltzmann machines. Simulations are provided to illustrate the method.

Keywords: Darwinian evolution; traits; population dynamics; information theory

MSC: 37N30; 37N40; 39-08

check for updates

Citation: Kwessi, E. On Using Relative Information to Estimate Traits in a Darwinian Evolution Population Dynamics. *Axioms* **2024**, *13*, 406. https://doi.org/10.3390/axioms13060406

Academic Editors: Cheng-Cheng Zhu and Hari Mohan Srivastava

Received: 30 April 2024
Revised: 11 June 2024
Accepted: 12 June 2024
Published: 16 June 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The theory of evolution was famously championed by Darwin [1] in a publication where he stated his theory of natural selection. This theory soon became known as Darwinian evolution theory and can be construed as a series of postulates whose aim is to understand dynamical changes in organisms' traits. In other words, understanding the mechanisms of survival or disappearance of a species entails understanding mechanisms by which species traits are passed on overtime to their offspring. From a statistical point of view, a species having multiple offspring is the realization of a random process by which a certain amount of information is passed on to an offspring. How much information or how relevant information (often found in the organism's genome) is passed on may determine the species viability overtime. Since the paper of Vincent et al. [2] on Darwinian dynamics and evolutionary game theory, there have been many studies related to Darwinian dynamics. In Ecology in particular, Ackley et al. [3] proposed a model for competitive evolutionary dynamics, and Cushing [4] established difference equations for population dynamics. A Susceptible-Infected Darwinian model with evolutionary resistance was discussed in Cushing et al. [5]. Models for competitive species were proposed by Elaydi et al. [6]. However, the literature is very sparse on how to estimate species' traits (information stored or passed on to offspring) from these models using readily available data.

Information theory can help design estimation methods for species traits given data. Before showing how to use information theory for such a purpose, let us mention that there are two approaches to information theory that are related, but meaning and capturing different aspects of it. Firstly, consider a given sample of data from a probability distribution that depends on a parameter. Fisher's information [7] represents the amount of information that the sample contains about the parameter. It can be calculated as the inverse of the sampling error which is the random discrepancy between an estimate and the estimated parameter, and which arises from the sampling process itself (e.g., due to the fact that the

population as a whole has not been considered). Fisher's information is important because it is the inverse measure of the precision of a point estimate of a parameter. Indeed, the higher Fisher's information, the more precise is the point estimate. On the other hand, when the information content (or message) of a distribution is under consideration, Shannon's entropy [8] is used. The main difference here is that it is assumed that the content is not a parameter and therefore is not unknown. Most importantly, there is no latency relative to the population involved, unlike a sample, which is only a partial representation of a population. From Shannon's entropy, one can define the relative entropy, which allows us to compare different distributions and thus discriminate between different types of contents of information. Even though they have different interpretations, we note that the two approaches to information are mathematically related. Indeed, Fisher's information is the Hessian matrix with respect to the parameter of the relative entropy or Kullback–Leibler divergence. In discrete or continuous Darwinian dynamics, the main assumption is that of a deterministic relationship between inputs and outputs. As we stated above, the transmission of traits from parents to offspring is, in fact, a random process and should be analyzed as such. This assumes that there is a stochastic relationship between inputs and outputs via a network of connections often referred to as weights.

More precisely, say $\mathbf{u}$ is the vector of inputs and $\mathbf{v}$ the vector of outputs with network connections matrix $\mathbf{W}$. If input and output samples $(\mathbf{u}_s, \mathbf{v}_s), 1 \leq s \leq S$ are both available, the stochastic relationship between inputs and outputs is described as $\mathbb{P}(\mathbf{v}|\mathbf{u}, \mathbf{W})$, which is the probability that input $\mathbf{u}$ generates output $\mathbf{v}$ when the weight matrix of the network is $\mathbf{W}$. In supervised learning, the goal is to match as closely as possible the input–output distribution $\mathbb{P}(\mathbf{v}|\mathbf{u}, \mathbf{W})$ with the probability distribution $\mathbb{P}(\mathbf{v}|\mathbf{u})$ related to the samples $(\mathbf{u}_s, \mathbf{v}_s)$, by adjusting the weights $\mathbf{W}$ in a feedforward manner. In unsupervised learning, it is assumed that output samples are not given and only input samples $(\mathbf{u}_s), 1 \leq s \leq S$ are given. However, using an appropriate energy function, one can obtain the joint probability distribution $\mathbb{P}(\mathbf{u}, \mathbf{v}, \mathbf{W})$ and infer from rules of probability that $\mathbb{P}(\mathbf{u}, \mathbf{W}) = \sum_{\mathbf{v}} \mathbb{P}(\mathbf{u}, \mathbf{v}, \mathbf{W})$. The goal in unsupervised learning is therefore to match $\mathbb{P}(\mathbf{u}, \mathbf{W})$ as closely as possible to the probability distribution $\mathbb{P}(\mathbf{u})$ related to samples $(\mathbf{u}_s), 1 \leq s \leq S$, by adjusting the weights $\mathbf{W}$ in a feedforward manner. Feedforward techniques require the minimization of an appropriate objective function, and since we are dealing with probability distributions, the Kullback–Leibler divergence seems to be the best choice of objective function.

In this paper, our focus is on using relative information or entropy to estimate trait parameters in a Darwinian population dynamics model. The procedure we propose is similar to what occurs in a Boltzmann machine. The remainder of the paper is organized as follows: In Section 2, we make a brief overview of Fisher's information theory, entropy and relative as it pertains to mathematical statistics. In Section 3, we discuss how to computationally estimate trait parameters of discrete Darwinian models in supervised and unsupervised environments by maximizing the relative information or Kullback–Leibler divergence. Finally, in Section 4, we make some concluding remarks.

2. Review of Information Theory

In this light review of information theory, we mention Fisher's information only for self-containment sake. As for Fisher's information and evolution dynamics, the reader can refer to [9] for a deeper analysis. For ease of analysis and comprehension, the assumptions and notations below are along the lines of that work.

2.1. Fisher's Information Theory

Let $G(x, \Theta)$ be the probability density of a random variable X, continuous or discrete on an open set $\mathscr{X} \times \Omega \subset \mathbb{R} \times \mathbb{R}^n$. Here $\Theta = (\theta_1, \theta_2, \cdots, \theta_n)$ is either a single parameter or a vector of parameters.

In the sequel, we consider the following assumptions on the function $G(x, \Theta)$:

A_1 : The support $\{x \in \mathscr{X} : G(x, \Theta) \neq 0\}$ of G is independent of Θ.

A_2 : $G(\cdot, \Theta)$ is nonnegative and $G \in L^1(\mathscr{X})$ for all $\Theta \in \Omega$.

A_3 : $G(x,\cdot) \in C^2(\Omega)$, the set of continuously and twice differentiable functions of Θ, for all $x \in \mathscr{X}$.

The first assumption is to discard from consideration the uniform distribution $G(x,\theta) = \frac{1}{\theta}$ whose support is $(0,\theta)$. The second and third assumptions allow the well-definiteness of $\lambda(x,\theta) = \ln(G(x,\theta))$, its first derivative or score function $g(x,\Theta) = \nabla_\theta \lambda(x,\Theta)$ and its second derivative $h(x,\Theta) = \nabla_\Theta g(x,\Theta) = \nabla^2_\Theta \lambda(x,\Theta)$. In the sequel, the expected value of a random variable X is denoted as $\mathbb{E}[X]$.

Definition 1. *Given a random variable X with density function $G(x,\Theta)$ satisfying A_1 and A_2, Fisher's information of X is defined as*

$$I(\Theta) = \mathbb{E}_X[(g(X,\theta))^2] = -\mathbb{E}_X[h(X,\Theta)] \,. \tag{1}$$

When Θ is vector of more than one coordinates, Fisher's information is a symmetric positive definite (thus invertible) matrix $I(\Theta) = (I_{kl}(\Theta))_{1\le k,l\le n}$, where

$$I_{kl} = \mathbb{E}_X\left[\frac{\partial^2 \lambda(X,\Theta)}{\partial\theta_k \partial\theta_l}\right], \quad \text{for } 1 \le k,l \le n \,. \tag{2}$$

Fisher's information $I(\Theta)$ represents the amount of information contained in an estimator of Θ, given data X.

2.2. Entropy and Relative Entropy

Let us start by recalling the following definition of Shannon's entropy and that of the Kullback–Leibler divergence or relative entropy.

Definition 2. *Let μ be a probability distribution defined on a sample space Ω. Then entropy of μ is given as*

$$H(\mu) = -\mathbb{E}_\mu[\ln[\mu]] = \begin{cases} -\sum\limits_{\theta\in\Omega} \mu(\theta)\ln(\mu(\theta)), & \text{if } \mu \text{ is discrete} \\ -\int_{\theta\in\Omega} \mu(\theta)\ln(\mu(\theta))d\theta, & \text{if } \mu \text{ is continuous} \end{cases} . \tag{3}$$

Definition 3. *Suppose μ and ν are two probability distributions defined on a sample space Ω. Then, the Kullback–Leibler divergence or relative information $D(\mu,\nu)$ of μ relative to a fixed ν is given as*

$$D_{KL}(\mu,\nu) = \mathbb{E}_\mu\left[\frac{\ln(\mu)}{\ln(\nu)}\right] = \begin{cases} \sum\limits_{\theta\in\Omega} \mu(\theta)\ln\left(\frac{\mu(\theta)}{\nu(\theta)}\right), & \text{if } \mu \text{ and } \nu \text{ are discrete} \\ \int_{\theta\in\Omega} \mu(\theta)\ln\left(\frac{\mu(\theta)}{\nu(\theta)}\right)d\theta, & \text{if } \mu \text{ and } \nu \text{ are continuous} \end{cases} . \tag{4}$$

There is an obvious connection between entropy and relative entropy:

$$D_{KL}(\mu,\nu) = -\mathbb{E}_\mu[\ln(\nu)] - H(\mu) \,.$$

We notice that when $\mu = \nu$, we have that $D_{KL}(\mu,\nu) = 0$. It is known that Fisher's information induces a Riemannian metric ([10,11]) defined on a statistical manifold, that is, a smooth manifold whose points are probability measures defined on a probability space. It therefore represents the "informational" discrepancy between measurements. As such, it is related to the Kullback–Leibler divergence (KL), used typically to assess the

difference between two probability distributions. Indeed, to see this, let $\theta, \theta_0 \in \Omega$. Let us use a second-order Taylor expansion in μ. Then we will obtain

$$\begin{aligned} D_{KL}(\mu,\nu) &= D_{KL}(\mu,\nu)\big|_{\mu(\theta)=\nu(\theta)} \\ &+ \nabla_\theta(D_{KL}(\mu,\nu))\big|_{\mu(\theta)=\nu(\theta)} + \frac{1}{2}(\theta-\theta_0)^T\mathscr{H}(\theta)(\theta-\theta_0) + \mathrm{O}((\theta-\theta_0)^3)\,. \end{aligned}$$

From the above observation, we have $D(\mu,\nu)\big|_{\mu=\nu} = 0$. We also have

$$\begin{aligned} \nabla_\theta(D_{KL}(\mu,\nu)) &= \frac{\partial}{\partial\theta}\mathbb{E}_\mu\left[\frac{\ln(\mu)}{\ln(\nu)}\right] \\ &= \mathbb{E}_\mu\left[\frac{\partial}{\partial\theta}\left(\frac{\ln(\mu)}{\ln(\nu)}\right)\right] \\ &= \mathbb{E}_{\partial\mu}\left[\frac{\ln(\mu)}{\ln(\nu)}\right] + \mathbb{E}_\nu\left[\frac{\partial}{\partial\theta}\left(\frac{\mu}{\nu}\right)\right]. \end{aligned}$$

It follows that $\nabla_\theta(D_{KL}(\mu,\nu))\big|_{\mu(\theta)=\nu(\theta)} = 0$. We conclude by noticing that the Hessian matrix $\mathscr{H}(\theta) = \nabla^2_\theta D_{KL}(\mu,\nu)\Big|_{\mu(\theta)=\nu(\theta)} = \mathbb{E}_\mu\left[\frac{\partial^2}{\partial\theta_k\partial\theta_l}\ln(\mu(\theta))\right] = I(\theta)$. Therefore, if $\mu = \mu(\theta)$ and $\nu = \mu(\theta_0)$ are infinitesimally close to each other, that is,

$$\mu(\theta) = \mu(\theta_0) + \sum_k \Delta\theta_k \frac{\partial}{\partial\theta_k}\mu(\theta)\Big|_{\theta=\theta_0}, \tag{5}$$

we have

$$D_{KL}(\mu,\nu) = \frac{1}{2}\sum_{kl}\Delta\theta_k\Delta\theta_l I_{kl}(\theta) + \mathrm{O}(\Delta\theta^3)\,. \tag{6}$$

There is an interesting discussion in the finite case in [12]. Indeed, suppose that Ω is a finite set with cardinality n. Let $\Theta = (\theta_1, \theta_2, \cdots, \theta_n)$ and let $\Lambda(\Omega)$ be the set of all probability measures on Ω, and $S_n(\Omega)$ be the simplex $S_n(\Omega) = \{\Theta \in \Omega : \sum_{i=1}^n \theta_i = 1\}$. There is a isometry $\varphi : \Lambda(\Omega) \to S_n(\Omega)$ defined as $\varphi(\mu) = (\mu(1), \mu(2), \cdots, \mu(n)) = (\theta_1, \theta_2, \cdots, \theta_n)$. In this case, Fisher's information components become

$$I_{kl}(\theta) = \mathbb{E}_\mu\left[\frac{\partial^2}{\partial\theta_k\partial\theta_l}\ln(\mu(\theta))\right] = \sum_{j=1}^n \theta_j \frac{1}{\theta_k}\delta_{kj}\frac{1}{\theta_l}\delta_{lj} = \frac{1}{\theta_k}\delta_{kl}\,,$$

where δ_{kl} is the Kronecker symbol. It follows that

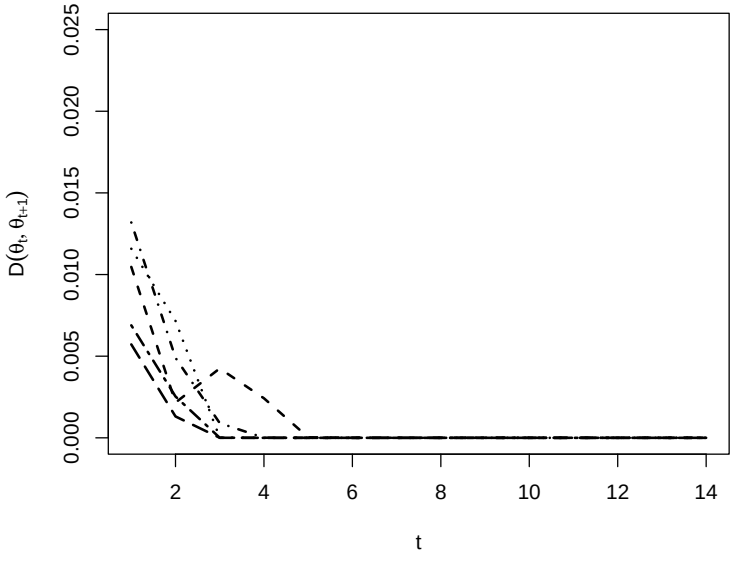
$$\frac{\partial^2 D(\mu,\mu_0)}{\partial\theta_k\partial\theta_l} = \frac{1}{\theta_k}\delta_{kl}\,.$$

The interpretation of the latter is Fisher's fundamental theorem (see [7]) and Kimura's maximal principle (see [13]) in terms of Fisher information: natural selection forms a gradient with respect to an information measure, and hence locally has the direction of maximal information increase. The rate of change of the mean fitness of the population is given by the information variance.

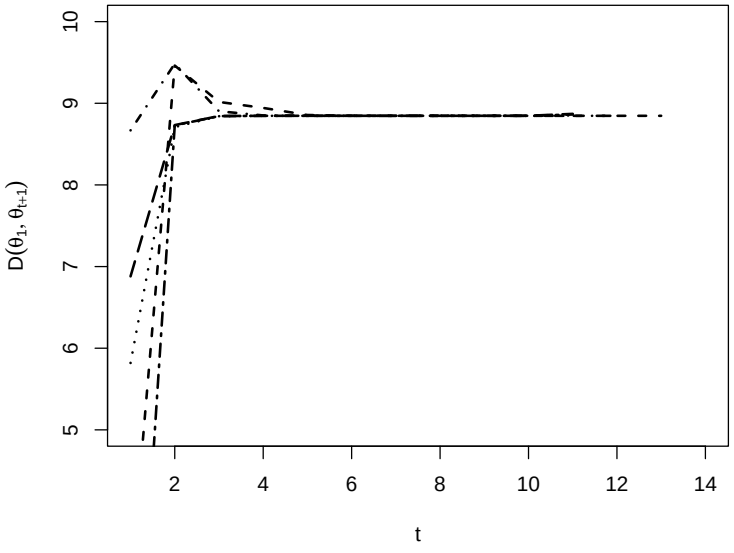
Let us highlight some of these facts by calculating the KL divergence $D_{KL}(\Theta_1, \Theta_t)$ between distributions $\nu = \mathbb{P}(\Theta_1)$ and $\mu = \mathbb{P}(\Theta_t)$ of Θ_1 and Θ_t, respectively, and $D_{KL}(\Theta_t, \Theta_{t+1})$ between the probability distributions $\nu = \mathbb{P}(\Theta_t)$ and $\mu = \mathbb{P}(\Theta_{t+1})$ of Θ_t and Θ_{t+1}, respectively, for $t = 1, 2, \cdots, n$. See Figure 1 below. We will use the same parameters as above with the same starting points for θ in the algorithm.

The minimization of $D_{KL}(\Theta_t, \Theta_{t+1})$ and the maximization of $D_{KL}(\Theta_1, \Theta_t)$ both occur when Fisher's information is maximal or when the variance of the estimator is minimal. From a dynamical system perspective, this occurs when the fixed point (e^{-1}, e) has been

attained. This means that the KL divergence can be used to determine the critical points of a discrete Darwinian model. Therefore, there is a dichotomy between the problems of minimizing $D(\Theta_t, \Theta_{t+1})$ and maximizing $D_{KL}(\Theta_1, \Theta_t)$. This amounts to matching the probability distributions $\mathbb{P}(\Theta_t)$ and $\mathbb{P}(\Theta_{t+1})$ as closely as possible. When this happens, we know from above that we will have $D_{KL}(\Theta_t, \Theta_{t+1}) \approx 0$, which in turn will mean that the fixed points of the discrete Darwinian model have been attained. To that end, we will discuss how to accomplish this with machine learning approaches, namely under supervised and unsupervised learning.



(**a**)



(**b**)

Figure 1. In (**a**) $D(\Theta_t, \Theta_{t+1})$; (**b**) $D(\Theta_1, \Theta_{t+1})$ for $t = 1, 2, \cdots, 15$. In the (**a**), convergence to zero means that the distributions of Θ_t and Θ_{t+1} are similar. In (**b**), convergence to about 8.7 means that the distributions of Θ_t are no more different from each other, making their KL divergence from the distribution of Θ_1 constant. In both cases, this is an indication that Fisher's information has been maximized or that the variances of the estimates are the same and very small.

3. Evolution Population Dynamics and Relative Information

3.1. Single Darwinian Population Model with Multiple Traits

Now suppose we are in the presence of one species with density x possessing n traits given by the vector $\Theta = (\theta_1, \theta_2, \cdots, \theta_n)$ and a vector $U = (u_1, u_2, \cdots, u_n)$. We will consider the following:

(H_1) $b(\Theta) = b_0 \exp\left(-\sum_{i=1}^{n} \frac{\theta_i^2}{2w_i^2}\right)$ is the joint distribution of the independent traits θ_i, each with mean 0 and variance w_i^2.

(H_2) $c_U(\Theta) = c_0 \exp\left(-\sum_{i=1}^{n} \kappa_i(\theta_i - u_i)\right)$.

(H_3) The density of x_t is given as $G(x, \Theta, U) = b(\Theta)\exp(-c_U(\Theta)x)$ at $t = 1$.

Under $H_1 - H_3$, we will consider the discrete dynamical system

$$\begin{cases} x_{t+1} &= x_t G(x_t, \Theta_t, U_t) \\ \\ \Theta_{t+1} &= \Theta_t + \Sigma g(x_t, \Theta_t, U_t) \end{cases}, \tag{7}$$

where

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}. \tag{8}$$

Remark 1. *We note that in the context of Darwinian population dynamics, $G(x_t, \Theta_t, U_t)$ is the population growth rate, $b(\Theta)$ represents the birth rate, while $c_U(\Theta)$ is the intra-specific competition function. Σ represents the covariance matrix of the distribution of traits among phenotypes of the species.*

In this section, we propose two approaches to estimate the traits vector Θ in a Darwinian model, using relative information.

3.2. Supervised and Unsupervised Learning

Supervised and unsupervised learning are two types of machine learning techniques whose ultimate aim is to learn the best possible connections between a set of inputs and outputs. In supervised learning, the inputs and outputs are presented to a system that learns the best possible connections between inputs and outputs. In an unsupervised learning environment, only inputs are presented and the system learns and describes the best possible outputs that would be related to the inputs. Let us note that the distribution of Θ_t depends on $X_t, K = (\kappa_1, \kappa_2, \cdots, \kappa_n), W = (w_1, w_2, \cdots, w_n)$, and $U = (u_1, u_2, \cdots, u_n)$.

In the sequel, we propose a machine learning approach for the estimation of trait vector Θ. We show in particular that depending on the amount of information available, machine learning approaches are possible. To that end, learning schemes for supervised and unsupervised learning are designed to estimate both W and K, which are the key parameters in the Darwinian model (7) proposed above. In fact, once the vectors W and K have been estimated, data available on X_t and the second equation in (7) are used to evaluate $\Theta_t^{\text{estimated}}$, the estimated value of Θ. In supervised learning, the values of $\Theta_t^{\text{estimated}}$ should match closely with their sample counterparts $\Theta_t^{\text{samples}}$. In unsupervised learning, in the absence of sample data $\Theta_t^{\text{samples}}$, the estimated values $\Theta_t^{\text{estimated}}$ should serve as actual values for Θ_t, the vector of traits.

3.2.1. Supervised Learning

Let n, T and M be given positive integers. In supervised learning, we are given a sample of inputs/outputs sample in the form

$$\left\{\left(X_t^{(m)}, \Theta_t^{(m)}\right), \quad t = 1, 2 \cdots, T; m = 1, 2 \cdots, M\right\},$$

where each $\Theta_t^{(m)}$ is a $1 \times N$ vector. This means that in this case, we already know the inputs that generate the solution of the system (7). The probability $\mathbb{P}(X_t|\Theta_t, W, K)$ of X_t given Θ_t, W, K is

$$\mathbb{P}(X_t|\Theta_t, W, K) = b_0 e^{-b(X_t, \Theta_t, W, K)}. \tag{9}$$

Therefore, using known techniques of conditional probabilities, we have

$$\mathbb{P}(\Theta_t|X_t, W, K) = \frac{\mathbb{P}(X_t|\Theta_t, W, K)\mathbb{P}(\Theta_t)}{Z}, \tag{10}$$

where

$$Z = \sum_{\Theta_t} \mathbb{P}(X_t|\Theta_t, W, K)\mathbb{P}(\Theta_t).$$

Since we are trying to estimate traits, in supervised learning, the goal will be to determine how well $\mathbb{P}(\Theta_{t+1}|X_{t+1}, W, K)$ matches $\mathbb{P}(\Theta_t|X_t)$. Since these are distributions, we use the Kulback–Leibler (KL) divergence

$$D_{KL}(\mathbb{P}(\Theta_t|X_t), \mathbb{P}(\Theta_{t+1}|X_{t+1}, W, K)) = -\sum_{\theta_t} \mathbb{P}(\Theta_t|X_t) \ln(\mathbb{P}(\Theta_{t+1}|X_{t+1}, W, K)) - H(\Theta_t|X_t), \tag{11}$$

where

$$H(\Theta_t|X_t) = -\sum_{\theta_t} \mathbb{P}(\Theta_t|X_t) \ln(\mathbb{P}(\Theta_t|X_t))$$

is the entropy of the distribution of $\Theta_t|X_t$. At the sample level for given M samples, the average of KL divergence is given as

$$\langle D_{KL}(\mathbb{P}(\Theta_t|X_t), \mathbb{P}(\Theta_{t+1}|X_{t+1}, W, K))\rangle = -\frac{1}{M}\left[\sum_{m=1}^{M} \ln\left(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)}, W, K)\right) + H\left[\Theta_t^{(m)}|X_t^{(m)}\right]\right]. \tag{12}$$

We minimize the quantity on the right-hand side of (12), using a stochastic gradient descent scheme, by updating the weight vectors W and K, respectively. Hence, the following result on supervised learning.

Theorem 1. *Suppose that X_t, W, and K are as above and that we have input/output sample $\left\{\left(X_t^{(m)}, \Theta_t^{(m)}\right), t = 1, 2, \cdots, T, m = 1, 2, \cdots, M\right\}$. The minimization process of the KL divergence can be achieved with a classic gradient ascent algorithm for weights W and K with an update scheme given as*

$$\begin{cases} w_{l,new} &= w_{l,old} + \alpha_W \dfrac{\partial \ln(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)}, W, K))}{\partial w_l} \\ \kappa_{l,new} &= \kappa_{l,old} + \alpha_K \dfrac{\partial \ln(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)}, W, K))}{\partial \kappa_l} \end{cases}, \tag{13}$$

where $\alpha_w, \alpha_\kappa > 0$ are the learning rates of W and K, respectively, and $\dfrac{\partial \ln(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)}, W, K))}{\partial x}$ for $x = w_l, \kappa_l$ is given in the Appendix A.

Corollary 1. *Under the assumptions of Theorem 1 above, the minimization process of the KL divergence can be achieved more efficiently with a stochastic gradient ascent algorithm for weights W and K with an update scheme given as*

$$\begin{cases} w_{l,new} & = w_{l,old} - \alpha_W \left[\dfrac{\partial b(X_{t+1}^{(m)}, \Theta_{t+1}^{(m)}, W, K)}{\partial w_l} - \dfrac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial w_l} \right] \\ \kappa_{l,new} & = \kappa_{l,old} - \alpha_K \left[\dfrac{\partial b(X_{t+1}^{(m)}, \Theta_{t+1}^{(m)}, W, K)}{\partial \kappa_l} - \dfrac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial \kappa_l} \right] \end{cases} . \tag{14}$$

where $\alpha_w, \alpha_\kappa > 0$ are the learning rates of W and K, respectively.

Remark 2. *One could use Jeffrey's prior as the probability distribution $\mathbb{P}(\Theta_t)$ of Θ_t if no other information about its distribution is given. In this case, $\mathbb{P}(\Theta_t) \propto \frac{1}{\sqrt{det(I(\Theta_t))}}$, where $det(A)$ is the determinant of the matrix A. We discussed in [9] that Fisher's information matrix for such a system depends on both W and K. This means that a Jeffrey's prior is advisable only for algorithm initialization; otherwise, from Equation (10), the updating scheme would have to be changed according to the dependence of $\mathbb{P}(\Theta_t)$ on both W and K. In fact, in this case, one would need to consider the respective partial derivatives $\frac{\partial \ln(\mathbb{P}(\theta_t))}{\partial x} = -\frac{1}{2}\frac{\frac{\partial}{\partial x} det(I(\theta_t))}{det(I(\theta_t))}$, for $x = w_l, \kappa_l$, which significantly increases the complexity of the problem.*

3.2.2. Unsupervised Learning

Now, we assume that X_t, W, and K are as above and only a sample of inputs $\left\{ X_t^{(m)}, m \in \{1, 2, \cdots, M\} \right\}$ is given. Since $\Theta_t^{(m)}$ is not given, we cannot use the learning scheme above. We have to design a method that only depends on $X_t^{(m)}, W$ and K. We assume that the joint probability $\mathbb{P}(X_t, \Theta_t, W, K)$ of X_t and Θ_t is

$$\mathbb{P}(X_t, \Theta_t, W, K) = \frac{1}{Z} e^{-b(X_t, \Theta_t, W, K)} , \tag{15}$$

where Z is the constant $Z = \sum_{X_t, \Theta_t} e^{-b(X_t, \Theta_t, W, K)}$.

In unsupervised learning, we would like to determine how well $\mathbb{P}(X_{t+1}, W, K)$ matches $\mathbb{P}(X_t)$. Since these are distributions, we use the Kulback–Leibler (KL) divergence

$$D_{KL}(\mathbb{P}(X_t), \mathbb{P}(X_{t+1}, W, K)) = -\sum_{X_t} \mathbb{P}((X_t)) \ln(\mathbb{P}(X_{t+1}, W, K)) - H(X_t) , \tag{16}$$

where $H(X_t) = -\sum_{X_t} \mathbb{P}(X_t) \ln(\mathbb{P}(X_t))$ is the entropy of the distribution of X_t. At the sample level, and for given M samples from the random variable X_t, the average of KL divergence is given as

$$\langle D_{KL}(\mathbb{P}(X_t), \mathbb{P}(X_{t+1}, W, K)) \rangle = -\frac{1}{M} \left[\sum_{m=1}^{M} \ln\left(\mathbb{P}(X_{t+1}^{(m)}, W, K) \right) + H\left(X_t^{(m)} \right) \right] . \tag{17}$$

Similarly, as above, we minimize the quantity on the right-hand side of (17), using a stochastic gradient descent scheme, by updating the weight vectors W and K, respectively. Hence, the following result is on unsupervised learning.

Theorem 2. *Suppose that X_t, W, and K are as above and that we have input/output sample $\left\{ X_t^{(m)}, t = 1, 2, \cdots, T, m = 1, 2, \cdots, M \right\}$. The minimization process of the KL divergence can*

be achieved with a classic gradient ascent algorithm for weights W and K with an update scheme given as

$$\begin{cases} w_{l,new} &= w_{l,old} + \alpha_W \dfrac{\partial \ln\left(\mathbb{P}(X_{t+1}^{(m)}, W, K)\right)}{\partial w_l} \\ \kappa_{l,new} &= \kappa_{l,old} + \alpha_K \dfrac{\partial \ln\left(\mathbb{P}(X_{t+1}^{(m)}, W, K)\right)}{\partial \kappa_l} \end{cases}. \tag{18}$$

where $\alpha_w, \alpha_\kappa > 0$ *are the learning rates of W and K, respectively, and* $\dfrac{\partial \ln\left(\mathbb{P}(X_{t+1}^{(m)}, W, K)\right)}{\partial x}$, *for* $x = w_l, \kappa_l$ *is given in the Appendix A.*

Corollary 2. *Under the hypotheses of Theorem 2 above, the minimization process of the KL divergence can be achieved more efficiently with a stochastic gradient ascent algorithm for weights W and K with an update scheme given as*

$$\begin{cases} w_{l,new} &= w_{l,old} - \alpha_W \left[\dfrac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial w_l} - \dfrac{\partial b(X_t^{(m)}, \Theta_t^{(m)}, W, K)}{\partial w_l} \right] \\ \kappa_{l,new} &= \kappa_{l,old} - \alpha_K \left[\dfrac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial \kappa_l} - \dfrac{\partial b(X_t^{(m)}, \Theta_t^{(m)}, W, K)}{\partial \kappa_l} \right] \end{cases}, \tag{19}$$

where $\alpha_w, \alpha_\kappa > 0$ *are the learning rates of W and K, respectively.*

Remark 3.

1. *We observe that since* $D_{KL}(\mathbb{P}(\Theta_t|X_t), \mathbb{P}(\Theta_{t+1}|X_{t+1}, W, K))$ *is an expectation with respect to* $\Theta_t|X_t$ *(see Definition 3 above), then the right-hand-side of Equation (12) is a Riemann sum for* $D_{KL}(\mathbb{P}(\Theta_t|X_t), \mathbb{P}(\Theta_{t+1}|X_{t+1}, W, K))$.
2. *Similarly,* $D_{KL}(\mathbb{P}(X_t, \mathbb{P}(X_{t+1}, W, K))$ *is an expectation with respect to the random variable* X_t, *therefore the right-hand-side of Equation (17) is one of its Riemann sum.*

Remark 4. *We note that there is a subtle difference between Equations (14) and (19), but an important one. In Equation (14), all the data are available for training, therefore all quantities can be calculated. In Equation (19), we are not given sample data* $\Theta_t^{(m)}$, *however, we can use the Darwinian difference Equation (7) to evaluate both* $\dfrac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial x}$ *and* $\dfrac{\partial b(X_t^{(m)}, \Theta_t^{(m)}, W, K)}{\partial x_l}$.

3.3. Simulations

In this simulation, we generate the data from system (7) as follows: we choose $T = 200, M = 350, N = 2, c_0 = 0.1, b_0 = 1.5$ and the vectors $W \sim 100 \times Unif(0.5, 1)$, $K \sim Unif(0.1), U \sim Unif(0, 10)$ and $\Theta = (1, 2, 3)$. From this we generate the samples

$$\left\{ \left(X_t^{(m)}, \Theta_t^{(m)} \right), t = 1, \cdots, T, m = 1, \cdots, M \right\}.$$

In Figure 2, we start by illustrating of the generated data, where the first column represents the dynamics of X_t and the second column represents the dynamics of $\Theta_t = (\theta_{1t}, \theta_{2t})$. To evaluate the accuracy of the method in supervised learning, we evaluate the percentage of correct calculation of vector parameters W and K as

$$\text{Correct Percentage} = \frac{\#\{\|\text{Parameter}^{\text{true}} - \text{Parameter}^{\text{estimated}}\| \leq \eta\}}{\#\text{ of samples}},$$

for a threshold parameter $\eta = 0.001$, where $\|\cdot\|$ represents a norm in $\mathbb{R}^N$. In Figure 3 below, we show the percentage of correct calculation for the given data under supervised learning.

Remark 5. *Let us observe that the learning techniques proposed above are similar to those of a Boltzmann machine. Indeed, we have*

$$\begin{aligned} G(X,\Theta) &= \frac{1}{b_0}e^{-b(X,\Theta,W,K)} \\ &= \frac{1}{b_0}exp\left(-\sum_{i=1}^{n}\frac{\theta_i^2}{2w_i^2} - c_0 X exp\left(-\sum_{i=1}^{n}\kappa_i(\theta_i - u_i)\right)\right) \\ &= \frac{1}{b_0}exp\left(-\frac{1}{2}\Theta^T\mathbf{W}\Theta - \Theta^T\mathbf{Z}X\right) \\ &= \frac{1}{b_0}exp(-E(X,\Theta))\,, \end{aligned}$$

where $E(X,\Theta) = \frac{1}{2}\Theta^T\mathbf{W}\Theta + \Theta^T\mathbf{Z}$, $\Theta = (\theta_1,\theta_2,\cdots,\theta_n)$, $i = 1,2,\cdots,n$, *and* $\mathbf{W}$ *is the diagonal matrix* $Diag\left(\frac{1}{w_1^2},\frac{1}{w_2^2},\cdots,\frac{1}{w_n^2}\right)$ *and* $\mathbf{Z}$ *is a matrix such that the threshold value is* $\Theta^T\mathbf{Z}X = c_0 X exp\left(-\sum_{i=1}^{n}\kappa_i(\theta_i - u_i)\right)$. *Written this way,* $G(X,\Theta) = \frac{1}{b_0}e^{-E(X,\Theta)}$ *can be seen as an energy function along the lines of a Boltzmann machine where the constant* b_0 *can be chosen such that* $b_0 = \sum_{X,\Theta} exp(-E(X,\Theta))$, *see for instance [14]. However, with* $G(X,\Theta)$ *as above, the Darwinian system is strictly not a Boltzmann machine since the diagonal terms of the matrix W are nonzero and the off-diagonal terms are zero. In fact, in a Boltzmann machine, the opposite is true, that is, the diagonal terms are zero and the off-diagonal terms are nonzero.*

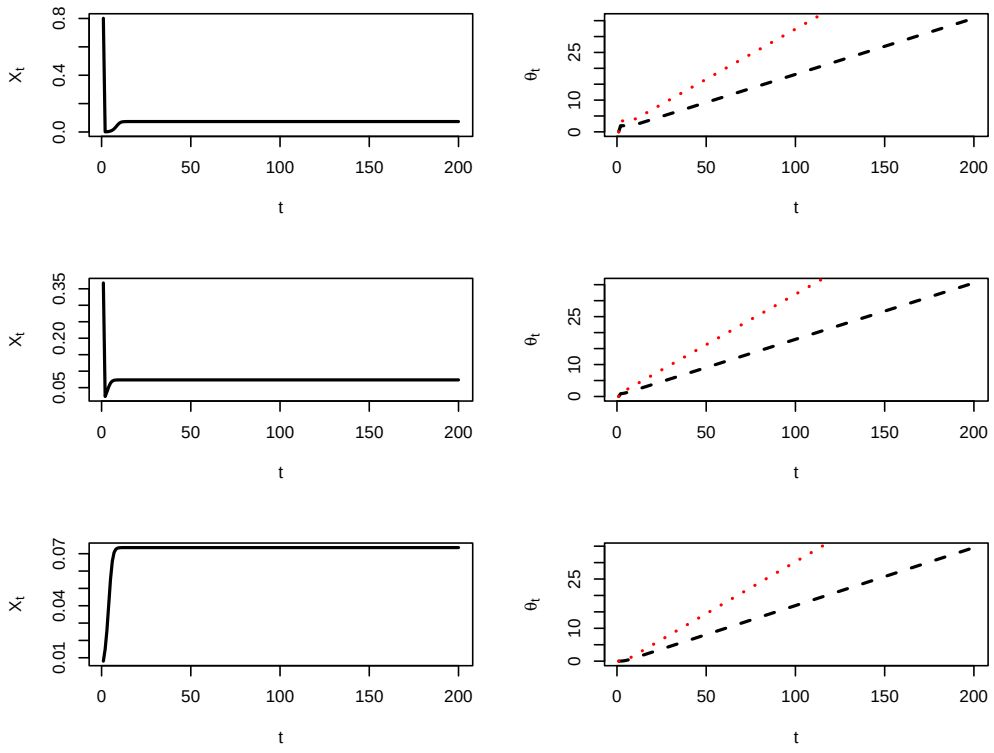
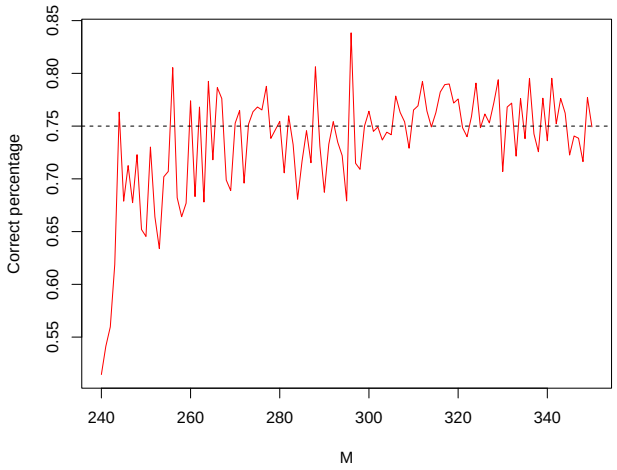
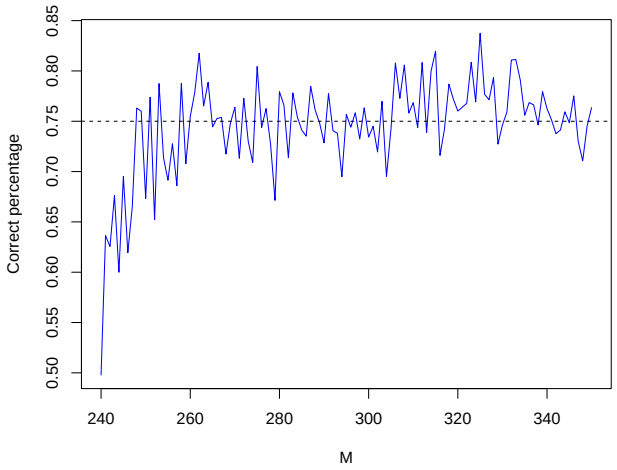


Figure 2. In the **left** column of this figure, represented is $X_t^{(m)}$ for $m = 1,2,3$ and $t = 1,2\cdots,200$. The **right** column represents the samples θ_{1t} and θ_{2t} respectively. From a dynamical system point of view, this is a situation where there is no fixed point (x,Θ) even though there is a fixed point for x_t.



(**a**)



(**b**)

Figure 3. (**a**) Percentage of correct learning for the vector W as a function of the number of samples used. The fluctuations are normal and are due to the random selection of samples at each iteration. We observe that despite fluctuating, the percentage of correct learning increases with the number of samples and hovers around 75%. This is quite remarkable given that all the system's parameters are randomly reshuffled at each epoch. (**b**) Percentage of correct learning for the vector K. The percentage of correct learning for K also hovers around 75% which shows in general that supervised learning can be quite effective at learning the key parameters in a Darwinian evolution dynamics model.

4. Conclusions

In this paper, we have shown how to estimate trait parameters in a Darwinian evolution dynamics model with one species and multiple traits under supervised and unsupervised learning. The procedure can be implemented using a regular gradient or a stochastic gradient descent. We have shown the similarity between the proposed procedure and Boltzmann machine learning, even though the type of energy function is not the same as that of a Boltzmann machine in the strictest sense of the term. The techniques proposed in this paper could certainly be adaptable to readily available data. This is to say, this is a proof of concept meant to kickstart the conversation on how to bring modern techniques of estimation to important problems of evolution theory with much-needed mathematical rigor.

Funding: This research was funded by The American Mathematical Society and Simmons Foundation, grant number AMS-SIMMONS-PUI-23028GR.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: I would like to acknowledge Cleves Epoh Nsali Ewonga for invaluable comments that enhanced the quality of this manuscript.

Conflicts of Interest: The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Appendix A.1. Proof of Theorem 1

Proof. We recall that

$$\mathbb{P}(X_t|\Theta_t, W, K) = \frac{1}{b_0} e^{-b(X_t,\Theta_t,W,K)} . \tag{A1}$$

Therefore, using the definition of condition probability, we have from Equation (A1) that

$$\mathbb{P}(\Theta_t|X_t, W, K) = \frac{\mathbb{P}(X_t|\Theta_t, W, K)\mathbb{P}(\Theta_t)}{Z} , \tag{A2}$$

where

$$Z = \sum_{\Theta_t=\theta_t} \mathbb{P}(X_t|\Theta_t = \theta_t, W, K)\mathbb{P}(\Theta_t = \theta_t) .$$

For simplicity, we write $Z = \sum_{\Theta_t} \mathbb{P}(X_t|\Theta_t, W, K)\mathbb{P}(\Theta_t)$.

We recall that Z is also the marginal $\mathbb{P}(X_t, W, K) = \sum_{\Theta_t} \mathbb{P}(X_t|\Theta_t, W, K)\mathbb{P}(\Theta_t)$.

Therefore, it follows that

$$\ln(\mathbb{P}(\Theta_t|X_t, W, K)) = \ln(\mathbb{P}(X_t|\Theta_t, W, K)) + \ln(\mathbb{P}(\Theta_t)) - \ln\left(\sum_{\Theta_t} \mathbb{P}(X_t|\Theta_t, W, K)\mathbb{P}(\Theta_t)\right) . \tag{A3}$$

Now, we fix $\Theta_t = \Theta_*$ and $X_t = X_*$. For a given $l = 1, 2, \cdots, n$ and $x = w_l$ or $x = \kappa_l$, it follows from (A3) that

$$\frac{\partial \ln(\mathbb{P}(\Theta_*|X_*, W, K))}{\partial x} = \frac{\frac{\partial \mathbb{P}(X_*|\Theta_*, W, K)}{\partial x}}{\mathbb{P}(X_*|\Theta_*|, W, K)} - \frac{\sum_{\Theta_t} \frac{\partial}{\partial x}\mathbb{P}(X_*|\Theta_t, W, K)\mathbb{P}(\Theta_t)}{\sum_{\Theta_t} \mathbb{P}(X_*|\Theta_t, W, K)\mathbb{P}(\Theta_t)} . \tag{A4}$$

From Equation (A1), we obtain

$$\frac{\partial \mathbb{P}(X_*|\Theta_*, W, K)}{\partial x} = -\frac{\partial b(X_*, \Theta_*, W, K)}{\partial x}\mathbb{P}(X_*|\Theta_*, W, K) . \tag{A5}$$

The latter implies that

$$\frac{\frac{\partial \mathbb{P}(X_*|\Theta_*, W, K)}{\partial x}}{\mathbb{P}(X_*|\Theta_*|, W, K)} = -\frac{\partial b(X_*, \Theta_*, W, K)}{\partial x} . \tag{A6}$$

And also that

$$\sum_{\Theta_t} \frac{\partial}{\partial x}\mathbb{P}(X_*|\Theta_t, W, K)\mathbb{P}(\Theta_t) = -\sum_{\Theta_t} \frac{\partial b(X_*, \Theta_t, W, K)}{\partial x}\mathbb{P}(X_*|\Theta_t, W, K)\mathbb{P}(\Theta_t) . \tag{A7}$$

Using the definition of conditional probability, marginal, and Equation (A7), we have

$$\begin{aligned}\frac{\sum_{\Theta_t}\frac{\partial}{\partial x}\mathbb{P}(X_*|\Theta_t,W,K)\mathbb{P}(\Theta_t)}{\sum_{\Theta_t}\mathbb{P}(X_*|\Theta_t,W,K)\mathbb{P}(\Theta_t)} &= \sum_{\Theta_t}\frac{\frac{\partial}{\partial x}\mathbb{P}(X_*|\Theta_t,W,K)\mathbb{P}(\Theta_t)}{\sum_{\Theta_t}\mathbb{P}(X_*|\Theta_t,W,K)\mathbb{P}(\Theta_t)}\\ &= -\sum_{\Theta_t}\frac{\partial b(X_*,\Theta_t,W,K)}{\partial x}\frac{\mathbb{P}(X_*|\Theta_t,W,K)\mathbb{P}(\Theta_t)}{\mathbb{P}(X_*,W,K)}\\ &= -\sum_{\Theta_t}\frac{\partial b(X_*,\Theta_t,W,K)}{\partial x}\frac{\mathbb{P}(X_*,\Theta_t,W,K)}{\mathbb{P}(X_*,W,K)}\\ &= -\sum_{\Theta_t}\frac{\partial b(X_*,\Theta_t,W,K)}{\partial x}\mathbb{P}(\Theta_t|X_*,W,K)\,.\end{aligned}$$

Consequently, Equation (A4) becomes

$$\frac{\partial\ln(\mathbb{P}(\Theta_*|X_*,W,K))}{\partial x} = -\frac{\partial b(X_*,\Theta_*,W,K)}{\partial x} + \sum_{\Theta_t}\frac{\partial b(X_*,\Theta_t,W,K)}{\partial x}\mathbb{P}(\Theta_t|X_*,W,K)\,. \quad \text{(A8)}$$

To summarize, respectively, for $x = w_l$ and $w = \kappa_l$, we have

$$\begin{aligned}\frac{\partial\ln(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)},W,K))}{\partial w_l} &= -\frac{\partial b(X_{t+1}^{(m)},\Theta_{t+1}^{(m)},W,K)}{\partial w_l} + \sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial w_l}\mathbb{P}(\Theta_t|X_{t+1}^{(m)},W,K)\\ \frac{\partial\ln(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)},W,K))}{\partial \kappa_l} &= -\frac{\partial b(X_{t+1}^{(m)},\Theta_{t+1}^{(m)},W,K)}{\partial \kappa_l} + \sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial \kappa_l}\mathbb{P}(\Theta_t|X_{t+1}^{(m)},W,K)\,.\end{aligned} \quad \text{(A9)}$$

A classical gradient procedure will update the weights W and K by moving oppositely to the gradient of the average sample KL divergence, for selected learning rates α_W and α_K, as follows:

$$\begin{cases} w_{l,new} &= w_{l,old} + \alpha_W\dfrac{\partial\ln(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)},W,K))}{\partial w_l}\\ \kappa_{l,new} &= \kappa_{l,old} + \alpha_K\dfrac{\partial\ln(\mathbb{P}(\Theta_{t+1}^{(m)}|X_{t+1}^{(m)},W,K))}{\partial \kappa_l}\end{cases}. \quad \text{(A10)}$$

Since from Equation (A9), we have to run through all values of Θ_t, this procedure may not be suitable for large networks. With a stochastic gradient procedure, the quantity $\sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial x}\mathbb{P}(\Theta_t|X_{t+1}^{(m)},W,K)$ is replaced with a single realization $\frac{\partial b(X_{t+1}^{(m)},\Theta_t^{(m)},W,K)}{\partial x}$. Hence, we obtain a stochastic gradient scheme

$$\begin{cases} w_{l,new} &= w_{l,old} - \alpha_W\left[\dfrac{\partial b(X_{t+1}^{(m)},\Theta_{t+1}^{(m)},W,K)}{\partial w_l} - \dfrac{\partial b(X_{t+1}^{(m)},\Theta_t^{(m)},W,K)}{\partial w_l}\right]\\ \kappa_{l,new} &= \kappa_{l,old} - \alpha_K\left[\dfrac{\partial b(X_{t+1}^{(m)},\Theta_{t+1}^{(m)},W,K)}{\partial \kappa_l} - \dfrac{\partial b(X_{t+1}^{(m)},\Theta_t^{(m)},W,K)}{\partial \kappa_l}\right]\end{cases}, \quad \text{(A11)}$$

for carefully chosen constants α_w and α_K. □

Appendix A.2. Proof of Theorem 2

Proof. The joint probability distribution $\mathbb{P}(X_t, \Theta_t, W, K)$ of X_t and Θ_t given W, K is

$$\mathbb{P}(X_t, \Theta_t, W, K) = \frac{1}{Z} e^{-b(X_t, \Theta_t, W, K)}, \tag{A12}$$

where Z is the constant $Z = \sum_{X_t, \Theta_t} e^{-b(X_t, \Theta_t, W, K)}$.
We also have the marginal of X_t given W and K as

$$\mathbb{P}(X_t, W, K) = \sum_{\Theta_t} \mathbb{P}(X_t, \Theta_t, W, K), \tag{A13}$$

and the condition probability

$$\mathbb{P}(\Theta_t | X_t, W, K) = \frac{\mathbb{P}(X_t, \Theta_t, W, K)}{\mathbb{P}(X_t, W, K)}, \tag{A14}$$

Now, we fix $X_t = X_*$. Let $l = 1, 2, \cdots, n$ and $x = w_l$ or $x = \kappa_l$. From Equation (A13) and logarithm differentiation, we have that

$$\frac{\partial}{\partial x} \ln(\mathbb{P}(X_*, W, K)) = \frac{\sum_{\Theta_t} \frac{\partial}{\partial x} \mathbb{P}(X_*, \Theta_t, W, K)}{\mathbb{P}(X_*, W, K)} \tag{A15}$$

Using the quotient rule in Equation (A12), we obtain

$$\begin{aligned}
\frac{\partial}{\partial x} \mathbb{P}(X_*, \Theta_t, W, K) &= \frac{\partial}{\partial x} \left(\frac{e^{-b(X_*, \Theta_t, W, K)}}{Z} \right) \\
&= \frac{\frac{\partial}{\partial x} e^{-b(X_*, \Theta_t, W, K)} \, Z}{Z^2} \\
&- \frac{e^{-b(X_*, \Theta_t, W, K)} \left(\sum_{X_t, \Theta_t} \frac{\partial}{\partial x} e^{-b(X_t, \Theta_t, W, K)} \right)}{Z^2} \\
&= -\frac{\partial b(X_*, \Theta_t, W, K)}{\partial x} \frac{e^{-b(X_*, \Theta_t, W, K)}}{Z} \\
&+ \frac{e^{-b(X_*, \Theta_t, W, K)}}{Z} \frac{\left(\sum_{X_t, \Theta_t} \frac{\partial b(X_t, \Theta_t, W, K)}{\partial x} e^{-b(X_t, \Theta_t, W, K)} \right)}{Z} \\
&= -\frac{\partial b(X_*, \Theta_t, W, K)}{\partial x} \mathbb{P}(X_*, \Theta_t, W, K) \\
&+ \mathbb{P}(X_*, \Theta_t, W, K) \left(\sum_{X_t, \Theta_t} \frac{\partial b(X_t, \Theta_t, W, K)}{\partial x} \mathbb{P}(X_t, \Theta_t, W, K) \right)
\end{aligned}$$

It follows from the above and Equation (A15) that

$$\begin{aligned}
\frac{\partial}{\partial x}\ln\Big(\mathbb{P}(X_{t+1}^{(m)},W,K)\Big) &= \frac{\sum_{\Theta_t}\frac{\partial}{\partial x}\mathbb{P}(X_{t+1}^{(m)},\Theta_t,W,K)}{\mathbb{P}(X_{t+1}^{(m)},W,K)}\\
&= \sum_{\Theta_t}\frac{\frac{\partial}{\partial x}\mathbb{P}(X_{t+1}^{(m)},\Theta_t,W,K)}{\mathbb{P}(X_{t+1}^{(m)},W,K)}\\
&= -\sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial x}\,\frac{\mathbb{P}(X_{t+1}^{(m)},\Theta_t,W,K)}{\mathbb{P}(X_{t+1}^{(m)},W,K)}\\
&+ \sum_{\Theta_t}\frac{\mathbb{P}(X_{t+1}^{(m)},\Theta_t,W,K)}{\mathbb{P}(X_{t+1}^{(m)},W,K)}\left(\sum_{X_t,\Theta_t}\frac{\partial b(X_t,\Theta_t,W,K)}{\partial x}\mathbb{P}(X_t,\Theta_t,W,K)\right)\\
&= -\sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial x}\,\mathbb{P}(\Theta_t|X_{t+1}^{(m)},\Theta_t,W,K)\\
&+ \frac{\sum_{\Theta_t}\mathbb{P}(X_{t+1}^{(m)},\Theta_t,W,K)}{\mathbb{P}(X_{t+1}^{(m)},W,K)}\left(\sum_{X_t,\Theta_t}\frac{\partial b(X_t,\Theta_t,W,K)}{\partial x}\mathbb{P}(X_t,\Theta_t,W,K)\right)\\
&= -\sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial x}\,\mathbb{P}(\Theta_t|X_{t+1}^{(m)},\Theta_t,W,K)\\
&+ \frac{\mathbb{P}(X_{t+1}^{(m)},W,K)}{\mathbb{P}(X_{t+1}^{(m)},W,K)}\left(\sum_{X_t,\Theta_t}\frac{\partial b(X_t,\Theta_t,W,K)}{\partial x}\mathbb{P}(X_t,\Theta_t,W,K)\right)\\
&= -\sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial x}\,\mathbb{P}(\Theta_t|X_{t+1}^{(m)},\Theta_t,W,K)\\
&+ \sum_{X_t,\Theta_t}\frac{\partial b(X_t,\Theta_t,W,K)}{\partial x}\mathbb{P}(X_t,\Theta_t,W,K)\,.
\end{aligned}$$

To summarize, respectively, for $x = w_l$ and $w = \kappa_l$, we have

$$\begin{cases}
\frac{\partial}{\partial w_l}\ln\Big(\mathbb{P}(X_{t+1}^{(m)},W,K)\Big) &= -\sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial w_l}\mathbb{P}(\Theta_t|X_{t+1}^{(m)},\Theta_t,W,K)\\
&\quad+\sum_{X_t,\Theta_t}\frac{\partial b(X_t,\Theta_t,W,K)}{\partial w_l}\mathbb{P}(X_t,\Theta_t,W,K)\\
\frac{\partial}{\partial \kappa_l}\ln\Big(\mathbb{P}(X_{t+1}^{(m)},W,K)\Big) &= -\sum_{\Theta_t}\frac{\partial b(X_{t+1}^{(m)},\Theta_t,W,K)}{\partial \kappa_l}\mathbb{P}(\Theta_t|X_{t+1}^{(m)},\Theta_t,W,K)\\
&\quad+\sum_{X_t,\Theta_t}\frac{\partial b(X_t,\Theta_t,W,K)}{\partial \kappa_l}\mathbb{P}(X_t,\Theta_t,W,K)
\end{cases} \tag{A16}$$

A classical gradient procedure will update the weights W and K by moving oppositely to the gradient of the average sample KL divergence, for selected learning rates α_W and α_K, as follows:

$$\begin{cases}
w_{l,new} &= w_{l,old} + \alpha_W\dfrac{\partial\ln\Big(\mathbb{P}(X_{t+1}^{(m)},W,K)\Big)}{\partial w_l}\\
\kappa_{l,new} &= \kappa_{l,old} + \alpha_K\dfrac{\partial\ln\Big(\mathbb{P}(X_{t+1}^{(m)},W,K)\Big)}{\partial \kappa_l}
\end{cases}. \tag{A17}$$

The classical approach requires to run through all values of X_t and Θ_t, again not very suitable for large networks. With a stochastic gradient procedure, the quantity

$\sum_{\Theta_t} \frac{\partial b(X_{t+1}^{(m)}, \Theta_t, W, K)}{\partial x} \mathbb{P}(\Theta_t | X_{t+1}^{(m)}, W, K)$ is replaced with a single random realization $\frac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial x}$, whereas the quantity $\sum_{X_t, \Theta_t} \frac{\partial b(X_t, \Theta_t, W, K)}{\partial \kappa_l} \mathbb{P}(X_t, \Theta_t, W, K)$ is replace with a single random realization $\frac{\partial b(X_t^{(m)}, \Theta_t^{(m)}, W, K)}{\partial x}$. Hence, we obtain a stochastic gradient scheme

$$\begin{cases} w_{l,new} & = w_{l,old} - \alpha_W \left[\frac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial w_l} - \frac{\partial b(X_t^{(m)}, \Theta_t^{(m)}, W, K)}{\partial w_l} \right] \\ \kappa_{l,new} & = \kappa_{l,old} - \alpha_K \left[\frac{\partial b(X_{t+1}^{(m)}, \Theta_t^{(m)}, W, K)}{\partial \kappa_l} - \frac{\partial b(X_t^{(m)}, \Theta_t^{(m)}, W, K)}{\partial \kappa_l} \right] \end{cases}, \tag{A18}$$

for carefully chosen constants α_w and α_K. □

References

1. Darwin, C. *On the Origin of Species by Means of Natural Selection*; John Murray: London, UK, 1859.
2. Vincent, T.L.; Vincent, T.L.S.; Cohen, Y. Darwinian dynamics and evolutionary game theory. *J. Biol. Dyn.* **2011**, *5*, 215–226. [CrossRef]
3. Ackleh, A.S.; Cushing, J.M.; Salceneau, P.L. On the dynamics of evolutionary competition models. *Nat. Resour. Model.* **2015**, *28*, 380–397. [CrossRef]
4. Cushing, J.M. Difference equations as models of evolutionary population dynamics. *J. Biol. Dyn.* **2019**, *13*, 103–127. [CrossRef] [PubMed]
5. Cushing, J.M.; Park, J.; Farrell, A.; Chitnis, N. Treatment of outocme in an si model with evolutionary resistance: A darwinian model for the evolutionary resistance. *J. Biol. Dyn.* **2023**, *17*, 2255061. [CrossRef] [PubMed]
6. Elaydi, S.; Kang, Y.; Luis, R. The effects of evolution on the stability of competing species. *J. Biol. Dyn.* **2022**, *16*, 816–839. [CrossRef] [PubMed]
7. Fisher, R.A. On the mathematical foundation of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser.* **1922**, *222*, 594–604.
8. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 623–656. [CrossRef]
9. Kwessi, E. Information theory in a darwinian evolution population dynamics model. *arXiv* **2024**, arXiv:2403.05044.
10. Shun-ich, A.; Horishi, N. *Methods of Information Geometry*; chapter Chentsov Theorem and Some Historical Remarks; Oxford University Press: Oxford, UK, 2000; pp. 37–40.
11. Dowty, J.G. Chentsov theorem for exponential families. *Inf. Geom.* **2018**, *1*, 117–135. [CrossRef]
12. Harper, M. Information geometry and evolution game theory. *arXiv* **2009**, arXiv:0911.1383. [CrossRef]
13. Kimura, M. On the change of population fitness by natural selection. *Heredity* **1958**, *12*, 145–167. [CrossRef]
14. Hinton, G.E.; Sejnowski, T.J. Optimal perceptual inference. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 19–23 June 1983; pp. 448–453.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.