Eigen Artificial Neural Networks

Francisco Yepes Barrera
I-CON Srl, Verona (Italy)

paco.yepes@godelia.org

***Abstract*—This work has its origin in intuitive physical and statistical considerations. The problem of optimizing an artificial neural network is treated as a physical system, composed of a conservative vector force field. The derived scalar potential is a measure of the potential energy of the network, a function of the distance between predictions and targets.**

Starting from some analogies with wave mechanics, the description of the system is justified with an eigenvalue equation that is a variant of the Schrõdinger equation, in which the potential is defined by the mutual information between inputs and targets. The weights and parameters of the network, as well as those of the state function, are varied so as to minimize energy, using an equivalent of the variational theorem of wave mechanics. The minimum energy thus obtained implies the principle of minimum mutual information (MinMI). We also propose a definition of the work produced by the force field to bring a network from an arbitrary probability distribution to the potential-constrained system. At the end of the discussion we expose a recursive procedure that allows to refine the state function and bypass some initial assumptions.

The results demonstrate how the minimization of energy effectively leads to a decrease in the average error between network and target predictions.

***Index Terms*—Artificial Neural Networks optimization, variational techniques, Minimum Mutual Information Principle, Wave Mechanics, eigenvalue problem.**

I. Introduction

This paper analyzes the problem of optimizing artificial neural networks (ANNs), ie the problem of finding functions $y(\mathbf{x};\Gamma)$, dependent on matrixes of input data $\mathbf{x}$ and parameters Γ, such that, given a target $\mathbf{t}$ make an optimal mapping between $\mathbf{x}$ and $\mathbf{t}$.

The starting point of this article is made up of some well-known theoretical elements:

1) The training of an artificial neural network consists in the minimization of some error function between the output of the network $y(\mathbf{x};\Gamma)$ and the target $\mathbf{t}$. In the best case it identify the global minimum of the error; in general it finds local minima. The total of minimums forms a discrete set of values.
2) The passage from a prior to a posterior or conditional probability, that is the observation or acquisition of additional knowledge about data, implies a collapse of the function that describes the system: the conditional probability calculated with Bayes' theorem leads to distributions of closer and more localized probabilities than prior ones [2].
3) Starting from the formulation of the mean square error produced by an artificial neural network and considering a set C of targets t_k whose distributions are independent

$$p(\mathbf{t}|\mathbf{x}) = \prod_{k=1}^{C} p(t_k|\mathbf{x})$$

$$p(\mathbf{t}) = \prod_{k=1}^{C} p(t_k)$$

with $p(t_k|\mathbf{x})$ the conditional probability of t_k given $\mathbf{x}$ and $p(t_k)$ the marginal probability of t_k, it can be shown that

$$\left\langle (y_k - t_k)^2 \right\rangle \geq \left\langle (\langle t_k|\mathbf{x}\rangle - t_k)^2 \right\rangle \qquad \text{(I.1)}$$

being $\langle t_k|\mathbf{x}\rangle$ the expected value or conditional average of t_k given $\mathbf{x}$, and the equal valid only at the optimum. In practice, any trial function $y_k(\mathbf{x};\Gamma)$ leads to a quadratic deviation with respect to t_k greater than that generated by the optimal function, $y'_k = y_k(\mathbf{x};\Gamma')$, corresponding to the absolute minimum of the error, since this represents the conditional average of the target, as demonstrated by the known result [2]

$$y'_k = \langle t_k|\mathbf{x}\rangle \qquad \text{(I.2)}$$

These three points can be directly related to three theoretical elements at the base of wave mechanics [8]:

1) Any physical system described by the Schrõdinger equation and constrained by a scalar potential $V(\mathbf{x})$ leads to a quantization of energy values, which constitute a discrete set of real values.
2) A quantum-mechanical system is formed by the superposition of a series of states

described by the Schrõdinger equation, corresponding to as many eigenvalues. The observation of the system causes the collapse of the wave function on one of the states, being only possible to calculate the probability of obtaining the different eigenvalues.

3) When it is not possible to analytically obtain the true wave function Ψ' and the true energy E' of a quantum-mechanical system, it is possible to use trial functions Ψ, with eigenvalues E, dependent on a set Γ of parameters. In this case we can find an approximation to Ψ' and E' varying Γ and taking into account the variational theorem

$$\left\{ E = \frac{\int \Psi^* \hat{H} \Psi \, d\mathbf{x}}{\int \Psi^* \Psi \, d\mathbf{x}} \right\} \geq E'$$

Regarding point 3, we can consider the condition (I.1) as an equivalent of the variational theorem for artificial neural networks.

II. Treatment of the optimization of artificial neural networks as an eigenvalue problem

The analogies highlighted in Section I suggest the possibility of dealing with the problem of optimizing artificial neural networks as a physical system. Analysis attempts using models from mathematical physics are not new [9]. The analogies are studied in this work to understand if it is possible to model the ANNs optimization problem with eigenvalue equations, as happens in the physical systems modeled by the Schrõdinger equation. This model allows to define the energy of the network, a concept already used in some types of neural networks, such as the Hopfield networks in which Lyapunov or energy functions can be derived for binary elements networks allowing a complete characterization of their dynamics . We will generalize the concept of energy for any type of ANN.

Suppose we can define a conservative force generated by the set of targets $\mathbf{t}$, represented in the input space $\mathbf{x}$ with a vector field, being N the dimensionality of $\mathbf{x}$. In this case we have a scalar function $V(\mathbf{x})$, called potential, which depends exclusively on the position and which is defined as

$$\mathbf{F} = -\nabla V(\mathbf{x}) \tag{II.1}$$

which implies that the potential of the force at a point is proportional to the potential energy possessed by an object at that point due to the presence of force. The negative sign in the equation (II.1) means that the force is directed towards the target, where the force is maximum and the potential is minimal, so $\mathbf{t}$ generates an attractive force that attracts the *bodies* immersed in the field, represented by the average predictions of the network, with an intensity proportional to a function of the distance between $y(\mathbf{x};\Gamma)$ and $\mathbf{t}$.

The equation (I.2) highlights how, at the optimum, the output of an artificial neural network is an approximation to the conditional averages or expected values of the targets given the input $\mathbf{x}$. Both $\mathbf{x}$ and $\mathbf{t}$ are given by the problem, with average values that do not vary over time. We can therefore hypothesize a stationary system and an eigenvalue equation independent of time, having the same structure as the Schrõdinger equation

$$-\epsilon \nabla^2 \Psi + \zeta V'(\mathbf{x}) \Psi = E \Psi \tag{II.2}$$

with Ψ the state function of system (network), $V(\mathbf{x}) = \zeta V'(\mathbf{x})$ a scalar potential, E the network energy, ϵ a multiplicative constant and a ζ a variational parameter. ζ seems necessary since the equation (II.2) does not arise from a true physical system, so the relative values between the first and second terms of the first member are unknown. Preliminary calculations show that for some problems the value of $V'(\mathbf{x})$ can be very small compared to the first term. We will consider that $\zeta \in \mathbb{N}$ and is dimensionless.

The equation (II.2) implements a parametric model for the ANNs in which the optimization consists in minimizing, on average, the energy of the network, function of $y(\mathbf{x};\Gamma)$ and $\mathbf{t}$, modeled by appropriate probability densities and a set of variational parameters Γ. The working hypothesis is that the minimization of energy through a parameter-dependent trial function that makes use of the variational theorem (I.1) leads, using an appropriate potential, to a reduction of the error committed by the network in the prediction of $\mathbf{t}$.

III. The potential $V(\mathbf{x})$

A function that satisfies all the requirements exposed to be used as potential is the mutual information, $I(\mathbf{t},\mathbf{x})$ [14], which is a positive quantity. In this case, the minimization of energy through a variational state function that satisfies the equation (II.2) implies the principle of minimum mutual information (*MinMI*) [4, 6, 7, 15], equivalent to the principle of maximum entropy (*MaxEnt*) [3, 10, 11, 13]. The scalar potential depends only on the vector $\mathbf{x}$ and for C targets

becomes[1]

$$\begin{aligned} V(\mathbf{x}) &= \zeta \sum_{k=1}^{C} \int I_k(t_k, \mathbf{x})\, dt_k \\ &= \zeta \sum_{k=1}^{C} \int p(t_k|\mathbf{x}) p(\mathbf{x}) \ln\left(\frac{p(t_k|\mathbf{x})}{p(t_k)}\right) dt_k \end{aligned} \tag{III.1}$$

The equation (III.1) assumes a superposition principle, similar to the valid one in the electric field, in which the total potential is given by the sum of the potentials with respect to each of the C targets of the problem.

Considering that the network provides an approximation to the target t_k given by a deterministic function $y_k(\mathbf{x};\Gamma)$ with a noise ε_k, $t_k = y_k + \varepsilon_k$, and considering that the error ε_k is normally distributed with mean zero, the conditional probability $p(t_k|\mathbf{x})$ can be written as [2]

$$p(t_k|\mathbf{x}) = \frac{1}{(2\pi\chi_k^2)^{1/2}} \exp\left\{-\frac{(y_k - t_k)^2}{2\chi_k^2}\right\} \tag{III.2}$$

Note that χ_k is the standard deviation of $y_k(\mathbf{x};\Gamma)$ and $\chi_k = \chi_k(\mathbf{x})$, so $\chi_k \notin \Gamma$. To be able to integrate the differential equation (II.2) we will consider the vector $\vec{\chi}$ constant. We will see at the end of the discussion that it is possible to obtain an expression for χ_k dependent on $\mathbf{x}$, which allows us to derive a more precise description of the potential.

We also write unconditional probabilities for inputs and targets as Gaussians to simplify the mathematical treatment

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{(2\pi)^{N/2}|\sum|^{1/2}} \times \\ &\quad \exp\left\{-\frac{1}{2}(\mathbf{x} - \vec{\mu})^T \sum{}^{-1} (\mathbf{x} - \vec{\mu})\right\} \\ p(t_k) &= \frac{1}{(2\pi\theta_k^2)^{1/2}} \exp\left\{-\frac{(t_k - \rho_k)^2}{2\theta_k^2}\right\} \end{aligned} \tag{III.3}$$

Considering the absence of correlation between the N input variables, the probability $p(\mathbf{x})$ is reduced to

$$\begin{aligned} p(\mathbf{x}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right\} \\ &= \prod_{i=1}^{N} \mathcal{N}(\mu_i; \sigma_i^2) \end{aligned} \tag{III.4}$$

with, in this case, $|\Sigma|^{1/2} = \prod_{i=1}^{N} \sigma_i$, representing with $\mathcal{N}(\mu_i;\sigma_i^2)$ the Gaussian with mean μ_i and variance σ_i^2 relative to the component x_i of the vector $\mathbf{x}$. The equations (III.3) e (III.4) introduce in the model a statistical description of the problem starting from the observed data, through the set of constants $\vec{\rho}$, $\vec{\theta}$, $\vec{\mu}$ e $\vec{\sigma}$.

The integration of the equation (III.1) over t_k gives

$$V(\mathbf{x}) = \zeta \prod_{i=1}^{N} \mathcal{N}(\mu_i; \sigma_i^2) \sum_{k=1}^{C} \left(\alpha_k y_k^2 - \beta_k y_k + \gamma_k\right) \tag{III.5}$$

[1] When not specified, we will implicitly assume that integrals extend to all space in the interval $[-\infty : \infty]$.

with

$$\alpha_k = \frac{1}{2\theta_k^2},\ \beta = \frac{\rho_k}{\theta_k^2},\ \gamma_k = \frac{\rho_k^2 + \chi_k^2}{2\theta_k^2} - \ln\frac{\chi_k\sqrt{e}}{\theta_k} \tag{III.6}$$

Mutual information in the potential (III.1) is expressed in nats. We will call the units of energy calculated from (II.2) nats of energy or *enats*.

It is known that a linear combination of Gaussians can approximate an arbitrary function. Using a base of dimension P we can write the following expression for $y_k(\mathbf{x};\Gamma)$

$$y_k(\mathbf{x};\mathbf{w}) = \sum_{p=1}^{P} w_{kp}\phi_p(\mathbf{x}) + w_{k0} \tag{III.7}$$

with

$$\begin{aligned} \phi_p(\mathbf{x}) &= \exp\left\{-\xi_p \left\|\mathbf{x} - \vec{\omega}_p\right\|^2\right\} \\ &= \prod_{i=1}^{N} \exp\left\{-\xi_p (x_i - \omega_{pi})^2\right\} \end{aligned} \tag{III.8}$$

and w_{k0} the bias term for the output unit k. The equations (III.7) e (III.8) propose a model of neural network of type RBF (Radial Basis Function), which contain a single hidden layer and allow to facilitate the calculation given the complexity of the model.

Taking into account the equations (II.1), (III.5) e (III.7) the components of the force, F_i, are given by

$$\begin{aligned} &F_i = \zeta \prod_{i=1}^{N} \mathcal{N}(\mu_i;\sigma_i^2) \sum_{k=1}^{C} \left\{\alpha_k \tfrac{x_i-\mu_i}{\sigma_i^2} w_{k0}^2 + \right. \\ &2\alpha_k w_{k0} \sum_{p=1}^{P} w_{kp} \left(2\xi_p (x_i - \omega_{pi}) + \tfrac{x_i-\mu_i}{\sigma_i^2}\right) \phi_p + \\ &2\alpha_k \sum_{p=1}^{P}\sum_{q=1}^{P} w_{kp} w_{kq} \xi_q (x_i - \omega_{qi}) \phi_p \phi_q + \\ &2\alpha_k \sum_{p=1}^{P}\sum_{q=1}^{P} w_{kp} w_{kq} \xi_p (x_i - \omega_{pi}) \phi_p \phi_q + \\ &\alpha_k \sum_{p=1}^{P}\sum_{q=1}^{P} w_{kp} w_{kq} \tfrac{x_i-\mu_i}{\sigma_i^2} \phi_p \phi_q - \\ &\beta_k \tfrac{x_i-\mu_i}{\sigma_i^2} w_{k0} + \gamma_k \tfrac{x_i-\mu_i}{\sigma_i^2} - \\ &\left.\beta_k \sum_{p=1}^{P} w_{kp} \left(2\xi_p (x_i - \omega_{pi}) + \tfrac{x_i-\mu_i}{\sigma_i^2}\right) \phi_p \right\} \end{aligned}$$

In physical conservative fields, work, W, is defined as the minus difference between the potential energy of a body subject to the forces of the field and that possessed by the body at a reference point, $W = -\Delta V(\mathbf{x})$. In some types of central force fields, as in the electrostatic or gravitational cases, the reference point is located at an infinite distance from the source where, given the dependence of V on $\frac{1}{r}$, the potential energy is zero.

Since in the discrete case the mutual information is limited superiorly from the minimum among the marginal entropies, $h(\mathbf{x})$ and $h(\mathbf{t})$,[2] given that the distribution with maximum entropy is the uniform one, U, and that the reference point

[2] We use h in lower case as discrete entropy to distinguish it from H, which in this work is used as a symbol of the Hamiltonian operator and of the integrals H_{mn}.

against which to calculate the potential difference is arbitrary, we can propose the following definition of work[3]

$$W = C\zeta h(U) - \int V(\mathbf{x})\, d\mathbf{x} \qquad \text{(III.9)}$$

For $W > 0$, the equation (III.9) explains the work, in enats, carried out by the forces of the field to pass from a neural network that realizes uniformly distributed predictions to a network that realizes an approximation to the density $p(\mathbf{t}|\mathbf{x})$.

IV. The State Equation

A dimensional analysis of the potential (III.5) shows that the term $\alpha_k y_k^2 - \beta_k y_k + \gamma_k$ is dimensionless. Thus, the units of the potential are determined by the factor $|\Sigma|^{-1/2}$. To maintain the dimensional coherence in the equation (II.2) we multiplied the first term of the first member by the factor $\frac{\sigma_{\mathbf{x}}^2}{|\Sigma|^{1/2}}$, where[4]

$$\sigma_{\mathbf{x}}^2 \nabla^2 = \sum_{i=1}^{N} \sigma_i^2 \frac{\partial^2}{\partial x_i^2}$$

$\sigma_{\mathbf{x}}^2$ cannot be a constant factor independent of the single components of $\mathbf{x}$ since in general every x_i has its own units and its own variance.

Given the variational parameter ζ in the second term of the first member of the equation (II.2), we can without losing generality multiply the first term by $\frac{1}{2}$, obtaining $\epsilon = -\frac{\sigma_{\mathbf{x}}^2}{2|\Sigma|^{1/2}}$. The Hamiltonian operator

$$\begin{aligned} \hat{H} \quad = \quad & \hat{T} + \hat{V} = -\tfrac{\sigma_{\mathbf{x}}^2}{2|\Sigma|^{1/2}} \nabla^2 + \\ & \zeta \textstyle\prod_{i=1}^{N} \mathcal{N}(\mu_i; \sigma_i^2) \times \\ & \textstyle\sum_{k=1}^{C} \left(\alpha_k y_k^2 - \beta_k y_k + \gamma_k\right) \end{aligned}$$

is real, linear and hermitian, and has the same structure as that used in the Schrõdinger equation. Hermiticity stems from the condition that the average value of energy is a real value, $\langle E \rangle = \langle E^* \rangle$.[5] $\hat{T}$ and $\hat{V}$ represent the operators related respectively to the kinetic and potential components of the Hamiltonian.

The final state equation is

$$\begin{aligned} E\Psi \quad = \quad & -\tfrac{\sigma_{\mathbf{x}}^2}{2|\Sigma|^{1/2}} \nabla^2 \Psi + \\ & \zeta \textstyle\prod_{i=1}^{N} \mathcal{N}(\mu_i; \sigma_i^2) \times \\ & \textstyle\sum_{k=1}^{C} \left(\alpha_k y_k^2 - \beta_k y_k + \gamma_k\right) \Psi \end{aligned} \qquad \text{(IV.1)}$$

Wanting to make an analogy with wave mechanics, we can say that the equation (IV.1) describes the motion of particle of mass $|\Sigma|^{1/2}$ subject to the potential (III.5). $\sigma_{\mathbf{x}}^2$, as happens in quantum mechanics with the Planck constant, has the role of a scale factor: the phenomenon described by the equation (IV.1) is relevant in the range of variance for each single component x_i of the vector $\mathbf{x}$.

We discussed the role of the operator $\hat{V}$: its variation in the space $\mathbf{x}$ implies a force that is directed towards the target where $V(\mathbf{x})$ is minimum and $\mathbf{F}$ is maximum. The operator $\hat{T}$ contains the divergence of a gradient in the space $\mathbf{x}$ and represents the flow density of the Ψ gradient, being a measure of the deviation of the state function at a point with respect to the average of the surrounding points. The role of ∇^2 in the equation (IV.1) is to introduce information about Ψ curvature. In neural networks a similar role is found in the use of the Hessian matrix, calculated in the space of weights, in conventional second order optimization techniques.[6]

Starting from the expected energy value obtained from the equation (IV.1)[7]

$$E = \frac{\int \Psi^* \hat{H} \Psi \, d\mathbf{x}}{\int \Psi^* \Psi \, d\mathbf{x}} \qquad \text{(IV.2)}$$

assuming a base of dimension D for the trial function

$$\Psi(\mathbf{x}) = \sum_{d=1}^{D} c_d \psi_d \qquad \text{(IV.3)}$$

with the basis functions developed in a similar way to what we did for y_k in the equations (III.7) and (III.8)

$$\psi_d(\mathbf{x}) = \prod_{i=1}^{N} \exp\left\{-\lambda_d (x_i - \eta_{id})^2\right\} \qquad \text{(IV.4)}$$

and considering the coefficients independent of each other, $\frac{\partial c_m}{\partial c_n} = \delta_{mn}$, the Rayleigh-Ritz method leads to the linear system

$$\sum_n \left[(H_{mn} - S_{mn} E)\, c_n\right] = 0 \qquad \text{(IV.5)}$$

with

$$H_{mn} = \int \psi_m^* \hat{H} \psi_n \, d\mathbf{x} \qquad \text{(IV.6)}$$

$$S_{mn} = \int \psi_m^* \psi_n \, d\mathbf{x} \qquad \text{(IV.7)}$$

[3] The treatment of the text makes considerations on the discrete case since the differential entropy can be negative.

[4] This setting makes it possible to incorporate $|\Sigma|^{-1/2}$ into the value of E, but in the continuation we will leave it explicitly indicated.

[5] In this article we only use real functions, so the hermiticity condition is reduced to the symmetry of the $\mathbf{H}$ and $\mathbf{S}$ matrixes.

[6] In this case, as we will later show, the second derivatives are calculated in the space of the weights and not in the space $\mathbf{x}$.

[7] Although all the functions used in this work are real, we will make their complex conjugates explicit in the equations, as is usual in the wave mechanics formulation.

To obtain a nontrivial solution the determinant of the coefficients have to be zero

$$\det\left(\mathbf{H}-\mathbf{S}E\right)=0 \qquad \text{(IV.8)}$$

which leads to D energies, equal to the size of the base (IV.3). The D energy values E_d represent an upper limit to the first true energies $E_d^{'}$ of the system. The substitution of every E_d in (IV.5) allows to calculate the D coefficients c of Ψ relative to the state d. The lowest value among E_d represents the global optimum of the problem or *fundamental state* that leads, in the hypotheses of this article, to the minimum or global error of the neural network in the prediction of the target $\mathbf{t}$, while the remaining eigenvalues can be interpreted as local minima. It can be shown that the eigenfunctions obtained in this way form an orthogonal set. The variational method we have discussed has a general character and can be applied, in principle, to artificial neural networks of any kind, not bound to any specific functional form for y_k.

The proposed model assumes a change of paradigm with respect to some known methods of optimizing neural networks, such as the gradient descent, which carry out a search in the parameter space, in particular the set of weights $\mathbf{w}$, through the search for a expression for $\frac{\partial \mathrm{E_r}}{\partial \mathbf{w}}$ with E_r a form of error of the neural network. In this article the variables of the problem, and the search in the relative space, are the input $\mathbf{x}$ with $\mathbf{w}\in\Gamma$ a set of variational parameters.

Using the equations (III.7) and (III.8) and taking into account the constancy of $\vec{\chi}$, the integrals (IV.6) and (IV.7) have the following expressions

$$\begin{aligned} S_{mn} &= \left(\tfrac{\pi}{\lambda_n+\lambda_m}\right)^{\frac{N}{2}}\times \\ &\quad \prod_{i=1}^{N}\exp\left\{-\tfrac{\lambda_m\lambda_n}{\lambda_n+\lambda_m}(\eta_{im}-\eta_{in})^2\right\}\end{aligned}$$

$$\begin{aligned} H_{mn} &= -\tfrac{\lambda_m\lambda_n}{|\sum|^{1/2}(\lambda_n+\lambda_m)^2}S_{mn}\times \\ &\sum_{i=1}^{N}\sigma_i^2\left[2\lambda_m\lambda_n(\eta_{im}-\eta_{in})^2-\lambda_n-\lambda_m\right]- \\ &\Lambda\sum_{k=1}^{C}\gamma_k+\Lambda\sum_{k=1}^{C}\beta_k w_{k0}+ \\ &\sum_{k=1}^{C}\beta_k\sum_{p=1}^{P}w_{kp}\Omega-\Lambda\sum_{k=1}^{C}\alpha_k w_{k0}^2- \\ &2\sum_{k=1}^{C}\alpha_k w_{k0}\sum_{p=1}^{P}w_{kp}\Omega- \\ &\sum_{k=1}^{C}\alpha_k\sum_{p=1}^{P}\sum_{q=1}^{P}w_{kp}w_{kq}\Phi\end{aligned}$$

with

$$\begin{aligned}\Lambda &= \prod_{i=1}^{N}\tfrac{\sqrt{2\pi}\sigma_i}{\sqrt{2\sigma_i^2(\lambda_n+\lambda_m)+1}}\times \\ &\exp\left\{-\tfrac{2\sigma_i^2(\eta_{in}-\eta_{im})^2\lambda_m\lambda_n}{2\sigma_i^2(\lambda_n+\lambda_m)+1}\right\}\times \\ &\exp\left\{-\tfrac{(\eta_{in}-\mu_i)^2\lambda_n+(\eta_{im}-\mu_i)^2\lambda_m}{2\sigma_i^2(\lambda_n+\lambda_m)+1}\right\}\end{aligned}$$

$$\begin{aligned}\Omega &= \prod_{i=1}^{N}\left[\tfrac{\sqrt{2\pi}\sigma_i}{\sqrt{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}}\times\right. \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1\right)\xi_p\omega_{pi}^2}{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{\tfrac{\left(4\sigma_i^2\eta_{in}\lambda_n 4\sigma_i^2\eta_{im}\lambda_m 2\mu_i\right)\xi_p}{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\eta_{in}^2\lambda_n+2\sigma_i^2\eta_{im}^2\lambda_m+\mu_i^2\right)\xi_p}{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\eta_{in}^2-4\sigma_i^2\eta_{im}\eta_{in}\right)\lambda_m\lambda_n}{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{2\sigma_i^2\eta_{im}^2\lambda_m\lambda_n}{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\eta_{in}^2\lambda_n-2\mu_i\eta_{in}\lambda_n+\mu_i^2\lambda_n}{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\left.\exp\left\{-\tfrac{\left(\eta_{im}^2-2\mu_i\eta_{im}+\mu_i^2\right)\lambda_m}{2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\right]\end{aligned}$$

$$\begin{aligned}\Phi &= \prod_{i=1}^{N}\left[\tfrac{\sqrt{2\pi}\sigma_i}{\sqrt{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}}\times\right. \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1\right)\xi_q\omega_{qi}^2}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{\tfrac{\left(4\sigma_i^2\xi_p\omega_{pi}4\sigma_i^2\eta_{in}\lambda_n 4\sigma_i^2\eta_{im}\lambda_m 2\mu_i\right)\xi_q\omega_{qi}}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\xi_p\omega_{pi}^2+2\sigma_i^2\eta_{in}^2\lambda_n+2\sigma_i^2\eta_{im}^2\lambda_m+\mu_i^2\right)\xi_q}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1\right)\xi_p}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{\tfrac{\left(4\sigma_i^2\eta_{in}\lambda_n+4\sigma_i^2\eta_{im}\lambda_m+2\mu_i\right)\xi_p\omega_{pi}}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\eta_{in}^2\lambda_n+2\sigma_i^2\eta_{im}^2\lambda_m+\mu_i^2\right)\xi_p}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\left(2\sigma_i^2\eta_{in}^2-4\sigma_i^2\eta_{im}\eta_{in}+2\sigma_i^2\eta_{im}^2\right)\lambda_m\lambda_n}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\exp\left\{-\tfrac{\eta_{in}^2\lambda_n-2\mu_i\eta_{in}\lambda_n+\mu_i^2\lambda_n}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\times \\ &\left.\exp\left\{-\tfrac{\left(\eta_{im}^2-2\mu_i\eta_{im}+\mu_i^2\right)\lambda_m}{2\sigma_i^2\xi_q+2\sigma_i^2\xi_p+2\sigma_i^2\lambda_n+2\sigma_i^2\lambda_m+1}\right\}\right]\end{aligned}$$

The number of variational parameters of the model, n_Γ, is

$$n_\Gamma = C(P+1)+(N+1)P+(D+2)N+D+2C+1 \qquad \text{(IV.9)}$$

The energies obtained by the determinant (IV.8) allow to obtain a system of equations resulting from the condition of minimum

$$\frac{\partial E_d}{\partial \Gamma}=0 \qquad \text{(IV.10)}$$

The system (IV.10) is implicit in χ_k, $\Gamma=\Gamma(\chi_k)$, and must be solved in an iterative way, as χ_k depends on y_k which in turn is a function of Γ.

V. Interpretation of the State Function

The model we have proposed contains two main weaknesses: 1) the normality of the marginal densities $p(\mathbf{x})$ and $p(\mathbf{t})$; 2) the constancy of the vector $\vec{\chi}$. The following discussion tries to resolve the second point.

Similarly to wave mechanics we can interpret the square module of Ψ as a probability. In this

case, the Laplacian operator in equation (IV.1) models a probability flow. Given that we have obtained Ψ from a statistical description of the known targets, we can assume that $|\Psi|^2$ represents the conditional probability of $\mathbf{x}$ given $\mathbf{t}$, subject to the set of parameters Γ

$$p(\mathbf{x}|\mathbf{t},\Gamma) = |\Psi(\mathbf{x})|^2 \tag{V.1}$$

The equation (V.1) is related to the conditional probability $p(\mathbf{t}|\mathbf{x})$ through the Bayes theorem

$$p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})} \tag{V.2}$$

Since we considered the C targets independent, using the expressions (III.2) and (V.1) into (V.2), separating variables and integrating over t_k, assuming that at the optimum is satisfied the condition $\theta_k > \chi_k$, we have

$$\frac{|\Psi(\mathbf{x})|^2}{p(\mathbf{x})} = \prod_{k=1}^{C} \sqrt{\frac{2\pi}{\theta_k^2 - \chi_k^2}}\theta_k^2 \exp\left\{\frac{(y_k(\mathbf{x}) - \rho_k)^2}{2\left(\theta_k^2 - \chi_k^2\right)}\right\}$$

which leads to an implicit equation in χ_k. For networks with a single output, $C = 1$, we have

$$\chi_{(\tau+1)} = \sqrt{\theta^2 - 2\pi\theta^4 \frac{p(\mathbf{x})^2}{\left|\Psi_{(\tau)}\right|^4} \exp\left\{\frac{(y-\rho)^2}{\theta^2 - \chi_{(\tau)}^2}\right\}} \tag{V.3}$$

With Ψ, y and χ functions of $\mathbf{x}$. The equation (V.3) allows in principle an iterative procedure which, starting from the constant initial value $\chi_{(0)}$ which leads to a state function $\Psi_{(0)}$, through the resolution of the system (IV.5) permits to calculate successive corrections of Ψ.

VI. Results

The resolution of the system (IV.5) requires considerable computational powers. For this reason the minimum energy was calculated in an approximate way with a genetic algorithm (GA), on an Intel 6-Core i7-8750H MT MCP processor. The equations have been treated symbolically with the Computer Algebra System maxima.[8]

The test problem comes from the Statlib repository.[9] It is a synthetic dataset made up of 3848 records, generated by David Coleman, referred to for convenience as POLLEN and which represents geometric and physical characteristics of pollen grain samples. It consists of 5 variables: the first three are the lengths in the directions x (ridge), y (nub) and z (crack), the fourth is the weight and the fifth is the density, the latter being the target of the problem. In our model they represent, respectively, x_1, x_2, x_3, x_4 and t_1. The choice of this problem lies in the fact that the data were generated with Gaussian distributions with low correlations, and is therefore close to the initial assumptions of the model for $\mathbf{x}$ and $\mathbf{t}$. Tables I and II show the general statistics of the dataset.

The characteristics of the genetic algorithm have been described in a previous paper [1]. This is a steady-state GA, with a generation gap of one or two, depending on the operator applied. The population has binary coding and implements a fitness sharing mechanism [5] to allow speciation and avoid premature convergence, according to the equations

$$E_l^{'} = E_l \sum_m \varphi(d_{lm}) \tag{VI.1}$$

$$\varphi(d_{lm}) = \begin{cases} 1 - \left(\frac{d_{lm}}{R}\right)^v & \Rightarrow \quad d_{kl} < R \\ 0 & \Rightarrow \quad d_{kl} \geq R \end{cases}$$

being $\varphi(d_{lm})$ a function of the diversity between individuals l and m, d_{lm} the Hamming distance and R the niche radius within which individuals are considered similar. Niche sharing implements a correction to energy calculated based on the similarity between the individual l and the rest of the population. The more similar it is, the greater the value of $\varphi(d_{lm})$, penalizing the energy in the equation (VI.1) since we are minimizing.

The decoding of the genotype implements the code of Gray to avoid discontinuities in the binary representation. The transformation between the binary representations, b, and Gray, g, for the i-th bit, considering numbers composed of n bits numbered from right to left, with the most significant bit on the left, is given by

$$g_i = \begin{cases} b_i & \Rightarrow \quad i = n \\ b_{i+1} \otimes b_i & \Rightarrow \quad i < n \end{cases}$$
$$b_i = \begin{cases} g_i & \Rightarrow \quad i = n \\ b_{i+1} \otimes g_i & \Rightarrow \quad i < n \end{cases}$$

with $\otimes$ the XOR operator.

The GA uses four operators: crossover, mutation, uniform crossover and internal crossover, and performs a search in the space of the computed energies according to the equation (IV.2), but simultaneously realizes a search in the space of the operators through the use of two additional bits in the genotype of each individual of the population. This allows a dynamic choice of the probabilities of each operator at each moment of the calculation, according to the fraction of elements of the population that encode for each of the four possibilities. The initial population is randomly generated.

The procedure for assessing an individual consists of the following steps:

[8] http://maxima.sourceforge.net/

[9] http://lib.stat.cmu.edu/datasets/

Table I
POLLEN DATASET, GENERAL FEATURES. THE TABLE SHOWS THE MEANS (μ, ρ), STANDARD DEVIATIONS (σ, θ), SKEWNESS AND KURTOSIS (THE REFERENCE FOR NORMALITY IS 0) OF THE ORIGINAL DATA AND NORMALIZED DATA

	Original data		Normalized data		Skewness	Kurtosis
Var	μ / ρ	σ / θ	μ / ρ	σ / θ		
x_1	-3.637e-03	6.398	0.0418	0.2863	-0.130	-0.057
x_2	1.597e-04	5.186	-0.0257	0.3082	0.072	-0.311
x_3	3.103e-03	7.875	0.0178	0.2551	-0.057	-0.158
x_4	4.237e-03	10.004	-0.0252	0.2876	0.109	-0.163
t_1	1.662e-04	3.144	0.0512	0.2745	0.110	0.192

Table II
DATASET POLLEN, CORRELATION MATRIX

	x_1	x_2	x_3	x_4	t_1
x_1	1.00	0.13	-0.13	-0.90	-0.57
x_2	0.13	1.00	0.08	-0.17	0.33
x_3	-0.13	0.08	1.00	0.27	-0.15
x_4	-0.90	-0.17	0.27	1.00	0.24
t_1	-0.57	0.33	-0.15	0.24	1.00

1) the values of the n_Γ parameters are generated within certain prefixed ranges through the application of one of the operators;
2) the network output, y_k, is generated for each element of the dataset. This set of values allows us to calculate χ_k;
3) the $D \times D$ elements of the matrices $\mathbf{H}$ and $\mathbf{S}$ are computed by means of the integrals (IV.6) and (IV.7);
4) the determinant (IV.8) is calculated;
5) the system (IV.5) is solved.

The result is the energy value, E, and the D coefficients c of Ψ.

Before the execution of the tests, a preprocessing of the dataset was performed, normalizing $\mathbf{x}$ and t within the range [-1:1]. 15 calculations were conducted, each consisting of 10 concurrent processes sharing the best solution found. In each calculation the set of lower energy solutions found in the previous calculations were introduced. The values of the n_Γ parameters have been varied within certain pre-established ranges, identified through a preliminary test campaign. The reference ranges are shown in Table III. Table IV shows the reference values of the parameters used in the genetic algorithm.

For each element of the population, in addition to the energy value, has been calculated the square error percentage of the neural network [1, 12]

$$E_r = \frac{100}{s(t_{max} - t_{min})^2} \sum_s (y_s - t_s)^2$$

with s the number of dataset records and $(t_{max} - t_{min})^2 = 4$ the normalization interval used.

Some of the parameters in the Table IV deserve some observation:

Table III
VALUES OF THE MODEL AND RANGE OF VARIABILITY OF THE n_Γ VARIATIONAL PARAMETERS AND OF THE DATA $\mathbf{x}$ AND t_1 OF THE DATASET

Variable	Value
C	1
D	8
N	4
P	10
ζ	[1:$(2\pi)^N$]
$\mathbf{x}, t_1$	[-1:1]
λ, ξ	[0:4]
w	[0:4]
η, ω	[0:1]

Table IV
REFERENCE VALUES OF THE GENETIC ALGORITHM

Variable	Value
Population	100
Point mutation probability	0.005
R	[0:0.9]
υ	1
Chromosome length	2174 bits
Calculation cycles	[5000:6000]

- $\upsilon = 1$ implies the so-called *triangular niche sharing*;
- R has a considerable influence on the results and was chosen for each of the 10 concurrent processes of each calculation according to the criterion $R_i = (i-1)/10$, $i = 1, \ldots, 10$, with i the process number. This allows to avoid arbitrary choices since R can be dependent on the nature of the problem;
- ξ was chosen in the interval [0:4], which includes the value given by a heuristic RBF rule which proposes for the standard deviation of the associated Gaussian, $\sigma_\xi = \sqrt{\frac{1}{2\xi}}$, the reference value $2\bar{d}_\omega$, with $\bar{d}_\omega$ the average value between the centroids of the functions ϕ_p of the equation (III.8). Considering an estimate of $\bar{d}_\omega = 1$ we get $\xi = 0.125$. The range $\xi \in [0 : 4]$ is equivalent to $\sigma_\xi \in [0.354 : \infty]$. The same criterion has also been used for the vector $\vec{\lambda}$.

The data was divided into two parts, a set of training (2886 records) and a set of testing (962 records). Technically, this subdivision is not nec-

Table V
RESULTS OF THE GENETIC ALGORITHM FOR THE PARAMETERS OF THE BASIS FUNCTIONS OF THE NETWORK y_k

ξ	P \ N	ω_1	ω_2	ω_3	ω_4
3.4013120	ω_1	-0.764254	0.487696	0.828128	-0.309709
2.8162860	ω_2	0.722523	-0.829504	0.884602	-0.440280
3.9395730	ω_3	-0.977770	-0.860840	-0.385800	0.694424
1.1004576	ω_4	0.563102	0.138821	0.801561	0.928019
0.1681290	ω_5	0.829662	0.126966	-0.345363	-0.546116
0.8617840	ω_6	-0.621750	-0.468552	-0.684078	0.462411
0.2275410	ω_7	-0.969937	0.801109	-0.634125	-0.436867
3.2266080	ω_8	0.424373	0.804881	0.896162	-0.218709
3.4019790	ω_9	-0.756753	-0.440174	-0.688590	0.101702
3.6097500	ω_{10}	0.937970	0.675499	0.980483	-0.884515

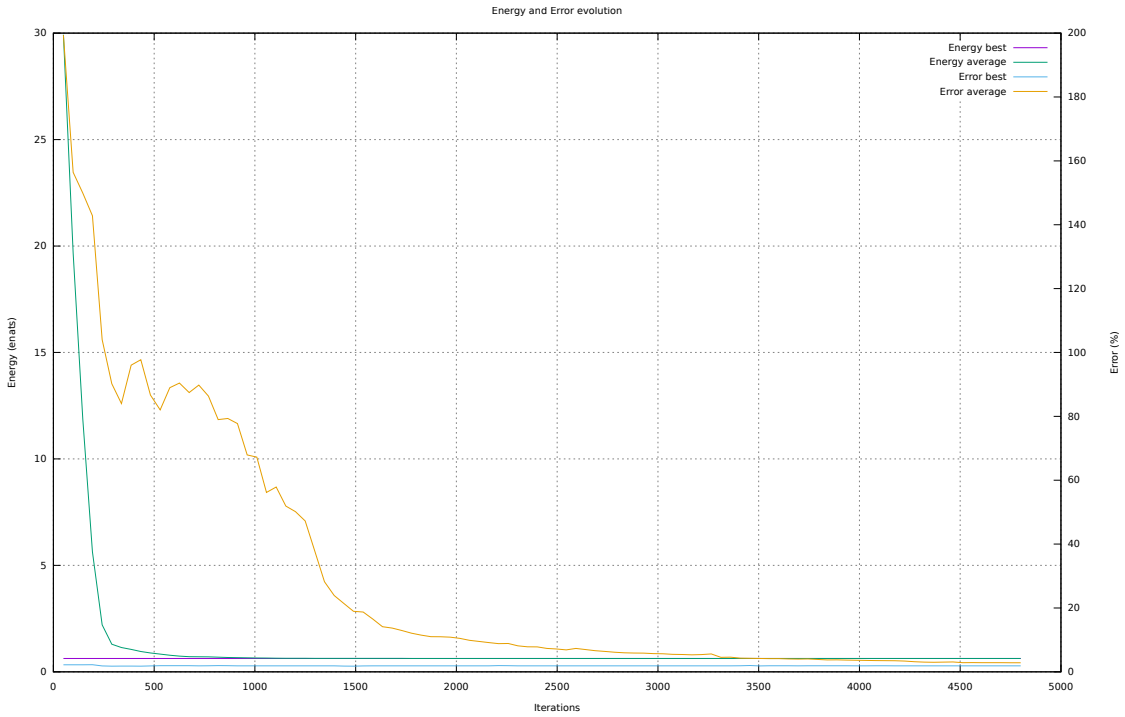

Figure VI.1. Evolution of the genetic algorithm that produced the solution with lower energy

Table VI
RESULTS OF THE GENETIC ALGORITHM FOR NETWORK WEIGHTS

P \ C	w_1
w_0	-0.262204
w_1	-0.654650
w_2	0.270344
w_3	-0.479204
w_4	0.291074
w_5	-1.255660
w_6	-0.103152
w_7	2.376997
w_8	-0.905523
w_9	0.479075
w_{10}	1.621334

essary, since the two data partitions are generated by the same distribution and the characterization of the problem in the model is given exclusively by the value of the constants $\vec{\rho}$, $\vec{\theta}$, $\vec{\mu}$ and $\vec{\sigma}$.

The variational parameters of the best solution are reported in Tables V, VI and VII. The final results of the calculation, including error and energy, are shown in Table VIII. Figure VI.1 shows the evolution of error and energy (in the lower and average versions of the population) of the calculation that generated the lower energy solution, which shows how the minimization of energy leads to a decrease of the error committed by the net in the target prediction. The final error value for training (0.945%) and testing (0.951%) partitions is particularly significant given the low number of basis functions used in the definition of Ψ and y_k.

VII. CONCLUSIONS

In this work we have developed a model for artificial neural networks based on an analogy with

Table VII
RESULTS OF THE GENETIC ALGORITHM FOR THE COEFFICIENTS AND PARAMETERS OF THE BASIS FUNCTIONS OF THE STATE FUNCTION Ψ

c	λ	P \ N	η_1	η_2	η_3	η_4
-1.282546	0.110080	η_1	-0.391103	0.5176760	-0.572174	-0.302664
-1.438387	0.120303	η_2	-0.668461	0.857889	-0.470183	-0.626793
-0.419667	0.186642	η_3	-0.679090	0.564954	-0.457736	-0.676878
1.357698	0.100000	η_4	-0.552879	0.782675	-0.552136	-0.562775
0.233801	0.100029	η_5	-0.134230	0.284860	-0.598276	0.211275
-0.559200	0.194570	η_6	-0.661250	0.672405	-0.335728	-0.348150
0.027306	0.100185	η_7	0.173759	-0.846671	-0.848185	-0.514533
2.092270	0.159235	η_8	-0.635414	0.671673	-0.428168	-0.496624

Table VIII
RESULTS OF THE GENETIC ALGORITHM FOR THE NETWORK y_k WITH LOWER ENERGY

Variable	Value
α_1 (train)	26.064396
β_1(train)	2.748593
γ_1(train)	-0.448009
χ_1(train)	0.194494
α_1 (test)	25.537031
β_1 (test)	2.375004
γ_1 (test)	-0.419022
χ_1 (test)	0.195101
ζ	1
E_r (train)	0.945%
E_r (test)	0.951%
E (train)	0.627771 enats
E (test)	0.621814 enats

a physical-quantum mechanical system. One of the advantages of this approach is the possibility, potentially, of using wave mechanics techniques in their study. An example is the generalized Hellmann-Feynman theorem

$$\frac{\partial E_d}{\partial \Gamma} = \int \Psi^* \frac{\partial \hat{H}}{\partial \Gamma} \Psi \, d\mathbf{x}$$

whose validity needs to be demonstrated in this context, but whose use seems justified since it can be demonstrated by assuming exclusively the normality of Ψ and the hermiticiy of $\hat{H}$. Its applicability could help in the calculation of the system (IV.10).

It is necessary to carry out a systematic test campaign to verify the results obtained. These tests are currently underway and will be the subject of a subsequent work. The preliminary results obtained on a set of selected problems coming from the Statlib and UCI[10] repositories confirm the validity of the model.

REFERENCES

[1] Francisco Yepes Barrera. Búsqueda de la estructura óptima de redes neurales con algoritmos genéticos y simulated annealing. verificación con el benchmark proben1. *Inteligencia Artificial, Revista Iberoamericana de IA*, 11(34):41–61, 2007.

[2] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[3] Alex Finnegan and Jun S. Song. Maximum entropy methods for extracting the learned features of deep neural networks. *PLOS Computational Biology*, 13(10):e1005836, October 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005836.

[4] Jeffrey D. Fitzgerald, Lawrence C. Sincich, and Tatyana O. Sharpee. Minimal Models of Multidimensional Computations. *PLOS Computational Biology*, 7(3): e1001111, March 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1001111.

[5] Lan Gao and Youwei Hu. Multi-target matching based on niching genetic algorithm. *JCSNS International Journal of Computer Science and Network Security*, 6(7A), July 2006.

[6] Amir Globerson and Naftali Tishby. The Minimum Information Principle for Discriminative Learning. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 193–200, Arlington, Virginia, United States, 2004. AUAI Press. ISBN 978-0-9749039-0-3. event-place: Banff, Canada.

[7] Amir Globerson, Eran Stark, Eilon Vaadia, and Naftali Tishby. The minimum information principle and its application to neural code analysis. *Proceedings of the National Academy of Sciences of the United States of America, PNAS*, 106(9), march 2009.

[8] Ira N. Levine. *Quantum chemistry*. Pearson, Boston, seventh edition edition, 2014. ISBN 978-0-321-80345-0.

[9] Javier R. Movellan and James L. McClelland. Learning Continuous Probability Distributions with Symmetric Diffusion Networks. *Cognitive Science*, 17(4):463–496, October 1993. ISSN 03640213. doi: 10.1207/s15516709cog1704-1.

[10] https://archive.ics.uci.edu/ml/index.php

[10] Joseph C. Park and Salahalddin T. Abusalah. Maximum Entropy: A Special Case of Minimum Cross-entropy Applied to Nonlinear Estimation by an Artificial Neural Network. *Complex Systems*, 11, 1997.

[11] Carlos A. L. Pires and Rui A. P. Perdigao. Minimum Mutual Information and Non-Gaussianity Through the Maximum Entropy Method: Theory and Properties. *Entropy*, 14(6):1103–1126, June 2012. ISSN 1099-4300. doi: 10.3390/e14061103.

[12] Lutz Prechelt. Proben1 - a set of neural network benchmark problems and benchmarking rules. Technical Report 21/94, Fakültat für Informatik, Universität Karlsruhe, 76128 Karlsruhe, Germany, September 1994.

[13] Zhang Xiaodong. Evaluation model and simulation of basketball teaching quality based on maximum entropy neural network. page 5, 2014.

[14] Dongxin Xu. *Energy, entropy and information potential for neural computation*. PhD thesis, University of Florida, 1999.

[15] Yan Zhang, Mete Ozay, Zhun Sun, and Takayuki Okatani. Information Potential Auto-Encoders. *arXiv:1706.04635 [cs, math, stat]*, June 2017. arXiv: 1706.04635.