entropy

MDPI

Article

Learning by Population Genetics and Matrix Riccati Equation

Sergei Kozyrev

Steklov Mathematical Institute of Russian Academy of Sciences, Gubkina St. 8, 119991 Moscow, Russia; kozyrev@mi-ras.ru

Abstract: A model of learning as a generalization of the Eigen's quasispecies model in population genetics is introduced. Eigen's model is considered as a matrix Riccati equation. The error catastrophe in the Eigen's model (when the purifying selection becomes ineffective) is discussed as the divergence of the Perron–Frobenius eigenvalue of the Riccati model in the limit of large matrices. A known estimate for the Perron–Frobenius eigenvalue provides an explanation for observed patterns of genomic evolution. We propose to consider the error catastrophe in Eigen's model as an analog of overfitting in learning theory; this gives a criterion for the presence of overfitting in learning.

Keywords: learning theory; statistical mechanics; evolution theory

PACS: 02.30.Hq

MSC: 68T05; 92D15

1. Introduction

In the present paper we discuss the relation of three different areas, such as statistical physics, learning theory, and theory of biological evolution. This relationship has been already widely discussed in the literature. In particular, the relationship between the theory of evolution and learning theory was mentioned by A.Turing [1] (learning is the minimization of risk and biological evolution is optimization of fitness). The relationship between statistical physics and evolutionary theory was discussed in [2,3]. The consideration of biological evolution as a learning problem for functional programming was discussed in [4–8] and various aspects of relation of statistical physics, learning and evolution was considered in [9–11]. In this paper, we consider the application of population genetics in learning theory.

Universal patterns of genome evolution found in genomics [2,3] were discussed by E. Koonin as a manifestation of the Gibbs distribution of a model of "interacting gas of genes". Here, we discuss a model of population genetics given by a matrix Riccati equation, which is a generalization of the Eigen's quasispecies model [12]. Patterns of genome evolution [2] in our model correspond to known estimates of the Perron–Frobenius eigenvalue. Eigen's "error catastrophe" for this model takes the form of divergence of the Perron–Frobenius eigenvalue of the matrix Ricatti equation in the limit of large matrices. Error catastrophe describes the regime of ineffective purifying selection in population genetics in the case of high mutation rates. From the point of view of learning theory, the "error catastrophe" describes the transition to overfitting in the corresponding learning model.

Biological evolution was compared with the statistical physics of disordered systems (or spin glass theory, in particular, frustration in biology and spin glasses were mentioned) in [9]; these authors also discuss the relation of evolution and learning [10,11]. Let us note that in [10,11], solvability of learning problems in evolution was taken for granted and here we address exactly this problem of solvability (in the form of problem of overfitting in learning).

Citation: Kozyrev, S. Learning by Population Genetics and Matrix Riccati Equation. *Entropy* **2023**, *25*, 348. https://doi.org/10.3390/e25020348

Academic Editors: Wilson A. Zuniga-Galindo and Adam Lipowski

Received: 21 December 2022
Revised: 26 January 2023
Accepted: 12 February 2023
Published: 14 February 2023

The exposition of this text is as follows. In Section 2 of this text we discuss the Eigen's quasispecies model and error catastrophe; introduce our generalization of this model (a kind of matrix Riccati model); describe the error catastrophe as divergence of the Perron–Frobenius eigenvalue of the matrix Riccati model in the limit of large matrices; and describe from this point of view known patterns of genomic evolution. In Section 3, we introduce a population genetics-type learning model and discuss the relation of the error catastrophe to overfitting in learning. In Appendix A (the Appendices), the Perron–Frobenius theorem, matrix Riccati equations, and basic definitions of statistical learning theory are discussed.

2. Generalization of Eigen's Model in Population Genetics

The Eigen's model in population genetics. Here we consider the Eigen's quasispecies model following [12] (for a discussion of relation to other models of population genetics see [13], in particular, Moran's model was introduced in [14]). We investigate a family of different "genotypes" with populations $x_i \geq 0$, $i = 1, \ldots, n$; the total population is normed: $\sum_{i=1}^{n} x_i = 1$.

The following system of equations [12] describes the dynamics, where Q is a matrix with positive matrix elements. The non-linear term describes the competition of genotypes

$$\frac{d}{dt}x_i(t) = \sum_{j=1}^{n} Q_{ij}x_j(t) - E(t)x_i(t), \quad E(t) = \sum_{i,j=1}^{n} Q_{ij}x_j(t). \tag{1}$$

Diagonal matrix elements of Q describe reproduction rates of genotypes, off-diagonal matrix elements describe mutation rates.

The analog of the Eigen's model with discrete time is as follows

$$x_i(t+1) = \frac{1}{E(t)} \sum_{j=1}^{n} Q_{ij}x_j(t), \quad E(t) = \sum_{i,j=1}^{n} Q_{ij}x_j(t).$$

In Eigen's paper [12], genotypes are enumerated by strings of characters. The length of the genome is denoted by ν, the size of the alphabet is denoted by k (in particular for nucleotides $k = 4$). The fidelity of reproduction of a single nucleotide is q, $0 < q < 1$. In this case, the accuracy of reproduction of a genome is $R_{ii} = q^{\nu}$. Mutation rates in Eigen's model are equal to

$$Q_{ij} = \epsilon^{d(i,j)} R_{ii}, \quad \epsilon = \frac{q^{-1} - 1}{k - 1}. \tag{2}$$

Here, $d(i, j)$ is the number of different nucleotides in the i-th and j-th genotypes (called the Hamming distance). The reproduction and mortality rates of the i-th genotype are denoted P_i and D_i correspondingly, which gives for diagonal matrix elements $Q_{ii} = P_i R_{ii} - D_i > 0$.

The Eigen's model is a matrix Riccati Equation (A2) (see Appendix A.2), where $A_4 = Q$, $A_1 = 0$, $A_3 = 0$, $A_2 = \mathbf{e}^{\dagger} Q$, and $\mathbf{e}$ is a vector with unit coordinates. By the Perron–Frobenius theorem, see Appendix A.1, the dynamics of this model reduces to convergence to a stationary solution (which for Eigen's model is called quasispecies) given by the Perron–Frobenius eigenvector corresponding to the Perron–Frobenius eigenvalue of matrix Q (the largest eigenvalue of a matrix with positive matrix elements).

Error threshold. Eigen considered the behavior of the stationary solution of (1) (the quasispecies) depending on the mutation rate in the frameworks of perturbation theory by small mutation rates (i.e., by small off-diagonal matrix elements of Q). Let us denote I the most fit genotype for the PF eigenvector of Q (i.e., the sequence with the maximal population). Then, if the matrix Q is diagonal (there are no mutations), one has $x_I = 1$ and $x_i = 0$, $i \neq I$. If mutations are small but non-zero (first-order perturbation of the stationary solution), one has $x_i = x_i^{(0)} + x_i^{(1)}$, $x_I^{(0)} = 1$, for $i \neq I : x_i^{(0)} = 0$, first-order corrections are given by

$$(Q_{ii} - Q_{II})x_i^{(1)} + Q_{iI} = 0, \quad x_i^{(1)} = \frac{Q_{iI}}{Q_{II} - Q_{ii}}. \tag{3}$$

The correction $x_I^{(1)}$ satisfies

$$x_I^{(1)} + \sum_{i \neq I} x_i^{(1)} = 0,$$

therefore

$$x_I^{(1)} = -\sum_{i \neq I} \frac{Q_{iI}}{Q_{II} - Q_{ii}}. \tag{4}$$

Expressions (3) and (4) give coordinates of the PF vector for the stationary solution of (1) in the first order of perturbation theory by small mutation rates.

The correction $x_I^{(1)}$ is small if the series for the rates of mutations is small

$$\sum_{i \neq I} Q_{iI}, \tag{5}$$

and if there are no small denominators in the above Formula (4), i.e., $Q_{II} - Q_{ii} > \delta > 0$ for some δ, this condition is sufficient.

For Eigen's choice of mutation rates (2), this series can be estimated by a geometric progression. Therefore, if the reproduction accuracy q is not close to one, this series will be large (actually we discuss finite but long progressions). The regime when the stationary state (the quasispecies) for Eigen's model loses localization is called the error catastrophe. The corresponding error catastrophe mutation rate separates regimes of effective and ineffective purifying selection in population genetics.

It is easy to see that $\sum_i Q_{iI}$ is the estimate (A1) for the Perron–Frobenius eigenvalue for matrix Q (and $x_i = Q_{iI}$, $i \neq I$ are estimates of coordinates of the PF vector in the first order of perturbation theory, if we ignore denominators in (3)). Therefore, *Eigen's model is a variant of the matrix Riccati equation and error catastrophe is the divergence of the Perron–Frobenius eigenvalue of the model in the limit of large matrices*.

Generalization of Eigen's model. Let us introduce a generalization of Eigen's model. We consider a space of possible genotypes and a set of possible mutations $E = [e_1, \ldots, e_n]$. Here, e_s are not necessarily point mutations, mutations may include duplications, insertions, deletions, etc. Let us put in correspondence to a mutation e_s a weight $w(e_s) > 0$ as "evolutionary effort" to produce the mutation. The Boltzmann factor $e^{-\alpha w(e_s)}$ ($\alpha > 0$ is a parameter of the kind of inverse temperature for mutations) is the analog of the mutation rate for a single mutation $1 - q$ in Eigen's model. We define the transition rate from genotype i to genotype j in the model of population genetics under consideration as

$$Q_{ji} = \sum_{p: i \to j} e^{-\alpha \sum_{k \in p} w(e_s(k))}, \tag{6}$$

where summation over p runs over paths $p : i \to j$ of generation of j from i and summation over k runs over mutations along the path p (i.e., k-th mutation at the path p is e_s). This sum over k weights of mutations is the analog of the Hamming distance in (2), the summation over paths takes into consideration retinal evolution (possibility to access j from i taking mutations in different order).

We define diagonal matrix elements Q_{ii} by the functional R, which describes fitness ($\beta > 0$ is the inverse temperature for selection, temperatures for selection and mutations can be different)

$$Q_{ii} = e^{-\beta R[i]}. \tag{7}$$

Then, we define the model of population genetics by using equations of the Eigen's model (1) with more general mutation and survival matrix (6), (7) and more general family of mutations, this allows us to explain patterns of genomic evolution (9) and (10); see the discussion below.

The condition for effective purifying selection for this model is the condition of convergence of the estimate (A1) of the Perron–Frobenius eigenvalue of the model, if we exclude in this estimate the diagonal matrix element and reduce the corresponding series to (5), we get

$$Z = \sum_j \sum_{p:i\to j} e^{-\alpha \sum_{k\in p} w(e_s(k))}. \tag{8}$$

This expression has the form of a statistical sum over iterated mutations. Here, i is the starting point of evolution (the ancestral genome). Critical phenomena for this statistical sum (transition between convergence and divergence of (8) depending on the inverse temperature α) describe the transition between regimes of effective and ineffective purifying selection in population genetics (the error catastrophe). Let us note that all possible mutations give contributions to this expression. Even if only point mutations are taken into account, this gives a contribution of order of the length of a genome. Therefore, to keep (8) small, the mutations rates should be sufficiently low. This observation also puts limitations on learning without overfitting, see the discussion in Section 3.

Laws of genomic evolution, population genetics, and statistical physics. Let us consider two examples of genomic evolution discussed in [2,3] and show that patterns of genomic evolution can be considered as a manifestation of the statistical sum (8).

Orthologous proteins in different species are related by common origin. For such proteins the logarithm of amino acid substitution frequency is distributed according to normal law. Let us consider for orthologous proteins the evolution by random independent amino acid substitutions with probability of substitution $A \to B$ depending only on amino acids A, B. The coordinates of the PF vector, by perturbation theory (3), can be estimated by mutation rates (6), which gives for the coordinates

$$e^{-\alpha \sum_k E_k}, \tag{9}$$

where E_k are weights of mutations in the process of protein generation from the ancestor (summation with respect to k is the summation along the path of evolution). For independent mutations (this is the assumption that the evolution is neutral) we obtain the lognormal distribution for protein occurrences in the orthologous family (coordinates of the PF vector).

Genes in the same genome generated by duplication events are called paralogous. Let us consider evolution by gene duplication, each duplication corresponds to a contribution in (6) (evolutionary effort) E. Then, for a family of N paralogous genes, the "evolutionary effort" contribution is NE; thus, the expression for a coordinate of the PF vector corresponding to a family of N paralogous genes will be equal to

$$e^{-\alpha NE}, \tag{10}$$

i.e., one obtains the degree distribution for sizes N of families of paralogs.

Therefore, the statistical sum (8) explains known patterns of genomic evolution. For discussion of these patterns, E.V. Koonin conjectured [2,3] an idea of "interacting gas of genes", i.e., the evolution of genomes should be explained by the Gibbs distribution of some model of statistical physics with interaction of genes. This hypothetical model was also called "the third evolutionary synthesis". In [9–11], following these ideas, the relation between statistical physics, learning, and evolution was discussed. In these papers, the authors applied the approach of [15,16], where evolution phenomena were explained using the structure of the fitness landscape (i.e., the diagonal part of the selection–mutation matrix in the above model).

In our approach, the mutation off-diagonal part of this matrix was applied, universal genomic evolution patterns follow from the universality of the mutation matrix (6). From the point of view of learning theory, see the next Section, the universal form of mutation rates looks like universal regularization in learning. The above generalization of Eigen's

model (6) and (7) can be considered as a possible candidate for the "interacting gas of genes" model (the corresponding Gibbs distribution is (8), or the PF eigenvalue estimate).

3. Learning Theory and Population Genetics

Learning is the minimization of the risk functional (or loss functional) R over the hypothesis space $\mathcal{F}$ of the system, see, in particular, [17] and Appendix A.3. Analogy between learning and Darwinian evolution, i.e., between minimization of risk and optimization of fitness, was mentioned by A. Turing [1] in 1950, now this idea attracts attention [4,5,8,10,11]. From this point of view, it is natural to apply in evolution theory different ideas of learning theory and vice versa. In particular, regularization, an important idea in learning, looks promising for evolution theory, as was shown at the end of the previous Section, universal regularization (by mutation rates) in learning problems of evolution gives universal distributions in genomics. The analogy between evolution and learning goes even further—in a discussion by R. Fisher [18], selection was considered as a random phenomenon (random weather conditions, etc.). These arguments look similar to the statistical learning theory (where training data are randomly chosen, see Appendix A.3). In a more standard discussion of evolution theory, training data are considered as fixed (selection is fixed).

The analogy between optimization and selection can be considered as an application in the learning of Darwinism, or the first evolutionary synthesis. Modern evolution theory is the population genetics, or the second evolutionary synthesis. In population genetics, an ensemble (population) instead of a single object is considered. One of the central achievements of population genetics is the explanation of purifying selection. Purifying selection was discussed by R. Fisher [18]; it is related to competition of different genotypes and prevents degradation of the fitness by mutations. Learning in population genetics can be defined as the convergence of the population of hypotheses to a peak around the minimum of the risk functional. The transition from learning a single hypothesis to learning a population of hypotheses is analogous to the transition from mechanics to statistical mechanics (where ensembles are studied). Using this analogy, we propose to discuss error catastrophe, or transition to ineffective purifying selection, as a model of overfitting in learning.

We formulate the learning model as the analog of the considered above generalization of Eigen's model. Let $x_f(t) \geq 0$, $\sum_f x_f(t) = 1$ be a normed distribution on the hypothesis space $\mathcal{F}$ of the learning system (the "space of genotypes" or hypotheses). Let us consider the analog of mutations in genetics—a list of partially defined maps $E = [e_1, \dots, e_n]$, $e_s : \mathcal{F} \to \mathcal{F}$ of the hypothesis space. Hypotheses are generated from the initial hypothesis (in biology, ancestral genome) by an iterated application of hypothesis transformation operations (in biology, mutations).

The model of learning by population genetics is given by the following matrix Riccati equation (an analog of (1))

$$\frac{d}{dt}x_f(t) = \sum_g Q_{fg}x_g(t) - x_f(t)\sum_{f,g} Q_{fg}x_g(t) \tag{11}$$

where mutation rates Q_{fg}, $f \neq g$ have the form (6) used in the discussed above generalization of Eigen's model with the defined above mutations (hypotheses transformations) $E = [e_1, \dots, e_n]$ and corresponding weights $w(e_s) > 0$ (efforts to perform mutations), diagonal matrix elements are given by (7), where R is the risk functional of the learning problem under consideration.

The discrete time analog of (11) is

$$x_f(t+1) = \frac{\sum_g Q_{fg}x_g(t)}{\sum_{f,g} Q_{fg}x_g(t)}.$$

One of the central problems in learning theory is overfitting, which is a strong dependence of the results of learning on the training sample—if the learning system is too

complex it can overreact to small details of the data, hence a large subset of the hypothesis space contribute to learning. Therefore, overfitting is related to high entropy of the hypothesis space, to control overfitting a regularization is applied, see, in particular, VC theory [17].

In the above model of learning by population genetics, overfitting can be considered an error catastrophe, or transition to the regime of ineffective purifying selection, or divergence of the statistical sum (8), which gives the estimate for the Perron–Frobenius eigenvalue of the model as a matrix Riccati equation (actually, the estimate of the PF eigenvalue minus the diagonal matrix element in (A1)). Divergence of (8) means that the large subset of the space of genomes contribute to the population, and selection can not isolate the most fit genotype. This implies the divergence of (8) due to the large entropy of the hypothesis space $\mathcal{F}$ (the Boltzmann factor in (6) decays slowly with additional mutations). Convergence of (8) is provided if this decay is fast; this can be considered a regularization in the learning problem.

The condition of convergence of the statistical sum (8) can be satisfied for a wide choice of learning models and sufficiently low temperatures (large α). This condition is much less restrictive than the condition of the finite VC dimension in VC theory. This gives a criterion of the presence of overfitting in a population genetics-type learning problem. This criterion is a thermodynamic type effect and can be understood only if an ensemble (population) of learning systems is considered. The author does not claim that this statement about control of overfitting in learning by population genetics is mathematically proven, the idea is to exploit physical (and biological) intuition in the learning theory.

In the above discussion, we considered the fixed risk functional R (i.e., fixed training sample). In principle, one can vary the sample (use test data instead of training data); this will modify the risk functional R (selection) and diagonal matrix elements (7), but will not change the off-diagonal matrix elements (6) (mutations) and the statistical sum (8); hence, predictions on overfitting will be the same.

Relation to "complexity as energy". The theory of "complexity as energy" was discussed by Yu.I. Manin [19]. In this approach, the Gibbs distribution with the Hamiltonian equal to the Kolmogorov complexity was applied to explain the power Zipf's law of word frequency distribution in texts. The sum of weights of hypothesis generation operations in (6) can be discussed as a weighted upper bound for the Kolmogorov complexity of generation of a hypothesis. Therefore, statistical sum (8) is an example (of approximation) of the "complexity as energy" approach.

Relation to GAN. Generative Adversarial Network (or GAN) is a learning model which works by contest of two neural networks, generator and discriminator [20]. Modification and competition of networks at each step of the contest can be considered as analogs of mutation and selection correspondingly. From this point of view, this looks like a kind of the predator–prey model with mutations; moreover, GANs are described by minimax models similar to evolutionary game theory [21]. It is a general opinion of biologists that the predator–prey competition accelerates evolution greatly and this can be considered as an explanation why GANs are very successful. It looks like different models of population genetics might contribute to learning theory, in particular, the generalization of Eigen's model introduced in the present paper. One can also mention genetic algorithms as an example of applications of biological ideas in learning.

Summary. In the present paper, we introduced a generalization of the Eigen's model in population genetics and described the error catastrophe (transition to ineffective purifying selection) as a divergence of the Perron–Frobenius eigenvalue of the mutation–selection matrix of the model. The introduced model explains known patterns of genomic evolution. We propose to consider this population genetics model as a model of learning, where: the learning model is an ensemble (population) of learning models (a distribution on the hypothesis space); the risk functional in learning is described by fitness in population genetics; mutation rate matrix in population genetics corresponds to a set of hypothesis transformation operations and corresponding matrix of transformation (mutation) rates;

learning reduces to convergence of the population of hypotheses to a peak around the minimum of the risk functional. Then, overfitting in learning can be described as the error catastrophe in population genetics, this criterion of overfitting can be understood only using ensembles (populations) of hypotheses.

Funding: This work was performed at the Steklov International Mathematical Center and supported by the Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-15-2022-265).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Appendix A.1. Perron–Frobenius Theorem

Let $A = (a_{ij})$ be a square matrix with positive matrix elements, then:

(1) The largest in modulus eigenvalue r (Perron–Frobenius eigenvalue) is real and positive;
(2) This eigenvalue is simple (non-degenerate);
(3) There exists an eigenvector (Perron–Frobenius vector), corresponding to r with strictly positive coordinates, all other eigenvectors do not have this property;
(4) $\lim_{k\to\infty} \frac{A^k}{r^k} = P$, where P is a projection to the Perron–Frobenius vector;
(5) Eigenvalue r satisfies inequalities

$$\min_i \sum_j a_{ij} \le r \le \max_i \sum_j a_{ij}. \tag{A1}$$

For a matrix with non-negative matrix elements, the analogous properties are satisfied (these properties can be obtained as limits of the above properties, in particular the highest eigenvalue can be degenerate and some coordinates of the corresponding eigenvector can be zeros).

Appendix A.2. Matrix Riccati Equation

In [22], an approach to analysis of texts based on matrix Riccati equations is discussed. Namely, the matrix is considered

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix},$$

where A_4 is a $(N-1)\times(N-1)$–matrix, A_1 is a number, A_2 and A_3 correspondingly are row and column of length $N-1$.

The corresponding map of the projective space P^{N-1} is investigated

$$A : y_m = \begin{pmatrix} 1 \\ x_m \end{pmatrix} \mapsto y_{m+1} = \begin{pmatrix} 1 \\ x_{m+1} \end{pmatrix}, \quad x_{m+1} = \frac{A_3 + A_4 x_m}{A_1 + A_2 x_m}.$$

This discrete time dynamical system (iteration of the map above) can be considered as a discretization of the matrix Riccati equation

$$\frac{d}{dt}x(t) = A_3 + A_4 x(t) - x(t)A_1 - x(t)A_2 x(t). \tag{A2}$$

The corresponding flow converges to a stationary point defined by the Perron–Frobenius theorem (under corresponding constraints for matrix A).

Appendix A.3. Basic Definitions of the Statistical Learning Theory [17]

Learning theory discusses extracting patterns from data. In particular, the definition of a supervised learning problem is as follows: let us consider a training sample (i.e., a set of labeled data) $z_i = (x_i, y_i)$, $x_i \in X$, $y_i \in Y$. We have to find a function (hypothesis) $f : X \to Y$ in the hypothesis space $\mathcal{F}$ related to the training sample.

Let $p(x, y)$ be a probability distribution in $X \times Y$. To evaluate a hypothesis, we will consider the loss (or risk) function $V(f(x), y)$ with non-negative values. The expected risk functional is defined as an average of the risk function

$$R[f] = \int_{X \times Y} V(f(x), y) p(x, y) dx dy.$$

The problem of statistical learning is to find a hypothesis that gives the minimum of the risk functional

$$f = \arg\min_{h \in \mathcal{F}} R[h].$$

The empirical risk functional

$$R_{\text{emp}}[f, \text{data}] = \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i),$$

where $\text{data} = \{z_i\} = \{(x_i, y_i)\}$, $i = 1, \ldots, n$ is a training sample generated by the probability distribution, $p(x, y)$, should approximate the expected risk functional.

The supervised learning problem is to find the optimal hypothesis defined by the training sample

$$f[\text{data}] = \arg\min_{h \in \mathcal{F}} R_{\text{emp}}[h, \text{data}].$$

Classification problem. Let $y_i = 0, 1$, a hypothesis f belongs to a family of characteristic functions (i.e., $f(x) = 0$ or $f(x) = 1$), and the risk function $V(f(x), y) = |f(x) - y|$ is as follows: it is equal to zero for $f(x) = y$ and to one otherwise. In this case, the empirical risk functional is given by the average number of errors of the risk function at the training sample:

$$R_{\text{emp}}[f, \text{data}] = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - y_i|.$$

Problem of overfitting. The following situation is possible: it is possible that the optimal hypothesis $f[\text{data}]$ computed for the training sample data will give high values of the empirical risk functional if we would compute this functional for a control sample data$'$ different from data. This phenomenon is called overfitting. According to the Vapnik–Chervonenkis theory (or VC theory) [17], overfitting is explained by high entropy of the hypothesis space.

A general approach to control overfitting is the application of regularization: one can add a non-negative regularizing term depending on the hypothesis to the functional of empirical risk

$$H[f, \text{data}] = R_{\text{emp}}[f, \text{data}] + Reg[f], \tag{A3}$$

then, the learning problem will have the form

$$f_{Reg}[\text{data}] = \arg\min_{h \in \mathcal{F}} H[h, \text{data}]. \tag{A4}$$

If a regularization term corresponds to some potential well in the hypothesis space, the optimization problem will be restricted to this well. In this way, the entropy of the hypothesis space will be limited.

References

1. Turing, A.M. Can machines think? Computing Machinery and Intelligence. *Mind* **1950**, *49*, 433–460. [CrossRef]
2. Koonin, E.V. *The Logic of Chance: The Nature and Origin of Biological Evolution*; FT Press: Upper Saddle River, NJ, USA, 2012.
3. Koonin, E.V. Are There Laws of Genome Evolution? *PLoS Comput. Biol.* **2011**, *7*, e1002173. [CrossRef] [PubMed]
4. Kozyrev, S.V. Biology as a constructive physics. *p-Adic Numbers Ultrametr. Anal. Appl.* **2018**, *10*, 305–311. . S2070046618040076. [CrossRef]
5. Kozyrev, S.V. Learning problem for functional programming and model of biological evolution. *p-Adic Numbers Ultrametr. Anal. Appl.* **2020**, *12*, 112–122.
6. Kozyrev, S.V. Genome as a functional program. *Lobachevskii J. Math.* **2020**, *41*, 2326–2331. https://doi.org/10.1134/S1995080220120173.
7. Kozyrev, S.V. Is genome written in Haskell? *Lobachevskii J. Math.* **2021**, *42*, 2359–2364. [CrossRef]
8. Kozyrev, S.V. Learning theory and population genetics. *Lobachevskii J. Math.* **2022**, *43*, 1417–1424. [CrossRef]
9. Wolf, Y.I.; Katsnelson, M.I.; Koonin, E.V. Physical foundations of biological complexity. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E8678–E8687. [CrossRef]
10. Vanchurin, V.; Wolf, Y.I.; Koonin, E.V.; Katsnelson, M.I. Thermodynamics of Evolution and the Origin of Life. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2120042119. [CrossRef] [PubMed]
11. Vanchurin, V.; Wolf, Y.I.; Katsnelson, M.I.; Koonin, E.V. Towards a Theory of Evolution as Multilevel Learning. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2120037119. [CrossRef] [PubMed]
12. Eigen, M.; McCaskill, J.; Schuster, P. Molecular Quasi-Species. *J. Phys. Chem.* **1988**, *92*, 6881–6891. [CrossRef]
13. Wilke, C.O. Quasispecies theory in the context of population genetics. *Bmc Evol. Biol.* **2005**, *5*, 44. [CrossRef]
14. Moran, P.A.P. Random processes in genetics. *Math. Proc. Camb. Philos. Soc.* **1958**, *54*, 60–71. [CrossRef] [PubMed]
15. Sella, G.; Hirsh, A.E. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 9541–9546. [CrossRef]
16. Barton, N.H.; Coe, J.B. On the application of statistical physics to evolutionary biology. *J. Theor. Biol.* **2009**, *259*, 317–324. [CrossRef] [PubMed]
17. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995. [CrossRef] [PubMed]
18. Fisher, R.A. *The Genetical Theory of Natural Selection*; The Clarendon Press: Oxford, UK, 1930.
19. Manin, Y.I. Complexity vs energy: theory of computation and theoretical physics. *J. Phys. Conf. Ser.* **2014**, *532*, 012018. [CrossRef].
20. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680. [CrossRef]
21. Maynard Smith, J. *Evolution and the Theory of Games*; Cambridge University Press: Cambridge, UK, 1982.
22. Manin, Y.; Marcolli, M. Semantic Spaces. *Math. Comput. Sci.* **2016**, *10*, 459–477. [CrossRef].