Learning theory, population genetics and matrix Riccati equations

S.V. Kozyrev, Steklov Mathematical Institute

Analogies between ideas from three areas —
statistical physics,
machine learning theory,
biological Darwinian evolution.

Population genetics type model of learning — effectiveness of purifying selection is related to absence of overfitting in learning theory.

V.N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.
R.A. Fisher, The Genetical Theory of Natural Selection, The Clarendon Press, Oxford, UK, 1930.
E. V. Koonin, The Logic of Chance: The Nature and Origin of Biological Evolution, FT Press, 2012.

Perron–Frobenius theorem. Let $A = (a_{ij})$ be a square matrix with strictly positive matrix elements, then:
1) The maximal in modulus eigenvalue r (Perron–Frobenius eigenvalue) is real and strictly positive;
2) This is simple (non-degenerate) eigenvalue;
3) The corresponding to r eigenvector (Perron–Frobenius eigenvector) can be chosen to have strictly positive coordinates, all other eigenvectors can not be chosen in this way;
4) $\lim_{k\to\infty} \frac{A^k}{r^k} = P$, P is the projection to Perron–Frobenius eigenvector;
5) Perron–Frobenius eigenvalue r satisfies

$$\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}. \tag{1}$$

For matrices with non-negative matrix elements analogous properties can be obtained as limits of above properties (in particular Perron–Frobenius eigenvalue can be degenerate and some coordinates of the corresponding eigenvector can be zeros).

Semantic analysis of texts using **Matrix Riccati equation**

Yuri Manin, Matilde Marcolli, Semantic Spaces, Mathematics in Computer Science, **10**, *459–477 (2016), arXiv:1605.04238*

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

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

The map of a projective space P^{N-1}

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

defines a dynamic system with discrete time (by iteration of the map) which is a discretization of the matrix Riccati equation

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

The Riccati flow converges to a fixed point of the dynamical system by Perron–Frobenius theorem (if we choose A properly).

Eigen's quasispecies model in population genetics

M. Eigen, J. McCaskill, and P. Schuster, Molecular Quasi-Species, J. Phys. Chem. **92**, *6881–6891 (1988).*

A family of different "genotypes" with populations $x_i \geq 0$, $i = 1, \ldots, n$, is considered, the total population $\sum_{i=1}^{n} x_i = 1$. Genotypes correspond to strings of characters (nucleotides). ν — length of the line (the number of nucleotides in the genome), k — number of characters in the alphabet (for nucleotides $k = 4$). The accuracy of reproduction for a single nucleotide is q, $0 < q < 1$, then the accuracy of reproduction of a sequence of length ν will be equal $R_{ii} = q^{\nu}$. Mutation rates are

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

where $d(i, j)$ is the Hamming distance between the i-th and j-th genotypes (the number of different nucleotides in these genotypes). $Q_{ii} = P_i R_{ii} - D_i > 0$, P_i and D_i — reproduction and mortality rates.

The population dynamics is given by the matrix Riccati equation

$$\frac{d}{dt}x_i(t) = \sum_{j=1}^{n} Q_{ij}x_j(t) - E(t)x_i(t), \quad E(t) = \sum_{i,j=1}^{n} Q_{ij}x_j(t), \quad (3)$$

off-diagonal elements Q_{ij}, $i \neq j$ (mutation rates), diagonal matrix elements Q_{ii} describe reproduction rates.
Subtraction of $E(t)x_i(t)$ describes the mortality due to competition. The total population $\sum_{i=1}^{n} x_i = 1$ is conserved.
Analogue for discrete time:

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

The dynamics at large times by Perron–Frobenius theorem — convergence to highest eigenvector of Q, this is the **quasispecies** surviving in evolution.

Perturbation theory by small mutation rates Q_{ij}.
Let I be the sequence with the maximum population for Perron–Frobenius vector of Q. In the first order of perturbation theory $x_i = x_i^{(0)} + x_i^{(1)}$, $x_I^{(0)} = 1$, for $i \neq I : x_i^{(0)} = 0$,

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

Q_{iI} approximate coordinates of the PF vector in the first order of perturbation theory (if we ignore denominators).
A sufficient condition for smallness of the correction $x_I^{(1)}$ — smallness of series $\sum_{i \neq I} Q_{iI}$ for rates of mutations $I \to i$. Moreover

$$\sum_i Q_{iI} \tag{5}$$

is an estimate of the PF eigenvalue for Q by (1).

Error threshold separates the regimes of existence of well defined quasispecies (Perron–Frobenius vector of Q is a localized peak in the sequence space) and of *error catastrophe* where PF vector loses localization. By (5) error catastrophe is the transition to divergence for the PF eigenvalue.

If $\sum_i Q_{il}$ converges the model predicts an effective purifying selection — the quasispecies will be localized on a small group of closely related sequences. Stabilizing (purifying) selection works due to competition between genotypes in population genetics.

Eigen's model is a variant of a matrix Riccati equation, the main quasispecies is the Perron–Frobenius eigenvector, and the error catastrophe is the divergence of the Perron–Frobenius eigenvalue in the limit of large matrices.

Generalization of Eigen's model. Set of possible mutations $E = [e_1, \ldots, e_n]$; mutations e_s may include duplications, point mutations, insertions, deletions etc.
$w(e_s) > 0$ is a weight ("evolutionary effort") to produce e_s.
$\alpha > 0$ — inverse temperature for mutations.
Boltzmann factor $e^{-\alpha w(e_s)}$ is the analogue of the mutation rate for a single mutation $1 - q$ in Eigen's model. Mutation rate $i \to j$:

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

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 of weights of mutations is the analogue of the Hamming distance in (2), the summation over paths takes in consideration retinal evolution (possibility to access j from i taking mutations in different order).

Diagonal matrix elements Q_{ii} are defined by the functional R which describes fitness ($\beta > 0$ is the inverse temperature for selection)

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

The model of population genetics is defined by equations of the Eigen's model (3) with more general mutation and survival matrix (6), (7).
Condition for effective purifying selection is the convergence of the estimate of Perron–Frobenius eigenvalue (a statistical sum over mutations)

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

Here i is the starting point of evolution (the ancestral genome). Critical phenomena (transition between convergence and divergence of Z depending on the inverse temperature α) describe the transition between regimes of effective and ineffective purifying selection in population genetics.

Lognormal distribution in protein evolution. Orthologous proteins are proteins in various organisms related by the same origin. Distribution of the logarithm of frequencies of amino acid substitutions in orthologous proteins is close to the normal.

Evolution by a set of independent random amino acid substitutions in a protein. The coordinates of the PF vector, by perturbation theory (4), can be estimated by mutation rates (10), which gives

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

where E_k are weights of mutations in the process of generation of the protein from the ancestor. If mutations are independent (the evolution is neutral) this implies a log-normal distribution for frequencies of proteins in orthologous family (coordinates of the PF vector).

Power-law distribution of the sizes of families of paralogous genes. Genes in the genome generated by evolutionary duplication events are called paralogous.
Evolution by gene duplications with weights E (evolutionary effort). A family of N paralogous genes will correspond to N equal weights E. For neutral evolution model describing the process of gene duplication a coordinate of the PF vector corresponding to a family of N paralogous genes is

$$e^{-\alpha NE}.$$

This gives a power-law distribution depending on the size N of the family of paralogous genes.

By E.V.Koonin patterns of genomic evolution should be described by a Gibbs distribution for "interacting gas of genes", actually population genetics type model (3), (6), (7) works, the Gibbs distribution is given by mutation rates.

Statistical learning theory

Learning is extracting patterns from data. Supervised learning: 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. We assume the existence of an unknown joint probability distribution $p(x, y)$ in $X \times Y$. To evaluate the hypothesis, the loss (risk) function $V(f(x), y)$ taking non-negative values is used, the expected risk functional is

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

The problem of statistical learning: to find a hypothesis that minimizes the risk functional

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

Since $p(x, y)$ is unknown, the empirical risk functional is used

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

where $\text{data} = \{z_i\} = \{(x_i, y_i)\}$, $i = 1, \ldots, n$ is a training sample. The learning problem — find the optimal hypothesis depending on the training sample

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

Example: classification. Learning problem where $y_i = 0, 1$, f belongs to some family of characteristic functions (i.e. $f(x) = 0$ or $f(x) = 1$), the loss function $V(f(x), y) = |f(x) - y|$ equals zero if $f(x) = y$ and equals one otherwise, the empirical risk is equal to average number of errors for the training sample:

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

Problem of overfitting — it may happen that substitution of the objective function $f[\mathrm{data}]$ computed using the training sample data into the empirical risk functional for the control sample data' will give high value of empirical risk. Vapnik–Chervonenkis theory (or VC–theory) claims that overfitting is related to high entropy of the hypothesis space.
To control overfitting, regularization is used: adding a (non-negative) regularizing term to the empirical risk functional

$$H[f, \mathrm{data}] = R_{\mathrm{emp}}[f, \mathrm{data}] + Reg[f], \tag{8}$$

optimal hypothesis with regularization

$$f_{Reg}[\mathrm{data}] = \arg\min_{h \in \mathcal{F}} H[h, \mathrm{data}]. \tag{9}$$

The regularizing term limits the entropy of the hypothesis space (the part of this space where the regularization is small).

Darwinian evolution — analogue of learning (Turing, 1950).

Population genetics versus Darwinism — an ensemble (population) is considered.

Learning by population genetics — the convergence of population of hypotheses to a peak around minimum of the risk functional.

Population Genetics Type Learning Model —
analogue of considered above generalization of Eigen's model.

Regime of ineffective purifying selection (error catastrophe) — overfitting in learning.

Divergence of the Perron–Frobenius eigenvalue for Eigen's like model in the limit of large matrices.

Phase transition for statistical sum over mutations
(divergence of the statistical sum for high mutation rates).

The hypothesis space $\mathcal{F}$. Let $x_f(t) \geq 0$, $\sum_f x_f(t) = 1$ be a normalized distribution in the hypothesis space.
Family of hypothesis transformation operations (analogue of mutations in genetics) is the list of partially defined mappings $E = [e_1, \ldots, e_n]$, $e_s : \mathcal{F} \to \mathcal{F}$, weights $w(e_s) > 0$ (efforts to perform transformations).
Hypotheses are generated from the initial hypothesis (in biology, ancestral genome) by iterated application of hypothesis transformation operations (in biology, mutations).
Matrix of mutation and survival rates in population genetics model

$$Q_{gf} = \sum_{p:f\to g} e^{-\alpha \sum_{k\in p} w(e_s(k))}, \quad Q_{ff} = e^{-\beta R[f]}, \qquad (10)$$

R is the risk functional of the learning problem.

The model of learning by population genetics is given by the matrix Riccati equation (an analogue of Eigen's model (3))

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

The discrete time analogue is:

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

The condition of effective purifying selection is the convergence of the estimate for Perron–Frobenius eigenvalue

$$Z = \sum_g \sum_{p:f\to g} e^{-\alpha \sum_{k\in p} w(e_s(k))}. \qquad (12)$$

Here f is the initial point of learning (the ancestral genome).

Overfitting in learning — it is not possible to separate correct and incorrect hypotheses — analogue of ineffective purifying selection in population genetics.

Error threshold — transition between convergence and divergence of (12) depending on the inverse temperature α — critical phenomenon for this statistical sum.

Regularization — small temperature (large α) — smaller part of the hypothesis space contributes to the PF eigenvalue and the PF eigenvector.

Above condition of convergence of the PF eigenvalue (12) is much easier to satisfy compared to the restrictions of VC-theory (finite VC-dimension). This gives a criterion of 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.

Conclusion.

Population genetics type model in machine learning theory (a kind of matrix Riccati equation). Learning can be described by competition of hypotheses ("genotypes"), where hypotheses are transformed by "mutations".

Condition for effectiveness of purifying selection in population genetics corresponds to absence of overfitting in learning and is given by convergence of Perron–Frobenius eigenvalue of the mutation rate matrix in the limit of large matrices.

The PF eigenvalue has the form of a statistical sum over mutations, error threshold (which separates regimes of effective and ineffective purifying selection) looks like critical phenomenon for this statistical sum.

This statistical sum describes experimentally observed patterns of evolution of genomes. Universal regularization (by mutation rates) in learning problems of evolution gives universal distributions in genomics.