Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 53-2 (2020) 1385–1390

Early History of Machine Learning

Alexander L. Fradkov *

* *Institute for Problems in Mechanical Engineering, Russian Academy of Sciences, Saint-Petersburg (e-mail: Alexander.Fradkov@gmail.com)*

Abstract: Machine learning belongs to the crossroad of cybernetics (control science) and computer science. It is attracting recently an overwhelming interest, both of professionals and of the general public. In the talk a brief overview of the historical development of the machine learning field with a focus on the development of mathematical apparatus in its first decades is provided. A number of little-known facts published in hard to reach sources are presented.

Keywords: Machine Learning, Neural Networks, Separation Theorems, Convex Optimization

1. INTRODUCTION

Machine learning is an area related to both cybernetics and computer science (or Control Science and Computer Science) [1], attracting recently an overwhelming interest both of professionals and of the general public. In the last few years thanks to successes of computer science (the emergence of GPUs, leading to significant improvements in the performance of computers and development of special software, allowing to work with big data) machine learning is often attributed to computer science. However, historically, learning algorithms that provide convergence and sufficient convergence rate of the learning process arose within ybernetics/control. Below a look of a control theorist at the history of machine learning is presented. The author is a mathematician by education and has some experience in pattern recognition and control based on adaptation and learning in the 1960s.

2. EARLY YEARS

The origin of machine learning in its modern sense is usually associated with the name of the psychologist Frank Rosenblatt from Cornell University, who, based on ideas about the work of the human nervous system, created a group that built a machine for recognizing the letters of the alphabet Rosenblatt (1957, 1959, 1960). The machine, called the "perceptron" by its creator, used both analog and discrete signals and included a threshold element that converted analog signals into discrete ones. It became the prototype of modern artificial neural networks (ANN), and the model of its learning was close to the models of animal and human learning developed in psychology, see Bush & Mosteller (1951). Rosenblatt himself performed the first mathematical studies of the perceptron Rosenblatt (1959). However, the Novikoff theorem Novikoff (1962), which gives the conditions for the convergence of a perceptron learning algorithm in a finite number of steps, has become better known.

⋆ Supported by the Government of Russian Federation, project 08-08

[1] Cybernetics is understood here as control theory in a broad sense, including system identification, control, optimization and other disciplines under research within control community.

3. THE 1960s: FIRST BOOM OF MACHINE LEARNING

3.1 Deterministic approaches

In early 1960s several groups were engaged in the design and testing of learning recognition systems Widrow (1961); Bongard (1961); Braverman (1962). General statements of pattern recognition problem were proposed in 1963-1964 Abramson *et al*, (1963); Widrow (1964) in the groups of Mark Aizerman (including E.Braverman and L.Rozonoer) and Alexander Lerner (including V. Vapnik and A.Chervonenkis) in Moscow Institute of Control Science and in the group of Vladimir Yakubovich (including V.Fomin, A.Gelig and a number of younger researchers) in Leningrad State University LSU (currently St. Petersburg State University). In the papers Vapnik and Lerner (1963); Vapnik and Chervonenkis (1964) and Yakubovich (1963, 1965, 1966) deterministic approaches were developed, while in Aizerman *et al* (1964a,b) probabilistic statement of the machine learning problem was suggested.

Let us consider the developed algorithms of learning in more detail using a traditional simple example. Let each object (image) $X_k, k = 1, 2, ...N$ shown to the machine be encoded by a set of real numbers - values of some features: $X_k = (X_{k1}, X_{k2}, ..., X_{kn})^T$ and it is required to train the machine to recognize to which one of two classes: A or B belongs the presented new object $X*$. Let for certainty X_k belongs to class A for $k = 1, 2, ...k_A$ and class B for $k = k_A + 1, ..., N$. That is, the objects have a membership function $y(X)$ such that $y(X_k) = 1$ for $k = 1, 2, ...k_A$, where $1 < k_A < N$ and $y(X_k) = -1$ for $k = k_A + 1, ..., N$. It is assumed that the convex hulls of vectors X_k belonging to different classes do not intersect, so that these sets can be separated by a hyperplane. It means that there are a vector of weights $w = (w_1, w_2, ..., w_n)^T$ and a number w_0, such that $W(X_k) > 0$ for $k = 1, 2, ...k_A$ and $W(X_k) < 0$ at $k = k_A + 1, ..., N$, where $W(X) = w^T X + w_0$. Thus, the problem is reduced to the approximation of the function based on its values on a finite set. Gradient-type algorithms for perceptron learning construct some of the separating hyperplanes, and, as was shown in Novikoff (1962); Yakubovich (1963),

10.1016/j.ifacol.2020.12.1888

under some conditions perform it in a finite number of steps.

Let us make the transition to the conjugate space of weight vectors (w, w_0). Then the problem is transformed into a dual problem of finding the intersection of a finite number of half-spaces $(w, w_0) : y(X_k)(w^T X_k + w_0) > 0, k = 1, 2, ...N$. The dual problem can also be solved by different methods. If the objects are presented to machine sequentially, then a class of gradient-type recurrent algorithms can be used. They look as follows

$$w_{k+1} = w_k - \gamma_k y(X_k) X_k, w_{0,k+1} = w_{0,k} - \gamma_k y(X_k), \quad (1)$$

and require knowledge of only one image coordinates at each step. Different approaches have been proposed to select the size γ_k of the steps. In particular, it is possible to project the current vector of weights onto the boundary hyperplane at each iteration if the presented inequality is not satisfied, i.e. if the current object is classified incorrectly Yakubovich (1965, 1966) Since the weights are not corrected if the presented inequality is satisfied, the inequality defined a deadzone for the algorithm in the space of weights.

Another approach is to choose a vector of weights $w*$ in such a way that the corresponding hyperplane $\{x : w*^T x + w_0 = 0\}$ is a supporting hyperplane to the convex hull of the available set of vectors $y(X_k)X_k, k = 1, 2, ...N$, so that the minimum distance from it to the convex hull of classes is maximal. This idea formed the basis of the celebrated *support vector machine (SVM)* method. It is mentioned in many sources that this idea was first proposed in the work Vapnik and Chervonenkis (1964), which had an English translation and has become widely known. It is a historical fact, however, that in the same year 1964, an employee of the laboratory of theoretical cybernetics of LSU Boris Kozinets published another fairly simple recurrent algorithm converging to an optimal supporting hyperplane Kozinets (1964). Recurrent property of the algorithm means that it is based on processing objects (images) one by one, depending on availability (unlike nonrecurrent algorithms that need to have all the objects in the memory to process them).

Another algorithm, called MDM, was proposed at LSU by V. N. Malozemov, V. F. Demyanov and D. Mitchell Mitchell *et al* (1974). It is based on the formulation of the problem as minimax optimization and the use of nonsmooth optimization methods developed in the group of V.F.Demyanov working mostly on nonsmooth optimization problems. In 1965-66 V.A. Yakubovich developed a systematic approach to the mathematical theory of pattern recognition, called "the method of recurrent goal inequalities" Yakubovich (1965, 1966). It is based on the reduction of the problem to the solution of a system of inequalities constructed for a given purpose of functioning of the system and allows one to find a solution to an infinite number of previously not shown inequalities. This allowed Yakubovich with coauthors to apply the approach to solving problems of learning regulators for dynamical systems under uncertainty, i.e. to solve problems of adaptive control Yakubovich (1972). A number of application problems have been solved: training in handwriting recognition, aerial photographs, signal isolation from noise, description and analysis of scenes Kozinets *et al*, (1966); Kharichev *et al*, (1973).

Another powerful method was proposed in 1964 (again in LSU!) by young mathematician Lev Bregman (he was 23 years old in 1964). Bregman proposed a recurrent algorithm for finding a point in the intersection of the finite number of convex sets in a Hilbert space. The problem was motivated by an application problem of city planning that had nothing to do with learning. The Bregmans method consisted in evaluation of consecutive projections onto the nearest point of the set taken in the cyclic order. Weak convergence of the projections to the intersection of all sets was proved in the paper Bregman (1965) (the paper was recommended to publication by Leonid Kantorovich, future Nobel prize winner).

In his next paper Bregman (1966) Bregman proposed a useful functional transformation. Let $f(x)$ be a strictly convex twice differentiable function, $x \in R^n$. Let $D(x, y) = f(x) - f(y) - (gradf(y), x - y)$, where $gradf(x)$ is the gradient of the function $f(x)$. Function $D(x, y)$ turns out to be convenient to use for the convergence proofs as the Lyapunov function candidate. It was later called *Bregman divergence*. In the paper Bregman (1967) Bregman extended his previous results to present an elegant framework for convex optimization. It was later widely used for machine learning, clasterization, image deblurring, image segmentation, data reconstruction, etc. The terms *Bregman divergence* and *Bregman method* are now widely used: the number of the papers in the journals indexed in Scopus that have those terms (and related terms *Bregman projection*, *Bregman iteration*, etc. in the title exceeds 700 in October, 2019. As for the paper Bregman (1967) itself it has more than 1250 citations in Scopus. It is interesting that in the paper by Gubin *et al* (1967) a similar problem was studied and a number of results on strong convergence and convergence rate in Hilbert space were obtained without use of Bregman divergence. Also algorithms with incomplete relaxation were proposed and convergence in a finite number of steps was established as well as some applications to Chebyshev approximation and optimal control. The results of the paper by Gubin *et al* (1967) are also used by many authors in machine learning and related areas of optimization: it has about 500 citations in Scopus.

3.2 Stochastic approaches

The above mentioned methods are based upon deterministic approach where an uncertainty is modeled as an element of a bounded set. However many papers are devoted to statistical approach which roots can be traced back to statistical formulations of communications and decision theory Marill & Green (1960); Widrow (1959).

The most systematic framework based on average risk minimization was developed by Yakov Tsypkin Tsypkin (1966, 1971) who has brilliantly demonstrated that it encompasses a large number of algorithms proposed by different authors as special cases. Although the idea of average risk minimization appeared earlier in operation research and was introduced into control of uncertain systems by Feldbaum (1960), Tsypkin was the first who proposed a unified approach to adaptation and learning us-

ing stochastic approximation machinery. Parameter choice and convergence results then follow from the results on stochastic approximation obtained earlier in mathematical statistics. After the Tsypkins works the stochastic approximation has become a standard instrument in study of adaptation and learning algorithms. An important unifying idea introduced by Tsypkin (Tsypkin 1966, 1968) is in the formulation of the adaptation and learning problem in terms of a performance index which is the average of the cost function $Q(x, w)$:

$$J(w) = \int_X Q(x, w)p(x)dx = E_x Q(x, w)$$

That is the problem is formulated as

$$min_w J(w)$$

and its solution may be based on stochastic gradient algorithms:

$$w[n] = w[n-1] - \gamma[n]\nabla Q(x[n], w[n-1])$$

Choosing the cost function in appropriate way allowed the author to design different classes of algorithms described previously in the literature as well as a number of new ones. Convergence of the algorithms can be established based on the stochastic approximation scheme under conditions of convexity and bounded growth of $J(w)$, as well as the so called Robbins-Monro conditions on the sequence of $\gamma[n]$, namely

$$\gamma[n] > 0, \sum_n \gamma[n] = \infty, \sum_n \gamma[n]^2 < \infty.$$

4. TWO WINTERS OF AI AND SECOND BOOM OF MACHINE LEARNING

In 1969 the book by M.Minsky and S.Papert Minsky & Papert (1969) was published where some limitations for complexity of the problem that can be solved by perceptrons were established. Namely, the authors emphasized that perceptrons cannot represent some logical functions like XOR or NXOR. As a result, very little research was done in this area until about the 1980s.The book has triggered the reduction of funding of AI research in the world for more than two decades. This period was later called the first winter of AI. Nevertheless, study of more complex learning algorithms was still continuing.

Further studies of structures and learning abilities of multilayer neural networks took place in 1970-1980s. In 1980 Kunihiko Fukushima proposed a hierarchical multilayered convolutional neural network, known as neocognitron Fukushima (1980). A significant impact was made by the invention of backpropagation learning algorithm by several authors in mid-eighties Rumelhart *et al* (1986a,b), although its initial ideas were proposed still in the early 1960s, see Dreyfus (1990); Widrow & Lehr (1990); Werbos (1990). Compared to a standard gradient descent, which updates all the parameters with respect to error simultaneously, backpropagation first propagates the error term at output layer, back to the layer at which parameters need to be updated, and then uses the chain rule to update parameters with respect to the propagated error. Some drawbacks of backpropagation were discovered, see Brady *et al* (1989); Gori & Tesi (1992). Particularly it may fail in the case when the classes cannot be linearly separated. An intensive advertisement of the success of backpropagation and other computational advances produced great hope for future successes. However real successes were less that the expected ones and the investments in the area of machine learning decreased again in early 1990s. This period is sometimes called the *second winter of AI*.

Nevertheless an interest in studying neural networks as an instrument of machine learning was growing in the 1990s. A real breakthrough was produced by the paper Cortes & Vapnik (1995) where a new version of the SVM algorithm for the general non-separable case was introduced under the name *support-vector networks*. It has been widely used and been shown to be effective in practice. The algorithm and its theory have had a profound impact on theoretical and applied machine learning and inspired research on a variety of topics Mohri *et al* (2014). The paper Cortes & Vapnik (1995) has got more than 22000 citations in Scopus by 2019.

5. THE GOLD RUSH OF MACHINE LEARNING

The beginning of the first decade of the XXI century turned out to be a turning point in the history of ML, and this is explained by three synchronous trends, which together gave a noticeable synergistic effect. The first trend is Big Data. Amount of the data has became so big that new approaches were brought to life by practical necessity rather than by curiosity of scientists.

The second trend is to reduce the cost of parallel computing and memory. This trend was discovered in 2004, when Google unveiled its MapReduce technology, followed by its open-source counterpart Hadoop (2006), and together they made it possible to distribute the processing of huge amounts of data between simple processors. At the same time, Nvidia made a breakthrough in the GPU market: if earlier in the gaming segment it could compete with AMD/ATI, then in the segment of GPUs that can be used for machine learning purposes, it proved to be a monopolist. And at the same time, the cost of RAM has significantly decreased, which opened the possibility to work with large amounts of data in memory and, as a result, there are numerous new types of databases, including NoSQL. Finally, in 2014, the Apache Spark framework for distributed processing of unstructured and weakly structured data appeared, which was convenient for the implementation of machine learning algorithms.

The third trend is development of the new algorithms of *deep machine learning*, inheriting and developing the idea of perceptron in combination with a successful scientific PR-campaign. After many years of study of multilayer neural networks a concept a technology of *deep neural networks (DNN)* was born. It is believed that the term *deep learning* was proposed in 1986 by Rina Dechter Dechter (1986) although the history of its appearance is apparently more complicated. A detailed and preferably objective analysis of the events of this period is still waiting for its researcher. Different points of view on the deep learning can be found, e.g. on websites Foote (2017); Kurenkov (2015); Wang & Raj (2017).

By the middle of the last decade, a critical mass of knowledge in the field of DNN was accumulated, and, as always

in such cases, someone has broken away from the peloton. In this case, the leader was Geoffrey Hinton, a British scientist who continued his career in Canada. Since 2006, he and his colleagues began to publish numerous articles on DNN, including papers in the popular multidisciplinary journal Nature. This earned him a lifetime fame as a classic. Around him a strong and cohesive community has formed that worked for several years "in the invisible mode". Its members called themselves "Deep Learning Conspiracy" or even "Canadian Maffia". A leading trio was formed: Ian LeCun, Yehoshua Benjo and Geoffrey Hinton, they are also called LBH (LeCun & Bengio & Hinton). LBH's exit from the underground was well prepared and supported by Google, Facebook and Microsoft. Andrew Ng, who worked at MIT and Berkeley and is now the head of artificial intelligence research at Baidu lab, worked extensively with LBH. He linked deep learning to GPUs. Finally Ian LeCun, Yehoshua Benjo and Geoffrey Hinton were awarded with Turing Award in 2018 "for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing". The leader of the trio Geoffrey Hinton has more than 300,000 citations in Google Scholar by November, 2019. Quite a number of books and textbooks, are devoted to machine learning, see e.g. Bishop (2006); Mohri *et al* (2014); Theodoridis (2015), including Vapnik (1998) that has more than 83,000 citations in Google Scholar. However books are becoming obsolete very rapidly and the best way to feel the state-of-the-art is to participate in one of flagman conferences, e.g. NeuIPS or ICML. Canada has become the Mecca of ML: it will host NeurIPS in 2019 and 2020.

6. DISCUSSION

Looking through the several decades of the history of adaptation and learning it is seen that during the first few decades those two areas were very close to each other and a fruitful cooperation between them can be observed. However the dramatic development of the machine learning during the last two decades leads to the divergence of the two sister fields which is an issue of discussions in the control community. However control related views have been published only for the more narrow areas of iterative learning Bristow *et al* (2006) and reinforcement learning Recht (2018).

This paper is an attempt to present a general control related view of the issue in the historical perspective. Attention is focused mainly on the first decades of the history of machine learning. The possibility of more tight interaction between adaptive control and machine learning in future is still unclear although some papers coauthored by both experts in control and experts in machine learning are published and well accepted. E.g. paper by Belhumeur *et al* (1997) has more than 7000 citations after 20 years, while paper by Wright *et al* (2009) got more that 5400 citations after ten years.

Looking into the past one can see a few success stories of applying the ideas well known in optimization and control area to machine learning. One of them is application of multistep tuning algorithms to increase the convergence rate of gradient type algorithms. E.g. acceleration of convergence by "heavy ball" method or averaging is known in optimization and control since the 1960s Polyak (1964); Tsypkin (1971); Nesterov (1983); Polyak & Juditsky (1992). However 300 out of about 320 citations of the seminal paper Polyak (1964) were made in the last five years.

Looking into the future we see new expansion of the field. The top international conference on ML 'Neural Information Processing Systems' (NeurIPS, formely NIPS, neurips.cc) had more than 8000 participants in 2018, while second top conference 'International Conference on Machine Learning' (ICML, icml.cc) got more than 6000 participants in 2019.

What is a challenge for control theorists is that there are very few rigorous mathematical proofs in a new deep learning 'tsunami'. Since machine learning researchers needed means to compare the effectiveness of their methods and it was hard to find solid mathematical means to test performance of the numerous proposed methods [2] , the standard datasets of training and testing sets that could be used to evaluate machine learning algorithms were proposed.

The author would be happy if this paper would encourage the readers to seek for new links between "old good" adaptation and learning theory of control/cybernetics and new paradigms and trends of *machine learning* and *deep neural networks*. Strengthening such links would be of mutual benefit for both fields.

REFERENCES

Abramson N., D. Braverman and G. Sebestyen (1963). Pattern recognition and machine learning IEEE Transactions on Information Theory. V.: 9, Is.: 4.

Aizerman M.A. (1963). The problem of training an automaton to perform classification of input situations (pattern recognition), Theory Self-Adapt. Contr. Syst. Proc. IFAC Symp. 2nd 1963.

Aizerman, M.A., Braverman, E.M., and Rozonoer, L.I. (1964a). Theoretical foundations of the potential function method in the problem of training automata to classify input situations, Automat. Remote Contr. (USSR) 25 (6).

Aizerman, M.A., Braverman, E.M., and Rozonocr, L.I. (1964b) The probabilistic problem of training automata to recognize patterns and the potential function method, Automat. Remote Contr. (USSR) 25 (9).

Belhumeur P.N., Hespanha J.P. and D.J. Kriegman (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection IEEE Transactions on Pattern Analysis and Machine Intelligence, V.: 19, Is.: 7.

Bishop C.M. (2006).Pattern Recognition and Machine Learning. Springer.

Bongard M. M. (1961). Simulation of the recognition process on a digital computing machine. Biophysics, vol. 4, No. 2.

Brady M.L., Raghavan R., and J. Slawny. Back propagation fails to separate where perceptrons succeed. IEEE Transactions on Circuits and Systems, 36(5):665–674, 1989.

[2] Perhaps the only exclusion is the Vapnik-Chervonenkis theory of uniform convergence of the frequencies to the probabilities produced the notion of Vapnik-Chervonenkis (VC) dimension.

Braverman, E.M. (1962). The experiments with training a machine to recognize patterns, Automat. Remote Contr. 23 (3).

Bregman L.M. (1966) Relaxation method of finding a point common to some given convex sets and its use in solving optimization problems Doklady Akademii Nauk SSSR Vol.: 171 Issue: 5 pp.: 1019–1023.

Bregman, L.M. (1965) Use of consecutive projection for finding a common point of convex sets Doklady Akademii Nauk SSSR Volume: 162 Issue: 3 pp. 487–490.

Bregman L.M. (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics 7(3), pp. 200–217.

Bristow D.A., Tharayil M., G. Alleyne A.G. (2006) A Survey of Iterative Learning Control." IEEE Control Systems Magazine, Is.1, pp.96–114.

Bush, R. R., and Mosteller, F. (1951). A mathematical model for simple learning. Psychological Review, 58(5), 313–323.

Cortes C, and Vapnik, V.(1995). Support-vector networks. Machine Learning. V.: 20 Is.: 3, pp. 273–297, 1995.

Dechter R. (1986) Learning while searching in constraint-satisfaction problems. AAAI-86 Proceedings, pp.179–183.

Dreyfus, S. E. (1990). "Artificial Neural Networks, Back Propagation, and the Kelley-Bryson Gradient Procedure". Journal of Guidance, Control, and Dynamics. 13 (5): 926-928.

Feldbaum, A. A. (1960). Dual Control Theory, Automation and Remote Control , Vol. 21, Parts I, II, April 1961, pp. 874-880, and May 1961, pp. 1033–1039.

Foote K.D. (2017) A Brief History of Deep Learning. on February 7, 2017 https://www.dataversity.net/brief-history-deep-learning/

Fukushima K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4): 193–202.

Gori M. and A. Tesi. (1992). On the problem of local minima in backpropagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(1):76–86.

Gubin L.G., Polyak B.T., Raik E.V. (1967). The method of projections for finding the common point of convex sets. USSR Computational Mathematics and Mathematical Physics 7(6), pp. 1–24.

Kharichev V.V., Schmidt A.A. and V. A. Yakubovich (1973). New problem in pattern recognition. Autom. Remote Control, V.34, Is. 1, 98–109.

Kozinets B.N. (1964) On one algoritm for learning a linear perceptron. In Vichislitelnaia Tekhnika i Voprosi Programirovania, Vol. 3. Leningrad State University Press, Leningrad. (In Russian.)

Kozinets, B.N; Lantsman, R.M; Yakubovich V.A. Use Of Electron Computers In Criminalistics For Differentiation Between Very Similar Handwritings. Doklady Akademii Nauk SSSR V.: 167 Is.: 5 pp. 1008–1011 : 1966.

Kurenkov A. (2015). A 'Brief' History of Neural Nets and Deep Learning https://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/

T. Marill and D. M. Green (1960). Statistical Recognition Functions and the Design of Pattern Recognizers. IRE Transactions on Electronic Computers. V.: EC-9, Is.: 4.

M. Minsky, and S. Papert (1969) Perceptrons: An Introduction to Computational Geometry MIT Press, Cambridge, MA, USA.

Mitchell B.F., Dem'yanov V.F., Malozemov V.N. (1974) Finding The Point Of A Polyhedron Closest To The Origin. SIAM J Control 12(1), pp. 19-26 (Translated from: Vestnik Of Leningrad Univerity. 1971. Is. 19, pp. 38-45.)

Mohri M., Rostamizadeh A. and Talwalkar A. (2014) Foundations of Machine Learning, MIT Press.

Nesterov Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 27(2), 372-376.

Novikoff, A. B. (1962). On convergence proofs on perceptrons. Symposium on the Mathematical Theory of Automata, Polytechnic Institute of Brooklyn 12, 615–622.

Polyak, B.T. (1964). Some methods of speeding up the convergence of iteration methods, USSR Computational Mathematics and Mathematical Physics 4(5), . 1–17.

Polyak, B.T. and A.B. Juditsky. (1992) Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization 30(4), . 838–855.

Recht B. (2018) A Tour of Reinforcement Learning: The View from Continuous Control. `arXiv:1806.09460v2 [math.OC] 10 Nov 2018.`

Rosenblatt F.(1957). The Perceptron, A Perceiving and Recognizing Automaton, Project Para Report No. 85-460-1, Cornell Aeronautical Laboratory (CAL).

Rosenblatt F. (1959). Two Theorems of Statistical Separability in the Perceptron. Symposium of the Mechanisation of Thought Processes. National Physical Laboratory, Teddington, UK, Nov. 1958, Vol I, H.M. Stationery Office, London.

Rosenblatt, F. (1960). Perceptron Simulation Experiments. Proc. Inst. Radio Engineers V. 18, 3 March 1960, 301–309.

Rumelhart D.E., Hinton G.E. and R. J. Williams (1986a) Learning internal representations by error propagation, in Parallel Distributed Processing, vol. 1, ch. 8, D. E. Rumelhart and J. L. McClelland, Eds., Cambridge, MA: M.I.T. Press, 1986.

Rumelhart D.E., Hinton G.E. and R. J. Williams (1986b) "Learning representations by back-propagating errors". Nature. 323 (6088), 533-536.

Theodoridis S. (2015). Machine Learning. A Bayesian and Optimization Perspective, Elsevier.

Tsypkin, Y.Z. (1966) Adaptation, Training And Self-Organization In Automatic Systems. Automation And Remote Control, V. 27, Is.: 1, pp. 16–51.

Tsypkin Ya.Z. *Adaptation and learning in automated systems* NY, Academic Press, 1971 (Russian original 1968).

Vapnik V.N. (1998) The Nature of Statistical Learning Theory. Springer.

Vapnik V. N. and A. Ya. Lerner. Recognition of Patterns with help of Generalized Portraits, Avtomat. i Telemekh., 24:6 (1963), 774-780.

Vapnik, V. and Chervonenkis, A. (1964). On a class of perceptrons. Automat. Remote Control 25, 103-109.

Wang H., and B. Raj (2017) On the Origin of Deep Learning arXiv:1702.07800v4 [cs.LG] 3 Mar 2017

Werbos P.J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10):1550–1560.

Widrow, B. (1959). Adaptive sampled-data systems. A statistical theory of adaptation, WESCON Conv. Rec. 3 (4).

Widrow, B. (1961). Self-adaptive discrete systems, Theory Self Adapf. Contr. Syst. Proc. IFAC Symp. 1st 1961.

Widrow, B. (1964). Pattern recognition and adaptive control, IEEE Trans. Appl. Ind. 83 (74).

Widrow B. and M.Lehr. (1990) 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. Proceedings IEEE, V. 78, Is.9, 1990.

Wright J., Yang A.Y., Ganesh A., Sastry S., Yi Ma (2009). Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence V.: 31, Is.: 2.

Yakubovich, V.A. (1963). Machines that learn to recognize patterns. In Metodi Vichisleniy, Vol. 2. Leningrad State University Press, Leningrad. (In Russian. To be translated from the reprint: Vestnik of Saint-Petersburg University. Mathematics, 4, 2019).

Yakubovich, V.A. (1965). Certain general theoretical principles in the design of learning pattern recognition systems, Part I. In Vichislitelnaia Tekhnika i Voprosi Programirovania, Vol. 4. Leningrad State University Press, Leningrad. (In Russian.)

Yakubovich, V.A. (1966). Recurrent finitely convergent algorithms for solving systems of inequalities. Sov. Math. Doklady 1966, V. 7, pp. 300–304.

Yakubovich V.A. On a method of adaptive control under conditions of great uncertainty. Prepr. 5th World Congress IFAC (Paris). 1972. V.37. N3. pp.1–6.