University of Groningen

The Statistical Physics of Learning Revisited

Biehl, Michael

Published in:
Brain-Inspired Computing

DOI:
10.1007/978-3-030-82427-3_10

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Biehl, M. (2021). The Statistical Physics of Learning Revisited: Typical Learning Curves in Model Scenarios. In K. Amunts, L. Grandinetti, T. Lippert, & N. Petkov (Eds.), *Brain-Inspired Computing : 4th International Workshop, BrainComp 2019, Revised Selected Papers* (pp. 128-142). (Lecture Notes in Computer Science ; Vol. 12339). Springer. https://doi.org/10.1007/978-3-030-82427-3_10

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 04-06-2022

The Statistical Physics of Learning Revisited: Typical Learning Curves in Model Scenarios

Michael Biehl(✉)

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence,
University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands
m.biehl@rug.nl
http://www.cs.rug.nl/~biehl

Abstract. The exchange of ideas between computer science and statistical physics has advanced the understanding of machine learning and inference significantly. This interdisciplinary approach is currently regaining momentum due to the revived interest in neural networks and deep learning. Methods borrowed from statistical mechanics complement other approaches to the theory of computational and statistical learning. In this brief review, we outline and illustrate some of the basic concepts. We exemplify the role of the statistical physics approach in terms of a particularly important contribution: the computation of typical learning curves in student teacher scenarios of supervised learning. Two, by now classical examples from the literature illustrate the approach: the learning of a linearly separable rule by a perceptron with continuous and with discrete weights, respectively. We address these prototypical problems in terms of the simplifying limit of stochastic training at high formal temperature and obtain the corresponding learning curves.

1 Introduction

At least two major developments have led to the regained popularity of machine learning in general and neural networks in particular [1–6]. Most importantly, the ever-increasing availability of training data from various domains and contexts have made possible the training of very powerful systems such as deep neural networks [4–6]. At the same time, the computational power necessary for the data driven adaptation and optimization of such systems, has become available.

Several concepts that had been developed earlier, some of them even decades ago, could be realized and applied successfully in practice only recently. Examples and further references can be found in [4–6]. In addition, novel computational techniques and important modifications of the considered systems have contributed to this success. This includes the use of pre-trained networks, sophisticated regularization techniques, weight sharing in convolutional neural networks, or the use of alternative activation functions [4–8].

While the relevance and success of the methods are widely recognized, several authors note that the theoretical understanding does not yet parallel the

© The Author(s) 2021
K. Amunts et al. (Eds.): BrainComp 2019, LNCS 12339, pp. 128–142, 2021.
https://doi.org/10.1007/978-3-030-82427-3_10

practical advances of the field, see for instance [9–13] in the context of deep learning. It is certainly desirable to strengthen and put forward the theoretical investigation of machine learning processes in general and deep learning in particular. The development of novel concepts and the design and optimization of practical training prescriptions would greatly benefit from better theoretical understanding. This concerns, for instance, mathematical and statistical foundations, the dynamics of training, and insights into the expected generalization ability of learning systems.

Concepts borrowed from statistical mechanics have been applied in many areas beyond the scope of traditional physics. In particular, analytical and computational approaches developed for the study of complex physical systems can be exploited within computer science and statistics. A prominent example is the use of Markov chain Monte Carlo methods [14], which exploit mathematical analogies between stochastic optimization and the statistical physics of systems with many degrees of freedom. Similarly, analytical methods which had been developed for the analysis of disordered systems [15], have been applied in this context.

A somewhat surprising and very inspiring analogy was pointed out by John Hopfield [16]: the conceptual similarity of simple dynamical neural networks with models of disordered magnetic materials [15]. It attracted considerable interest in neural networks and related systems within the physics community. Initially, the analysis of thermal equilibrium states in so-called attractor neural networks was in the center of interest [1,16,17]. However, the same concepts were applied successfully to the investigation of learning and synaptic plasticity in neural networks. Elizabeth Gardner's pioneering work [18,19] paved the way for the theory of learning in a large variety of machine learning scenarios, including the supervised training of feedforward neural networks, methods of unsupervised data analysis, and more general inference problems, see [20–22] for reviews.

A variety of analytical tools and modelling frameworks have been developed and applied successfully to, for instance, the study of supervised learning in the context of regression and classification tasks. Mostly, relatively simple and shallow feedforward neural networks have been analysed [20–22]. Frequently, these training processes are modelled in the frame of a student and teacher scenario. There, a specific neural network, the teacher, is assumed to define the target task, e.g. a classification scheme. A student network is trained from a set of examples provided by the teacher and parameterizes a data driven hypothesis about the target rule. This allows to explicitly control the complexity of the target rule and of the learning system in the model. Moreover the performance of the trained system can be quantified in terms of its similarity and agreement with the teacher network. Learning can be interpreted as the stochastic optimization of many degrees of freedem, which motivates possible training algorithms based on statistical mechanics ideas. Also, analytical tools for the study of large systems in (formal) thermal equilibrium situations can be used, which describe the model in terms of a few macroscopic quantities, only. Frequently, these so-called order parameters appear naturally when analysing student teacher scenarios.

Ultimately, methods developed in the theory of disordered systems allow for the investigation of typical properties of the learning system. This concerns, for instance, the computation of learning curves as an outcome of the stochastic training process on average over the assumed randomness in the example data.

The successful applications of these concepts include, among other relevant topics, the highly interesting phenomenon of symmetry breaking phase transitions which result in discontinuous learning curves: Frequently, the success of training is found to depend critically on the number of available examples or other model parameters [20–25]. Currently, the interest in this type of analysis is gaining momentum again in the context of deep learning and other popular learning paradigms, see [26–31] for examples.

In the following section, we briefly outline and illustrate the statistical physics of student teacher scenarios in supervised learning. We present two variants of a simple and by now classical example: the learning of a linearly separable rule with a perceptron network with continuous or discrete weights, respectively. The perceptron has been discussed extensively in the literature and serves as a prototypical system for the understanding of machine learning processes, see e.g. [1,20–22]. For the sake of brevity, we focus on a particularly simplifying approach, the consideration of stochastic training in the limit of high (formal) temperature. It was introduced and applied to perceptron training in [22]. Despite its conceptual simplicity and mathematical ease, this example illustrates the basic concepts very well and yields non-trivial results and insights into the learning process.

This contribution is based on a tutorial talk at the Workshop on Brain Inspired Computing, BrainComp 2019. It is obviously far from providing a complete overview of the statistical physics of learning. The intention is to attract the reader's attention in terms of selected example applications of the approach and to provide references as a starting point for further exploration of this highly relevant area of research.

2 Statistical Physics of Learning: Learning Curves

Typically, the statistical physics based computation of learning curves in supervised learning proceeds along the following steps:

1) A student and teacher scenario is defined, which parameterizes the target rule and fixes the complexity of the student hypothesis.
2) It is assumed that training examples and test instances are generated according to a specific input density, while target labels are provided by the teacher network.
3) The study of large systems in the thermodynamic limit allows to describe systems in terms of relatively few macroscopic quantities or order parameters.
4) The outcome of stochastic training processes is interpreted as a formal thermal equilibrium, in which thermal averages can be considered.

5) An additional disorder average over a randomly generated set of training data is performed in order to obtain typical results independent of the actual training set.

The following sections illustrate the above points in the context of learning a linearly separable rule [20–22], before two concrete example scenarios are analysed in Sect. 2.6.

2.1 Learning a Linearly Separable Rule: Student and Teacher

We consider the supervised learning of a linearly separable classification of N-dimensional data. In our model, the target rule is defined through a teacher perceptron with fixed weight vectors $\mathbf{w}^* \in \mathbb{R}^N$ and output

$$S^*(\boldsymbol{\xi}) = \text{sign}\left[\mathbf{w}^* \cdot \boldsymbol{\xi}\right] = \pm 1 \quad \text{for any } \boldsymbol{\xi} \in \mathbb{R}^{\mathrm{N}}. \tag{1}$$

Here, the feature vector $\boldsymbol{\xi}$ represents N numerical inputs to the system and S^* corresponds to the correct output. The teacher weight vector parametrizes an $(N-1)$-dim. hyperplane which separates positive from negative responses.

We note that the norm $|\mathbf{w}^*|$ of the weights is irrelevant for the perceptron response (1). Throughout the following, we therefore consider normalized teacher weights with $\mathbf{w}^* \cdot \mathbf{w}^* = N$.

In the learning scenario, information about the rule is only available in the form of a data set which comprises P examples:

$$\mathbb{D} = \{\boldsymbol{\xi}^\mu, S^*(\boldsymbol{\xi}^\mu)\}_{\mu=1,2,\ldots,P} \,. \tag{2}$$

Here we assume that the labels $S^{*\mu} = S^*(\boldsymbol{\xi}^\mu)$ provided in $\mathbb{D}$ are reliable and represent the rule (1) faithfully. We refrain from considering corruption by different forms of noise, for simplicity, and refer the reader to the literature for the corresponding extensions of the analysis [20, 21].

A second simple perceptron serves as the student network in our model. Its adaptive weights $\mathbf{w} \in \mathbf{R}^N$ parameterize a linearly separable function

$$S(\xi) = \text{sign}\left[\mathbf{w} \cdot \boldsymbol{\xi}\right]. \tag{3}$$

The weight vector $\mathbf{w}$ is chosen in a data-driven training process which is based on the available data $\mathbb{D}$ and corresponds to the student hypothesis about the unknown target. As a consequence of the invariance

$$\text{sign}\left[\,(\lambda \mathbf{w}) \cdot \boldsymbol{\xi}\right] = \text{sign}[\,\mathbf{w} \cdot \boldsymbol{\xi}] \quad \text{for arbitrary } \lambda > 0$$

we will also consider normalized student weights with $\mathbf{w} \cdot \mathbf{w} = N$ in the following.

2.2 The Density of Input Data

In realistic learning situations it is expected that the density of input features is correlated with the actual task to a certain extent. In real world classification problems, for instance, one would expect a more or less pronounced cluster structure which reflects the class memberships already. Clustered or more generally structured input densities have been considered in the statistical physics literature, see [26] for a recent discussion and further references. Here, however, we follow the most frequent approach and resort to the simplifying assumption of an isotropic input density which generates input vectors independently. In a sense, this constitutes a worst case in which the only information about the target rule is contained in the assigned training labels $S^*(\boldsymbol{\xi})$, while no gap or region of low density in feature space marks the class boundaries.

Specifically, we assume that components of example vectors $\boldsymbol{\xi}^\mu$ in $\mathbb{D}$ consist of independent, identically distributed (i.i.d.) random quantities with means and covariances

$$\left\langle \xi_j^\mu \right\rangle \,=\, 0, \qquad \left\langle \xi_j^\mu \xi_k^\nu \right\rangle \,=\, \delta_{\mu\nu}\, \delta_{jk} \tag{4}$$

with the Kronecker symbol $\delta_{mn}\!=\!1$ if $m\!\neq\! n$ and $\delta_{mm}\!=\!0$.

2.3 Generalization Error and the Perceptron Order Parameter

The performance of a given weight vector $\mathbf{w}$ in the student teacher model can be evaluated with respect to a test input $\boldsymbol{\xi} \notin \mathbb{D}$. If we assume that the test input follows the same statistics as the training examples, i.e.

$$\langle \xi_j \rangle \,=\, 0, \qquad \langle \xi_j \xi_k \rangle \,=\, \delta_{jk}, \tag{5}$$

we can define the so-called generalization error as the expectation value

$$\epsilon_g(\mathbf{w},\mathbf{w}^*) = \langle \epsilon\left(S(\boldsymbol{\xi}, S^*(\boldsymbol{\xi}))\right\rangle \quad \text{where} \quad \epsilon(S,S^*) = \begin{cases} 1 \text{ if } S \neq S^* \\ 0 \text{ else,} \end{cases} \tag{6}$$

serves as a binary error measure. Hence, the generalization error quantifies the probability for disagreement between student and teacher for a random input vector. It is instructive to work out ϵ_g explicitly under the assumption of i.i.d. inputs. To this end, we consider the arguments of the threshold function in student and teacher perceptron:

$$x = \mathbf{w}\cdot\boldsymbol{\xi}/\sqrt{N} \;\text{ and }\; x^* = \mathbf{w}^*\cdot\boldsymbol{\xi}/\sqrt{N}.$$

Assuming that the random input vector $\boldsymbol{\xi}$ satisfies Eq. (5), x and x^* correspond to sums of N random quantities. By means of the Central Limit Theorem (CLT) there density is given by a two-dim. Gaussian, which is fully specified by first and second moments. These can be obtained immediately as

$$\langle x\rangle = \langle x^*\rangle = 0, \quad \left\langle x^2\right\rangle = \frac{1}{N}\sum_{i,j} w_i\, w_j\, \langle \xi_i \xi_j\rangle \;=\; \frac{\mathbf{w}^2}{N} = 1, \quad \left\langle (x^*)^2\right\rangle = \frac{(\mathbf{w}^*)^2}{N} = 1$$

$$\text{and} \quad \langle x\, x^*\rangle = \frac{1}{N}\sum_{i,j} w_i\, w_j^*\, \langle \xi_i \xi_j\rangle \;=\; \frac{\mathbf{w}\cdot\mathbf{w}^*}{N} \equiv R, \tag{7}$$

where we have exploited the normalization of weight vectors. The covariance $\langle x x^*\rangle$ is given by the scalar product of student and teacher weights. The moments

(7) fully specify the two-dimensional normal density $P(x, x^*)$ and we obtain the generalization error as the probability of observing $xx^* < 0$:

$$\epsilon_g(\mathbf{w}, \mathbf{w}^*) = \left[\int_{-\infty}^{0} \int_{0}^{\infty} + \int_{0}^{\infty} \int_{-\infty}^{0} \right] P(x, x^*) dx dx^* = \frac{1}{\pi} \arccos(R). \quad (8)$$

This result can be obtained immediately by an intuitive argument: The probability for a random vector $\boldsymbol{\xi}$ to fall into the hypersegments between the hyperplanes defined by $\mathbf{w}$ and $\mathbf{w}^*$ is directly given by $\angle(\mathbf{w}, \mathbf{w}^*)/\pi$ which corresponds to the right hand side of Eq. (8).

In the following, the overlap $R = \mathbf{w} \cdot \mathbf{w}^*/N$ plays the role of an order parameter. This macroscopic quantity summarizes essential properties of the N microscopic degrees of freedom, i.e. the adaptive student weights w_j. It is also the central quantity in the following analysis of the training outcome.

2.4 Training as a Stochastic Process and Thermal Equilibrium

The outcome of any practical training process will clearly depend on the actual choice of an algorithm and its parameters that is used to infer a suitable weight vector $\mathbf{w}$ from a given data set $\mathbb{D}$. Generically, the training process is guided by a cost function, such as the quadratic deviation of the student output from the target in regression systems or the number of incorrect responses in a classification problem.

Frequently, gradient based methods can be used for the optimization of continuous weights $\mathbf{w} \in \mathbb{R}^N$, often incorporating some form of noise as in the popular stochastic gradient descent. The search for optimal weights in a discrete space with, e.g., $\mathbf{w} \in \{-1, +1\}^N$ could be performed by means of a Metropolis Monte Carlo method, as an example.

The degree to which the system is forced to approach the actual minimum of the cost function is controlled implicitly or explicitly in the training algorithm. Example control parameters are the learning rate in gradient descent or the temperature parameter in Metropolis like schemes. In the statistical physics approach to learning, this concept is taken into account by considering a formal thermal equilibrium situation as outlined below.

In the context of the perceptron student teacher scenario we consider a cost function of the form

$$H(\mathbf{w}) = \sum_{\mu=1}^{P} \epsilon(S^\mu, S^{*\mu}) \quad \text{with} \quad S^\mu = \text{sign}[\mathbf{w} \cdot \boldsymbol{\xi}^\mu], \quad S^{*\mu} = \text{sign}[\mathbf{w}^* \cdot \boldsymbol{\xi}^\mu]. \quad (9)$$

With the binary error measure of Eq. (6), the cost function represents the number of disagreements between student and teacher for a given data set.

Without referring to a particular training prescription we can describe the outcome of suitable stochastic procedures in terms of a Gibbs-Boltzmann density of weight vectors

$$P_{eq}(\mathbf{w}) = \frac{e^{-\beta H(\mathbf{w})}}{Z} \quad \text{with } Z = \int d\mu(\mathbf{w}) \; e^{-\beta H(\mathbf{w})}. \quad (10)$$

It describes a canonical ensemble of trained networks in thermal equilibrium at formal inverse temperature $\beta = 1/T$. The cost function $E(\mathbf{w})$ plays the role of the energy of state $\mathbf{w}$ and the normalization Z is known as the partition function. The measure $d\mu(\mathbf{w})$ is implicitly understood to incorporate restrictions of the N-dimensional integration such as the normalization $\mathbf{w}^2 = N$. Similarly, Z can be written as a sum over all possible weight configurations for systems with $\mathbf{w} \in \{-1, +1\}^N$.

In the limit $\beta \to \infty, T \to 0$, only the groundstate with minimal energy can be observed in the ensemble, as any other state will have an exponentially smaller P_{eq}. On the contrary, for $\beta \to 0, T \to \infty$, the energy becomes irrelevant and every state P_{eq} can occur with the same probability. In general, the parameter β controls the mean energy of the system which can be written as a thermal average of the form

$$\langle H \rangle_\beta \;=\; \int d\mu(\mathbf{w})\, H(\mathbf{w})\, \frac{e^{-\beta H(\mathbf{w})}}{Z} \;=\; -\frac{\partial}{\partial \beta} \ln Z. \tag{11}$$

Quite generally, thermal averages can be written as appropriate derivatives of the so-called free energy $F = -\frac{1}{\beta} \ln Z$, which is also in the center of the following analysis. Introducing the microcanonical entropy $S(E)$ we can rewrite

$$Z = \int dE\, e^{-\beta E + S(E)} \quad \text{where} \quad S(E) \;=\; \ln \int d\mu(\mathbf{w})\, \delta[H(\mathbf{w}) - E] \tag{12}$$

with the Dirac delta-function $\delta[\ldots]$. For large systems in the thermodynamic limit $N \to \infty$ we assume that entropy and energy are extensive, i.e. that $S = N\, s$ and $E = N\, e$ with $e, s = \mathcal{O}(1)$. A saddle-point integration yields

$$\lim_{N\to\infty} \left(-\ln Z/N\right) \;=\; \beta F/N \;=\; \beta\, e - s(e) \tag{13}$$

where the right hand side $\beta F/N$ has to be evaluated in its minimum with respect to e for a given β.

2.5 Disorder Average and High-Temperature Limit

The consideration of a formal thermal equilibrium in the previous section refers to a particular data set $\mathbb{D}$, since the energy function $H(\mathbf{w})$ is defined with respect to the given example data. In order to obtain typical results independent of the particularities of a specific data set, an additional average over randomly generated $\mathbb{D}$ has to be performed.

In the simplest case, we consider data sets which comprise P independent vectors $\boldsymbol{\xi}^\mu$ with i.i.d. components that obey (4). Hence the corresponding density factorizes over examples $\mu = 1, 2, \ldots P$ and components $j = 1, 2, \ldots N$ of the feature vectors in $\mathbb{D}$.

The randomness in $\mathbb{D}$ can be interpreted as an external disorder which determines the actual energy function $H(\mathbf{w})$ and the corresponding thermal equilibrium. In addition to the thermal average discussed in the previous section, the associated quenched average is denoted as $\langle \ldots \rangle_{\mathbb{D}}$. Quantities of interest have to be studied in terms of appropriate averages of the form $\langle \langle \ldots \rangle_\beta \rangle_{\mathbb{D}}$ which can be derived from the quenched free energy

$$\langle F \rangle_{\mathbb{D}} = - \, \langle \ln Z \rangle_{\mathbb{D}} / \beta.$$

The computation of $\langle \ln Z \rangle_{\mathbb{D}}$ is, in general, quite involved and requires the application of sophisticated methods such as the replica trick [1,15,20–22].

We refrain from discussing the conceptual difficulties and mathematical subtleties of the replica approach. Instead we resort to a very much simplifying limit, which has been presented and discussed in [22]. In the extreme setting of learning at high formal temperature with $\beta \to 0$, the so-called annealed approximation

$$\langle \ln Z \rangle_{\mathbb{D}} \approx \ln \langle Z \rangle_{\mathbb{D}}$$

becomes exact and can be exploited to obtain the typical training outcome [20–22]. Note that in this limit also

$$\langle Z \rangle_{\mathbb{D}} = \left\langle \int d\mu(\mathbf{w}) e^{-\beta H(\mathbf{w})} \right\rangle_{\mathbb{D}} = \int d\mu(\mathbf{w}) e^{-\beta \langle H(\mathbf{w}) \rangle_{\mathbb{D}}} \quad \text{with}$$

$$\langle H(\mathbf{w}) \rangle_{\mathbb{D}} = \sum_{\mu=1}^{P} \left\langle \epsilon(S^\mu, S^{*\mu}) \right\rangle_{\mathbb{D}} = P \, \epsilon_g . \rangle_{\mathbb{D}}. \tag{14}$$

Here we make use of the fact that the i.i.d. random examples in $\mathbb{D}$ contribute the same average error which is given by ϵ_g. It is expressed as a function of the order parameter R in Eq. (8). We can now perform a saddle point integration in analogy to Eqs. (12, 13) to obtain

$$\lim_{N \to \infty} (- \ln \langle Z \rangle_{\mathbb{D}} / N) \; = \; \beta \langle F \rangle_{\mathbb{D}} / N \; = \frac{\beta P}{N} \epsilon_g(R) - s(R). \tag{15}$$

Again, the right hand side has to be evaluated in its minimum, now with respect to the characteristic order parameter R of the system. The entropy term

$$s(R) = \frac{1}{N} \ln \int d\mu(\mathbf{w}) \delta[\mathbf{w} \cdot \mathbf{w}^* - NR] \tag{16}$$

can be obtained analytically by an additional saddle point integration making use of the integral representation of the δ-function [1,20–22]. Since $s(R)$ depends on potential constraints on the weight vectors as represented by $d\mu(\mathbf{w})$, we postpone the computation to the following sections.

In order to obtain meaningful results from the minimization with respect to R in Eq. (15), we have to assume that the number of examples P scales like

$$P = \alpha \, N / \beta \quad \text{with } \alpha = \mathcal{O}(1). \tag{17}$$

Obviously, P should be proportional to the number N of adaptive weights in the system, which is consistent with an extensive energy. In addition, P has to

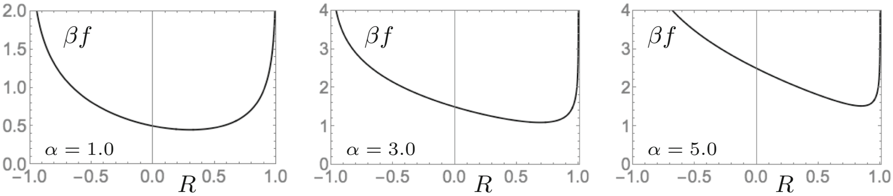

Fig. 1. The quenched free energy βf as a function of the order parameter R in the training scenario with spherical student and teacher perceptron, cf. Sect. 2.6. From left to right, the rescaled numbers of examples are $\alpha = 1.0, 3.0$ and 5.0, respectively.

grow like β^{-1} in the high temperature limit. The weak role of the energy in this limit has to be compensated for by an increased number of example data. In layman's terms: *"Almost nothing is learned from infinitely many examples"*. This also makes plausible the identification of the energy with the generalization error. The space of possible input vectors is sampled so well that training set performance and generalization behavior become indistinguishable.

Finally, the quenched free energy per weight, $f = \langle F \rangle_{\mathbb{D}}/N$ of the perceptron model in the high temperature limit has the form

$$\beta f \;=\; \alpha \, \epsilon_g(R) \;-\; s(R), \tag{18}$$

where α plays the role of an effective temperature parameter, which couples the number of examples and the formal temperature of the training process. These quantities cannot be varied independently within the simplifying limit $\beta \to 0$ in combination with $P/N \propto \beta^{-1}$.

2.6 Two Concrete Examples

Despite the significant simplifications and scaling assumptions, it is possible to obtain non-trivial, interesting results also in the high temperature limit. Very often, more sophisticated approaches, such as the replica method or the annealed approximation for finite training temperatures, confirm the results for $\beta \to 0$ qualitatively. Therefore, the simplified treatment has often been used to obtain first, useful insights into the qualitative properties of various learning scenarios. In this brief review, we restrict the discussion to two well-known results for simple model situations. Both concern the training of a simple perceptron in a student teacher scenario. Originally the models were treated in [22] and they have been revisited in several reviews, for instance, [20,21]. We reproduce the results here as particularly illustrative examples for the statistical physics approach to learning.

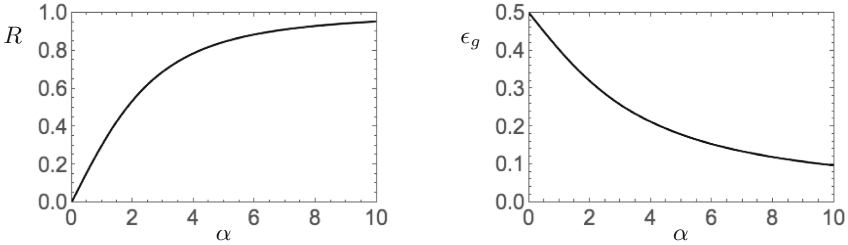

Fig. 2. Typical learning curves of the perceptron with continuous weights in the student teacher scenario, see Sect. 2.6. The left panel shows $R(\alpha)$, the right panel displays the corresponding generalization error $\epsilon_g(\alpha)$.

The Perceptron with Continuous Weights

Here we consider a student teacher scenario where the student weight vector $\mathbf{w} \in \mathbb{R}^N$ is normalized ($\mathbf{w}^2 = N$) but otherwise unrestricted.

The generalization error as a function of the student teacher overlap R is given in Eq. (8). The corresponding entropy, Eq. (16), can be obtained by means of a saddle point integration. Alternatively, one can interpret $e^N s$ as the volume of an $(N-1)$-dimensional hypersphere in weight space with radius $\sqrt{1-R^2}$, see [21] for the geometrical argument. One obtains

$$s(R) = \frac{1}{2}\ln(1-R^2) + const., \tag{19}$$

where the additive constant does not depend on R. Apart from such irrelevant terms, we obtain the quenched free energy in the limit $\beta \longrightarrow 0$ as

$$\beta f \; = \; \alpha \, \frac{1}{\pi} \arccos R - \frac{1}{2}\ln(1-R^2). \tag{20}$$

In absence of training data, $\alpha = 0$, the maximum of the entropy term in $R = 0$ governs the behavior of the system. In the high-dimensional feature space, the student weight vector is expected to be orthogonal to the unknown $\mathbf{w}^*$.

The free energy is displayed in Fig. 1 for three different values of α. As α is increased, we observe that the minimum of βf is found in larger, positive values of R, reflecting the knowledge about the rule as inferred from the set of examples.

The student teacher overlap $R(\alpha)$ that corresponds to the minimum of βf is displayed in Fig. 2 (left panel). In this simple case, it can be obtained analytically from the necessary condition for the presence of a minimum:

$$\frac{\partial \beta f}{\partial R} = 0 \quad \Rightarrow \quad R(\alpha) = \frac{\alpha}{\sqrt{\alpha^2 + \pi^2}}. \tag{21}$$

By means of Eq. (8) this result translates into a learning curve $\epsilon_g(\alpha)$, which is shown in the right panel of Fig. 2. One can show that large training sets facilitate perfect generalization with

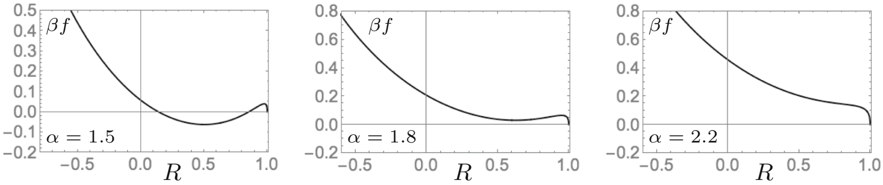

Fig. 3. The quenched free energy βf as a function of the order parameter R in the training scenario with Ising student and teacher perceptron, cf. Sect. 2.6. In the leftmost panel the rescaled numbers of examples is $\alpha = 1.5 < \alpha_c$, where $R = 1$ constitutes a local minimum while a state with $0 < R < 1$ is thermodynamically stable. In the center panel with $\alpha = 1.8 > \alpha_c$, here perfect generalization with $R = 1$ corresponds to the global minimum. The rightmost panel displays βf for $\alpha = 2.2 > \alpha_d$ where $R = 1$ constitutes its only minimum.

$$R(\alpha) \approx 1 - \frac{\pi^2}{2\alpha^2} \quad \text{and} \quad \epsilon_g(\alpha) \approx \frac{1}{\alpha} \quad \text{for} \quad \alpha \to \infty. \tag{22}$$

It is interesting to note that the basic asymptotic α-dependences are recovered in the more sophisticated application of the annealed approximation or the replica formalism [22]. Obviously, an explicit temperature dependence and the correct prefactors cannot be obtained in the simplifying limit.

The Perceptron with Discrete Weights

As an interesting exercise we also revisit the model with discrete student weights [22]. The term Ising perceptron has been coined for the model with weights $\mathbf{w} \in \{-1, 1\}^N$ [21,22]. Note that the assumed normalization $\mathbf{w}^2 = N$ is trivially satisfied. Moreover, the generalization error is also given by Eq. (8) since its derivation does not depend on details of the weight space.

The corresponding entropy can be obtained by a simple counting argument: In order to obtain an overlap $\sum_j w_j w_j^* = NR$, a number of $N(R+1)/2$ components must satisfy $w_j = w_j^*$ while for $N(R-1)/2$ we have $w_j = -w_j^*$. The associated entropy of mixing is given by the familiar form

$$s(R) = -\left(\frac{1+R}{2}\right) \ln\left(\frac{1+R}{2}\right) - \left(\frac{1-R}{2}\right) \ln\left(\frac{1-R}{2}\right). \tag{23}$$

The resulting free energy (18) as a function of R is displayed in Fig. 3 for three different values of α.

For all $\alpha > 0$, βf displays a local minimum in $R = 1$ with $f(R = 1) = 0$. For small α, however, a deeper minimum can be found with an overlap $0 < R < 1$. This is exemplified for $\alpha = 1.5$ in the leftmost panel of Fig. 3. The global mininum of βf determines the thermodynamically stable state of the system.

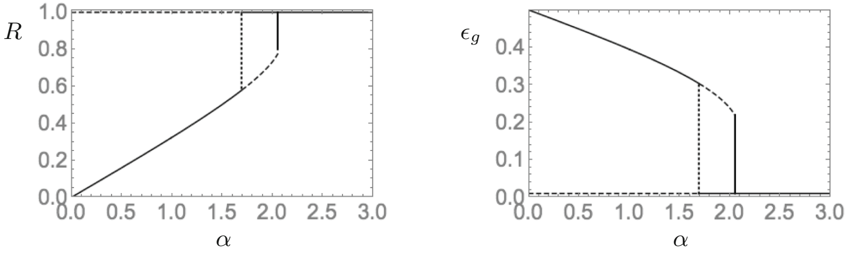

Fig. 4. Learning of the Ising perceptron with discrete weights in the student teacher scenario, see Sect. 2.6. The left panel shows $R(\alpha)$, the right panel displays the corresponding generalization error $\epsilon_g(\alpha)$. States corresponding to local minima of βf are marked by dashed lines, while solid lines mark the thermodynamically stable global minima. Vertical dotted and solid lines correspond to the critical $\alpha_c \approx 1.69$ and $\alpha_d \approx 2.08$, respectively.

For training sets with α larger than a critical value $\alpha_c \approx 1.69$, the state with $R = 1$ constitutes the global minimum. A competing configuration with $R < 1$ persists as a local minimum, but becomes unstable for $\alpha > \alpha_d \approx 2.08$, see the center and rightmost panel of Fig. 3.

The learning curves $R(\alpha)$ and $\epsilon_g(\alpha)$ reflect the specific α-dependence of βf in terms of a discontinuous phase transition. In Fig. 4, the solid lines mark the thermodynamically stable state in terms of $R(\alpha)$ (left panel) and $\epsilon_g(\alpha)$ (right panel). Dashed lines correspond to local minima of βf and the characteristic values α_c and α_d are marked by the dotted and solid vertical lines, respectively.

The essential findings of the high temperature treatment do carry over to the training at lower formal temperatures, qualitatively [21,22]. Most notably, the system displays a freezing transition to perfect generalization. Furthermore, the first order phase transition scenario will have non-trivial effects in practical training. The existence of metastable, poorly generalizing states can delay the success of training significantly. Related hysteresis effects with varying α have been observed in Monte Carlo simulations of the training process, see [21] and references therein.

3 Summary and Conclusion

This brief review merely discusses one goal of the statistical physics of learning: the computation of typical learning curves in clear-cut model scenarios. This type of results provide basic insight into relevant mechanisms and phenomena which play a role in practical machine learning setups as well. The framework provides a workshop in which to analyse, put forward and optimize training algorithms. Moreover it offers the possibility to systematically compare different adaptive systems, network architectures etc.

The classical examples discussed in this short tutorial concern merely the simplest models, i.e. the learning of a linearly separable rule with a percetpron network. The presentation is furthermore restricted to the particularly simplifying limit of training at high temperature.

In the literature, numerous studies of more complex adaptive systems, such as layered neural networks or support vector machines can be found. Similarly, models of unsupervised learning and related problems of data analysis and inference have been analysed. Among the many interesting extensions, we mention only the study of symmetry breaking phase transitions in feedforward layered neural networks.

The analysis of more realistic training at low formal temperatures requires a much more involved mathematical treatment. A thorough discussion thereof would be clearly beyond the scope of this brief introduction to the field. Indeed, the theory of learning has had a very fruitful impact on the development and understanding of sophisticated methods for the analysis of disordered systems in general.

Apart from the equilibrium approach discussed here, statistical physics also provides the tools to analyse non-equilibrium situations. This has helped to study the dynamics of learning in a very similar fashion. The resulting insights directly link to popular practical training prescriptions such as the popular stochastic gradient descent.

Recently, the statistical physics of learning is being rediscovered and has gained popularity again in the context of deep learning. A better theoretical understanding of this successful machine learning framework is highly desirable. Currently, many researchers revisit the statistical physics perspective to learning, aiming a fundamental insights into design and performance of, for instance, deep layered networks. A brief discussion of recent developments, challenges and open questions, as well as further references can be found in [32].

The author is convinced that the revival of the area will contribute significantly to the further development of machine learning and data analysis in general.

References

1. Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison-Wesley (1991)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. SSS, Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
3. Bishop, C.: Pattern Recognition and Machine Learning. Cambridge University Press, Cambridge (2007)
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
5. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
6. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
7. Saitta, L., Giordana, A., Cornuéjols, A.: Phase Transitions in Machine Learning, 383 p. Cambridge University Press (2011)
8. Rynkiewicz, J.: Asymptotic statistics for multilayer perceptrons with ReLU hidden units. In: Verleysen, M. (ed.) Proceedings European Symposium on Artificial Neural Networks (ESANN), 6 p. d-side publishing (2018)

9. Marcus, G.: Deep learning: a critical appraisal. http://arxiv.org/abs/1801.00631. Accessed 23 Apr 2018
10. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: Proceedings of the 6th International Conference on Learning Representations ICLR (2017)
11. Martin, C.H., Mahoney, M.W.: Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. Computing Research Repository CoRR, eprint 1710.09553 (2017).http://arxiv.org/abs/1710.09553
12. Lin, H.W., Tegmark, M., Rolnick, D.: Why does deep and cheap learning work so well? J. Stat. Phys. **168**(6), 1223–1247 (2017)
13. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P.: Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. **11**, 625–660 (2010)
14. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087 (1953)
15. Mezard, M., Parisi, G., Virasoro, M.: Spin Glass Theory and Beyond. World Scientific (1986)
16. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Nat. Acad. Sci. USA **79**(8), 2554–2558 (1982)
17. Amit, D.J., Gutfreund, H., Sompolinsky, H.: Storing infinite numbers of patterns in a spin-glass model of neural networks. Phys. Rev. Lett. **55**(14), 1530–1533 (1985)
18. Gardner, E.: Maximum storage capacity in neural networks. Europhys. Lett. **4**(4), 481–486 (1988)
19. Gardner, E.: The space of interactions in neural network models. J. Phys. A Math. General **21**(1), 257–270 (1988)
20. Engel, A., Van den Broeck, C.: Statistical Mechanics of Learning, 342 p. Cambridge University Press (2001)
21. Watkin, T.L.H., Rau, A., Biehl, M.: The statistical mechanics of learning a rule. Rev. Mod. Phys. **65**(2), 499–556 (1993)
22. Seung, H.S., Sompolinsky, H., Tishby, N.: Statistical mechanics of learning from examples. Phys. Rev. A **45**, 6065–6091 (1992)
23. Kinzel, W.: Phase transitions of neural networks. Philos. Mag. B **77**(5), 1455–1477 (1998)
24. Opper, M.: Learning and generalization in a two-layer neural network: the role of the Vapnik-Chervonenkis dimension. Phys. Rev. Lett. **72**, 2113 (1994)
25. Herschkowitz, D., Opper, M.: Retarded learning: rigorous results from statistical mechanics. Phys. Rev. Lett. **86**, 2174 (2001)
26. Goldt, S., Mézard, M., Krzakala, F., Zdeborová, L.: Modelling the influence of data structure on learning in neural networks. eprint 1909.11500v1[stat.ML] (2019). http://arxiv.org/abs/1909.11500
27. Cocco, S., Monasson, R., Posani, L., Rosay, S., Tubiana, J.: Statistical physics and representations in real and artificial neural networks. Phys. A Stat. Mech. Its Appl. **504**, 45–76 (2018)
28. Kadmon, J., Sompolinsky, H.: Optimal architectures in a solvable model of deep networks. In: Lee, D.D. Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NIPS 29), pp. 4781–4789. Curran Associates Inc. (2016)

29. Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems (NIPS 27), pp. 2933–2941. Curran Associates Inc. (2014)
30. Pankaj, M., Lang, A.H., Schwab, D.: An exact mapping from the variational renormalization group to deep learning. arXiv repository [stat.ML], eprint 1410.3831v1 (2014). https://arxiv.org/abs/1410.3831v1
31. Sohl-Dickstein, J., et al.: Deep unsupervised learning using non-equilibrium thermodynamics. Proc. Mach. Learn. Res. **37**, 2256–2265 (2016)
32. Biehl, M., Caticha, N., Opper, M., Villmann, T.: Statistical physics of learning and inference. In: Verleysen, M. (ed.) Proceedings European Symposium on Artificial Neural Networks (ESANN), pp. 501–509. d-side Publishing (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.