BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF MECHANICAL ENGINEERING

FAKULTA STROJNÍHO INŽENÝRSTVÍ

INSTITUTE OF PHYSICAL ENGINEERING

ÚSTAV FYZIKÁLNÍHO INŽENÝRSTVÍ

INTERCONNECTION OF RESTRICTED BOLTZMANN MACHINE METHOD WITH STATISTICAL PHYSICS AND ITS IMPLEMENTATION IN THE PROCESSING OF SPECTROSCOPIC DATA

POPIS RESTRICTED BOLTZMANN MACHINE METODY VE VZTAHU SE STATISTICKOU FYZIKOU A JEHO NÁSLEDNÉ VYUŽITÍ VE ZPRACOVÁNÍ SPEKTROSKOPICKÝCH DAT

MASTER'S THESIS

DIPLOMOVÁ PRÁCE

AUTHOR | Bc. Jakub Vrábel
AUTOR PRÁCE

SUPERVISOR | Ing. Pavel Pořízka, Ph.D.
VEDOUCÍ PRÁCE

BRNO 2019

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav fyzikálního inženýrství
Student:	**Bc. Jakub Vrábel**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Fyzikální inženýrství a nanotechnologie
Vedoucí práce:	**Ing. Pavel Pořízka, Ph.D.**
Akademický rok:	2018/19

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Popis Restricted Boltzmann machine metody ve vztahu se statistickou fyzikou a jeho následné využití ve zpracování spektroskopických dat

Stručná charakteristika problematiky úkolu:

Algoritmy strojového učení (především tzv. Hluboké učení) jsou bytostně prostoupeny fyzikálními zakonitostmi. Fyzika je v tomto ohledu využita nejen jako inspirace ke konstrukci, ale i jako nástroj k pochopení funkce daných modelů učení. Restricted Boltzmann Machines (RBM) je jedním z algoritmů umělých neuronových sítí bez učitele. Z historického vývoje, v roce 2006 Hinton (Hinton 2006) významně upozornil na neuronové sítě zavedením hlubokého učení. Aplikace RBM sahají od redukce dimenze přes klasifikaci až po feature learning, atd.

Zpracování spektroskopických dat je předmětem výzkumu na rozmezí fyziky, matematiky a statistiky. Vícerozměrná data, naměřená např. využitím techniky Spektroskopie laserem buzeného plazmatu (Laser–Induced Breakdown Spectroscopy, LIBS), představují, co do objemu a komplexní struktury, náročný úkol pro klasické lineární algoritmy. Z toho důvodu je využití neuronových sítí hojně prosazováno i v oblasti spektroskopie. Optimalizace učícího algoritmu založeného na RBM v aplikaci laserové spektroskopie je pak unikátním a doposud neprozkoumaným řešením.

Cíle diplomové práce:

1. Rešerše methods vycházejících ze statistické fyziky a jejich podobnost s Hlubokým učením, především RBM.
2. Přímé využití RBM pro redukci dimenze spektroskopických dat (emisní spektra z LIBS).
3. Srovnání schopností RBM a běžně používaných metod (PCA, …).

Seznam doporučené literatury:

HINTON, G. E. Reducing the Dimensionality of Data with Neural Networks. Science. 313(5786), 504-507 (2006).

MIZIOLEK, A. W., PALLESCHI, V. a SCHECHTER, I. Laser Induced Breakdown Spectroscopy. Cambridge: Cambridge University Press (2006).

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2018/19

V Brně, dne

L. S.

prof. RNDr. Tomáš Šikola, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

ABSTRACT

In this work, connections between statistical physics and machine learning are studied with emphasis on the most basic principles and their implications. Also, the general properties of spectroscopic data are revealed and used beneficially for improving automatized processing of the data. In the beginning, the partition function of a Boltzmann distribution is derived and used to study the Ising model utilizing the mean field theory approach. Later, the equivalence between the Ising model and the Hopfield network (machine learning model) is shown, along with an introduction for machine learning in general. At the end of a theoretical part, Restricted Boltzmann Machine (RBM) is obtained from the Hopfield network. Suitability of applying RBM to the processing of spectroscopic data is discussed and revealed by utilization of RBM to dimension reduction of the data. Results are compared to the standard tool (Principal Component Analysis), with discussing possible further improvements.

KEYWORDS

Machine Learning, LIBS, Spectroscopic Data, Artificial Neural Networks, Deep Learning, Restricted Boltzmann Machine, RBM, Dimension Reduction, Statistical Physics.

ABSTRAKT

Práca sa zaoberá spojeniami medzi štatistickou fyzikou a strojovým učením s dôrazom na základné princípy a ich dôsledky. Ďalej sa venuje obecným vlastnostiam spektroskopických dát a ich zohľadnení pri pokročilom spracovaní dát. Začiatok práce je venovaný odvodeniu partičnej sumy štatistického systému a štúdiu Isingovho modelu pomocou "mean field" prístupu. Následne, popri základnom úvode do strojového učenia, je ukázaná ekvivalencia medzi Isingovým modelom a Hopfieldovou sieťou - modelom strojového učenia. Na konci teoretickej časti je z Hopfieldovej siete odvodený model Restricted Boltzmann Machine (RBM). Vhodnosť použitia RBM na spracovanie spektroskopických dát je diskutovaná a preukázaná na znížení dimenzie týchto dát. Výsledky sú porovnané s bežne používanou Metódou Hlavných Komponent (PCA), spolu so zhodnotením prístupu a možnosťami ďalšieho zlepšovania.

KLÍČOVÁ SLOVA

Strojové učenie, LIBS, spektroskopické dáta, Neuronové siete, hlboké učenie, RBM, redukcia dimenzie.

VRÁBEL, Jakub *Interconnection of Restricted Boltzmann machine method with statistical physics and its implementation in the processing of spectroscopic data*: master's thesis. Brno: Brno University of Technology, Faculty of Mechanical Engineering, Institute of Physical Engineering, 2019. 85 p. Supervised by Ing. Pavel Pořízka, Ph.D.

DECLARATION

I declare that I have elaborated my master's thesis on the theme of "Interconnection of Restricted Boltzmann machine method with statistical physics and its implementation in the processing of spectroscopic data" independently, under the supervision of the master's thesis supervisor and with the use of technical literature and other sources of information which are all quoted in the thesis and detailed in the list of literature at the end of the thesis.

As the author of the master's thesis I furthermore declare that, concerning the creation of this master's thesis, master's thesis, I have not infringed any copyright. In particular, I have not unlawfully encroached on anyone's personal copyright and I am fully aware of the consequences in the case of breaking Regulation § 11 and the following of the Copyright Act No 121/2000 Vol., including the possible consequences of criminal law resulted from Regulation § 152 of Criminal Act No 140/1961 Vol.

Brno ....

(author's signature)

Acknowledgement

In the first place, I would like to thank my parents and family for support during my studies. Also, I want to thank my supervisor Dr. Pavel Pořízka, for suggestions to the thesis and introducing me to the scientific work. Furthermore, I am thankful to colleagues (DP, EK, JK) from the group for fruitful discussions and cooperation on various projects.

Bc. Jakub Vrábel

CONTENTS

INTRODUCTION

During the last years, Machine Learning (ML) obtained exponential growth of its popularity. Everything begins with silicon boom a few decades ago when computers were created. Hand to hand with computers, sensing of various events (experiments, industrial measurements, weather, etc.) became much easier and digitalized. Nowadays, nearly everything is sensed and measurements are stored. This results in a huge amount of data, which has to be analyzed. Modern data are rather different from data which were assumed by statisticians during the creation of classical statistics. While in classical statistics, there is a relatively small amount of measurements with only one or few variables. Now we have data containing many more measurements and number of variables can be comparable to the number of samples. These high-dimensional data require completely new methodologies for analysis.

Machine learning is a big family of methods, while part of them are suitable for this usage. Machine learning is generally understood as a study of mathematical models or computational algorithms which are built (learned) on available (training) data using computers. Models could be used for regression, classification, dimensionality reduction and other specific tasks. The diagram on Figure 1 shows ML as part of Artificial Intelligence (AI) and further divides ML models to supervised and unsupervised. According to the state of art research in ML field, one of the most promising models is Neural Networks (NN) and especially Deep Neural Networks (DNN). DNN has achieved numerous records in challenging tasks (image classification, text classification, speech recognition, classification of scientific data, game playing, etc.) competing to other methods. Further description of NN and DNN will be presented later. As it was stated, ML is a different approach to classical statistical analysis. However, Its important part holds on classical statistics. This strong connection is more visible from the Bayesian perspective. Alongside with classical statistics, there is also a strong connection of the ML with statistical physics. ML takes not only inspiration from statistical physics approaches, but many of ML methods has its roots directly in statistical physics.

Following subsections serves just as a brief illustration of this connection and every important part will be studied more detailed later.

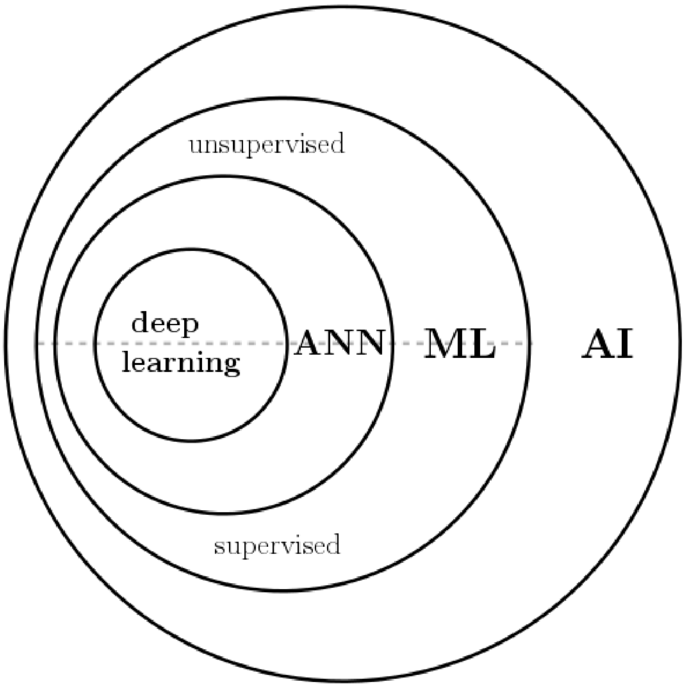

Fig. 1: Venn diagram with Machine Learning relations.

Why we want to apply stat. phys point of view to ML?

There is a great amount of knowledge as a result of intensive research in statistical physics since the time of Boltzmann and Gibbs to the modern era and its numerous applied sub-parts. Statistical physics, in a nutshell, is just a study of collective behavior of the system composed of many parts. This collective behavior appears only while there is a huge number of parts present and is mostly quite different from the behavior of its single or few parts. Under the concept of the system, we can understand many things as a collection of atoms (for example in solid material), collection of spins (in Ising model), etc. but also more non-physical or abstract terms as information theoretical bits, neurons (in Neural Network). This abstraction of statistical physics is building a bridge between communities and making possible to better understand why and how ML algorithms work.

How we can use ML models in the processing of spectroscopic data?

Spectroscopy in its general meaning is a collection of methods studying the interaction of electromagnetic radiation with matter [1]. Result of such spectroscopic measurement is a spectrum where we scope dependence of intensity (or the number of photons/counts on the detector) on a specific wavelength. Despite considerable differences in each spectroscopic methods, there are a common features in spectra from the data point of view. Spectroscopic data have usually high dimension (number of variables = wavelength resolution) and sparsity. Sparsity means that there is a relatively small amount of information with respect to insignificant parts of the spectra (consisting of background and noise). Sparsity is nearly always originated of high dimensionality, but spectroscopic data are special also due to other property. Peaks which are present in spectra are correlated by itself in the data sense. This means that the surrounding values of the peak are dependent on its central value. Using these properties, It is obvious that dimensionality reduction can be performed with only small losses in information, theoretically. In spectroscopy, we often would like to separate our measurements according to the material of samples, for example. If we provide labeled (chemical elements, molecules, materials) dataset of spectra, we can build a model for further classification of unlabeled data. This type of analysis is called supervised classification and it is easy to imagine a lot of applications. Machine learning methods suitable for this task are for example Neural Networks, Support Vector Machines, and many more. While technical possibilities are increasing and spectroscopic analysis became much faster (kHz frequency of measurements) we are obtaining millions of spectra. This amount of data is not possible to explore by looking at each spectrum separately. Also storing and handling such datasets is challenging. So other important applications of ML in spectroscopy are revealing: dimensionality reduction and visualization of high dimensional data.

1 STATISTICAL PHYSICS

This section serves only as a summary of essential principles and ideas of statistical physics, but cannot provide a proper introduction to the subject. While some parts here are exactly and hierarchically derived from the first principles, others may be ripped out of context and serve more like a definition. Unfortunately, the range of this thesis cannot cover the whole subject and for further information, reader is advised to great courses [2–4]. As foundations of statistical physics largely rely on principles and terminology of theoretical mechanics, It can be also useful to review basics from Landau's course on the topic [5]. Approach presented in this chapter is based on [6] and [4].

To begin explanation of what statistical physics is, we should start from historical motivation. The main goal of statistical physics is the study of collective behavior of systems containing a large number of particles. Even while, laws of statistical physics were historically developed for classical mechanics, they can be generalized and hold even for quantum systems. To describe the behavior of a mechanical system completely, we have to solve the corresponding number of equations of motion, depending on degrees of freedom. Dealing with common problems in macroscopic bodies, we are facing a huge number of particles (e.g. 10^{22} atoms 1 cm cube of Cu lattice). We can easily conclude impossibility of solving comparable number of equations and also specifying initial conditions for every particle. A remarkable fact is that in contrary to intuition taken from previous consideration, with a rising number of particles the complexity of system properties is not increasing. However, we observe completely new behavior of the system, arisen from a very high number of its particles. This new collective behavior cannot be explained in purely mechanical terms but can be treated by statistical physics approach. [4]

In statistical physics, objects of our interest are macroscopic *systems* consisting of a large number of various particles. Generally, there is no restriction on the type of these particles. Most common are of course atoms, spins, bits and many more. We can describe the system using *phase space*, a concept from classical mechanics, where one uses n generalized coordinates q_i and corresponding velocities $\dot{q}_i$, where index i is representing degrees of freedom. In practice for building theory, it is more convenient to use momenta p_i than velocity, at least because of conservation law. State of the system is described by a point in $2ni$ dimensional phase space and evolution of the system is described by its phase trajectory (line in phase space). A small part of system w.r.t. itself, but still possibly macroscopic, we call *subsystem.* If we consider the whole system as closed, i.e. it cannot interact with other systems, the subsystem is not closed. According to an application, the subsystem can be exchanging energy or even particles with the rest of the system through various complicated interactions. Let's imagine subsystem as a small region of phase space $\Delta\vec{p}\Delta\vec{q}$ called *phase volume*. During the time, the system is evolving and the phase trajectory will pass through this region many times. We define the probability of finding subsystem in this phase volume and corresponding state as

$$w = \lim_{T\to\infty} \frac{\Delta t}{T}, \tag{1.1}$$

where Δt is time section of the subsystem being in mentioned phase space region and T is total time. Infinitesimal element of phase volume $\mathrm{d}\phi$ is defined as

$$\mathrm{d}\phi = \mathrm{d}\vec{q}\, \mathrm{d}\vec{p} = \mathrm{d}q_1\, \mathrm{d}q_2 ... \mathrm{d}q_i\, \mathrm{d}p_1\, \mathrm{d}p_2 ... \mathrm{d}p_i = \mathrm{d}q_i\, \mathrm{d}p_i, \tag{1.2}$$

where last equality is fulfilled by using Einstein's sum convention. Finally, we can define probability dw as

$$\mathrm{d}w = \rho(p_i, q_i)\, \mathrm{d}q_i\, \mathrm{d}p_i, \tag{1.3}$$

where $\rho(p_i, q_i)$ is *probability density* of probability distribution. It is representing density number of states inside phase space element $\mathrm{d}w$. According to general requirements on distribution function, the normalization condition

$$\int \rho\, \mathrm{d}q_i\, \mathrm{d}p_i = 1, \tag{1.4}$$

has to be fulfilled. Integration is taken all over the phase space.

For calculating average values of dynamical variables, we usually scope system for during long period and obtain time average. However, J. W. Gibbs brought a neat solution in which time averages are replaced by so-called *ensemble averages*, also known as *thermal averages*. The ensemble of a system is a virtual group of many identical systems. The number of virtual systems in ensemble is selected with respect to the number of accessible states of such physical system. All systems in the ensemble are equivalent and hold for conditions required by the original system. In this construction, we assume that averages taken over the ensemble can correctly substitute time averages of a single (original) system after a sufficiently long time. Justification of this equality is a point of interest in *ergodic theory*. While equality holds for many common cases, it was not yet proven for all mechanical systems.

$$\langle A \rangle \overset{?}{=} \bar{A} \tag{1.5}$$

Equation 1.5 is representing this ergodic problem, where LHS is ensemble average and RHS time average defined as

$$\langle A \rangle \equiv \int A(q_i, p_i) \rho(p_i, q_i)\, \mathrm{d}\phi, \tag{1.6}$$

$$\bar{A} \equiv \lim_{\tau\to\infty} \frac{1}{\tau} \int\limits_{t_0}^{t_0+\tau} f(q_i(t)p_i(t))\,\mathrm{d}t = \lim_{\tau\to\infty} \frac{1}{\tau} \int\limits_{0}^{\tau} f(q_i(t)p_i(t))\,\mathrm{d}t. \tag{1.7}$$

There is an important assumption in time averaging, that statistical distribution of the dynamic variable is independent of the initial state. This is satisfied after a sufficiently long time passed.

In contrary to the probabilistic nature of systems consisting of many particles, we know experimentally that dynamical variables of macroscopic bodies in stationary state are measured practically constant with only small fluctuations. This follows from the shape of probability density function ρ which have a very sharp peak at average value. The peak is sharper the larger macroscopic body we study. This basic principle will be explained in further sections. The state of a system where its physical quantities can be described by their mean values is called *statistical equilibrium*. If we interact with the system in time t_0 and change its state, it will relax to equilibrium. Dynamics of these transitions to equilibrium state is outside of scope of this thesis.

Before moving forward, we have to note special case of statistical independence between subsystems. Subsystems were defined as smaller parts of the system, which are generally not closed. If we consider only short periods of time and also fact that subsystem can still form a macroscopic body, we can think of them as "quasi-closed"or weakly interacting with surrounding space. [4] In this case of short time periods, subsystems are considered as statistically independent. That is, the state of one subsystem is independent on the state of other and its probabilities. From probability theory and statistical independence assumption we get

$$\rho_{12}\,\mathrm{d}q^{(1,2)}\,\mathrm{d}p^{(1,2)} = \rho_1\,\mathrm{d}q^{(1)}\,\mathrm{d}p^{(1)}\rho_2\,\mathrm{d}q^{(2)}\,\mathrm{d}p^{(2)}, \tag{1.8}$$

$$\rho_{12} = \rho_1\rho_2. \tag{1.9}$$

So probability density of combined system (from 2 subsystems) is a direct product of subsystems prob. density.

1.1 Liouville's theorem

Equation of the continuity may be written in form

$$\frac{\partial\rho}{\partial t} + \mathrm{div}(\rho\vec{v}) = 0, \tag{1.10}$$

where divergence can be generalized for more dimensions as

$$\sum_{i=1}^{2s} \frac{\partial(\rho v_i)}{\partial x_i} = 0. \tag{1.11}$$

Upper bound of summation is $2s$ because of using full phase space and then x_i are generalized "coordinates" q and p and corresponding generalized velocities v_i are $\dot{q}$ and $\dot{p}$. So for our case, we obtain

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{s} \left[\frac{\partial(\rho \dot{q}_i)}{\partial q_i} + \frac{\partial(\rho \dot{p}_i)}{\partial p_i} \right] = 0, \tag{1.12}$$

where we carry out derivations and get

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{s} \left[\dot{q}_i \frac{\partial \rho}{\partial q_i} + \rho \frac{\partial \dot{q}_i}{\partial q_i} + \dot{p}_i \frac{\partial \rho}{\partial p_i} + \rho \frac{\partial \dot{p}_i}{\partial p_i} \right] = 0. \tag{1.13}$$

Now, we are using Hamilton canonical equations to obtain

$$\dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i}; \qquad \dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i} \tag{1.14}$$

where $\mathcal{H}$ is the Hamiltonian of the subsystem. Due to the interchangeability of second partial derivative, it is easy to find that

$$\frac{\partial \dot{q}_i}{\partial q_i} = \frac{\partial^2 \mathcal{H}}{\partial p_i \partial q_i} = -\frac{\partial \dot{p}_i}{\partial p_i}. \tag{1.15}$$

After this substitution, the terms with ρ in equation 1.13 are canceling each other and we get

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i = 0 \tag{1.16}$$

We know that $\rho = \rho(t, q_i, p_i)$ so Equation 1.16 is clearly total differential of ρ. Finally, we have

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = 0, \tag{1.17}$$

which is called *Liouville's theorem*. It says that probability density is constant and the probability of systems is behaving as incompressible fluid and has great importance in statistical mechanics as we will see in the following section.

1.2 Energy

From Liouville's theorem follows that probability density ρ is constant with time. So in case of the closed system, it can only be a function of variables, which are constant in time. In classical mechanics, we call these variables *integrals of motion*.

If we consider previously derived equality for probability density function $\rho_{12} = \rho_1\rho_2$, is has to be fulfilled also for its logarithm

$$\log \rho_{12} = \log \rho_1 + \log \rho_2. \tag{1.18}$$

Therefore, it is obvious that the logarithm of probability density is an additive quantity. Then logarithm of ρ can be a function, depending only on additive integrals of motion. Fortunately, there are exactly seven additive integrals of motion (energy, momentum and angular momentum):

$$\rho = \rho(E, \vec{p}, \vec{L}). \tag{1.19}$$

Even more, we can always select reference frame moving and rotating with the system center of gravity. So finally, we have left only one quantity, probability density is depending on - the energy:

$$\rho = \rho(E). \tag{1.20}$$

This observation has far-reaching consequences and reveals the unique role of energy in statistics. It is possible to describe an equilibrium state of the macroscopic body, consisting of a very large number of degrees of freedom, with just its energy. We will use this principle even in machine learning part of this work, which on first sight may look completely different from presented material from physics.

1.2.1 Density of states

With previous observations about the constant value of phase volume and only energy dependence of probability density, we may conclude that also phase volume is depending only on the energy of system $\phi = \phi(E)$. Let's define dimensionless phase volume as

$$\mathrm{d}\Gamma \equiv \frac{\mathrm{d}\phi}{(2\pi\hbar)^s}. \tag{1.21}$$

Normalizing coefficient $(2\pi\hbar)^s$ represents the size of one state in s dimensions. However, we put it here just as a definition, its shape can be exactly derived from quantum mechanical considerations. Finally, *density of states* γ is

$$\gamma(E) \equiv \frac{\mathrm{d}\Gamma}{\mathrm{d}E}. \tag{1.22}$$

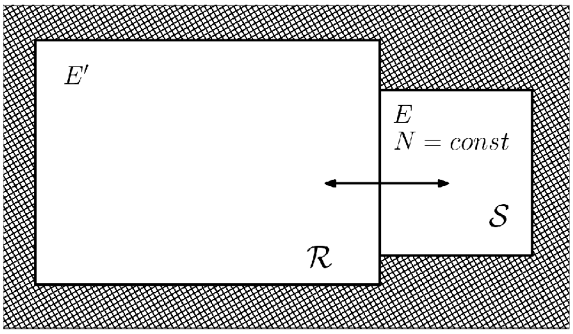

Fig. 1.1: Gibbs canonical ensemble.

1.2.2 Gibbs canonical ensemble

Let's define a system $\mathcal{S}$ consisting of N particles and energy E which is allowed to interact (exchange energy) with its surroundings (reservoir) $\mathcal{R}$. Such a system is schematically shown in Figure 1.1. Exchange of particles is forbidden, so $N = const = N_0$. As a simplification, we do not take into account the surface effects of the system, that means we consider the energy of system surface small w.r.t. the total energy of the system: $E_{\partial V} \ll E$. From the first law of thermodynamics [7] we have

$$\mathrm{d}U = \delta Q - \delta A + \delta U_N, \tag{1.23}$$

where $\mathrm{d}U$ is internal energy of a system ($\mathrm{d}U$ is total differential), δQ heat supplied to the system, δA net work done by the system and δU_N internal energy of particle exchange. In case of heat differential form δQ, it is possible to find integration factor to create total differential. This integration factor is reciprocal temperature and created total differential is *entropy* $\mathrm{d}S$ defined as

$$\mathrm{d}S = \frac{1}{T}\delta Q. \tag{1.24}$$

For our purpose, we may restrict net work done by the system just to mechanical work

$$\delta A = F\,\mathrm{d}l = pS'\,\mathrm{d}l = p\,\mathrm{d}V, \tag{1.25}$$

here F is a force, $\mathrm{d}l$ is infinitesimal length element in direction of applied force, p is pressure, S' is surface and V is volume taken by system particles. If we have the system consisting only of one type particles, particle exchange internal energy is

$$\delta U_n = \mu\delta N \tag{1.26}$$

μ is called chemical potential. However, μ is not potential from a mathematical point of view, so it is just a factor of energy change. N is again the number of particles.

In previously considered system (Gibbs canonical ensemble), there is not allowed exchange of particles with reservoir, therefore we may write first law of thermodynamics as

$$\mathrm{d}U = T\,\mathrm{d}S - p\,\mathrm{d}V \tag{1.27}$$

According to previous observations, probability of finding system and its surroundings in the state of specific energy follows from energy additivity and phase space multiplicativity as

$$\mathrm{d}w_{\text{total}} = \rho(E_{\text{total}})\,\mathrm{d}\Gamma_{\text{total}} = \rho(E + E')\,\mathrm{d}\Gamma\,\mathrm{d}\Gamma', \tag{1.28}$$

but from Equation 1.9 we know that $\rho(E+E') = \rho(E)\rho(E')$. Then there is only one possibility of function dependance fulfilling all these conditions - the exponential function:

$$\rho(E) = e^{\alpha-\beta E}. \tag{1.29}$$

Constants α and β will be found from thermodynamics (limit case of statistical physics). The average value of energy is determined by

$$U = \int E\,\mathrm{d}w = \int E\rho(E)\,\mathrm{d}\Gamma = \int E\rho(E)\gamma(E)\,\mathrm{d}E. \tag{1.30}$$

If we compute differential for this equation, we may compare it to the first law of thermodynamics (Equation 1.27). In this case, the only external parameter for each system in the ensemble is volume V. Thus we may write

$$\mathrm{d}U = \int \frac{\partial E}{\partial V}\,\mathrm{d}(V)\rho\,\mathrm{d}\Gamma + \int E\,\mathrm{d}(\rho)\,\mathrm{d}\Gamma, \tag{1.31}$$

where brackets after differential symbol d() are used to distinguish between total differential made by differentiation, from integration variables. Using supplementary relation for work dW done by force F acting on infinitesimal length element dl, pressure p, area A and volume V:

$$\mathrm{d}W = F\,\mathrm{d}l = pA\,\mathrm{d}l = p\,\mathrm{d}V, \tag{1.32}$$

the first term of equation 1.31 may be rewritten to the form: $-\int p\,\mathrm{d}V\rho\,\mathrm{d}\Gamma$. For second term, we use trick of rewriting energy E using expected shape for probability density ρ (Eq. 1.29):

$$\rho(E) = e^{\alpha-\beta E} \Longrightarrow E = \frac{\alpha}{\beta} - \frac{1}{\beta}\log\rho. \tag{1.33}$$

Putting all together, we have

$$\mathrm{d}U = -\int p\,\mathrm{d}(V)\rho\,\mathrm{d}\Gamma + \frac{\alpha}{\beta}\int \mathrm{d}(\rho)\,\mathrm{d}\Gamma - \frac{1}{\beta}\int \mathrm{d}(\rho)\log\rho\,\mathrm{d}\Gamma. \tag{1.34}$$

Clearly in the first term we have formula for average value of the pressure, which is equal to the "macroscopic"pressure. In second term, order of integration and differentiation may be reversed. While integral of probability density over phase space is equal to 1, its differential has to be zero. For the last (third) term we use well known "per partes"formula $f\,\mathrm{d}g = \mathrm{d}(fg) - g\,\mathrm{d}f$. Rearranged Eq. 1.34 is

$$\mathrm{d}U = -p\,\mathrm{d}(V) \;-\frac{1}{\beta}\mathrm{d}(\int \rho \log \rho \,\mathrm{d}\Gamma) + \frac{1}{\beta}\int \rho\,\mathrm{d}(\log \rho)\,\mathrm{d}\Gamma. \tag{1.35}$$

In the last term, we do differentiation of logarithm and we have $\int \rho \frac{1}{\rho}\,\mathrm{d}(\rho)\,\mathrm{d}\Gamma$.

Again, after changing the order of differentiation and integration, the term is equal to 0. Finally, we have a relation for the differential of internal energy derived only from statistical physics consideration, which is possible to compare with well experimentally proven relation of thermodynamics (Eq. 1.27):

$$\mathrm{d}U = -p\,\mathrm{d}(V) \;-\frac{1}{\beta}\mathrm{d}(\int \rho \log \rho \,\mathrm{d}\Gamma). \tag{1.36}$$

Using this comparison, it is obvious that factor β has to be equal to the reciprocal temperature up to a multiplicative parameter (constant) k_B

$$\beta = \frac{1}{k_\mathrm{B}T} \tag{1.37}$$

and differentiated part of the second term is relation for *entropy*

$$S = -k_\mathrm{B}\int \rho \log \rho \,\mathrm{d}\Gamma. \tag{1.38}$$

Parameter k_B is called *Boltzmann constant* and its value can be obtained only by doing an experiment. A requirement for experimental determination of one constant parameter is a common sign of all well-defined physical theories (quantum theory, classical electrodynamics, general theory of relativity, ...). The physical interpretation of k_B is heat capacity of one degree of freedom in the system. This intuitively follows from further considerations in many statistical physics texts [4], but unfortunately not this work. Obtained entropy relation serves as a guideline to understanding what entropy is. After an integration, we may see that entropy is equal to the average value of the logarithm of probability density (multiplied by constant). This relation (especially in the discrete case, where we replace ρ for probability w) is similar to Boltzmann's definition of entropy:

$$S = k_\mathrm{B} \log P, \tag{1.39}$$

P being a probability. While this formula is engraved on his famous memorial in Vienna, he never wrote this equation in present form. Also, the Boltzmann constant k_B (with the formula 1.39) was introduced by Planck [7].

There is still unrevealed constant α left in relation for the probability density ρ in Equation 1.29. Now we use Equation 1.38 for entropy and rewrite logarithm in terms of ρ expectation ($\rho = e^{\alpha-\beta E}$).

$$S = -k_{\mathrm{B}} \int \rho \log \rho \, \mathrm{d}\Gamma \tag{1.40}$$

$$S = -k_{\mathrm{B}}\alpha + k_{\mathrm{B}}\beta \int E\rho \, \mathrm{d}\Gamma. \tag{1.41}$$

Integral in the second term is clearly an average value of energy U:

$$S = -k_{\mathrm{B}}\alpha + k_{\mathrm{B}}\beta U. \tag{1.42}$$

Rearranging previous equation (using $\beta = 1/(k_{\mathrm{B}}T)$ and expressing α there is

$$\alpha = \frac{-S + k_{\mathrm{B}}\beta U}{k_{\mathrm{B}}} = \frac{U - TS}{k_{\mathrm{B}}T} = \frac{F}{k_{\mathrm{B}}T}. \tag{1.43}$$

Last equality follows from thermodynamics, where F is *Gibbs free energy* (thermodynamic potential) defined as $F \equiv U - TS$. Now we have got both constants of previous expectation on function, describing probability density in Gibbs canonical ensemble:

$$\rho(E) = e^{\beta(F-E)}. \tag{1.44}$$

The same derivation would be possible in a discrete case, where energies are quantized and integration is replaced by sum. Then, probability for finding system in state with energy E_n is

$$w_n(E) = e^{\beta(F-E_n)}. \tag{1.45}$$

1.2.3 Partition function

Of course, sum of all partial probabilities has to be equal to 1

$$\sum_n w_n = 1 \Longrightarrow \sum_n e^{\beta(F-E_n)} = 1, \tag{1.46}$$

where term independent on n can be taken out of summation and moved to RHS of the equation:

$$\sum_n e^{\beta E_n} = e^{-\beta F} \Longrightarrow F = -k_{\mathrm{B}}T \log\left(\sum_n e^{-\beta E_n}\right) \tag{1.47}$$

In the end, there is a definition of the Gibbs free energy in the sense of statistical physics. More importantly, term inside logarithm is called *partition function* Z

$$Z \equiv \sum_n e^{\beta E_n}; \qquad \text{or} \qquad Z \equiv \int e^{\beta E} \, \mathrm{d}\Gamma \tag{1.48}$$

and plays a crucial role in the whole framework. Importance and detailed meaning of the partition function is revealed in the following section about Boltzmann distribution. However, even now it can be said that, if we are able to compute the partition function of the statistical system, then we know nearly everything about the system through simple relation $F = -k_{\mathrm{B}}T \log Z$. With knowledge of the Gibbs free energy, one can easily compute entropy, internal energy, and pressure (state equation).

1.2.4 Boltzmann distribution

In practice, many classical systems may be modeled using Gibbs canonical ensemble. We call that these systems are described by *Boltzmann statistics* or *Boltzmann distribution*. However, the second term is incorrect from a strictly mathematical point of view, where distribution is defined as integral from probability density. In physics, authors sometimes use term distribution function or statistical distribution function with the meaning of probability density (especially in older works [4]) and the reader should carefully decide its meaning from the context. There are also other names used for Boltzmann distribution in various areas, in mathematics it is Gibbs measure, in statistics log-linear model and in machine learning they use term softmax.

Historically this "distribution"was derived by Boltzmann using a different approach to presented one (through Gibbs canonical ensemble). As we will see later, it can be also derived from a completely different approach in information theory as the most probable distribution in a case when there is not any *prior information* about the system. It is the Boltzmann distribution and mentioned fact, what is connecting such distinct research areas as statistical physics and machine learning (or generally information science). However, it is not only one connection, later there will be revealed more nontrivial connections.

There would be also the possibility to start text about machine learning with a definition of Boltzmann distribution, out of nowhere, and build everything else on this "axiom". On the contrary, I believe that this common approach is dismissing many interesting consequences and limitations, which are clear after detailed ab initio derivation. Just to follow up, let's define the probability of finding a system in the state with energy E as

$$P_\beta(n) = \frac{e^{-\beta E(n)}}{Z_\beta}, \tag{1.49}$$

where factor $\beta \equiv \frac{1}{k_{\mathrm{B}}T}$, where k_{B} is Boltzmann constant and T temperature. The numerator in RHS of Equation 1.49 is called *Boltzmann factor* and Z_β is the partition function defined as

$$Z_\beta = \sum_n e^{-\beta E(n)}. \tag{1.50}$$

In this form, it is easy to see that partition function is a sum over all Boltzmann factors, serving as a norm for probability. For the continuous case, the sum is replaced with integration (generally, the system of interest may have part of energy levels discrete and rest continuous).

While the description of a physical system may seem easy when we "only"have to compute partition function and everything else is obtained from it, in practice, it is not that straightforward. Such a system can have infinitely many energy states which have to be taken into account. As a result of those complications, there are only a few systems where we can find an analytical expression for the partition function. Systems in higher dimensions (or) with various interactions can be studied only numerically or with proper approximations in field theory. It is worth to mention that in quantum systems, energy levels are degenerate and thus Boltzmann factor have to be appended by degeneration factor g as $P_\beta(n) = g_n \exp\left(-\beta E(n)\right)/Z_\beta$. As an example of systems described by Boltzmann statistics we note ideal gas, particles in an external field, ..., but also interestingly a dilute plasma (which is central topic and source of spectroscopic information in LIBS). Last mentioned example reveals really interesting connections inside presented work. We are using statistical physics method to improve and understand machine learning algorithms, which are behaving in correspondence with Boltzmann statistics. Afterward, we use those ML algorithms to process spectroscopic data originated from the process, guided by the same Boltzmann statistics.

It should be emphasized that there exist different statistics describing systems above limitations specified for Gibbs canonical ensemble. In further sections, we briefly mention statistics describing systems beyond Boltzmann statistic, but we will not derive them in detail as in the previous case.

1.2.5 Grand canonical ensemble

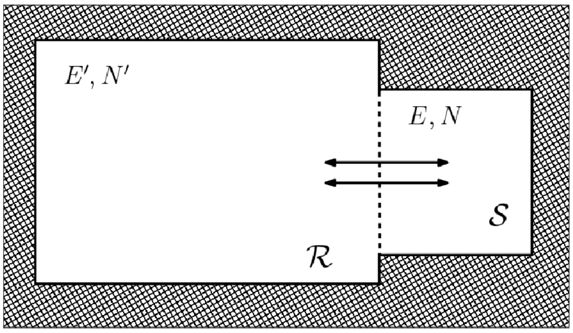

Fig. 1.2: Grand canonical ensemble.

Grand canonical ensemble is another model situation of equilibrium statistical mechanics. In contrast to Gibbs canonical ensemble, exchange of particles between system and reservoir is allowed. Rest of assumptions required for the former ensemble are still demanded. After exhaustive derivation, similar to the Gibbs ensemble, we obtain

$$w_{n,N} = e^{\beta(\Omega - E_{nN} + \mu N)} \quad ; \qquad \rho_N = e^{\beta(\Omega - E_N + \mu N)}, \tag{1.51}$$

for continuous or discrete case, respectively. As was mentioned before, μ is the chemical potential, N is an actual number of particles in the system and Ω is *grand-canonical potential.* Ω has a similar meaning to Gibbs free energy in Gibbs canonical ensemble. It is easy to see that

$$\Omega = -k_{\mathrm{B}} T \log \Xi, \tag{1.52}$$

where Ξ is *grand-canonical partition function* or shortly *grand sum* defined as

$$\Xi = \sum_{n,N} e^{-\beta E_{nN} + \beta \mu N} \quad ; \qquad \Xi = \sum_{n} \int e^{-\beta E_N + \beta \mu N} \, \mathrm{d}\Gamma_N. \tag{1.53}$$

Clearly, the role of the partition function is the same as before.

1.2.6 Statistical description of identical particles

In quantum theory, one cannot distinguish identical particles from each other. That means, if we interchange 2 identical particles, the system remains the same. Only different states of the system can be distinguished. According to the mentioned mechanism, quantum systems are described in a different way according to the nature of its particles. There is *Bose-Einstein distribution* describing the behavior of *bosons*, which have symmetrical wave functions in its coordinates. Second is *Fermi-Dirac distribution* for *fermions* with antisymmetrical wave functions. [8]. Derivation of the corresponding statistics is based on a principle, where, due to the property of identical particles, we have to count all permutations in partition function as only one state.

In Fermi-Dirac distribution, Pauli's exclusion principle has to hold. The mean number of particles in the i-th state is

$$\bar{N}_i = \frac{1}{\exp[\beta(\varepsilon_i - \mu)] + 1}, \tag{1.54}$$

where ε_i is energy of the i-th state. This type of a function is in mathematics called *logistic function* or (especially in machine learning) *sigmoid*. In low-temperature limit ($\beta \to \infty$), we may easily inspect that all states are filled with one particle (actually there are 2 with opposite spins) until reaching state $\varepsilon_F = \mu$ called Fermi energy. Every state with higher energy than Fermi energy is vacant. Most prominent application for this statistics is the behavior of electrons in metals.

In the case of bosons, they are not guided by Pauli's principle so there is not any restriction on the number of particles in one state. In Bose-Einstein distribution, the mean number of particles in the i-th state is

$$\bar{N}_i = \frac{1}{\exp[\beta(\varepsilon_i - \mu)] - 1}. \tag{1.55}$$

There is a condition on chemical potential $\mu \leq \varepsilon_0$ (following from the partition function convergence). Again, we may construct low-temperature limit ($\beta \to \infty$) and inspect, that every particle is in ground energy state. This behavior is a sign of the state of matter called *Bose-Einstein Condensate*. Another example of B-E statistics are photons.

Apparently, the two mentioned distributions are different only in sign ± before 1 inside denominator. Considering a system with higher distances between particles (dilute gas), not affecting each other or system with high temperatures and a great number of accessible energy states, we could suppose a "classical" behavior described by the Boltzmann statistics. This assumption will be confirmed by the following steps. To obtain the same limit dependence of both statistics, we have to get rid of term ± 1 in the combined equation:

$$\bar{N}_i = \frac{1}{\exp[\beta(\varepsilon_i - \mu)] \pm 1}. \tag{1.56}$$

This can be done if exponential part of the denominator is much larger in comparison with 1, which means that $\bar{N}_i \ll 1$. In such case

$$\bar{N}_i \sim \frac{1}{\exp[\beta(\varepsilon_i - \mu)]}, \tag{1.57}$$

what is classical Gibbs distribution and even more if we forbid particle exchange ($\mu = 0$) there is the Boltzmann distribution

$$\bar{N}_i \sim \frac{1}{\exp(\beta \varepsilon_i)}. \tag{1.58}$$

At this point, we have covered all important topics of equilibrium statistical physics of particles (except fluctuations) and we may move slightly to statistical physics of fields, where we only gather the most important topics and tools usable for machine learning.

2 STATISTICAL PHYSICS OF FIELDS

In the previous chapter, we have studied the behavior of macroscopic systems composed of a huge number of particles. As it was discussed, classic-mechanical approach (computing equations of motion) is impossible considering the number of particles. Solution for this problem was to focus on the collective behavior of particles and regularities which have appeared. Many useful results were obtained, but we have dealt mostly with non-interacting systems in an equilibrium state. However, in the real world, we experience problems with various interactions between system particles, overreaching limitations of the statistical-mechanical approach. *Field Theory* became one of the most useful and precise physical theory of all time with applications in Quantum Theory and also Statistical Physics, enabling to deal with much more complex systems. Field Theory approach is many times based on outstanding ideas as using symmetries of the system and locality of interactions. Approximations taken with consideration of these ideas may rapidly simplify the description of the problem in comparison to ab initio approach, which could be often impracticable.

Basic course on Statistical Physics of Fields usually covers *phase transitions*, *criticality*, *fluctuations*, *renormalization group*, and other branches, not covered in this thesis. We introduce only the concept of the *Lattice Systems*, *Mean Field Theory*, and slightly cover more general *variational free energy approach.* For a deeper insight to the topic is recommended to study book by M.Kardar [9] or introductory/review text [10].

2.1 Lattice models

Lattice models of statistical physics play a crucial role in various areas of interest, ranging from condensed matter physics to theoretical physics. As the name suggests, they are defined at lattice (e.g. atoms in crystalic structure) in contrast to continuous models. Their great importance is based on computability of complex physical systems consisting of many particles. In a few cases, there are exactly solvable models, besides to perturbatively solvable models. Another application is in computational physics where they serve as discretization tools for continuous problems. Especially for our topic, 2D Ising model reveals the connection between energy-based machine learning models and statistical physics, what is exactly the topic of presented work.

As a lattice model, we understand model defined on graph $G = (V, E)$ where V are vertices and E edges (see figure 2.1). Most commonly G is regular D dimensional square (or hyper-cubic) lattice, but we can imagine even more general geometries. [10]

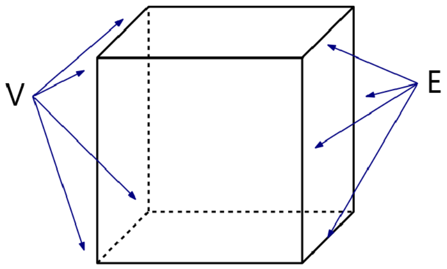

Fig. 2.1: Graph of lattice model.

In the vertex points, there are placed *statistical degrees of freedom* (e.g. discrete spins or more generally continuous variables). Spins are interacting with each other, but commonly interaction is restricted w.r.t. position (to nearest neighbors). We assign an energy functional (Hamiltonian) $\mathcal{H} = E/(k_B T)$ to every spin configuration depending on interaction strength and action of external fields. If our model lives in dimension $D \geq 2$ we observe phase transition at T_C. However, even if we have restricted interaction to short ranges, correlation length between spins can reach long distances close to T_C. More about critical behavior of the statistical systems could be found in [9].

2.1.1 Ising model

Most representative candidate of lattice models is the *Ising model* (fig. 2.2). Hamiltonian for 2 dimensional Ising model is defined as

$$\mathcal{H} = -\frac{1}{2} \sum_{(i,j)\in E} \frac{1}{2} J_{i,j} s_i s_j - \sum_{i \in V} h_i s_i \tag{2.1}$$

where $s_i = \pm 1$ stands for a spin, J_i is a coupling constant and h_i is external magnetic field acting on element i. The sum is taken only over the nearest neighbors. Ising model is exactly solvable only for $1D$ and $2D$, while in $2D$ there is required $h = 0$. Despite non-analytical solutions for $D > 2$ and nonzero h, there is good understanding of model behavior based on approximations from field theory.

To examine properties of the Ising model we have to compute its partition function and correlation functions. Probability of finding system in specific configuration follows Boltzmann distribution

$$P_\beta(S) = \frac{e^{-\beta \mathcal{H}(\mathcal{S})}}{Z_\beta}, \tag{2.2}$$

here factor $\beta \equiv \frac{1}{k_\mathrm{B} T}$, where k_B is Boltzmann constant and T temperature. Z_β is

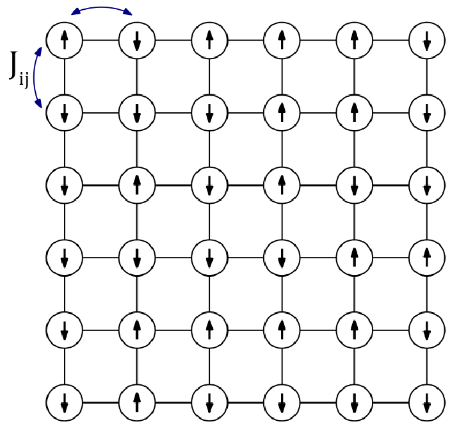

Fig. 2.2: Ising model in 2D.

partition function defined as

$$Z_\beta = \sum_S e^{-\beta\mathcal{H}(\mathcal{S})} \tag{2.3}$$

Ising model can be further generalized for a case where coupling (J_{ij}) and external field (h_i) are no more constant. We call this system *spin glass* and it will be mentioned in connection with Machine Learning.

2.2 Mean Field Theory (MFT)

We describe Mean Field Theory on an example of the Ising model, taken from [10]. As was stated, it is impossible to analytically compute partition function for the 2D Ising model in the external field. Suppose 2D Ising model with constant external field h and constant coupling J_0. The energy of specific spin configuration ($\mathbf{s}$) is

$$E(\mathbf{s}) = -J_0 \sum_{i\sim j} s_i s_j - h \sum_i s_i, \tag{2.4}$$

where summation is over pairs of neighbor sites.

For selected spin inside Ising model, MFT is replacing interactions with neighbors by placing it into an *effective field* created by neighbor spins. For our 2D case, mean field is $h + 4J_0\bar{m}$, where $\bar{m}$ is local magnetization of the spin ($\bar{m} = \langle s \rangle$). That can be expressed by replacing the first sum in equation 2.4 by by $4\bar{m}\sum_i s_i$. The energy of spin configuration is then transformed to a sum of independent terms

$$E(\mathbf{s}) \to \bar{E}(\mathbf{s}) = \sum_i (4J_0\bar{m} + h)s_i = -(h + \Delta h)\sum_i s_i. \tag{2.5}$$

Now, when we got rid of interaction, partition function could be computed for a single spin as

$$Z_{s_i} = \sum_{s_i=\pm 1} e^{-\beta s_i(h+\Delta h)} = 2\text{cosh}(\beta(h+\Delta h)). \tag{2.6}$$

For a system consisting of N spins we have $Z^{MF}_{[s]} = (Z_{s_i})^N$. Local mean energy may be defined as $e_i \equiv (4J_0\bar{m} + h)s_i$. Using the Boltzmann distribution we obtain a local magnetization

$$\bar{m} = \frac{e^{-\beta e(+)} - e^{-\beta e(-)}}{e^{-\beta e(+)} + e^{-\beta e(-)}} = \tanh(\beta(4J_0\bar{m} + h)). \tag{2.7}$$

2.2.1 Variational Free Energy Minimization

Mean Field Theory described above is only a special case of more general approach by Feynman and Bogoliubov, using variational free energy [11]. Previous results of MFT may be also obtained by approximating complex distribution $P(x) = \exp[-\beta E(x)]/Z$ by simpler distribution $Q_\theta(x) = \exp[-\beta E_\lambda(\mathbf{x})]/Z_Q$. Such an approximation is done by adjusting parameters θ. We need to define the *relative entropy* (or *Kullback-Leibler divergence*) as

$$D_{KL}(Q||P) = \sum_x Q_\theta(x) \ln \frac{Q_\theta(x)}{P(x)}. \tag{2.8}$$

It describes a similarity of two probability distributions (more exactly, it measure information loss by using approximation Q instead of true probability distribution P in *bits* or *nats*). $D_{KL}(Q||P) = 0$ if and only if $Q_\theta = P$, else $D_{KL} \geq 0$. Using definition 2.8 and taking $\beta = 1$, we obtain

$$\begin{aligned} D_{KL}(Q||P) &= \sum_x Q_\theta(x) \log Q_\theta(x) - \sum Q_\theta(x) \log P(x) \\ &= S(Q_\theta) - \sum Q_\theta(x)[-E(x) - \log Z] \\ &= \langle E(x) \rangle_Q - S(Q_\theta) + \log Z, \end{aligned} \tag{2.9}$$

where $S(Q_\theta)$ is the *Shannon's entropy*, $\langle\ \rangle_Q$ is an average over distribution Q and we may notice last term, which is well-known "true"*free energy* $\beta F \equiv -logZ$ (derived previously in discussion of Boltzmann distribution). Using equation 2.9, *variational free energy* F_θ is defined in relation

$$\beta F_\theta = D_{KL}(Q||P) + \beta F, \tag{2.10}$$

$$\beta F_\theta \equiv \beta \langle E(x) \rangle_Q - S_Q = \sum_x Q_\theta(x) \log \frac{Q_\theta(x)}{\exp[-\beta E(x)]}. \tag{2.11}$$

Equation 2.10 is implying the key idea of this approach. By varying parameters θ in order to minimize βF_θ, we improve our approximation Q of true distribution P (because of minimizing Kullback-Leibler divergence). This result will be extremely useful for consideration of unsupervised neural networks and their possibility to "learn"probability distribution, studied in chapter 5.

By minimizing variational free energy function of Ising model with respect to variational parameters **a**, we would obtain (for details see [11]) the same equations as were derived by simple mean field theory approach in section 2.2. However, now we use more general form

$$a_m = \beta \left(\sum_n J_{mn}\bar{x} + h_m \right), \tag{2.12}$$

$$\bar{x}_n = \tanh(a_n). \tag{2.13}$$

To identify the new notation with 2D Ising example, the field $a_m = \beta(4J_0\bar{m} + h)$ (the factor 4 originated from summation) and obviously $\bar{x}_n = \bar{m} = \langle s \rangle$. Satisfying equations 2.12 and 2.13, extremization of variational free energy F_θ is guaranteed. Generally, stationary points of F_θ could be also maximum or saddle (not only minimum). This could be treated by *asynchronous* updating of parameters. [11]

3 SPECTROSCOPIC DATA

In the introduction, we have defined *spectroscopy* in general as a collection of methods studying the interaction of electromagnetic radiation with matter. Since this term could cover big amount of techniques with considerably distinct underlying physics and processes, it is difficult to define some common properties of spectroscopic methods. However, the result of any spectroscopic measurement is *spectrum*. For example, the spectrum may represent a number of detected photons on specific wavelength over a defined range of wavelengths, but also a number of detected particles with specific energy (in case of *spectrometry*). Generally, it describes the dependency of "something"on "something". Even for such uncertain definition, we may collect some "data-related"properties of *spectroscopic data*, which holds for almost all types of spectra. Probably the most common imagination about spectroscopic method would be *emission spectroscopy* or *absorption spectroscopy*. In our approach, we suppose spectra of emission spectroscopy, but it can be easily generalized to other types. Here we present basic properties of spectroscopic data (note that some undefined terms are used for brevity, but will be properly explained in Appendix about LIBS):

- *high-dimensionality*

  Dimension of spectroscopic data is dependent on the resolution of a spectrometer, used for measurement. It is not unusual to have tens of thousands variables in spectral data, requiring special approaches to their processing.

- *sparsity*

  In spectral data, we usually observe peak-like structures (*spectral lines*) of known shape and positions. This structure is material-specific and offers valuable information. However, lines are covering relatively minor part of wavelength range and they are surrounded by "unimportant"information (noise, continuum, ...). A trained spectroscopic specialist will surely distill only important features, but in the case of automatized spectra processing, this became a problem. The simple computational model cannot easily recognize what is important and what is noise, special techniques are required to make it possible. This motivation is one of the cornerstones of whole Machine Learning as we study in the corresponding chapter. Also, the computational time required to process such a high dimensional data is growing rapidly. To sum it up, finding a universal tool for recognizing important parts of spectra and suppressing noise would be beneficial to automatized processing of spectra.

- *redundancy of spectral information*

  Considering *atomic emission spectroscopy* (for example), there are usually many *spectral lines* corresponding to one element present in a single spectrum. If our goal is to decide the presence of a specific element in the measured sample, we don't have to identify all lines, but one or two (for confirmation) is enough. This fact implies that spectral data are highly redundant for specific tasks (classification, detection of elements, ...), and can be used with advantage to improve analysis of spectra.

Different type of redundancy is present inside a single spectral line. As we derive later, line is not infinitely thin, but have peak-shape determined by line-broadening mechanisms. From the data point of view, there are many variables corresponding to a single peak, which are correlated together. This needless extent of variables can be represented by only a few variables (central position/wavelength of the line, total intensity, and width of Voigt line profile - discussed later). Amount of necessary variables is determined by task, we aim to solve. For determination of the presence of a specific element in the sample, only one variable is enough, while for quantification we naturally need more information. Property of spectral redundancy is motivating to use dimension reduction techniques for spectral data.

Respecting simple properties of data, generally valid for most spectroscopic techniques, it is theoretically possible (and beneficial) to extract only important information about spectral lines and drop surrounding positions with only noise. Complementary to this, we may ”shrink”peaks to single intensity values (keeping aside information about shape and width of peak) for reducing dimension even more. In practice (for example of simplest classification), we may end with only a few resulting variables representing the presence or absence of selected elements. This is a huge reduction of dimension from tens of thousands of variables to just a few reliable elements (binary values for presence/absence). Of course, the described process is commonly employed by a spectroscopic specialist, without ambiguity. But considering the automatization of spectra processing, finding correct and interpretable method is very challenging and matter of active research.

In appendix, we describe underlying physical processes of spectra creation for representative *atomic emission spectroscopy* method - Laser-Induced Breakdown Spectroscopy (LIBS). Naturally, those processes are method-specific and cannot be taken as valid for other methods generally. LIBS was selected as an example to describe complex processes responsible for the shape of measured spectroscopic data. So, even when some following mechanisms may not be generalizable to other spectroscopic techniques, is important to remind that previously mentioned properties are valid in general.

3.1 Laser-Induced Breakdown Spectroscopy (LIBS)

LIBS is an analytic spectroscopic technique for obtaining the elemental composition of the sample. High power laser pulse, focused on the target, is used for sample analysis. Concisely, the process of laser-matter interaction consists of few following stages. Firstly, a small volume of material is heated, evaporated and atomized – microplasma is created (see Figure 3.1).

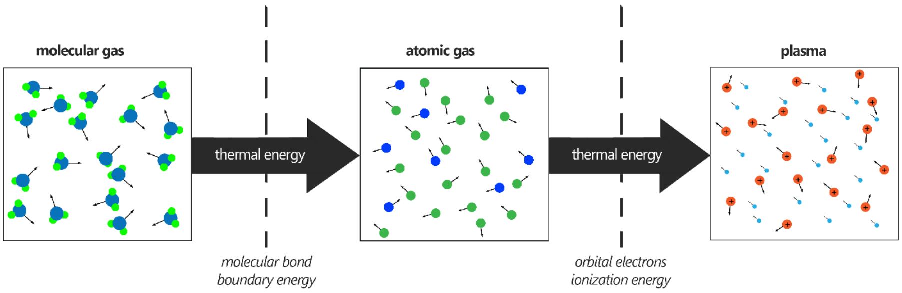

Fig. 3.1: Schematic representation of plasma creation process. After evaporation of material due to the applied electromagnetic field, molecular gas is formed. Addition of thermal energy to the system results in the formation of atomic gas and further after ionization of atoms, the creation of plasma.

Exact process is dependent on used wavelength and time, for which radiation is emitted. For clarity, let's suppose the case of nanosecond pulsed laser (this means, that photons are radiated during timescale ranging in few nanoseconds). After plasma creation, radiation is still ”on”and atoms are excited to higher energy levels. In the plasma plume, there are competing various mechanisms of energy transfer and dissipation. At this stage, plasma mostly consists of free electrons, ions, atoms, but still possibly some molecular structures. Common temperature of such laser-induced plasma could be cca 10 000 - 20 000 K. Plasma plume is naturally expanding to surrounding space and during this expansion, plasma is cooling down. During this process, electrons are recombining back to atoms and binded excited electrons undergo radiative transitions to lower energy levels. Photons created during these transitions have specific wavelengths, depending on the energy difference of corresponding levels. In LIBS we collect radiation of plasma and guide it to the spectrometer. Result of LIBS measurement is a line spectrum, what can be though as a “chemical fingerprint” of examined material. One of the greatest advantages of LIBS comparing to other techniques of similar interests (ICP-MS, LA-ICP-MS,. . .) is the speed of the analysis (currently up to 1 kHz). This feature makes it suitable for mapping bigger surfaces with a lateral resolution in the order of tens of micrometers. Analysis by LIBS usually doesn't require any sample preparation and cost practically nothing. Thus, it is possible to provide LIBS analysis at great distances (up to 30 meters) and outside the lab. The great advantage of LIBS is also a possibility to measure samples in liquid or gas phase. Applications of LIBS ranging from the metal industry, environmental studies, geology to space exploration. Nowadays, there is also a growing interest in biological application of LIBS, which resulted in many technological improvements. For example, special techniques for lowering the size of the laser spot and preserving the adequate environment for samples during analysis can rapidly improve obtained results.

4 MACHINE LEARNING

It has already been mentioned that in contrast with traditional statistical analysis of data, Machine Learning is aiming primarily on prediction and not to estimation of some parameters describing data. Even more, methods of ML are more suitable for use in case of high-dimensional data. Of course, there are many aspects those two fields have in common, such as using an observable $\boldsymbol{x}$ representing the system of interest. This system is guided by a generating process (or model) with parameters $\boldsymbol{w}$ and then the probability of observing a system in state $\boldsymbol{x}$ is given by conditional probability $p(\boldsymbol{x}|\boldsymbol{w})$. If we perform an experiment on the system, we measure a set of observables $\boldsymbol{X}$, which are used to fit a model of parameters $\hat{\boldsymbol{w}}$. Surely, a model obtained from fitting data and real model are generally distinct due to noise and error originated from the measurement. Thus we are searching for parameters that maximize the probability of observing $\boldsymbol{X}$ as $\hat{\boldsymbol{w}} = \text{argmax}_w p(\boldsymbol{X}|\boldsymbol{w})$. Final remark to the difference of ML and statistical analysis is that while in case of statistical analysis one is considering the accuracy of $\hat{\boldsymbol{w}}$ used for estimation, in ML interest is placed to the possibility of the model to predict behavior for new observation, also called generalizability of the model. To reach a goal of ML (good generalizability of model), different and "new"approaches has to be used. New approaches and techniques to deal with large high-dimensional datasets raised from many different fields as computer science, statistics, biology and importantly physics. While ML is a young and fast developing branch, it may sometimes rely on more empirical observation and formal mathematical proofs may be lacking. However, similar cases are well known also in physics, where we may have good intuition and empirical results for the theory, but internally it is inconsistent (e.g. QM Dirac's equation, path integrals, Ising model for higher dimensions and many more). Fortunately, physics have good potential with helping to define and understand the behavior of ML models. To mention parts where ML strongly rely on physics, there are notoriously known examples as Monte-Carlo methods, variational methods and finally so-called energy-based models of ML which are topic of main interest in the thesis. The way of using physics to the understanding of ML is searching for connections between structures and using abstraction to explain its behavior. Besides the influence of physics to ML, surely there are other contributing fields and approaches to ML which are not presented here. Greatest importance will be taken to describe basic cornerstones of ML as-is the process of learning in 2 most basic cases (supervised and unsupervised) generally. The task of supervised learning is usually a classification of data or regression, while unsupervised learning is more abstract and looking for some patterns or new regularities in original data. After defining basic concepts, we move to a subset of ML which is Artificial Neural Networks, leaving many important topics and subbranches untouched. To obtain a consistent and complete understanding of ML, the reader is encouraged to see monograph by C.M.Bishop [12] or recent well-written introduction to the topic by P.Mehta et al. [13]. Our presentation of the topic in the thesis is partially inspired and hierarchically consistent with the second mentioned reference.

4.1 Model and Cost Function

Our treatment starts with the procedure of ML-related analysis. Firstly suppose that there is dataset $\boldsymbol{X}$ originated from some generating process, which we would like to learn about. To learn about the process, we build a parametric model $g(\boldsymbol{w})$ based on our knowledge of the generating process. If we have a model, there is a need for some figure of merit describing how well can model describe observed data $\boldsymbol{X}$. This figure of merit is called *cost function* $\mathcal{C}(\boldsymbol{X}, g(\boldsymbol{w}))$. A simple example, for the cost function is often used squared error (which is a good metric for low dimension, but as we will see later problematic in higher dimensions). Learning process could be defined as the changing of model parameters to minimize the cost function. Standard approach of learning arbitrary ML model is a separation of dataset $\boldsymbol{X}$ disjunct subsets, training $\boldsymbol{X}_{\text{train}}$, validation $\boldsymbol{X}_{\text{valid}}$ and test $\boldsymbol{X}_{\text{test}}$. Model training is done on the training and validation data (sometimes collectively called train data $\boldsymbol{X}_{\text{train}'}$), while test data are used to obtain final performance. To compare the performance of model on the training set with performance on the test set, we introduce in-sample error and out-sample error respectively. In-sample error is just value of cost function for model with best fit, $E_{\text{in}} = \mathcal{C}(\boldsymbol{X}_{\text{train}'}, g(\hat{\boldsymbol{w}}))$, where $\hat{\boldsymbol{w}} = \text{argmin}_w \mathcal{C}(\boldsymbol{X}_{\text{train}}, g(\boldsymbol{w}))$. Similarly, out-sample error is defined as $E_{\text{out}} = \mathcal{C}(\boldsymbol{X}_{\text{test}}, g(\hat{\boldsymbol{w}}))$. It is obvious that E_{in} is generally always lower than E_{out}, because of the way how the model was trained. During the learning process, the model could become *overtrained*, which means that in-sample error is lowering while the out-sample error is growing rapidly. Overtraining is an unwanted effect preventing the model from good generalizability. In further sections, we closer discuss mechanisms of how to control and prevent overtraining in the learning process.

In order to clarify abstract definition of ML model, we demonstrate mentioned terms on a simple model of polynomial regression, where cost function will be the basic squared error. Let's suppose dataset obtained from some generating process (1D function $f(x_i)$ in this case) affected by additive Gaussian noise. Thus data are described by equation

$$y_i = f(x_i) + \eta_i, \tag{4.1}$$

where η_i is normally distributed additive noise with an average value equal to zero and variance σ.

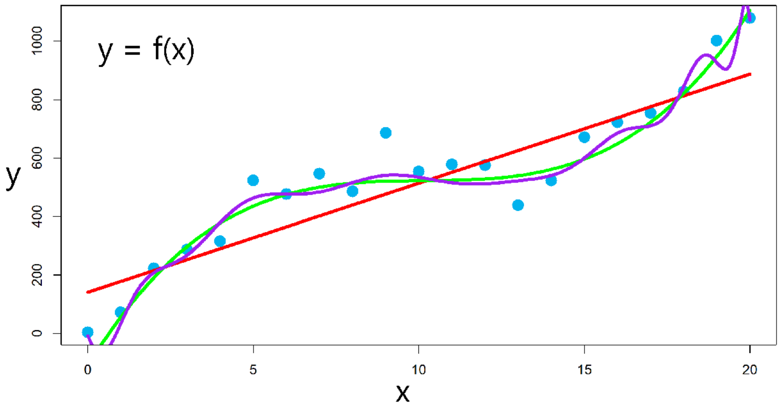

Fig. 4.1: Noisy observation of data produced by 1D function with constant sampling and fitted polynomials of different orders.

In the picture 4.1, we may see such observation of noisy data with fitted polynomials of 1st, 3rd and 30th order. We use this example for an explanation of the difference between fitting and prediction. It is clearly visible that fitting by line (1st order polynomial) is taking into account the only global trend of data, but ignores the true shape of function and noise. This behavior is well expected, due to the simplicity of a model with only one parameter. While the fitting performance of the linear model is really poor (computed squared error would be high), its predictive power is not worst (in comparison with high-order polynomials). Considering the polynomial of 3rd order, both fitting and predictive performance is improved. Even more, in the presented example, it seems that 3rd order polynomial is best for regression of data and thus best for revealing generative process (function). Last curve representing fitting by 30th polynomial obtained the best result in fitting (lowest squared error), but it is easy to see that the predictive power of such model is worst. This is a nice example of overtraining or overfitting in this case. To conclude, the function used for dataset generation was a cubic polynomial with additive noise superposed. Thus, it is not surprising that the best predictive power was obtained by model consisting of 3 parameters. This discovery is easily extendible to other ML models (more complicated than polynomial regression). If the generating process of data is simple, using a model with many parameters will tend to overfit and capture the noise of the data. This simple statement acquired big importance especially in Artificial Neural Networks, where the model is consisting millions of parameters and special techniques to prevent overfitting has to be applied. However, it is worth to mention that while considering large dataset containing a huge number of observations, the role of noise and overfitting is restrained by good statistical behavior. More extensive treatment of polynomial regression could be found in the mentioned work [13], where authors also discuss the role of the number of observations and model complexity to errors (in-sample and out-sample). Figure 4.2 is introducing so-called *bias* and *variance* of ML model. Generally, it can be said that we are looking for a model with high bias and lowest possible variance, which of course comes for a price. Bias is limiting the best possible performance for the case of having infinitely many observations (training data) and variance is setting fluctuations in performance.

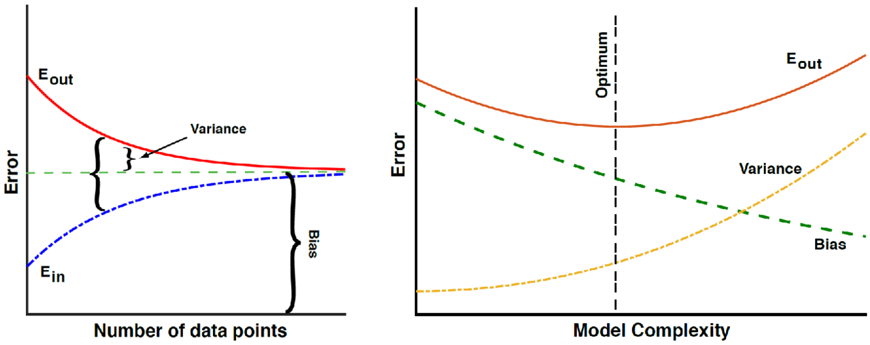

Fig. 4.2: (left)Error dependence on number of observations. (right) Error as a function of model complexity. The figure is introducing competing bias and variance of the ML model, with optimal value for model complexity. ([13])

4.1.1 Gradient Descent

Previously, we have mentioned essential aspects of ML analysis, data division, model building, and model-parameter dependence. Now, let's move to description of learning process itself, considering multi-parametric model $g(\hat{\boldsymbol{w}})$ with cost function $\mathcal{C}(\boldsymbol{X}, g(\hat{\boldsymbol{w}}))$. Training (learning) of a model is done by minimizing the cost function for observations $\boldsymbol{X}$. The actual state of the model could be imagined as a point in multi-dimensional parametric space (similar to phase space of statistical physics). Obviously, it is impracticable to compute cost function for every point in space for models containing a large number of parameters. But if we imagine this or consider the model of only a few parameters, we obtain multi-dimensional landscape of cost function with many local minima, saddles, and complicated structure. A similar problem is well known in statistical physics - spin glass theory, where energy landscapes are considered especially for saddle points. Back to ML, optimization of the cost function in such glassy landscape is clearly challenging problem requiring a nontrivial approach. Even more, while considering datasets with multi-categorical inputs and varying samples inside the category, is extensively difficult to capture this complexity inside simple model.

Define *energy* function of a model as $E(\boldsymbol{\theta}) = \mathcal{C}(\boldsymbol{X}, g(\hat{\boldsymbol{w}}))$, what is just another nomenclature for the cost function. One of the reasons for using term energy is also additivity of this variable. Thus, for example in polynomial regression discussed above, the energy function is a sum of mean squared error for all data points i. Generally

$$E(\boldsymbol{\theta}) = \sum_{i=1}^{n} e_i(\boldsymbol{X}_i, \boldsymbol{\theta}), \tag{4.2}$$

e_i being mean squared error of i-th data point. Generalizing gradient operator of differential calculus to more dimensions ($\nabla f = \frac{\partial f}{\partial x^i} g^{ij} \hat{\boldsymbol{e}}_j$), we have a tool of finding local minima in energy landscapes (formally gradient is defined as the direction of biggest growth, what can be easily fixed by including minus sign). In ML we call *gradient descent* (GD) an algorithm which iteratively translates model in parameter space to lower energy. Initial state of the model $\boldsymbol{\theta}_0$, is changing according to

$$\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0 - \boldsymbol{v}_t, \tag{4.3}$$

where one iteration is

$$\boldsymbol{v}_t = \eta_t \nabla_\theta E(\boldsymbol{\theta}_t), \tag{4.4}$$

η_t being so-called *learning rate* which is regulating the size of translation step in parametric space. Using a small learning rate will ensure convergence to a local minimum of the energy function, but the price for it is extensive computational time. Middle-to-big learning rates may cause oscillation around minima or even divergence, respectively. Thus, it is clear that the setting of acceptable learning rate is crucial stability of gradient descent algorithm and so for learning a model. LeCun et al. in their work [14] had shown that the optimal learning rate could be found as

$$\eta_{opt} = \left(\frac{\partial^2 E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right)^{-1}. \tag{4.5}$$

This derivation was done by using insight from more complicated (w.r.t. gradient descent) Newton's optimization method. According to used learning rate, 4 learning regimes of model exist (see Figure 4.3). As was mentioned earlier, we should try to keep the learning rate at optimal value, because too low or high learning rate is preventing the model from successful training.

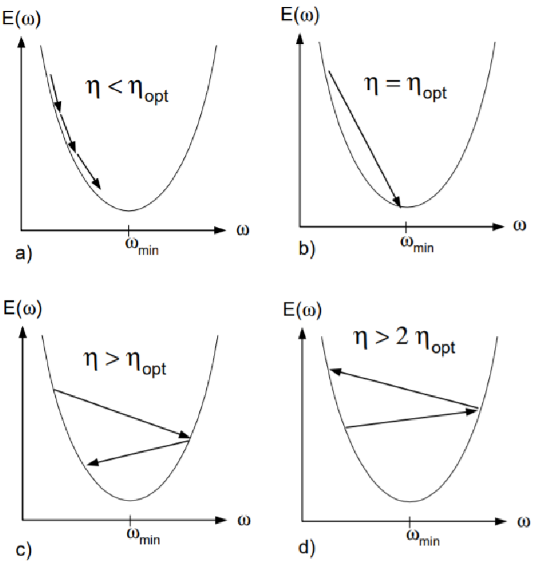

Fig. 4.3: Convergence of a model to a local minimum for various learning rates. The energy function is 1D quadratic potential. (taken from [14])

Now, after obtaining a tool for learning a model, limitations of such an algorithm should be noted. It is clear, that the gradient descent method is finding only local minima of the energy landscape, which can be really poor for overall performance of a model. Even more, it is a deterministic algorithm, so it ends up in the same minimum for specific initial conditions ($\boldsymbol{\theta}_0$). Other problems of GD are the computational cost of gradients computing for large datasets, strong dependence on learning rate, isotropy of learning rate and the possibility of exponential times for escaping saddle points. These limitations are restricting the GD method from use in multi-parametric models for treating large datasets. However, there exist many improvements of the method to make it capable deal with such problems, especially in Neural Networks.

One of the improved GD methods is *Stochastic Gradient Descent* (SGD) and as the name suggest there is stochasticity introduced. This is done by restricting sum in equation 4.2 for $i \in B_j$, where B_k is a *minibatch* of the original dataset. Thus, the original dataset is divided into K mini-batches of equivalent size ($k = 1, 2, ..., K$), *batch size* M being n/K, where n is the number of total observations (data-points in the presented example). Then the gradient of energy function is

$$\nabla_\theta E'(\boldsymbol{\theta}) = \sum_{i \in B_k}^{M} \nabla_\theta e_i(\boldsymbol{X}_i, \boldsymbol{\theta}), \tag{4.6}$$

where $\nabla_\theta E'$ is gradient over minibatch of the energy function. Rest of algorithm is identical to GD ($\boldsymbol{v}_t = \eta_t \nabla_\theta E'(\boldsymbol{\theta}_t)$; $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0 - \boldsymbol{v}_t$). As a result, SGD is replacing total gradient (considering all samples) with only approximative gradient and thus introducing stochasticity to the way of searching for the minimum of the energy landscape. Advantages of SGD are the following: preventing from being stuck in a local minimum, increasing the speed of calculation and regularization of ML model (preventing the model from overtraining), which will be mentioned later.

Another practice of improving SGD algorithm is "adding momentum"

$$\boldsymbol{v}_t = \gamma \boldsymbol{v}_{t-1} \eta_t \nabla_\theta E'(\boldsymbol{\theta}_t), \tag{4.7}$$

where γ is a momentum parameter ($0 \leq \gamma \leq 1$). This type of algorithm is called *gradient descent with momentum* (GDM) and is clear that for $\gamma = 1$ we got back to SGD. A simple analogy from physics can be found in the motion of a particle (mass m) in viscous medium (drag μ). After a detailed analysis of the analogy (see ref. [13]) we may found that momentum parameter is proportional to the mass of the particle and thus effectively generates inertia. Inertia helps to gain speed for the algorithm in flat parts of the energy landscape and oppositely reduce oscillations and smoothen the trajectory in changing parts of the landscape.

There are also other more-advanced methods, usually based on using higher order moments of the gradient. As an example of a widely used algorithm, ADAM optimizer is the method using the first and second moment of the gradient to adaptively set the learning rate for different parameters, thus introducing anisotropy to parameter space. Learning rate is adapted proportionally to the signal-to-noise ratio, which is extremely beneficial in ignoring of small fluctuations in energy landscape and focusing to the general trend. In ADAM there are still basic features as memory (inertia) or stochasticity included.

4.1.2 Maximum Likelihood Estimation (MLE)

. MLE is the method used to estimate parameters of a model for some known fixed data. It is based on maximizing the *likelihood function* $p(\boldsymbol{X}|\boldsymbol{w})$, describing the probability of observing the data $\boldsymbol{X}$ for varying parameters $\boldsymbol{w}$ of some *prior* distribution $p(\boldsymbol{w})$. Thus in MLE, we choose parameters which maximize likelihood (or equivalently log-likelihood) of the observed data [13]:

$$\hat{\boldsymbol{w}} = \mathrm{arg}_{\boldsymbol{w}}\mathrm{max}\ \log\ p(\boldsymbol{X}|\boldsymbol{w}). \tag{4.8}$$

For this task, the Bayes theorem is used for obtaining a *posterior distribution* $p(\boldsymbol{W}|\boldsymbol{X})$. Even more, we need partition function and tool for drawing samples from posterior distributions (usually done by Markov Chain Monte Carlo methods) [13].

4.1.3 Logistic Regression

Until this point, we have discussed properties of models with continuous outputs. However, many tasks of ML are dealing with categorically tagged data (classification tasks). Common applications of ML classification are: assigning spectra to specific materials, distinguishing pictures of cats and dogs, classifying critical phase of Ising model and many more. Logistic Regression is a basic representant of models with categorical output (binary). In spite of it's relative "simplicity"(or commonness) it is still widely used inside of complicated modern Deep Learning models.

Simplest approach for obtaining binary output from continuous input would be setting of some threshold or rescaling values and using signum function (defined as $f(x) = \mathrm{sign}(x) \equiv 1$ for $x \geq 0$ and $\mathrm{sign}(x) \equiv 0$ for $x < 0$). This type of classifier is called "hard", also known as *perceptron* in ML community. In contrary, Logistic Regression is a representative of "soft"classifiers. The output from logistic regression is interpreted as a probability of data sample bmx_i belonging to a category $y_i = \{0, 1\}$ as

$$P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{x}_i^T \boldsymbol{W}}}, \tag{4.9}$$

$$P(y_i = 0|\boldsymbol{x}_i, \boldsymbol{\theta}) = 1 - P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta}), \tag{4.10}$$

where $\boldsymbol{\theta}$ are model parameters. Equation 4.9 is well known in statistical physics, for system consisting of two energy levels.

There is possibility to generalize logistic regression to multi-category classification model, called also *softmax regression.* Cost function to optimize in logistic regression is *cross-entropy* (see [11] and [10]).

4.2 Artificial Neural Networks (ANN)

ANN are one of the most used tools in modern ML. Common way of introducing ANN is a statement that inspiration for their design came from the human brain and its possibility to learn. However, the compatibility of this statement with modern observations in neuroscience is up to a discussion, but it is satisfactory enough for our considerations.

Common architecture of such neural network consists of neurons (or nodes) grouped in the input layer, hidden layers and output layer (see Figure 4.4). The input layer, consisting of nodes is representing training data, which are fed to the network, sample by sample. Roughly speaking, hidden layers transform inputs to output data of the different shape.

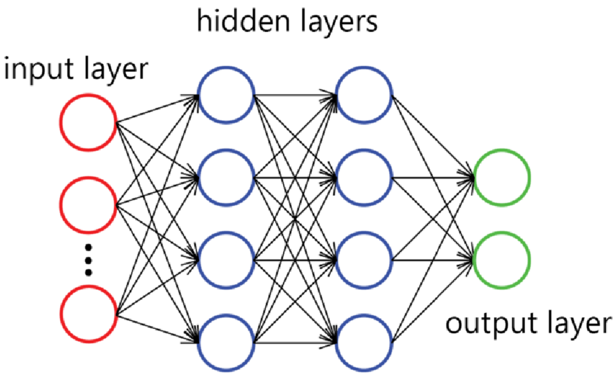

Fig. 4.4: Example of the basic architecture of Artificial Neural Network consisting input layer, hidden layers and output layer with 2 neurons.

Nodes in each layer are connected to every node in the following and previous layer, but not with other nodes in the same layer. A number of hidden layers may vary and while there are more hidden layers present, we speak about *Deep Learning*. This name, originated by Hinton [15], was successfully used to rebrand and awake older field of ANN. Of course, reborn of the ANN was caused primarily by the appearance of the new approaches and efficient algorithms and not just by name. Back to the architecture of ANN, through the connections between neurons (or layers), simple mathematical operations (parametrized multiplication, sum, and nonlinear mapping) on inputs are evaluated and passed to the next layer. The whole mechanism can be learned by a technique called backpropagation which is described in appendix. Learning of ANN can be though as showing labeled inputs or patterns to network with adjusting its parameters to obtain specific outputs. After learning procedure, the network can be used for example to classify unknown data – this is the case of *supervised learning* discussed later. There is also another big group (or groups) of ANN which in not totally compatible with the presented introduction. These are networks used for *unsupervised learning* (or Reinforcement Learning - not described in thesis), in this work treated separately later.

A building block of ANN is a *neuron*, computational unit with $n + 1$ input connections and one output connection (see Figure 4.5).

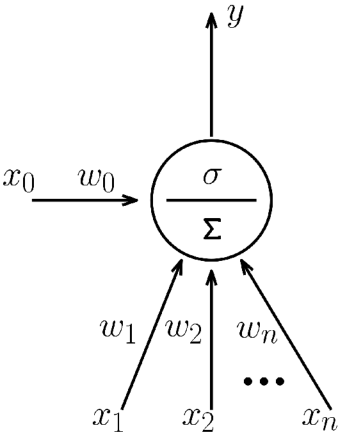

Fig. 4.5: Schematic drawing of neuron with $n+1$ inputs, weighted sum operation and nonlinear function before the output y.

Input connections are supplying data (values) x_i and multiplying them by weight factor w_i. There is one special input x_0 called *bias*. Inside of neuron, those values are summed up and passed to *activation function* σ. Mathematically, for the k-th neuron, we have

$$z^{(k)} = \sum_i w_i^{(k)} x_i^{(k)} + b^(k), \tag{4.11}$$

$$y^{(k)} = \sigma(z^{(k)}), \tag{4.12}$$

where $z^{(k)}$ is a weighted sum of inputs and $y^{(k)}$ is output from the neuron. The output is calculated using the activation function, which is a non-linear function. Various types of activation function are used, while most common are perceptron (historically), sigmoid function and ReLU (have Rectified Linear Unit), all shown in Figure 4.6.

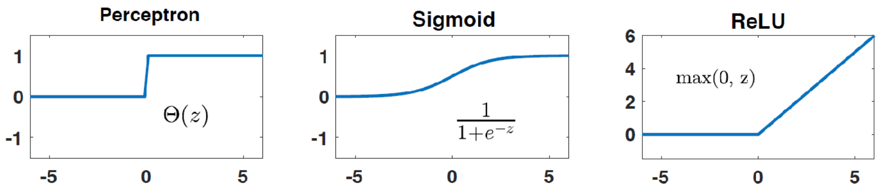

Fig. 4.6: Graphs of common activation functions. (taken from [13])

Considering the graph of functions, differences between them are obvious. First two mentioned have problems with saturation for large input values and treatment for this problem came in the form of ReLU. However, sigmoids as activation function are still used at specific places of neural network or with proper normalization. As it was mentioned, learning is usually done with Gradient Descent methods which have a different effect on different activation functions. Obviously, GD learning cannot be used for perceptrons, due to non-zero derivative only at zero. Sigmoids are continuous well-behaved functions, but suffer from *vanishing gradients* effect. This is caused by saturation of function for larger values, where learning is rapidly slowed or stopped. Treatment for vanishing gradients came in the form of ReLU function, which is now one of most commonly used activation function even despite discontinuity at zero. Selection of correct activation function or its combination depends on a specific application. Obviously, for classification tasks, the output layer has to consist of functions with possible categorical output as sigmoids, while in regression task we require some linear function at the single neuron of the output layer. However, there is not any exact or optimized way how to choose correct architecture because of dependence on a huge number of parameters, preventing to perform standard optimization. Of course, there are some recommended approaches, but often experience and heuristic are used.

4.2.1 Deep Learning

There exist a theorem, stating that arbitrary continuous function defined on $\mathbb{R}^n$ can be approximated by single layer ANN with only modes requirements on activation functions [16]. However, a number of neurons in that single layer network has to be sufficiently large for obtaining required accuracy, which leads to impractible computation. For a great advantage, it was found that increasing number of layers and reducing number of neurons in a single layer could bring demanded representative power and lower computational requirements drastically. While the depth of the network is raising, also its complexity is growing rapidly and computing gradients for learning is quite challenging. That was probably one of the reasons for a succession of failures in attempts of applying ANN to ML task in early days and also the reason of great success in the modern era of ANN, where we have good approaches for this task. Algorithm suitable for computing gradients in deep networks is called *Backpropagation* and was mostly popularized by G. Hinton [17], but originally was found even before him.

4.3 Regularization techniques

Problems with overtraining of ML models were repeatedly mentioned through the whole chapter and here we focus on special techniques developed to overcome this unwanted effect of learning. Collective names for those techniques is *regularization*. To refresh motivation and explanation of why models tend to overtrain, the example of polynomial regression from section 4.1 will serve well. Models with the number of parameters comparable to the number of data samples tend to overfit and describe intrinsic noise of the data, which is every time present. In extreme case, where the number of parameters is way higher than the number of samples ($p \gg n$), models cannot learn. This is exactly the reason for poor performance of ANN to tasks, where only small datasets are available (but excellent for huge datasets). Most common forms of regularization are surely L_2 penalty(*Ridge regression*) and L_1 penalty(*LASSO*), which were firstly used in linear regression and still serve even for more complex models. Formally, methods are adding some penalty to regression problem, resulting in the following way of searching for model parameters. In Ridge-Regression we add L_2 norm to the cost function (least square loss function) and so L_1 norm for LASSO Regression. The basic idea behind LASSO is reducing complexity (by putting some parameter weights equal to zero) and Ridge is reducing the variance of model while increasing its bias. Unfortunately, detailed treatment of those methods is beyond the scope of this work, but an interested reader could find more in work [12] or [18].

In practice of Neural Networks training, basic regularization technique was already mentioned in the form of Stochastic Gradient Descent. It was the property of stochasticity, that prevents the model from overfitting. To complement SGD, *Dropout* method could be used. As the name suggests, we "drop"some amount of randomly selected neurons from the deep network at every SGD step. After this step, neurons are recovered and new ones are selected for dropping. Generally, this method is suppressing correlations between hidden neurons. Besides improving generalization of the model, dropout also reduces the number of parameters to learn and thus rapidly shorten training time. [19]

The widely used technique, that is achieving regularization in a slightly different way is *Batch Normalization*. The goal of Batch Normalization is to keep activations of neurons around zero mean to restrict vanishing (or exploding) of gradients due to saturation of activation functions. This is done by normalizing inputs to the network by subtracting it's mean and dividing by variance (of a batch). Speed of learning is also enhanced due to well-behaving gradients. [20]

4.4 Supervised Learning

Until this point, every consideration of ML was valid for so-called *supervised learning*. However, as we will see in further sections, part of the supervised learning framework is usable in more general *unsupervised learning*. To sum up the properties and meaning of supervised learning, the most significant fact is, that we are dealing with labeled data and usually we know what we are looking for. For example in classification tasks, the model is trained on dataset separated to categories, trying to distinguish between them. Afterward, unlabeled data may be passed through the model to obtain information whether they belong to some category or not. In tasks of regression, we know a continuous output value for specific training input and model tries to predict this value for unknown data. In both mentioned cases, the generalization of the model is a key to success. In other words, we have to make sure that the model will be able to make valid predictions for previously unseen data samples. This requirement is sometimes not easy to achieve, but some techniques for doing so were mentioned before. The main problem of supervised learning is a lack of the labeled data and difficulties in creating such labels. Usually, it takes a huge amount of time, labeling data by humans and also their performance is not always 100% correct.

It is important to mention, that there exists a special type of Neural Networks achieving state-of-art performance on (not only) image-related classification tasks. This type is called *Convolutional Neural Networks* (CNN), but because it is not explicitly related to the goal of presented work, we are not covering this topic. However, at least motivation for the development of this type of networks should be mentioned. In physics, we understand the importance of symmetries inside physical laws or systems since Noether [21]. Similar case took place in ML classification problems. As an example, If we try to recognize some objects inside an image, the position of the object is not important to the decision of its category. Thus, there is translational invariance in data. Basic ”fully-connected networks”cannot take this symmetry as an advantage, while identical objects placed to different parts of the image are distinct inputs for them. This problem was recognized by ML community and successfully implemented as CNN. [13]

4.5 Unsupervised Learning

A more general version of learning is *unsupervised learning*, where the objective is to find some patterns or new regularities in unlabelled data. Objectives of unsupervised learning are really broad and it is hard to set some borders. Most common tasks belonging to this category are cluster analysis, dimensionality reduction, learning of probability distributions, repairing data, distilling information from noise and much more. In the following sections, we are describing the basics of the topic, but surely not covering all aspects of unsupervised learning. Actually, only parts necessary to reveal connections with Statistical Physics and understand results of the thesis are covered.

4.5.1 Principal Component Analysis (PCA)

PCA is an unsupervised technique commonly used to a reduction of dimension and visualization of high-dimensional data. Historically, it was invented by Pearson in the era when datasets were much less-dimensional and extensive. Nowadays, a big portion of the data-related analysis is starting with using PCA. This great spread of the method resulted in the implementation of PCA to almost every environment or programming language used for data processing. So performing PCA on data and obtaining results could be simply done by executing a single line of code. This simplicity of usage is sometimes resulting in a misunderstanding of the method, not-rarely seen in publications.

Formally, PCA is just a linear transformation (rotation) of the original data to a new coordinate system. Rotations are described by special orthogonal matrices with determinant equal to one. However, the rotation in the PCA method is not taken arbitrary, but with conditions. The goal of PCA is to create new variables (directions) as a linear combination of original variables with few conditions. First is to keep total variance the same and second is a restriction on new variables to be uncorrelated between themselves. This can be also seen as representing data in a new orthonormal basis. New variables (*principal components*) are in the form

$$\begin{aligned} Y_1 &= c_{11}X_1 + c_{12}X_2 + ... + c_{1p}X_p = \mathbf{c_1^T X} \\ Y_2 &= c_{21}X_1 + c_{22}X_2 + ... + c_{2p}X_p = \mathbf{c_1^T X} \\ &\vdots \\ Y_p &= c_{p1}X_1 + c_{p2}X_2 + ... + c_{pp}X_p = \mathbf{c_p^T X}, \end{aligned} \tag{4.13}$$

where c_{ii} are coefficients (elements of transformation matrix) and X_i are original variables. We have presented a case where the number of observables n is equal to the dimension of data d (represented by square matrix $p \times p$) what is rarely satisfied in real problems. Even while $n \neq d$ PCA could be computed, but the number of principal components is equal to the lower number of n and d. Moreover, in PCA we put components (Y_i) in order according to the variance, which is each component describing. There is a task of PCA during the search for principal components. Variance of first component (Y_1) has to be maximized while satisfying normalization condition $\mathbf{c_1^T c_1} = 1$. Mathematically this is a problem of finding extrema with constraints, leading to the method of Lagrange multipliers. After solution of this problem

$$\text{var}(Y_1) = \mathbf{c_1^T}\text{cov}(\mathbf{X})\mathbf{c_1}, \tag{4.14}$$

we obtain result

$$\text{var}(Y_1) = \lambda_1, \tag{4.15}$$

λ being *eigenvalue* of covariance matrix $\text{cov}(\mathbf{X})$ and $\mathbf{c_1}$ its *eigenvector*. This process is repeated for obtaining the rest of principal components. Due to ordering according to the variance explained in components, dimensionality reduction is performed by using only a reduced number of components. In practice, most of the variance contained in the dataset is covered in only the first few components.

4.5.2 Dimension and Clustering

As was discussed throughout the thesis, we mostly deal with high dimensional data. With increasing of the dimension, many problems start to occur and intuition from the real world is no longer viable. To mention some of the consequences, proximity and basic distance between high dimensional data makes no longer sense. This means if we repeatedly compute distance (e.g. using L_2 norm) of 2 pairs of random points in high-dimensional space, obtained values will be nearly identical, no matter to "real"proximity of the points [22]. Data become sparse and geometry of space is counterintuitive. For uniformly distributed high-dimensional data can be proven, that they live mostly near the edge of the space, what is in opposite to everyday low-dimensional experience. However, real data describing our objects of interest usually lives in lower-dimensional spaces embedded to the original one. We may try to extract those important dimensions and throw away unimportant noise or redundant features. Various *dimension reduction* techniques are used to perform this task with different performance and usability range. During the reduction process, we would like to keep distances between points the same or at least proportional to the original space, what is impossible for most cases. Examining PCA as rotation, it preserves distance only until dimension reduction (throwing away components with lower variance).

Dealing with unsupervised data, it would be beneficial into separate them to some groups according to their common properties. While there exist many clustering methods usable even for high-dimensional data, in practice, clustering is employed after reduction of dimension. Reason for this sequentiality is mostly because of distances, which has to be computed. Well-known representative is *Hierarchical Clustering* (HC), where "clusters"starts as single data points and in every iteration are agglomerated with other close clusters up to the specific distance. This is repeated until we obtain one big cluster consisting of all data. Result of this method is a hierarchical structure of data proximity, which could be visualized in the form of *dendrogram* [13]s. Hierarchical Clustering is a widely used method for its good interpretability and possibility to select correct threshold distance from dendrogram. There is interesting equivalence between HC and *Persistent Homology* with zeroth Betti numbers, the tool of the way general *Topological Data Analysis* [23].

5 INTERCONNECTION OF ML WITH STATISTICAL PHYSICS

Connection of physics with ML can be dated to Shannon, by defining the entropy of information in his seminal paper [24] . This complex work was recognized even by physicists, resulting in publications by Jaynes [25], where he developed the principle of maximum entropy and also shown that statistical mechanics could be obtained from more general statistical (Bayesian) inference. Such a deep connection, reaching to the roots of both branches have many consequences and ways of demonstration. Surely, we cannot cover all of them and for a more general view on the topic, we suggest to study [13].

The presentation offered in this work is focusing to reach the goal (*Restricted Boltzmann Machine*) by revealing connections between *Lattice Models* and *Hopfield Networks* and their possibility to minimize energy.

5.1 Hopfield Networks

Neural Networks can be separated into two general categories according to connections between neurons. Until this point, we were describing networks with one-way connections, called *feedforward networks.* Now, our interest is moved to *feedback networks* with connections working in two-ways 5.1.

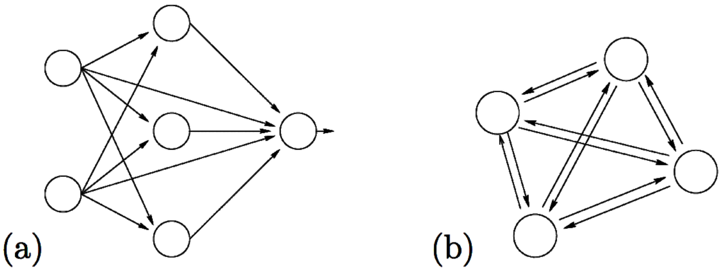

Fig. 5.1: a) Feedforward network with one-way connections; b) Feedback network with two-way connections. (taken from [11])

Hopfield Networks are feedback networks with fully interconnected neurons and symmetrical two-way connections between neurons. Hopfield Networks as one of the building stones of modern unsupervised learning are well known also in the community of statistical physics. They are used famously as *associative memory* (or biological memory) and as solvers for optimization problems. The principle of associative learning is captured by *Hebbian learning*, stating that if there are neurons positively correlated, in the network, the weights of their connection is increased. Imaginable example (taken from D.MacKay's book [11] is a functionality of idealized "brain" detecting smell of banana along with another stimulus (yellow color, taste, ...). Each stimulus is represented by a neuron in the network and weights between them are increasing during the learning process. Later, if only one of those correlated neurons is active, also other neurons are activated afterward. This procedure is known as *pattern completion* and could be used for various error-correction tasks or data reparation. Is obvious, that Hebbian learning is an unsupervised process, producing associative memory. In general, there are more variations and complications of Hopfield network, but basically, we may imagine neurons only with activations 1, -1 and activation functions as hard thresholds. Weights and other properties have the same meaning as in previously described NN. For complicated versions of HN we have to take into account order and time-dependency of activations updates, the stability of system, normalization, and capacity.

There is a scalar value associated with continuous HN, called energy E, which have identical shape as *spin glass* model of statistical physics. This energy function is obtained by generalization of Ising model energy function (putting J_{mn} and h_n non-constant)

It can be found that stable HN will converge to minimum of variational free energy, same as spins of Ising model (or spin glass) will align to arrangement with minimal energy (depending on pairwise interaction J_{mn} and field h_n). With the activity rule (output) $x_n = \tanh(a_n)$, HN is approximating probability distribution associated with this energy function, taking the form of Boltzmann distribution. Here we may observe intimate relation between most general spin glasses and Hopfield Networks, opening connections between ML and physics.

5.2 Boltzmann Machines

We have stated that Hopfield Network minimizes the variational free energy function and can be viewed as approximating probability distribution (of Boltzmann form) dependent on energy [11]. The core idea behind creating Boltzmann Machines was to implement this probability distribution into the network, forcing Hopfield Network to be stochastic. This is done by *Gibbs sampling*, Markov Chain Monte-Carlo technique for a sampling probability distribution. Implementing Gibbs sampling, we obtain activity rules of Boltzmann machine:

$$\begin{gathered}\text{set } x_i = +1 \text{ with probability } \frac{1}{1+e^{-2a_i}} \\ \text{else set } x_i = -1.\end{gathered} \tag{5.1}$$

Boltzmann Machine learns a probability distribution (of input data) by adjusting weights in a way such that the generative model ($P(\boldsymbol{x}|\boldsymbol{w})$) is well matched to this probability distribution. Principle of generative models and generating samples are closely discussed later. As it was mentioned in the case of Hopfield Network, Boltzmann Machine is viable to cover simple correlations between neurons (x_i and x_j of single inputs), but fails in covering higher-order correlations [11]. For real-world high-dimensional data, we may surely expect the presence of higher order correlations and thus the idea of Boltzmann Machines have to be improved. Solution to this task may be reached by introducing *latent variables*. General importance and usage of models with latent variables (also called hidden variables) is reviewed in work [13]. We focus on a single method using latent variables with an additional condition on connections between specific neurons.

5.3 Restricted Boltzmann Machine (RBM)

Restricted Boltzmann Machine (Figure 5.2) is a *generative model* (artificial neural network) with latent variables (hidden layer) where interaction (connections between neurons) are only between visible and hidden layer, but not between neurons inside one layer. The Energy function of this arrangement with binary units is of the form

$$E(\mathbf{v},\mathbf{h}) = -\sum_i a_i v_i - \sum_\mu b_\mu h_\mu - \sum_{i\mu} w_{i\mu} v_i h_\mu, \tag{5.2}$$

where $\mathbf{v}$ is configuration of visible units v_i, $\mathbf{h}$ is configuration of hidden units h_μ and $w_{i\mu}$ are weights of connections between them.

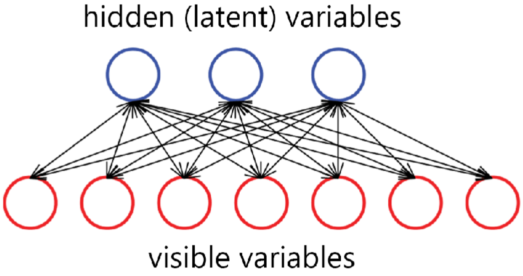

Fig. 5.2: Restricted Boltzmann Machine energy-based generative model with two-way connections and symmetric weights.

Introduction of latent variables to RBM is justified by so-called *Hubbard-Stratonovich transformation*, where visible units are decoupled by latent variables [13]. That means, complex interaction between visible units are now described by hidden units with possibly simpler structure. RBM is an energy-based model of form

$$p(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z'}, \tag{5.3}$$

where we recognize Boltzmann distribution. Step back to Hopfield Model (and so Ising) may be done by integrating out (or marginalize over) latent variables

$$p(\mathbf{v}) = \int \mathrm{d}\mathbf{h} p(\mathbf{v}, \mathbf{h}) = \frac{e^{-E_{\mathrm{HOP}}}}{Z}. \tag{5.4}$$

RBM is trained by using Maximum Likelihood Estimation procedure 4.1.2, where the cost function (the negative log-likelihood function) is minimized by SGD (see section 4.1.1) [13]. An effective implementation for training the RBM - *contrastive divergence*, was developed by Hinton [15]. Constrastive divergence is improved version of Gibbs sampling, allowing to lower number of necessary iterations for converging to equilibrium distribution. More details about the method and training process is presented in [13] or [12].

Last thing to mention is an existence of more versions of the RBM according to allowed values in visible and hidden layer. Most basic type, Bernouli-Bernouli RBM have binary values in both (visible and hidden units). Another type is Gaussian-Gaussian RBM with continuous values, and also mixed RBMs [13].

6 DIMENSIONALITY REDUCTION OF THE SPECTROSCOPIC DATA

To refresh motivation stated at the beginning of the chapter 3, we would like to exploit sparsity and redundancy of spectral data to lower dimension in a specific way, keeping only important information and dropping noise or non-unique information. The idealized method, which meets all mentioned conditions, could be used to substitute or complement work of trained spectroscopic specialist partially. The necessity of automatization in data processing arises especially for big datasets, which we measure on a daily basis. With the rapid improvement of instrumental capabilities, measurements (elemental mapping of the surface) could be done up to kHz repetition rate frequency, resulting in millions of spectra to process. Especially for heterogeneous samples, is not possible to inspect spectra-by-spectra manually, to select important parts of spectra or do other analyses. This chapter is presenting and interpreting results obtained by application of Restricted Boltzmann Machine method to dimension reduction of spectroscopic data. The functioning of RBM was studied in chapter 5, and it seem as (at least theoretically) ideal candidate for dimension reduction of spectroscopic data, taking into account mentioned properties. We may imagine "idealized"RBM model, exploiting the sparsity of data by inactive connection (low weights) of units from unimportant spectral regions and redundancy by a correspondence of all related visible units to the single hidden unit. If we were able to build such a model, the effective reduction of dimension would be realized without loss of any important spectral information.

Surely, such an idealized model is impracticable in reality, but even its approximation could be useful and able to compete with common approaches. Performance of RBM in dimension reduction will be evaluated in comparison with most common PCA model. PCA as an "evergreen"of data analysis, especially dimension reduction, is widely applied tool with great performance and interpretability. The reason for searching alternative method to standard PCA is linearity of the method and computational time. Since PCA is a linear model, it cannot cover complex non-linear dependencies in data. Keeping in mind strong non-linearities in spectra originating processes (due to material constants and matrix effect), a simple linear model must have significant limitations of use. While a comparison of 2 substantially distinct methods is difficult task, partially empirical evidence is included in the discussion of results.

Firstly, we try to reduce the dimension of a big spectroscopic dataset by both methods and later reconstruct "original"spectra with losing some information. As a figure of merit we use absolute value of the distance between original and reconstructed spectra, but also a ”visual”consideration of important structures. In the end, the possibility of generating new ”unseen”spectra by RBM is explored.

6.1 Samples, experiment, data

To demonstrate the performance of both algorithms, a unique dataset was designed and measured. This dataset is containing LIBS spectra from 138 samples in total separated to 12 categories according to dominant mineral composition (e.g. Hematite). The samples are OREAS certified soil samples cast into gypsum for more convenient handling. For each sample in the dataset, there are 5000 spectra available. In one class (e.g. Hematite - Fig. 6.1) we have n samples with similar chemical composition. However, specific concentrations are varying in some range, so the resulting spectra are different. In addition to this, 2/3 of samples are produced as a mixture of the selected sample with some random part of any other class in ratio 3:1.

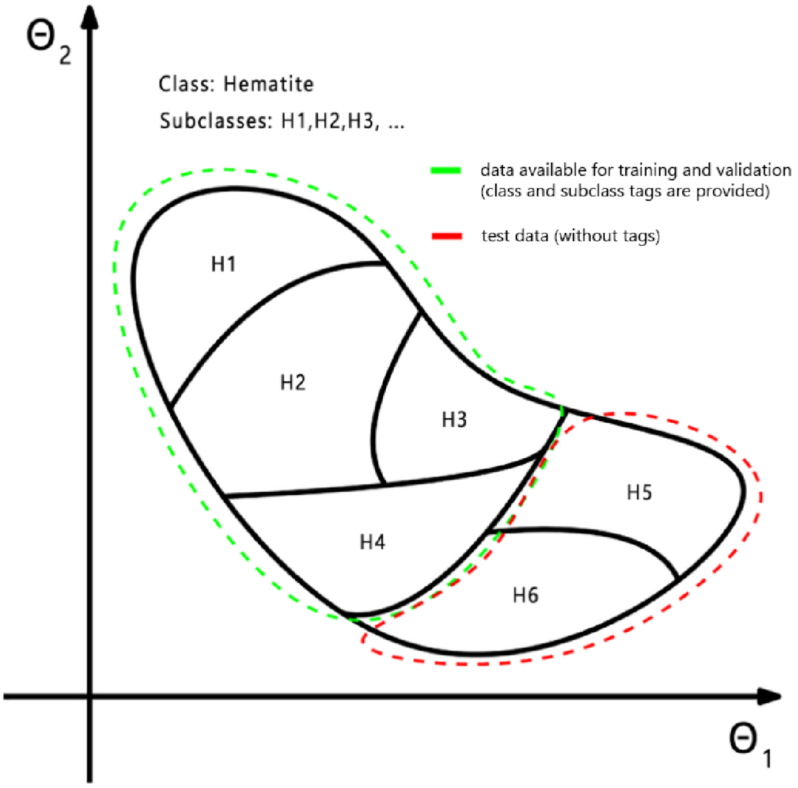
Θ2
Class: Hematite
Subclasses: H1,H2,H3, ...
🟩 data available for training and validation (class and subclass tags are provided)
🟥 test data (without tags)
H1
H2
H3
H4
H5
H6
Θ1

Fig. 6.1: Representation of a dataset class in artificial parametric space.

This dataset could be used for classification task after dividing to 2 subsets, training and test. We have selected 100 samples to serve for training a model and the remaining 38 was kept for test purpose. It should be noted that such dataset is unique in LIBS community (and also in spectroscopy in general) due to its size and number of samples. It is worth to mention that this big dataset is difficult to handle with and effective dimension reduction could provide significant improvement in posterior processing of data.

Dataset is freely accessible online with more detailed description at repository webpage [26]. The exact composition and categorical information cannot be provided currently (due to running contest using similar data), but will be published shortly at the mentioned webpage.

6.1.1 Measurement

Measurement of the samples was provided by Sci-Trace instrument (Atomtrace, CZ), modular system suitable for complex LIBS analysis. Sci-Trase was equipped by a special interaction chamber, described in [27]. As a light source, we have used Q-Switch FPSS Nd:YAG laser Solar LQ-529a, with the wavelength 532 nm and pulse duration 10 ns. Plasma radiation was collected by BK7 plano-convex (Thorlabs), focal length 75 mm, AR 350-700. The Echelle spectrometer Andor Mechelle 5000 (resolution $\lambda/\Delta\lambda = 5000$) was used to detect spectra with modified software, allowing higher speed of measurement. Rest of components were identical or similar to the mentioned setup [27].

Energy of laser was 20 mJ and gate delay was 1 μs. Exposition time was kept on default setting 50 μs. Map of 75x75 points (spacing 30 μm) was created on surface of each sample and spectrum was obtained from each point (one shot - one point). This resulted in 5625 spectra per sample, but first 625 was deleted to ensure good stability of the system.

6.1.2 Application of the RBM

For an initial demonstration of the method, the RBM model was trained on part of the original dataset. Used dataset consisted of spectra from 30 samples divided into 2 classes equally. Thus the partially similar structure of spectra belonging to one class was guaranteed. There was used 1000 spectra per sample, 30000 spectra in total. Original spectral dimension was 10000 wavelength values. Each spectrum was normalized by unit vector normalization (UVN).

For the computation, scripts with code in R and Python languages were created and are available online [26].

Gaussian-Gaussian RBM model was designed with a single hidden layer, consisting of 100 neurons. Learning rate was set to 0.01 and spectra were fed to model by minibatches (100 spectra each). Visible states were sampled with zero mean and

$\sigma = 0.7$. After basic optimization, it was deduced that 4 epochs are enough (epochs are number of the repetitive pass of all data).

After the training process, random spectra were selected and reconstructed by model. We have obtained similar results for all tested spectra and example is shown in figure 6.2. Progress of training is plotted on figure 6.3 a), and schematic diagram of reconstruction on figure 6.3 b).

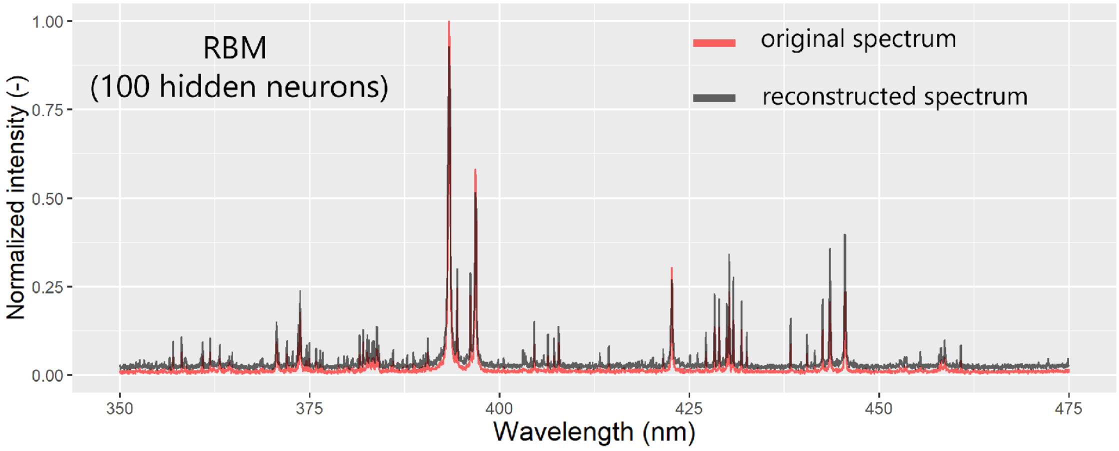

Fig. 6.2: Randomly selected spectra, reconstructed by using RBM model with 100 hidden neurons.

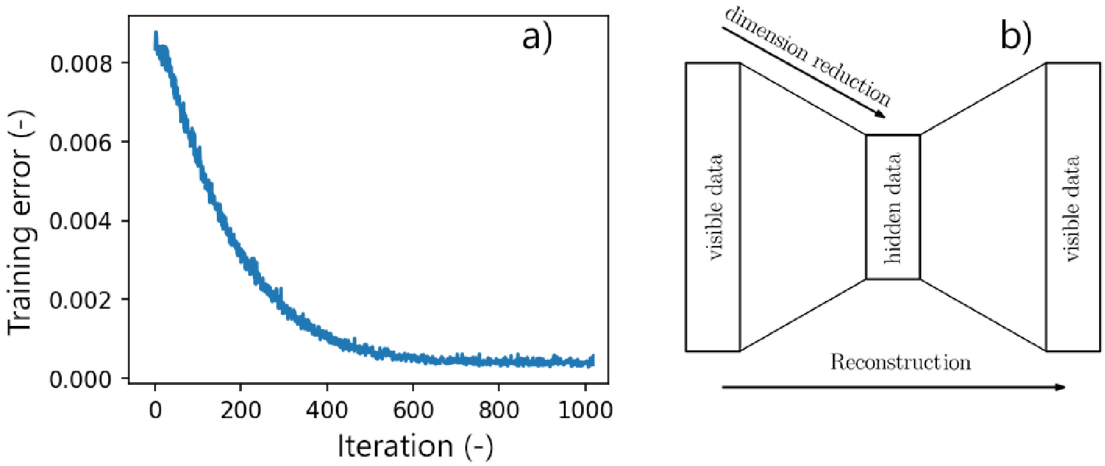

Fig. 6.3: a) Training error during learning procedure of RBM, b) Diagram of reconstruction procedure.

Reconstruction of the spectra could be considered as successful. General features as line positions and ratios are well preserved, but the intensity is slightly modified.

Also a "background"or bias is higher than in original spectra, however, this could be easily treated by performing UVN normalization on the result. We may conclude, that the dimension of the dataset was effectively reduced and possible reconstruction of spectra provide satisfying results.

Comparison of performance

We have found that RBM is suitable for dimension reduction of spectroscopic data, but it is desirable to compare the method with something well known. For the comparison, we use Principal Component Analysis method, a linear method described in section 4.5.1.

Slightly different data were used for this comparison, reduced on size, because of the high computational cost of PCA algorithm. This time, spectra from 100 samples were used (100 spectra for each sample, 10000 in total), divided to 12 categories. Such parameters imply an increased complexity of the dataset.

PCA was employed on the dataset and 6 different numbers of the components were kept to provide the reduction of the dimension. This was done in a similar way to article I have published earlier [28]. Thus dimension of the original data (10000 wavelength values) was projected to lower dimension (5 - 30) and later reconstructed back.

RBM was trained 6 times with a different number of hidden neurons (5 - 30). Learning rate was 0.01 for all cases, batch size 100, and the number of epochs was 5 (except the model with only 5 neurons, where 10 epochs were used). After the training, all data were reconstructed in the same way as was shown before.

As the figure-of-merit, we have selected the absolute distance (L1 norm) of the reconstructed spectrum to the original one at each point (wavelength). To obtain representative values for whole dataset, one spectrum was selected for each sample in dataset and evaluated. So, the distance of 100 reconstructed spectra to their original ones was computed. Later, the mean value at each point was taken and the result is shown in the following figures, for RBM 6.4 and for PCA 6.5.

While exploring the performance of the RBM on this task, we may observe relatively higher distance (error) of reconstructed spectra to the original one in comparison to PCA results. However, the position of the peaks in the RBM result is corresponding to the lines frequently present in the dataset. As we have stated, reconstruction error in the intensity of the spectral line is not a big problem, which could be easily treated by normalization. Red line in the figure 6.4 is showing the biggest error observed in the corresponding PCA model (model with the same dimension reduction). Even while the PCA reconstruction error was lower, there are several weaknesses

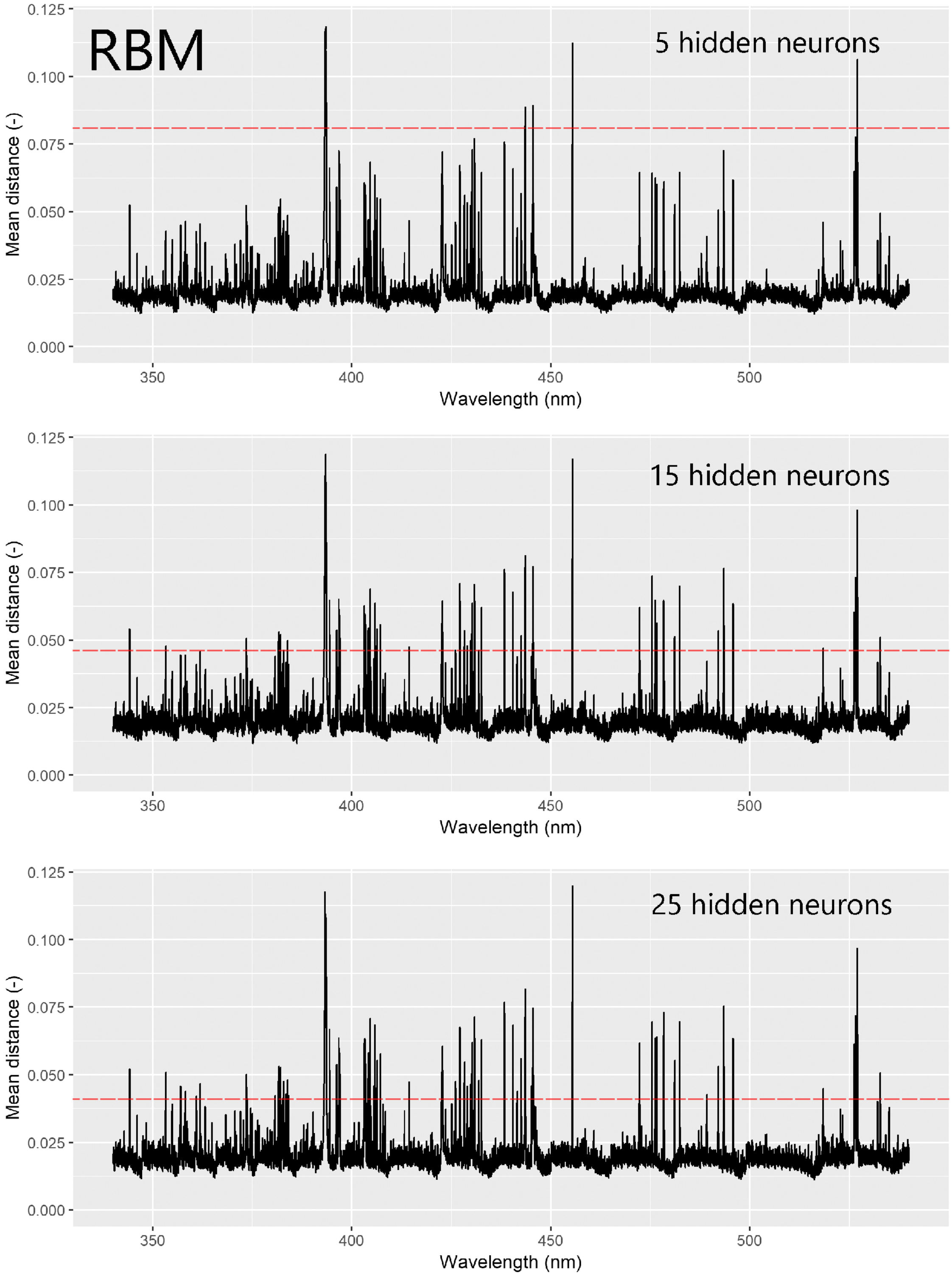

Fig. 6.4: Mean absolute distance between original and reconstructed spectra by RBM model.

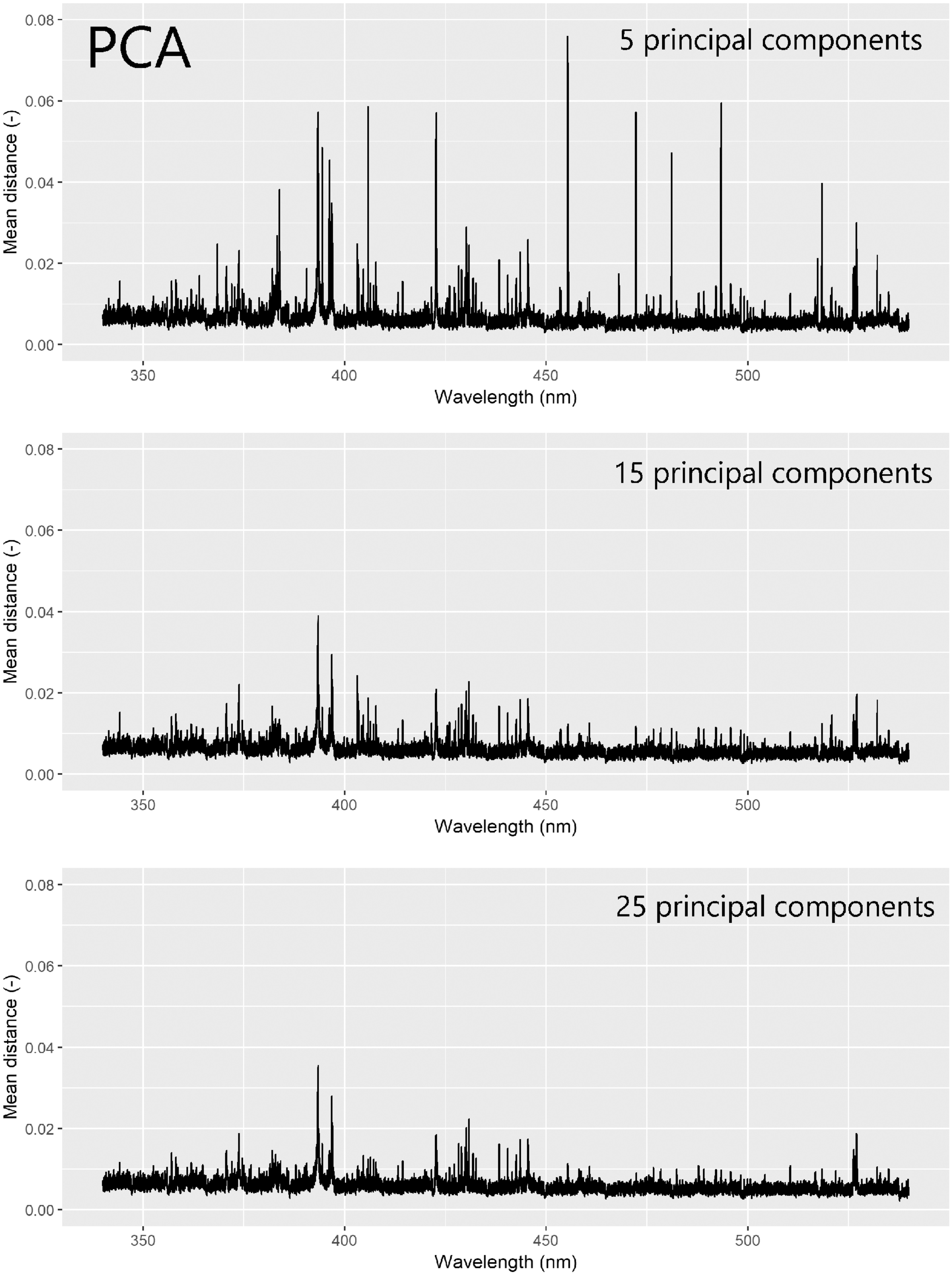

Fig. 6.5: Mean absolute distance between original and reconstructed spectra by PCA model.

of the PCA approach. The first big disadvantage of PCA is extensive computational cost. This makes the method unusable for really ”big datasets”. The problem is based in the PCA algorithm, where we need to have all data at once. Thus, we are limited with memory of the computer and some statistical reduction of objects is necessary. In the case of RBM, we learn model by minibatches, so huge datasets could be easily used for training. Even more, it is desired to use as many as possible spectra for building the RBM (this is generally valid for all neural network models). It also means that this comparison was bit unfair to RBM, while smaller amount of the data was used. Also, approximate computational time was differing rapidly, for PCA it was cca 3 hours (in R) and for RBM, a single model took around 10 minutes (in Python). Only basic accessible processor was used for the computation (usage of the graphics card would make RBM training even faster).

This is a good starting point for future research and improvement of RBM. As it was mentioned, the bias of the RBM reconstruction could be improved by further normalization, Dimension could be reduced more effectively using deep-structure, where more hidden layers are introduced to the model. Such a structure is also suitable for direct classification of the data, just by replacing activation functions in the final layer.

Generating unseen spectra

We have defined the RBM as an energy-based generative model. The possibility to generate new spectra follows from sampling a probability distribution, which was learned by the model during the training process. The sampling is provided by Gibbs sampling technique (a detailed description of the technique is provided in [11]), in a simplified way just transforming visible units to hidden and back to the visible, given weights and probabilities. We have explored this possibility using the RBM model trained on the first mentioned dataset (figure 6.2).

In the start of a generative process, a vector of only zero values is given to the input layer of the pre-trained network. Then a specific number of Gibbs sampling steps is performed between input and hidden layer. Finally, we may inspect the result in visible layer with probability given by the spectra examples from the training process. However, this new spectrum is not a copy of any original training spectra, It is a completely new unseen spectrum. Here we show (Fig. 6.6) a randomly generated spectrum corresponding to a class presented in the figure 6.2.

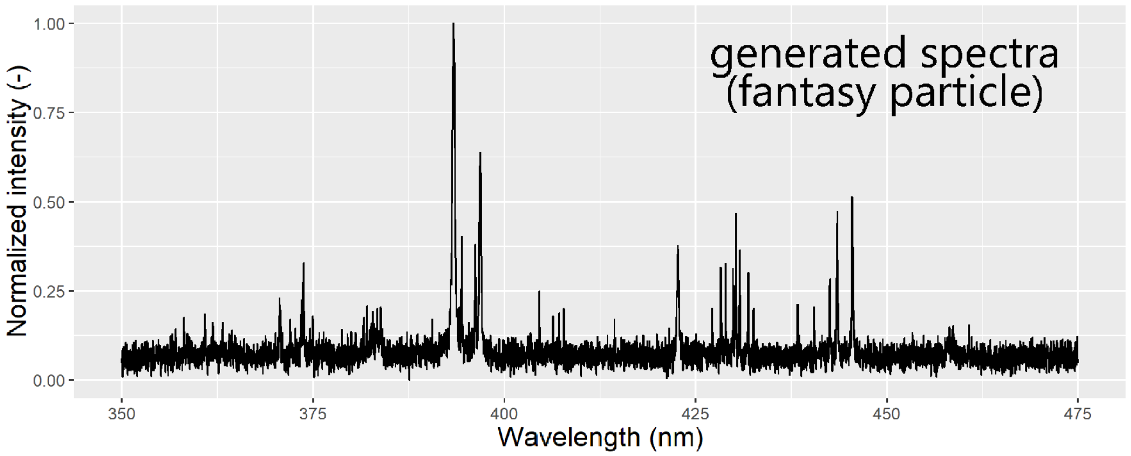

Fig. 6.6: Generated spectrum from the RBM model learned on 30000 spectra divided to 2 categories.

Those generated "samples"of data are also called *fantasy particles* and they are valuable for the inspection of the model, repairing of corrupted data, and also for the cases where not many inputs are available.

6.1.3 Further plans

An extensive exploration of the application of the RBM model to the processing of spectroscopic data was presented, but obviously there are many improvements possible. As was already mentioned, building a deeper structure of the network (*Deep Boltzmann Machine*, consisting of more hidden layers), implementing supervised features, and visualization of the network features would improve the whole methodology.

7 CONCLUSION

This work aimed to reveal connections between statistical physics and machine learning (ML), which at first sight may seem like completely distinct branches of science. I have started with developing essential tools of statistical physics from ab initio approach. After obtaining the partition function of a Boltzmann distribution, we have studied the Ising model of spins in 2D. Mean field approach, and later, more general variational free energy approach helped to show the behavior of the Ising model. On the Ising model example was shown, that by minimizing the variational free energy, a Kullback-Leibler divergence is lowered. The Kullback-Leibler divergence measures a dissimilarity of two probability distributions and thus could be used to describe (and ensure) learning process of ML algorithms. A strong connection between machine learning and statistical physics appeared after observing equivalence between the Ising model and Hopfield network, a basic model of unsupervised machine learning. Also, an introduction to the machine learning, in general, was provided. Basic principles originated in physics, but used in ML were mentioned (symmetries, locality, free energy, ...). From the Hopfield network, by adding stochasticity and restricting specific connections, Restricted Boltzmann Machine model was derived. Introduction of latent variables as one of RBM property was discussed to deal with higher-order correlations in the data.

A considerable part of the thesis studied general properties of spectroscopic data as sparsity and redundancy in spectral lines. Spectra from Laser-Induced Breakdown Spectroscopy (LIBS) were selected as representants for spectroscopic data, and physical processes standing behind their creation were described (in appendix). According to mentioned properties, generally valid for most types of spectroscopic data, RBM method was deduced as a good candidate to deal with sparsity and redundancy of the data.

In the practical part, an extensive unique spectra dataset was created. The dataset consisted of 138 samples, where for each sample, 5000 spectra were measured. Samples were related together partially, forming 12 distinct categories. Such a dataset is suitable for challenging classification tasks, but due to the excessive number of measurements, the use of advanced classification algorithms is limited. This problem was treated by the effective reduction of dimension, using RBM. Performance of the RBM was evaluated by comparison to a commonly used method - Principal Component Analysis (PCA). Comparison between methods was done on a smaller part of the original dataset, due to computational requirements of PCA. The dimension of the data was reduced from original 10000 values to a way lower dimension (5-30). Each spectrum was projected to this lower dimension and later reconstructed. In some aspects, the PCA reached better performance (reconstruction error, interpretability of reduced dimension, ...) but failed in other aspects, where RBM dominated (learning time, extensibility of the method, ...). RBM was proved to be suitable for dimension reduction of spectroscopic data, but further exploration of the method is desired.

Beside of dimension reduction, RBM offers a possibility to generate new data (spectra) from the learned probability distribution of original data. This feature has potential applications in repairing of incomplete spectra or transfer of libraries between spectrometers. Generation of spectra was demonstrated by the RBM model learned on 30000 spectra. Results were discussed and generally evaluated. To provide all computations, several scripts with code were created in R and Python language.

The objectives of the thesis were accomplished, but there is still space to improve the whole methodology. Since handling with such a huge amount of data is highly non-trivial and computationally-expensive, optimization of the method is difficult to provide and will be the goal of my continuing research.

LIST OF APPENDICES

A BACKPROPAGATION METHOD

For learning a Neural Network we still rely on "general truths"described in previous sections discussing ML as a whole. Cost function still has to be selected and minimized using the gradient descent method. Thus, backpropagation is just a method to overcome the computational difficulty of computing gradients in the complex and interconnected parametric model as a deep neural network is. In this section, we reveal basics of the backpropagation algorithm.

In the start, we have to select proper cost function (called also loss function or energy function). For regression task and continuous data, we use basic L_2 norm (also possibly L_1 norm)

$$E(\boldsymbol{w}) = \frac{1}{N} \sum_i (y_i - \hat{y}_i(\boldsymbol{w})), \tag{A.1}$$

where y_i is a real value (a category in classification) of the data and $\hat{y}_i(\boldsymbol{w})$ is prediction dependent on parameters $\boldsymbol{w}$. Classification tasks and categorical data requires different treatment and cross-entropy is most commonly used (see the section about logistic regression). In case of data separated to more than two categories $y \in \{0, 1, 2, ..., M-1\}$, strategy one-versus-all is used defining $y_{im} = 1$ if $y_i = m$, otherwise $y_{im} = 0$. Then we have *categorical cross-entropy*

$$E(\boldsymbol{w}) = -\sum_{i=1}^{n} \sum_{m=0}^{M-1} y_{im} \log \hat{y}_{im}(\boldsymbol{w}) + (1 - y_{im} \log[1 - \hat{y}_{im}(\boldsymbol{w})]). \tag{A.2}$$

Actual learning starts with feeding input data vector (representing one sample) to a deep neural network with parameters (weights) selected randomly in the range $-1 \leq w_i \geq 1$. Because we are dealing with supervised learning, we know what desired output is (specific value or category). After passing data through the network, "comparison"of output with desired output is provided by evaluating corresponding cost function. At this point, we try to lower cost function but the explicit computation of gradient over all parameters would be extremely costly. Taking into account the structure of a network, backpropagation is using rules of partial differentiation for the cost function. Before we define proper notation and provide a formal derivation of backpropagation, we sketch the procedure intuitively.

We want to scope how sensitive is cost function to small changes in every parameter of the model, final activation function depends on. After feeding data sample to network and passing it to output, the error of the final layer is easily computed by definition of the cost function (comparing to desired value). At this point, we would like to make gradient descent step to lower cost function, so we need to compute the gradient of cost function dependent on model parameters. Error in final layer could be propagated through connections to every neuron in the network and related to partial derivatives of corresponding parameters. Obtaining all partial derivatives, we may construct the gradient and lower the cost function.

Formally, we have a neural network consisting of layers $l = \{1, 2, ..., L\}$, connection weight between k-th neuron of $l-1$ layer with j-th neuron of layer l is w^l_{jk}. The bias of j-th neuron of l-th layer is b^l_j. Finally, the activation function of neuron in the l-th layer may be expressed as a weighted sum of activations in previous layer $l-1$ passed through activation function σ

$$a^l_j = \sigma\left(\sum_k w^l_{jk} a^{l-1}_k + b^l_j\right) = \sigma(z^l_j). \tag{A.3}$$

Denoting the sum as z^l_j, activation a^l_j is a function of z^l_j, which is a further function of bias b^l_j, weight w^l_{jk} and activation of previous layer a^{l-1}_k. In the output layer, we may compare real output to desired one, by computing cost function C. However, it is obvious that this value of the cost function C is also dependent on activations of all previous layers. What we want to seek is, how the cost function will change upon a small change of any dependent parameter. It is useful to denote generalized error (change of cost function with respect to weighted sum z^l_j) as

$$\delta^l_j = \frac{\partial C}{\partial z^l_j} = \frac{\partial C}{\partial a^l_j}\sigma'(z^l_j), \tag{A.4}$$

where second equality follows from simple chain rule. Similarly, we may construct a change of cost function with respect to bias

$$\frac{\partial C}{\partial b^l_j} = \frac{\partial C}{\partial z^l_j}\frac{\partial z^l_j}{\partial b^l_j} = \delta^l_j, \tag{A.5}$$

provided by $\partial z^l_j / \partial b^l_j = 1$. There is a dependence on weight for cost function left. From previous definitions follows

$$\frac{\partial C}{\partial w^l_{jk}} = \frac{\partial C}{\partial z^l_j}\frac{\partial z^l_j}{\partial w^l_{jk}} = \delta^l_j a^{l-1}_k. \tag{A.6}$$

The error can be propagated to layer l inside the network, while we know that it depends on activations in following layer $l+1$ and using chain rule we derive

$$\frac{\partial C}{\partial z^l_j} = \sum_k \frac{\partial C}{\partial z^{l+1}_j}\frac{\partial z^{l+1}_j}{\partial z^l_j} = \sum_k \delta^{l+1}_k \frac{\partial z^{l+1}_k}{\partial z^l_j} = \sum_k \delta^{l+1}_k w^{l+1}_{kj}\sigma'(z^l_j). \tag{A.7}$$

Using Equations A.4-A.7 we may backpropagate error to each neuron of the deep network. With knowledge of errors, the gradient of cost function w.r.t. all model parameters ($\partial C/\partial b^l_j$, $\partial C/\partial w^l_{jk}$) is easily computed and gradient descent may be used for learning. [29]

B LIBS: PHYSICS, INSTRUMENTATION AND APPLICATIONS

Here we present fundamentals of Laser-Induced Breakdown Spectroscopy, from laser-matter interaction, through plasma processes, emission, the shape of spectral lines to instrumentation and applications. Emphasis is placed on the physical description of processes responsible for spectra shape and structure. Basic plasma diagnosis methods and approaches are also mentioned.

B.1 Laser-Matter Interaction

In this section, we discuss basic properties of interactions between high power coherent light beam (laser) with a solid state of matter. Since this topic is so extensive and non-trivial, we provide just a brief explanation of the most important facts. Due to strong non-linearities in many material parameters, surrounding environment and wavelength dependence of this process, we have to restrict our focus just to nanosecond (and slightly covered femtosecond) laser sources with energies tens to hundreds mJ, focused to spot of radius 10-100 μm. Mentioned parameters of laser radiation are forming a beam of sufficient *flux density* (or irradiance) (GW/cm^2), well above examined solid state material breakdown threshold (threshold of the gaseous or liquid matter is generally higher than a solid state). In case of interaction such beam with material, laser-induced breakdown takes place and there is so-called *ablation* of material. Theories or models describing this process for gasses are *multiphoton ionization* (dominating at low pressures) and *collisional cascade ionization* (higher pressures) [30]. However, in a solid structure, one has to take into account more complex threshold dependence. Generation of plasma in solid structure is delayed due to phonon excitations of the lattice and its transfer to heat. For example in metals, conduction electrons receive energy through *inverse Bremsstrahlung* effect and release it to a phonon system.

In the ablation process, there is a small amount of material (up to tens of nanograms) transformed to plasma and some amount ejected around crater border. If the irradiance is below threshold value, material from the bulk sample is not removed. However, there can be some minor desorption of individual atoms from the surface.

Considering the case of the nanosecond laser pulse, we assume that energy is absorbed just by the surface of the sample [31] and due to diffusion there is heating of bulk material. In solids, the penetration depth of radiation δ_p with definition as reciprocal attenuation coefficient α (Beer-Lambert law), $\alpha = 4\pi n''/\lambda$, where n'' is imaginary part of the refractive index of the material. The intensity of laser beam inside solid material is described as

$$I(z) = (1-R)I_0e^{-az}, \tag{B.1}$$

where z is the distance from surface on an axis parallel to the beam, I_0 is the laser intensity and R the reflectivity of the surface. For the visible wavelengths, the coefficient α is smaller in comparison to heat diffusion length $d_{dif} \approx (\tau\kappa)^{1/2}$, with τ being laser pulse duration and κ thermal diffusivity. $\kappa = K/\rho C$, where K is thermal conductivity, ρ density and C specific heat per unit mass. [32] Through a lasting supply of energy, the material is melted and further evaporated. At this point (in simplified sense), the system reached the gas phase and its treatment was described above. With rising temperature, also pressure is increasing and as it was already mentioned, part of the liquid matter is ejected to the open space. Schematic representation of this process is plotted on Figure B.1.

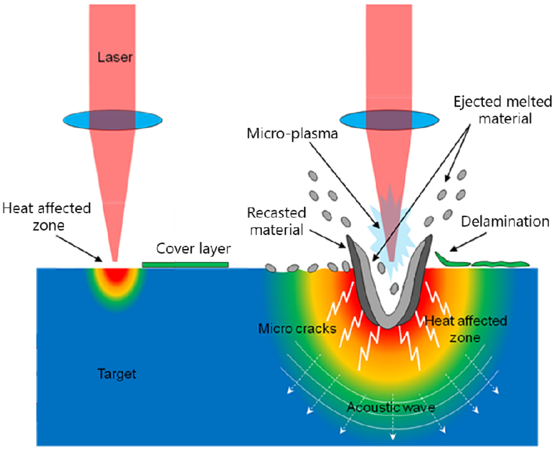

Fig. B.1: Effect of high power pulsed (ns) laser on solid-state target. (taken and edited from [33])

In contrast with nanosecond ablation, the nature of femtosecond laser ablation process is fundamentally different. Diffusion length in this regime becomes comparable to the absorption length. In such a short pulse with terawatt power, affected volume of material could be rapidly ionized and ejected from the surface by *Coulomb explosion*. [34] Comparison of ablation craters produced by both regimes is shown at Figure B.2.

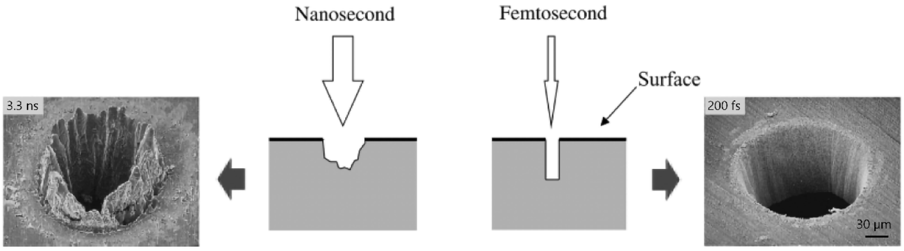

Fig. B.2: Comparison of ablation craters produced by nanosecond and femtosecond laser pulse. (taken and edited from [35])

To add even more complexity to the problem, there are more laser beam profiles used. The most common *Gaussian profile* has an advantage in manipulation (focusing, guidance) of the beam, while *flat top* profile has better-defined boundary of the crater and so spatial and axial resolution. Besides of the diffraction limit for smallest possible crater diameter, the character of a laser beam (pulse time, wavelength, profile) is also contributing to total spot size and so to the resolution of the method. A more general and extensive study of laser ablation dependence on various parameters and its properties is presented in work [36].

B.2 Laser Induced Plasma (LIP)

In the previous section, we have reviewed the basics of LIP formation mechanism. However, investigation of LIP properties could be started also by a different point of view, ignoring the process of its creation and focusing just on its behavior with given plasma parameters. While plasma is considered as a statistical system, important parameters are temperature, electron density, volume, and pressure. With some knowledge about plasma composition and surrounding environment, its dynamical evolution and properties are almost completely determined. Again, due to diverse external conditions, it is not possible to expect analytic solution simply covering the general case, but there are many eligible models describing plasma evolution (volume expansion, temperature evolution, interaction of internal particles and more).

Plasma parameters may be determined experimentally, but there are several difficulties according to fast dynamics. To shortly review experimental techniques suitable for plasma diagnosis: electron density may be measured by *shadowgraphy* or *Schlieren method* or indirectly from emission lines, the electron temperature is determined indirectly from emission lines (this part will be covered extensively in following sections), the volume could be taken from fast imaging techniques.

Now, the most important aspects of LIP, necessary for further considerations, will be described:

- *plasma expansion*
  As it was mentioned earlier, expansion of plasma to background gas is complex phenomenon mostly due to various material and environmental characteristics. Consequently, its theoretical description and simulation are going hand to hand with experimental observations. There are more valid approaches on how to describe this temporal evolution of plasma plume, ranging from fluid-dynamic models to Monte Carlo methods. Work of M. Capitelli et al. **??** offers a review of suitable methods and deeper theoretical description of LIP expansion. To summarize most important observations, LIP expansion can reach relatively high velocities (up to $10^4 m/s$) accompanied by electron density and temperature lowering. The lifetime of such plasma is approximately 2-3 μs. Maybe the most general and interpretable approach to plasma expansion modeling is using Navier-Stokes equations (Fluid-dynamics) for a multispecies gas with using the symmetry of a problem (taking into account effects as viscosity and diffusion). Even simplified variation of fluid-dynamic code (Euler equations) could be used if it is not necessary to cover the full range of plasma lifetime and pressure of the surrounding environment is low.
- *optical depth* In the further section we will discuss emission of plasma, where absorption of light by plasma is not intrinsically taken into account. If we want to fit our models to match with reality, plasmas have to be thought of as optically thick. Fortunately, this fact doesn't imply that our simplified theory about plasma emission is not correct, but it has to be fixed sometimes to match real-world situations. The most reliable consequence of optically thick plasma existence is self-absorption of lines. This effect is observed for *resonance lines*, which are lines corresponding to the transition of some excited state i to ground state of an atom or ion. A self-absorbed line has different intensity and shape in comparison to the basic theory of radiation. An extreme case of self-absorption is called self-reversion or splitting of a spectral line. Such lines are not suitable for quantitative analysis or plasma diagnostics. There was extensive work done to include optical depth into models describing plasma emission (see publication [37]).
  Therefore, the lines might be self-absorbed in the case of an optically thick plasma. In addition, temperature inhomogeneities exist along the line of sight of observation leading to self-reversed lines
- *LTE condition*
  To ensure complete thermodynamic equilibrium, the optical depth of plasma must be large for all wavelengths, thus radiation cannot escape. Unfortunately,

due to a rapid expansion of laser-induced plasma and its complex dynamics with nontrivial internal processes, temperature gradients are present and conditions for thermal equilibrium cannot be fulfilled. However, Planck's law could be valid at least locally. In such case a *Local Thermodynamic Equilibrium* takes place, satisfying specific conditions. The existence of LTE is guaranteed when radiation processes are negligible in comparison to collision processes. Also, collision processes should be balanced by its converse [38]. Under LTE, populations of atomic states are still described by Boltzmann distribution and Saha equation is also valid. To ensure the presence of LTE in plasma, McWhirter criterion for electron density has to be fulfilled, which is

$$n_e \geq 1.6 \cdot 10^12\sqrt{T}(\Delta E)^3 \quad (\mathrm{cm}^{-3}), \tag{B.2}$$

n_e being electron density. Limits and justifications of this criterion are studied in publication [39].

Plasmas reaching beyond LTE model could be described by *Coronal model* or *Collisional-radiative model.* First mentioned is dealing with plasmas of low electron density, where the optical thickness is small for all wavelengths and the collisional rate is small with respect to the spontaneous decay rate. Coronal model is usually of no interest in LIBS or LIP considerations. The second one, Collisional-radiative (C-R) model forms sort of transition between the Coronal model and LTE. This model takes into account every collisional process possible, but only two radiative (spontaneous decay and radiative recombination). In C-R model, the population of each energy level is described by a differential equation consisting of all transition processes. [40] Simulations of LIP using C-R model is providing many important insights about plasma dynamics and spectra simulations, but sometimes could be difficult to compute. There are as many differential equations, as is the number of accessible energy levels. Due to this complication, close energy levels could be grouped together and form only a few level systems as was suggested by Gornushkin in work [41].

B.3 LIP Emission

As it was mentioned before, LIP radiates light during its lifetime. Those photons originating from various processes are carrying much useful information about plasma itself. Plasma emission can be divided into radiation of free electrons and bounded electrons. While free electrons usually have a continuous spectrum, bounded electrons have discrete energy states and so the spectrum is consisting of well-known peaks (spectral lines). In spectroscopic measurement, we obtain complex data where both processes are present with some additional instrument noise. Treatment of data is considered in section B.4.1.

Time evolution of plasma is described in Figure B.3. In the first nanoseconds after the end of a laser pulse, the temperature of the plasma is highest resulting in many collisions and highest radiation intensity. In this region, most of the atoms are ionized and free electrons radiation dominates over bounded electrons. This phenomenon is also caused by the optical density of plasma and shielding by dense electron gas. There are more non-trivial effects playing the role as *ionization potential depression*, resulting in the non-existence of suitable energy levels for commonly observable transitions.

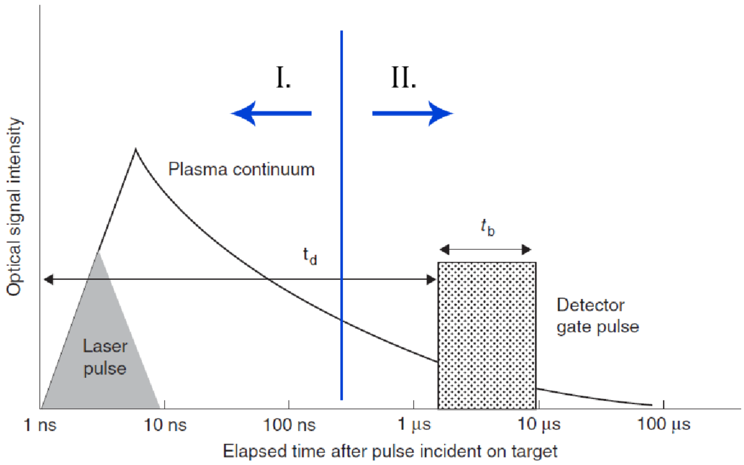

Fig. B.3: Time evolution of Laser-Induced Plasma, heuristically separated to 2 regions. Region I is representing time-range with the dominance of free electron radiation and hardly recognizable lines, while region II is usually suitable for spectra measurement and obtaining elemental information about the sample. (taken and edited from [35])

It is well known that accelerated charged particles emit light. Those accelerations may occur due to the presence of external fields as a magnetic or electric field (always present inside plasmas). According to [38], most representative mechanisms of free electron radiation are *cyclotron radiation* and *Bremsstrahlung*.

Cyclotron radiation is an effect taking place for example when a particle is moving perpendicular to external magnetic field B ($\vec{v} \cdot \vec{B} = 0$). If we investigate motion equation (Lorentz) for this event, we shall see that the trajectory of a particle is spiral. Charged particles experiencing acceleration emits a directive beam of light with a specific frequency. Relativistic variation to cyclotron radiation (using electrons as particles) is called *synchrotron radiation* and has a broad range of use in spectroscopy, computed tomography (as a source of light with good coherence and frequency range) and more. While the frequency range of such emitted light depends on broadening mechanism as Doppler or relativistic, in plasma there are even more significant broadening effects due to strongly varying magnetic field B. Thus, on the contrary to the expected result of cyclotron radiation as series of narrow spectral lines. Bremsstrahlung is a German word for "braking radiation"describing collisional effects between the electron and charged particles. In this interaction, we recognize 2 different situations:

- *free-free transition*
  This occurs when electron after its collision with ion remains unbounded, but its energy and direction of the motion is generally changed. The total energy of electron after the collision is greater than zero. It is easy to see that this type of collision results in a continuous spectrum, while there is not any quantization of free electron energy levels.
- *free-bound transition (recombination)*
  In this interaction, colliding electron is bound to ion and rest of electron energy can be emitted as a photon or transferred to heat. The total energy of electron will be lower than zero. Thus there are both types, radiative and non-radiative free-bound transitions. Regardless of bounded electron states are quantized, the spectrum is again continuous, because there is not any restriction on the initial state of an electron.

It should be noted that for our purpose, electron-electron collisions may be omitted since their contribution to radiation is negligible for non-relativistic plasma. In LIBS practice, the continuum of free electrons is rarely used and it is considered as some parasitic effect. However, it may contain some useful information about temperature and electron density of the plasma and further possibilities are still a matter of research.

From the spectroscopic point of view, bounded electron radiation of plasma is the most important effect. Characteristic spectral lines are products of electron transition between two energy levels in atom or ion. Wavelength λ (related to frequency as $\nu = c/\lambda$) is dependent on the energy difference between those levels as $h\nu_{ij} = E_i - E_j$, h being Planck's constant, ν_{ij} photon frequency and E_i (E_j) energy of upper (lower) level of transition. Note that i will be used as index of higher level and j as lower.

In plasma, we are dealing with a huge number of atoms and ions experiencing various transitions. As it was mentioned before, the probability of finding an atom (or ion) at a specific energy level is guided by Boltzmann distribution. Then, number density of particles $n(E)$ (atoms or ions of one specific chemical element) in state with energy E is proportional to Boltzmann factor multiplied by *degeneration factor* g

$$n(E) \propto g e^{-\beta E}. \tag{B.3}$$

To change the proportionality symbol in the equation to equality, the partition function is needed. However, there is another trick or possibility of how to treat this problem. For the ratio of two populations corresponding to upper and lower level of transition (for atoms/ions representing specific element inside plasma) we have

$$\frac{n(E_i)}{n(E_j)} = \frac{g_i \exp\left(-\beta E_i\right)}{g_j \exp\left(-\beta E_j\right)}. \tag{B.4}$$

Let's introduce coefficients A_{ij}, B_{ij} and B_{ji}, called *Einstein's coefficients* representing probabilities of specific transitions. A_{ij} is a probability that spontaneous transition from level i to j take place for a unit time. Let's also note $\rho(\nu)$ as the energy density of electromagnetic radiation acting on the particle (dependent on frequency of light ν). Then probability (per unit time) of absorption a photon by a particle (= atom or ion) is noted as $B_{ji}\rho(\nu_{ij})$. The last coefficient stays for stimulated emission, where the probability of this event is $B_{ij}\rho(\nu_{ij})$. In thermal equilibrium, the energy density of radiation is given by blackbody radiation as

$$\rho(\nu) = \frac{8\pi h\nu^3}{[\exp(h\nu/T) - 1]c^3}. \tag{B.5}$$

To satisfy equilibrium condition, rate of atoms making the transition from level i to j has to be equal to the rate of vice versa transitions. Then we obtain an equation for this *detailed balance principle*

$$(A_{ij} + B_{ij}\rho)n_i = B_{ji}\rho n_j. \tag{B.6}$$

Rearranging the previous equation to obtain dependency for ρ we have

$$\rho = \frac{A_{ij}}{(N_j/N_i)B_{ji} - B_{ij}} = \frac{A_{ij}}{(g_j/g_i)\exp(h\nu_{ij}/T)B_{ji} - B_{ij}}. \tag{B.7}$$

If the principle of detailed balance has to be fulfilled (for all temperatures) using the relationship for blackbody radiation, it is possible if and only if

$$A_{ij} = \frac{8\pi h\nu_{ij}^3}{c^3}B_{ij} \tag{B.8}$$

and

$$g_i B_{ij} = g_j B_{ji}. \tag{B.9}$$

While those coefficients are related to atoms itself, Equations B.8 and B.9 has to hold independently on thermal equilibrium.

It is worth to note that in comparison with a high number of accessible energy levels inside atoms or ions, transitions may occur only between specific levels guided by selection rules. Restrictions produced by selection rule (total angular momentum: $\triangle J = \pm 1$) are clearly visible from Grotrian diagram (Figure B.4), where only allowed transitions are between adjacent columns.

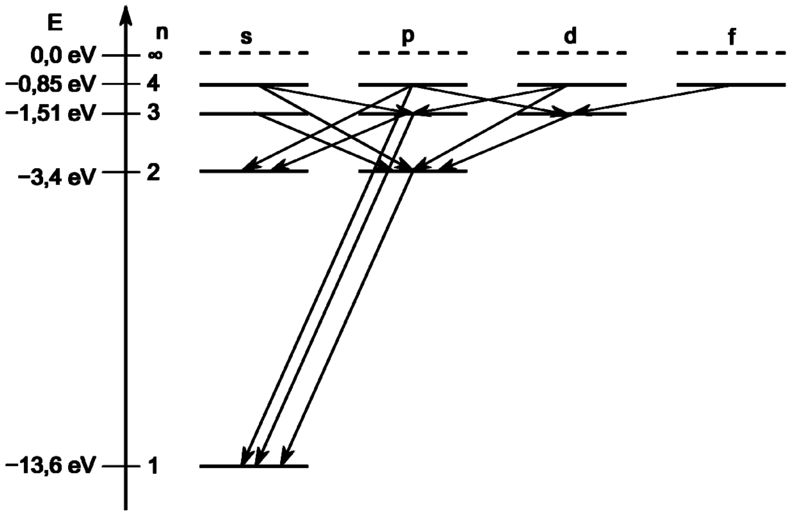

Fig. B.4: Grotrian diagram for atomic Hydrogen with allowed transitions. (taken from

B.3.1 Spectral Lines

Here we continue with a description of LIP emission in the form of bounded electron transitions (spectral lines). This section is based on works [38], [32] and [42]. While considerable part presented in this section is heavily dependent on facts and formalism described in sections B.2 and B.3, we recommend to study those simultaneously. Let's imagine laser-induced plasma under LTE conditions, with negligible optical depth, produced from target consisting of known elemental composition. Also, let's suppose that ablation process was stoichiometric, which means that elemental composition inside plasma plume is the same as in the sample. In such case, measured intensity of a spectral line corresponding to a transition from upper energy level i to lower level j is given by

$$\overline{I_{ij}} = F A_{ij} n_i, \tag{B.10}$$

where F is function aggregating every experimental aspect (spectrometer sensitivity, light collection efficiency and more), A_{ij} is Einstein's coefficient for spontaneous emission and n_i is population density of an upper state. It is remarkable that such a simple equation can describe spectra originated from complex mechanisms inside the plasma. However, there are further restrictions (with previously mentioned ones) for the validity of the equation, firstly plasma has to be optically thin to prevent self-absorption. Second one is non-importance of stimulated decay inside plasma. Even if the LIP satisfy all conditions, the computing population of the upper level is not that straightforward. Expressing population density in Equation B.10, there is

$$\overline{I_{ij}} = FA_{ij}C^l g_i \frac{e^{-E_i/(k_\mathrm{B}T)}}{Z_k^l(T)}, \tag{B.11}$$

C^l being normalized concentration of element l, g_i degeneracy of i-th level, $e^{-E_i/(k_\mathrm{B}T)}$ Boltzmann factor and $Z_k^l(T)$ Partition function of element l with ionization k. Equation B.11 is inside LIBS community often called "Boltzmann equation", but unfortunately that is inappropriate convention because this name stands for famous *Boltzmann transport equation* describing statistical behavior of non-equilibrium systems. Later in this work, we call Equation B.11 as *Boltzmann formula* for line intensity, which respects its meaning and derivation strictly from equilibrium statistical mechanics. Experimental term of Boltzmann formula F could be obtained for example using calibration lamp (with known emissivity), while other terms as A_{ij}, E_i, g_i are collected in spectroscopic databases (e.g. NIST [43]). There are two more terms (depending on temperature) to be determined, Boltzmann factor and Partition function. First mentioned is trivial to calculate if the temperature of the plasma is known (methods for plasma temperature measurements are studied below), but the computation of Partition function is problematic. A common way for obtaining partition function is using the NIST database, while correctness of obtained values is up to a discussion, depending on chosen element and temperature. If we reverse the Boltzmann formula, we were given a powerful tool for obtaining the elemental concentration of measured sample according to experimentally measured line intensities and temperatures. This possibility of quantitative analysis is reviewed in section B.4.2.

General description and derivation of partition function was mentioned in section 1.2.3, but here we shortly investigate some specific challenges of partition function computation related to LIP. As was mentioned, partition function values obtained from the NIST database are generally lower to "real"or correct values. This is caused by non-presence of all accessible energy states of atom or ion in the database or only partial information. Importance of this variation is growing for heavier elements or experimentally unexplored ones. Seemingly best approach would be analytical computation of all energy states obtained as a solution of Schrödinger equation. But as could be easily verified, just for Hydrogen such a sum would diverge. This behavior may be treated by setting an upper limit to a sum, called *cutoff criterion*. There can be various selections for cutoff criterions as ionization energy, Bohr radius (highest energy level counted in sum is one corresponding to a specific semi-classical Bohr radius) or Debye length. While ionization energy would be a good criterion for a single particle system, in plasma there are complex interactions between particles present and then also remaining mentioned criterions serve just as an approximation [44]. An interesting method suitable for LIBS is so-called few level approximation for partition function suggested by G. Colonna and M. Capitelli in their work [45]. In few level approximation scheme (e.g. 3 level), the partition function is summed from 3 terms. The first term being a ground state with respective degeneracy function (statistical weight) G_0. The second one is lumped energy level $\overline{\varepsilon_1}$ consisting of low energy levels and the last third is lumped of high energy levels:

$$Z(T) = G_0 + G_1 \exp(-\overline{\varepsilon_1}/k_{\mathrm{B}}T) + G_2 \exp(-\overline{\varepsilon_2}/k_{\mathrm{B}}T). \tag{B.12}$$

Authors had compared results obtained by this approximation and concluded that errors are under 10% maximally.

Among discussed theoretical intensity of a spectral line, there is another important feature - shape of the line. Width of a spectral line cannot be infinitesimal due to the nature of fundamental physical laws. Its shape is the result of more simultaneous mechanisms, taking place during plasma evolution. Let's review most fundamental mechanisms of line broadening:

- *Natural broadening*
  This mechanism is the most fundamental one and follows directly from Heisenberg's uncertainty principle. The lifetime of the excited state is finite, thus there is uncertainty in energy as

  $$\Delta E \geq \frac{h}{2\pi\tau}, \tag{B.13}$$

  τ being the lifetime of an atomic excited state before it undergoes radiative transition. Lifetime could be defined as

  $$1/\tau = 2\sum_j A_{ij}. \tag{B.14}$$

  Result of a Natural line broadening is shape described by Lorentzian curve

  $$I(\nu) = I(\nu_0)\frac{1}{1 + [(\nu - \nu_0)2\pi\tau]^2}. \tag{B.15}$$

  A common way of describing broadened lines is FWHM (Full width at half maximum) value $\nu_{1/2}$, which is for natural broadening

  $$\Delta\nu_{1/2} = 1/\pi\tau. \tag{B.16}$$

- *Doppler broadening*
  Doppler broadening is caused by Doppler shift of moving particle. Considering the Maxwellian distribution of particle velocity inside the plasma, this broadening results in Gaussian profile

  $$I(\nu) = I(\nu_0)\exp\left[\frac{-(\nu - \nu_0)^2 c^2}{2v_{ta}^2 \nu_0^2}\right], \tag{B.17}$$

where v_{ta}^2 is squared velocity obtained from the kinetic energy of emitting atom a. FWHM for Gaussian profile is

$$\Delta\nu_{1/2} = \nu_0(v_{ta}/c)(2\ln 2)^{1/2}. \tag{B.18}$$

Practically, in time ranges commonly used for LIBS analysis, Doppler broadening forms an only minor contribution to line width and could be neglected.

- *Stark broadening*
  Another mechanism to study is Stark broadening, also called *pressure broadening.* As the second name is indicating, it is caused by collisions inside the plasma. Emission of a colliding particle is perturbed by the presence of electric field. Thus, the energy level of the particle is perturbed and so the wavelength is shifted. For hydrogen-like atoms, linear Stark effect is taking place ($\Delta\nu \propto E$). But for other atoms, the Stark effect is quadratic, where its computation is much complicated and outside of the scope of this thesis. A detailed study of the Stark effect is provided by Griem in his glorious work [46]. The FWHM of the Stark broadened line (in case of neutral atoms or singly charged ions), with neglecting the ion-contribution, can be expressed as

  $$\Delta\lambda_{1/2} = \frac{\Delta\nu_{1/2}}{\nu_0}\lambda_0 = 2W\left(\frac{n_e}{n_r}\right), \tag{B.19}$$

  W being Stark electron-impact broadening parameter (a weak function of temperature), n_e electron density and n_r reference electron density (typically 10^{16} cm^{-3} for neutral atoms and 10^{17} cm^{-3} for singly charged ions [47]). For conversion of FWHM from frequency dependence to wavelength, simple relation $\Delta\lambda/\lambda = \Delta\nu/\nu$ was used.
- *Instrumental broadening*
  The last one to mention is instrumental broadening due to the finite resolution of the spectrometer. It can be determined experimentally, using a calibration lamp and it's resulting in Gaussian shape.
  Finally, if 2 independent profiles (or mechanisms) taking place similarly, the resulting profile is a convolution of both. Convolution of two Gaussian profiles is again Gaussian and the same is for 2 Lorentzian. But for the convolution of Gaussian with Lorentzian, we obtain a new profile called *Voigt profile* (see Figure B.5), which is the actual shape of spectral lines.

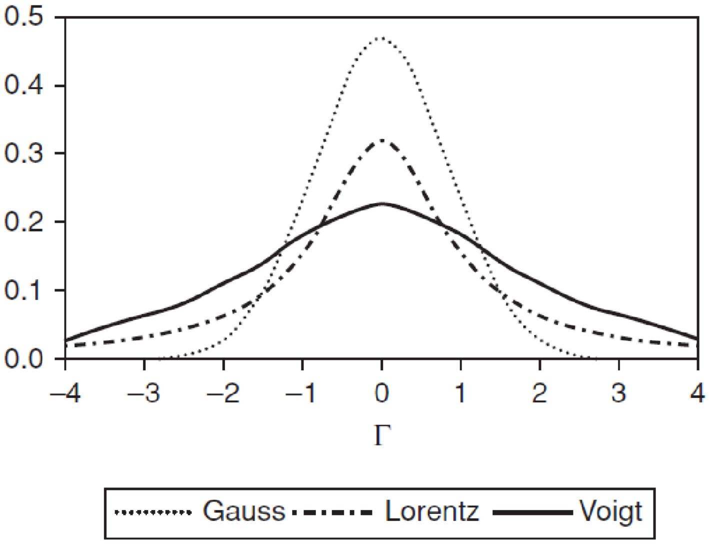

Fig. B.5: Line profiles resulting from various broadening mechanisms. Voigt profile results from a convolution of Gaussian profile with Lorentz profile. (taken from [35])

B.3.2 Plasma temperature estimation

One of the simplest methods to obtain temperature is the two-line method. In this method, we use the trick mentioned in Equation B.4, where the ratio of 2 populations was taken. However, in two-line temperature measurement, we take ratio of 2 integrated intensities for lines corresponding to the same element with identical ionization. Applying the trick to Equation B.11 and expressing it for T, we have

$$T = \frac{E_i - E_m}{k_\mathrm{B} \ln\left(\frac{I_{mn} g_i A_{ij}}{I_{ij} g_m A_{mn}}\right)}. \tag{B.20}$$

Thus, we got rid of partition function dependence and other parameters could be easily found in spectroscopic databases.

Maybe the most common tool for temperature estimation is *Boltzmann plot method*. For building a Boltzmann plot, we linearize Equation B.11 and rearrange it to

$$\ln \frac{I_{ij}}{g_i A_{ij}} = \ln \frac{FC^l}{Z_k^l(T)} - \frac{E_i}{k_\mathrm{B} T}. \tag{B.21}$$

Now, we are able to construct plot (Figure B.6) for more lines of the same element and identical ionization. Temperature is determined by linear regression of points inside the plot as a slope ($-1/k_\mathrm{B}T$) of the line. Logarithm term at RHS of Equation B.21 is just a constant term of linear regression and thus not affecting the slope. For obtaining the desired accuracy, it is important to select more lines with similar upper energy levels. Also, spectrometer should be calibrated to provide consistent sensitivity across the wavelength range.

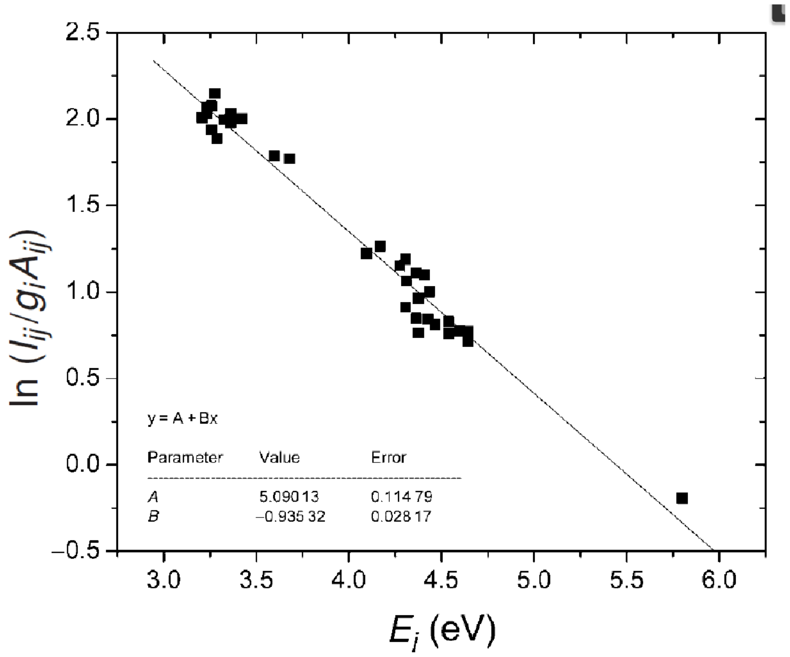

Fig. B.6: Boltzmann plot of the neutral iron lines observed in the LIBS spectrum of an aluminum alloy sample. The line intensity values have been determined as the integral area of the best fitting analytical function. The values have been corrected for the wavelength response of the system. The resulting excitation temperature is $1.24 \cdot 10^4 K \pm 3\%$ (taken from [48])

There exist natural extension for combining atomic lines of an element with its ionic lines. This method is called *Saha-Boltzmann plot* and the core idea is Saha equation (determining relative populations of ionizations i adn $i+1$ as

$$\frac{n_e n_{i+1}}{n_i} = \frac{g_{i+1}}{g_i} \left[\frac{2m^3}{h^3} \left(\frac{2\pi T}{m} \right)^{3/2} \right] e^{-\chi_i/T}. \tag{B.22}$$

For details about Saha equation, please see [38]. Saha-Boltzmann plot allows us to include ionic lines of the element besides atomic lines and thus improve the performance of the model. More advanced utilization of multi-elemental Saha-Boltzmann plot was studied in work [49], making possible to use lines from distinct elements at once.

B.4 Instrumentation and Experiments

As it was mentioned in the introduction for LIBS section, it is a relatively easy utilizable method without any complicated instrumental requirements. However, a high power pulsed laser source is necessary with some guidance optics and spectrometer. Most commonly used are nanosecond Nd:YAG solid-state lasers, operating at 1064 nm (or higher harmonics). In the past few years, usage of femtosecond lasers is obtaining growing attention.

There is a wide range of suitable spectrometers, cost ranging from hundreds of euros to tens of thousands euro. Due to this broad price range, also properties of such spectrometers differ rapidly. State-of-art spectrometers are offering good time synchronization and gating, variable detectors, sensitivity, possibility to work under a noble gas atmosphere and many more. For utilizing plasma diagnostics experiments, it is necessary to use spectrometer with gating possibility (gating time range up to nanoseconds). Low-cost spectrometers are applicable to some specific tasks, lacking requirements on some mentioned properties. Also for simple qualitative analysis, where time synchronization and resolution are not that important, low-cost spectrometers are good option. LIBS applicable spectrometers could be further divided to two most frequently used types:

- *Czerny-Turner configuration*

  Czerny-Turner configuration of a spectrometer is the most common setup using relatively simple parts. It consists of a slit where light enters the spectrometer, then is reflected and collimated by a concave mirror to grating (groove density could range from 100 to 4800 grooves per millimeter). In the last step, light is diffracted from grating and later reflected and focused by the concave mirror to the detector. In the place of the detector, spectra are registered with resolution depending on groove density of grid and also camera resolution. In Czerny-Turner setup, usually, first diffraction order is measured. The wavelength range could cover almost all visible wavelengths at once, but for higher wavelengths overlapping of first and second diffraction order may occur. There is a trade-off between spectral resolution and the range of covered wavelengths.
- *Echelle configuration*

  In echelle configuration of the spectrometer, there are two dispersive elements (in comparison with only one grating in Cz.-T.). First grating is dispersing light in a similar manner to Cz.-T., but only high diffraction orders are collected and guided further. In case of using higher diffraction orders, there is intense overlapping between individual orders. This overlap is treated by using special "echelle"grating, dispersing light in an orthogonal direction to previous dispersion. Separated orders are forming a 2-dimensional pattern which is focused to detector (usually CCD or CMOS camera). The main advantage of echelle spectrometer is simultaneous high (relatively) spectral resolution and the large range of covered wavelengths. However, the best obtainable resolution is not as good as in case of Czerny-Turner configuration, for many applications it is beneficial to have a large spectral bandpass and high spectral resolution at once. The more detailed description of mentioned spectrometers and more configurations suitable for LIBS can be found in [35] or [48].

The two most important mentioned instruments have to be synchronized and controlled by the operator. The device providing this functionality is *Digital Delay Generator* and sometimes could be implemented inside spectrometer or laser itself. Modern DDGs offer the possibility to synchronize more lasers and spectrometers at once with a precision below a nanosecond. Of course, all functionality is operated through PC through various automatized environments.

In addition to the most necessary equipment for LIBS analysis, there are many possible improvements enlarging experimental possibilities or enhancing the user-instrument experience. Maybe the most reliable example is motorized 3-axis stage or manipulator, essential for surface chemical mapping or precise depth profiling.

B.4.1 Basic data processing

Hand to hand with technological improvements in LIBS instrumentation, speed of measurement is increasing and thus the number of produced spectra is raising rapidly. Measurement frequency had reached 1 kHz, enabling to obtain chemical maps from large areas (cm^2) in minutes, resulting in millions of spectra. Such huge numbers of spectra are not possible to inspect by spectroscopic specialist one by one and new approaches are emerging. Methods of *multi-variate data analysis* (MVDA) and Machine Learning are applied to processing of spectra, which can be (in spectroscopic applications) jointly called *chemometry*.

The routine spectroscopic analysis starts with spectra inspecting and assigning peaks to corresponding elements and ionizations, using databases like NIST. Nowadays, it is possible to assign spectra automatically with the help of various tools. Such assigned spectra provide valuable qualitative analysis of the sample with relatively good sensitivity up to ppm (not for every element).

More advanced way of data processing is required in qualitative analysis. The method is sometimes called semi-quantitative because of strong matrix effect and other complications. However, if we have a good set of calibration samples with matching matrices to unknown sample, quantification is carried out with sufficient precision (for many practical tasks in analytical chemistry). But still, it should be emphasized that the biggest advantage of the LIBS method is the speed of analysis and no need for sample preparation. A more detailed review of possibilities and limitations of LIBS quantification is provided in last section dedicated to the applications. In the case where a calibration set is not available, quantitative analysis became challenging. There is a group of methods dealing with this kind of problem representatively called *Calibration-Free LIBS* (CF-LIBS) firstly described in the work [50]. For CF-LIBS, more advanced instrumental equipment (spectrometer with gating possibility and suitable resolution in time and spectra) is essential. Before actual quantification, plasma parameters as temperature and electron density are determined (possibly by methods mentioned in the previous text).

Using modern LIBS instruments with a high repetitive rate of measurements, MVDA algorithms are coming to hand. In large datasets with varying spectra, it is not possible to make visualization or assignation of lines in the usual way. *Principal Component Analysis* (PCA) method became really popular and valuable for visualization and dimensionality reduction of high-dimensional data (spectra). We present this method more closely in section 4. A simple procedure for sorting or classification of high-dimensional spectral data could be done as follows: 1) Carry out PCA analysis on data to visualize important peaks (loadings plot, score plot), reduce dimension by keeping only a few PCs. 2) Cluster analysis, filtering of data, normalization. 3) The classification provided on clusters, testing performance.

B.4.2 LIBS Applications

In previous sections of this chapter, theoretical background and cornerstones of the LIBS method were presented, while applications were mentioned just marginally. For more than 50 years of existence, LIBS was utilized in numerous unique tasks, where other methods were not reliable and also served as a complementary method in other cases. Applicability of LIBS is bounded to the biggest advantages of the method. It is clear that for applications requiring below micrometer resolution and sensitivity up to PPB, LIBS won't serve the best. However, if the speed, cost and sample preparation are priorities, LIBS is the first option. Great success has been obtained by LIBS in 2D elemental mapping for geological and paleoclimate applications. Maps of several squared centimeters were measured on various minerals with resolution up to tens of micrometers. Obtaining precise elemental composition from each spot is enabling advanced geochemical analysis, now possible with unbeatable speed of measurement. [51]

Besides geological mapping, there is a huge potential for LIBS in the biological mapping of plants (toxicology) or soft tissues (heavy metals distribution).

Moving out from mapping, there are applications in environmental monitoring (soils, air, water), automotive and industry (depth profiling, classification of metals) and many more, ending with space exploration (ChemCam). LIBS applications are well described in classic literature [32, 35, 48].

BIBLIOGRAPHY

[1] Douglas Skoog. *Principles of instrumental analysis*. Thomson Brooks/Cole, Belmont, CA, 2007.

[2] Charles Kittel and Herbert Kroemer. *Thermal physics*. W.H. Freeman, San Francisco, 1980.

[3] Charles Kittel. *Elementary statistical physics*. Dover Publications, Mineola, N.Y, 2004.

[4] L. D. Landau and E.M. Lifshitz. *Statistical physics*. Pergamon Press, Oxford,New York, 1969.

[5] L. D. Landau and E.M. Lifshitz. *Mechanics*. Pergamon Press, Oxford New York, 1976.

[6] Petr Kulhanek. *Vybrane kapitoly z teoreticke fyziky*. AGA, Praha, 2016.

[7] Arnold Sommerfeld. *Thermodynamics and statistical mechanics*. Academic Press, New York, 1964.

[8] P. A. M. Dirac. *The principles of quantum mechanics*. Clarendon Press, Oxford, 1958.

[9] Mehran Kardar. *Statistical physics of fields*. Cambridge University Press, Cambridge New York, 2007.

[10] Denis BERNARD. Statistical field theory and applications: An introduction for (and by) amateurs, 2018. URL: https://www.phys.ens.fr/~dbernard/Publications/QFT_STAT_2018_vnew.pdf.

[11] David MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK New York, 2003.

[12] Christopher Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.

[13] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 2019. URL: http://www.sciencedirect.com/science/article/pii/S0370157319300766, doi:https://doi.org/10.1016/j.physrep.2019.03.001.

[14] Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, pages 9–50, London, UK, UK, 1998. Springer-Verlag. URL: http://dl.acm.org/citation.cfm?id=645754.668382.

[15] G E Hinton and R R Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504 LP – 507, jul 2006. URL: http://science.sciencemag.org/content/313/5786/504.abstract, doi:10.1126/science.1127647.

[16] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989. URL: https://doi.org/10.1007/BF02551274, doi:10.1007/BF02551274.

[17] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. URL: https://doi.org/10.1038/323533a0, doi:10.1038/323533a0.

[18] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 78–, New York, NY, USA, 2004. ACM. URL: http://doi.acm.org/10.1145/1015330.1015435, doi:10.1145/1015330.1015435.

[19] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. URL: http://arxiv.org/abs/1207.0580, arXiv:1207.0580.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL: http://arxiv.org/abs/1502.03167, arXiv:1502.03167.

[21] Emmy Noether. Invariant variation problems. *Transport Theory and Statistical Physics*, 1(3):186–207, 1971. URL: https://doi.org/10.1080/00411457108231446, arXiv:https://doi.org/10.1080/00411457108231446, doi:10.1080/00411457108231446.

[22] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. pages 420–434, 2001. URL: http://link.springer.com/10.1007/3-540-44503-X{_}27, arXiv:0812.0624, doi:10.1007/3-540-44503-X_27.

[23] Peter Bubenik. Statistical topological data analysis using persistence landscapes. pages 1–26, 2012. URL: http://arxiv.org/abs/1207.6437, arXiv:1207.6437.

[24] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 7 1948. URL: https://ieeexplore.ieee.org/document/6773024/, doi:10.1002/j.1538-7305.1948.tb01338.x.

[25] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957. URL: https://link.aps.org/doi/10.1103/PhysRev.106.620, doi:10.1103/PhysRev.106.620.

[26] Dataset. https://github.com/JVrabel/Diploma_thesis_attachements.

[27] J. Novotný, M. Brada, M. Petrilak, D. Prochazka, K. Novotný, A. Hrdlička, and J. Kaiser. A versatile interaction chamber for laser-based spectroscopic applications, with the emphasis on laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 101:149 – 154, 2014. URL: http://www.sciencedirect.com/science/article/pii/S0584854714001773, doi:https://doi.org/10.1016/j.sab.2014.08.004.

[28] J. Vrábel, P. Pořízka, J. Klus, D. Prochazka, J. Novotný, D. Koutný, D. Paloušek, and J. Kaiser. Classification of materials for selective laser melting by laser-induced breakdown spectroscopy. *Chemical Papers*, October 2018. URL: https://doi.org/10.1007/s11696-018-0609-1, doi:10.1007/s11696-018-0609-1.

[29] M.A. Nielsen. Neural networks and deep learning, 2019. URL: http://neuralnetworksanddeeplearning.com/.

[30] Magesh Thiyagarajan and Shane Thompson. Optical breakdown threshold investigation of 1064 nm laser induced air plasmas. *Journal of Applied Physics*, 111(7):073302, 2012. URL: https://doi.org/10.1063/1.3699368, arXiv:https://doi.org/10.1063/1.3699368, doi:10.1063/1.3699368.

[31] W. Svendsen, O. Ellegaard, and J. Schou. Laser ablation deposition measurements from silver and nickel. *Applied Physics A*, 63(3):247–255, Sep 1996. URL: https://doi.org/10.1007/BF01567877, doi:10.1007/BF01567877.

[32] Sergio Musazzi. *Laser-induced breakdown spectroscopy : theory and applications.* Springer, Berlin, Heidelberg, 2014.

[33] Laser ablation figure. http://www.celia.u-bordeaux1.fr/spip.php?article857.

[34] Richard E Russo, Xianglei Mao, Haichen Liu, Jhanis Gonzalez, and Samuel S Mao. Laser ablation in analytical chemistry—a review. *Talanta*, 57(3):425 – 451, 2002. URL: http://www.sciencedirect.com/science/article/pii/S003991400200053X, doi:https://doi.org/10.1016/S0039-9140(02)00053-X.

[35] David Cremers. *Handbook of laser-induced breakdown spectroscopy*. Wiley, A John Wiley & Sons, Ltd, Publication, Chichester, West Sussex, 2013.

[36] Lieselotte Blankenburg. *Laser microanalysis*. Wiley, New York, 1989.

[37] Igor Gornushkin, C.L. Stevenson, B.W. Smith, N Omenetto, and J.D. Winefordner. Modeling an inhomogeneous optically thick laser induced plasma: A simplified theoretical approach. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 56:1769–1785, 09 2001. doi:10.1016/S0584-8547(01)00254-3.

[38] I. H. Hutchinson. *Principles of plasma diagnostics*. Cambridge University Press, Cambridge New York, 2002.

[39] G. Cristoforetti, A. De Giacomo, M. Dell'Aglio, S. Legnaioli, E. Tognoni, V. Palleschi, and N. Omenetto. Local Thermodynamic Equilibrium in Laser-Induced Breakdown Spectroscopy: Beyond the McWhirter criterion. *Spectrochimica Acta*, 65:86–95, January 2010. doi:10.1016/j.sab.2009.11.005.

[40] Boris Vodar, Jacques Romand, and Nicole Damany. Some aspects of vacuum ultraviolet radiation physics / edited by boris vodar, nicole damany, jacques romand. *SERBIULA (sistema Librum 2.0)*, 05 2019.

[41] Igor Gornushkin, Reto Glaus, and Lev Nagli. Stimulated emission in aluminum laser-induced plasma: kinetic model of population inversion. *Appl. Opt.*, 56(3):695–701, Jan 2017. URL: http://ao.osa.org/abstract.cfm?URI=ao-56-3-695, doi:10.1364/AO.56.000695.

[42] Andrzej Miziolek. *Laser-induced breakdown spectroscopy (LIBS) : fundamentals and applications*. Cambridge University Press, Cambridge, UK New York, 2006.

[43] A. Kramida, Yu. Ralchenko, J. Reader, and and NIST ASD Team. NIST Atomic Spectra Database (ver. 5.6.1), [Online]. Available: https://physics.nist.gov/asd [2015, April 16]. National Institute of Standards and Technology, Gaithersburg, MD., 2018.

[44] D. Bruno, M. Capitelli, C. Catalfamo, and A. Laricchiuta. Cutoff criteria of electronic partition functions and transport properties of atomic hydrogen thermal plasmas. *Physics of Plasmas*, 15(11):112306, 2008. URL: https://doi.org/10.1063/1.3012566, arXiv:https://doi.org/10.1063/1.3012566, doi:10.1063/1.3012566.

[45] Gianpiero Colonna and Mario Capitelli. A few level approach for the electronic partition function of atomic systems. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 64:863–873, 09 2009. doi:10.1016/j.sab.2009.07.002.

[46] Hans Griem. *Principles of Plasma Spectroscopy*. Cambridge University Press, Cambridge, 1997.

[47] Joaquin Juan Camacho, Jakub Vrabel, Sadia Manzoor, Luis Vicente Pérez-Arribas, Deseada Díaz, and Jorge O Caceres. Spatiotemporal diagnostics of laser induced plasma of potassium gallosilicate zeolite. *J. Anal. At. Spectrom.*, pages –, 2019. URL: http://dx.doi.org/10.1039/C9JA00052F, doi:10.1039/C9JA00052F.

[48] Andrzej Miziolek. *Laser-induced breakdown spectroscopy (LIBS) : fundamentals and applications*. Cambridge University Press, Cambridge, UK New York, 2006.

[49] J.A Aguilera and C. Aragón. Multi-element saha–boltzmann and boltzmann plots in laser-induced plasmas. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 62(4):378 – 385, 2007. URL: http://www.sciencedirect.com/science/article/pii/S0584854707000924, doi:https://doi.org/10.1016/j.sab.2007.03.024.

[50] A Ciucci, M Corsi, Vincenzo Palleschi, S Rastelli, A Salvetti, and E Tognoni. New procedure for quantitative elemental analysis by laser-induced plasma spectroscopy. *Applied Spectroscopy - APPL SPECTROSC*, 53:960–964, 08 1999. doi:10.1366/0003702991947612.

[51] Jorge Caceres, Frédéric Pelascini, V Motto-Ros, Samuel Moncayo, Florian Trichard, G Panczer, A Marín-Roldán, Juncal Cruz, Ismael Coronado, and Javier Martin-Chivelet. Megapixel multi-elemental imaging by laser-induced breakdown spectroscopy, a technology with considerable potential for paleoclimate studies. *Scientific Reports*, 7:5080, 07 2017. doi:10.1038/s41598-017-05437-3.