Charles University in Prague

Faculty of Mathematics and Physics

and

Institute of Computer Science

Academy of Sciences of the Czech Republic

Ph.D. Thesis

Learning from Data based on Approximation of Functions by Neural Networks

Terezie Šámalová

2008

Advisor: RNDr. Věra Kůrková, DrSc.

To my family

As required by Charles University, I hereby declare that I wrote the thesis on my own and that I have included all the sources of information which I used to the references. I authorize Charles University to lend this document to other institutions or individuals for academic and research purposes.

Burnaby and Prague, 2008

Terezie Šámalová

Acknowledgments

This work would not have seen the light in the presented form were it not for many people and circumstances which I can hardly mention here in full. But I would like to use this opportunity to mention some of them. I would like to thank to my advisor Věra Kůrková for pointing to the subject and for stimulating guidance, though due to my long-term stay abroad she could not supervise the thesis as she would have wished to. Cooperation with Petra Kudová-Vidnerová added an interesting practical dimension to my theoretical research, but I would like to thank her mainly for her friendly and reliable attitude at all times. Discussions with my colleagues from Academy of Sciences (Jiří Šíma, Roman Neruda, Martin Holeňa, Petr Savický, Jiří Nedoma and others) helped me with the difficult task of orienting in the scientific world. My thanks go to my friend Stanislav Hencl for helpful comments on some analysis issues. I would like to thank to many of my friends for their encouragement towards finishing the thesis, to name at least some, to Patricia Martínková and her family for supplying me with good mood and much more, to Matt DeVos for his saint-like tolerance, to Zuzana Haniková, Jirka Hanika, Katka Šidlofová, Marie Rucká, Zuzana Kerr, Hardy Hartman, Libor Mašíček and others for being there. And most of all I would like to thank to all my family for their long-long-lasting support. To my parents Jana and Pavel Šidlofovi for their faith in my abilities and for them being great grandparents, to my brother Petr Šidlof for encouragement, proof-reading of the thesis and practical home-land assistance and to my children Toníček and Prokůpek for maintaining my smile. Last of all I cannot thank enough to my ”substitute advisor" when abroad my husband Robert for not giving up on me in science and out of it.

Preface

In this thesis we analyse approximation abilities of one-hidden-layer feedforward neural networks and from theoretical point of view propose and analyse some practically applicable algorithms. In Chapter 1 we explain the model of feedforward neural network and present known universal approximation results applicable to this model. We overview the state of the art in approximation theory with stress on one-hidden-layer feedforward neural networks to introduce the reader to problems addressed in Chapter 2. Here we build on results of Maurey, Jones and Barron and Kůrková et al. on rates of approximation achievable in convex hulls and extend some of these results to a more general setting. These results were published in [S03a] and in [S08]. In Chapter 3 we use results of Darken et al. on rates of approximation to infer probabilistic algorithm for neural networks. It is shown that this algorithm can be used for learning and also for pruning big networks with theoretically guaranteed error of approximation. Results in this chapter were presented in cooperation with Robert Šámal in [SS08]. Chapter 4 comes closer to practical applications proposing special types of kernel-based neural networks. These have been tested on practical applications in cooperation with Petra Kudová-Vidnerová [KS06] and [KS05b]. Intrinsic trouble of kernel-type networks is high number of hidden units. We suggest one possible pruning method by applying probabilistic algorithm proposed in Chapter 3. These ideas were presented as part of [SS08]. Testing on practical applications is running in cooperation with Petra Vidnerová.

Contents

Chapter 1

Introduction to Neural Networks and Universal Approximation

In this chapter we start by introducing the concept of neural networks in Section 1.1. In Section 1.2 we overview two basic and widely used models of neural networks – the multilayer perceptron and radial basis function (RBF) network which we study further in our work. In Section 1.3 we present universal approximation results applicable to our objective. Section 1.4 is basically an introduction to the next chapter, presenting the concept of rates of approximation and overviewing some relevant results.

1.1 Neural Networks

There is no universally accepted definition of what a neural network and theory of neural networks is or should be. Basically it is agreed that neural network theory is a collection of models of computation very loosely based on biological motivations. Neural network is a highly parallel distributed processor that is able to store experimental knowledge making it available for use. Knowledge is acquired in learning process and stored as strength of inter-neuron connections (synaptic weights) – here lies the resemblance to brain (see for example Haykin [Ha94]).

Now for some more mathematical formulation: Neural networks deal with problems where we are given a set of data $z = \{(x_i, y_i); i = 1, \dots, N\} \subseteq \mathbb{R}^d \times \mathbb{R}^m$ of N inputs and corresponding outputs. Input $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ is processed in some way obtaining the output $y = (y_1, \dots, y_m) \in \mathbb{R}^m$. We assume that the process is given by some mathematical function

$$y = F(x),$$

where F could be quite complicated. We even cannot expect to be able to compute the unknown function F exactly. Another thing to keep in mind is that the given data might be noisy. We do not address the issue here because it generally doesn't affect our considerations but techniques discussed in Chapter 4 can deal with the problem very elegantly.

So we try to choose our candidate G for the unknown function F from some parameterized set of functions using a given set of examples $E \subseteq Z$ (some inputs x and associated correct outputs y). The set E should help us choose the right parameters. Neural network models can be considered as particular choices of classes of functions $G(x, w)$, where $w \in \mathbb{R}^k$ are the parameters, accompanied by various rules and specific procedures for choosing optimal parameters. Most neural network models have some training rule to learn parameters w from a set of examples E. There are many different models of neural networks, for an overview of some basic types see for example [Sa02, SiNe96].

Neural networks are a powerful and general means for representing non-linear mappings from several input to several output variables where the form of the mapping is adjusted by a number of parameters. Nonlinearity of neural network models is of great advantage but there is a cost. Determining values of the parameters is a problem of nonlinear optimization and tends to be computationally complicated. Finding efficient algorithms is of great importance since utility of any model depends on them.

Our work does not deal with concrete means to find values of parameters. Instead, we try to understand, what are the abilities of neural networks: In Chapter 2 we investigate the ability of networks (of a certain type) with a given number of neurons to *approximate* a given function. In Chapter 3 we address the same question from an algorithmic point of view. In particular, we propose an algorithm to "prune" a network with too many units. Finally, in Chapter 4 we consider the problem of learning from data as described above. While this is a different problem than approximation, it is reasonable to expect that good approximation properties will imply the ability to learn well. Indeed, we will be able to use methods of Chapter 3. We explain Tikhonov regularization – a successful approach to the problem of learning from data. We report on computer experiments using this theory (with use of specific kernels) and explain how to use pruning (as described in Chapter 3) to improve upon it.

1.2 One-hidden-layer Feedforward Neural Network Models

In this section we will introduce two basic models of networks that we will deal with in our work. There are numerous others but these two are the most used and best understood from mathematical point of view. In our further work we concentrate only on one-hidden-layer networks since unfortunately (or luckily) two- and more-hidden-layer networks exhibit rather different approximation behaviour than one-hidden-layer networks, see for example [MaPi99].

1.2.1 Model of Perceptron

Model of multilayer perceptron works with basic *unit* – the *formal neuron*. It is a simplified mathematical model of biological neuron. Formal neuron with weights $(w_1, \dots, w_d)$ obtains its input $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, computes *weighted sum* $\sum_{i=1}^d w_i x_i$, compares it to *threshold* θ and computes final output using *activation function* σ:

$$y = \sigma\left(\sum_{i=1}^{d} w_i x_i - \theta\right)$$

There are many standard types of activation functions:

- Heaviside function $\sigma(t) = \vartheta(t) = \chi_{[0,\infty)}(t)$ (the characteristic function of interval $[0, \infty)$)

- piecewise linear function

$$\sigma(t) = \begin{cases} 0 & t \leq -1 \\ \frac{t+1}{2} & -1 \leq t \leq 1 \\ 1 & t \geq 1 \end{cases},$$

- logistic sigmoid

$$\sigma(t) = \frac{1}{1+e^{-t}},$$

- hyperbolic tangens (equal to the logistic sigmoid)

$$\sigma(t) = \frac{\tanh(t/2)}{2} + \frac{1}{2},$$

- Gaussian sigmoid

$$\sigma(t) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{t} e^{-y^2/2} dy,$$

- arctan sigmoid

$$\sigma(t) = \frac{1}{\pi} \arctan(t) + \frac{1}{2}.$$

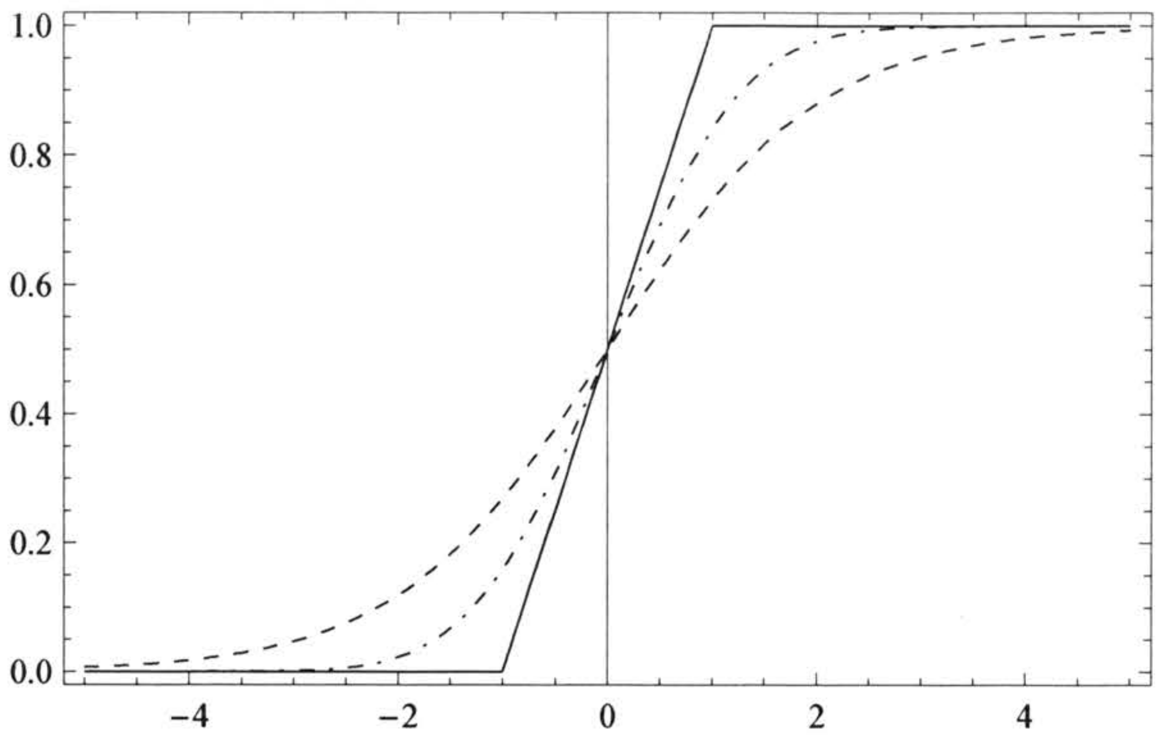

Figure 1.1: Sigmoidal functions.

The listed activation functions are all *sigmoidal* (by this we mean that $\lim_{x\to-\infty}\sigma(x) = 0$, $lim_{x\to\infty}\sigma(x) = 1$ and σ is nondecreasing; we do not require continuity). These are the most commonly used and investigated activation functions, but generally, any activation function is conceivable. The choice, however, has implications on approximation properties of the model. On the other hand Theorem 2.2.9 shows that all sigmoidal activation functions behave the same with respect to the known convergence estimates in $\mathcal{L}_p$ spaces.

Multilayer feedforward perceptron is specified as fixed architecture consisting of finite number of successive *layers*. Each layer has finite number of units (formal neurons). Generally, activation function σ is chosen at the beginning and is the same for all units. Each unit in $(i+1)$-st layer connects to all neurons in the previous i-th layer.

The connections have weights assigned to them, w_{ijk} is weight of connection between k-th neuron in $(i-1)$-st layer and j-th neuron in i-th layer. All neurons have thresholds – θ_{ij} is the threshold in j-th unit of i-th layer.

Multilayer feedforward perceptron (see Figure 1.2) computes its output in a series of discrete steps. Computation starts in *input layer* (values $x_{0,j}$ are assigned according to input). Neurons propagate computation through the network from layer to layer until *output layer* is reached. Value of the j-th unit in the i-th layer is computed as follows:

$$x_{i,j} = \sigma\left(\sum_k w_{ijk}x_{i-1,k} - \theta_{ij}\right).$$

As we already mentioned, the number of layers and units therein remain fixed and so does also σ. Weights and thresholds are parameters of the system and are adapted to find the best approximating function G to the desired function F. This is done in a process called

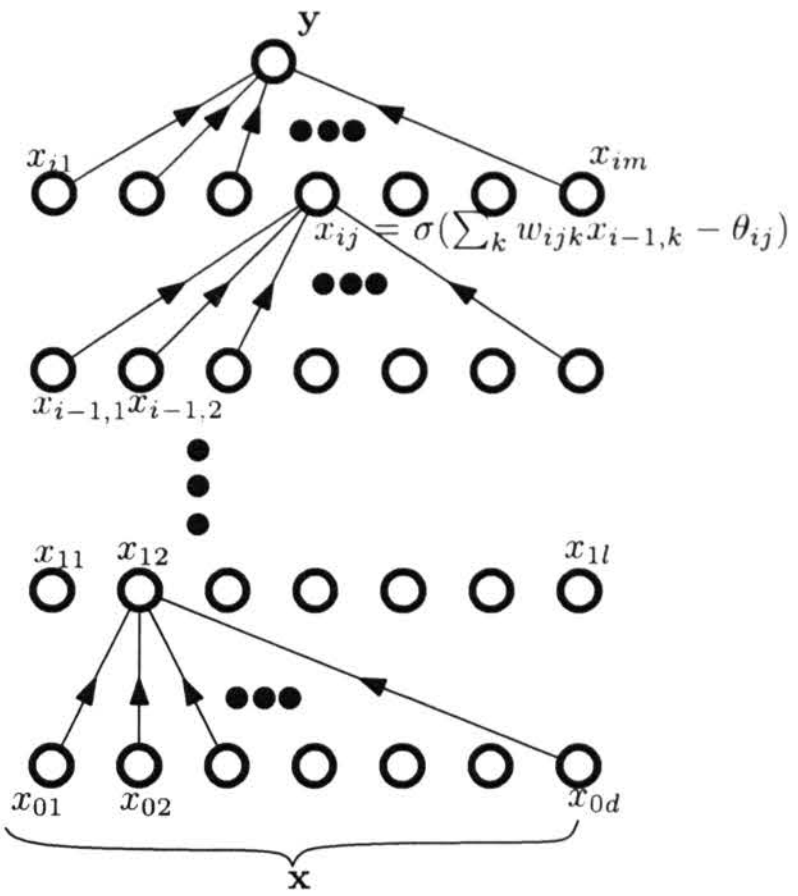

Figure 1.2: Multilayer feedforward perceptron.

learning or *training*. The basic *learning algorithm* for multilayer feedforward perceptron is called *backpropagation*. It is a gradient descent method based on backward propagation of error and is used in majority of all applications of neural networks. In its basic form it uses first derivative of activation function. For details see for example [SiNe96].

Multilayer perceptrons are usually classified according to the number of *hidden layers*, that is the number of layers excluding input and output layer. We will deal only with one-hidden-layer models. Zero hidden layer model is no longer used except for problems of linear separation since it cannot do more. Zero hidden layer model computes

$$x_{1,j} = \sigma\left(\sum_{k=1}^{d} w_{jk} x_{0,k} - \theta_j\right), j = 1, \ldots, m.$$

We can see that this function is constant along parallel hyperplanes, see Figure 1.4 in Section 1.3. This argument along with the fact that no learning algorithm was known for networks with hidden layers was used in the sixties (M. Minsky and S. Papert [MiPa69]) to discredit neural networks. It was shown later that hidden layers (one is enough) allow for approximating arbitrarily well and learning algorithm (backpropagation) was found (Rumelhart, Hinton, Williams [RHW86]).

We will further consider one-hidden-layer perceptron with minor modifications – we simplify to only single output networks and we do not apply activation function and threshold in the output layer. This presents no real restriction. In fact from mathematical point of view applying activation function in output layer does not bring anything new and may introduce restrictions, single output basically only simplifies notation. Thus for d-dimensional input and n units in hidden layer we have the output of the network computed as:

$$y = \sum_{i=1}^{n} c_i \sigma \left(\sum_{j=1}^{d} w_{ij} x_j - \theta_i \right),$$

where x_j are inputs, w_{ij} is the weight between the j-th unit in the input layer and the i-th unit in hidden layer, θ_i is the threshold at the i-th unit of hidden layer and c_i is the weight between the i-th unit in hidden layer and output. We can equivalently write

$$y = \sum_{i=1}^{n} c_i \sigma(w_i \cdot x - \theta_i).$$

This is the type of schema we will mainly concentrate upon.

For completeness we include the formula for more hidden layer perceptron – we iterate the one hidden layer model. So two hidden layer perceptron with n units in first hidden layer and s units in second hidden layer computes:

$$y = \sum_{k=1}^{s} d_k \sigma \left(\sum_{i=1}^{n} c_{ik} \sigma(w_{ik} \cdot x - \theta_{ik}) - \gamma_k \right).$$

As stated above further we address only one-hidden-layer perceptron.

1.2.2 Model of RBF Network

Now we introduce another very well-known and used model of network - *radial basis function networks* or RBF networks. Our task has been made easy as we can apply most of the section above. RBF network architecture is exactly the same as in one hidden layer perceptron. We fix number of units in layers. Each unit in higher layer connects to all units in previous layer. Input units simply pass input values on to the hidden layer. Hidden layer RBF units compute their outputs using radial basis function and pass their result to the linear (summing) output unit.

The RBF unit has again d-dimensional real input $x = (x_1, \ldots, x_d)$ and one real output y. We also apply activation function σ as in the perceptron model but here we apply it to something else than weighted and shifted sum, namely we put

$$y = \sigma \left(\frac{\|x - w\|}{\theta} \right). \tag{1.1}$$

Here $w = (w_1, \ldots, w_d)$ is now interpreted no more as weights but rather as coordinates of *center* of the unit and θ plays the role of *width* of the unit. The norm applied is usually Euclidean. The most commonly used activation function in the RBF model is:

$$\sigma(t) = e^{-\frac{t^2}{\beta^2}}, \beta \geq 0 \qquad \text{(Gaussian)}.$$

Functions defined by (1.1) are called *radial symmetric*, as the value of the function only depends on distance from some center. Functions that are not radial symmetric are used as well.

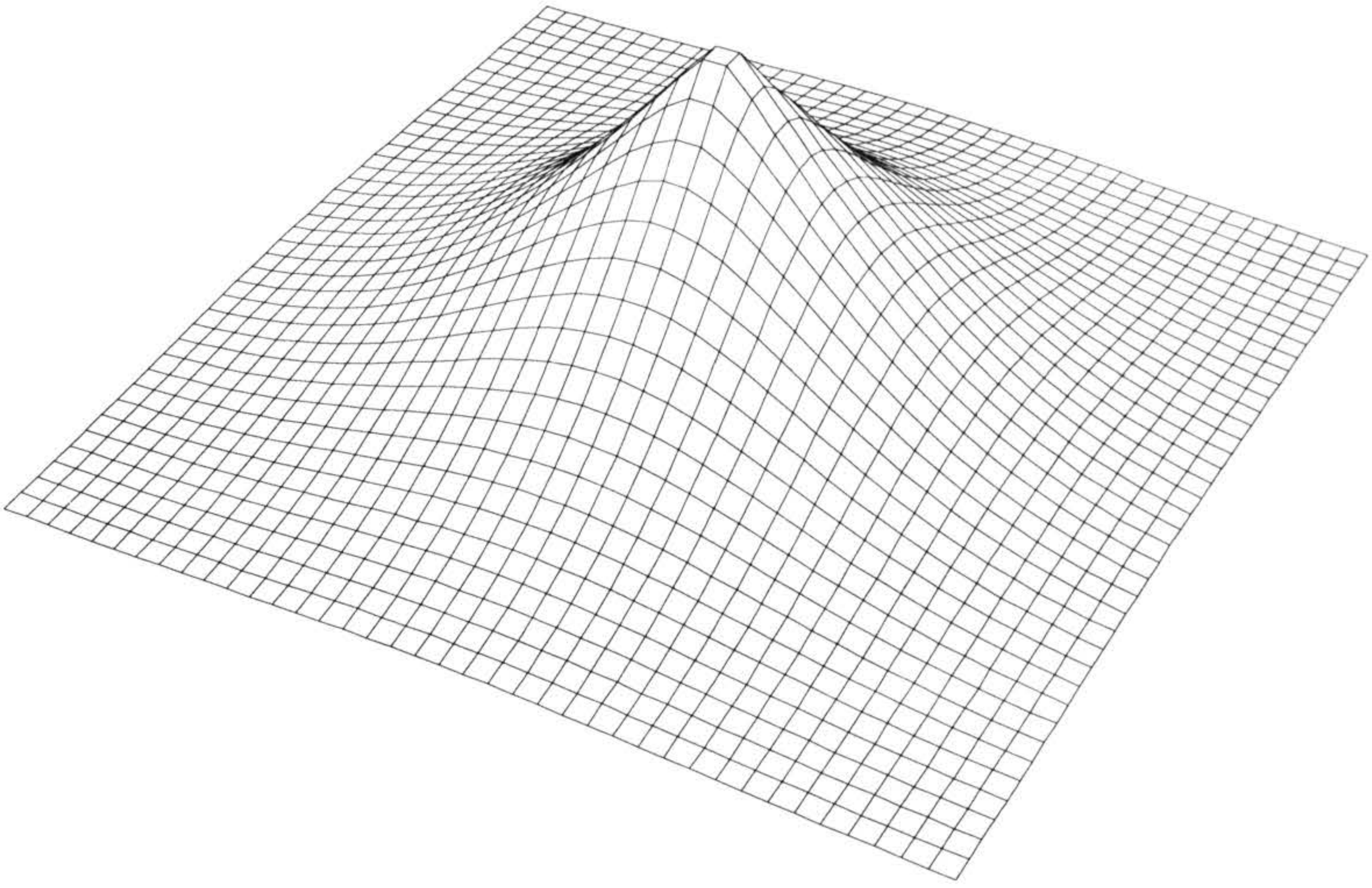

Figure 1.3: Gaussian activation function on $\mathbb{R}^2$.

Summing up, an RBF network with d inputs and n units in the hidden layer computes:

$$y = \sum_{i=1}^{n} c_i \sigma \left(\frac{\sqrt{\sum_{j=1}^{d} (x_j - w_{ij})^2}}{\theta_i} \right),$$

where w_{ij} is the j-th coordinate of the center of the i-th hidden unit, θ_i is width of the i-th unit and c_i is weight applied to the i-th hidden unit.

Although this is the most common model, in many models the norm is omitted and most properties remain unchanged. Then we have:

$$y = \sigma\left(\frac{x - w}{\theta}\right), \tag{1.2}$$

which can be rewritten as $y = \sigma(\alpha x + \beta)$, where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$. Note that here (in contrast to the previous section) σ is a function from $\mathbb{R}^d$ to $\mathbb{R}$.

RBF networks can learn similarly as perceptrons using backpropagation. It is also possible to take advantage of meaning of concrete parameters (centers, widths) and introduce new learning methods as *self-organization* or *genetic algorithms*. We will not go into more detail on this subject as it generally does not affect the topic of our work.

It is easily seen that RBF networks are working on local basis (their activation functions are localised around the centres w – see Figure 1.3) and thus certain kinds of tasks are much easier for them to handle than they are for perceptrons.

1.3 Universal Approximation

Universal approximation of a model is of crucial importance as it answers the question whether reasonable functions can be approximated by the model at all. We consider the set $\mathcal{M}$ of functions on $\mathbb{R}^d$ computable by our model. We call the model to have the property of *universal approximation* if $\mathcal{M}$ is dense in the space $C(\mathbb{R}^d)$ of continuous functions on $\mathbb{R}^d$ in the topology of *uniform convergence on compacta*:
For any $f \in C(\mathbb{R}^d)$, any compact subset $K \subset \mathbb{R}^d$ and any $\varepsilon > 0$ there exists $g \in \mathcal{M}$ such that

$$\max_{x \in K} |f(x) - g(x)| < \varepsilon.$$

The norm of uniform convergence on compacta is very strong. If μ is any nonnegative finite Borel measure with support in some compact set K, then $C(K)$ is dense in $\mathcal{L}^p(K, \mu)$ for any $1 \leq p < \infty$. Thus density results extend also to these spaces, the approximation is "universal". In some cases only the easier task of approximation in $\mathcal{L}^p$-norm is considered, then we talk about density in these spaces.

Universal approximation alone, however, does not ensure nice practical approximation properties of the chosen schema, it only says the schema is able to approximate with arbitrary precision. It does not talk about efficiency (in the case of neural networks how many units in the network are needed to achieve a given precision of approximation).

The universal approximation question for the one-hidden-layer models we have chosen (perceptron and RBF) has been studied by many authors (see for example surveys in [Pi99], [SiNe96]). We will present only the most general results relevant to our study.

1.3.1 Universal Approximation for One-hidden-layer Perceptron

For one-hidden-layer perceptron the solution was given in Leshno, Lin, Pinkus and Schocken [LLPS93]. For a continuous activation function the necessary and sufficient condition on universal approximation is that it is not a polynomial. The authors also consider conditions on non-continuous activation functions and possibilities to restrict weights and thresholds.

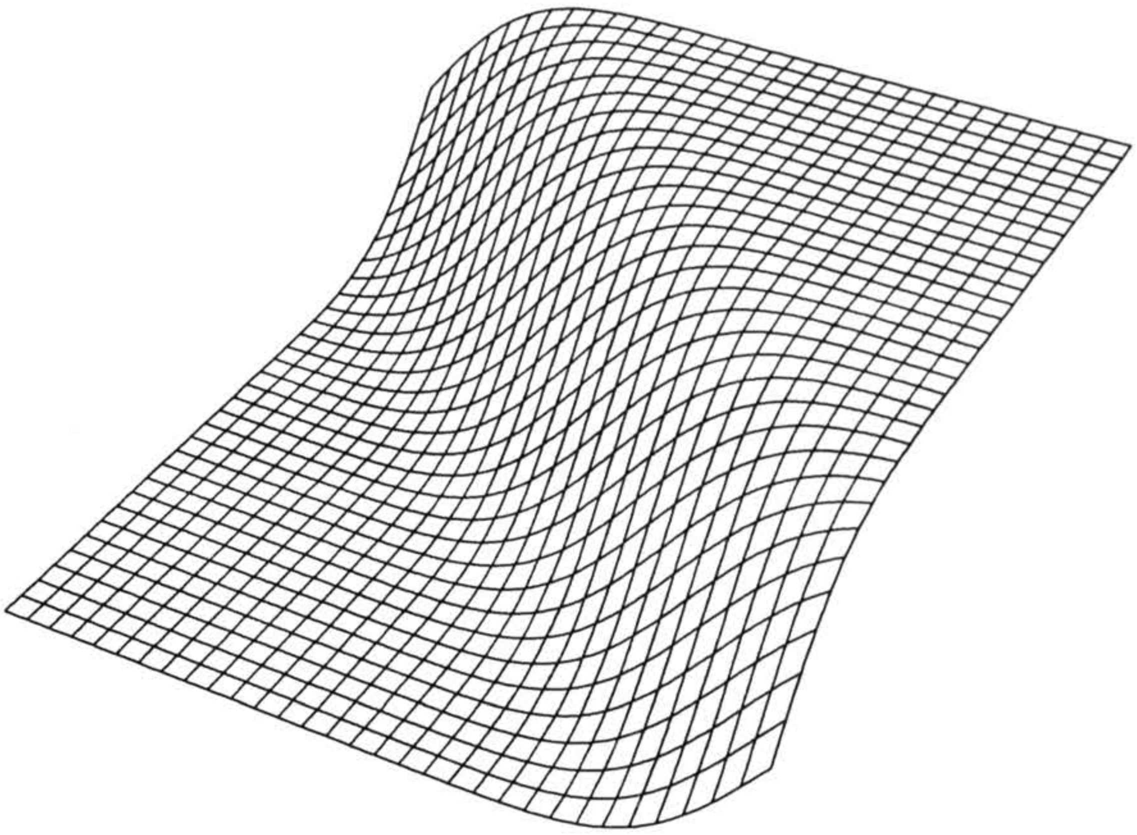

Figure 1.4: Ridge function $x \mapsto \sigma(a \cdot x)$ where $a, x \in \mathbb{R}^2$ and σ is the logistic sigmoid.

Theorem 1.3.1 (Universal approximation – characterization for σ continuous [LLPS93]) *Let $\sigma \in C(R)$. Then*

$$\mathcal{M}(\sigma) = \operatorname{span}\{\sigma(w \cdot x - \theta) : \theta \in \mathbb{R}, w \in \mathbb{R}^d\}$$

is dense in $C(R^d)$ in the topology of uniform convergence on compacta if and only if σ is not a polynomial.

Remark 1.3.2 (Preprocessing [LiPi94]) *The Theorem 1.3.1 is still valid for preprocessed input $x = (x_1, \ldots, x_d) \to h(x) = (h_1(x), \ldots, h_k(x))$ for given fixed $h_j \in C(\mathbb{R}^d), j = 1, \ldots, k$, if and only if h separates points. Here we define*

$$\mathcal{M}_h(\sigma) = \operatorname{span}\{\sigma(w \cdot h(x) - \theta) : w \in \mathbb{R}^k, \theta \in \mathbb{R}\}.$$

We very briefly outline the proof of Theorem 1.3.1. First, one uses the notion of *ridge functions* – functions $g : \mathbb{R}^d \to \mathbb{R}$ given by $g(x) = g_1(a \cdot x)$ for some $g_1 : \mathbb{R} \to \mathbb{R}$ (see Figure 1.4). Result of Vostrecov and Kreines [VoKr61] implies that linear combinations of ridge functions given by continuous g_1s are dense in $C(\mathbb{R}^d)$ in the topology of uniform convergence on compacta. Then one shows that the set $\mathrm{span}\{\sigma(at+b) : a, b \in \mathbb{R}\}$ is dense in $C(\mathbb{R})$ for σ continuous nonpolynomial. (For details, see [Pi99].) We only state here the result of Vostrecov and Kreines [VoKr61]: it is very elegant and may have further use in the theory of neural networks.

Theorem 1.3.3 (Ridge functions are dense in $C(\mathbb{R}^d)$ [VoKr61]) *The set of ridge functions*

$$\mathcal{R}(A) = \mathrm{span}\{g(a \cdot x); g \in C(\mathbb{R}), a \in A\}$$

is dense in $C(\mathbb{R}^d)$ in the topology of uniform convergence on compacta if and only if there is no nontrivial homogeneous polynomial that vanishes on A.

Pinkus et al. further extend their result to noncontinuous σ but some smoothness demands remain:

Theorem 1.3.4 (Universal approximation for σ integrable [LLPS93]) *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be bounded and Riemann-integrable on every finite interval (set of discontinuities of σ has Lebesgue measure zero). Then $\mathrm{span}\{\sigma(ax-b) : a \in \mathbb{R}, b \in \mathbb{R}\}$ is dense in $C(R)$, unless σ is a polynomial almost everywhere.*

By similar arguments as in proof of Theorem 1.3.1 and by Vostrecov and Kreines we can obtain universal approximation in $C(\mathbb{R}^d)$.

Pinkus et al. also present some results granting universal approximation for limited sets of weights and thresholds (again, see [Pi99] for details). Similar techniques as were used to prove universal approximation for the model give also the following interpolation result:

Theorem 1.3.5 (Interpolation with continuous σ [Pi99]) *Let $\sigma \in C(R)$ be not a polynomial. For any k distinct points $\{x_i\}_{i=1}^k \subset \mathbb{R}^d$ and associated outputs $\{y_i\}_{i=1}^k \subset \mathbb{R}$ there exist $\{w_j\}_{j=1}^n \subset \mathbb{R}^d$ and $\{c_j\}_{j=1}^n, \{\theta_j\}_{j=1}^n \subset \mathbb{R}$ such that*

$$\sum_{j=1}^{n} c_j \sigma(w_j \cdot x_i - \theta_j) = y_i, i = 1, \ldots, k.$$

If $\mathrm{span}\{\sigma(x-b) : b \in \mathbb{R}\}$ is dense in $C(\mathbb{R})$ then we can choose $\{w_j\}_{j=1}^n \subset S^{d-1}$.

Universal approximation property for one-hidden-layer perceptron clearly implies the same property for any perceptron type network. Note, that networks with more layers not only give universal approximation of the model but also yield some interesting approximation properties – see for example [ChLM94] or [MaPi99] for two-hidden-layer perceptron.

1.3.2 Universal Approximation for RBF Networks

Universal approximation of RBF networks has been studied by many authors but the most well known result comes from Park and Sandberg [PaSa91], [PaSa93]. They showed that if the radial basis activation function used in hidden layer is bounded and integrable on $\mathbb{R}^d$ and the integral is not zero, then an RBF network (as defined above in the more general sense – Equation (1.2)) can approximate any function in $\mathcal{L}^p(\mathbb{R}^d)$ with respect to the $\mathcal{L}^p$ norm for $1 \leq p < \infty$ arbitrarily well. Similar result is valid for activation function σ integrable, continuous and either with non-zero integral over $\mathbb{R}^d$ or radial symmetric with respect to Euclid norm ($\|x\| = \|y\| \Rightarrow \sigma(x) = \sigma(y)$). In this case RBF network can approximate any function on $C(X)$ for X compact subset of $\mathbb{R}^d$ arbitrarily well. We present these results here as restated in [SiNe96]:

Theorem 1.3.6 (RBF dense in $\mathcal{L}^1(\mathbb{R}^d)$, necessary and sufficient condition [PaSa91]) *Let* $\sigma : \mathbb{R}^d \to \mathbb{R}$. *Then* $\operatorname{span}\{\sigma\left(\frac{x-w}{\theta}\right), w \in \mathbb{R}^d, \theta \in \mathbb{R}\}$ *is dense in* $\mathcal{L}^1(\mathbb{R}^d)$ *if and only if* σ *is integrable and*

$$\int_{\mathbb{R}^d} \sigma(x)\,\mathrm{d}x \neq 0.$$

We can even fix θ *(all units have the same width) and the result still holds.*

Theorem 1.3.7 (RBF dense in $\mathcal{L}^p(\mathbb{R}^d)$ [PaSa91]) *Let* $p \geq 1$ *and let* $\sigma : \mathbb{R}^d \to \mathbb{R}$ *be such that* $\sigma \in \mathcal{L}^p(\mathbb{R}^d)$ *and*

$$\int_{\mathbb{R}^d} \sigma(x)\,\mathrm{d}x \neq 0.$$

Then $\operatorname{span}\{\sigma\left(\frac{x-w}{\theta}\right), w \in \mathbb{R}^d, \theta \in \mathbb{R}\}$ *is dense in* $\mathcal{L}^p(\mathbb{R}^d)$.

Theorem 1.3.8 (RBF dense in $C(X)$ [PaSa93]) *Let* $X \subseteq \mathbb{R}^d$ *compact. Let* σ *be integrable continuous and let either*

$$\int_{\mathbb{R}^d} \sigma(x)\,\mathrm{d}x \neq 0$$

or σ *be radial symmetric with respect to Euclid norm. Then* $\operatorname{span}\{\sigma\left(\frac{x-w}{\theta}\right), w \in \mathbb{R}^d, \theta \in \mathbb{R}\}$ *is dense in* $C(X)$.

Liao, Fang and Nuttle (2003) [LFN03] extend this result in the respect that integrability of the activation function is not required – they claim the same conclusion for $\sigma \in C^\infty(\mathbb{R}^n)$ that is not a polynomial. Inspection of the proof, however, reveals that they in fact are requiring σ to be analytic.

1.4 Rates of Approximation

Rates of approximation will be studied in detail in Chapter 2, here we give an overview of basic notions and of one type of results.

As we already mentioned, universal approximation of a model only guarantees existence of arbitrary good approximations – it does not say anything about how "efficient" these approximations are. To measure this quality of the approximation schema we introduce *rates of approximation.*

Generally we are working in a normed space of functions X trying to approximate a function $f \in X$. We have a family of manifolds $\{\mathcal{M}_n\}_{n=1}^{\infty}$, so that $\mathcal{M}_n \subseteq X$ and $\mathcal{M} = \bigcup_n \mathcal{M}_n$ is dense (in some sense) in X and

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_n \subset \cdots$$

Thus the distance to our desired function f from $\mathcal{M}_n$ decreases with increasing n and approximation in $\mathcal{M}_n$ can get arbitrarily close to f provided we use sufficiently large n.

Since computational time needed to find approximation in $\mathcal{M}_n$ is increasing with n it is of great interest to know the rate of convergence to zero of distance of f from $\mathcal{M}_n$ as a function of n – the *rate of approximation.*

There are generally two ways in which this problem has been approached. Either we consider standard smoothness classes (Sobolev spaces, etc.) and estimate worst case error of approximation by perceptron from functions in this class (Pinkus, Maiorov, DeVore, Howard, Micchelli, Meir, Petrushev and others).

A different approach is, given approximation schema $\{\mathcal{M}_n\}_{n=1}^{\infty}$, to try to find classes of functions which are well approximated by it (Jones, Barron, Hornik, Maurey, Makovoz, Darken, Donahue, Gurvits, Sontag, Stinchcombe, White, Kainen, Kurkova, Kreinovich and others).

In our further work we will pursue the second approach and so we will not elaborate on it here and postpone overview of concrete results to Chapter 2. Here we will mention only a few results from the first group. For further details see for example [Pi99].

1.4.1 Rates of Approximation in Standard Smoothness Classes

For a given activation function σ let us define

$$\mathcal{M}_n(\sigma) = \{\sum_{i=1}^{n} c_i \sigma(w_i \cdot x - \theta_i) : c_i, \theta_i \in \mathbb{R}, w_i \in \mathbb{R}^d\}.$$

Universal approximation has been proven for this schema for σ continuous non-polynomial, see Theorem 1.3.1. Now we are concerned with error of this approximation and asymptotic properties of the error. As in the proof of Theorem 1.3.1, ridge functions are of help. We define

$$\mathcal{R}_n = \{\sum_{i=1}^{n} g_i(a_i \cdot x) : a_i \in \mathbb{R}^d, g_i \in C(\mathbb{R}), i = 1, \ldots, n\}.$$

Obviously for any σ continuous we have $\mathcal{M}_n(\sigma) \subseteq \mathcal{R}_n$, thus we can estimate the error of approximation using $\mathcal{M}_n$ by

$$\inf_{g \in \mathcal{M}_n(\sigma)} \|f - g\|_X \geq \inf_{g \in \mathcal{R}_n} \|f - g\|_X.$$

In 1999 Maiorov [Ma99] proved lower bound for the error of approximation by ridge functions.

Theorem 1.4.1 (Lower bound on rates of approximation for perceptron [Ma99]) *Let $m \geq 1$ and $d \geq 2$. Then for each n there exists an $f \in \mathcal{W}^{m,2}(B^d)$; $\|f\|_{m,2} \leq 1$ for which*

$$\inf_{g \in \mathcal{R}_n} \|f - g\|_2 \geq C n^{-m/(d-1)}.$$

Here the positive constant C is independent of f and n, B^d denotes the unit ball in $\mathbb{R}^d$.

It can be shown [MMR99] that the set of functions f for which the lower bound holds is in fact of large measure, in other words this is not only a worst case result. We see that there are functions in $\mathcal{W}^{m,2}$ that cannot be approximated by ridge functions with better rates than $n^{-m/(d-1)}$, in other words if m (the smoothness) does not grow with the dimension d of the input space, we get so-called curse of dimensionality. This aspect will be discussed in more detail at the end of Section 2.4.

Theorem 1.4.1 gives a lower bound also for $\inf_{g \in \mathcal{M}_n} \|f - g\|_2$ on B^d since the $\mathcal{M}_n(\sigma)$ consists of ridge functions of a certain type (the "one-dimensional" function defining the ridge function is restricted to shifts of a given activation function). The question now is: how relevant are these bounds? It has been shown already in [Ma99] that the bound in Theorem 1.4.1 is asymptotically tight. Moreover, Maiorov and Pinkus [MaPi99] show that there exists an activation function that is sigmoidal, strictly increasing and $C^\infty(\mathbb{R})$ and for which the lower bound is attained. The activation function they use, while infinitely smooth, is somewhat artificial. For the logistic sigmoid $\sigma(t) = 1/(1 + e^{-t})$ a slightly larger lower bound was shown [MM00]. Yet more importantly, a slightly larger bound ($Cn^{-m/d}$) than that in Theorem 1.4.1 is an *upper* bound for all sufficiently nice activation functions (including the logistic sigmoid). This was shown by Mhaskar [Mh96], see also [Pi99] for a nice exposition.

So we see that when approximating any function $f \in \mathcal{W}^{m,p}$ for which $\|f\|_{m,p} \leq 1$ by the perceptron schema with smooth activation function, we always have (for nonpolynomial functions) rate of approximations $O(n^{-m/d})$, while we never can achieve better rate than $\Omega(n^{-m/(d-1)})$.

Theorem 1.4.2 (Upper bound on rates of approximation for perceptron [Mh96]) *Let I be an open interval. Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is such that $\sigma \in C^{\infty}(I)$ and σ is not a polynomial on I. Then for each $p \in [1, \infty]$, $m \geq 1$ and $d \geq 2$*

$$\sup_{f \in \mathcal{W}^{m,p}(B^n); \|f\|_{m,p} \leq 1} \quad \inf_{g \in \mathcal{M}_n(\sigma)} \|f - g\|_p \leq C n^{-m/d},$$

for some constant C independent of n.

Maiorov [Ma03] proved similar bounds for radial functions:

Theorem 1.4.3 (Lower and upper bound on rates of approximation for RBF [Ma03]) *Let $d \geq 2$, $m > 0$ and n be natural numbers. Then there exist weights $w_1, \cdots, w_n$ on the unit sphere S^{d-1} such that*

$$\sup_{f \in \mathcal{W}^{m,2}; \|f\|_{m,2} \leq 1} \quad \inf_{g \in \mathcal{R}^{rad}_{w_1, \cdots, w_n}} \|f - g\|_2 \leq c_1 n^{-\frac{m}{d-1}}$$

and for any $w_1, \cdots, w_n \in \mathbb{R}^d$ we have lower estimate on error:

$$\sup_{f \in \mathcal{W}^{m,2}; \|f\|_{m,2} \leq 1} \quad \inf_{g \in \mathcal{R}^{rad}_{w_1, \cdots, w_n}} \|f - g\|_2 \geq c_2 n^{-\frac{m}{d-1}},$$

where $\mathcal{R}^{rad}_A = \operatorname{span}\{g(\|x - w\|^2), w \in A \subseteq \mathbb{R}^d, g \in C(\mathbb{R})\}$, c_1, c_2 are constants depending only on m and d and $\|.\|$ is Euclidean.

A different approach to the study of rates of approximation (the "second group" mentioned above) will be discussed in the next chapter.

Chapter 2

Estimates on Rates of Approximation by Neural Networks Using Integral Representations

In this chapter we address a crucial question of interest when building a neural network: how precisely can we approximate a given function using a limited number of units. We proceed along the lines initiated by Barron in the respect that we try to derive classes of functions that are well approximated by some specific approximation schema $\{\mathcal{M}_k\}_{k=1}^{\infty}$.

In Section 2.1 we first review the pioneering work of Maurey [Ps81], Jones [Jo92], and Barron [Ba93], and the extension by Darken, Donahue, Gurvits, and Sontag [DDGS93]. Then we show how these results were utilized by Kůrková, Kainen, and Kreinovich [KKK97] who (implicitly) used so-called $\mathcal{G}$-variation (explicitly defined in [Ku97]) of the function f to be approximated. We will see that bounded $\mathcal{G}$-variation is a sufficient (though not necessary) condition for good rates of approximation.

In Section 2.2 we first present results of [KKK97] where bounds on $\mathcal{G}$-variation are obtained for functions in the form of integral representation using continuous or Heaviside functions. We then extend their results to more general function spaces. We do not require continuity of the functions involved in the integral representation; we also present simpler proof of the estimate from [KKK97]. The obtained improvements enable more direct and more general application of results of Maurey, Jones, Barron and Darken et al. giving approximation error rate of order $O(n^{1/q})$ for one-hidden-layer networks with n hidden units. Here q is a constant depending on the "type" of the involved function space, but not on the "dimension". E.g., if we are dealing with functions in $\mathcal{L}^p(\mathbb{R}^d)$ $(1 < p < 2)$ then $q = p/(p-1)$ is the conjugate exponent; in particular, q does not depend on d (note that for high d we may obtain large constant in the $O(\cdot)$, though, [KHS98]). Using $\mathcal{L}^p$ spaces for $p \neq 2$ is of practical interest, as by using $\mathcal{L}^p$-norm for $1 < p < 2$ one can cope better with functions with peaks, which are probably errors in measurement, so-called outliers [Re83, HaBu88]. We also present an interesting property of $\mathcal{G}$-variation for neural network approximation schema with sigmoidal activation functions – we show that the presented estimates on approximation rates cannot distinguish between different activation functions (Theorem 2.2.9).

For our estimates on $\mathcal{G}$-variation we need to have function f represented in form of an integral representation. In Section 2.3 we listed examples of functions where such integral representation exists. We generalise integral representation of function by using measure instead of weights. This enables us to provide in Section 2.3.2 explanation and justification of the metaphor "neural network with continuum many neurons", which is used in [KKK97] to motivate special type of integral representation of functions. By an application of Helly's theorem on w^* sequential compactness we get that, in a proper setting such representation is equivalent to a limit of "classical" finite neural networks.

In Section 2.4 we combine the three previous sections and thus provide a few concrete estimates on rates of approximation of the type: If a function is "smooth enough" then it can be approximated by one-hidden-layer neural network with n units with rate of approximation of $O(n^{1/q})$. We also discuss possibilities to weaken the smoothness assumptions.

Results presented in this section have been published in [S08, S03a, S03b].

2.1 Rates of Approximation in Banach Spaces

A general topic (not only) in mathematics is, how to approximate some complicated object using limited resources. To be more specific, we have a Banach space X of functions, and a set $\mathcal{G} \subseteq X$ of functions we are allowed to use for approximation of a given function $f \in X$, while we want to use as few functions from $\mathcal{G}$ as possible.

In Section 2.1.1 we show results from approximation theory that provide good rates of approximation for function f in the closure of convex hull of the set of approximating functions. In Section 2.1.2 we show how the above results have been reformulated in a more explicit form taking into account relationship between the set of approximating functions and the function to be approximated. We will see that the condition of f being in convex hull is sufficient (though not necessary) for the existence of efficient approximations of f. On the other hand, if f is merely in the closed linear span, the rates of approximation of f may be arbitrary bad (Corollary 2.1.8).

2.1.1 Approximations in the Closed Convex Hull

A frequent approach in approximation theory is to iteratively construct a sequence of approximants f_n to a function f, where at each step we add an appropriate element of $\mathcal{G}$:

$$f_{n+1} = f_n + g, \qquad g \in \mathcal{G}\,. \tag{2.1}$$

Here, g is chosen to minimize the norm $\|f_{n+1} - f\|$ (or to make it close to $\inf\{\|(f_n + g) - f\| : g \in \mathcal{G}\}$). A natural setting for this is when X is a Hilbert space. Huber [Hu85] conjectured that for *projection pursuit regression* (which corresponds to $\mathcal{G}$ consisting of all ridge functions) this method always produces a sequence f_n converging to f. This was affirmatively resolved by Jones [Jo87]. However, the convergence can in general be very slow.

In a subsequent work [Jo92] Jones studies approximations when slightly more general iterative step is allowed – instead of adding some $g \in \mathcal{G}$ to the previous function, we take a convex combination:

$$f_{n+1} = \alpha f_n + (1-\alpha) g, \qquad g \in \mathcal{G}, \alpha \in [0,1] \tag{2.2}$$

where g and α are chosen (approximately) optimal. Somewhat surprisingly, this modification significantly increases the speed of convergence:

Theorem 2.1.1 (Maurey-Jones-Baron – Iterative rates in Hilbert sp. [Ba93, Jo92, Ps81])

Let $\mathcal{G}$ be a set of functions, subset of a Hilbert space $\mathcal{H}$ of functions on $\mathbb{R}^d$. Suppose f is in $\operatorname{cl}\operatorname{conv}\mathcal{G}$, and that for every $g \in \mathcal{G}$ we have $\sqrt{\|g\|^2_{\mathcal{H}} - \|f\|^2_{\mathcal{H}}} \le \rho$ for some constant $\rho \in \mathbb{R}$. Then it is possible to find a sequence $\{f_n\}$ satisfying

$$\|f - f_n\|_{\mathcal{H}} \le \frac{\rho}{\sqrt{n}},$$

by using the recurrence (2.2), *when the functions g and numbers α are chosen sufficiently close to the optimum. Observe that we have $f_n \in \operatorname{conv}_n \mathcal{G}$.*

Note that ρ does not depend on n, however it depends on $\mathcal{H}$, $\mathcal{G}$, and, in particular, on f. We'll present estimates of this constant later (Theorem 2.1.5 and 2.1.6), the dependence on f is actually the topic of the rest of the chapter. The above result in slightly weaker version is attributed to Maurey by Pisier [Ps81]. Barron [Ba93] was the first who noticed that it is applicable to neural networks. He also provides the improved bound: instead of $\frac{1}{\sqrt{n}} \sup_{g \in \mathcal{G}} \|f - g\|_{\mathcal{H}}$ (Maurey) he obtains $\frac{1}{\sqrt{n}} \sup_{g \in \mathcal{G}} \sqrt{\|g\|^2_{\mathcal{H}} - \|f\|^2_{\mathcal{H}}}$, which in the natural applications is lower.

We feel obliged to comment here that the title of the theorem (Iterative rates) is slightly misleading but for a good reason. Maurey's proof of the theorem is in fact probabilistic but we retain the title iterative to stress that an iterative proof is possible as this is interesting from algorithmic point of view. We follow this approach also in titles of further theorems.

The above result was extended in various ways. The strongest result obtained in this direction is due to Makovoz [Mk96]. He replaces the bound $\rho/\sqrt{n}$ by $\varepsilon_n(\mathcal{G})/\sqrt{n}$, where $\varepsilon_n(\mathcal{G}) \in (0, \rho]$ depends on $\mathcal{G}$ and on n:

$$\varepsilon_n(\mathcal{G}) = \inf\{\varepsilon > 0 : \mathcal{G} \text{ can be covered by at most } n \text{ sets of diameter} \leq \varepsilon\}\,.$$

When $\mathcal{G}$ is finite-dimensional, $\varepsilon_n(\mathcal{G}) = o(1)$ (as $n \to \infty$), so this is a stronger result. Consider the particular case where $\mathcal{G}$ corresponds to neural networks with Heaviside activation functions, with inputs in $\mathbb{R}^d$. In this case $\mathcal{G} = \mathcal{G}_\eta = \{\eta(a \cdot x + b), a \in \mathbb{R}^d, b \in \mathbb{R}\}$. This yields [Mk96] an improved bound on the error of the n-term approximation, namely $O(1/n^{\frac{1}{2}+\frac{1}{2d}})$. We will not pursue this direction, as our particular interest is on the case of large d, where the improvement is only slight. The drawback of Makovoz' approach is that it does not yield existence of approximants that can be computed in an iterative manner, as in (2.2).

Let us pause here to explain the dependence on the "dimension of the data". In early proofs of universal approximation property of neural networks, the "amount of work" needed for efficient approximation of a function on $\mathbb{R}^d$ seemed to depend exponentially on the dimension d. This so-called "curse of dimensionality" is obviously a major obstacle in applications of neural networks, as many interesting applications are intrinsically multi-dimensional. Theorem 2.1.1 tells us, that for a fixed function f, space $\mathcal{H}$ and approximating functions $\mathcal{G}$, the error of approximation decreases fast with n, the number of approximants. This is certainly useful (and in particular proves superiority of the approximation schema (2.2) over (2.1)), the dimension, however, is "cursed" in more ways than this. One other problem is, that with increasing dimension of inputs, we are likely to see larger constants ρ in Theorem 2.1.1. In [KHS98] a sequence of functions is presented, where ρ grows exponentially with the dimension. Yet another problem is met when we consider the algorithmic point of view. The amount of work to do "elementary operations" (estimating the norm, scalar product, etc.) with functions on $\mathbb{R}^d$ grows exponentially with d. This can be remedied by using more sophisticated numerical methods (as Monte Carlo), however. We will address some of these issues in the next chapter.

It is natural to ask, whether the above-mentioned result can be generalized to arbitrary Banach spaces. Not only this is an interesting question in itself, it was motivated by the fact, that spaces $\mathcal{L}^p$ (for $p < 2$) possess better approximation properties than $\mathcal{L}^2$: namely, they can cope better with an "error in measurement" of the function to be approximated [Re83, HaBu88].

In Darken et al. [DDGS93] this question was addressed in a great detail. It is shown, that

Theorem 2.1.1 can be extended to any Banach space with unit ball that is not too "pointed" – namely to any *uniformly smooth* space. We say that a Banach space X has *modulus of smoothness* ϱ if $\varrho : [0, \infty) \to [0, \infty)$ is a function given by

$$\varrho(r) := \sup_{\|f\|_X = \|g\|_X = 1} \Big(\frac{\|f + rg\|_X + \|f - rg\|_X}{2} - 1 \Big)$$

(the supremum is taken over all $f, g \in X$ of unit norm). It is easy to observe that $\varrho(r) \leq r$ for any Banach space, and that in a Hilbert space $\varrho(r) = \sqrt{1 + r^2} - 1 = O(r^2)$ (as $r \to 0$). A Banach space is termed *uniformly smooth* if $\varrho(r) = o(r)$ (as $r \to 0$). This is in particular satisfied for $\mathcal{L}^p$ spaces with $1 < p < \infty$, the modulus of smoothness is (see [DDGS93])

$$\varrho(r) \leq \begin{cases} r^p/p & \text{if } 1 < p \leq 2 \\ \frac{p-1}{2} r^2 & \text{if } 2 \leq p < \infty. \end{cases}$$

Darken et al. [DDGS93] prove a result about approximating functions in Banach spaces based on modulus of smoothness of these spaces (see Theorem 3.1.1 where numerical issues are discussed). This theorem applied to $\mathcal{L}^p$ spaces yields Theorem 2.1.2. It also turns out, that the convex combination in (2.2) can be chosen so that $\alpha = n/(n+1)$.

Theorem 2.1.2 (Rates in $\mathcal{L}^p$ spaces – iterative [DDGS93]) *Let $\mathcal{G}$ be a bounded subset of an $\mathcal{L}^p$-space X ($1 < p < \infty$), with $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ given. Put $q = p/(p-1)$ and let $\rho > 0$ be such, that $\|f - g\| \leq \rho$ for all $g \in \mathcal{G}$. Then for every $\varepsilon > 0$ there is a sequence $\{g_n\} \subset \mathcal{G}$ such that the sequence $\{f_n\} \subset \operatorname{conv}\mathcal{G}$ defined by*

$$f_1 = g_1, \qquad f_{n+1} = \frac{n}{n+1} f_n + \frac{1}{n+1} g_n$$

satisfies

$$\|f - \operatorname{conv}_n \mathcal{G}\| \leq \|f - f_n\| \leq \frac{2^{1/p}(\rho + \varepsilon)}{n^{1-1/p}} \big(1 + \frac{(p-1)\log_2 n}{n}\big)^{1/p} \qquad \text{if } 1 < p \leq 2$$

and

$$\|f - \operatorname{conv}_n \mathcal{G}\| \leq \|f - f_n\| \leq \frac{(2p-2)^{1/2}(\rho + \varepsilon)}{n^{1/2}} \big(1 + \frac{\log_2 n}{n}\big)^{1/2} \qquad \text{if } 2 \leq p < \infty.$$

When we lift the condition to construct the approximants iteratively, it is possible to get somewhat better bounds. The improvement is only in the constant factor – in this case, however, the result is tight for $p \in (1, 2]$; for $p > 2$ it is still "only" asymptotically tight. Theorem 2.1.3 is obtained by a different approach than Theorem 2.1.2 – by using the probabilistic method in Banach spaces. We will go into more detail in Chapter 3 when we will discuss algorithmic consequences of this approach.

Theorem 2.1.3 (Rates in $\mathcal{L}^p$ spaces – probabilistic [DDGS93]) *Let $\mathcal{G}$ be a bounded subset of an $\mathcal{L}^p$-space X ($1 < p < \infty$), with $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ given. Put $q = \max\{p/(p-1), 2\}$ and let $\rho > 0$ be such, that $\|f - g\| \leq \rho$ for all $g \in \mathcal{G}$. Then for all n*

$$\|f - \operatorname{conv}_n \mathcal{G}\| \leq \frac{\rho C_p}{n^{1/q}}.$$

Here $C_p = 1$ if $p \leq 2$ and $C_p = \sqrt{2}\big(\Gamma(\frac{p+1}{2})/\sqrt{\pi}\big)^{1/p}$ for $p > 2$. For large p, $C_p \sim \sqrt{p/e}$.

Further we go into more detail regarding the constants that appear in the presented estimates.

2.1.2 Approximation Rates using $\mathcal{G}$-variation

The results of Jones and of Darken et al. were used by Kůrková [Ku97, Ku03], Kůrková, Kainen, and Kreinovich [KKK97] and Kůrková, Kainen and Vogt [KKV07]. Kůrková [Ku97] exhibited a natural way to obtain functions f and system of functions $\mathcal{G}$, such that $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ and the constant ρ from the previous section can be estimated. As we will build on and extend their results, we explain them now in some detail.

Consider a set $\mathcal{G}$ of functions, a bounded subset of a Banach space X. For convenience, we will assume that $g \in \mathcal{G}$ implies $-g \in \mathcal{G}$.[1] A function $f \in X$ can be approximated arbitrarily well by a linear combination of elements of $\mathcal{G}$ if and only if $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$. To apply the results of [Jo92, DDGS93] we need a set $\mathcal{G}'$ such that $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}'$. As

$$\operatorname{cl}\operatorname{span}\mathcal{G} = \operatorname{cl}\bigcup_{c>0}\operatorname{conv} c\mathcal{G}$$

we may try to put $\mathcal{G}' = c\mathcal{G}$ for some c. To this end, we follow Kůrková [Ku97][2] and define *$\mathcal{G}$-variation* as the Minkowski functional of the set $\operatorname{cl}\operatorname{conv}\mathcal{G}$. Note that $\mathcal{G} = -\mathcal{G}$ implies $\operatorname{conv}\mathcal{G} = \{\sum_i c_i g_i : g_i \in \mathcal{G}, c_i \geq 0, \sum_i c_i \leq 1\}$. Consequently, the set $\operatorname{cl}\operatorname{conv}\mathcal{G}$ is convex, bounded, *balanced* (that is, $h \in \operatorname{conv}\mathcal{G}$ and $|a| \leq 1$ implies $ah \in \operatorname{conv}\mathcal{G}$) and closed. Thus, we may put

$$\|f\|_{\mathcal{G}} = \inf\{c > 0 \mid f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}\} \tag{2.3}$$

and we get a norm on the subspace $\{f : \|f\|_{\mathcal{G}} < \infty\}$. Note that (2.3) defines $\|f\|_{\mathcal{G}} = \infty$ if we do not have $f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}$ for any c. This will certainly happen if $f \notin \operatorname{cl}\operatorname{span}\mathcal{G}$, in which case we can not get arbitrary close approximations. It will also happen when $f \in \operatorname{cl}\operatorname{span}\mathcal{G} \setminus \operatorname{span}\mathcal{G}$. As the next example shows, this may occur even for "reasonable" f and $\mathcal{G}$.

[1] This will simplify some of the expressions and is satisfied for the practically interesting applications.

[2] who did extend the concept of *variation with respect to half-spaces* introduced in [Ba92].

Example 2.1.4 (Infinite $\mathcal{G}$-variation) *Consider $X = \ell^2$ and let $\{e_k\}_{k=1}^{\infty}$ be the orthonormal basis ($e_k = (0, \dots, 0, 1, 0, \dots)$ is the sequence with 1 exactly at the k-th place). Then we put $f = (1/k)_{k \geq 1}$, $f_n = (1, 1/2, \dots, 1/n, 0, \dots)$, and $\mathcal{G} = \{\pm e_k \mid k \geq 1\}$. It is easy to see that f_n is the best approximant to f in $\operatorname{span}_n \mathcal{G}$ and that $\|f - f_n\|_2 \to 0$. However, $\|f_n\|_{\mathcal{G}} = \sum_{i=1}^{n} \frac{1}{i} \sim \log n$, and so $\|f\|_{\mathcal{G}} = \infty$.*

In the example above, the error of approximation of f by combination of n terms is $\sqrt{\sum_{k>n} \frac{1}{k^2}} = O(1/\sqrt{n})$, so one may think, that the assumption on finite $\mathcal{G}$-variation is not that crucial. However, this is not the case, as we will show later, in Theorem 2.1.7.

We need one more definition to describe the approach of [Ku97, Ku03]. Let $s_{\mathcal{G}} = \sup\{|g| : g \in \mathcal{G}\}$. Recall that we are assuming $\mathcal{G}$ to be bounded, so $s_{\mathcal{G}} < \infty$. A consequence of the definition of $\|f\|_{\mathcal{G}}$ is that $f \in \operatorname{cl}\operatorname{conv} c\,\mathcal{G}$ for $c \geq \|f\|_{\mathcal{G}}$. If $c = \|f\|_{\mathcal{G}}$, and $g \in c\,\mathcal{G}$, then the following bound holds:

$$\|f - g\| \leq \|f\| + \|g\| \leq 2\sup\{\|h\| : h \in c\mathcal{G}\} = 2 s_{\mathcal{G}} \|f\|_{\mathcal{G}} \,.$$

Consequently, one gets the following corollaries of results of Jones/Barron and of Darken et al., as stated by Kůrková [Ku97, Ku03], see also [KKK97, KKV07].

Theorem 2.1.5 (Rates with $\mathcal{G}$-variation – iterative [Ku97]) *Let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$ and let $\mathcal{G}$ be a bounded subset of $\mathcal{H}$. Let us denote $s_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \|g\|_{\mathcal{H}}$. Then, for every $f \in \operatorname{cl}\operatorname{span} \mathcal{G}$ with finite $\|f\|_{\mathcal{G}}$ and for every natural number n the following holds:*

$$\|f - \operatorname{span}_n \mathcal{G}\|_{\mathcal{H}} \leq \frac{\sqrt{(s_{\mathcal{G}}\|f\|_{\mathcal{G}})^2 - \|f\|_{\mathcal{H}}^2}}{\sqrt{n}}.$$

Theorem 2.1.6 (Rates with $\mathcal{G}$-variation – probabilistic [DDGS93, Ku03]) *Let $\mathcal{G}$ be a bounded subset of an $\mathcal{L}^p$-space X ($1 < p < \infty$) and $s_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \|g\|_p$. Let $f \in \operatorname{cl}\operatorname{span} \mathcal{G}$ have finite $\|f\|_{\mathcal{G}}$. Then for every n*

$$\|f - \operatorname{span}_n \mathcal{G}\|_p \leq \frac{2 C_p s_{\mathcal{G}} \|f\|_{\mathcal{G}}}{n^{1-1/t}} \,,$$

where $t = \min\{p, 2\}$, $C_p = 1$ if $p \leq 2$ and $C_p = \sqrt{2}\big(\Gamma(\frac{p+1}{2})/\sqrt{\pi}\big)^{1/p}$ for $p > 2$. For large p, $C_p \sim \sqrt{p/e}$.

Note that in both the theorems above one could instead of $\|f - \operatorname{span}_n \mathcal{G}\|$ write the more accurate expression $\|f - \operatorname{conv}_n c\mathcal{G}\|$, where $c = \|f\|_{\mathcal{G}}$. This actually gives a stronger result: we do not need to use the entire span of $\mathcal{G}$ to attain good approximation. This is interesting also from the numerical point of view: it will not happen that we need to work with big numbers to approximate small ones as convex combinations work with $c_i \in (0, 1)$, $\sum c_i = 1$.

Further one would like to estimate $\|f\|_{\mathcal{G}}$ in concrete instances of approximation schemas; we do this for neural networks in the next section. Before that, we discuss assumptions in the above estimates on rate of convergence.

Analogues of Theorem 2.1.6 are false in many spaces of interest, including $C[0,1]$ and $\mathcal{L}^1[0,1]$. By Theorem 2.3 and 2.4 in [DDGS93], in such spaces we may see arbitrary slow convergence even for elements of $\mathcal{G}$-variation equal to 1. We complement this by showing that the same happens in ℓ^p spaces for elements of infinite $\mathcal{G}$-variation. (Note that by the obvious embedding this yields the same result for $\mathcal{L}^p$ spaces as well.)

Theorem 2.1.7 (Slow rate of approximation) *Suppose $1 < p < \infty$ and let $(a_n)_{n=0}^{\infty}$ be a sequence of real numbers decreasing to 0 so that the sequence (a_n^p) is convex (that is, $a_{n-1}^p + a_{n+1}^p \geq 2a_n^p$). Then there is a set $\mathcal{G} \subseteq \ell^p$ and an element $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ so that*

$$\|f - \operatorname{span}_n \mathcal{G}\|_p = a_n.$$

So we have $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ and the rate of convergence is a_n.

This is in particular possible for $a_n = 1/n^\alpha$ (for any $\alpha > 0$), and $a_n = 1/\log^k n$ (where $\log^k$ denotes k-times iterated logarithm). More generally, we may have $a_n = 1/g(n)$ whenever g is a concave increasing function with limit ∞.

Proof: We let $\mathcal{G} = \{\pm e_i : i \geq 0\}$. Put $b_n = (a_n^p - a_{n+1}^p)^{1/p}$ for $n \geq 0$ and $f = (b_0, b_1, b_2, \dots)$. As $a_n \geq a_{n+1}$, the numbers b_n are well-defined and an easy computation shows $\|f\|_p = a_0$, in particular f is in ℓ^p. Convexity of the sequence (a_n^p) implies that $b_n^p = a_n^p - a_{n+1}^p$ is decreasing, so the element of $\operatorname{span}_n \mathcal{G}$ closest to f is $f_n = (b_0, \dots, b_{n-1}, 0, 0, \dots)$. Now

$$\|f - f_n\|_p = \sum_{i \geq n}(a_n^p - a_{n+1}^p) = a_n\,,$$

as claimed.

For the specific examples of sequences (a_n): $(x^{-p\alpha})'' = p\alpha(p\alpha+1)x^{-p\alpha-2}$ is positive for $x > 0, \alpha > 0$, so $x^{-p\alpha}$ is a convex function, thus $n^{-p\alpha}$ is a convex sequence. If $a_n = 1/g(n)$ then $a_n^p = 1/g(n)^p$. The first derivative (replacing again n by a continuous variable) is

$$\left(\frac{1}{g(x)^p}\right)' = \frac{-pg'(x)}{g(x)^{p+1}}\,,$$

which is an increasing function (as g' is decreasing and g increasing), thus a_n^p is convex as required. Computing the first derivative of the iterated logarithm reveals that it is a concave function, which finishes the proof. □

The convexity assumption in the above theorem may be a bit misleading. It is in fact possible to do away with this assumption, if we only want a lower bound on the error of approximation.

Corollary 2.1.8 (Slow rate of approximation – lower bound) *Suppose $1 < p < \infty$ and let $(a_n)_{n=0}^{\infty}$ be a strictly decreasing sequence of real numbers converging to 0. Then there is a set $\mathcal{G} \subseteq \ell^p$ and an element $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ so that*

$$\|f - \operatorname{span}_n \mathcal{G}\|_p \geq a_n.$$

Proof: We find a sequence $a'_n \geq a_n$ so that a'_n decreases to 0 and $(a'_n)^p$ is convex. This is a standard exercise in analysis, we may for example take $a'_n = \max\{a_n, (2(a'_{n-1})^p - (a'_{n-2})^p)^{1/p}]\}$. Then we apply Theorem 2.1.7 for (a'_n). □

2.2 Properties of $\mathcal{G}$-variation

So far we showed several results describing how efficiently can we approximate a function f using functions in some set $\mathcal{G}$, provided $f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}$ holds for some constant c. (Recall that cl and "approximation" are to be understood with respect to some Banach space that contains f and $\mathcal{G}$.) Several questions come up. Given $\mathcal{G}$, for which functions f such finite constant c exists? How can we estimate it?

In Theorem 2.1.7 we have shown that for elements of $\operatorname{cl}\operatorname{span}\mathcal{G}$, the rate of approximation can be arbitrarily bad; this happens if the $\mathcal{G}$-variation is infinite. Perhaps surprisingly, the situation turns out to be different for systems $\mathcal{G}$ corresponding to neural networks – such systems $\mathcal{G}$ are sufficiently rich, so that $\|f\|_{\mathcal{G}}$ is finite for large class of functions f.

In Section 2.2.1 we start with the general set-up and with bounds for the $\mathcal{G}$-variation due to Kůrková et al. [KKK97]. In Sections 2.2.2 and 2.2.3 we follow up with developing bounds for $\mathcal{G}$-variation that are applicable in a more general setting of Banach spaces. Finally, in Section 2.2.4 we clarify the dependence on the activation function.

2.2.1 $\mathcal{G}$-variation: Continuous / Heaviside Activation Functions

To answer the questions regarding existence and finiteness of c in the expression $c\mathcal{G}$ we have to be more specific as to the task investigated: We consider one-hidden-layer neural networks, which consist of interconnected computational units with activation functions depending on parameters and input variables: Consider a function $\varphi(x, a) : H \times A \to \mathbb{R}$, where $x \in H$ are inputs and $a \in A$ parameters, $H \subseteq \mathbb{R}^d$, $A \subseteq \mathbb{R}^k$. For $a \in A$ we let $\varphi_a = \varphi(\cdot, a)$ be the function parametrized by a. One-hidden-layer network with n units of type φ computes a function of d variables of the form:

$$f(x) = \sum_{i=1}^{n} w_i \varphi_{a_i}(x),$$

where $w_i \in \mathbb{R}$, $a_i \in A$, and $x \in H$. More specifically, in the case of neural networks we would typically let

$$\varphi(x,a) = \sigma(w \cdot x + \theta), \quad \text{where } a = (w,\theta) \in \mathbb{R}^{d+1}\,. \tag{2.4}$$

for perceptron–type networks or

$$\varphi(x,a) = \sigma\big(\frac{x-w}{\theta}\big), \quad \text{where } a = (w,\theta) \in \mathbb{R}^{d+1} \tag{2.5}$$

for RBF networks.

Following Kůrková [Ku03] and Kůrková et al. [KKK97], we extend this notion to a "continuum of hidden units". That is, we consider functions with integral representation

$$f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a\,, \tag{2.6}$$

with a weight function $w : A \to \mathbb{R}$. (We will discuss the relation between such generalized neural networks and ordinary neural networks later, in Section 2.3.2.)

We wish to apply the results of the previous section to a more specific case of the set $\mathcal{G}$, namely we put

$$\mathcal{G} = \{\pm\varphi_a : a \in A\}\,.$$

In the case that φ is given by (2.4), we use $\mathcal{G}_\sigma$ to denote this particular set $\mathcal{G}$. We will consider the set $\mathcal{G}$ as a subset of various spaces of functions from H to $\mathbb{R}$. The results to follow show in various circumstances how function f can be approximated by convex combinations of elements of $\mathcal{G}$. In view of Theorem 2.1.5 and 2.1.6, this amounts to estimating $\mathcal{G}$-variation of f, that is $\|f\|_{\mathcal{G}}$. (A surprising phenomenon is that $\|f\|_{\mathcal{G}_\sigma}$ does in fact not depend on σ, for a large class of functions σ. This first appeared implicitly in [KKK97], we will prove this more generally as Theorem 2.2.9.) The following result appears as Corollary 2.3 in [KKK97] (without explicitly using the term $\mathcal{G}$-variation).

Theorem 2.2.1 ($\mathcal{G}$-variation for continuous activation functions [KKK97]) *Let d, k be positive integers, $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$ compact sets. Finally, let $w \in C(A)$, $\varphi \in C(H \times A)$ and $\mathcal{G} = \{\pm\varphi_a\}$.*

Let $f \in C(H)$ be represented as

$$f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a\,. \tag{2.7}$$

Then $f \in \mathrm{cl}_C \operatorname{conv} c\,\mathcal{G}$, where $c = \int_A |w(a)|\,\mathrm{d}a$. Using our previous terminology, $\|f\|_{\mathcal{G}} \le \|w\|_1$.

We have to give some remarks here: The theorem speaks about supremum norm. Careful reader might have noticed, that theorems on rates of approximation (Section 2.1.1) do not hold for this norm. However, when the measure of H is finite (as it is when H is compact), then closure in $C(H)$ is contained in the closure in $\mathcal{L}^p(H)$. Consequently, the above theorem can be combined with the results in Section 2.1.1.

Note that in [KKK97] a slightly sharper version is presented, with $c = \int_{A_\varphi} |w(a)| \,\mathrm{d}a$, where A_φ is the set of $a \in A$, such that $\varphi(x,a) \neq 0$ for some $x \in H$. We present here the shorter statement, because for natural choices of activation function φ we have $A_\varphi = A$.

Without going into details, the main idea of the proof of Theorem 2.2.1 in [KKK97] is using the definition of Riemann integral to approximate an integral by a sum. In Section 2.2.2 we generalize this result to bounded functions in $\mathcal{L}^p$. To prove that, we will use Luzin's theorem to approximate a measurable function by a continuous function (and Fubini's theorem to deal with the error of the approximation). In Section 2.2.3 we will use more abstract functional-analytic approach to generalize this result even further, with easier (though non-constructive) proofs.

Theorem 2.2.1 was further extended in [KKK97] to $\varphi(x,a)$ given by the Heaviside function ($\vartheta(x) = 1$ for $x \geq 0$, and $\vartheta(x) = 0$ otherwise):

Theorem 2.2.2 ($\mathcal{G}$-variation for Heaviside activation functions [KKK97]) *Let d be a positive integer, $A \subseteq S^{d-1} \times \mathbb{R}$, where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$. Let H be a compact subset of $\mathbb{R}^d$ and let $f \in C(H)$ be any function that can be represented as*

$$f(x) = \int_A w(\mathbf{e}, b)\vartheta(\mathbf{e}\cdot\mathbf{x} + b)\,\mathrm{d}(\mathbf{e}, b)$$

where $w \in C(S^{d-1} \times \mathbb{R})$ is compactly supported and $supp(w) \subseteq A$. Then $f \in \mathrm{cl}_C \operatorname{conv} c\mathcal{G}_\vartheta$, where $c = \int_A |w(\mathbf{e}, b)|\,\mathrm{d}(\mathbf{e}, b)$. Using our previous notation, $\|f\|_{\mathcal{G}_\vartheta} \leq \|w\|_1$.

2.2.2 $\mathcal{G}$-variation in $\mathcal{L}^p$ Spaces

In this section we shall use the Luzin's theorem (see Theorem C.2.7 in the Appendix) to extend Theorem 2.2.1 of Kůrková, Kainen and Kreinovich [KKK97] to a more general setting, where the activation functions need no longer be continuous. This also in a sense generalizes Theorem 2.2.2: we do not prove that some function is in cl_C (closure in the supremum norm), but only that it is in $\mathrm{cl}_{\mathcal{L}^p}$ (closure in the $\mathcal{L}^p$-norm), which is, however, what we will use later to obtain rates of approximation using results in Section 2.4.

Theorem 2.2.3 ($\mathcal{G}$-variation in $\mathcal{L}^p$ spaces) *Let k, d be positive integers, let $p \in [1, \infty)$. Consider sets $A \subseteq \mathbb{R}^k$ and $H \subseteq \mathbb{R}^d$ of finite measure, that is $\lambda_k(A) < \infty$ and $\lambda_d(H) < \infty$.*

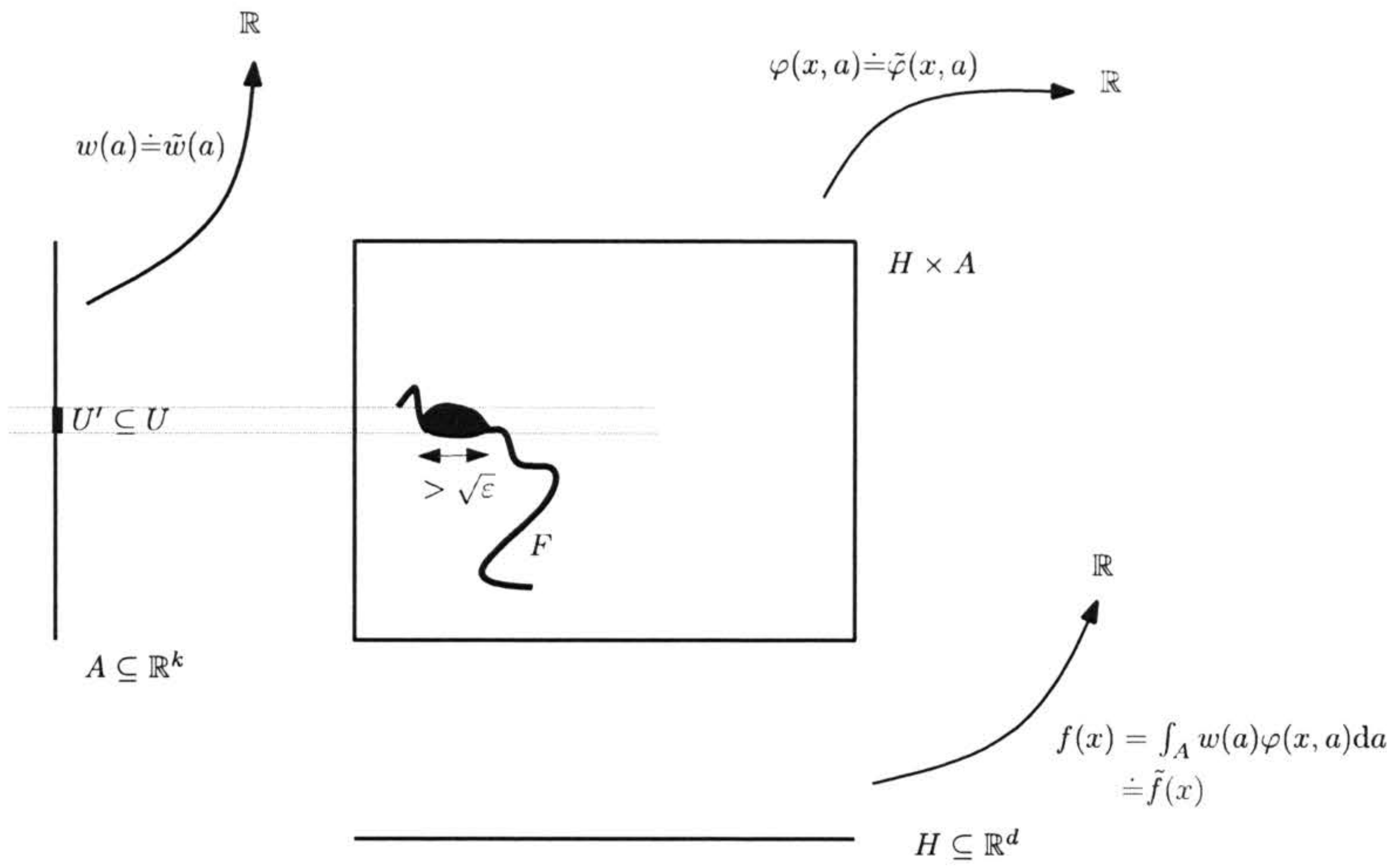
ℝ
$w(a)\doteq\tilde{w}(a)$
$\varphi(x,a)\doteq\tilde{\varphi}(x,a)$
ℝ
$H \times A$
$U' \subseteq U$
$> \sqrt{\varepsilon}$
F
ℝ
$A \subseteq \mathbb{R}^k$
$f(x) = \int_A w(a)\varphi(x,a)\mathrm{d}a$
$\doteq \tilde{f}(x)$
$H \subseteq \mathbb{R}^d$

Figure 2.1: Illustration of proof of Theorem 2.2.3.

Consider functions $w \in \mathcal{L}^1(A, \lambda_k)$ and $\varphi \in \mathcal{L}^p(H \times A, \lambda_{d+k})$ such that there exists $b \in \mathbb{R}$ so that

- *$|w| \leq b$ holds λ_k-almost everywhere on A and*
- *$|\varphi| \leq b$ holds λ_{d+k}-almost everywhere on $H \times A$.*

Put $f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a$ and $\mathcal{G} = \{\pm\varphi(\cdot, a) \mid a \in A\} \subseteq \mathcal{L}^p(H)$. Then

$$\|f\|_{\mathcal{G}} \leq \|w\|_1.$$

Note: an almost everywhere bounded function on a set of finite measure is clearly in $\mathcal{L}^p$ for any p. We still included the condition $\varphi \in \mathcal{L}^p$ and $w \in \mathcal{L}^1$ to indicate the "right" spaces to consider these functions in. In particular, let us emphasize that in the definition of $\|f\|_{\mathcal{G}}$ the $\mathcal{L}^p$ norm is used. See also the end of Section 2.2.3, where the achieved results are stated in terms of bounds on certain operators.

Proof: In the definition of $\|f\|_{\mathcal{G}}$, the underlying space (in our case $\mathcal{L}^p$) is used. Thus, the statement we are proving can be equivalently rewritten as follows

$$f(x) \in \mathrm{cl}_p \operatorname{conv}\{c\varphi(x,a) : a \in A, |c| \leq \|w\|_1\}.$$

We will prove that f can be approximated arbitrarily well (in $\mathcal{L}^p$-norm) by functions of the type

$$\sum_i c_i\varphi(x, a_i) \qquad \text{with } a_i \in A,\ c_i \geq 0, \text{ and } \textstyle\sum_i c_i \leq \|w\|_1.$$

To achieve this we first approximate functions w and φ by continuous functions (using a version of Luzin's theorem) and then apply Theorem 2.2.1.

We start, however, by showing that we can restrict to the case when A and H are compact. A finite measure subset of $\mathbb{R}^n$ can be approximated arbitrarily closely by its compact subsets (Lemma 15.3 and Theorem 26.1 of [LuMa95]). So let us choose $\varepsilon > 0$ and find compact sets A^{cp}, H^{cp}, so that $A^{cp} \subseteq A$, $H^{cp} \subseteq H$, $\lambda_k(A \setminus A^{cp}) < \varepsilon$, and $\lambda_d(H \setminus H^{cp}) < \varepsilon$. We apply the theorem for sets A^{cp}, H^{cp} instead of A, H. That is, we put

$$f^{cp}(x) = \int_{A^{cp}} w(a)\varphi(x, a)\,\mathrm{d}a$$

for $x \in H^{cp}$ and find an approximation of f^{cp} in $\mathcal{L}^p(H^{cp})$ by a function $\widetilde{f} = \sum_i c_i\varphi(x, a_i)$, where $a_i \in A^{cp}$, $c_i \geq 0$, and $\sum_i c_i \leq \|w\|_{\mathcal{L}^1(A^{cp})} \leq \|w\|_{\mathcal{L}^1(A)}(= \|w\|_1)$. We can demand

$$\|f^{cp} - \widetilde{f}\|_{\mathcal{L}^p(H^{cp})} < \varepsilon$$

and we only need to observe, that $\widetilde{f}$ (extended to H) is close to f. Clearly, $|f(x) - f^{cp}(x)| < \varepsilon \cdot b^2$ whenever $x \in H^{cp}$. For any $x \in H$, we have $\max\{|f(x)|, |\widetilde{f}(x)|\} < b\|w\|_1$. Together we have

$$\begin{aligned}
\|f - \widetilde{f}\|^p_{\mathcal{L}^p(H)} &= \int_H |f(x) - \widetilde{f}(x)|^p\,\mathrm{d}x \\
&= \int_{H^{cp}} |f(x) - \widetilde{f}(x)|^p + \int_{H\setminus H^{cp}} |f(x) - \widetilde{f}(x)|^p \\
&\leq \|(f - f^{cp}) + (f^{cp} - \widetilde{f})\|^p_{\mathcal{L}^p(H^{cp})} + \int_{H\setminus H^{cp}} (|f(x)| + |\widetilde{f}(x)|)^p \\
&\leq \left(\|f - f^{cp}\|_{\mathcal{L}^p(H^{cp})} + \|(f^{cp} - \widetilde{f})\|_{\mathcal{L}^p(H^{cp})}\right)^p + \int_{H\setminus H^{cp}} (|f(x)| + |\widetilde{f}(x)|)^p \\
&\leq \left(\varepsilon \cdot b^2 \cdot \lambda_d(H^{cp}) + \varepsilon\right)^p + (2b \cdot \|w\|_1)^p \cdot \varepsilon \\
&= O(\varepsilon), \text{ as } \varepsilon \text{ tends to } 0.
\end{aligned}$$

Thus we will assume further on, that A, H are compact sets.

Let us fix an $\varepsilon > 0$, we may assume that $\varepsilon < 1$. Using part (iii') of Theorem C.2.7 we find a continuous function $\widetilde{w}$ on A and a set $E \subseteq A$ such that

$$\widetilde{w} = w \text{ on } A \setminus E,\ |\widetilde{w}| \leq b \text{ on } A, \text{ and } \lambda_k(E) < \varepsilon\,.$$

Similarly, we find function $\widetilde{\varphi}$ on $H \times A$ and a set $F \subseteq H \times A$ such that

$$\widetilde{\varphi} = \varphi \text{ on } (H \times A) \setminus F,\ |\widetilde{\varphi}| \leq b \text{ on } H \times A, \text{ and } \lambda_{d+k}(F) < \varepsilon\,.$$

In order to apply Theorem 2.2.1 we need to define another small "exceptional set" to describe where our approximation fails, namely the set of such a's that for many x's the functions φ and $\widetilde{\varphi}$ differ on (a, x). To be precise, put

$$U' = \left\{a \in A \mid \lambda_d\{x; (a,x) \in F\} > \sqrt{\varepsilon}\right\} \cup E\,.$$

By an application of Fubini's theorem we get that $\lambda_k(U') < \sqrt{\varepsilon} + \varepsilon$. Continuity of measure implies that we can choose an open set $U \supseteq U'$ such that $\lambda_k(U) < 2\sqrt{\varepsilon}$ (recall that $\varepsilon < 1$). Finally we define

$$\widetilde{f}(x) = \int_{A \setminus U} \widetilde{\varphi}(x,a)\widetilde{w}(a)\,\mathrm{d}a$$
$$\widetilde{\mathcal{G}} = \{\pm\widetilde{\varphi}(\cdot, a) \mid a \in A \setminus U\}\,.$$

Next we use Theorem 2.2.1 for the functions $\widetilde{w}$, $\widetilde{\varphi}$, and $\widetilde{f}$, set $\widetilde{\mathcal{G}}$ and with the set $A \setminus U$ in place of A. We conclude that

$$\|\widetilde{f}\|_{\widetilde{\mathcal{G}}} \leq \|\widetilde{w}\|_1\,.$$

This means that there is $n \in \mathbb{N}$, and $c_i \in \mathbb{R}$, $a_i \in A \setminus U$ $(i = 1, \dots, n)$ such that $\sum_{i=1}^n |c_i| \leq \|\widetilde{w}\|_1$, and for the function $\widetilde{f}_1$ defined by

$$\widetilde{f}_1(x) = \sum_{i=1}^n c_i \widetilde{\varphi}(x, a_i)$$

we have $|\widetilde{f}(x) - \widetilde{f}_1(x)| < \varepsilon$ for all $x \in H$. We use these parameters to define our desired approximant, f_1:

$$f_1(x) = \sum_i c_i \varphi(x, a_i)\,.$$

We have $\|\widetilde{w}\|_1 = \int_{A\setminus U} |\widetilde{w}| = \int_{A\setminus U} |w| \leq \int_A |w| = \|w\|_1$. To finish the proof, we need to establish an upper bound on $\|f - f_1\|_p$. To this end, we first use the triangle inequality

$$\|f - f_1\|_p \leq \|f - \widetilde{f}\|_p + \|\widetilde{f} - \widetilde{f}_1\|_p + \|\widetilde{f}_1 - f_1\|_p\,.$$

Now we deal with these three terms one by one.

(A) $\|\widetilde{f} - \widetilde{f}_1\|_p$
We know that $|\widetilde{f}(x) - \widetilde{f}_1(x)| < \varepsilon$ on H, thus $\|\widetilde{f} - \widetilde{f}_1\|_p < \varepsilon\lambda_d(H)$.

(B) $\|\widetilde{f}_1 - f_1\|_p$

Observe first, that

$$\begin{aligned} |\widetilde{f}_1(x) - f_1(x)| &= \Big|\sum_i c_i\big(\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big)\Big| \\ &\leq \sum_i |c_i|\big|\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big| . \end{aligned}$$

Due to the bounds on φ and $\widetilde{\varphi}$, each of the absolute values in the last sum is at most $2b$ for every $x \in H$. Moreover, each of these absolute values is equal to 0 for most of $x \in H$, namely up to a set of measure $\sqrt{\varepsilon}$ (recall that $a_i \notin U$). Now, we have

$$\begin{aligned} \|\widetilde{f}_1 - f_1\|_p^p &= \int_H |\widetilde{f}_1 - f_1|^p \\ &= \int_H |\widetilde{f}_1 - f_1|^{p-1}|\widetilde{f}_1 - f_1| \\ &\leq \int_H (2b\|w\|_1)^{p-1}|\widetilde{f}_1 - f_1| \\ &\leq (2b\|w\|_1)^{p-1} \int_H \sum_{i=1}^n |c_i|\big|\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big| \,\mathrm{d}x \\ &= (2b\|w\|_1)^{p-1} \sum_{i=1}^n |c_i| \int_H \big|\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big| \,\mathrm{d}x \end{aligned}$$

According to the previous paragraph, we can bound each of the integrals in the last sum by $2b\sqrt{\varepsilon}$, yielding

$$\begin{aligned} \|\widetilde{f}_1 - f_1\|_p^p &\leq (2b\|w\|_1)^{p-1} \sum_i |c_i| \cdot 2b\sqrt{\varepsilon} \\ &\leq (2b\|w\|_1)^{p-1} \|\widetilde{w}\|_1 \cdot 2b\sqrt{\varepsilon} \\ &\leq (2b\|w\|_1)^{p-1} \|w\|_1 \cdot 2b\sqrt{\varepsilon} \\ &= (2b\|w\|_1)^p \sqrt{\varepsilon} . \end{aligned}$$

(C) $\|f - \widetilde{f}\|_p$

Here we proceed similarly as in part (B):

$$\begin{aligned} |f(x) - \widetilde{f}(x)| &= \left| \int_{A\setminus U} (w(a)\varphi(x, a) - \widetilde{w}(a)\widetilde{\varphi}(x, a)) \,\mathrm{d}a + \int_U w(a)\varphi(x, a) \,\mathrm{d}a \right| \\ &\leq \int_{A\setminus U} \big|w(a)\varphi(x, a) - \widetilde{w}(a)\widetilde{\varphi}(x, a)\big| \,\mathrm{d}a + \int_U \big|w(a)\varphi(x, a)\big| \,\mathrm{d}a \end{aligned}$$

We will use that both $|w(a)\varphi(x, a)|$ and $|\widetilde{w}(a)\widetilde{\varphi}(x, a)|$ are at most b^2, the set U is small, and $|w(a)\varphi(x, a) - \widetilde{w}(a)\widetilde{\varphi}(x, a)|$ is "usually" zero. In particular, $|f(x) - \widetilde{f}(x)| \leq 2b^2\lambda_k(A)$.

$$\begin{aligned}
\|f-\widetilde{f}\|_p^p &= \int_H |f-\widetilde{f}|^p \\
&= \int_H |f-\widetilde{f}|^{p-1}|f-\widetilde{f}| \\
&\le \int_H (2b^2\lambda_k(A))^{p-1}|f-\widetilde{f}| \\
&\le (2b^2\lambda_k(A))^{p-1}\int_H\left(\int_{A\setminus U}|w(a)\varphi(x,a)-\widetilde{w}(a)\widetilde{\varphi}(x,a)|\,\mathrm{d}a\right. \\
&\qquad \left.+\int_U |w(a)\varphi(x,a)|\,\mathrm{d}a\right)
\end{aligned}$$

Next we use Fubini's theorem – note that we integrate a nonnegative measurable function:

$$\begin{aligned}
&\le (2b^2\lambda_k(A))^{p-1}\left(\int_{H\times(A\setminus U)}|w(a)\varphi(x,a)-\widetilde{w}(a)\widetilde{\varphi}(x,a)|\,\mathrm{d}(x,a)\right. \\
&\qquad \left.+\int_{H\times U}|w(a)\varphi(x,a)|\,\mathrm{d}(x,a)\right) \\
&\le (2b^2\lambda_k(A))^{p-1}\left(\int_F|w(a)\varphi(x,a)-\widetilde{w}(a)\widetilde{\varphi}(x,a)|\,\mathrm{d}(x,a)+\int_{H\times U}b^2\,\mathrm{d}(x,a)\right) \\
&\le (2b^2\lambda_k(A))^{p-1}\left(\varepsilon 2b^2+\lambda_d(H)2\sqrt{\varepsilon}\,b^2\right)
\end{aligned}$$

By combining (A), (B), and (C) we see that we can choose ε small enough to get as good approximation as desired. □

We have proven $\|f\|_{\mathcal{G}} \le \|w\|_1$ for f computed by one-hidden-layer neural network with $\mathcal{L}^\infty$ (almost everywhere bounded) activation function. Together with Theorem 2.1.6 we derive rates of approximation for this approximation schema:

Corollary 2.2.4 (Rates in $\mathcal{L}^p$) *Let k, d be positive integers, A a compact subset of $\mathbb{R}^k$ and H a compact subset of $\mathbb{R}^d$. Let $w \in \mathcal{L}^1(A,\lambda_k)$ and $\varphi \in \mathcal{L}^p(H\times A,\lambda_{d+k})$ for some $1<p<\infty$. Additionally let w and φ be bounded almost everywhere on A and $H\times A$ respectively. Let $\mathcal{G}=\{\varphi(\cdot,a) : a\in A\}$ be bounded and $s_{\mathcal{G}}=\sup_{\varphi\in\mathcal{G}}\|\varphi\|_p$. Let f be any function that can be represented as $f(x)=\int_A w(a)\varphi(x,a)\,\mathrm{d}a$. Then*

$$\|f-\operatorname{span}_n\mathcal{G}\|_p \le \frac{2C_p s_{\mathcal{G}}\|w\|_1}{n^{1-1/t}},$$

where $t=\min\{p,2\}$ and $C_p=1$ if $p\le 2$ and $C_p=\sqrt{2}\big(\Gamma(\frac{p+1}{2})/\sqrt{\pi}\big)^{1/p}$ for $p>2$.

As in Theorems 2.1.5 and 2.1.6 one could write instead of $\|f - \operatorname{span}_n \mathcal{G}\|_p$ the more accurate expression $\|f - \operatorname{conv}_n c\mathcal{G}\|_p$, with $c = \|w\|_1$.

2.2.3 Estimates of $\mathcal{G}$-variation via Hahn-Banach Theorem

In this section we provide a generalization (and also an alternative proof) of the result of the previous section. We extend Theorems 2.2.1 and 2.2.3 (the estimate of $\mathcal{G}$-variation) to more general Banach spaces in place of $C(K)$, resp. $\mathcal{L}^p \cap \mathcal{L}^\infty$. We also generalize the integral formula to employment of signed measures.

The proof of the main result in this section is in fact shorter than previously presented proofs of weaker results. This is achieved by using more advanced tools from functional analysis. A slight drawback is that this approach (relying on Hahn-Banach theorem) is no longer constructive: given formula $f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a$, the previous proofs suggested a technique to really obtain a sequence of convex combinations that converge to f. The functional-analytic approach, on the other hand, only proves that such a sequence exists. This, however, has no implications for our present considerations; we will revisit this issue in Chapter 3.

The main tool we will use in this section is the following version of geometric Hahn-Banach theorem [Lax02, LuMa95].

Theorem 2.2.5 (Geometric Hahn-Banach [Lax02]) *Let X be a Banach space, consider $x \in X$ and $T \subseteq X$. Then $x \in \operatorname{cl}\operatorname{conv} T$, unless there is a functional $\ell \in X^*$ and $z \in \mathbb{R}$ such that*

$$\ell(x) > z \quad \textit{and} \quad \ell(t) < z \textit{ for every } t \in T\,. \tag{2.8}$$

We also refer the reader to the appendix for description of dual spaces C^*, $(\mathcal{L}^p)^*$.

First we present an alternative proof (a generalization) of Theorem 2.2.1.

Theorem 2.2.6 ($\mathcal{G}$-variation for continuous activation functions using measure) *Let d, k be positive integers, $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$ compact sets. Suppose ν is a signed Radon measure on A. Finally, let $\varphi \in C(H \times A)$ and $\mathcal{G} = \{\pm\varphi_a\}$.*

Let the function $f \in C(H)$ be represented as $f(x) = \int_A \varphi(x,a)\,\mathrm{d}\nu(a)$.

Then $f \in \operatorname{cl}_C \operatorname{conv} c\mathcal{G}$, where $c = \|\nu\|$ is the norm of ν. Using our previous terminology, $\|f\|_{\mathcal{G}} \leq \|\nu\|$.

Proof: Let (P, N) be a Hahn decomposition for the measure ν. That is, A is the disjoint union of P and N, and $\nu(E) \geq 0$ (resp. ≤ 0) whenever $E \subseteq P$ (resp. $E \subseteq N$). Define the

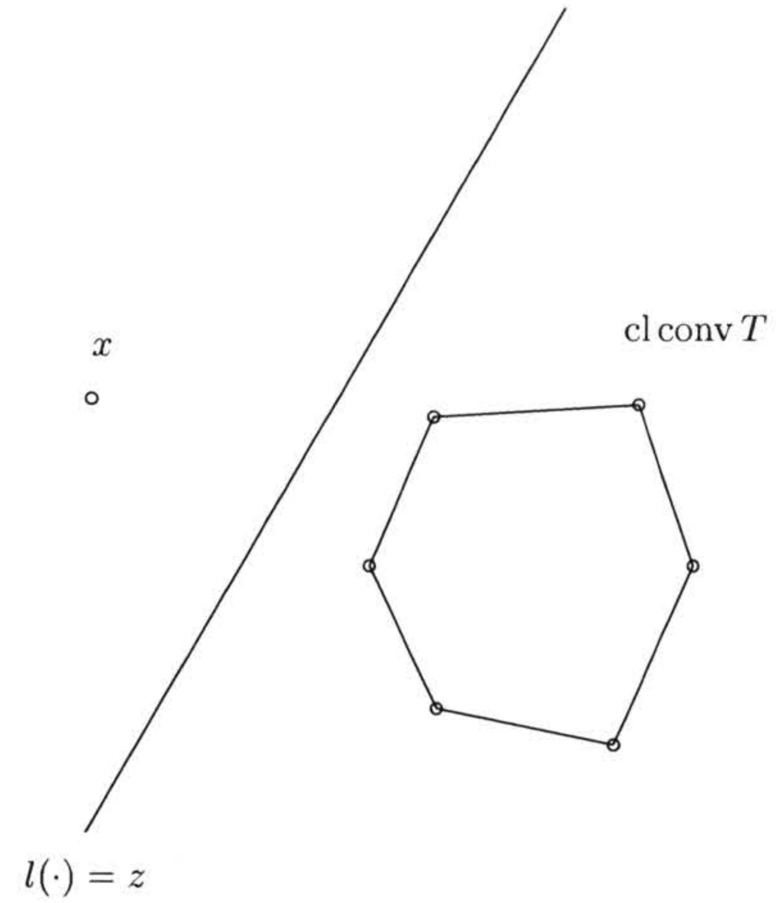

Figure 2.2: An illustration of the geometric Hahn-Banach theorem.

function s by

$$s(a) = \begin{cases} +1 & \text{for } a \in P \\ -1 & \text{for } a \in N \end{cases}$$

that is, $s(a)$ is the "sign of ν at a". In particular, we have $c = \|\nu\| = \int_A s(a)\,\mathrm{d}\nu(a)$.

If $c = 0$ then $\nu(E) = 0$ for any set E, thus $f(x) \equiv 0$ and the assertion is true. So we may assume $c > 0$; note that $c = \|\nu\| < \infty$, as ν is a signed measure.

We need to prove that $f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}$. Suppose the contrary; according to the geometric Hahn-Banach theorem (Theorem 2.2.5), there is a constant z, and a functional $\ell \in C(H)^*$ such that (2.8) is true with $x = f$ and $T = c\mathcal{G}$. Let μ be the signed measure defining ℓ as in (C.1). We have $\ell(f) > z$ and for every a

$$\ell(\pm c\varphi_a) = \pm c \int_H \varphi_a \,\mathrm{d}\mu < z\,. \tag{2.9}$$

By definition,

$$\begin{aligned} \ell(f) &= \int_H f\,\mathrm{d}\mu \\ &= \int_H \int_A \varphi(x,a)\,\mathrm{d}\nu(a)\,\mathrm{d}\mu(x)\,. \end{aligned}$$

Note, that $\varphi(x, a)$ is a continuous function and it is only integrated over a compact set. So, the integral of the absolute value is finite and, obviously, both ν and μ are σ-finite (they are even finite). Thus we can use Fubini's theorem to get

$$\begin{aligned}\ell(f) &= \int_A \int_H \varphi(x, a)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(a) \\ &= \frac{1}{c}\int_A s(a) \int_H s(a)c\varphi(x, a)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(a)\,.\end{aligned}$$

Next, we use (2.9) and the definition of s

$$\begin{aligned}&\le \frac{1}{c}\int_A s(a)z\,\mathrm{d}\nu(a) \\ &= \frac{z}{c}\|\nu\| \\ &= z\,.\end{aligned}$$

This contradiction finishes the proof. □

As a corollary, we obtain an alternative proof for Theorem 2.2.1 (that appears as Corollary 2.3 in [KKK97]). Actually, we obtain a stronger version, as we do not require w to be continuous.

Corollary 2.2.7 ($\mathcal{G}$-variation for continuous activation functions (weaker assumptions)) *Let d, k be positive integers, $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$ compact sets. Let $w \in \mathcal{L}^1(A)$, $\varphi \in C(H \times A)$ and $\mathcal{G} = \{\pm\varphi_a\}$.*

Finally let the $f \in C(H)$ be represented as $f(x) = \int_A w(a)\varphi(x, a)\,\mathrm{d}a$.

Then $f \in \mathrm{cl}_C \operatorname{conv} c\mathcal{G}$, where $c = \int_A |w(a)|\,\mathrm{d}a$. Using our previous terminology, $\|f\|_{\mathcal{G}} \le \|w\|_1$.

Proof: We define a signed measure ν by letting for any Lebesgue-measurable $E \subseteq A$

$$\nu(E) = \int_E w(a)\,\mathrm{d}a\,.$$

We easily get

$$\|\nu\| = \int_A |w(a)|\,\mathrm{d}a = \|w\|_1$$

and

$$f(x) = \int_A w(a)\varphi(x, a) = \int_A \varphi(x, a)\,\mathrm{d}\nu(a)\,.$$

It only suffices to apply Theorem 2.2.6 and we conclude. □

Now we present a theorem bounding $\mathcal{G}$-variation for $\mathcal{L}^p$ activation function.

Theorem 2.2.8 ($\mathcal{G}$-variation for $\mathcal{L}^p$ activation functions using measure) *Let d, k be positive integers, let $p \in (1, \infty)$. Consider sets $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$, and a signed Radon measure ν on A. Let φ be a measurable function such that there is $b \in \mathbb{R}$ so that for any $a \in A$ the function $\varphi_a = \varphi(\cdot, a)$ is in $\mathcal{L}^p(H, \lambda_d)$, and $\|\varphi_a\|_p \leq b$. Put $\mathcal{G} = \{\pm\varphi_a, a \in A\}$.*

Let the function f be represented as $f(x) = \int_A \varphi(x, a)\,\mathrm{d}\nu(a)$ (that is, the integral exists for almost every x).

Then $f \in \mathrm{cl}_{\mathcal{L}^p} \operatorname{conv} c\,\mathcal{G}$, where $c = \|\nu\|$ is the norm of ν. Using our previous terminology, $\|f\|_{\mathcal{G}} \leq \|\nu\|$.

Proof: We proceed similarly as in the proof of Theorem 2.2.6. Again, we use Hahn decomposition of ν to define $s(a)$ as the "sign of ν at a" and put $c = \|\nu\| = \int s(a)\,\mathrm{d}\nu(a)$. We first remark that f is in $\mathcal{L}^p(H, \lambda_d)$: For any linear functional l in $(\mathcal{L}^p)^*$ we will derive in (2.10) that $l(f)$ is finite. This shows, that f is an element in $(\mathcal{L}^p)^{**}$, and as $\mathcal{L}^p$ is reflexive, we see that indeed $f \in \mathcal{L}^p$.

If $f \notin \mathrm{cl} \operatorname{conv} c\,\mathcal{G}$ then, using Theorem 2.2.5 again, there is an $\ell \in (\mathcal{L}^p)^*$ such that

$$\ell(f) > z > \ell(\pm c\varphi_a)$$

for some z and all $a \in A$. Let $\psi \in \mathcal{L}^q$ (with $1/p + 1/q = 1$) be the representant of ℓ. Similarly as before, we get

$$\begin{aligned}\ell(f) &= \int_H f\psi \\ &= \int_H \int_A \varphi(x, a)\psi(x)\,\mathrm{d}\nu(a)\,\mathrm{d}\lambda(x)\end{aligned}$$

To apply Fubini's theorem, we observe that $\varphi(x, a)\psi(x)$ is a measurable function, and that $\int_H |\varphi(x, a)\psi(x)| \leq \left(\int_H |\varphi(x, a)|^p\right)^{1/p} \left(\int_H |\psi(x)|^q\right)^{1/q} = \|\varphi_a\|_p \|\psi\|_q$ (Hölder inequality for $|\varphi_a|$ and $|\psi|$). Consequently,

$$\int_H \int_A |\varphi\psi| \leq \|\nu\| \cdot b \cdot \|\psi\|_q \tag{2.10}$$

and we may use Fubini's theorem to obtain

$$\begin{aligned}\ell(f) &= \int_A \int_H \varphi(x, a)\psi(x)\,\mathrm{d}\lambda(x)\,\mathrm{d}\nu(a) \\ &= \frac{1}{c} \int_A s(a) \int_H s(a) c\varphi(x, a)\psi(x)\,\mathrm{d}\lambda(x)\,\mathrm{d}\nu(a) \\ &\leq \frac{1}{c} \int_A s(a) z\,\mathrm{d}\nu(a) \\ &= z\end{aligned}$$

Again, by Hahn-Banach Theorem we found a contradiction. □

Note, that we get Theorem 2.2.3 as a corollary. Also we recovered a version of a result of [KKK97] (Theorem 2.2.2). We do not obtain that f is in the closure in the supremum norm. This, however, is not needed to apply the Maurey-Jones-Barron theorem or any of the other theorems on rates of approximation.

The technique of using Hahn-Banach theorem can be applied to other spaces as well – all we need is to have elements of the dual to that space "behave nicely with respect to integration", that is, some version of Fubini's theorem holds. Natural candidates to consider in this setting would be Sobolev spaces, leading to a simultaneous approximation of a function and its derivatives. We will not elaborate on this topic, as it was already researched by Hornik et al. [HSW89].

Before we end this section, let us remark that we can express the obtained results in terms of functional analysis. Define an operator T_φ by

$$T_\varphi(\nu) = \int_A \varphi(\cdot, a)\, \mathrm{d}\nu(a)\,.$$

We consider T_φ as an operator from $M(A)$ (the space of all signed measures on A) to a subspace of $C(H)$ (or $\mathcal{L}^p(H)$, etc.) with the norm $\|\cdot\|_{\mathcal{G}}$ (the subspace consists of functions of finite $\|\cdot\|_{\mathcal{G}}$-norm). Then the above results say that the operator norm of T_φ is at most (in fact exactly) equal to 1.

2.2.4 Surprising Property of $\mathcal{G}$-variation

In this short section we are going to prove a surprising property of the $\mathcal{G}_\sigma$-variation, namely its independence of σ for a large class of activation functions.

We will assume σ is a sigmoidal function (that is $\lim_{x\to-\infty}\sigma(x) = 0$, $\lim_{x\to\infty}\sigma(x) = 1$ and σ is nondecreasing). Note that we do not require continuity: after all, from practical perspective, the easiest functions to evaluate are step functions, that is linear combinations of characteristic functions of intervals.

If we consider $\mathcal{G}_\sigma$ as a subset of $\mathcal{L}^p(H)$ (for a compact set H) then the $\mathcal{G}_\sigma$-variation $\|f\|_{\mathcal{G}_\sigma}$ does not depend on σ. A version of this result appears implicitly in [KKK97]. However, there σ was assumed to be either continuous, or the Heaviside function.

Theorem 2.2.9 ($\mathcal{G}_\sigma$-variation independent of σ) *Suppose $1 < p < \infty$, let $H \subseteq \mathbb{R}^d$ be a compact set and $f \in \mathcal{L}^p(H)$. Then there is $c_f \in [0, \infty]$ so that for any sigmoidal function σ we have*

$$\|f\|_{\mathcal{G}_\sigma} = c_f\,.$$

(Recall that sigmoidal function means a function $\mathbb{R} \to \mathbb{R}$ that is nondecreasing with limits in $\pm\infty$ being 0 and 1; we do not demand continuity.)

Proof: We put $c = \|f\|_{\mathcal{G}_\vartheta}$ and show that for any sigmoidal function σ we have $\|f\|_{\mathcal{G}_\sigma} = c$. To this end, we prove an auxiliary claim first, reducing to a question about functions of one real variable. Then we utilize this claim by letting either σ_1 or σ_2 be the Heaviside function ϑ.

Claim Let σ_1, σ_2 be two sigmoidal functions so that for each finite interval $J \subseteq \mathbb{R}$

$$\sigma_2(t) \in \operatorname{cl}\operatorname{conv}\{\sigma_1(rt+s) : r, s \in \mathbb{R}\},$$

where the closure is taken in $\mathcal{L}^p(J)$. Then for any function $f \in \mathcal{L}^p(H)$ we have $\|f\|_{\mathcal{G}_{\sigma_1}} \leq \|f\|_{\mathcal{G}_{\sigma_2}}$.

Indeed, by definition of the $\mathcal{G}$-variation, there are functions $f_{apx}(x)$ that are arbitrarily close to $f(x)$ (in $\mathcal{L}^p(H)$-norm), and that are of form

$$f_{apx} = \sum_i c_i \sigma_2(a_i \cdot x + b_i), \qquad \sum_i |c_i| \leq \|f\|_{\mathcal{G}_{\sigma_2}}.$$

If the assumptions of the Claim are satisfied, then we can approximate each of $\sigma_2(t)$ by a finite convex combination $g_i(t) = \sum_j k_{i,j}\sigma_1(r_{i,j}t + s_{i,j})$ in $\mathcal{L}^p(J)$ for a finite interval J containing $\cup_i\{a_i \cdot x + b_i : x \in H\}$. If we put $E(t) = |g_i(t) - \sigma_2(t)|^p$, we get

$$\int_H E(a_i \cdot x + b_i)\,\mathrm{d}x \leq \frac{B}{|a_i|}\int_J E(t)\,\mathrm{d}t.$$

Here B is the upper bound on $\lambda_{d-1}(H_{a_i,c})$, where $H_{a_i,c} = \{x \in H : a_i \cdot x = c\}$ are the sections of H. In particular, B can be chosen as a constant depending only on H. So we can for any given $\varepsilon > 0$ find functions g_i so that $\|g_i(a_i \cdot x + b_i) - \sigma_2(a_i \cdot x + b_i)\|_{\mathcal{L}^p(H)} < \varepsilon$. By triangle inequality it follows that $\|f_{apx}(x) - \sum_i c_i g_i(a_i \cdot x + b_i)\|_{\mathcal{L}^p(H)} < \varepsilon$, too.

Also $\sum_i\sum_j |c_i k_{i,j}| = \sum_i |c_i| \sum_j k_{i,j} = \sum_i |c_i| \leq \|f\|_{\mathcal{G}_{\sigma_2}}$, which finishes the proof of the claim.

(A) $\|f\|_{\mathcal{G}_\sigma} \leq \|f\|_{\mathcal{G}_\vartheta}$ for any f (This part appears in [KKK97], we repeat the simple argument for reader's convenience.) According to the Claim, we only need to observe, that for any M, $\|\sigma(Nt) - \vartheta(t)\|_{\mathcal{L}^p([-M,M])}$ tends to zero as $N \to \infty$. To observe this, we only need for any $\varepsilon > 0$ choose N so large that $\sigma(\varepsilon \cdot N) > 1 - \varepsilon$ and $\sigma(-\varepsilon \cdot N) < \varepsilon$. Then

$$\begin{aligned}\|\sigma(Nt) - \vartheta(t)\|^p_{\mathcal{L}^p([-M,M])} &\leq \int_{[-\varepsilon,\varepsilon]} |\sigma(Nt) - \vartheta(t)|^p + \int_{[-M,-\varepsilon]\cup[\varepsilon,M]} |\sigma(Nt) - \vartheta(t)|^p \\ &\leq 2\varepsilon \cdot 1 + 2M \cdot \varepsilon^p\end{aligned}$$

A choice of arbitrarily small ε finishes the proof.

(B) $\|f\|_{\mathcal{G}_\vartheta} \leq \|f\|_{\mathcal{G}_\sigma}$ for any f Now we pursue with the more surprising part of the proof. We will actually prove something stronger than required by the Claim. Namely, for any $\varepsilon > 0$ there is a function g of form

$$g(t) = \sum_{i=1}^{k} c_i \vartheta(t - b_i) \tag{2.11}$$

so that $|g(t) - \sigma(t)| \leq \varepsilon$ for every $t \in \mathbb{R} \setminus \{b_1, \ldots, b_k\}$. This clearly implies that g and σ are close in $\mathcal{L}^p$ norm on any set of finite measure.

We will construct g by inductively finding points b_i. We will rely heavily on the result from first-year analysis: all points of discontinuity of a nondecreasing function σ are jumps, that is, for any x there exists $\sigma(x_-)$ (the limit from the left) and $\sigma(x_+)$ (the limit from the right).

To start the process, we put $b_0 = -\infty$ and $h_0 = 0$. Now, whenever b_i was defined (and $b_i < \infty$), we put

$$b_{i+1} = \sup\{x \in \mathbb{R} : \sigma(x) \leq h_i + \varepsilon\}$$

and

$$h_{i+1} = \sigma((b_{i+1})_+) .$$

Now for any $y > b_{i+1}$ we have $\sigma(y) > h_i + \varepsilon$, so in particular $h_{i+1} = \sigma((b_{i+1})_+) \geq h_i + \varepsilon$. This implies our process ends after at most $\lfloor 1/\varepsilon \rfloor + 1$ steps (recall that σ is bounded by 1). When we reach $b_{k+1} = \infty$, we define g by (2.11) with $c_i = h_i - h_{i-1}$.

Next, observe that for any $i = 0, 1, \ldots, k$ we have $\sigma((b_{i+1})_-) \leq h_i + \varepsilon$ (by definition of b_{i+1}. As for $t \in [b_i, b_{i+1})$ the constructed function $g(t)$ is equal to $\sum_{1 \leq j \leq i}(h_j - h_{j-1}) = h_i$, we see that $|g(t) - \sigma(t)| \leq \varepsilon$ unless t is one of the points b_i. Thus we conclude that (B) holds as well and this finishes the proof. □

Let us now comment about implications of the above result. The result applies whenever we want to estimate the rate of convergence in $\mathcal{L}^p$ norm, using results of Maurey, Jones, Barron, and Darken et al. As far as these estimates are concerned, all sigmoidal functions are of equal utility. Let us mention some limitations for practical applications, though:

- In part (A) of the above proof ("any σ is at least as good as the Heaviside function") we need to use large multiplicative coefficients, which is not numerically feasible.

- It says nothing about convergence in the supremum norm. (For supremum norm the analog of Maurey-Jones-Barron theorem is false [DDGS93]. See the discussion preceding Theorem 2.1.7 for more details.)

- To elaborate further on the previous point, we can extend part (B) of the above proof to get the following simple bound for estimates in the supremum norm: If $\|f - \operatorname{span}_n \mathcal{G}_\sigma\| \leq \varepsilon$ then $\|f - \operatorname{span}_N \mathcal{G}_\vartheta\| \leq 2\varepsilon$ with $N = n/\varepsilon$.

- Theorem 2.2.9 implies equality of *bounds* on the rate of convergence. It is quite possible, that for problems of practical interest, convergence will be faster (but perhaps not for all activation functions). This question deserves further study.

2.3 Integral Representations

As we have seen in the previous section we needed integral representation of the function f to estimate its $\mathcal{G}$-variation. Thus a natural question is: when does such a representation exist?.

In Section 2.3.1 we present several specific examples of functions where integral representation is known to exist. In Section 2.3.2 we discuss relationship between integral representation and neural network with number of units going to infinity.

2.3.1 Integral Representations for Specific Classes of Functions

In this section we present known integral representations for specific types of function f.

A. Absolutely continuous functions Let us consider one-dimensional functions first. Let f be an absolutely continuous function on $[a, b]$. It is known (see, e.g., Corollary 23.5 of [LuMa95]) that f' exists almost everywhere as a function in $\mathcal{L}^1[a, b]$. Moreover,

$$f(x) = f(a) + \int_a^x f'(t)\,\mathrm{d}t\,.$$

Assume now that $f(a) = 0$ and recall that $\vartheta(x)$ is the Heaviside function ($\vartheta(x) = 1$ if $x \geq 0$, $\vartheta(x) = 0$ otherwise). Then the above formula can be expressed as

$$f(x) = \int_a^b f'(t)\vartheta(x - t)\,\mathrm{d}t\,. \tag{2.12}$$

B. Integral representation of $f(x)$ based on Poisson's theorem / inverse Radon transform To apply the above mentioned types of bounds we need function f expressed in the form of an integral as in (2.7). To this end, the following consequence of Poisson's theorem of potential theory was proved in [KKK97]. (The same result, but only for functions in the Schwartz space, is obtained in [Ito91] using inverse Radon transform [He99].

In [KKV06] a variant of Theorem 2.3.1 (for functions of *weakly controlled decay*) is proved and in [KKV07] this is utilized to find bounds on $\mathcal{G}$-variation in terms of the Sobolev norm.)

Let D_e be the operator of directional derivative in the direction given by e, that is $D_e f(y) = \lim_{h \to 0} \frac{f(y+h \cdot e)-f(y)}{h}$. For a positive integer d, $D_e^{(d)}$ is d-fold iteration of D_e. Note, that if f is C^d, that is the partial derivatives of order at most d exist and are continuous, then one can use the partial derivatives to express all directional derivatives. Finally, $H_{eb} = \{y \in \mathbb{R}^d : y \cdot e + b = 0\}$.

Theorem 2.3.1 (Integral representation in $C^d(\mathbb{R}^d)$ [KKK97]) *For every odd positive integer d every compactly supported function $f \in C^d(\mathbb{R}^d)$ can be represented as*

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f(y) \, \mathrm{d}y \Big) \vartheta(e \cdot x + b) \, \mathrm{d}b \, \mathrm{d}e$$

where $a_d = \frac{(-1)^{(d-1)/2}}{2(2\pi)^{d-1}}$.

Thus from Theorems 2.1.6, 2.1.5, and 2.2.2 it follows that if $f \in C^d(\mathbb{R}^d)$, then it can be approximated efficiently by neural networks with Heaviside activation functions, that is with rate $O(\frac{1}{n^{1-1/p}})$ in the space $\mathcal{L}^p$, $1 < p \leq 2$ and with rate $O(\frac{1}{\sqrt{n}}$ for $p > 2$. We can get the same conclusions with somewhat weaker assumptions. Namely, instead of requiring d continuous derivatives, we only ask for weak derivatives (as members of an $\mathcal{L}_p$ space):

Theorem 2.3.2 (Integral representation in $W^{d,p}(\Omega)$) *Let d be an odd integer, $p > 1$ and let $\Omega \subseteq \mathbb{R}^d$ be a bounded open set with a C^1 boundary. Then every $f \in \mathcal{W}^{d,p}(\Omega)$ can be represented as*

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f(y) \, \mathrm{d}y \Big) \vartheta(e \cdot x + b) \, \mathrm{d}b \, \mathrm{d}e$$

where $a_d = \frac{(-1)^{(d-1)/2}}{2(2\pi)^{d-1}}$.

Proof: Let f be a function in $\mathcal{W}^{d,p}$. It is known [Lan93] that we can find functions $f_n \in C^\infty(\Omega)$ such that $\|f_n - f\|_{d,p} < 1/n$. For f_n we know the formula

$$f_n(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f_n(y) \, \mathrm{d}y \Big) \vartheta(e \cdot x + b) \, \mathrm{d}b \, \mathrm{d}e \qquad (2.13)$$

from several sources ([He99], [Ito91], Theorem 2.2.1, [KKV07]). It remains to show, how we can derive the same formula for f itself.

By definition of $\mathcal{W}^{d,p}$-norm we easily conclude that $D_e^{(d)} f_n \to D_e^{(d)} f$ in $\mathcal{L}_p$-sense. For a given $\varepsilon > 0$ let us choose n so that $|D_e^{(d)} f_n - D_e^{(d)} f| < \varepsilon$. Then we have for each $e \in S^{d-1}$

$$\begin{aligned}
&\left| \int_{\mathbb{R}} \Big(\int_{H_{eb}} (D_e^{(d)} f_n(y) - D_e^{(d)} f(y)) \,\mathrm{d}y \Big) \vartheta(e \cdot x + b) \,\mathrm{d}b \right| \\
&\leq \int_{\mathbb{R}} \Big(\int_{H_{eb}} |D_e^{(d)} f_n(y) - D_e^{(d)} f(y))| \,\mathrm{d}y \Big) \vartheta(e \cdot x + b) \,\mathrm{d}b \\
&\qquad\qquad \leq \int_{\Omega} |D_e^{(d)} f_n(y) - D_e^{(d)} f(y))| \,\mathrm{d}y
\end{aligned}$$

and by the power mean inequality we get that for some C depending only on the measure of Ω

$$\begin{aligned}
&\leq C \|f_n - f\|_{d,p} \\
&\leq C\varepsilon
\end{aligned}$$

Consequently, the right-hand side of (2.13) for f and for f_n differ by at most $a_d \lambda_{d-1}(S^{d-1}) C\varepsilon$. The difference of the left-hand sides of (2.13) can be estimated too.

$$\|f_n - f\|_{C(\Omega)} \leq C' \|f_n - f\|_{d,p} \leq c_1 \varepsilon \,.$$

Here we are using the Sobolev inequality; c_1 depends only on *d*, *p*, and Ω. It follows, that there is a constant $c_2 > 0$ such that for each $\varepsilon > 0$ the representation (2.13) holds for f with the error at most $c_2\varepsilon$. Letting $\varepsilon > 0$ go to 0 finishes the proof. □

C. Wavelets For set $\mathcal{G}$ obtained from functions of RBF type (2.5), the theory of wavelets is of use. The basic result there is the following. Let σ be an L_2 function with $\|\sigma\|_2 = 1$, such that $\int \frac{|\hat{\sigma}(a)|^2}{|a|} \,\mathrm{d}a$ is finite (such σ is called a wavelet). Under suitable conditions (which we will not describe here in detail) one has

$$f = \int w_{a,b} \sigma \left(\frac{x-b}{a} \right) \mathrm{d}(a,b) \,,$$

where $w_{a,b}$ are suitable "weights". For more details, any book about wavelets, e.g. [Bl98] can be of use.

D. Integral representation of $f(x)$ based on Fourier transform Another approach to bounds on $\|f\|_{\mathcal{G}}$ (although without this notation) is due to Barron [Ba93]. Let $B \subseteq \mathbb{R}^d$ be bounded. Let $\Omega_{B,\rho}$ be the set of all functions $f : B \to \mathbb{R}$ such that

1. For some complex-valued measure $\hat{F}(\,\mathrm{d}\omega)$ and for any $x \in B$
$$f(x) = f(0) + \int (e^{i\omega\cdot x} - 1)\hat{F}(\,\mathrm{d}\omega)\,.$$
2. We have $\int |\omega|_B F(\,\mathrm{d}\omega) \leq \rho$. Here $F(\,\mathrm{d}\omega)$ denotes the magnitude distribution of $\hat{F}(\,\mathrm{d}\omega)$ from part 1, and $|\omega|_B = \sup_{x\in B} |x \cdot \omega|$.

Examples of such functions include functions f for which the Fourier transform $\hat{f}$ exists, the inverse Fourier transform produces f, and $\omega\hat{f}(\omega)$ is integrable. Many more examples (positive definite functions, functions in C^s where $s = \lfloor d/2 \rfloor + 2$, etc.) are listed in [Ba93].

Theorem 2.3.3 (Integral representation based on Fourier transform [Ba93]) *Let σ be any sigmoidal function, let $f \in \Omega_{B,\rho}$. Then $\|f(x) - f(0)\|_{\mathcal{G}_\sigma} \leq \rho$. Consequently, $f(x) - f(0)$ can be approximated well by elements in $\mathcal{G}_\sigma$:*
$$\|(f(x) - f(0)) - \mathrm{conv}_n\, \rho\mathcal{G}_\sigma\|_2 \leq \frac{\rho}{\sqrt{n}}\,.$$

To compare with Theorem 2.2.2, this method is more widely applicable; it does not yield an explicit formula for $f(x)$, though.

2.3.2 Networks with Continuum Many Units

The term *neural network with continuum many units* was metaphorically used in [KKK97] to describe a function in integral representation
$$f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a \tag{2.14}$$
where it is to be understood that for every a we have a function $\varphi(\cdot, a)$ as the activation function of one neuron; we take this neuron with weight $w(a)$. This concept enabled an interesting application of results of Section 2.1.1. It is not clear, however, what is the relation between the class of functions representable as (2.14), functions realizable by finite neural networks, that is expressible as
$$f(x) = \sum_{i=1}^{n} c_i\varphi(x, a_i) \tag{2.15}$$
and functions that can be approximated by finite networks.

In this section we will try to clarify these relationships. To this end, we extend the notion of neural network with continuum many neurons even further. For any signed measure ν on A, we consider the function
$$f(x) = \int_A \varphi(x,a)\,\mathrm{d}\nu(a)\,. \tag{2.16}$$

Any function representable by (2.14) can be represented as (2.16), when ν has density $w(a)$.

We recall (and introduce) some notation. We have a continuous function φ on $H \times A$. As in previous sections, we put $\mathcal{G} = \{\pm\varphi(\cdot, a), a \in A\}$. In this notation, finite neural networks compute functions in $\operatorname{span} \mathcal{G}$. As we want to restrict the size of the weights towards the output neuron, it makes more sense to consider functions in $\operatorname{conv} c\mathcal{G}$ for some real c. Functions that can be approximated by such bounded finite networks are those in $\operatorname{cl} \operatorname{conv} c\mathcal{G}$. Finally, we let $I(c, \mathcal{G})$ denote the set of functions f that can be represented as (2.16) for some measure ν on A with $\|\nu\| \leq c$.

It is obvious that $\operatorname{conv} \mathcal{G} \subseteq \operatorname{cl} \operatorname{conv} \mathcal{G}$ and $\operatorname{conv} c\mathcal{G} \subseteq I(c, \mathcal{G})$. Less obvious is the relationship between $\operatorname{cl} \operatorname{conv} \mathcal{G}$ and $I(c, \mathcal{G})$:

Theorem 2.3.4 (Sum ⇒ Integral) *Let $\varphi \in C(H \times A)$, H and A compact subsets of $\mathbb{R}^d$ and $\mathbb{R}^k$, respectively. Then for every real c*

$$\operatorname{cl} \operatorname{conv} c\mathcal{G} \subseteq I(c, \mathcal{G})\,.$$

Explicitly, every function that can be approximated by functions of form $\sum_{i=1}^{m} c_i \varphi(x, a_i)$ for $\sum_{i=1}^{m} |c_i| \leq c$ can be expressed as $f(x) = \int \varphi(x, a) \,\mathrm{d}\nu(a)$ for some signed measure ν of norm at most c.

Proof: Let f be a function in $\operatorname{cl} \operatorname{conv} c\mathcal{G}$, and choose a sequence f_n converging to f. We can write $f_n = \sum_{i=1}^{m_n} c_{n,i} \varphi(a_{n,i}, x)$, where $\sum_i |c_{n,i}| \leq c$. We let ν_n be the weighted counting measure, that is for any set $E \subseteq A$ we put

$$\nu_n(E) = \sum_{i: a_{n,i} \in E} c_{n,i}\,.$$

Recall that the space $M(A)$ of signed measures on A is the dual to $C(A)$. As $C(A)$ is separable, Helly's theorem implies that the ball of radius c in $M(A)$ is w^*-sequentially-compact. This in particular implies that there is a measure ν and a subsequence ν_{n_k} converging to ν in the w^* topology (as ν is w^*-limit of measures with norms at most c its norm is at most c as well). This in turn means that for every function $g \in C(A)$ we have $\int g \,\mathrm{d}\nu_{n_k} \to \int g \,\mathrm{d}\nu$. We apply this for $g = \varphi(\cdot, a)$ for every $x \in H$. We obtain

$$f_{n_k}(x) = \int_A \varphi(x, a) \,\mathrm{d}\nu_{n_k}(a) \to \int_A \varphi(x, a) \,\mathrm{d}\nu(a)\,.$$

As $\lim_n f_n(x) = f(x)$ by our choice of f_n, this finishes the proof. □

In Theorem 2.2.6 we showed the converse to the above theorem: if a function $f(x)$ is in form (2.16) then it is a limit of functions of form (2.15). This tells us that certain functions can be approximated well. Moreover, combination of Theorem 2.2.6 and 2.3.4 concludes our intention to compare these two ways to extend the notion of neural networks to infinity.

2.4 Applications

In this section we combine results regarding rates of approximation (Section 2.1), $\mathcal{G}$-variation (Section 2.2) and integral representation (Section 2.3) to derive practically applicable results.

We start with a result that appears already in [KKK97].

Corollary 2.4.1 (Approximation for $C^d(\mathbb{R}^d)$ functions [KKK97]) *Let d be an odd positive integer and $f \in C^d(\mathbb{R}^d)$ a compactly supported function. Let σ be a continuous sigmoidal function. Then there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_2 \leq \frac{C}{\sqrt{n}}\,.$$

Corollary 2.4.2 (Approximation for $C^d(\mathbb{R}^d)$ functions in $\mathcal{L}_p$, gen. sigm. function) *Let $1 < p < \infty$, let d be an odd positive integer and $f \in C^d(\mathbb{R}^d)$ a compactly supported function. Let σ be a nondecreasing sigmoidal function (not necessarily continuous). Then there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p \leq \frac{C}{n^{1-1/t}}\,,$$

where $t = \min\{p, 2\}$.

Proof: By Theorem 2.3.1 we have integral representation of f using Heaviside functions:

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f(y)\,\mathrm{d}y \Big) \vartheta(e \cdot x + b)\,\mathrm{d}b\,\mathrm{d}e\,.$$

Thus by Theorem 2.2.2 we obtain bounded $\mathcal{G}_\vartheta$-variation, bound given by integral of directional derivatives:

$$\mathcal{G}_\vartheta \leq \int_{S^{d-1}} \int_{\mathbb{R}} a_d \Big| \int_{H_{eb}} D_e^{(d)} f(y)\,\mathrm{d}y \Big| \,\mathrm{d}b\,\mathrm{d}e).$$

Using Theorem 2.2.9 we observe that $\mathcal{G}_\vartheta(f) = \mathcal{G}_\sigma(f)$ for any sigmoidal activation function σ. Now it remains to use Theorem 2.1.6, we observe that $\mathcal{G}_\sigma$ is $\mathcal{L}^p$-bounded on support of f and having shown that $\mathcal{G}_\sigma(f)$ is finite we conclude the proof. □

Corollary 2.4.3 (Approximation for $\mathcal{W}^{d,p}(\mathbb{R}^d)$ functions in $\mathcal{L}_p$, gen. sigm. function) *Let $1 < p < \infty$, let d be an odd positive integer, let $\Omega \subseteq \mathbb{R}^d$ be a bounded open set with a C^1 boundary and consider an $f \in \mathcal{W}^{d,p}(\Omega)$. Let σ be a nondecreasing sigmoidal function (not necessarily continuous). Then there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p \leq \frac{C}{n^{1-1/t}}\,,$$

where $t = \min\{p, 2\}$.

Proof: By Theorem 2.3.2 we have integral representation of f using Heaviside functions:

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f(y)\, \mathrm{d}y \Big) \vartheta(e \cdot x + b)\, \mathrm{d}b\, \mathrm{d}e$$

(the derivatives are taken in the weak sense). Thus by Theorem 2.2.8 we find that the $\mathcal{G}_\vartheta$-variation is bounded. The bound is given by integral of directional derivatives:

$$\mathcal{G}_\vartheta \leq \int_{S^{d-1}} \int_{\mathbb{R}} a_d \Big| \int_{H_{eb}} D_e^{(d)} f(y)\, \mathrm{d}y \Big| \,\mathrm{d}b\, \mathrm{d}e\,.$$

In [KKV06] this computation is carried on to provide an upper bound on $\mathcal{G}_\vartheta$ variance in terms of Sobolev norm (even Sobolev seminorm) $\|\cdot\|_{d,1}$. As Ω is of finite measure, this implies a bound $\|f\|_{\mathcal{G}_\vartheta} = O(\|\cdot\|_{d,p})$. Using Theorem 2.2.9 we observe that $\mathcal{G}_\vartheta(f) = \mathcal{G}_\sigma(f)$ for any sigmoidal activation function σ. Now it remains to use Theorem 2.1.6, we observe that $\mathcal{G}_\sigma$ is $\mathcal{L}^p$-bounded on support of f and having shown that $\mathcal{G}_\sigma(f)$ is finite we conclude the proof. □

By using Theorem 2.2.9 instead of results in [KKK97], we can weaken the assumption on σ – we do not need σ continuous, it is enough, if σ is nondecreasing and bounded. More disagreeable, though, are the assumptions required on f, which are perhaps too strong for applications. We mostly care about rather large d, so we need f to be very smooth. Next, we will discuss the possibilities to weaken this requirement. First, we will see that for $d = 1$ such weakening is possible. (Similar result for σ being the Heaviside function is suggested in [KKV06].)

Theorem 2.4.4 (Rates for absolutely continuous functions) *Let f be an absolutely continuous function on $[a, b]$. Let σ be any sigmoidal function (not necessarily continuous). Then there is a constant C so that*

$$\|f - \mathrm{span}_n\, \mathcal{G}_\sigma\|_p \leq \frac{C}{n^{1-1/p}}\,.$$

Proof: We represent $f(x)$ in form (2.12) (Section 2.3.1, part A). Theorem 2.2.8 implies that $\|f\|_{\mathcal{G}_\vartheta} \leq \|f'\|_1$ (which we know is finite). From Theorem 2.2.9 we know that $\|f\|_{\mathcal{G}_\sigma} = \|f\|_{\mathcal{G}_\vartheta}$, so it remains to use Theorem 2.1.6. □

We see that we lowered the smoothness assumption – we require f to be absolutely continuous, instead of being C^1. However, it is possible to weaken the assumptions on f even further and at the same time improve the approximation, at least in the one-dimensional case. (This result may be known in the analysis community, we have been unable to find it in the literature, though.)

Theorem 2.4.5 (Rates for bounded variation functions) *Let f be a bounded variation function on $[a,b]$. Then*

$$\|f - \operatorname{span}_n \mathcal{G}_\vartheta\|_\infty \le \frac{\|f\|_{BV[a,b]}}{n-1}\,.$$

If σ is any sigmoidal function (not necessarily continuous) then we have for any $p \in (1,\infty)$ and a constant $c = c(a,b,p)$

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p \le \frac{c\|f\|_{BV[a,b]}}{n-1}\,.$$

Proof: It is known from calculus (see, e.g., Theorem 1.2 in Section X.1 of [Lan93]) that a bounded variation function can be expressed as a difference of two nondecreasing functions, $f = f_1 - f_2$ in such a way, that $\|f\|_{BV[a,b]} = d_1 + d_2$, where $d_i = f_i(b) - f_i(a)$. Using the technique in the proof of Theorem 2.2.9 (part (B)) we approximate $f_i(x) - f_i(a)$ by a function $g_i(x)$, which is a linear combination of n_i shifts of the Heaviside function ϑ, so that $n_i \le \lfloor \frac{d_i}{\varepsilon} \rfloor$ and for all but finitely many values of x we have $0 \le \big(f_i(x) - f_i(a)\big) - g_i(x) \le \varepsilon$. Consequently,

$$|\big(g_1(x) - g_2(x) + f(a)\big) - f(x)| \le \varepsilon$$

for all but finitely many values of x. We may realize addition of $f(a)$ as one extra Heaviside function, so we found an approximation using $n_1 + n_2 + 1 \le \frac{d_1+d_2}{\varepsilon} + 1$ Heaviside functions and achieved an $\mathcal{L}^\infty$ error ε.

The second assertion follows immediately by approximating $\vartheta(t)$ by $\sigma(Nt)$ for N large enough. □

We remark that a weaker version of the above theorem (with the usual rate of convergence $O(1/n^{1-1/p})$ in $\mathcal{L}^p$-norm) could be proved also using Theorem 2.2.8: if f is a bounded variation function on $[a,b]$ and μ_f the corresponding Riemann-Stieltjes measure, then we have the following formula (Proposition 1.8 in Section X.1 of [Lan93])

$$f(x) - f(a) = \int_a^x 1 \,\mathrm{d}\mu_f\,,$$

whenever f is continuous at both a and x. As bounded variation function is continuous at all but finitely many points, we can indeed apply Theorem 2.2.8.

Next, we will consider the case of larger d.

In Theorem 2.3.2 we decreased the differentiability requirement – instead of existence of continuous d-fold derivatives as in 2.4.1 and 2.4.2 we only require that d weak derivatives exist (and are bounded in the $\mathcal{L}_p$ norm). This may not seem as a tremendous improvement. On the other hand, in this setting we have the following result that presents a limit on how much can we weaken the assumptions on the function to be approximated.

Theorem 2.4.6 (Good rates $\Longrightarrow$ many weak derivatives) *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function. Suppose that for each function $f \in \mathcal{W}^{m,2}(B^d)$ (where B^d is the unit ball in $\mathbb{R}^d$) there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_2 \leq \frac{C}{\sqrt{n}} \, .$$

Then $m \geq (d-1)/2$.

Proof: By Theorem 1.4.1 there is a function $f \in \mathcal{W}^{m,2}(B^d)$ such that

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_2 \geq C' n^{-m/(d-1)} \, .$$

So we have $\frac{C}{\sqrt{n}} \geq C' n^{-m/(d-1)}$. Considering the limit as $n \to \infty$ finishes the proof. □

We finish this discussion by mentioning the connection with Theorem 1.4.2 from Chapter 1. Indeed, this theorem actually gives better bounds on $\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p$ than the results of this chapter. The drawback, however, is that we need to use linear combinations with unbounded coefficients. (The inspection of the proof, as presented in [Pi99] shows that, indeed, unbounded coefficients are crucial for the proof.) This renders the result useless for practical applications: we can find good approximation of f in form

$$\sum_{i=1}^{n} c_i \sigma(a_i \cdot x + b_i) \tag{2.17}$$

for small n, but at the expense of using large coefficients c_i. Consequently, we need to do the computations with a high precision – which only shows that n is not an appropriate measure of complexity of the expression (2.17). This problem is partially avoided by using convex combinations (or, rather, combinations with bounded sum of the coefficients). However, a detailed study of the numerical issues involved remains to be done.

Similar results as in Theorem 2.4.4 can be easily derived for wavelets and Baron's representation – paragraphs C and D in Section 2.3.1.

2.5 Conclusion

In this chapter we studied properties of approximations of functions using convex combinations. Results of Maurey, Jones and Barron and of Darken et al. show that, when applicable, such convex combinations yield good rates of approximation (independent of input dimension).

Further study of constants that appear in these rates bring the notion of $\mathcal{G}$-variation (as defined in [Ku97]). To maintain the mentioned rates when approximating a function f by functions from $\mathcal{G}$ we have seen that fininte $\mathcal{G}$-variation of f is needed. Pursuing this idea Kůrková [Ku97] shows that for continuous approximating functions in $\mathcal{G}$ for f representable as integral of these functions weighted by a continuous function $\mathcal{G}$-variation is finite. She proves this result also for Heaviside activation functions. We extend these results to $\mathcal{L}^p$ almost everywhere bounded activation functions and weights by a constructive proof (Theorem 2.2.3) and nonconstructively using Hahn-Banach Theorem to continuous or $\mathcal{L}^p$ activation functions and weights represented by any signed measure (Theorems 2.2.6 and 2.2.8). We further investigate the notion of $\mathcal{G}$-variation and show that for f with infinite $\mathcal{G}$-variation we can have arbitrarily slow convergence of approximation (Theorem 2.1.7). A surprising result comes from Theorem 2.2.9 - we show that the presented rates of approximation cannot distinguish between sigmoidal functions.

As mentioned above, all the presented results require f to representable as an integral. In Section 2.3 we overview known results towards this direction and also show that integral representation is a necessary condition for f to be approximable with good rates of approximation by convex sums of continuous activation functions (Theorem 2.3.4).

In Section 2.4 we combine the above mentioned and present a few instances of theoretical bounds on rates of approximation for specific functions f showing how to easily derive corollaries of the type using results of previous sections. One more interesting and less obvious result of this section is the less optimistic information presented in Theorem 2.4.6 – if we wish to have good rates of approximation for function f we have to demand it to have many weak derivatives.

Chapter 3

Algorithmic Aspects

As we already mentioned in Chapter 2, along the Barron-type investigation there are two established ways to estimate rates of approximation in Hilbert (and more generally Banach) spaces. The first one (pioneered by Jones [Jo92] and Barron [Ba93], and generalized to Banach spaces by Darken et al. [DDGS93]) analyzes the process of iteratively improving the approximating function. The intrinsic feature of this approach is, that to get the n-th approximating function, f_n, we first compute f_0, f_1, ..., f_{n-1}. This makes each of the iterative steps easier to manage, but it also may constitute unnecessary work.

The other approach (established by Maurey [Ps81], and also pursued by Barron [Ba93], Darken et al. [DDGS93], and Makovoz [Mk96]) is nonconstructive: it is possible to show that there exists a good approximant using the so-called probabilistic method; that is in appropriately chosen probability space, the probability of getting a good approximant is nonzero.

In this chapter we extend this second approach – we suggest a randomized algorithm and compare it to the iterative approach. To this end, we also analyze complexity of the iterative algorithm, a task, which seems not to be addressed in existing literature.

By its nature the randomized algorithm is very well suited to prune networks with too many units to obtain smaller ones with still good approximation abilities (probability estimates on error with respect to error of original big network can be expressed). We propose how to use the randomized algorithm to prune any kernel-method attained network (see Chapter 4) to obtain reasonable number of hidden units. Experiments towards this end are running in cooperation with Petra Kudová-Vidnerová on kernel-based neural network schemas (for more details see Chapter 4).

We start by analyzing the iterative algorithm in Section 3.1. In Section 3.2 we propose and study the randomized algorithm. In Section 3.3 we briefly discuss the complexity of various "subprocedures" used in the algorithms under study. Finally, in Section 3.4 we compare the suggested algorithms in various situations.

Results presented in this chapter came from cooperation with Robert Šámal and have been published in [SS08].

3.1 Iterative Algorithms

In this section we analyze time complexity of the algorithm suggested by Jones [Jo92] and Barron [Ba93], and further analyzed by Darken et al. [DDGS93]. In these papers, the authors concentrate on the estimate of approximating abilities of some approximation schemas (in particular neural networks), without addressing the amount of computations needed to achieve such approximation. In this section we try to clarify this point. We start by specifying the algorithm under study – or, more precisely, the algorithms, as there are (at least) two distinct ways to organize the computations. Then we do a "high-level" estimate of complexity, postponing the discussion of how some standard tasks can be implemented to Section 3.3.

Let us recall basic assumptions from Chapter 2. We consider approximations of a function f using convex combinations of functions from $\mathcal{G}$, a subset of some function space X (which is a Banach space). We are given n, the upper bound on the number of elements of $\mathcal{G}$ we are allowed to use. The key idea is to use the following iterative transformation.

$$\begin{aligned} f_0 &= 0 \\ f_{k+1} &= \alpha f_k + (1-\alpha) g_k, \qquad g_k \in \mathcal{G}, \alpha \in [0,1] \end{aligned} \tag{3.1}$$

We will use a result that appears as Theorem 3.5 in [DDGS93]:

Theorem 3.1.1 (Iterative rates in smooth Banach spaces [DDGS93]) *Let X be a uniformly smooth Banach space having modulus of smoothness $\rho(u) \leq \gamma u^t$ with $t > 1$. Let $\mathcal{G}$ be a bounded subset of X and let $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ be given. Select $\rho > 0$ such that $\|f - g\| \leq \rho$ for all $g \in \mathcal{G}$ and fix $\eta > 0$. We choose sequences $\{f_k\} \subset \operatorname{conv}\mathcal{G}$ and $\{g_k\} \subset \mathcal{G}$ recursively so that*

1. $f_1 \in \mathcal{G}$
2. $F_k(g_k - f) \leq \frac{2\gamma}{k^{t-1}\|f_k - f\|^{t-1}}\left((\rho+\eta)^t - \rho^t\right) =: \delta_k$
3. $f_{k+1} = \frac{k}{k+1} f_k + \frac{1}{k+1} g_k$,

where F_k is peak functional for $f_k - f$ (we terminate the procedure if $f_k = f$). Then we have

$$\|f_k - f\| \leq \frac{(2\gamma t)^{1/t}(\rho+\eta)}{k^{1-1/t}}\big(1 + \frac{(t-1)\log_2 k}{2tk}\big)^{1/t}. \tag{3.2}$$

Note that the error bound presented here holds in every step and thus also for an n given in advance by the "enemy". There are two natural ways to find an appropriate function g_k in (3.1).

1. Using repeated optimization. The first approach is to use (3.1) as a way to reduce dimensionality of an optimization problem (it was apparently meant so by Jones, who originally proposed this method). That is, instead of choosing all functions to form the approximation at the same time, we only decide about one function at a time. This amounts to n times solving a (multidimensional) optimization problem, to find good enough g_k: in view of Theorem 3.1.1 this means to find g_k for which $F_k(g_k - f) \leq \delta_k$. Here F_k is the peak functional for $f_k - f$, that is a linear continuous functional of unit norm for which $F_k(f_k - f) = \|f_k - f\|$. We denote the time needed for finding such g_k by $t_{opt}(\mathcal{G}, \delta_k)$. We will discuss t_{opt} further in Section 3.3.2. In (3.3) a lower bound on δ_k is presented; it yields $\delta_k = \Omega(1/k^{1-1/t})$. Based on these considerations we have the following algorithm and time analysis.

Algorithm 3.1.2 (Iterative from scratch)

Given: f, $\mathcal{G}$, n

```
1. $f_0 = 0$
2. For $k = 0$ to $n - 1$:
    • Let $F_k$ be a peak functional for $f_k - f$.
    • Find $g_k \in \mathcal{G}$ so that $F_k(g_k - f) < \delta_k$, where $\delta_k$ is as in Theorem 3.1.1.
    • Put $f_{k+1} = \frac{k}{k+1} f_k + \frac{1}{k} g_k$
```

Output: f_n
Time: $\sum_{k=1}^{n-1}(t_{opt}(\mathcal{G}, \delta_k) + O(1)) = O(n \cdot t_{opt}(\mathcal{G}, \Omega(\frac{1}{n^{1-1/t}})))$

2. Using an auxiliary approximation. The second approach (inspired by the proofs in [Jo92, Ba93, DDGS93]) is to first find an auxiliary approximation of f, possibly using a large number of terms: For $i = 1, \dots, N$ we find $h_i \in \mathcal{G}$ and c_i ($c_i \geq 0$, $\sum_{i=1}^{N} c_i = 1$) and put

$$f_{apx} = \sum_{i=1}^{N} c_i h_i \,.$$

We let ε_{apx} denote the error of this approximation, that is $\varepsilon_{apx} = \|f - f_{apx}\|$. We will discuss the possibilities to obtain such approximation later, in Section 3.3.3. Meanwhile, we just let $t_{apx}(\mathcal{G}, \varepsilon_{apx}, N)$ denote the time needed to find such f_{apx}. Next, in each iteration of (3.1) we choose for g_k one of the functions h_i, $i = 1, \dots, N$. We will do so to make $F_k(g_k - f) < \delta_k$. Let us pause for a while to analyze when is this possible. Due to linearity of F_k we have

$$F_k(g_k - f) = F_k(g_k - f_{apx}) + F_k(f_{apx} - f)\,.$$

The second term is at most $\|F_k\| \cdot \|f_{apx} - f\| = \varepsilon_{apx}$. The first term is "zero in the average":

$$\sum_{i=1}^{N} c_i F_k(h_i - f_{apx}) = F_k(\sum_{i=1}^{N} c_i h_i) - \sum_{i=1}^{N} c_i F_k(f_{apx}) = 0\,.$$

Consequently, there is at least one h_i for which $F_k(h_i - f_{apx}) \leq 0$, therefore $F_k(h_i - f) \leq \varepsilon_{apx}$. We will let g_k be this h_i (or one of them, if there is more than one such h_i).

Theorem 3.1.1 states that we need $F_k(g_k - f) \leq \delta_k$, thus it is sufficient to ensure $\varepsilon_{apx} \leq \delta_k$. As we do not know in advance the size of $\|f - f_k\|$, we need a lower bound on δ_k, that can be evaluated before we start the algorithm. Using the estimate $(\rho + \eta)^t - \rho^t \geq \eta t \rho^{t-1}$ for $t > 1$ and the inductive bound on $\|f - f_k\|$, we get that

$$\begin{aligned} \delta_k &\geq 2\gamma t \eta \left(\frac{\rho}{k \frac{(2\gamma t)^{1/t}(\rho+\eta)\left(1+\frac{(t-1)\log_2 k}{2tk}\right)^{1/t}}{k^{1-1/t}}} \right)^{t-1} \\ &= (2\gamma t)^{1/t} \frac{\eta}{k^{1-1/t}} \big(\frac{\rho}{\rho+\eta}\big)^{t-1} \big(1 + \frac{(t-1)\log_2 k}{2tk}\big)^{1/t-1} =: d_k\,. \end{aligned} \tag{3.3}$$

Thus, we will choose f_{apx} such as to make $\varepsilon_{apx} \leq d_{n-1}$. Hence the assumptions of Theorem 3.1.1 will be satisfied, and so we get approximation f_n as guaranteed by this theorem.

Algorithm 3.1.3 (Iterative from auxiliary approximation)

Given: f, $\mathcal{G}$, n

1. *Choose $\eta > 0$ (smaller η yields better approximation, as given by Theorem 3.1.1). Let the modulus of smoothness of the given space be $\varrho(u) \leq \gamma u^t$.*

2. *Put $\varepsilon_{apx} = (2\gamma t)^{1/t} \frac{\eta}{(n-1)^{1-1/t}} \big(\frac{\rho}{\rho+\eta}\big)^{t-1} \big(1 + \frac{(t-1)\log_2(n-1)}{2t(n-1)}\big)^{1/t-1}$*

3. *For $i = 1, \ldots, N$ find functions $h_i \in \mathcal{G}$, and coefficients $c_i \geq 0$ with $\sum_{i=1}^{N} c_i = 1$ so that*
$$\big|f - \sum_{i=1}^{N} c_i h_i\big| < \varepsilon_{apx}\,.$$

4. *Put $f_0 = 0$.*

5. *For k from 0 to $n - 1$:*
   - *Find $i \in \{1, \ldots, N\}$, so that $F_k(h_i - f) \leq \delta_k$, where δ_k is as in Theorem 3.1.1.*
   - *Put $g_k = h_i$.*

- *Put* $f_{k+1} = \frac{k}{k+1} f_k + \frac{1}{k+1} g_k$.

Output: *f_n satisfying the estimate* (3.2)
Time: $t_{apx}(\mathcal{G}, \varepsilon_{apx}, N) + nNt_{int}$

Two remarks are in place.

How to compute peak functionals In a general Banach space, the existence of a peak functional for every nonzero element is only guaranteed nonconstructively, by means of Hahn-Banach theorem. In many concrete spaces, however, peak functionals are given by a simple integral formula (which is why we denote time to evaluate peak functionals by t_{int}, the time needed to enumerate an integral). For instance, in a Hilbert space, peak functional corresponding to f is given by scalar product with $f/\|f\|$. More generally, if f is an element of the $\mathcal{L}^p$ space, then we may put $\psi(t) = sgn(f(t))\frac{|f(t)|^{p-1}}{\|f\|^{p-1}}$. It is easy to verify that F given by $F(g) = \int g(t)\psi(t)\,\mathrm{d}t$ is a peak functional corresponding to f.

An improvement for Hilbert spaces In Hilbert spaces we may utilize the simple form of the peak functionals to improve time complexity of the above algorithm. A simple way to calculate the required peak functionals (that is, scalar products) is to compute N scalar products at each iteration. This requires time nNt_{int}; we can do better, though.

Before we start our iteration process, we can compute the N scalar products $\langle h_i, f\rangle$ and also $\langle f, f\rangle$. We also store all the scalar products $(\langle h_i, f_k\rangle$ for $i = 1, \dots, N$ and for the "current value" of k. After each step, when f_{k+1} is found using (3.1), we recalculate these scalar products. We assume that all scalar products $\langle h_i, h_j\rangle$ are easy to compute, or precomputed, so the computation of $\langle g_i, f_k\rangle$ only takes $O(N)$ operations with numbers, but no scalar products. In this way, all the scalar products are computed in time $(N+1)t_{int} + O(nN)$.

3.2 Randomized Algorithms

In this section we are going to follow closely the proof of Theorem 3.2.1 (more general version of 2.1.3) as it appears in Darken et al. [DDGS93]. At the end of it, rather than using the basic observation "there is always something at least as good as the average", we use the Markov inequality (Theorem C.2.8). This extra step has useful algorithmic consequences, which we state as Corollary 3.2.3.

Recall that $\mathcal{L}^p$ spaces for $p \geq 1$ are of Rademacher type $\min\{p, 2\}$.

Theorem 3.2.1 (Probabilistic estimate of error in Banach spaces (based on [DDGS93])) *Let X be Banach space of Rademacher type t, $1 \leq t \leq 2$. Suppose $\mathcal{G} \subset X$, $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ and let $\rho > 0$ be such that for every $g \in \mathcal{G}$ we have $\|g - f\| \leq \rho$.*

Assume we have an approximation of f of the form

$$f_{apx} = \sum_{i=1}^{N} c_i h_i, \qquad \sum_{i=1}^{N} c_i = 1, \quad c_i \geq 0, \quad h_i \in \mathcal{G}$$

so that $\|f - f_{apx}\| < \varepsilon$. We pick $g_1, \ldots, g_n$ from $h_1, \ldots, h_N$ independently at random, so that $\Pr[g_j = h_i] = c_i$ for each i, j.

Then the average

$$g_{apx} = \tfrac{1}{n}\sum_{i=1}^{n} g_i$$

is (with high probability) a reasonable approximation, namely

$$\Pr[\|g_{apx} - f\| > Q^{1/t} \cdot (\frac{C(\rho+\varepsilon)}{n^{1-1/t}} + \varepsilon)] < \frac{1}{Q}.$$

Here C is a constant depending on X but independent of n, and $Q > 1$ is a parameter specifying the trade-off between quality of approximation and time needed to find it.

Proof: We follow the proof of [DDGS93] altering it slightly at the end of the proof. Put $\Delta = f - f_{apx}$, so that $\|\Delta\| < \varepsilon$.

Take g_j to be a sequence of independent random variables on X every one of them taking value h_i with probability c_i. Then for any $\beta \in (0,1)$

$$\begin{aligned}
E\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\|^t &= \frac{1}{n^t} E\left\|\sum_{j=1}^{n}(f - g_j - \Delta) + n\Delta\right\|^t \\
&\leq \frac{1}{n^t} E\left((1-\beta)\frac{\|\sum_{j=1}^{n}(f - g_j - \Delta)\|}{1-\beta} + \beta n\frac{\|\Delta\|}{\beta}\right)^t \\
&\leq \frac{1}{n^t}\left[(1-\beta)E\left(\frac{\|\sum_{j=1}^{n}(f - g_j - \Delta)\|}{1-\beta}\right)^t + \beta\left(\frac{n\|\Delta\|}{\beta}\right)^t\right] \\
&= \frac{1}{n^t(1-\beta)^{t-1}} E\left\|\sum_{j=1}^{n}(f - g_j - \Delta)\right\|^t + \frac{1}{\beta^{t-1}}\|\Delta\|^t \qquad (3.4)
\end{aligned}$$

This is true since $x \mapsto x^t$ is a convex function on $\mathbb{R}^+$ for $1 \leq t \leq 2$. Since the range of g_j has finitely many values, g_j is a Radon random variable. As the Rademacher type of X is t, by Proposition 9.11 in [LeTa91] we have that there is a constant $C = C(X)$ so that

$$E\left\|\sum_{j=1}^{n}(f - g_j - \Delta)\right\|^t \leq C^t \sum_{j=1}^{n} E\|f - g_j - \Delta\|^t. \tag{3.5}$$

It is also true that for every j

$$\begin{aligned} E\|f - g_j - \Delta\|^t &= \sum_{i=1}^{N} c_i\|f - h_i - \Delta\|^t \\ &\leq \sum_{i=1}^{N} c_i(\|f - h_i\| + \|\Delta\|)^t \\ &< \sum_{i=1}^{N} c_i(\rho + \varepsilon)^t \\ &= (\rho + \varepsilon)^t. \end{aligned} \tag{3.6}$$

Combining (3.4), (3.5) and (3.6) we get

$$E\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\|^t < \frac{C^t(\rho+\varepsilon)^t}{n^{t-1}(1-\beta)^{t-1}} + \frac{\varepsilon^t}{\beta^{t-1}}.$$

At this point we divert from the proof in [DDGS93]. The above equation is valid for every $\beta \in (0, 1)$. A simple calculus shows that the right hand side is minimized when $\beta = (1 + \frac{C(\rho+\varepsilon)}{\varepsilon n^{1-1/t}})^{-1}$. This value of β yields

$$E\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\|^t < \left(\frac{C(\rho+\varepsilon)}{n^{1-1/t}} + \varepsilon\right)^t.$$

Thus, by Markov inequality C.2.8 we have

$$\begin{aligned} &\Pr\left[\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\|^t \geq Q \cdot \left(\frac{C(\rho+\varepsilon)}{n^{1-1/t}} + \varepsilon\right)^t\right] \\ &= \Pr\left[\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\| \geq Q^{1/t} \cdot \left(\frac{C(\rho+\varepsilon)}{n^{1-1/t}} + \varepsilon\right)\right] \\ &\leq \frac{1}{Q}. \end{aligned}$$

□

Theorem 3.2.1 leads to the following algorithm.

Algorithm 3.2.2 (Probabilistic from auxiliary approximation)

Given: *f, $\mathcal{G}$, n*

1. *Pick $\eta > 0$ and put $\varepsilon_{apx} = \frac{\eta}{n^{1-1/t}}$. The value of η affects the quality of approximation; as we will see below, for $\mathcal{L}_p$ spaces $(1 \leq p \leq 2)$ it is reasonable to make η comparable with ρ, or $\sup\{\|g - f\| : g \in \mathcal{G}$.*

2. *Find for $i = 1, \ldots, N$ functions $h_i \in \mathcal{G}$, and coefficients $c_i \geq 0$ with $\sum_{i=1}^{N} c_i = 1$ so that*
$$|f - \sum_{i=1}^{N} c_i h_i| < \varepsilon_{apx}$$

3. *For j from 1 to n choose randomly $t_j \in \{1, \ldots, N\}$ with $\Pr[t_j = i] = c_i$.*

4. *Put $g_{apx} = \frac{1}{n} \sum_{i=1}^{n} h_{t_j}$*

5. *If $\|f - g_{apx}\| \geq Q^{1/t}\big(\frac{C(\rho+\varepsilon_{apx})}{n^{1-1/t}} + \varepsilon_{apx}\big)$ go back to step 2.*

Output: g_{apx}
Expected time: $t_{apx}(\mathcal{G}, \varepsilon_{apx}, N) + (O(n \log N) + nt_{int})\frac{Q}{Q-1}$.

Corollary 3.2.3 (Error of probabilistic algorithm) *Algorithm 3.2.2 gives an approximation of f with an error at most*

$$Q^{1/t}\frac{C\rho + \eta(1 - C/n^{1-1/t})}{n^{1-1/t}}$$

after an expected number of $\frac{Q}{Q-1}$ repetitions.

Proof: By Theorem 3.2.1 the probability of failure in Step 5 is at most $1/Q$, so we can bound the number of repetitions by a geometric random variable with probability of success $1 - 1/Q$. The expectation of this random variable is $\frac{Q}{Q-1}$. □

Remark We have kept C in the estimates to preserve generality. However, in the most interesting case (from the practical point of view), which is $\mathcal{L}_p$ with $p \in [1, 2]$, we have $C = 1$ (see [DDGS93], the discussion above Corollary 2.6).

Now we set out to estimate the running time of Algorithm 3.2.2. As in the previous section, we will postpone the discussion of some "sub-procedures" to Section 3.3.

The second step has time complexity $t_{apx}(\mathcal{G}, \varepsilon_{apx}, N)$. Then we need to pick a random number n times with the correct distribution. The standard way to do this is to define $b_i =$

$\sum_{j \leq i} c_j$, so that $0 = b_0 < b_1 < \cdots < b_N = 1$. Then we pick a uniformly random number $u \in [0, 1]$ and find i such that $u \in (b_{i-1}, b_i)$; we let the next t_j be equal to i. To be able to tell which interval u belongs to, we need to sample it with sufficient precision. When we do this "adaptively", that is we only generate as many bits as needed to tell which interval will u end up in, we need on average

$$\sum_{i=1}^{N} c_i(1 + \log_2 \frac{1}{c_i}) \leq 1 + \log_2 \sum_{i=1}^{N} c_i \frac{1}{c_i} = 1 + \log_2 N$$

random bits (we have used concavity of log and the Jensen inequality (C.2.6) for numbers $1/c_i$). Altogether Step 3 takes time $O(n \log N)$.

For Step 4 and 5 we need time $O(n \cdot t_{int})$, as we need to compute the sum in Step 4 for every point, where we will be evaluating f_{apx} when checking the norm $\|f - f_{apx}\|$ in Step 5.

Altogether, we obtain the expected time complexity

$$t_{apx}(\mathcal{G}, \varepsilon_{apx}, N) + (O(n \log N) + nt_{int})\frac{Q}{Q-1} .$$

In Section 3.4 we will discuss how this compares to the estimates obtained for the iterative algorithm.

Next we proceed with a variant of the above results for functions for which an integral representation is known. The following theorem is based on ideas from [DDGS93].

Theorem 3.2.4 (Probabilistic error in Banach spaces, int. repr.) *Let $H \subseteq \mathbb{R}^d$ be open and let $X = \mathcal{L}_p(H)$. Put $t = \min\{p, 2\}$. Let A be a subset of $\mathbb{R}^k$. Suppose $\varphi_a = \varphi(\cdot, a)$ is in X for each $a \in A$. Put $\mathcal{G} = \{\pm\varphi_a : a \in A\} \subset X$ and assume that we have an integral representation of f of form*

$$f(x) = \int_A w(a)\varphi(x, a)\,\mathrm{d}a \qquad \int_A |w(a)| = 1 .$$

Let $\rho > 0$ be such that for every $g \in \mathcal{G}$ we have $\|g - f\| \leq \rho$.

We pick $g_1, \ldots, g_n$ from $\mathcal{G}$ independently at random, so that $|w(a)|$ is the density of the probability of choosing $\pm\varphi(\cdot, a)$. We take $+\varphi(\cdot, a)$ for $w(a) > 0$ and $-\varphi(\cdot, a)$ for $w(a) < 0$.

Then the average $g_{apx} = \frac{1}{n}\sum_{i=1}^{n} g_i$ is (with high probability) a reasonable approximation, namely

$$\Pr\left[\|g_{apx} - f\| > Q^{1/t} \cdot \frac{C\rho}{n^{1-1/t}}\right] < \frac{1}{Q} .$$

Here $C > 0$ is a constant depending on X but independent of n, and $Q > 1$ is a parameter specifying the trade-off between quality of approximation and time needed to find it.

Proof: Take g_j to be a sequence of independent random variables on X every one of them taking value h_i with probability c_i.

$$E\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\|^t = \frac{1}{n^t} E\left\|\sum_{j=1}^{n}(f - g_j)\right\|^t \tag{3.7}$$

The range of g_j is a subset of $\mathcal{L}^p(H)$, which is a separable Banach space. ($C_0(H)$ is dense in $\mathcal{L}^p(H)$ (for $1 \le p < \infty$) – Corollary 19.20 in [Js03]. Then we use the Weierstrass theorem to approximate continuous functions by polynomials and we observe that it is enough to consider polynomials with rational coefficients.) Consequently, g_j is a Radon random variable [LeTa91, p.38]. As the Rademacher type of X is t, by Proposition 9.11 in [LeTa91] we have that there is a constant $C = C(X)$ so that

$$E\left\|\sum_{j=1}^{n}(f - g_j)\right\|^t \le C^t \sum_{j=1}^{n} E\|f - g_j\|^t. \tag{3.8}$$

It is also true that for every j

$$\begin{aligned} E\|f - g_j\|^t &= \int_A |w(a)| \|f \pm \varphi(x,a)\|^t \\ &\le \int_A |w(a)| \rho^t \\ &= \rho^t. \end{aligned} \tag{3.9}$$

Combining (3.7), (3.8) and (3.9) we get

$$E\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\|^t < \frac{C^t \rho^t}{n^{t-1}}.$$

Thus, by Markov inequality C.2.8 we have

$$\begin{aligned} &\Pr\left[\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\|^t \ge Q \cdot (\frac{C\rho}{n^{1-1/t}})^t\right] \\ &= \Pr\left[\left\|f - \frac{1}{n}\sum_{j=1}^{n} g_j\right\| \ge Q^{1/t} \cdot (\frac{C\rho}{n^{1-1/t}})\right] \\ &\le \frac{1}{Q}. \end{aligned}$$

□

Theorem 3.2.4 leads to the following algorithm.

Algorithm 3.2.5 (Probabilistic from integral representation)

Given: f, φ, n

1. *Find an integral representation for f of form*
$$f(x) = \int_A w(a)\varphi(x,a)\,,$$
so that $\int_A |w(a)| = 1$.
2. *For j from 1 to n choose randomly $a_j \in A$ with the density of probability $|w(a)|$.*
3. *Put $g_j = \varphi_{a_j}$ if $w(a_j) > 0$, and $g_j = -\varphi_{a_j}$ if $w(a_j) \leq 0$.*
4. *Put $g_{apx} = \frac{1}{n}\sum_{i=1}^{n} g_j$*
5. *If $\|f - g_{apx}\| \geq Q^{1/t}\big(\frac{C\rho}{n^{1-1/t}}\big)$ go back to step 2.*

Output: g_{apx}
Expected time: $O(ns_{gen}(w)d \cdot t_{int}\frac{Q}{Q-1})$

For discussion of s_{gen} see below.

Corollary 3.2.6 (Error of integral probabilistic algorithm) *Algorithm 3.2.5 gives an approximation of f with an error at most*
$$Q^{1/t}\frac{C\rho}{n^{1-1/t}}$$
after an expected number of $\frac{Q}{Q-1}$ repetitions.

The proof is the same as for Corollary 3.2.3.

In Algorithm 3.2.5 we need to generate a according to a non-uniform distribution, the density of the probability being $|w(a)|$. A general approach to do this is the acceptance-rejection method (developed by von Neumann). The basic idea is as follows. We find C so that $|w(a)| \leq C$ for each a. For simplicity assume first that $vol(A) = 1$ and $\int_A |w(a)| = 1$. We generate $a \in A$ and $u \in [0,1]$ uniformly at random, and accept a if $u \leq \frac{|w(a)|}{C}$ (that is, we accept a with probability proportional to $|w(a)|$). If we reject the generated a, we repeat the procedure. It is easy to see, that the expected number of steps needed to generate one a according to $|w(a)|$ is C: thus this basic approach is inefficient if $|w(a)|$ varies widely. In such case we try to find a simple upper bound $CW(a)$ on $|w(a)|$ (so that $\int_A W(a) = 1$) and generate a with density $W(a)$, rather than uniformly. For the details we refer the reader to Section 5.1.2 of [We00]. If we are given w such that $I = \int_A |w(a)| \neq 1$, we just find (estimate) I (for example by Monte Carlo integration, see Section 3.3.1) and then we will be generating a according to $|w(a)|/I$ and then put $g_i = \pm I\varphi(\cdot, a)$.

The conclusion is that our ability to efficiently generate with density $|w(a)|$ depends on our knowledge of the function w. We will use $s_{gen}(w)$ to denote the number of steps needed to generate one a with probability density $|w(a)|$.

The estimate of the running time of Algorithm 3.2.5 is derived similarly as that of Algorithm 3.2.2. There are two differences: instead of finding an approximating function f_{apx}, we need to get the integral representation – more precisely, we will be repeatedly evaluating $w(a)$.

When the approach of [KKK97] is used to obtain the integral representation, we may use the discussion in Section 3.3.3, where we estimate the time to evaluate $w(a)$ by $d \cdot t_{int}$. Also we need to estimate $\int_A |w(a)|$, which takes about $d \cdot (t_{int})^2$.

Altogether, we obtain the expected time complexity

$$d \cdot t_{int}^2 + (ns_{gen}(w)d \cdot t_{int} + n \cdot t_{int})\frac{Q}{Q-1} = O(ns_{gen}(w)d \cdot t_{int}\frac{Q}{Q-1}).$$

In the above equation we needed to compare t_{int} and ns_{gen}: from Section 3.3.1 we have $t_{int} \sim \varepsilon^{-2}$, while $n \sim \varepsilon^{-\frac{1}{1-1/t}}$. We see that for $t = 2$ we have $t_{int} \sim n$, if t gets closer to 1 (arguably the most interesting case) n gets larger. Moreover s_{gen} is at least 1. This allows us to simplify the time complexity as above.

In Section 3.4 we will discuss how this compares to the other algorithms.

3.3 Numerical Issues

In this section we will discuss numerical issues that arose when estimating running times of the algorithms proposed in the previous two sections. We do not claim any originality in this section, and also, given the limited space, we hardly scratch the surface of the extensive field of research, which numerical analysis presents. Our purpose here is to shortly overview relevant known results in order to be able to somewhat understand the running times of the studied algorithms. We will deal with t_{int} in Section 3.3.1, t_{opt} in Section 3.3.2, and t_{apx} in Section 3.3.3.

Our model of computation is a simple one: we assume, that variables in our algorithms hold one real number, and that we can perform basic operations with real numbers at unit cost. Thus, we avoid discussing in detail numerical stability of the presented algorithms and also the effect of required precision is treated only partially.

3.3.1 Numerical Integration

In the discussed algorithms we need to compute numerically various integrals. In this section we start by an overview of known methods for estimating integrals. Then we show how to extend the idea of Monte-Carlo integration for the situation when the integral depends on parameter. We will include an estimate of time complexity of these methods. This will in the end lead to an estimate on t_{int}.

A. Classic methods Finding numerical estimates of definite integrals is a central topic in numerical analysis. The classical approach to estimate $I = \int_S h(x)\,\mathrm{d}x$ is as follows: We divide the region S into small parts (by a "dense regular grid"). On each of them we approximate h by polynomials and integrate those. The next claim summarizes such results.

Claim 3.3.1 (Classic numerical integration [Numerical Basics]) *Suppose $h \in C^r(S)$, where $S \subseteq \mathbb{R}^d$. We can estimate $\int_S h(x)$ using values $h(x_i)$ in some (carefully chosen) points $x_1, \ldots, x_N$. with error at most*

$$O(N^{-\frac{r}{d}})$$

(the constant in $O(\cdot)$ depends on $\max_{x\in S} |f^{(r)}(x)|$ and on the volume of S).

We see another case of the "curse of dimensionality": to achieve error bounded by ε, we need to take $N = \Omega(\varepsilon^{-d/r})$. One way out of this would be to consider functions defined in $\mathbb{R}^d$ that are d-times continuously differentiable, but this may be a too strict requirement. Another common method is to use "sparse grids" – i.e., we use more points to sample the function h in that parts of S, where h varies more. This method yields error $O(\frac{(\log N)^{(d-1)(r+1)}}{N^r})$, where, again, N is the number of sample points and r the regularity of f. When d is large (as is usually the case in applications of neural networks), the standard approach is to use random sampling, which we discuss in the next section.

B. Monte-Carlo methods of integration We start by an overview of the basic Monte-Carlo method. For more details, see e.g. [We00]. In the same setting as above, let $x_1, \ldots, x_N$ be chosen uniformly at random from S. Put

$$E = vol(S)\frac{1}{N}\sum_{i=1}^{N} h(x_i).$$

As expectation of $h(x_i)$ is $I/vol(S)$, we get that E is an unbiased estimator for I. To find how reliable results E yields, one uses tail estimates for sums of independent random variables. For example, by using the Chebyshev and Chernoff-Hoeffding bound C.2.1 we can easily get the following.

Claim 3.3.2 (Monte-Carlo integration, see, e.g., [CuSm01]) *Let h, S, and E be as above. Then the expected value of E is $\int_S h(x)$. Moreover:*
(a) If h is in $\mathcal{L}^2(S)$ and $\sigma^2 = \frac{1}{vol(S)}\int_S (h\, vol(S) - \int_S h(x))^2$, then for any $\eta > 0$ we have

$$\Pr[|E - \int_S h| > \frac{\eta}{\sqrt{N}}] \le \frac{\sigma^2}{\eta^2}\,.$$

(b) If h is in $\mathcal{L}^\infty(S)$ and M is such that $|h(x)vol(S) - \int_S h(s)\,\mathrm{d}s| \le M$ for almost every $x \in S$, then for any $\eta > 0$ we have

$$\Pr[|E - \int_S h| > \frac{\eta}{\sqrt{N}}] \le 2e^{-\frac{\eta^2}{2M^2}}\,.$$

As a side-remark, we note that in the literature about the application side of Monte-Carlo integration, the method is frequently justified by mentioning the Central Limit Theorem. Indeed, Central Limit Theorem gives that the distribution of $(E - \int_S h)\sqrt{N}$ converges to normal distribution. This yields bound similar to that in part (b) for any function h, that satisfies assumptions of the Central Limit Theorem, in particular the Lyapunov condition is satisfied for $h \in \mathcal{L}^{2+\varepsilon}(S)$ for any $\varepsilon > 0$. However, this only yields information about the limit distribution. If we want a bound for some particular N, we need to utilize the Berry–Esseen theorem: it estimates the speed of convergence to the normal distribution, but the guaranteed bound is only $O(1/\sqrt{N})$.

Claim 3.3.2 allows us to quickly estimate the various integrals we need for time complexity estimates for our algorithms. If we want to get the "probable" error ε, it suffices to use $N \ge C(1/\varepsilon)^2$, with the constant C depending on the function to be integrated (and also on the probability with which we want to get the desired precision). As we do not care about constants in our analysis, we summarize this as $t_{int}(\varepsilon) = O(1/\varepsilon^2)$.

3.3.2 Optimization

In this section we briefly touch the problem of optimization, as we are using it in one of our iterative algorithms. We can give only a glimpse of this interesting subject, for a more thorough treatment we suggest for instance [BoVa04].

In Algorithm 3.1.2 we needed to find a function for which a certain functional is small. It would be hard to be searching for an arbitrary function, fortunately we are looking for functions in specific forms. In the perceptron model we are searching for an optimal ridge function $\sigma(a \cdot x + b)$, where σ is the given activation function, $a, x \in \mathbb{R}^d$, $b \in \mathbb{R}$. This means that we are actually minimizing a function of $d + 1$ real variables (a, b).

Generally, suppose we are seeking for a minimum of a function h, defined on $A \subseteq \mathbb{R}^{d'}$; we assume that d' is large. There is a plethora of branches of the optimization theory, where this question is studied under various assumptions (h is linear, quadratic, convex, ...). Unfortunately, none of these assumptions is applicable. The only general approach that guarantees finding the global optimum, is to sample h at all points of a dense enough grid, where what is dense enough depends on some bounds on derivatives of h (or its modulus of continuity). Obviously, this is not practical even for moderately large d', as the number of sample points scales as $m^{d'}$. (Here m is the number of division points in one dimension, which depends on the size of the region we are optimizing over, and on the modulus of continuity of the optimized function.) Thus, we need to leave the realm of guaranteed methods. Most methods used in practice are actually searching for local minimum. Then, a variety of methods (multiple runs of the algorithm with random starting points, simulated annealing, etc.) is used to get a chance of finding the global minimum. Obviously, without some knowledge of the function h, nothing can be said, and we can only attempt for a reasonably good heuristic.

A better understood question is how to search for the local minimum. The basic way is the gradient method:[1] we start at a random point $a = a_0$, compute the gradient $\nabla h(a)$ (assuming it exists), and move against the gradient, by a (carefully guessed) distance. In Algorithm 3.1.2 we need to optimize the peak functional $F_k(g)$. As we discussed after the algorithm $F_k(g) = \int g(x)\psi(x)\,\mathrm{d}x$ for some function ψ. In the notation of this section, the point $a = (a_1, \dots, a'_d) \in A$ corresponds to parameters we use to describe the function g, $h(a)$ is the integral $\int g\psi$. Under mild assumptions the gradient $\nabla h(a)$ is the integral $\int (\nabla g)\psi$. (Here $\nabla g = (\frac{\partial}{\partial a_i} g)_i$ is a collection of functions that describe how g changes if we change the parameters a.) It follows that single step of gradient method will take time $d' \cdot t_{int}$.

The speed of convergence depends on both the chosen variant of this method, and on properties of h. One of the estimates for the number of repetitions (under the assumption of strong convexity of h) is as follows, see Chapter 9 in [BoVa04]. We let κ denote an upper bound on

[1] This is the usual way to carry on the backpropagation algorithm, the common way to train a feedforward neural network.

the condition number of the matrix of the second partial derivatives $\nabla^2 h(a)$. We let $a^* \in A$ be the point where the minimum of h is attained (for strongly convex h such point exists and is unique). Finally, $\varepsilon > 0$ is the required precision, that is we want to find a so that $h(a) - h(a^*) \leq \varepsilon$. The number of iterations of the gradient method to achieve this is at most

$$\frac{\log \frac{h(a_0)-h(a^*)}{\varepsilon}}{\log(1-\frac{1}{\kappa})}. \tag{3.10}$$

Unfortunately, the condition number can be rather large and, what is worse, we need to minimize a function that is not convex. In this view, the time t_{opt} needed to find a minimum of a function, is hard to evaluate. A reasonable estimate is

$$t_{opt} \sim d' \cdot t_{int} \cdot s_{opt} \,,$$

where $s_{opt} = s_{opt}(\varepsilon)$ (the required number of steps) is estimated by the formula (3.10).

3.3.3 Approximation

In this section we discuss several ways to obtain a "starting approximation" of a function $f(x)$ we are trying to estimate. Our point of view will be somewhat different from the one in Sections 3.1 and 3.2 in that we are not striving to get small number of approximating functions. Thus, we allow a somewhat larger number of functions (resulting in a larger time-complexity of obtaining the solution, but, perhaps, also making it easier to find a solution). Then we shall use algorithms in Sections 3.1 and 3.2 to get efficient approximations.

So suppose we are given a function $f(x)$, for $x \in H$. Let us assume that this function is given to us by means of values at sample points (the position of these sample points may or may not be under our control).

High-dimensional optimization One way to find the desired approximation is to choose N and optimize expression $\sum_{i=1}^{N} c_i g_i(x)$. We optimize by modifying the c_i's and the parameters describing the functions g_i. We choose N high enough, so that we can get good enough approximation, but small enough, so that the task is still computationally feasible. By choosing N larger, than will be the number n of the functions in the final approximation, we may hope to partially overcome one of the complications of the high-dimensional optimization – the fact, that optima can be hard to find. By choosing N large, we make sure that very good approximations exist, so our desired precision may not be so hard to achieve. If we use some variant of the gradient method to do the optimization, the amount of computation for one step is N times larger than when we optimize just one function. It is not clear how the number of steps is affected by having a large N. However, we may roughly estimate

$$t_{apx} \sim N \cdot t_{opt} \,.$$

Using integral representation It may be easier to first find an integral representation. (Then we will "sample from the integral", instead from the sum.) Two different approaches to obtain a relevant integral representation were pursued by Kůrková, Kainen and Kreinovich, and by Barron, see Section 2.3. We will concentrate on the approach of Kůrková, Kainen and Kreinovich, that is we will use Theorem 2.3.1. Using this result, we express $f(x)$ as an integral of form

$$f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}x\,.$$

Here, $\varphi(\cdot, a)$ is the Heaviside function in a direction and shift given by a.

We will discuss next, how computationally accessible this representation is. In Algorithm 3.2.5 we needed to evaluate $w(a)$ and $\int_A |w(a)|$. To get $w(a)$ we need to compute the d-th directional derivative of f and integrate it over a hyperplane. Numerical estimate of the d-th derivative takes time proportional to d, so we can evaluate $w(a)$ in time $O(dt_{int})$. To estimate $\int_A |w(a)|$, we will use Monte-Carlo integration, which takes time $O(d(t_{int})^2)$. In this formula, the two instances of t_{int} correspond to different integrals (one over x and the other over a), but in most applications these variables have similar dimension, so we neglect this small distinction.

Kernels Yet another way to obtain the starting approximation to the given function is using the Tikhonov regularization method and Reproducing Kernel Hilbert Spaces. This will be discussed in Chapter 3. In Section 4.3 we will see how the algorithms presented in this chapter may be used in such setting.

3.4 Comparison

In this section we will compare the algorithms under discussion: the iterative algorithms (either with repeated optimization, or using an auxiliary approximation) and the randomized algorithms (using an auxiliary approximation or an integral representation). In Table 3.1 we summarized what we found in the previous sections. However, the time complexities depend highly on the particular problem at hand. Indeed, the choice of functions used in the approximation, as well as the data we are trying to approximate affect, among else, the number of steps needed in the gradient method, which is usually the workhorse of the optimization "subprocedure". We make no attempt to analyze these subtleties; for this reason (and for the sake of brevity) we also suppress $\mathcal{G}$ from the expressions in Table 3.1. We will not try to provide universal conclusion, which of the algorithms is the best. Instead, we will try to address this question in different "scenarios". The scenarios will describe how we search for the "auxiliary" approximation f_{apx}.

Iterative 1 (rep. optimiz.)	$\sum_{k=1}^{n-1} t_{opt}(\frac{c_1}{k^{1-1/t}}) \sim n \cdot t_{opt}(\frac{c_1}{n^{1-1/t}})$
Itrative 2 (using aux. approx.)	$t_{apx}(\frac{c_2\eta}{n^{1-1/t}}, N) + nN \cdot t_{int}$
Randomized 1 (approx.)	$t_{apx}(\frac{c_3\eta}{n^{1-1/t}}, N) + \big(O(n \log N) + n \cdot t_{int}\big)\frac{Q}{Q-1}$
Randomized 2 (int.repr.)	$O(nd \cdot s_{gen}(w) \cdot t_{int} \cdot \frac{Q}{Q-1})$

Table 3.1: Time complexities of the discussed algorithms.

Disclaimer about N One piece of information that will be missing from our analysis is the size of N relative to n. The value of N depends on our choice, and in turns affects the complexity of the approximation. Roughly speaking, if N increases, it is more work to find the approximation, but the approximation will become an easier task (we know from Chapter 2 that the achievable error decreases with the number of terms in the approximation). As such, optimal choice of N depends on the task being solved and the used methods. From practical considerations, however, it is reasonable to expect that $n \leq N \leq n^k$ for some small constant k.

Iterative 2 vs. Randomized 1 In this paragraph we will (somewhat in contrary to the first paragraph) compare the algorithms Iterative 2 and Randomized 1 "universally". Recall that Q in the randomized algorithm describes the expected quality of the solution (we are ready to be satisfied with a solution Q times worse than what is guaranteed to exist). As this is a constant greater than 1, we can see, that when using auxiliary approximation the Randomized algorithm is always superior to Iterative 2. In the following, we will only discuss Iterative 1 and Randomized.

1. scenario: auxiliary approximation is found by high-dimensional optimization As discussed in Section 3.3.3, $t_{apx}(\mathcal{G}, \varepsilon, N) \sim N \cdot t_{opt}(\mathcal{G}, \varepsilon)$. Consequently,

$$t_{Iter1} \leq n \cdot t_{opt} < N \cdot t_{opt} \sim t_{apx} \leq t_{Rand_1} .$$

This suggests, that in this case we should be using the iterative algorithm, especially if N is very large. This is probably the expected result: after all, the idea behind the Iterative algorithm as suggested by Jones [Jo92] is to reduce the dimension of the problem.

2. scenario: an auxiliary approximation is given In this second scenario, we assume that we are given *some* approximation of a function $f(x)$, possibly using many functions. Our task is to find an approximation using less functions and achieving a reasonable error. In this setting, the first iterative algorithm is wasteful, as it cannot anyhow use the given approximation. Indeed, we certainly have $t_{opt} \sim dt_{int}s_{opt} \gg t_{int}$. Also, for the typical range of N we have $t_{opt} \gg \log N$. Consequently, in this scenario the randomized 1 algorithm provides an efficient method to improve a given approximation by "pruning it at random".

$$t_{Rand1} < t_{Iter1}$$

This will be utilized in Chapter 4: there we will consider situations where a natural approach yields sums with large number of terms N (Theorem 4.2.1).

3. scenario: auxiliary approximation is obtained from an integral representation In many situations we may have access to some sort of integral representation of the function under study. Several methods to obtain those were described in Section 2.3. We will use the second version of the rendomized algorithms in such case.

We will consider here an approach of this type, which was studied by Kůrková, Kainen and Kreinovich [KKK97], that is we will use Theorem 2.3.1. As discussed in Section 3.3.3 and 3.2, the time complexity of our algorithm is $O(ns_{gen}(w)d \cdot t_{int}\frac{Q}{Q-1})$ To compare this to the iterative algorithm, recall from Section 3.3.2 that $t_{opt} \sim d \cdot t_{int}s_{opt}$. Thus we are comparing

$$nd \cdot t_{int}s_{gen}(w)\frac{Q}{Q-1} \qquad \text{with} \qquad nd \cdot t_{int}s_{opt}\,,$$

or $s_{gen}(w)\frac{Q}{Q-1}$ with s_{opt}. Now s_{gen} can be as small as 1 (and even in the simplistic approach it is at most $(\max w(a))/\int_A |w(a)|$). In particular, (in our simple analysis, at least) it does not depend on the required precision. On the other hand, s_{opt} is much larger than 1, and it depends on the required error (although only by a factor $\log 1/\varepsilon$). Most importantly, we may run into troubles with finding a local minimum that is not global. Thus, we may conclude that when the integral representation is available, the randomized algorithm that utilizes it (Algorithm 3.2.5) is better than the iterative one.

$$t_{Rand2} < t_{Iter1}$$

3.5 Conclusions

In this chapter we proposed a new algorithm to find neural networks with small number of units and reasonable approximation error. (Equivalently, we are seeking an approximation of a given function by a sum with each term being one of a given set of approximating functions, where we want this sum to have a small number of terms.) Roughly speaking, the algorithm utilizes an auxiliary approximation (using a large number of units) and "prunes" this approximation by taking only some of these units. This selection process is done randomly, and, as shown in Corollary 3.2.3 it provides an approximation with close-to-optimal error in a small expected number of steps.

We analyzed two scenarios, where this algorithm is useful. One is when we are given an approximation and the goal is to use a smaller number of units without increasing the error too much. (This combines well with the techniques that we will discuss in Chapter 4.) Another scenario, where the proposed algorithm is better than the usual one (which uses repeated optimization), is when our function admits an integral representation. We utilize a result of Kůrková, Kainen and Kreinovich [KKK97] (Theorem 2.3.1): given $f \in C^d(\mathbb{R}^d)$, there is an integral representation using Heaviside units with explicitly given weights.

Finally, let us remark, that the suggested randomized algorithm works without modification for other models of neural networks (such as RBF networks).

Chapter 4

Pruning Solutions of Specific Regularization Problems

In this chapter we show a natural way to use randomized algorithm for pruning a too rich network. We consider the approximation task formulated as regularized minimization problem with kernel-based stabilizer. These approximation tasks exhibit easy derivation of solution to the problem in the shape of linear combination of kernel functions (see for example [SchSl02]) that can be interpreted as one-hidden-layer feedforward network schemas. For basic quadratic error functional parameters of these networks can be computed from a linear system; however, these networks tend to have too many hidden units (in fact as many as was the number of data). There are many ways around this problem (see for example [SchWe06]). We propose a solution that uses randomized algorithm (3.2.2) from Chapter 3. This is probably the most interesting result of this chapter.

We also mention concrete applications of a special type of kernels proposed by the author – sum kernels and product kernels. As part of our joint work [KS05b], [KS06] Petra Kudová-Vidnerová tested these new schemas on real-life data and compared them to classical solutions. Experiments on pruning by randomized algorithm are running.

In Section 4.1 we briefly introduce theory of reproducing kernel Hilbert spaces (RKHS) following [Ar50]. This is utilized in further sections: In Section 4.2 learning from data using Tikhonov regularization employing kernels is presented. As has been mentioned in various detail in many works existence, uniqueness and form of solution to the basic problem can be proven. In Section 4.3 we present how to use the randomized algorithm to yield approximations that use less terms than the above derived one. Section 4.4 shows applications of some concrete regularization networks, part of the section comes from joint work with Petra Kudová-Vidnerová [KS05b], in cooperation with her we are also working on experiments employing randomized algorithm to these specific schemas.

Results presented in this chaper have been published in [SS08, KS06, KS05b, S04a].

4.1 Reproducing Kernel Hilbert Spaces

Reproducing kernels proved themselves very useful in approximation theory. We introduce the basic concept and properties in Section 4.1.1. In Sections 4.1.2 and 4.1.3 we show two special types of composite kernels which we will use later in Section 4.4 to derive concrete approximation schemas.

4.1.1 Basic Properties

Reproducing Kernel Hilbert Spaces were formally defined by Aronszajn [Ar50], although related concepts were used even before. They find ample use in applications ranging from PDE's to machine learning, see e.g. a survey [SchWe06]. In this section we present the reader with definitions and a brief overview of the properties needed for our purposes. We follow [Ar50], where we also kindly refer the reader for more details and all of the proofs.

Given a Hilbert space $\mathcal{H}$ of functions (real or complex) defined on $\Omega \subset \mathbb{R}^d$ we let $\mathcal{F}_x$ denote the evaluation functional, that is $\mathcal{F}_x(f) = f(x)$. If for each $x \in \Omega$ the evaluation functional $\mathcal{F}_x$ is continuous, we say that $\mathcal{H}$ is a *Reproducing Kernel Hilbert Space*, shortly an *RKHS*. The term is suggested by the following important property: In an RKHS as above there exists a positive definite symmetric function $k : \Omega \times \Omega \to \mathbb{R}$ (*reproducing kernel* corresponding to $\mathcal{H}$) such that

(i) for any $f \in \mathcal{H}$ and $y \in \Omega$ the following reproducing property holds

$$f(y) = \langle f(x), k(x,y) \rangle \, ,$$

where $\langle \cdot, \cdot \rangle$ is scalar product in $\mathcal{H}$ and

(ii) for every $y \in \Omega$, the function $k_y(x) = k(x,y)$ is an element of $\mathcal{H}$.

We bring here some basic properties of RKHS.

Lemma 4.1.1 (Uniqueness of reproducing kernel [Ar50]) *Let $\mathcal{H}$ be a Hilbert space with a reproducing kernel k. Then k is unique.*

Proof: We include here this short proof (due to [Ar50]) to illustrate the reproducing property of reproducing kernels. Suppose we have two reproducing kernels k, k' and $k \neq k'$. Then for some y we have $0 < \|k_y - k'_y\|^2 = \langle k_y - k'_y, k_y - k'_y \rangle = \langle k_y - k'_y, k_y \rangle - \langle k_y - k'_y, k'_y \rangle$. By the property (i) in the definition of kernel, this equals $(k(y,y) - k'(y,y)) - (k(y,y) - k'(y,y)) = 0$, which is a contradiction. □

Lemma 4.1.2 (Existence of reproducing kernel [Ar50]) *Let $\mathcal{H}$ be a Hilbert space of functions on $\Omega \subset \mathbb{R}^d$. Suppose that all evaluation functionals $\mathcal{F}_x$ are continuous. Then there exists a reproducing kernel satisfying properties (i) and (ii) and from the reproducing property it follows that the kernel is positive definite. On the other hand from (i) and (ii) we conclude that evaluation functionals are continuous.*

Lemma 4.1.3 (Uniqueness of RHKS [Ar50]) *To every $k(x, y)$ satisfying the properties (i) and (ii) there corresponds one and only one Hilbert space $\mathcal{H}$ admitting k as a reproducing kernel.*

4.1.2 Sum of Reproducing Kernels

In this section we will present the notion of a sum of RKHS's, as described in [Ar50]. This will be utilized later, in Section 4.4.2. The sum of two RKHS's is defined simply as the space containing all sums of two functions, one from each of the RKHS's. Some care has to be taken when defining the norm (and scalar product) as there may be many sums that yield the same function.

For $i = 1, 2$ let F_i be an RKHS of functions on Ω, let k_i be the corresponding kernel and $\|.\|_i$ the corresponding norm. Consider the space $H = \{\{f_1, f_2\} \mid f_1 \in F_1, f_2 \in F_2\}$. with norm given by $\|\{f_1, f_2\}\|^2 = \|f_1\|_1^2 + \|f_2\|_2^2$.

We have to deal with duplicities. Consider the class F_0 of all functions f belonging to $F_1 \cap F_2$. We define $H_0 := \{\{f, -f\}; f \in F_0\}$. H_0 is a closed subspace of H and thus we can write $H = H_0 \oplus H'$, where H' is the subspace complementary to H_0. Now to every element $\{f_1, f_2\}$ of H there corresponds a function $f(x) = f_1(x) + f_2(x)$. This is a linear correspondence transforming H into a linear class of functions F. Elements of H_0 are transformed into zero functions and thus the correspondence between H' and F is one-to-one and has an inverse (for $f \in F$ we let $\{g_1(f), g_2(f)\}$ be the corresponding element in $H' \subseteq H$). We define norm on F by

$$\|f\|^2 = \|\{g_1(f), g_2(f)\}\|^2 = \|g_1(f)\|_1^2 + \|g_2(f)\|_2^2 . \tag{4.1}$$

Here we present a result from [Ar50] showing that to the class F with the above defined norm there corresponds a reproducing kernel $k = k_1 + k_2$.

Theorem 4.1.4 (Sum-kernel RKHS [Ar50]) *Let F_i be RKHS and k_i and $\|.\|_i$ the corresponding kernels and norms. Let F be as above with norm given by* (4.1)*. Then*

$$k(x, y) = k_1(x, y) + k_2(x, y)$$

is the kernel corresponding to F.

The same holds when F *is defined as the class of all functions* $f = f_1 + f_2$ *with* f_i *in* F_i *and norm* $\|f\|^2 = \min \|f_1\|_1^2 + \|f_2\|_2^2$, *the minimum taken over all decompositions* $f = f_1 + f_2$ *with* f_i *in* F_i.

It is easy to extend this theorem to sum of more than two spaces: $k(x,y) = \sum_{i=1}^{n} k_i(x,y)$. We call F a Sum-Kernel Reproducing Hilbert Space.

4.1.3 Product of Reproducing Kernels

Here we will consider product of Reproducing Kernel Hilbert Spaces, again following [Ar50]. For $i = 1, 2$ let F_i be an RKHS of functions on Ω_i, let k_i be the corresponding kernel. Consider the set of functions on $\Omega = \Omega_1 \times \Omega_2$ defined by

$$F' = \{\sum_{i=1}^{n} f_{1,i}(x_1) f_{2,i}(x_2) \mid n \in \mathbb{N}, f_{1,i} \in F_1, f_{2,i} \in F_2\}\,.$$

Clearly, F' is a vector space. We define a scalar product on F': Let f, g be elements of F' expressed as $f(x_1, x_2) = \sum_{i=1}^{n} f_{1,i}(x_1) f_{2,i}(x_2)$, and $g(x_1, x_2) = \sum_{j=1}^{m} g_{1,j}(x_1) g_{2,j}(x_2)$. We define

$$\langle f, g\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \langle f_{1,i}, g_{1,j}\rangle_1 \langle f_{2,i}, g_{2,j}\rangle_2$$

(with $\langle\cdot,\cdot\rangle_i$ denoting the scalar product in F_i). It is a routine to check that this definition does not depend on the particular form in which f and g are expressed and that the properties of scalar product are satisfied. We define norm on F' by $\|f\| = \sqrt{\langle f, f\rangle}$. Finally, let F be the completion of F'. It can be shown [Ar50] that the completion exists not only as an abstract Hilbert space but that F is in fact a space of functions on Ω. We call F the product of F_1 and F_2 and write $F = F_1 \otimes F_2$.

Theorem 4.1.5 (Product-kernel RKHS [Ar50]) *For* $i = 1, 2$ *let* F_i *be an RKHS on* Ω_i *with kernel* k_i. *Then the product* $F = F_1 \otimes F_2$ *on* $\Omega_1 \times \Omega_2$ *is an RKHS with kernel given by*

$$k((x_1, x_2), (y_1, y_2)) = k_1(x_1, y_1) k_2(x_2, y_2)\,, \tag{4.2}$$

where $x_1, y_1 \in \Omega_1$, $x_2, y_2 \in \Omega_2$.

4.2 Learning from Data in RKHS Spaces

As advertised at the beginning of this chapter we are interested in learning from data, i.e., we are given a set of data $z = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}$ and we want to "learn" from it how to obtain y from a given x. This means we want to find a function f so that $f(x_i) \doteq y_i$. We do not require exact equality as we may assume the given data are not precise.

However, even if we require $f(x_i) = y_i$ exactly, the problem to find f is still *ill-posed* – there are many solutions to it (see Figure 4.1). This issue is addressed in a variety of ways (Lagrange interpolation, least square interpolation, ...). The approach to the approximate problem we describe here was suggested by Tikhonov [TiAr77] and studied by a plethora of researchers (see for example [PoSm03, SchSl02, CuSm01, Gi98, PaSa93, Wa90]). We follow the most basic exposition of [GJP95].

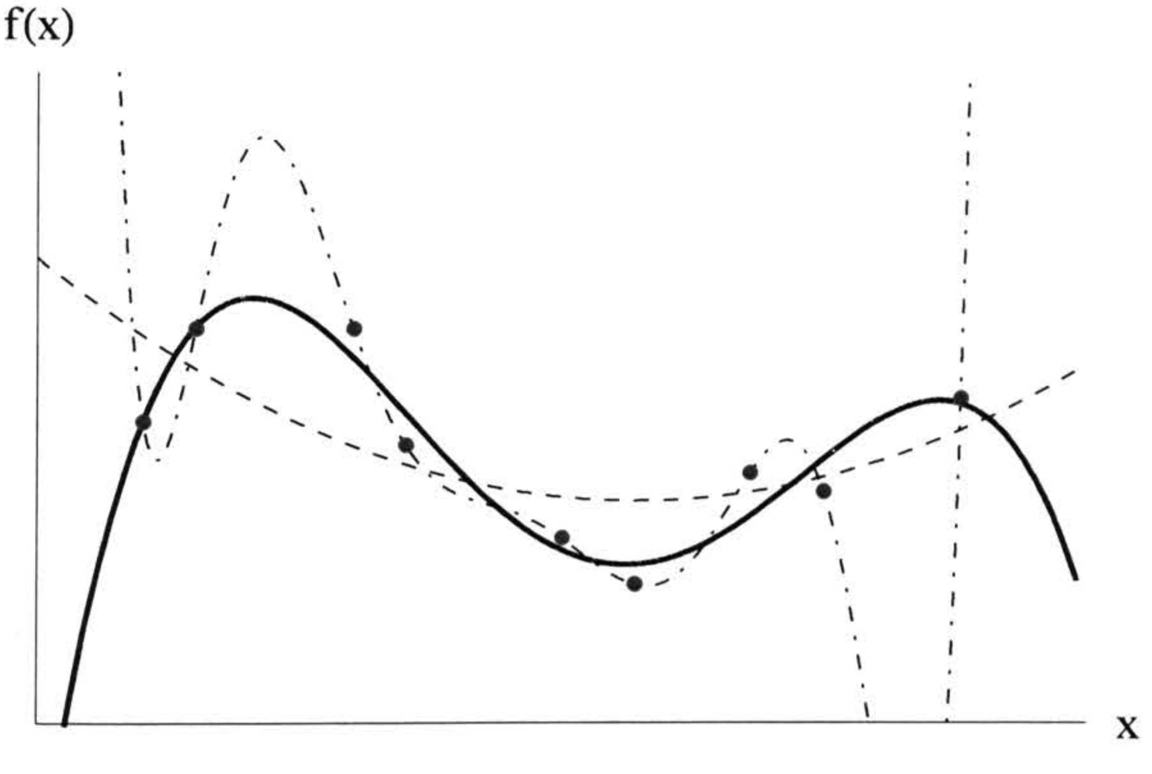

Figure 4.1: Illustration of the basic problem in learning. Data, an over-fitted function, a well-fitted function and a too much generalised function

We combine the goal of fitting the data with the goal to find a function with good "global properties" or complying with some a-priori knowledge. This can be partially achieved by appropriate choice of the space over which we are minimizing but we can do more. The common approach constructs a functional that quantifies the precision to which a given function fits the data and also the global properties; then we seek to minimize this functional over appropriately chosen set of functions.

Let us be more precise now. We are looking for a function $f : \Omega \to \mathbb{R}$ as a member of some function space X. (Thus the given data z are subset of $\Omega \times \mathbb{R}$ and we also assume $\Omega \subseteq \mathbb{R}^d$.) We let the functional $\mathcal{F}$ be defined by

$$\mathcal{F}(f) = \mathcal{E}_z(f) + \gamma\Phi(f) \tag{4.3}$$

where $\mathcal{E}_z$ is the error functional depending on the data $z = \{(x_i, y_i)\}_{i=1}^N$ and penalizing distance from the data, Φ is the regularization part – the *stabilizer* – penalizing "remoteness from the global property" and γ is the regularization parameter giving the trade-off between the two terms of the functional to be minimized.

The error functional is usually of the form $\mathcal{E}_z(f) = \sum_{i=1}^N V(f(x_i), y_i)$. We can easily infer sufficient conditions on V (for solvability) from proof of Theorem 4.2.1, but we shall not elaborate on it here. A typical example of the empirical error functional is the classical mean square error: $\mathcal{E}_z(f) = \frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2$ and we will restrict ourselves to it as it exhibits easy derivation of solution. In order for this problem to be well-defined, we would have to have the function f defined pointwise – which is not always the case (consider for example $X = \mathcal{L}^2(\Omega)$ where we may change the values of f on any set of measure 0). In the sequel we will assume, that the *evaluation functionals* $f \mapsto f(x)$ are not only well-defined but also continuous. In the important case when we want X to be a Hilbert space as well, we will be working with the reproducing kernel Hilbert spaces, as introduced in Section 4.1.

If X is an RKHS with norm $\|\cdot\|_k$, a useful choice for the regularization functional is $\Phi(f) = \|f\|_k^2$ (square of the norm), yielding

$$\mathcal{F}(f) = \mathcal{E}_z(f) + \gamma\|f\|_k^2 \,. \tag{4.4}$$

For such $\mathcal{F}$ it is possible to show the existence and uniqueness of the minimum and derive a simple form for this minimum. Many authors have addressed this topic for general or specific settings (see for example [Wa90, SchSl02, Gi98, PaSa93, PoSm03] and others).

Derivation of the shape of the solution to the regularized minimization problem is referred to as Representer theorem. For the proof of existence, uniqueness and form of solution to the most basic setting we ask the reader to refer to Appendix C.2.9. We presented the proof merely to acquaint the reader with the nice behaviour of the schema and partly also because we were unable to find these simple arguments put cleary elsewhere.

4.2.1 Representer Theorem

In this section we will discuss the solution to the problem of minimizing the functional in (4.4). This is a well-known question addressed for instance in [Wa90]. Our treatment follows the one sketched in [GJP95], which seems to be applicable to a wider class of functionals (4.3). Some details are omitted there, though, in particular it is not verified that the minimum exists. (Incidentally, the existence of minimum is not verified in [Wa90], either, although the proof presented there is easy to fix.) We present the proof in Appendix C, restated as Theorem C.2.9 (without any claim on being original) for the reader's convenience, as this result is the basis of the results in the later parts of this chapter. Also, the usage of RKHS's enables the proof to be short and iluminating.

Theorem 4.2.1 (Representer Theorem – basic version [KiWa71, SchSl02]) *Let $z = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}$ be given, let X be an RKHS with kernel k and norm $\| \cdot \|_k$. Let the functional $\mathcal{F}(f)$ be given by*

$$\mathcal{F}(f) = \frac{1}{N} \sum_{i=1}^N (f(x_i) - y_i)^2 + \gamma \|f\|_k^2$$

for a function $f \in X$ ($\gamma > 0$ is a constant).

Then there exists a unique function $f_0 \in X$ that minimizes $\mathcal{F}$. Moreover, f_0 is of form

$$f_0(x) = \sum_{i=1}^N c_i k(x, x_i)\,, \tag{4.5}$$

for real c_i's, where these can be found as

$$c_i = (K[x] + \gamma N I)^{-1} y_i,$$

where $K[x]_{i,j} = k(x_i, x_j)$ is Gramm matrix of kernel k with respect to vector x, I is $N \times N$ identity matrix.

Notice that here we consider only the simplest case where existence and uniqueness of minima can be proven easily. For more detail see for example [SchSl02] or [DRCDO05, Ku04], where (as opposed to proof presented in Appendix C.2.9) theory of inverse problems is used.

We remark that the same proof comes through also when the error part of the functional is $\frac{1}{N}\sum_{i=1}^N V(f(x_i) - y_i)$ for any continuous convex function V, however derivation of form of the solution will use derivative of V.

The form of solution was derived for example in [GJP95], we will use it in Section 4.4. Solutions have been derived also for schemas where uniqueness cannot be proven (regularization part Φ is not strictly quasiconvex). In that case we have to add functions from null space of Φ to the solution (see e.g. [GJP95]).

The solution derived above is very nice, since it resembles a neural network with $k(x, x_i)$ as the activation functions parameterized by the data points x_i.

The problem of the number of hidden units being too large to be implemented can be solved by variable-basis approximation using the obtained shape of the activation functions (see [KuSa05b]). Another approach is proposed and discussed in Section 4.3.

4.3 Randomized Pruning

In this section we will show how to combine results of Chapter 3 and 4. We saw how to apply regularization to find a function f that strikes a good balance between fitting the given data $z = \{(x_i, y_i)\}_{i=1}^N$ and possessing good global properties – we find f_0 so that $\mathcal{F}(f_0)$ is as small as possible ($\mathcal{F}$ is given by (4.4)). Theorem 4.2.1 claims that the (unique) optimal function with respect to these criteria is of a simple form (4.5), the only drawback being that the number of terms in the sum is equal to the number of data – which may be too many. We will show, how to "prune" this expression to get an expression using lower number of terms.

The basic idea is simple. We approximate the minimizing function f_0 using Algorithm 3.2.2. We only need to show that this does not change the value of $\mathcal{F}$ too much (here we will use again, that we are searching for an approximating function as a member of a certain Hilbert space). The regularization paradigm then suggests, that we will get a reasonable solution to the original problem.

Theorem 4.3.1 (Randomized pruning of kernel-based neural networks) *Let X be an RKHS with kernel k and norm $\|\cdot\|_k$ and according to Theorem 4.2.1 let $f_0 = \sum_{i=1}^N c_i k(x, x_i)$ be the unique solution of minimization problem*

$$\mathcal{F}(f) = \frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2 + \gamma\|f\|_k^2$$

based on data $z = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}$ and some a-priori knowledge.Then for any $\delta > 0$ there is $n = O(\frac{1}{\delta^2})$ for which we can find

$$g_{apx}^n = \tfrac{1}{n}\sum_{j=1}^n k(x, x_j'),$$

so that

$$\sqrt{\mathcal{F}(g_{apx}^n)} \le \sqrt{\mathcal{F}(f_0)} + \delta\,.$$

Each x_j' ($j = 1, \ldots, n$) is one of x_i ($i = 1, \ldots, N$) from the data z.

Proof: We use Algorithm 3.2.2 to obtain g_{apx}^n. We will utilize our analysis of the algorithm to derive the claim. To do so we introduce an auxiliary norm on X:

$$\|g\|_{aux} = \sqrt{\frac{1}{N}\sum_{i=1}^N (g(x_i))^2 + \gamma\|g\|_k^2}\,.$$

This is indeed a norm, and it corresponds to the scalar product

$$\langle g, h\rangle_{aux} = \frac{1}{N}\sum_{i=1}^N g(x_i)h(x_i) + \gamma\langle g, h\rangle_k\,.$$

For every $g \in X$ we have $\|g\|_{aux} \geq \sqrt{\gamma}\|g\|_k$. Consequently, a Cauchy sequence in $\|\cdot\|_{aux}$ is also Cauchy in $\|\cdot\|_k$, thus the completeness of $(X, \|\cdot\|_k)$ implies completeness of $(X, \|\cdot\|_{aux})$ – so we again have a Hilbert space. Consequently, by Corollary 3.2.3 we may efficiently obtain a simpler function g^n_{apx} so that $\|f_0 - g^n_{apx}\|_{aux} < \delta$ and g^n_{apx} uses $n = O(1/\delta^2)$ terms in the sum. Now by the triangle inequality for $\|\cdot\|_k$ and for the $\mathcal{L}_2$ norm we get that

$$\begin{aligned}
\sqrt{\mathcal{F}(g^n_{apx})} =& \sqrt{\sum_i (g^n_{apx}(x_i) - y_i)^2 + \gamma\|g^n_{apx}\|^2_k} \\
=& \sqrt{\sum_i \Big((g^n_{apx}(x_i) - f_0(x_i)) + (f_0(x_i) - y_i)\Big)^2 + \gamma\|(g^n_{apx} - f_0) + f_0\|^2_k} \\
\leq& \sqrt{\sum_i (g^n_{apx}(x_i) - f_0(x_i))^2 + \gamma\|g^n_{apx} - f_0\|^2_k} \\
&+ \sqrt{\sum_i (f_0(x_i) - y_i)^2 + \gamma\|f_0\|^2_k} \\
=& \|f_0 - g^n_{apx}\|_{aux} + \sqrt{\mathcal{F}(f_0)} \\
\leq& \sqrt{\mathcal{F}(f_0)} + \delta\,.
\end{aligned}$$

□

We see that $\mathcal{F}(g^n_{apx})$ is close to the optimal value $\mathcal{F}(f_0)$, so g^n_{apx} is a reasonable function to be learnt from the original data assuming that $\mathcal{F}$ is the right measure.

In the following remark we will try clarify the quality of g^n_{apx} as an approximation to the original underlying function.

Remark 4.3.2 (Kernel-based networks with feasible number of units) *Let $z = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}$ be given data sampled from an unknown underlying function f_{orig}, (that is, $f_{orig}(x_i) = y_i$ for $i = 1, \ldots, N$), let X be an RKHS with kernel k and norm $\|\cdot\|_k$ derived from a-priori knowledge on f_{orig}, in particular we assume that $f_{orig} \in X$.*

Let again the functional $\mathcal{F}(f)$ be given by

$$\mathcal{F}(f) = \frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2 + \gamma\|f\|^2_k$$

for a function $f \in X$ ($\gamma > 0$ is a constant). Let $f_0 \in X$ be the unique function that minimizes $\mathcal{F}$; according to Theorem 4.2.1 it is of the form

$$f_0(x) = \sum_{i=1}^N c_i k(x, x_i)\,, \tag{4.6}$$

for some real c_i's.

Let $\varepsilon_0 > 0$ be such that

$$\|f_{orig} - f_0\|_k < \varepsilon_0 .$$

If our choice of X and γ capture well the properties of f_{orig} the we can assume that ε_0 is small.

Then we can claim the following estimate

$$\|g^n_{apx} - f_{orig}\|_k < Q^{1/2} \cdot \Big(\frac{\rho + \varepsilon_0}{\sqrt{n}} + \varepsilon_0\Big).$$

Here $g^n_{apx} = \frac{1}{n}\sum_{j=1}^n k(x, x'_j)$, each x'_j ($j = 1, \dots, n$) is one of x_i ($i = 1, \dots, N$) from the data z; $Q > 0$, ρ is a positive constant depending on X and f_{orig} but not on n.

Now it is appropriate to comment on the constant ρ that appears in the above bound. Our reader might have noticed that analysis of ρ was the topic of most of Chapter 2. By Theorem 2.1.3 (upon which Corollary 3.2.3 is based) ρ is defined so that it holds: for every $g \in \mathcal{G}$ (here $\mathcal{G} = \{k(\cdot, x_1), \dots, k(\cdot, x_N)\}$ we have $\|g - f_0\|_k \leq \rho$ which can be further expressed in terms of $\mathcal{G}$-variation ($\|f_0\|_{\mathcal{G}} \leq \sum_{i=1}^N |c_i|$) and $s_{\mathcal{G}} = \max k(x_i, x_i)$. Heuristically ρ expresses how good our chosen kernel-based approximation schema is for approximating given f_{orig} and how good was our a-priori knowledge of f_{orig} that helped us create kernel k and further also the functional $\mathcal{F}$.

The proposed method seems to be of practical interest, in cooperation with Petra Vidnerová we work on experiments showing applicability in practical settings.

4.4 Specific Types of Kernels

In this section we try to come closer to practical results by presenting a few specific examples of kernels used in regularized minimization and mainly by proposing their natural combinations that may aid to tailor regularized minimization schemas better to specific situations.

4.4.1 Simple Kernels

Lef f be an $\mathcal{L}^1$ function on $\mathbb{R}^d$. In [GJP95] a special stabilizer based on the Fourier Transform was proposed:

$$\Phi_K(f) = \int_{\mathbb{R}^d} \frac{|\hat{f}(s)|^2}{\hat{K}(s)}\, \mathrm{d}m_d(s),$$

where $\hat{K} : \mathbb{R}^d \to \mathbb{R}_+$ is a symmetric function ($\hat{K}(s) = \hat{K}(-s)$) tending to zero as $\|s\| \to \infty$, $\hat{f}$ is Fourier transform of f and m_d is the normalized d-dimensional Lebesgue measure (m_d on $\mathbb{R}^d$ is definded as $\mathrm{d}m_d(x) = (2\pi)^{-d/2}\,\mathrm{d}\lambda(x)$). In this setting $1/\hat{K}$ is a low-pass filter.

Thus the functional $\mathcal{F}$ to be minimized is of the form:

$$\mathcal{F}(f) = \mathcal{E}_z(f) + \Phi_K(f) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 + \gamma \int_{\mathbb{R}^d} \frac{|\hat{f}(s)|^2}{\hat{K}(s)}\,\mathrm{d}m_d(s),$$

where $\gamma \in \mathbb{R}^+$. Now we show how to build an RKHS corresponding to the regularization part of our functional:

Let us define

$$k(x,y) = K(x-y) = \int_{\mathbb{R}^d} \hat{K}(t) e^{it\cdot x} e^{-it\cdot y}\,\mathrm{d}m_d(t).$$

For $k \in \mathcal{S}(\mathbb{R}^{2d})$ symmetric positive definite we obtain an RKHS X (using the classic construction, see [SchSl02], [Gi98], [Wa90]). We put $\langle f, g\rangle_X = \int_{\mathbb{R}^d} \frac{\hat{f}(s)\overline{\hat{g}(s)}}{\hat{K}(s)}\,\mathrm{d}m_d(s)$ and obtain the norm

$$\|f\|_X^2 = \int_{\mathbb{R}^d} \frac{|\hat{f}(s)|^2}{\hat{K}(s)}\,\mathrm{d}m_d(s).$$

This enables us to put X to be the completion of the set $\mathrm{span}\{k(x,.), x \in \mathbb{R}^d\}$ in the above norm. It is easy to check the reproducing property of k, resp. K on X, that is $\langle f(x), K(x-y)\rangle_X = f(y)$.

Special types of reproducing kernels and following RKHS are for example the well known

- Gaussian kernel $k_1(x,y) = e^{-\|x-y\|^2}$ with Fourier transform $\hat{K}_1(s) = e^{-\frac{\|s\|^2}{2}}$ or in one dimension
- kernel $k_2(x,y) = e^{-|x-y|}$ with Fourier transform $\hat{K}_2(s) = (1+s^2)^{-1}$.

The norm for this RKHS is of the form $\|f\|_{k_2}^2 = \int \frac{|\hat{f}|^2}{(1+s^2)^{-1}} = \|f\|_2^2 + \|f'\|_2^2$. So we see we obtain a Sobolev space $\mathcal{W}^{1,2}$. These and many more specific instances of kernels are presented for example in [SchWe06]. Here also feature maps for deriving specifically tailored kernels are discussed.

4.4.2 Composite Kernels

This section discusses two types of composite kernels proposed by the author and proposes their possible use in practical situations. Experiments have been done by Petra Vidnerová, se for example [KS06].

Sum of Kernels Here we will consider a more sophisticated type of kernels – sum of kernels introduced in Section 4.1.2. We will study two cases. First suppose that a-priori knowledge or analysis of data suggests to look for a solution in the form of a sum of two functions (for example data is generated from function influenced by two sources differing in frequency). We will use a kernel summed of two parts (employing Theorem 4.1.4) corresponding to high and low frequencies, in the easiest case two Gaussians of different widths:

$$k(x,y) = k_1(x,y) + k_2(x,y) = e^{-\frac{\|x-y\|^2}{d_1}} + e^{-\frac{\|x-y\|^2}{d_2}}$$

In this case we will consider regularized minimization schema of the form:

$$\mathcal{F}(f) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 + \gamma \int_{\mathbb{R}^d} \frac{|\hat{f}(s)|^2}{\widehat{(K_1 + K_2)}(s)}\,\mathrm{d}m_n(s). \tag{4.7}$$

Since we operate in an RKHS we can employ Representer theorem (for this case the simplified version of it 4.2.1) and obtain solution in the form of

$$f_0(x) = \sum_{i=1}^{N} c_i \left(e^{-\frac{\|x-x_i\|^2}{d_1}} + e^{-\frac{\|x-x_i\|^2}{d_2}}\right). \tag{4.8}$$

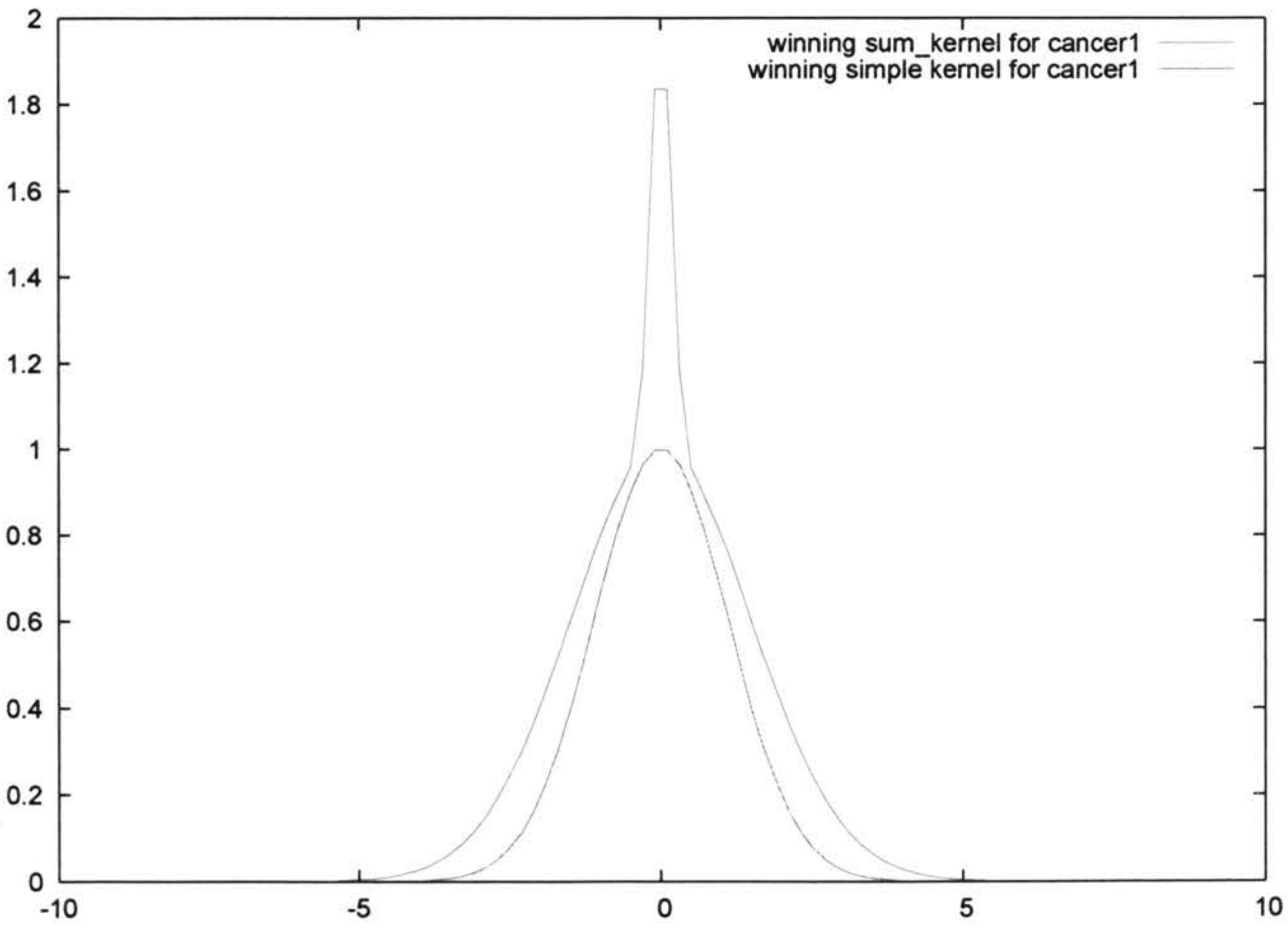

Figure 4.2: An example of the sum kernel (an optimal combination of widths for `cancer1` `data`).

This schema has been tested by Petra Vidnerová in [KS06] on data from [Proben], (see table 4.1, that shows also experiments done on product kernel networks discussed in the next paragraph). Figure 4.2 shows sum kernel as derived for specific task of `cancer1` from [Proben].

The experiments proved the schema to be promising for special types of tasks. Further experiments on improved schema (see section 4.3) are running.

Another conceivable task would be to approximate data with different distribution in the input space. Here again sum of kernels might be helpful if we use different kernels for different parts of the input space.

First, we present the following auxiliary lemma from [Ar50].

Lemma 4.4.1 (Restriction of kernel [Ar50]) *Let X be an RKHS of real-valued functions on Ω with k as kernel. Then function k_A defined by*

$$k_A(x,y) = \begin{cases} k(x,y) & \textit{if } x,y \in A, \\ 0 & \textit{otherwise;} \end{cases}$$

is a kernel for a space $X_A(\Omega) = \{f_A; f \in X \textit{ and } f_A(x) = f(x) \textit{ if } x \in A \textit{ and } f_A(x) = 0 \textit{ otherwise}\}$.

We may use this lemma for different sets $A \subseteq \Omega$. Then we can apply Theorem 4.1.4 for kernels gained in this way. Consequently, our kernels may look as follows:

Choose constants $d_1, \ldots, d_r > 0$ and partition $\mathbb{R}^d$ to sets $A_1, \ldots, A_r$. Then define

$$k_j(x,y) = K_j(x-y) = \begin{cases} e^{-d_j\|x-y\|^2} & x,y \in A_j \\ 0 & \text{otherwise.} \end{cases}$$

The outcomming sum kernel then is $k(x,y) = \sum_{j=1}^{r} k_j(x,y)$ and we can chose the basic regularization schema

$$\mathcal{F}(f) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i)-y_i)^2 + \gamma \int_{\mathbb{R}^d} \frac{|\hat{f}(s)|^2}{\widehat{\sum K_j}(s)} \, \mathrm{d}m_d(s).$$

Now by 4.2.1 again we obtain a form in which we will expect the solution:

$$f_0(x) = \sum_{i=1}^{N} c_i \sum_{j=1}^{r} k_j(x-x_i).$$

At the present we have unfortunately no experiments available to show performance of this type of sum kernel schema.

Let us proceed with another type of composite kernels:

	RN		SKRN$_1$		PKRN		SKRN$_2$	
Task	E_{train}	E_{test}	E_{train}	E_{test}	E_{train}	E_{test}	E_{train}	E_{test}
cancer1	2.28	**1.75**	0.00	1.77	2.68	1.81	2.11	1.93
cancer2	1.86	3.01	0.00	**2.96**	2.07	3.61	1.68	3.37
cancer3	2.11	2.79	0.00	**2.73**	2.28	2.81	1.68	2.95
card1	8.75	**10.01**	8.81	10.03	8.90	10.05	8.55	10.58
card2	7.55	**12.53**	0.00	12.54	8.11	12.55	7.22	13.03
card3	6.52	12.35	6.55	**12.32**	7.01	12.45	6.22	12.86
diabetes1	13.97	16.02	14.01	**16.00**	16.44	16.75	12.92	16.66
diabetes2	14.00	**16.77**	13.78	16.80	15.87	18.14	13.64	17.33
diabetes3	13.69	16.01	13.69	**15.95**	16.31	16.62	12.85	16.34
flare1	0.36	0.55	0.35	**0.54**	0.36	**0.54**	0.35	0.59
flare2	0.42	0.28	0.44	**0.26**	0.42	0.28	0.41	0.28
flare3	0.38	0.35	0.42	**0.33**	0.40	0.35	0.38	0.34
glass1	3.37	6.99	2.35	**6.15**	2.64	7.31	2.56	6.78
glass2	4.32	7.93	1.09	**6.97**	2.55	7.46	3.27	7.29
glass3	3.96	7.25	3.04	**6.29**	3.31	7.26	3.48	6.44
heart1	9.61	**13.66**	0.00	13.91	9.56	13.67	9.51	13.79
heart2	9.33	13.83	0.00	**13.82**	9.43	13.86	8.52	14.31
heart3	9.23	15.99	0.00	**15.94**	9.15	16.06	8.30	16.75
hearta1	3.42	4.38	0.00	**4.37**	3.47	4.39	3.20	4.45
hearta2	3.54	4.07	3.51	**4.06**	3.28	4.29	3.17	4.34
hearta3	3.44	4.43	0.00	4.49	3.40	4.44	3.37	**4.40**
heartac1	4.22	**2.76**	0.00	3.26	4.22	**2.76**	3.68	3.37
heartac2	3.50	3.86	0.00	**3.85**	3.49	3.87	2.99	3.97
heartac3	3.36	**5.01**	3.36	**5.01**	3.26	5.18	3.14	5.13
heartc1	9.99	16.07	0.00	**15.69**	10.00	16.08	6.50	16.07
heartc2	12.70	**6.13**	0.00	6.33	12.37	6.29	11.06	6.69
heartc3	8.79	12.68	0.00	12.38	8.71	12.65	9.91	**11.74**
horse1	7.35	**11.90**	0.20	**11.90**	14.25	12.45	7.66	12.62
horse2	7.97	15.14	2.84	**15.11**	12.24	15.97	6.84	15.70
horse3	4.26	**13.61**	0.18	14.13	9.63	15.88	8.56	15.24
soybean1	0.12	0.66	0.11	0.66	0.13	0.86	0.12	**0.64**
soybean2	0.24	**0.50**	0.25	0.53	0.23	0.71	0.19	0.54
soybean3	0.23	0.58	0.22	**0.57**	0.21	0.78	0.15	0.72

Table 4.1: Comparisons of errors on training and testing set for RN with Gaussian kernels and SKRN and PKRN.

Product of Kernels We consider the product of kernels introduced in Section 4.1.3. Suppose that a-priori knowledge of our data suggests to look for the solution as a member of product of two function spaces. In one dimension the data may be clustered faraway thus being suitable for approximation via narrow Gaussian kernels, in the other dimension the data is smooth, hence we will use broader Gaussian kernel. Employing Theorem 4.1.5 we obtain a kernel for the product space of the form: $k((x_1, x_2), (y_1, y_2)) = k_1(x_1, y_1) \cdot k_2(x_2, y_2) = e^{-\|x_1-y_1\|^2} \cdot e^{-\|x_2-y_2\|}$, where $x_1, y_1 \in \Omega_1$, $x_2, y_2 \in \Omega_2$.

Regularized minimization schema in this case is of the form:

$$\mathcal{F}(f) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 + \gamma \int_{\mathbb{R}^d} \frac{|\hat{f}(s)|^2}{\widehat{k_1 k_2}(s)} \, \mathrm{d}m_d(s). \tag{4.9}$$

Taking advantage of this being an RKHS we have the form of the solution to such a type of minimization:

$$f_0(x_1, x_2) = \sum_{i=1}^{N} c_i e^{-\|x_1-x_{i,1}\|^2} \cdot e^{-\|x_2-x_{i,2}\|}. \tag{4.10}$$

Product kernel network of this type was used to predict the flow rate on the Czech river Ploučnice [KS06]. Our goal was to predict the current flow rate from the flow rate and total rainfall from the previous date, i.e., we were approximating a function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

We have chosen the Gaussian function $e^{-\left(\frac{\|c-x\|}{d_i}\right)^2}$ for both kernel functions (k_1 and k_2), the kernels differ in the width d_i of the Gaussian. All parameters γ, d_1 and d_2 were estimated by crossvalidation. Outcomes of the experiments done by Petra Vidnerová are presented in Figure 4.3 and in Table 4.2.

The Table 4.2 compares the resulting errors of Product Kernel Regularization Network, Regularization Network (for detailed discussion see [K06]) and *conservative predictor*. Conservative predictor is a predictor saying that the value will be the same as it was yesterday, and in spite of its simplicity it is very successful on some tasks, including this one. We can see that the PKRN overperforms both the Regularization Network and Conservative Predictor. We can also see that PKRN shows higher parameter γ which suggest better generalisation.

Results of Section 4.3 on pruning were not used in these experiments. They are, however, the topic of our current cooperation with Petra Vidnerová.

4.5 Conclusion

In this section we have shown how to use randomized algorithm proposed in Section 3.2 in the specific case of neural networks derived from kernel-based regularized minimization schemas.

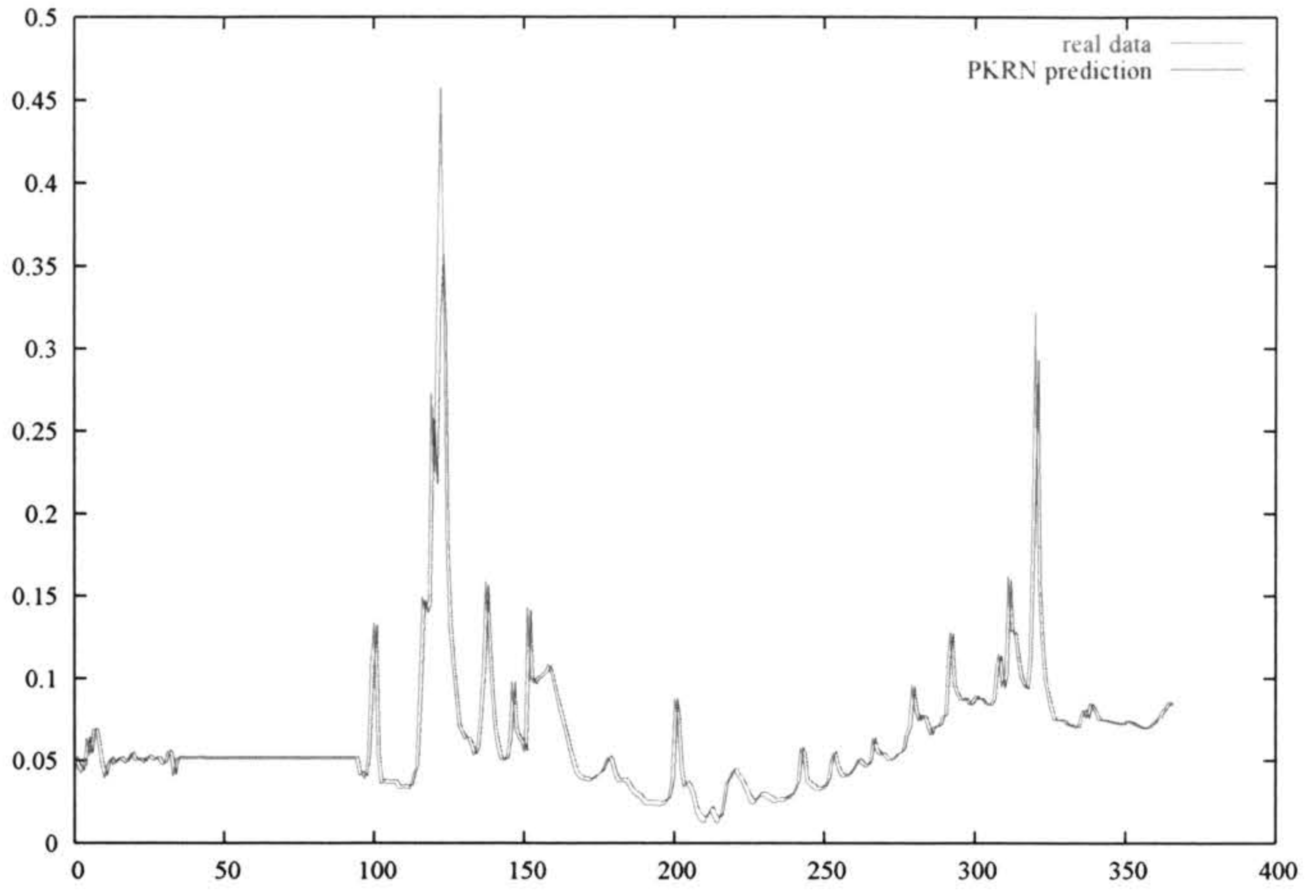

Figure 4.3: Prediction of flow rate on the river Ploucnice.

	PKRN	**RN**	**CP**
E_{train}	0.057	0.008	0.093
E_{test}	0.048	0.056	0.054
network parameters	$\gamma = 10^{-4}$ $d_1 = 0.8$ $d_2 = 1.9$	$\gamma = 1.48 \cdot 10^{-5}$ $d = 0.5$	

Table 4.2: Comparison of error of Product Kernel Regularization Network (PKRN), Regularization Network (RN) and Conservative Predictor (CP) on training set and testing set.

We overviewed briefly basic theory regarding RKHS spaces and regularized minimization that is applicable to our objective. Based on this theory and on results of Chapter 3 we show how to overcome the intrinsic problem of regularization-inferred neural networks – i.e. too high number of hidden units. We have derived bounds on approximation by so derived networks with a fixed (smaller) number of units.

We showed several special types of kernels used widely in practice and their combinations proposed by the author. We added a few illustrations of experiments done on these schemas by Petra Vidnerová. Unfortunately we cannot bring experiments on networks pruned by randomized algorithm as these are currently running.

Chapter 5

Conclusion

In this thesis we addressed many problems of interest when approximating functions by one-hidden-layer feedforward neural network. We derived conditions under which rates of approximation are applicable to this schema extending known results from continuous to $\mathcal{L}^p$ activation functions and proved related results regarding limitations of the presented approach.

Exploiting and modifying probabilistic proof of [DDGS93] we derived probabilistic algorithm that was proven to be theoretically very well suited to prune too rich neural network schemas. We presented analysis of time complexity of the proposed algorithm and compared it to related algorithms for deriving trained neural networks from given data.

We proposed how to apply the probabilistic algorithm to kernel-based neural networks that by definition suffer of too high number of hidden units. This schema is being tested on real data.

The work is based on results published by the author and cooperators ([S08, S04c, S04b, S04a, S03c, S03b, S03a, S02, S01, SS08, KS06, KS05a, KS05b]), the older of the results were published under the author's maiden name Šidlofová.

As grateful as the author is to all who read, re-read and commented on the thesis, she would like to stress that all mistakes therein are solely her own. This is even more so regarding her advisor Věra Kůrková who due to the circumstance of long-term stay abroad of the author could not supervise over the thesis as she would have wished to.

Appendix A

Suggested Future Research

During our work we came across many interesting questions that were unfortunately out of our time possibilities to pursue further on. As we find it a pity these to be forgotten entirely, we give a brief overview of them here in the hope of them appealing to someone better time-equipped:

- **Rates of convergence for sigmoidal functions** In 2.2.9 we have proven equality of theoretically known bounds on the rate of convergence for sigmoidal functions. It is quite possible, that for problems of practical interest, convergence will be faster (but perhaps not for all activation functions). This question deserves further study.
- **Numerical stability** In the end of Section 2.4 we discussed why we consider Theorem 1.4.2 not too relevant for practical applications, basically because we need to use high precision in all computations. It would be interesting to know, to what extend is this alleviated by using methods of Chapter 2. Consequently, we are interested in the numerical stability of the approximations obtained using the Maurey-Jones-Barron theorem.
- **Experiments for Probabilistic algorithm** In Section 3.2 probabilistic algorithm 3.2.2 has been proposed. Experiments testing the algorithm are running in cooperation with Petra Kudová-Vidnerová.
- **Numerical issues** Section 3.3 brings an overview of estimates necessary to assess time complexity of algorithms proposed in Sections 3.1, 3.2 but more detailed study was out of scope of this work and could be of practical interest.
- **Feature maps versus composite kernels** We did not discuss here the topic of feature maps for creating specifically tailored kernels [SchWe06], but this issue deserves further study mainly in connection with composite kernels proposed in Section 4.4.2.

Appendix B

Original Contributions

This thesis builds upon many results of other researchers, and I have found it necessary to provide a somewhat lengthy introduction into the subject to explain the context of the obtained results. As a consequence, it might be difficult at first glance to see, what are the original contributions of this work. For this reason, I prepared the following list of results, that are (to the best of my knowledge) original.

2 Estimates on Rates of Approximation by Neural Networks Using Integral Representations Several results about properties of $\mathcal{G}$-variation; application of Hahn-Banach theorem to estimate $\mathcal{G}$-variation for functions with integral representation in various settings; dependence on the activation function.

- Example 2.1.4: Infinite $\mathcal{G}$-variation
- Theorem 2.1.7: Slow rate of approximation
- Corollary 2.1.8: Slow rate of approximation - lower bound
- Theorem 2.2.3: $\mathcal{G}$-variation in $\mathcal{L}^p$ spaces
- Corollary 2.2.4: Rates in $\mathcal{L}^p$
- Theorem 2.2.6: $\mathcal{G}$-variation for continuous activation functions using measure
- Corollary 2.2.7: $\mathcal{G}$-variation for continuous activation functions (weaker assumptions)
- Theorem 2.2.8: $\mathcal{G}$-variation for $\mathcal{L}^p$ activation functions using measure
- Theorem 2.2.9: $\mathcal{G}_\sigma$-variation independent of σ
- Theorem 2.3.2: Integral representation in $W^{d,p}(\Omega)$

- Theorem 2.3.4: Sum $\Rightarrow$ Integral
- Corollary 2.4.2: Approximation for $C^d(\mathbb{R}^d)$ functions in $\mathcal{L}_p$, gen. sigm. function
- Corollary 2.4.3: Approximation for $\mathcal{W}^{d,p}(\mathbb{R}^d)$ functions in $\mathcal{L}_p$, gen. sigm. function
- Theorem 2.4.4: Rates for absolutely continuous functions
- Theorem 2.4.5: Rates for bounded variation functions
- Theorem 2.4.6: Good rates $\Longrightarrow$ many weak derivatives

3 Algorithmic aspects We suggest a randomized algorithm and compare it to the classical one.

- Algorithm 3.1.3: Iterative from auxiliary approximation
  – the algorithm is based on ideas of Jones (who basically proposed Algorithm Iterative from scratch 3.1.2)
- Theorem 3.2.1: Probabilistic error in Banach spaces (based on [DDGS93])
  – we modified proof of this theorem using Markov inequality, which enabled algorithmic consequences
- Algorithm 3.2.2: Probabilistic from auxiliary approximation
- Corollary 3.2.3: Error of probabilistic algorithm
- Theorem 3.2.4: Probabilistic error in Banach spaces, int. repr.
- Algorithm 3.2.5: Probabilistic from integral representation
- Corollary 3.2.6: Error of integral probabilistic algorithm
- Comparisons of algorithms 3.1.2, 3.1.3, 3.2.2 and 3.2.5 in Section 3.4

4 Pruning Solutions of Specific Regularization Problems We propose use of probabilistic algorithm for kernel-based regularization solutions.

- Theorem 4.3.1: Randomized pruning of kernel-based neural networks
- Remark 4.3.2: Kernel-based networks with feasible number of units
- Composite kernels - we proposed to use sum and product kernels for learning problems and suggested justification for doing so.

Appendix C

Mathematical Background and Symbols

C.1 Notation and Terminology

$\mathbb{N}, \mathbb{Z}, \mathbb{R}$ — Let $\mathbb{N}, \mathbb{Z}, \mathbb{R}$ denote the sets of *natural*, *integer*, and *real numbers*, respectively.

α — Let $\mathbb{Z}_+^n$ denote the set of all nonnegative multi-integers. For *multiindex* $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}_+^n$ we set $|\alpha| = \alpha_1 + \cdots + \alpha_n$ and $x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$.

Γ — *Gamma function* is a generalisation of factorial. Let $z \geq 0$ we put $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, \mathrm{d}t$.

$\lambda_m(A)$ — We will denote the *m-dimensional Lebesgue measure* of a set $A \subset \mathbb{R}^n$ by $\lambda_m(A)$. *Measurability* will be considered with respect to Lebesgue measure in some subset of $\mathbb{R}^n$.

$|\mu|$ — see total variation of measure

$\Omega()$ — See $o(), O(), \Omega()$

θ_{ij} — The *threshold* in the j-th unit of the i-th layer in perceptron or the width of the j-th unit of the i-th layer in RBF (i is again usually omitted).

σ — *activation function*

χ_I — Characteristic function of interval I.

absolutely continuous function — Led X, d be a metric space, $I \subseteq \mathbb{R}$ interval. Then $f : I \to X$ is *absolutely continuous* on I if for every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever a sequence of pairwise disjoint sub-intervals $[a_k, b_k]$ of I, $k = 1, \ldots, n$ satisfies $\sum_{k=1}^n |b_k - a_k| < \delta$ then $\sum_{k=1}^n d(f(b_k) - f(a_k)) < \varepsilon$.

almost everywhere

A property is valid *almost everywhere* if it is valid everywhere except on a set of measure 0.

analytic function

An *analytic function* is a function that is locally given by a convergent power series.

B^n

unit ball in $\mathbb{R}^n$: $B^n = \{x : \|x\|_2 = (x_1^2+,\dots,x_n^2)^{1/2} \leq 1\}$.

$B_r(\|.\|)$

Let X be a normed space. We denote by $B_r(\|.\|)$ the ball of radius r in norm $\|.\|$ i.e. $B_r(\|.\|) = \{x \in X; \|x\| \leq r\}$.

balanced set

A set X is *balanced* if $h \in X$ and $|a| \leq 1$ implies $ah \in X$.

Banach space

Banach space $(B, \|.\|)$ is any normed linear space that is complete in its norm.

bounded variation function

Let f be a real valued function on interval $[a, b] \subseteq \mathbb{R}$. Then f has *bounded variation* on $[a, b]$ if total variation of f on $[a, b]$ is finite. We define *total variation* of f on $[a, b]$ by $\sup_{(P \text{ partition of } [a,b])} \sum_{i=0}^{n_p-1} |f(x_{i+1}) - f(x_i)|$.

$C(X)$

For a topological space X, $C(X)$ denotes the set of all *continuous real-valued functions* on X and $\|.\|_C$ denotes the *supremum norm*

$$\|f\|_C = \sup_{x \in X} |f(x)|.$$

$C^k(X)$

For a topological space X, $k \in \mathbb{N} \cup \infty$, space $C^k(X)$ denotes the set of all *continuous real-valued functions with continuous derivatives up to order* k on X. For $k = \infty$ we call the functions *smooth*.

$C_C^\infty(X)$

$C_C^\infty(X)$ is space of infinitely differentiable functions with compact support in X.

$C^*(X)$

Let $X \subseteq \mathbb{R}^d$ compact. Then

$$\begin{aligned} C^*(X) = \{\ell : & \text{ exists signed Borel measure } \mu \text{ on } X \text{ with compact} \\ & \text{support such that } |\mu\| < \infty \text{ and} \\ & \ell(h) = \textstyle\int_X h \,\mathrm{d}\mu \text{ for every } h \in C(X)\} \\ & . \end{aligned} \tag{C.1}$$

$c\mathcal{G}$

Let $\mathcal{G}$ be a set and $c \in \mathbb{R}$. Then $c\mathcal{G} = \{cg : g \in \mathcal{G}\}$.

$\mathrm{cl}_{\|.\|} S$

$\|\cdot\|$-*closure* of a subset S of a normed space $X_{\|.\|}$ contains all limits (in the norm $\|.\|$) of elements of S (smallest closed subset containing S). Closure of the convex hull is called the *convex closure*.

condition number

Condition number of a matrix is the ratio of the largest and the smallest eigenvalue.

conv, conv_n

Convex hull of a set S is the set of all convex combinations of elements of S, i.e. $\mathrm{conv}\, S = \{\sum_{i=1}^{n} c_i s_i : n \in \mathbb{N}^+, s_i \in S, c_i \in \mathbb{R}^+, \sum_{i=1}^{n} c_i = 1\}$. Analogously we define convex hull of at most n elements of S: $\mathrm{conv}_n\, S = \{\sum_{i=1}^{n} c_i s_i : s_i \in S, c_i \in \mathbb{R}^+, \sum_{i=1}^{n} c_i = 1\}$.

convex funcional

A functional $\mathcal{F}$ is *convex* on a convex set $E \subseteq \mathrm{dom}\,\mathcal{F}$ if for all $f, g \in E$ and all $\lambda \in [0,1]$, $\mathcal{F}(\lambda f + (1-\lambda)g) \leq \lambda\mathcal{F}(f) + (1-\lambda)\mathcal{F}(g)$.

$D^\alpha(f)$

For $\alpha = (\alpha_1, \ldots, \alpha_d)$ multiindex we define *mixed partial derivative* $D^\alpha(f) = \frac{\partial^\alpha}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(f)$.

D_e

D_e is the operator of *directional derivative* in the direction given by e, that is $D_e f(y) = \lim_{h \to 0} \frac{f(y+h \cdot e) - f(y)}{h}$. For a positive integer k, $D_e^{(k)}$ is k-fold iteration of D_e.

dual space

Let X be a normed linear space. Then the space of bounded linear mappings from X to $\mathbb{R}$ is called *dual space* to X and denoted by X^*.

$E(\xi)$

expectation or *mean value* of random variable ξ

$$E(\xi) = \sum_{i=1}^{k} x_i P_\xi(x_i)$$

for finite probability space or

$$E(\xi) = \int_\Omega \xi \, \mathrm{d}P$$

for general probability space.

ess sup

Let (X, Σ, μ) be a measure space and $f : X \to \mathbb{R}$. Then *essential supremum* of f is defined as

$$\mathrm{ess\,sup}(f) = \inf\{a \in \mathbb{R} : \mu\{x : f(x) > a\} = 0\}.$$

Fourier transform

Let $f \in \mathcal{L}^1(\mathbb{R}^d)$. Then *Fourier transform* of f is defined as $\hat{f}(\omega) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} f(x) e^{-i(\omega \cdot x)} \, \mathrm{d}x$.

$\mathcal{G}$-variation

Let $\mathcal{G}$ be bounded subset of a Banach space X of functions. For convenience, we will assume that $g \in \mathcal{G}$ implies $-g \in \mathcal{G}$. Let $f \in X$. Then $\mathcal{G}$-*variation* of f is defined as $\|f\|_{\mathcal{G}} = \inf\{c > 0 : f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}\}$.

H_{eb}

Let $e \in \mathbb{R}^d$ and $b \in \mathbb{R}$ then $H_{eb} = \{y \in \mathbb{R}^d : y \cdot e + b = 0\}$ is a *hyperplane* in $\mathbb{R}^d$ given by e, b.

Hilbert space

Hilbert space is a Banach space in which the norm is given by an inner product $\langle .,. \rangle$, that is $\|x\| = \langle x, x \rangle^{1/2}$.

$I(c, \mathcal{G})$

Denotes the set of functions f that can be represented as $f(x) = \int_A \varphi(x,a)\,\mathrm{d}\nu(a)$, where ν is signed measure on A such that $\|\nu\| \leq c$, $\varphi \in \mathcal{G}$.

$\mathrm{d}m_d(x)$

See normalised Lebesgue measure.

$\mathcal{L}^p$

For $p \in [1, \infty]$ and X a space with measure μ, $\mathcal{L}^p(X)$ denotes the space of functions with finite $\mathcal{L}^p$-*norm*, which is defined by

$$\|f\|_p = \begin{cases} \sqrt[p]{\int_X |f(x)|^p \,\mathrm{d}\mu(x)} & 1 \leq p < \infty \\ \operatorname{ess\,sup}_{x \in X} |f(x)| & p = \infty \end{cases}$$

Note that for $\mathcal{L}^p$ to be a linear space we have to identify functions that are equal almost everywhere.

$\mathcal{L}^p_{loc}$

For $p \in [1, \infty]$ and X a measure space, $\mathcal{L}^p_{loc}(X)$ denotes the space of functions with finite $\mathcal{L}^p$-*norm* on every compact subset of X. The functions are called *locally integrable*.

$(\mathcal{L}^p)^*$

Let $p, q \geq 1$ be conjugated coefficients (i.e., $1/p + 1/q = 1$). Then

$$(\mathcal{L}^p)^* = \{\ell : \text{exists } k \in \mathcal{L}^q \text{ such that } \ell(h) = \textstyle\int kh\}. \tag{C.2}$$

m_d

see measure, normalised Lebesgue

measure, Borel

Measure μ on space X is *Borel* if all open subsets of X are μ-measurable.

measure, density

Let μ, ν be σ-finite measures such that whenever $\mu E = 0$ then also $\nu E = 0$. Then there exists a nonnegative function $f \in \mathcal{L}^1(\mu)$ so that

$$\nu(E) = \int_E f \,\mathrm{d}\mu,$$

for any E a μ-measurable set. Function f is called *density* of ν with respect to μ.

measure, Hahn decomposition

(P, N) is a *Hahn decomposition* of A for the measure μ if A is the disjoint union of P and N, and $\mu(E) \geq 0$ (resp. ≤ 0) whenever $E \subseteq P$ (resp. $E \subseteq N$).

measure, Lebesgue

For an arbitrary set $A \subset \mathbb{R}^n$ we define *outer Lebesgue measure* as

$$\lambda_n^* A = \inf\{\sum_{i=1}^{\infty} vol I_k : \bigcup_{i=1}^{\infty} I_k \supset A,\ I_k \text{open interval}\},$$

where interval $I \subset \mathbb{R}^n$ is a Cartesian product of n one-dimensional intervals $I = (b_1 - a_1) \times \cdots \times (b_n - a_n)$ and $vol I = (b_1 - a_1) \cdot \cdots \cdot (b_n - a_n)$. We say that a set $A \subset \mathbb{R}^n$ is *Lebesgue measurable* if $\lambda^* I = \lambda^*(A \cap I) + (A \setminus A)$ for every interval $I \subset \mathbb{R}^n$. On Lebesgue measurable functions outer measure λ^* is denoted by λ and called the *Lebesgue measure*.

measure, normalised Lebesgue [Ru91]

Following [Ru91] we define *normalized Lebesgue measure* m_d on $\mathbb{R}^d$ as $\mathrm{d}m_d(x) = (2\pi)^{-d/2}\,\mathrm{d}\lambda(x)$.

measure, norm

Let (P,N) be Hahn decomposition for measure μ. Define the function s by

$$s(a) = \begin{cases} +1 & \text{for } a \in P \\ -1 & \text{for } a \in N \end{cases}$$

that is, $s(a)$ is the "sign of μ at a". Then we have *norm of* μ $\|\mu\| = \int_A s(a)\,\mathrm{d}\mu(a)$.

measure, Radon

Let X be a locally compact space. We say that μ is *Radon measure* on X if Borel σ-algebra on X is μ-measurable, $\mu K < \infty$ for every compact $K \subset X$, $\mu G = \sup\{\mu K : K \subset G, K\text{compact}\}$ for every open $G \subset X$ and $\mu A = \inf\{\mu G : G \supset A, G\text{open}\}$ for every μ-measurable A.

measure, σ-finite

Measure μ on space X is called *σ-finite* if there exist sets $M_n \subseteq X$ such that $\mu(M_n) < \infty$ and $X = \bigcup_{n=1}^{\infty} M_n$.

measure, signed

Let $\mathcal{S}$ be σ-algebra of subsets of X. Set function $\mu : \mathcal{S} \to \mathbb{R}$ is a *signed measure* on X if $\mu\emptyset = 0$ and $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$ whenever A_n are pairwise disjoint.

measure, total variation

Let μ be signed measure on space X. Then $|\mu|$ denotes the *total variation* of μ and for a set $E \in X$ it holds that

$$|\mu|(E) = \mu^+ E + \mu^- E,$$

where all of $|\mu|$, μ^+, μ^- are positive measures.

modulus of smoothness

A Banach space X has *modulus of smoothness* ϱ if $\varrho : [0, \infty) \to [0, \infty)$ is a function given by

$$\varrho(r) := \sup_{\|f\|=\|g\|=1} \Big(\frac{\|f + rg\| + \|f - rg\|}{2} - 1 \Big).$$

normal space

Topological space X is *normal* if for any two disjoint closed subsets F_i of X there are neighbourhoods U_i of F_i that are also disjoint.

normed linear space

A *normed linear space* W is any vector space over $\mathbb{R}$ or $\mathbb{C}$ with a *norm* $\|.\|$, where for all $x, y \in W$, $\lambda \in \mathbb{R}$ (or $\mathbb{C}$).

1. $\|x\| \geq 0$ and $\|x\| = 0$ only if $x = 0$
2. $\|\lambda x\| = |\lambda| \|x\|$, and
3. $\|x + y\| \leq \|x\| + \|y\|$.

$o(), O(), \Omega()$

This notation describes *asymptotic behaviour*: Let $f(x), g(x)$ be real functions, $a \in \mathbb{R} \cup \{\infty\}$.

- $f(x)$ is $O(g(x))$ as $x \to a$ if and only if $\limsup_{x \to a} \left| \frac{f(x)}{g(x)} \right| < \infty$,
- $f(x)$ is $o(g(x))$ as $x \to a$ if and only if $\lim_{x \to a} \left| \frac{f(x)}{g(x)} \right| = 0$,
- $f(x)$ is $\Omega(g(x))$ as $x \to a$ if and only if $g(x)$ is $O(f(x))$.

peak functional

F is the *peak functional* for f if it is a linear continuous functional of unit norm for which $F(f) = \|f\|$.

positive definite function

Let B be a Banach space, $\Omega \subset B$ and let $f : \Omega \times \Omega \to \mathbb{R}$ be a symmetric function (that is $f(x, y) = f(y, x)$). Then f is *positive definite* if for any $a_1, \ldots, a_n \in \mathbb{C}$ and $t_1, \ldots, t_n \in \Omega$

$$\sum_{i,j=1}^{n} \overline{a}_i a_j f(t_i, t_j) \geq 0,$$

where $\overline{a}$ is complex adjoint of a.

quasiconvex, strictly quasiconvex

Let $E \subseteq \operatorname{dom} \mathcal{F}$. Functional $\mathcal{F}$ is *quasiconvex* if for all $f, g \in E$ and all $\lambda \in [0, 1]$ we have $\mathcal{F}(\lambda f + (1 - \lambda) g) \leq \max\{\mathcal{F}(f), \mathcal{F}(g)\}$. If the inequality is strict for all $f \neq g$ and $\lambda \in (0, 1)$, we say $\mathcal{F}$ is *strictly quasiconvex*.

$\mathcal{R}_{\mathcal{W}}^{rad}$

see Radial functions

$\mathcal{R}$ see Ridge functions

Rademacher sequence A *Rademacher sequence* $\{\varepsilon_i\}_{i=1}^n$ is a finite sequence of independent zero mean random variables taking values from $\{-1, 1\}$.

Rademacher type space Let X be Banach space. Space X is of *Rademacher type* t (with constant C) if for any Rademacher sequence $\{\varepsilon_i\}_{i=1}^n$ and any fixed finite sequence $\{f_i\}_{i=1}^n$ of elements of X it holds that

$$E\|\sum \varepsilon_i f_i\|^t \leq C \sum \|f_i\|^t.$$

Radial functions *Radial functions* are defined as follows: $\mathcal{R}_{\mathcal{W}}^{rad} = \mathrm{span}\{g(\|x - w\|^2), w \in \mathcal{W} \subseteq \mathbb{R}^d, g \in C(\mathbb{R})\}$.

Radon random variable Random variable ξ with values in a Banach space B is called *Radon random variable* if for each $\varepsilon > 0$ there is a compact set $K = K(\varepsilon)$ in B such that $\Pr\{\xi \in K\} \geq 1 - \varepsilon$.

reflexive space Banach space X is reflexive is second dual $(X^*)^*$ is again the original space X, i.e. $(X^*)^* = X$.

Ridge functions *Ridge functions* are defined as follows:

$$\mathcal{R} = \{\sum_{i=1}^{n} g_i(a^i \cdot x) : n \in \mathbb{N}, a^i \in \mathbb{R}^d, g_i \in C(\mathbb{R})\}.$$

By $\mathcal{R}_n$ we denote linear combinations of up to n functions.

S^{n-1} *unit sphere* in $\mathbb{R}^n$: $S^{n-1} = \{x : \|x\|_2 = (x_1^2 + \dots x_n^2)^{1/2} = 1\}$.

$s_{gen}(w)$ $s_{gen}(w)$ denotes the number of steps needed to generate one a with probability density $|w(a)|$.

s_{opt} $s_{opt} = s_{opt}(\varepsilon)$ denotes number of steps necessary for finding "ε-precision" minimum of function by gradient method (see 3.10).

Schwartz space We say that a C^∞ function on $\mathbb{R}^d$ belongs to the *Schwartz space* in $pD^\alpha f$ is a bounded function for any multiindex α and any polynomial p on $\mathbb{R}^d$.

separable space A topological space X is *separable* if it contains a countable dense subset.

sigmoidal function Function $\sigma : \mathbb{R} \to \langle 0, 1\rangle$ is *sigmoidal* if $\lim_{x\to-\infty} \sigma(x) = 0$, $\lim_{x\to\infty} \sigma(x) = 1$ and σ is nondecreasing.

span, span_n

finite linear combination of elements: Let S be a set (finite or infinite), then $\operatorname{span} S = \{\sum_{i=1}^{n} c_i s_i : s_i \in S, c_i \in \mathbb{R}, n \in \mathbb{N}\}$ and linear combination of at most n elements is denoted by $\operatorname{span}_n S = \{\sum_{i=1}^{n} c_i s_i; c_i \in \mathbb{R}, s_i \in S\}$.

strictly positive definite function

We call the function f *strictly positive definite* if it holds

$$\sum_{i,j=1}^{n} \overline{a}_i a_j f(t_i, t_j) \geq 0$$

and

$$\sum_{i,j=1}^{n} \overline{a}_i a_j f(t_i, t_j) = 0 \implies a_i = a_j = 0, i, j = 1, \ldots, N.$$

$\operatorname{supp} \mu$

Let μ be Radon measure on a space X. Then *support of* μ is

$$\operatorname{supp} \mu = X \setminus \bigcup \{G : G \text{ open}, \mu(G) = 0\},$$

i.e. the smallest closed set whose complement is of zero measure.

$\operatorname{supp}(f)$

Let $(X, \|.\|)$ be a normed space. For a function $f : X \to \mathbb{R}$ the *support* of f is defined by $\operatorname{supp}(f) = \mathrm{cl}_{\|.\|}\{x \in X; f(x) \neq 0\}$. If $\operatorname{supp}(f)$ is compact we call function f to be *compactly supported.*

$t_{apx}, t_{int}, t_{opt}$

$t_{apx}(\mathcal{G}, \varepsilon_{apx}, N)$ denotes time needed to find ε_{apx}-close approximation to f using at most N functions of $\mathcal{G}$. (Note that ε_{apx} and N are interconnected.)

By t_{int} we denote the time needed to enumerate a (d-dimensional) integral.

$t_{opt}(\mathcal{G}, \delta_k)$ denotes the time needed to find $g_k \in \mathcal{G}$ such that $F_k(g_k - f) \leq \delta_k$, where F_k is the peak functional for $f_k - f$ (see Theorem 3.1.1).

unbiased estimator

Let ξ be random variable. Then $\tilde{\xi}$ is and *unbiased estimator* if expectation $E(\tilde{\xi} - \xi) = 0$.

uniformly smooth

A Banach space is termed *uniformly smooth* if $\varrho(r) = o(r)$ as $r \to 0$, where ϱ is the modulus of smoothness.

variance

variance of random variable ξ

$$var(\xi) = E(\xi - E(\xi))^2.$$

w_{ijk} — *weight* of connection between k-th neuron in $i-1$-st layer and j-th neuron in i-th layer in perceptron or
k-th coordinate of centre of j-th unit in i-th layer in RBF (here i is usually omitted as it is 1).

$\mathcal{W}^{m,p}$ — *Sobolev space* $\mathcal{W}^{m,p}$: Let X be an open set in $\mathbb{R}^n$, let m be a natural number and let $1 \leq p \leq \infty$. The Sobolev space $\mathcal{W}^{m,p}(X)$ is defined to be completion of the set of all $C^\infty(X)$ functions defined on X such that for every multi-index α the mixed partial derivative $D^\alpha(f)$ is locally integrable and in $\mathcal{L}^p(X)$ for $|\alpha| \leq m$. Norm on this Sobolev space is defined as follows:

$$\|f\|_{m,p} = \begin{cases} \left(\sum_{0\leq\alpha\leq m} \|D^\alpha f\|_p^p\right)^{1/p}, & 1 \leq p < \infty \\ \max_{0\leq\alpha\leq m} \|D^\alpha f\|_\infty, & p = \infty \end{cases},$$

we use weak derivative where necessary.

w^* convergence — Let $\{x_n\}$ be a sequence in X^* a dual of some Banach space X. Then $\{x_n\}$ *w^*-converges* to $x \in X^*$, $x_n \overset{w^*}{\to} x$, if for any $t \in X$ it holds that $x_n(t) \to x(t)$.

weak convergence — Let $\{x_n\}$ be a sequence in X. Then $\{x_n\}$ *converges weakly* to $x \in X$, $x_n \rightharpoonup x$ (or $x_n \overset{w}{\to} x$), if for any $g \in X^*$ from dual space it holds that $g(x_n) \to x$.

weak derivative — Let $u, v \in \mathcal{L}^1_{loc}(U)$, where $U \subseteq \mathbb{R}$ is open. For a multiindex α we say that v is *α-th weak derivative* of u if for each $\xi \in C^\infty_C(U)$ we have

$$\int_U u D^\alpha \xi = (-1)^{|\alpha|} \int_U v\xi\,.$$

weakly sequentially compact set — Set E is *weakly sequentially compact* if any sequence in E has a weakly converging subsequence.

weakly sequentially lower semicontinuous — Functional $\mathcal{F}$ is *weakly sequentially lower semicontinuous* if and only if $f_n \rightharpoonup f$ implies $\mathcal{F}(f) \leq \liminf_{n\to\infty} \mathcal{F}(f_n)$, where $f_n \rightharpoonup f$ stands for weak convergence.

x_{ij} — In context of neural networks x_{ij} is *value* in the j-th unit of the i-th layer.

C.2 Relevant Mathematical Theorems

Theorem C.2.1 (–Hoeffding bound) *Let ξ be a random variable on a probability space X with mean $E(\xi) = \mu$ and variance $\sigma^2(\xi) = \sigma^2$. If $|\xi(x) - E(\xi)| \leq M$ for almost all $x \in X$*

then for all $\varepsilon > 0$

$$\Pr_{x\in X^n}[|\frac{1}{n}\sum_{i=1}^{n}\xi(x_i) - \mu| \geq \varepsilon] \leq 2e^{-\frac{n\varepsilon^2}{2M^2}}$$

Theorem C.2.2 (Fubini's Theorem) *Let* (X,μ), (Y,ν) *be* σ*-finite measure spaces (there exist sets* $\{M_i\}$ *such that* $\mu(M_i) < \infty$ *and* $X = \bigcup_{i=1}^{\infty} M_i$*, (the same for* Y*) and* f *measurable function on* $X \times Y$*. If* $\int_{X\times Y} f$ *exists (*f *nonnegative or integrable), then*

$$\int_{X\times Y} f(x,y) = \int_X \left(\int_Y f(x,y)\,\mathrm{d}\nu(y)\right)\,\mathrm{d}\mu(x).$$

Theorem C.2.3 (Geometric Hahn-Banach) *Let* X *be a Banach space, consider* $x \in X$ *and* $T \subseteq X$*. Then* $x \in \operatorname{cl}\operatorname{conv} T$*, unless there is a functional* $\ell \in X^*$ *and* $z \in \mathbb{R}$ *such that*

$$\ell(x) > z \quad \text{and} \quad \ell(t) < z \text{ for every } t \in T\,. \tag{C.3}$$

Theorem C.2.4 (Helly's Theorem) *Let* X *be separable Banach space,* X^* *its dual space. Then the closed unit ball in* X^* *is* w^**-sequentially compact.*

Theorem C.2.5 (Hölder's inequality) *Let* (X,μ) *be a measure space. Let* $f \in \mathcal{L}^p(X)$ *and* $g \in \mathcal{L}^q(X)$*, where* $p, q \in (1,\infty)$, $\frac{1}{p} + \frac{1}{q} = 1$*. Then* $fg \in \mathcal{L}^1(X)$ *and*

$$\left|\int_X fg\,\mathrm{d}\mu\right| \leq \left(\int_X |f|^p\,\mathrm{d}\mu\right)^{1/p}\left(\int_X |g|^q\,\mathrm{d}\mu\right)^{1/q}.$$

Theorem C.2.6 (Jensen's Inequality) *Let* (X,μ) *be finite measure space (let* $\mu(X) = 1$*). For* f *real-valued integrable function on* X *and* ψ *measurable convex function on* $\mathbb{R}$ *then*

$$\psi\int_X f\,\mathrm{d}\mu \leq \int_X \psi(f)\,\mathrm{d}\mu$$

For the case of probability space, ξ *inegrable real-valued random variable we have:*

$$\psi(E(\xi)) \leq E(\psi(\xi)).$$

Theorem C.2.7 (Luzin) *Let* (P,μ) *be a locally compact space,* μ *a complete Radon measure, suppose* P *is* σ*-finite (for example the Lebesgue measure on* $\mathbb{R}^n$ *satisfies these assumptions). Let* f *be a measurable function on* P*.*

Then for any $\varepsilon > 0$ *exists a continuous function* $\widetilde{f}$ *on* P *and an open set* E *such that* $\mu E < \varepsilon$ *and*

$$\widetilde{f} = f \text{on } P \setminus E.$$

Furthermore, we may choose $\widetilde{f}$ *so that*

$$\sup_{x\in P\setminus E} |f(x)| = \sup_{x\in P\setminus E} |\widetilde{f}(x)|$$

Theorem C.2.8 (Markov–Chebyshev Inequality) *Let (X, μ) be measure space, f measurable real-valued function and $t, r > 0$. Then*

$$\mu(\{x \in X : |f(x)| \geq t\}) \leq \frac{1}{t^r} \int_X |f(x)|^r \, \mathrm{d}\mu.$$

For the case of probability space, ξ random variable we have:

$$P(|\xi| \geq t) \leq \frac{E(|\xi|^r)}{t^r}.$$

Theorem C.2.9 (Representer Theorem – basic version [KiWa71, SchSl02]) *Let $z = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}$ be given, let X be an RKHS with kernel k and norm $\|\cdot\|_k$. Let the functional $\mathcal{F}(f)$ be given by*

$$\mathcal{F}(f) = \frac{1}{N} \sum_{i=1}^N (f(x_i) - y_i)^2 + \gamma \|f\|_k^2$$

for a function $f \in X$ ($\gamma > 0$ is a constant).

Then there exists a unique function $f_0 \in X$ that minimizes $\mathcal{F}$. Moreover, f_0 is of form

$$f_0(x) = \sum_{i=1}^N c_i k(x, x_i)\,, \tag{C.4}$$

for some real c_i's.

Proof: We will use several basic results in approximation theory, see e.g. [Dn71, p. 7–15]:

Lemma C.2.10 (Existence of minimizing function [Dn71]) *Let $\mathcal{F}$ be a weakly sequentially lower semicontinuous functional defined on a weakly sequentially compact set E. Then $\mathcal{F}$ attains its minimum: there is $f_0 \in E$ such that $\mathcal{F}(f_0) = \inf_{f \in E} \mathcal{F}(f)$.*

Lemma C.2.11 (Uniqueness of minimizing function [Dn71]) *A strictly quasiconvex functional $\mathcal{F}$ can attain minimum over a convex set C at no more than one point.*

Lemma C.2.12 (Necessary condition for minimum [Functional analysis]) *Let the functional $\mathcal{F}$ defined on a set E in a Banach space X be minimized at a point $f_0 \in E$, with f_0 an interior point in the norm topology. For any $h \in X$, if $\mathcal{F}$ has a derivative $\mathrm{D}_h\mathcal{F}(f_0)$ at f_0 in direction h, then $\mathrm{D}_h\mathcal{F}(f_0) = 0$.*

Now we can prove easily existence, uniqueness and form of solution to the kernel-based minimization problem (4.4). Once more we would like to stress that this is not our result nor is it in any sense new (see for example [SchWe06, PoSm03, SchSl02, CuSm01, Gi98, Wa90] and others). The reason for stating it including the proof is that the ideas are simple and we consider them illuminating for understanding the role of kernel in learning and we will also need the theorem in our considerations (Theorem 4.3.1).

We start by proving that $\mathcal{F}$ attains minimum and to achieve this we will use Lemma C.2.10. Pick any $g \in X$. Certainly, any f for which $\mathcal{F}(f) \leq \mathcal{F}(g)$ needs to have $\|f\|_k \leq \sqrt{\mathcal{F}(g)/\gamma}$, so we only need to minimize $\mathcal{F}$ over the ball in X with radius $\sqrt{\mathcal{F}(g)/\gamma}$. As X is a Hilbert space, this ball is weakly compact (see, e.g., Theorem 7 in [Lax02]), thus also weakly sequentially compact. Thus all we need to do to assure existence of minimum is to prove that $\mathcal{F}$ is weakly sequentially lower semicontinuous. (Perhaps confusingly, weak continuity is stronger property than continuity, so although $\mathcal{F}$ is continuous, that is not enough.)

So let us consider a sequence of functions f_n weakly converging to some f. From the basic properties of RKHS's (Theorem 4.1.2) we know that the evaluation functionals $\Lambda_i : f \mapsto f(x_i)$ are continuous (and, obviously, linear). Thus, by definition of weak convergence, $\lim_{n\to\infty} \Lambda_i(f_n) = \Lambda_i(f)$. It follows, that the data part of the functional $\mathcal{F}$ is weakly sequentially lower semicontinuous.

The second part of $\mathcal{F}$ is a square of a norm. A norm in any Banach space is weakly sequentially lower semicontinuous (see, e.g., Theorem 5 in [Lax02]). It easily follows that so is the square of the norm and thus $\mathcal{F}$ itself. Consequently, $\mathcal{F}$ attains its minimum.

To show uniqueness of the minimum, we will use Lemma C.2.11. The functional $\mathcal{F}$ is actually strictly convex, but verifying strict quasiconvexity is somewhat easier.

The first part of $\mathcal{F}$ is a sum of N elements, each of which is a convex functional, as the (real) function $z \mapsto \frac{1}{N}(z - y_i)^2$ is convex.

For the second part, let f, g be two distinct elements of X, $s \in (0,1)$ and $t = 1 - s$. Put $M = \max\{\|f\|_k^2, \|g\|_k^2\}$. We have

$$\begin{aligned}\langle sf + tg, sf + tg\rangle &= s^2\langle f, f\rangle + st(\langle f, g\rangle + \langle g, f\rangle) + t^2\langle g, g\rangle \\ &\leq M^2(s^2 + 2st + t^2) \\ &= M^2\end{aligned}$$

(we have used Cauchy inequality to estimate $\langle f, g\rangle$ and $\langle g, f\rangle$). This proves quasiconvexity of the mapping $f \mapsto \|f\|_k^2$, to obtain strict quasiconvexity, we analyze when equality is achieved in the above formula. We need to have $\|f\|_k = \|g\|_k$, moreover (to have equality in the Cauchy inequality) f and g are collinear. This implies $f = \pm g$; $f = -g$ does not produce equality and $f = g$ is false by assumption.

Altogether, $\mathcal{F}$ is a sum of a convex functional and a strictly quasiconvex functional, so it is strictly quasiconvex itself. Thus we may apply Lemma C.2.11 (with $C = X$).

So we have shown that there exists a unique f_0 at which the minimum of $\mathcal{F}$ is attained. As a final step we derive the form of f_0. Lemma C.2.12 implies that for each h we have $\mathrm{D}_h\mathcal{F}(f_0) = 0$. A routine computation yields

$$\mathrm{D}_h\mathcal{F}(f) = \frac{1}{N}\sum_{i=1}^{N} 2(f(x_i) - y_i)h(x_i) + \gamma(\langle f, h\rangle + \langle h, f\rangle)\,.$$

We put $h(y) = k(y, x)$. The reproducing property yields $\langle f_0, h\rangle = f_0(x)$, and so $\langle h, f_0\rangle = \overline{f_0(x)} = f_0(x)$ (as f_0 was assumed to be real). Putting $c_i = -(f_0(x_i) - y_i)/(\gamma N)$ we obtaine the desired form (recall that k is symmetric)

$$f_0(x) = \sum_{i=1}^{N} c_i k(x, x_i)\,.$$

And we easily obtain the form of c_i: $c = (K[x] + \gamma N I)^{-1}y$, where $K[x]_{i,j} = k(x_i, x_j)$ is Gramm matrix of kernel k with respect to vector x, I is $N \times N$ identity matrix, c, y are vectors. □

Theorem C.2.13 (Tieze's extension) *If f is a continuous function on closed subset F of a normal topological space* [1] *X, then there exists a continuous function $\tilde{f}$ on X such that*

$$f = \tilde{f} \text{ on } F \text{ and } \sup_X |\tilde{f}| = \sup_F |f|$$

[1] Note that any metric space is normal.

Appendix D

Pictures

Due to its theoretical nature, this work contains only few pictures. In this appendix we try to make up for it a bit. Below are neural networks, as depicted by my sons
Toníček (on the left) and Prokůpek

Bibliography

[Ar50] N. Aronszajn: *Theory of Reproducing Kernels*, Transactions of the AMS, **68** (1950), no. 3, 337–404.

[Ba93] A. R. Barron: *Universal approximation bounds for superposition of a sigmoidal function*, IEEE Transactions on Information Theory, **39** (1993), 930–945.

[Ba92] A. R. Barron: *Neural net approximation*, Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems, (K. S. Narendra, Ed.), (1992), 69–72.

[Bl98] Ch. Blatter: *Wavelets: a primer*, A K Peters, (1998).

[BoVa04] S. Boyd, L. Vandenberghe: *Convex Optimization*, Cambridge University Press, (2004).

[ChLM94] C.K. Chui, X. Li, H.N. Mhaskar: *Neural networks for localized approximation*, Mathematics of Computation, **63** (1994), 607–623.

[CuSm01] F. Cucker, S. Smale: *On the Mathematical Foundations of Learning*. Bulletin of the American Mathematical Society **39** (2001), 1–49.

[Dn71] J. W. Daniel: *The Approximate Minimization of Functionals*, Prentice-Hall, (1971).

[DDGS93] C. Darken, M. Donahue, L. Gurvits, E. Sontag: *Rate of Approximation Results Motivated by Robust Neural Network Learning*, Proceedings of the 6th Annual ACM Conference on Computational Learning Theory, Santa Cruz, CA, (1993), 303–309.

[DRCDO05] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone: *Learning from Examples as an Inverse Problem*, MIT Press, The Journal of Machine Learning Research, **6** (2005), 883–904.

[DHM89] R. De Vore, R. Howard, C. Micchelli: *Optimal non-linear approximation*, Manuskripta Math. **63** (1989), 469–478.

[Gi98] F. Girosi: *An Equivalence between Sparse Approximation and Support Vector Machines*, Neural Computation **10** (1998), 1455–1480 (A.I. Memo No. 1606, MIT, 1997).

[GiAn93] F. Girosi, G. Anzellotti: *Rates of convergence for radial basis functions and neural networks*, Artificial Neural Networks for Speech and Vision, p.169–176, Chapman and Hall, (1993) (A.I. Memo No. 1288, MIT, 1995).

[GJP95] F. Girosi, M. Jones, T. Poggio: *Regularization Theory and Neural Networks Architectures*, Neural Computation, **7** (1995), 219–269.

[HaBu88] S.J. Hanson, D.J. Burr: *Mikowski-r back-propagation: learning in connectionist models with non-Euclidean error signals*, Neural Information Processing Systems, New York: American Institute of Physics, (1988), p. 348.

[Ha94] S. Haykin: *Neural Networks*, MacMillan, New York, (1994).

[He99] S. Helgason: *The Radon Transform*, Progrees in Mathematics, Birkhauser, (1999).

[Ho92] K. Hornik: *Approximation Capabilities of Multilayer Feedforward Networks*, Neural Networks, **4** (1991), 251–257.

[HSW89] K. Hornik, M. Stinchcombe, H. White: *Multilayer feedforward networks are universal approximators*, Neural Networks **2** (1989), 359–366.

[Hu85] P.J. Huber: *Projection Pursuit*, Annals of Statistics, **13** (1985), 435–475.

[Ito91] Y. Ito: *Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory*, Neural Networks **4** (1991), no. 3, 385–394.

[Jo92] L. K. Jones: *A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training*, Annals of Statistics, **20** (1992), no. 1., 608–613.

[Jo87] L. K. Jones: *On a conjecture of Huber concerning the convergence of projection pursuit regression*, Annals of Statistics **15** (1987), 880-882.

[Js03] J. Jost: *Postmodern Analysis*, Springer Verlag, (2003).

[KiWa71] G. Kimeldorf, G. Wahba: *Some results on Tchebycheffian spline functions*, J. Math. Anal. Applic., **33** (1971), 82–95.

[K06] P. Kudová: *Learning with Regularization Networks*, PhD. Thesis, MFF UK, Prague, (2006).

[K04] P. Kudová: *Kernel Based Regularization Networks and RBF Networks*, Doktorandský Den 2004, Paseky nad Jizerou, (2004).

[KS06] P. Kudová, T. Šámalová: *Sum and Product Kernel Regularization Networks*, Lecture Notes in Artificial Intelligence, **4029** (2006), 56–65.

[KS05b] P. Kudová, T. Šámalová: *Product Kernel Regularization Networks*, Icannga 2005, Coimbra, (2005).

[KS05a] P. Kudová, T. Šidlofová:: *Sum and Product Kernel Regularization Networks*, ICS AS CR, Technical Report, V-935, (2005).

[Ku04] V. Kůrková: *Learning from Data as an Inverse Problem*, proceedings of COMPSTAT 2004, Prague (CZ), Computational Statistics, Physica Verlag, (2004), 1377–1384.

[Ku03] V. Kůrková: *High-dimensional approximation and optimization by neural networks*, Chapter 4 in Advances in Learning Theory: Methods, Models and Applications, (Eds. J. Suykens et al.), IOS Press, Amsterdam, (2003), 69–88.

[Ku02] V. Kůrková: *Universality and Complexity of Approximation of Multivariable Functions by Feedforward Networks*, Softcomputing and Industry, Springer, (2002), 13–24.

[Ku97] V. Kůrková: *Dimension-independent rates of approximation by neural networks*, in Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, (K. Warwick and M. Kárný, eds.), Birkhäuser, Boston, (1997), pp. 261–270.

[KHS98] V. Kůrková, K. Hlaváčková, P. Savický: *Representations and rates of approximation of real-valued Boolean functions by neural networks*, Neural Networks **11** (1998), no. 4, 651–659.

[KKK97] V. Kůrková, P.C. Kainen, V. Kreinovich: *Estimates of the Number of Hidden Units and Variation with Respect to Half-Spaces*, Neural Networks, **10** (1997), 1061–1068.

[KKV07] V. Kůrková, P.C. Kainen, A. Vogt: *A Sobolev-Type Upper Bound for Rates of Approximation by Linear Combinations of Heaviside Plane Waves*, Journal of Approximation Theory, **147** (2007), no. 1, 1–10.

[KKV06] V. Kůrková, P.C. Kainen, A. Vogt: *Integral combinations of Heavisides*, ICS AS CR, Technical Report, 968, (2006).

[KuSa05b] V. Kůrková, M. Sanguinetti: *Learning with generalization capability by kernel methods with bounded complexity*, Journal of Complexity **21** (2005), 350–367.

[KuSa05a] V. Kůrková, M. Sanguinetti: *Error Estimates for Approximate Optimization by the Extended Ritz Method*, SIAM Journal on Optimization, **15** (2005), no. 2, 461–487.

[Lan93] S. Lang: *Real and Functional Analysis*, Springer, 1993.

[Lap] Lapack library: `http://www.netlib.org/lapack/`.

[Lax02] P. D. Lax: *Functional Analysis*, J. Wiley, (2002).

[LeTa91] M. Ledoux, M. Talagrand: *Probability in Banach spaces*, Springer-Verlag, Berlin, (1991).

[LLPS93] M. Leshno, V. Lin, A. Pinkus, S. Schocken: *Multilayer feedforward networks with a non-polynomial activation function can approximate any function*, Neural Networks, **6** (1993), 861–867.

[LFN03] Y. Liao, S.C. Fang, H. Nuttle: *Relaxed conditions for radial-basis function networks to be universal approximators*, Neural Networks, **16** (2003), No. 7, 1019–1028.

[LiPi94] V. Lin, A. Pinkus: *Approximation of multivariate functions*, Advances of Computational Mathematics, New Delhi, India, World Scientific, Singapore, (1994), 257–265.

[Lu02] J. Lukeš: *Zápisky z funkcionální analýzy*, Karolinum, Praha, (2002).

[LuMa95] J. Lukeš,J. Malý: *Measure and Integral*, Matfyzpress, Praha, (1995).

[Ma99] V. E. Maiorov: *On best approximation by ridge functions*, Journal of Approximation Theory, **99** (1999), no. 1, 68–94.

[Ma03] V. E. Maiorov: *On best approximation of classes by radial functions*, Journal of Approximation Theory, **120** (2003), 36–70.

[MM00] V. Maiorov and R. Meir: *On the Near Optimality of the Stochastic Approximation of Smooth Functions by Neural Networks*, Advances in Computational Mathematics, **13** (2000), No. 1, 79–103.

[MMR99] V. Maiorov, R. Meir, J. Ratsaby: *On the Approximation of Functional Classes Equipped with a Uniform Measure Using Ridge Functions*, Journal of Approximation Theory, **99** (1999), No. 1, 95–111.

[MaPi99] V. Maiorov, A. Pinkus: *Lower bounds for approximation by MLP neural networks*, Neurocomputing **25** (1999), 81–91.

[Mk96] Y. Makovoz: *Random approximants and neural networks*, Journal of Approximation Theory **85** (1996), 98–109.

[Mh96] H. N. Mhaskar, *Neural networks for optimal approximation of smooth and analytic functions*, Neural Computation, **8** (1996), 164–177.

[MiPa69] M.L. Minsky, S.A. Papert: *Perceptrons*, MIT Press, Cambridge, MA, (1969).

[PaSa93] J. Park, I. W. Sandberg: *Approximation and radial-basis-function networks*, Neural Computation, **5** (1993), 305–316.

[PaSa91] J. Park, I. W. Sandberg: *Universal approximation using radial-basis-function networks*, **3** (1991), 246–257.

[Pe99] P. P. Petrushev: *Approximation by ridge functions and neural networks*, SIAM Journal on Mathematical Analysis, **30** (1999), No. 1, 155–189.

[Pi85] A. Pinkus: *n-Widths in Approximation Theory*, Springer-Verlag, Berlin-Heidelberg, Germany, (1985).

[Pi99] A. Pinkus: *Approximation theory of the MLP model in neural networks*, Acta Numerica (1999), 143–195.

[Ps81] G. Pisier: *Remarques sur un resultat non publi'e de B. Maurey*, in Seminaire D'Analyse Fonctionnelle, 1980-1981, 'Ecole Polytechnique, Centre de Math'ematiques, Palaiseau, France (1981).

[PoSm03] T. Poggio, S. Smale: *The Mathematics of Learning: Delaing with Data*, Notices of the AMS, **50** (2003), no. 5, 536–544.

[Proben] L. Prechelt: *PROBEN1 – A Set of Benchmarks and Benchmarking Rules for Neural Network Training Algorithms*, Universitaet Karlsruhe, 21/94, (1994).

[Re83] W.J. Rey: *Introduction to Robust and Quasi-Robust Statistical Methods*, Springer-Verlag, Berlin, (1983).

[Ru91] W. Rudin: *Functional Analysis*, 2nd Edition, McGraw-Hill, NY, (1991).

[RHW86] D.E. Rumelhart, G.E. Hinton, R.J. Williams: *Learning internal representations by error propagation*, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1, MIT Press Cambridge, USA, (1986), 318–362.

[Sa02] W.S. Sarle: Editor of *Neural Network FAQ*, Usenet newsgroup comp.ai.neural-nets, URL: ftp://ftp.sas.com/pub/neural/FAQ.html.

[SchWe06] R. Schaback, H. Wendland: *Kernel techniques: From machine learning to meshless methods*, Acta Numerica, (2006), 543–639.

[SchSl02] B. Schölkopf, A.J. Smola: *Learning with Kernels*, MIT Press, Cambridge, Massachusetts, (2002).

[S08] T. Šámalová: *$\mathcal{G}$-variation in $\mathcal{L}^p$-spaces and integral representation*, ICS AS CR, Technical Report, V-1021, (2008).

[SS08] T. Šámalová, R. Šámal: *Pruning algorithms for one-hidden-layer feedforward neural networks* ICS AS CR, Technical Report, V-1022, (2008).

[S04a] T. Šidlofová: *Existence and Uniqueness of Minimization Problems with Fourier Based Stabilizers*, Compstat 2004, Prague, (2004).

[S04b] T. Šidlofová: *Kernel Based Regularization and Neural Networks*, Doktorandský Den 2004, Matfyzpres, (2004), 134–140.

[S04c] T. Šidlofová: *Minimization Problems with Fourier Based Stabilizers*, ICS AS CR, Technical Report, V-904, (2004).

[S03a] T. Šidlofová: *Bounds on Rates af Approximation by Neural Networks in Lp-spaces*, Artificial Neural Nets and Genetic Algorithms, SpringerVerlag, (2003), 23–27.

[S03b] T. Šidlofová: *Estimates of Rates of Approximation by Neural Networks in Lp-Spaces*, Kognícia, umelý život a počítačová inteligencia, ELFA, (2003), 365–368.

[S03c] T. Šidlofová: *Neural Network Approximation and Regularization Theory*, Doktorandský den 2003, Matfyzpress, (2003), 115–121.

[S02] T. Šidlofová: *Comparison of Radial and Neural Network Approximation*, Doktorandský den 2002, Ústav informatiky AV ČR, (2002), 1–7.

[S01] T. Šidlofová: *Some Estimates of Rates of Approximation by Neural Networks Using Integral Representations*, Doktorandský den 2001, Ústav informatiky AV ČR, (2001), 33–36.

[SiNe96] J. Šíma, R. Neruda: *Teoretické otázky neuronových sítí*, Matfyzpress, (1996).

[StWh90] M. Stinchcombe, H. White: *Approximating and learning unknown mappings using multilayerfeedforward networks with bounded weights*, Neural Networks, IJCNN International Joint Conference on Neural Networks, **3** (1990), 7–16.

[TiAr77] A. N. Tikhonov and V. Y. Arsenin. Solutions of Ill-posed Problems. W. H. Winston, Washington, D.C., (1977).

[VoKr61] B. A. Vostrecov, M. A. Kreines: *Approximation of continuous functions by superpositions of plane waves*, Dokl. Akad. Nauk SSSR **140**, 1237–1240 = Soviet Math. Dokl. **2** (1961), 1326–1329.

[Wa90] G. Wahba: *Spline Models for Observational Data*, Series in Applied Mathematics, **59**, SIAM, Philadelphia, (1990).

[We00] S. Weinzierl: *Introduction to Monte Carlo methods*, arXiv:hep-ph/0006269v1, (2000).

[Zi89] W.P. Ziemer: *Weakly Differentiable Functions*, Springer, (1989).