1

Introduction

This monograph is about *Ridge Functions*. A ridge function is any multivariate real-valued function

$$F : \mathbb{R}^n \to \mathbb{R}$$

of the form

$$F(x_1, \ldots, x_n) = f(a_1 x_1 + \cdots + a_n x_n) = f(\mathbf{a} \cdot \mathbf{x}),$$

where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ are the variables, f is a univariate real-valued function, i.e., $f : \mathbb{R} \to \mathbb{R}$, and $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ is a fixed vector. This vector $\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ is generally called the *direction*. In other words, a ridge function is a multivariate function constant on the parallel hyperplanes $\mathbf{a} \cdot \mathbf{x} = c$, $c \in \mathbb{R}$. It is one of the simpler multivariate functions. Namely, it is a superposition of a univariate function with one of the simplest multivariate functions, the inner product.

More generally, we can and will consider, for given d, $1 \leq d \leq n-1$, functions F of the form

$$F(\mathbf{x}) = f(A\mathbf{x}),$$

where A is a fixed $d \times n$ real matrix, and $f : \mathbb{R}^d \to \mathbb{R}$. We call such functions *Generalized Ridge Functions*. For $d = 1$, this reduces to a ridge function.

1.1 Motivation

We see specific ridge functions in numerous multivariate settings without considering them of interest in and of themselves. We find them, for example, as kernels in integral formulæ. They appear in the Fourier transform

$$F(\mathbf{w}) = \int_{\mathbb{R}^n} e^{-i(\mathbf{w} \cdot \mathbf{x})} f(\mathbf{x})\, d\mathbf{x},$$

and its inverse. We see them in the n-dimensional Radon transform

$$(R_{\mathbf{a}}f)(t) = \int_{\mathbf{a}\cdot\mathbf{x}=t} f(\mathbf{x})\, d\sigma(\mathbf{x}),$$

and its inverse. Here the integral is taken with respect to the natural hypersurface measure $d\sigma$. It is possible to generalize the Radon transform still further by integrating over $(n-d)$-dimensional affine subspaces of $\mathbb{R}^n$. In addition, we find them in the Hermite–Genocchi formula for divided differences

$$f[x_0, x_1, \ldots, x_n] = \int_{\Sigma_n} f^{(n)}(\mathbf{t}\cdot\mathbf{x})d\mathbf{t},$$

where Σ_n is the n-simplex in $\mathbb{R}^{n+1}_+$, i.e., $\Sigma_n = \{\mathbf{t} = (t_0, t_1, \ldots, t_n) : t_i \geq 0, \sum_{i=0}^n t_i = 1\}$. See, for example, de Boor [2005] for a discussion and history of this formula. They appear in multivariate Fourier series where the basic functions are of the form $e^{i(\mathbf{n}\cdot\mathbf{x})}$, for $\mathbf{n} \in \mathbb{Z}^n$. And also in partial differential equations where, for example, if P is a constant coefficient polynomial in n variables, then

$$P\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right) f = 0$$

has a solution of the form $f(\mathbf{x}) = e^{\mathbf{a}\cdot\mathbf{x}}$ if and only if $P(\mathbf{a}) = 0$.

Classes of ridge functions also play a fundamental role in various subjects. The term *ridge function* is rather recent. However, these functions had been considered for many years under the name of *plane waves*. See, for example, the well-known book of John [1955]. In that book are considered representations of multivariate functions using integrals whose kernels are specific "plane waves" and applications thereof to partial differential equations. Plane waves are also discussed by Courant and Hilbert [1962]. In general, linear combinations of ridge functions with fixed directions occur in the study of hyperbolic constant coefficient partial differential equations. As an example, assume that the (a_i, b_i) are pairwise linearly independent vectors in $\mathbb{R}^2$. Then the general "solution" to the homogeneous partial differential equation

$$\prod_{i=1}^r \left(b_i \frac{\partial}{\partial x} - a_i \frac{\partial}{\partial y}\right) F = 0,$$

where the derivatives are understood in the sense of distributions, are all functions of the form

$$F(x, y) = \sum_{i=1}^r f_i(a_i x + b_i y),$$

for (almost) arbitrary univariate functions f_i.

The term *ridge function* was coined in the 1975 paper by Logan and Shepp

[1975]. Seemingly, they were unaware of the previous terminology, and the term "ridge function" has now been fairly universally adopted. This was a seminal paper in computerized tomography. In tomography, or at least in tomography as the theory was initially constructed in the early 1980s, ridge functions were basic. The idea there was to try to reconstruct a given, but unknown, function $G(\mathbf{x})$ from the values of its integrals along certain parallel planes or lines. Logan and Shepp considered functions in the unit disk in $\mathbb{R}^2$ with given line integrals along parallel lines and a finite number of equally spaced directions. More generally, consider some nice domain K in $\mathbb{R}^n$, and a function G belonging to $L^2(K)$. Assume that for some fixed directions $\{\mathbf{a}^i\}_{i=1}^r$ we are given the values

$$\int_{K\cap\{\mathbf{a}^i\cdot\mathbf{x}=\lambda\}} G(\mathbf{x})\, d\sigma(\mathbf{x})$$

for each λ and $i = 1, \ldots, r$, where $d\sigma(\mathbf{x})$ is the natural measure on the hyperplanes $\{\mathbf{x} : \mathbf{a}^i\cdot\mathbf{x} = \lambda\}$. They (mis-)termed these values the *projections* of G along the hyperplanes $K \cap \{\mathbf{a}^i \cdot \mathbf{x} = \lambda\}$. Assume that we are given these values for each λ and $i = 1, \ldots, r$. What is a good method of reconstructing G based only on this information? It easily transpires, from basic orthogonality considerations, that the unique best $L^2(K)$ approximation

$$f^*(\mathbf{x}) = \sum_{i=1}^{r} f_i^*(\mathbf{a}^i \cdot \mathbf{x})$$

to G from the linear subspace

$$\mathcal{M}(\mathbf{a}^1, \ldots, \mathbf{a}^r) = \left\{ \sum_{i=1}^{r} f_i(\mathbf{a}^i \cdot \mathbf{x}) : \ f_i \text{ vary} \right\},$$

if such a best approximation exists, necessarily satisfies

$$\int_{K\cap\{\mathbf{a}^i\cdot\mathbf{x}=\lambda\}} G(\mathbf{x})\, d\sigma(\mathbf{x}) = \int_{K\cap\{\mathbf{a}^i\cdot\mathbf{x}=\lambda\}} f^*(\mathbf{x})\, d\sigma(\mathbf{x})$$

for each λ and $i = 1, \ldots, r$. That is, it has the same projections as G. Furthermore, since it is a best approximation from a linear subspace in a Hilbert space, its norm is strictly less than the norm of G, unless $f^* = G$. Thus, among all functions with the same data (projections) as G, this specific linear combination of ridge functions is the one of minimal $L^2(K)$ norm. In the unit disk in $\mathbb{R}^2$ with equally spaced directions, Logan and Shepp also give a more closed-form expression for f^*.

Ridge functions and ridge function approximations are also studied in statistics in the analysis of large multivariate data sets. There they often go under the name of *projection pursuit*, see, for example, Friedman and Stuetzle [1981], Huber [1985] and Donoho and Johnstone [1989]. Projection pursuit algorithms approximate a function G of n variables by functions of the form

$$\sum_{i=1}^{r} f_i(\mathbf{a}^i \cdot \mathbf{x}),$$

where both the directions $\mathbf{a}^i$ and the univariate functions f_i are variables. The idea here is to "reduce dimension" and thus bypass the "curse of dimensionality". Each $\mathbf{a}^i \cdot \mathbf{x}$ is considered as a projection of $\mathbf{x}$. The directions $\mathbf{a}^i$ are chosen to "pick out the salient features". The method of approximation, introduced by Friedman and Stuetzle [1981] and called projection pursuit regression (PPR), is essentially a stepwise greedy algorithm that, at its kth step, looks for a best (or good) approximation of the form $f_k(\mathbf{a}^k \cdot \mathbf{x})$ to the function $G(\mathbf{x}) - \sum_{i=1}^{k-1} f_i(\mathbf{a}^i \cdot \mathbf{x})$, as we vary over both the univariate function f_k and the direction $\mathbf{a}^k$.

Ridge functions appear in many neural network models. One of the popular models in the theory of neural nets is that of a *multilayer feedforward perceptron* (MLP) neural net with input, hidden and output layers. The simplest case (which is that of one hidden layer, r processing units and one output) considers, in mathematical terms, functions of the form

$$\sum_{i=1}^{r} \alpha_i \sigma(\mathbf{w}^i \cdot \mathbf{x} + \theta_i),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is some given fixed univariate function, $\theta_i \in \mathbb{R}$, and $\mathbf{w}^i \in \mathbb{R}^n \backslash \{\mathbf{0}\}$. In this model, which is just one of many, we are in general permitted to vary over the $\mathbf{w}^i$ and θ_i, in order to approximate an unknown function. Note that for each $\mathbf{w}$ and θ the function

$$\sigma(\mathbf{w} \cdot \mathbf{x} + \theta)$$

is a ridge function. Thus, a lower bound on the degree of approximation by such functions is given by the degree of approximation by linear combinations of ridge functions. See, for example, Pinkus [1999] and references therein for more on this problem.

Motivated by the previous two topics, and other considerations, Candès in his thesis Candès [1998], see also Candès [1999], introduced the theory of ridgelets. In essence, the set

$$\{\sigma(\mathbf{w} \cdot \mathbf{x} + \theta) \, : \, \mathbf{w} \in \mathbb{R}^n, \, \theta \in \mathbb{R}\}$$

is called the set of *ridgelets* generated by σ. Ridgelets generated by a σ are a subset of ridge functions. For a class of σ, Candès [1998], [1999], provides an integral representation for functions with an associated ridgelet kernel. He then discretizes this representation with an eye towards obtaining approximations that are constructive, qualitative and stable.

Even the restriction of ridge functions to polynomials leads to interesting questions. *Waring's Problem* asks whether every positive integer can be expressed as a sum of at most $h(m)$ mth powers of positive integers, where $h(m)$ depends only upon m. This problem was solved in the affirmative by Hilbert [1909]. The key result in his proof was the following: for given m and n, and $N = \binom{n-1+2m}{n-1}$, there exist $\mathbf{a}^i \in \mathbb{Z}^n$, $i = 1, \ldots, N+1$, and λ_i positive rationals, $i = 1, \ldots, N+1$, such that

$$(x_1^2 + \cdots + x_n^2)^m = \sum_{i=1}^{N+1} \lambda_i (\mathbf{a}^i \cdot \mathbf{x})^{2m},$$

see also Stridsberg [1912]. A lucid exposition of Waring's Problem, and elementary proof of this result may be found in Ellison [1971]. Waring's Problem has various generalizations. One of them, for example, is the following. Can each homogeneous polynomial of degree m in n variables be written as a linear combination of mth powers of r linear homogeneous polynomials, where r depends only on n and m, i.e., linear combinations of $(\mathbf{a} \cdot \mathbf{x})^m$, where $\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$? And if it can, what is then the minimal number $h(m, n)$ such that each homogeneous polynomial of degree m in n variables can be written as a linear combination of mth powers of $h(m, n)$ linear homogeneous polynomials? And what about the same question for general algebraic polynomials of degree at most m in n variables? That is, we wish to express each algebraic polynomial of degree at most m in n variables in the form

$$p(\mathbf{x}) = \sum_{i=1}^{r} q_i (\mathbf{a}^i \cdot \mathbf{x}),$$

where the q_i are univariate algebraic polynomials, and r is minimal. There is a rich literature, mainly in number theory, on this and related issues.

Ridge functions are also of interest to researchers and students of approximation theory. The basic goal in approximation theory is straightforward and fundamental. Approximate complicated objects by simpler objects. Recent years have witnessed a flurry of interest in approximation from different classes of multivariate functions. We have, for example, multivariate spline functions, wavelets, radial basis functions, and many other such classes. Among the class of multivariate functions, linear combinations of ridge functions are a class of simpler functions. The questions one asks are the fundamental questions of approximation theory. Can one approximate arbitrarily well (density)? How well can one approximate (degree of approximation)? How does one approximate (algorithms)?

In this monograph we review much of what is today known about ridge functions. We hope this whets the reader's appetite, as much still remains unknown.

1.2 Organization

These notes are organized as follows. In Chapters 2–4 we consider some of the very basic properties of finite linear combinations of ridge and generalized ridge functions. In Chapter 2 we ask what can be said about the smoothness of each ridge function component if a finite linear combination of them is smooth. For example, assume

$$F(\mathbf{x}) = \sum_{i=1}^{r} f_i(\mathbf{a}^i \cdot \mathbf{x}) \tag{1.1}$$

and $F \in C^k(\mathbb{R}^n)$. What, if anything, does this imply with regard to the smoothness of the f_i? In Chapter 3 we consider to what extent the representation of a function as a finite linear combination of ridge functions is unique. That is, how many fundamentally different ways are there to represent an F of the form (1.1) as a linear combination of a finite number of ridge functions? In Chapter 4 we study an inverse problem. Namely, given an F of the form (1.1) with known or unknown directions and unknown functions, is it possible to identify associated unknown directions and functions in the finite sum based on our knowledge of F? Definitive answers to all these questions are not known.

Algebraic and homogeneous polynomials are important in the study of ridge functions. In Chapter 5 we consider ridge functions that are polynomials and discuss a wide variety of associated problems. In particular, we study questions of linear independence, interpolation and spanning by linear combinations of $(\mathbf{a}\cdot\mathbf{x})^m$ in the space of homogeneous polynomials of degree m, as we vary over a subset of directions, ask similar questions for algebraic polynomials of degree m, and discuss Waring's Problem for real homogeneous and algebraic polynomials.

In Chapter 6 we consider various questions associated with the density of linear combinations of ridge functions with fixed and variable directions in the set of continuous functions on $\mathbb{R}^n$, in the topology of uniform convergence on compact subsets of $\mathbb{R}^n$.

Chapter 7 contains a discussion of the closure properties of finite linear combinations of ridge functions with given directions in different norms and domains, while Chapter 8 is concerned with the existence and characterization of best approximations from these same subspaces.

In Chapter 9 we survey approximation algorithms for finding best approximations from spaces of linear combinations of ridge functions. We consider approximations in the cases of both fixed and variable directions. The algorithms considered are all predicated on the notion that it is possible to find a best approximation from each of its component subspaces, i.e., sets of ridge functions with one direction.

In Chapter 10 we look at integral representations of functions where the kernel is a ridge function. In particular we consider an integral representation using an orthogonal decomposition in terms of Gegenbauer polynomials (from Petrushev [1998]), and an integral representation based upon ridgelets (as presented by Candès [1998]).

Chapters 11 and 12 are concerned with the problem of interpolation by finite linear combinations of ridge functions. In Chapter 11 we look at point interpolation, while in Chapter 12 we consider interpolation to data given on straight lines.

In most of the chapters we also consider the extent to which the results reported on can extend to generalized ridge functions.

Finally, the reference section is divided into two parts. The first section contains all works that are actually referenced in the text. In a futile attempt to provide the interested researcher with a complete overview of the subject we have included a supplemental list of references on ridge functions.

There are topics related to ridge functions that are not presented here. The most glaring omission is that of degree of approximation, i.e., estimates on the error of approximation when using linear combinations of ridge functions, and the understanding of which classes of functions are well approximated by linear combinations of ridge functions, and which classes are not well approximated by linear combinations of ridge functions. Different papers are devoted to various aspects of this problem. We wish to mention Oskolkov [1997], [1999a], Petrushev [1998], Maiorov [1999], Maiorov, Meir and Ratsaby [1999], Maiorov, Oskolkov and Temlyakov [2002] and Maiorov [2010a]. Most known error estimates for approximating by linear combinations of ridge functions do not provide for bounds that are better than those provided by the full space of algebraic polynomials they contain. That is, in $\mathbb{R}^n$ there are many choices of m directions for which the space of linear combinations of ridge functions with these directions are easily seen to contain all algebraic polynomials of degree $c_n m^{1/(n-1)}$, with a constant c_n independent of m. The error estimates for many different classical function spaces, when approximating by either linear combinations of m ridge functions or the algebraic polynomials they contain, are comparable. As this is the case, then why bother approximating by ridge functions? Ridge functions are undoubtedly better approximants for certain classes of functions. But for which classes of functions? An interesting example is due to Oskolkov [1999a]. He proved therein that, for harmonic functions in $\mathbb{R}^2$, approximation by ridge functions gives significantly better bounds than those provided by the associated algebraic polynomials. In addition, as has been pointed out by Candès and Donoho, see, for example, Candès and Donoho [1999], ridge functions with varying directions are well-adapted to handle singularities along $(n-1)$-dimensional hyperplanes. Nevertheless the full theory, in the opinion of the author, is still very much lacking. This is unfortunate as the problems are both interesting and important.

1.3 Notation

In this section we review some of the notation that will be used repeatedly in these notes. A *direction* is any non-zero vector in $\mathbb{R}^n$. For a given direction $\mathbf{a} = (a_1, \ldots, a_n)$, set

$$\mathcal{M}(\mathbf{a}) := \{f(\mathbf{a} \cdot \mathbf{x}) : f : \mathbb{R} \to \mathbb{R}\},$$

where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ are the variables and

$$\mathbf{a} \cdot \mathbf{x} = \sum_{i=1}^{n} a_i x_i$$

is the standard inner product on $\mathbb{R}^n$. Note that $\mathcal{M}(\mathbf{a})$ is an infinite-dimensional linear subspace, and since we are varying over all univariate functions f it immediately follows that

$$\mathcal{M}(\mathbf{a}) = \mathcal{M}(\mathbf{b})$$

for directions $\mathbf{a}$ and $\mathbf{b}$ if and only if $\mathbf{a} = \lambda \mathbf{b}$ for some $\lambda \in \mathbb{R}$, $\lambda \neq 0$. Thus we could assume that the directions $\mathbf{a}$ are chosen to be of norm 1 and also identify $\mathbf{a}$ with $-\mathbf{a}$. But there seems to be no particular advantage in such an assumption.

Given directions $\mathbf{a}^i$, $i = 1, \ldots, r$, we set

$$\mathcal{M}(\mathbf{a}^1, \ldots, \mathbf{a}^r) := \mathcal{M}(\mathbf{a}^1) + \cdots + \mathcal{M}(\mathbf{a}^r)$$

$$= \left\{ \sum_{i=1}^{r} f_i(\mathbf{a}^i \cdot \mathbf{x}) : f_i : \mathbb{R} \to \mathbb{R},\ i = 1, \ldots, r \right\}.$$

We will sometimes also use the following notation. For a set $\Omega \subseteq \mathbb{R}^n$ we let

$$\mathcal{M}(\Omega) := \operatorname{span}\{f(\mathbf{a} \cdot \mathbf{x}) : f : \mathbb{R} \to \mathbb{R},\ \mathbf{a} \in \Omega\}.$$

These are all linear spaces.

Similarly, for a given d, $1 \leq d \leq n-1$, and $d \times n$ matrices $A^1, \ldots, A^r$, we let

$$\mathcal{M}(A^1, \ldots, A^r) := \left\{ \sum_{i=1}^{r} f_i(A^i \mathbf{x}) : f_i : \mathbb{R}^d \to \mathbb{R},\ i = 1, \ldots, r \right\}.$$

Sometimes we will also let Ω_d denote a subset of $d \times n$ real matrices, and set

$$\mathcal{M}(\Omega_d) := \text{span}\{f(A\mathbf{x}) : A \in \Omega_d,\ f : \mathbb{R}^d \to \mathbb{R}\}.$$

In $\mathbb{R}^n$ we let B^n and S^{n-1} denote the unit ball and unit sphere, respectively. That is,

$$B^n := \{\mathbf{x} : \|\mathbf{x}\|_2 \le 1\}$$

and

$$S^{n-1} := \{\mathbf{x} : \|\mathbf{x}\|_2 = 1\},$$

where $\|\cdot\|_2$ is the usual Euclidean (ℓ_2) norm on $\mathbb{R}^n$.

We recall some standard multi-index notation. For $\mathbf{k} = (k_1, \ldots, k_n) \in \mathbb{Z}_+^n$, let $|\mathbf{k}| = k_1 + \cdots + k_n$ and $\mathbf{k}! = k_1! \cdots k_n!$. We have that

$$\binom{|\mathbf{k}|}{\mathbf{k}} := \frac{|\mathbf{k}|!}{\mathbf{k}!} = \frac{(k_1 + \cdots + k_n)!}{k_1! \cdots k_n!}$$

are the usual multinomial coefficients. Given $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{k} \in \mathbb{Z}_+^n$, we set

$$\mathbf{x}^{\mathbf{k}} := x_1^{k_1} \cdots x_n^{k_n}.$$

Let H_m^n denote the set of real homogeneous polynomials of degree m in n variables, i.e.,

$$H_m^n := \left\{ \sum_{|\mathbf{k}|=m} b_{\mathbf{k}} \mathbf{x}^{\mathbf{k}} : b_{\mathbf{k}} \in \mathbb{R} \right\}.$$

It is well-known that $\dim H_m^n = \binom{n-1+m}{n-1}$. In addition, let Π_m^n denote the set of all real algebraic polynomials of total degree at most m in n variables, i.e.,

$$\Pi_m^n := \left\{ \sum_{|\mathbf{k}|\le m} b_{\mathbf{k}} \mathbf{x}^{\mathbf{k}} : b_{\mathbf{k}} \in \mathbb{R} \right\}$$

or

$$\Pi_m^n = \bigoplus_{r=0}^{m} H_r^n.$$

It is easily verified that $\dim \Pi_m^n = \dim H_m^{n+1} = \binom{n+m}{n}$. By Π^n we denote the set of all algebraic polynomials of n variables, and by H^n the set of all homogeneous polynomials of n variables, i.e.,

$$H^n = \bigcup_{k=0}^{\infty} H_k^n.$$

For $\mathbf{k} \in \mathbb{Z}_+^n$, set

$$D^{\mathbf{k}} := \frac{\partial^{|\mathbf{k}|}}{\partial x_1^{k_1} \cdots \partial x_n^{k_n}} .$$

For any polynomial q of the form

$$q(\mathbf{x}) = \sum_{\mathbf{k}} a_{\mathbf{k}} \mathbf{x}^{\mathbf{k}},$$

where $a_{\mathbf{k}} \in \mathbb{R}$, we let

$$q(D) = \sum_{\mathbf{k}} a_{\mathbf{k}} D^{\mathbf{k}}$$

denote the associated constant coefficient partial differential operator. A simple calculation shows that, for $\mathbf{k} \in \mathbb{Z}_+^n$, $|\mathbf{k}| = m$, we have

$$D^{\mathbf{k}}(\mathbf{a} \cdot \mathbf{x})^\ell = \begin{cases} 0, & m > \ell \\ \frac{\ell!}{(\ell-m)!}\mathbf{a}^{\mathbf{k}}(\mathbf{a} \cdot \mathbf{x})^{\ell-m}, & m \leq \ell. \end{cases}$$

Thus, if $q \in H_m^n$, then

$$q(D)(\mathbf{a} \cdot \mathbf{x})^\ell = \begin{cases} 0, & m > \ell \\ \frac{\ell!}{(\ell-m)!}q(\mathbf{a})(\mathbf{a} \cdot \mathbf{x})^{\ell-m}, & m \leq \ell \end{cases} \tag{1.2}$$

and, in particular, for $q \in H_m^n$ we have

$$q(D)(\mathbf{a} \cdot \mathbf{x})^m = m!\, q(\mathbf{a}). \tag{1.3}$$

Furthermore, for $\mathbf{k}, \mathbf{j} \in \mathbb{Z}_+^n$, $|\mathbf{k}| = |\mathbf{j}| = m$, we also have

$$D^{\mathbf{k}}\mathbf{x}^{\mathbf{j}} = \delta_{\mathbf{k},\mathbf{j}}\mathbf{k}!, \tag{1.4}$$

where δ denotes the usual Dirac delta function. Finally, for $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{c} = (c_1, \ldots, c_n)$, let

$$D_{\mathbf{c}} := \sum_{k=1}^n c_k \frac{\partial}{\partial x_k}$$

denote differentiation in the direction $\mathbf{c}$. For any univariate function $f \in C^1(\mathbb{R})$ we have

$$D_{\mathbf{c}} f(\mathbf{a} \cdot \mathbf{x}) = (\mathbf{a} \cdot \mathbf{c}) f'(\mathbf{a} \cdot \mathbf{x}). \tag{1.5}$$

Notation is often a compromise and is not necessarily unconditionally exact. For example, in this monograph we will use $\mathbf{a}^i$ for vectors and A^i for matrices. The former is in boldface, while the latter is in italics. In addition the i is here an index and in neither case does it indicate a power. (The A^i are also not square matrices.) We also do not always differentiate between a function and its value at