Data Compression by Unsupervised Classification

Pötzelberger, Klaus; Strasser, Helmut

DOI:
10.57938/ea8b9a43-35db-4776-a28f-031c055d3873

Published: 01/01/1997

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

Document License:
Unspecified

Data Compression by Unsupervised Classification

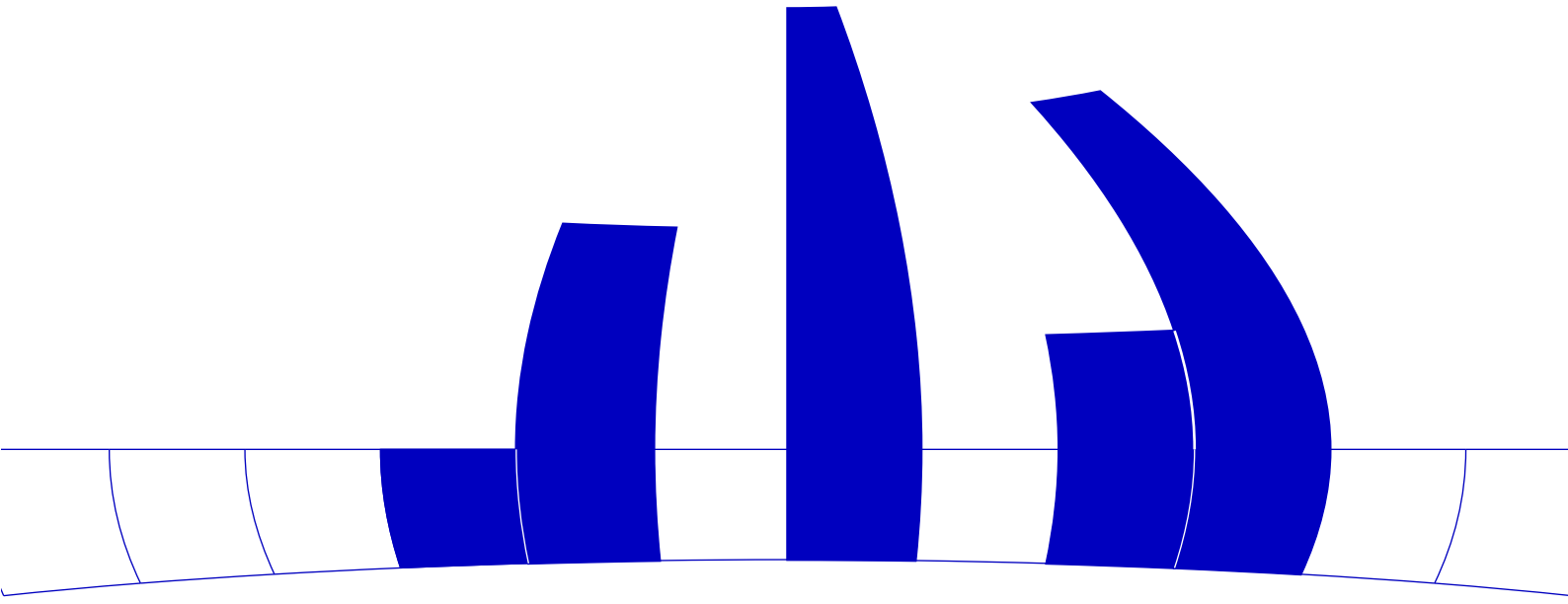
Klaus Pötzelberger, Helmut Strasser

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 52
November 1997

http://statmath.wu-wien.ac.at/



Data compression by unsupervised classification

Klaus Pötzelberger, Helmut Strasser
Department of Statistics, Vienna University of Economics and Business Administration, Augasse 2–6, A–1090 Vienna, Austria
E–mail: Klaus.Poetzelberger@wu-wien.ac.at, Helmut.Strasser@wu-wien.ac.at

Preprint, November 1997

Abstract: This paper deals with a general class of classification methods which are related both to vector quantization in the sense of Pollard, [12], as well as to competitive learning in the sense of Kohonen, [10]. The basic duality of minimum variance partitioning and vector quantization known from statistical cluster analysis is shown to be true for this whole class of classification problems. The paper contains theoretical results like existence of optima, consistency of approximate optima and characterization of local optima as fixpoints of a fix point algorithm. A fix point algorithm is proposed and its termination after finite time is proved for empirical distributions. The construction of a particular classification method is based on a statistical information measure specified by a convex function. Modifying this convex function gives room for suggesting a large variety of new classification procedures, e.g. of robust quantifiers.

1 Introduction

1.1 Some Background

Data compression is a central topic of nonparametric statistics. The most simple kind of data compression is a partition of the data range which gives a multinomial feature variable. Starting with multivariate data on different scales there is often no other possibility for data compression. Data compression by partitioning is called a classification.

[1] *AMS 1991 subject classifications.* Primary: 62H30, 62-07 secondary: 65U05, 68T05

[2] *Key words and phrases:* classification, vector quantization, principal point problem, convex function, information in a partition, fix point algorithm, robust quantifiers, neural networks, unsupervised learning

In the present paper we are interested in classifications which are not defined by a given feature variable. We are not dealing with optimal fit of partitions in order to forecast a feature variable. To put it into terms of learning algorithms we are not interested in supervised learning. Our topic is unsupervised learning. We want to find partitions which contain as much information as possible about the original data set. To put it into statistical terms our topic is cluster analysis.

There is a lot of recent literature on supervised and unsupervised learning. For completeness let us give as standard reference on supervised learning Devroye, Györfi und Lugosi, 1996, [4]. Basic references on unsupervised learning are Bock, 1974, [1], and Hartigan, 1975, [8]. Unsupervised learning in a wider sense is also covered by Conway und Sloane, 1993, [3] which deals with coding and number theory. A plenty of algorithms came up recently in connection with artificial neural networks. A classical reference is Kohonen, 1984, [10].

In this paper we do not seek to give an overview over the variety of methods for unsupervised learning of a partition. We will rather discuss a very special idea of partitioning a data set from the methodological point of view. This idea is based on the optimization of an objective function which is a measure of the information contained in the partition.

Our approach to the problem of data compression is closely related to the idea of vector quantization. Vector quantization, which has been considered e.g. by Pollard, [12] and [13], is also the subject of very recent papers like Cuesta-Albertos, Gordaliza and Matran, [9], and Bouton and Pages, [2]. The problem of vector quantization in the sense treated by Pollard, [12], contains as a special case the problem of finding a minimum variance partition (MVP). The method of data compression which is the subject of the present paper is also a generalization of the MVP-problem but goes into a different direction. We think that our approach has some advantages compared with vector quantization in the usual sense. We will try to explain this in the following.

1.2 How the problem arises

Let us start with some notation.

Let P be a probability distribution on $\mathbb{R}^d$ having finite first moments. This distribution P can be either a statistical model or the empirical distribution of a data set. Suppose that $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ is a partition of $\mathbb{R}^d$. Let $|\mathcal{B}|$ be the size of the partition $\mathcal{B}$, i.e. the number of sets in the partition. Consider the family of all partitions whose size does not exceed a given number m. It is our goal to find an optimal partition in that family, i.e. a partition which contains as most information

as possible with respect to the given distribution P. By $E(X|\mathcal{B})$ we denote the conditional expectation of the random variable $X(x) := x$, $x \in \mathbb{R}^d$ (cf. (2.1) and (2.2)).

The following is a list of optimization problems, each of which aims at finding partitions with nice properties. The relations between these optimization problems are the starting points both of usual vector quantization and of our approach to the problem.

(1.1) PROBLEM A: *(Minimal variance partition, MVP) Find a partition $\mathcal{B}$ with $|\mathcal{B}| = m$ such that*

$$\int ||X - E(X|\mathcal{B})||^2 \, dP = \text{Min !}$$

(1.2) PROBLEM B: *Find a partition $\mathcal{B}$ with $|\mathcal{B}| = m$ such that*

$$\int ||E(X|\mathcal{B})||^2 \, dP = \text{Max !}$$

(1.3) PROBLEM C: *(Principal point problem, PPP) Find points $a_1, \ldots, a_m \in \mathbb{R}^d$ such that*

$$G(a_1, \ldots, a_m) := \int \min_j ||X - a_j||^2 \, dP = \text{Min !}$$

and choose $\mathcal{B} = (B_1, \ldots, B_m)$ according to

$$B_j = \{x : j = \text{argmin}_k ||x - a_k||^2\}, \quad j = 1, 2, \ldots, m.$$

(1.4) PROBLEM D: *Find points $a_1, \ldots, a_m \in \mathbb{R}^d$ such that*

$$F(a_1, \ldots, a_m) := \int \max_j \left(a_j' X - ||a_j||^2/2\right) dP = \text{Max !}$$

and choose $\mathcal{B} = (B_1, \ldots, B_m)$ according to

$$B_j = \left\{x : j = \text{argmax}_k \left(a_k' x - ||a_k||^2/2\right)\right\}, \quad j = 1, 2, \ldots, m.$$

The following theorem is a basic and well-known fact from statistical cluster analysis.

(1.5) THEOREM *The optimization problems A, B, C and D are equivalent.*

Proof: The equivalence of A and B follows from the equation

$$\int ||X||^2 \, dP = \int ||X - E(X|\mathcal{B})||^2 \, dP + \int ||E(X|\mathcal{B})||^2 \, dP.$$

The equivalence of C and D is obvious from elementary algebra. The only non-trivial part is the equivalence of A and C which is a fundamental fact of cluster analysis (see e.g. Bock, [1]). □

Vector quantization deals with modifications of one or more of the optimization problems A to D. There are two main directions of possible generalizations.

The first and most popular generalization starts with problem C replacing the quadratic function $\Phi(x) = x^2$ by any other increasing function. In this way one obtains the following optimization problem.

(1.6) PROBLEM C*: *Find points* $a_1, \ldots, a_m \in \mathbb{R}^d$ *such that*

$$G(a_1, \ldots, a_m) := \int \min_j \Phi(||X - a_j||) \, dP = \text{Min !}$$

and choose $\mathcal{B} = (B_1, \ldots, B_m)$ *according to*

$$B_j = \{x : j = \text{argmin}_k \Phi(||x - a_k||)\}, \quad j = 1, 2, \ldots, m.$$

Opimization problems of the type C* are called generalized principal point problems. Such problems are considered by Pollard, [12] and [13], Flury, [7], Flury, Tarpey and Li, [6], Cuesta-Albertos, Gordaliza and Matran, [9]. Generalized PPPs are not related to a problem of kind A. Thus, the statistical interpretation given by the MVP problem cannot carried over from the quadratic function $\Phi(x) = x^2$ to the general case.

A different way of generalizing the vector quantization method has been used by Kohonen, [10]. Kohonen starts with problem D replacing it by the following optimization problem.

(1.7) PROBLEM D*: *(Kohonen problem) Find points* $a_1, \ldots, a_m \in \mathbb{R}^d$ *with* $||a_j|| \leq 1$, $j = 1, 2, \ldots, m$, *such that*

$$F(a_1, \ldots, a_m) := \int \max_j a_j' X \, dP = \text{Max !},$$

and choose $\mathcal{B} = (B_1, \ldots, B_m)$ *according to*

$$B_j = \{x : j = \text{argmax}_k a_k' x\}, \quad j = 1, 2, \ldots, m.$$

This optimization problem D* plays an important role for data compression models of associative memory. It is an interesting and important fact that for the Kohonen problem D* there is an equivalent statistical formulation.

(1.8) PROBLEM B*: *Find a partition* $\mathcal{B}$ *with* $|\mathcal{B}| = m$ *such that*

$$\int ||E(X|\mathcal{B})||\, dP = \text{Max !}$$

(1.9) THEOREM *The Kohonen problem D* is equivalent to problem B*.*

Proof: This is a special case of our equivalence theorem (2.12). □

Thus, the Kohonen problem D* is not only a modification of the classical problem D, but the equivalence of B and D can be carried over to B* and D*, too. This leads in a natural way to the question for the common mathematical background of B⇔D and B*⇔D*. This is one of the questions answered in the present paper.

The following definition contains a class of measures of the amount of information in a given partition with respect to a probability distribution.

(1.10) DEFINITION *Let* $f : \mathbb{R}^d \to (-\infty, \infty]$ *be a convex function. If* $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ *is a partition of* $\mathbb{R}^d$*, then*

$$I_f(\mathcal{B}, P) := \int f\Big(E(X|\mathcal{B})\Big)\, dP \tag{1.11}$$

is called the f-information of the partition $\mathcal{B}$ *with respect to the distribution P.*

We define a general optimization problem which covers both problems B and B*. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function.

(1.12) PRIMAL PROBLEM: *Find a partition* $\mathcal{B}$ *with* $|\mathcal{B}| = m$ *such that*

$$\int f(E(X|\mathcal{B}))\, dP = \text{Max !}$$

It is obvious that the optimization problems B and B* are special cases of the primal problem. Next we consider another optimization problem which covers D and D*. For this we require the concept of the conjugate convex function (Legendre transform).

(1.13) DEFINITION *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then*

$$f^c(a) := \sup_x (a'x - f(x)), \quad a \in \mathbb{R}^d,$$

is called the conjugate convex function of f.

(1.14) EXAMPLE In section 3 we will give several important examples of convex functions and their conjugates. For the moment let us illustrate the concept of conjugate convex functions by the two most basic cases:

$$f(x) = \frac{1}{2}||x||^2 \quad \Rightarrow \quad f^c(a) = \frac{1}{2}||a||^2,$$

and

$$f(x) = ||x|| \quad \Rightarrow \quad f^c(a) = \begin{cases} 0 & ||a|| \leq 1, \\ \infty & ||a|| > 1. \end{cases}$$

(1.15) DUAL PROBLEM: *Find points $a_1, \ldots, a_m \in \mathbb{R}^d$ such that*

$$F(a_1, \ldots, a_m) := \int \max_j \left(a_j'X - f^c(a_j)\right) dP = \text{Max !},$$

and choose $\mathcal{B} = (B_1, \ldots, B_m)$ according to

$$B_j = \left\{x : j = \text{argmax}_k \left(a_k'x - f^c(a_k)\right)\right\}, \quad j = 1, 2, \ldots, m.$$

It is again obvious that problems D and D* are special cases of the dual problem. The classical equivalence theorem of statistical cluster analysis can be viewed as the equivalence of the primal and the dual problem for the special convex function $f(x) = \frac{1}{2}||x||^2$. In the the present paper we will show that the known results from cluster analysis can be extended to fairly general convex functions and are thus valid also for the Kohonen problem and related problems.

1.3 Comments on convex functions

In the field of classification our information measure I_f is new as far as general convex functions are concerned. Let us justify the information measure by a few remarks. From the point of view of mathematical statistics a complete discussion of the information measure relies on connections to the theory of majorization, to the Bishop de Leeuw order relation between measures and on the theory of comparison of experiments. (Concerning these topics see Strasser, 1985, [15], and Torgersen, 1991, [16].) A thorough treatment of these connections will be given in a forthcoming paper. In the present paper we will give only some preliminary motivation of our interest in the f-information.

A first remark is concerned with Jensen's inequality. If $\mathcal{B}_1$ and $\mathcal{B}_2$ are two partitions and if $\mathcal{B}_2$ is finer than $\mathcal{B}_1$ (i.e. each set of $\mathcal{B}_1$ is a disjoint union of sets in $\mathcal{B}_2$), then we have $I_f(\mathcal{B}_1) \leq I_f(\mathcal{B}_2)$. Thus, any information gain by splitting sets in a given a partition increases the f-information.

The second remark deals with the relation to statistical cluster analysis. Maximizing $I_f(\mathcal{B})$ with $f(x) = ||x||^2$ is equivalent to minimizing the inner variance of the partition $\mathcal{B}$. However, what is the reason of our interest in generalizing the optimization problem from $f = ||.||^2$ to general convex functions f, and what are the interesting alternatives to $f = ||.||^2$?

We have already mentioned the Kohonen problem D* with $f(x) = ||x||$. This convex function increases much more slowly than the quadratic function. As a consequence the optimal partitions obtained by the Kohonen method are less influenced by the extreme parts of the underlying data distribution. Concerning the dual problem this effect is caused by a more rapid increase of the conjugate convex function f^c which plays here a role as a penalty function. Thus, a slower increase of the convex function f amounts to a more robust partitioning procedure. Robustness of a vector quantization procedure is a recent topic (cf. Cuesta-Albertos, Gordaliza and Matran, [9]).

Although the Kohonen problem with $f(x) = ||x||$ has nice properties regarding robustness the lack of differentiability may cause troubles. An interesting modification of the Kohonen approach has been proposed by Masters, [11], p.330. This author adds to the original data vectors a constant component, say $x^* = (1, x)$, and and then applies Kohonen's procedure. It is possible to show (and is shown below in Discussion (3.11) that such a procedure amounts to applying our primal and dual optimization problem with the convex function $f(x) = \sqrt{1 + ||x||^2}$. Thus, the modified Kohonen problem of Masters, [11], is another example where a special case of our approach turns out to be of recent interest.

The convex function $f(x) = \sqrt{1 + ||x||^2}$ is differentiable at zero and increases as

slowly as $||x||$. There are several further candidates of convex functions sharing these properties. An important case are those convex functions which are most popular in the field of robust statistics, namely

$$f(x) = \begin{cases} ||x||^2/2, & \text{if } ||x|| \leq c, \\ c||x|| - c^2/2, & \text{if } ||x|| > c. \end{cases} \tag{1.16}$$

The results of our paper are sufficiently general to cover convex functions of this kind. These and related examples are considered in detail in section 3.

1.4 Overview over the results

Our first result is the equivalence theorem for the primal and the dual problem (Theorem (2.12)). The theorem is the basis for the following results like the existence theorem (Theorem (2.14)), the consistency theorem (Theorem (2.17)), and for the construction and analysis of an algorithm.

We will present an algorithm which is an iterative method that increases the objective functions of both optimization problems until it ends up at a fixpoint (Definition (2.22)). Our main results on that fixpoint algorithm are concerned both with the dynamical properties of the iteration and with the characterization of the fixpoints. If P is the empirical distribution of a data set then the iteration stops after finitely many steps (Theorem (2.24)). The fixpoints of the procedure are the non-degenerate local maxima of the dual problem (Theorem (2.30)).

For general convex functions f our results are new. In particular, for the case $f(x) = ||x||$ our results are an analysis of the Kohonen problem D* which is popular in the field of artificial neural networks. In the case of the quadratic function $f(x) = ||x||^2$ the essential features of the results are well-known from statistical cluster analysis. However, our general results contain some improvements of the known facts for the quadratic function case. Thus, let us discuss some relations of our theorems to the literature.

Let $f(x) = ||x||^2$. Recall that in this case the primal problem is simply the problem of the minimum variance partition. The dual problem in this case is the principal point problem and the equivalence of these optimization problems is classical knowledge of statistical cluster analysis. A mathematically satisfactory presentation can be found in Bock, [1].

The existence theorem seems to be new also for the case of the quadratic function f. Consistency theorems for the principal point problem have been proved by Pollard,

[12]. In fact, Pollard considers the generalized PPP C* which only for the case of the quadratic function concerns the situation of our main results. But for this case our consistency theorem (2.17) requires considerably weaker conditions.

The fixpoint algorithm for the quadratic case has been proposed for the first time by Fix and Hodges, [5], (cf. Bock, [1]). As is pointed out by Bock, [1], the fixpoint algorithm may end up in periodic loops without stopping. Our version of the fixpoint algorithm, however, is slightly modified such that this complication cannot happen. Since our modification contains a possibly computer intensive part we show by theoretical arguments (Theorem (2.28)) that it is very seldom that one has to make use of the modification.

Some practical remarks

For practical purposes we have to ask how the fixpoint algorithm can be realized on computing machines and what is the quality of the attained fixpoints compared with the global maxima of the optimization problems.

Computing time depends mainly on the size of the data set. We are using implementations on Pentium Pro 200 PCs which work well on data sets up to, say, 10000 records and dimension 30. For considerably larger data sets a useful alternative to the fixpoint algorithm are adaptive algorithms (stochastic gradient methods). However, these gradient algorithms have as attractors the local maxima of the dual problem which according to our Corollary (2.30) are exactly the fixpoints of our iteration procedure.

There are no theoretical results on the quality of the those fixpoints which are typically attained by the fixpoint algorithm. Numerical experiments show that their quality may vary on a large scale. The quality depends on the starting configuration. The question for suitable starting values is presently answered only by experimental experiences. It seems to be true that for unimodal distributions starting with random samples gives good results. For multimodal distributions (several density clusters), however, a random sample as a starting configuration is not a universal recipe.

1.5 The organization of the paper

In section 2 we give a detailed overview over our theoretical results. The proofs are collected in section 4. Section 3 contains examples. In section 5 we provide some concepts of convex analysis for the reader's convenience.

2 Detailed results

In this section we present a detailed and mathematical rigorous overview over our results together with a discussion of leading ideas and relations to the literature. The proofs of our results are collected in section 4.

2.1 The primal optimization problem

Let P be a probability distribution on $\mathbb{R}^d$ having finite first moments. Suppose that $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ is a partition of $\mathbb{R}^d$. Let $|\mathcal{B}|$ be the size of the partition $\mathcal{B}$, i.e. the number of sets in the partition. We may summarize the information contained in the distribution P by computing the mean

$$E(X|B_j) := \frac{1}{P(B_j)} \int_{B_j} X \, dP, \quad j = 1, 2, \ldots, m, \tag{2.1}$$

for every set B_j of the partition. Here, the symbol X denotes that random variable which is defined by $X(x) := x$, $x \in \mathbb{R}^d$. The further statistical analysis is then no longer based on the original random variable X but on the so-called conditional expectation

$$E(X|\mathcal{B}) := \sum_{j=1}^{m} E(X|B_j) \, 1_{B_j}. \tag{2.2}$$

This is a $\mathcal{B}$-measurable function whose values are the means of the sets in the partition.

Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a convex function. In section 1, (1.11), we defined the f-information

$$I_f(\mathcal{B}, P) := \int f\Big(E(X|\mathcal{B}) \Big) \, dP$$

of the partition $\mathcal{B}$ with respect to the distribution P. If it is clear which distribution P is under consideration we will denote for simplicity $I_f(\mathcal{B}) := I_f(\mathcal{B}, P)$. The f-information may be written as a finite sum

$$I_f(\mathcal{B}) = \sum_{j=1}^{m} P(B_j) \, f\Big(E(X|B_j) \Big), \tag{2.3}$$

whence it is clear that it is always well-defined but may be infinite.

Recall the notation

$$I_f^m := \sup_{|\mathcal{B}|=m} I_f(\mathcal{B})$$

It is our goal to analyse this optimization problem which is our primal optimization problem.

The results of this paper are based on the following assumptions on the convex function f. Here $K(f) = \{x \in \mathbb{R}^d : f(x) < \infty\}$ denotes the domain of the convex function f and $K(f)^\circ$ the interior of this set.

(2.4) ASSUMPTION
1. *f is lower semicontinuous.*
2. $E(||X||) < \infty$.
3. *$f(X)$ is P-integrable.*
4. $P(K(f)^\circ) = 1$.

Some comments on these assumptions are given in section 4, Remark (4.1).

2.2 The dual optimization problem

The key for the analysis of the primal optimization problem is the fact that there exists an equivalent but simpler optimization problem which is more accessible. This second optimization problem will be called the dual optimization problem. The first basic result will be the equivalence of the primal and dual problem (Theorem (2.12)). For this recall the concept of the conjugate convex function (Definition (1.13)). For the reader's convenience we have collected some basic facts on convex functions and their conjugates in section 5. The domain of the conjugate convex function f^c will play a fundamental role and therefore we reserve for it the special notation $A := K(f^c)$.

Let us define the objective function of the dual problem.

(2.5) DEFINITION *Suppose that $E(||X||) < \infty$ and let $a \in A^m$. We define*

$$F_m(a) := \int \max_{1 \le k \le m} \left(a_k' X - f^c(a_k) \right) dP. \tag{2.6}$$

The dual optimization problem aims at maximizing the function F_m on the set A^m. The integrand of the function F_m is the pointwise maximum of m linear functions. Considering the sets B_j where the j-th of those linear functions is maximal gives a partition of $\mathbb{R}^d$. Let us give a particular name to partitions of this kind. The name is a matter of convenience and does not claim to be standard terminology.

(2.7) DEFINITION *A partition* $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ *is called a segmentation if there exists* $\boldsymbol{a} \in A^m$ *such that*

$$x \in B_j \quad \Rightarrow \quad a_j' x - f^c(a_j) = \max_{1 \le k \le m} \left(a_k' x - f^c(a_k) \right). \tag{2.8}$$

Let $\mathcal{S}(\boldsymbol{a})$ *be the family of all segmentations associated with* $\boldsymbol{a}$ *and let* $\mathcal{S}_m = \bigcup_{\boldsymbol{a} \in A^m} \mathcal{S}(\boldsymbol{a})$.

It is clear that segmentations are partitions of a very special kind. Segmentations consist of convex polyhedra. For practical purposes it is important to note that segmentations can be completely described by m affine linear functions. Thus, segmentations require only very little memory in computing machines. Moreover, modifying a segmentation amounts to modifying m linear functions. If we want to construct a partitioning algorithm then it is a great advantage if we may reduce our attention to segmentations instead of considering general partitions.

For the special case $f(x) = \|x\|^2$ the dual optimization problem is the well-known principle point problem (cf. Bock, [1], Satz 15.1) which plays a major role in cluster analysis and coding theory. Let us have a closer look to this case.

(2.9) EXAMPLE Suppose that $f(x) = \|x\|^2/2$. Then we have $A = \mathbb{R}^d$ and it is easy to see that

$$F_m(\boldsymbol{a}) = \int \|X\|^2/2 \, dP - \int \min_{1 \le k \le m} \|X - a_k\|^2/2 \, dP. \tag{2.10}$$

A partition $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ satisfying (2.8) is then characterized by

$$x \in B_j \quad \Rightarrow \quad \|x - a_j\| = \min_{1 \le k \le m} \|x - a_k\|. \tag{2.11}$$

Such partitions are called minimum distance partitions (Bock, [1], p. 165), or Voronoi partitions (z.B. Conway und Sloane, [3]).

2.3 The basic theorems

Basic theoretical results of this paper are the equivalence theorem, the existence theorem and the consistency theorem. Let us start with the formulation of the equivalence theorem.

(2.12) THEOREM *(Equivalence theorem)*
Suppose that assumptions (2.4) are fulfilled. Then the following assertions are valid:

1. *The objective functions of the primal and the dual optimization problem have the same suprema:*

$$\sup_{\mathbf{a}\in A^m} F_m(\mathbf{a}) = \sup_{|\mathcal{B}|=m} I_f(\mathcal{B}). \tag{2.13}$$

2. *If $\mathbf{a} \in A^m$ is an optimum of the dual optimization problem then any segmentation $\mathcal{B} \in \mathcal{S}(\mathbf{a})$ is an optimum of the primal optimization problem.*

As a result of the equivalence theorem the optimization of the f-information can be performed within the class of segmentations and by means of maximizing F_m. For the classical case of the quadratic function $f = ||.||^2$ the equivalence theorem can be found in Bock, [1], Satz 15.1.

The second basic assertion is the existence theorem. For empirical distributions P the existence theorem is trivial since there are only finitely many partitions of the supporting set of P. But in general the existence of an optimal partition is not trivial and there are simple examples showing that the conditions (2.4) cannot be omitted.

(2.14) THEOREM *(Existence theorem)*
Suppose that the assumptions (2.4) *are satisfied. Then for every $m \in \mathbb{N}$ there exists an optimal partition, i.e. the primal and the dual optimization problem have solutions.*

There is a consequence of the proof of the existence theorem which is worth to be isolated. It is concerned with properties of the sequence I_f^m and shows that typically this sequence is strictly increasing.

(2.15) COROLLARY *Suppose that the assumptions (2.4) are satisfied. Define*

$$m^* := \min\{m \in \mathbb{N} : I_f^m = I_f^{m+1}\}. \tag{2.16}$$

The following assertions are valid:

1. *Either $m^* = \infty$ and $(I_f^m)_{m\in\mathbb{N}}$ is strictly increasing, or $I_f^{m^*} = I_f^m$ for all $m > m^*$.*

2. *If $m \leq m^*$, then all sets in an optimal partition have positive P-measure.*

3. *The case $m^* < \infty$ happens iff there is a partition $\mathcal{B}$ of size $|\mathcal{B}| = m^*$ such that f is affine linear on each set of the partition.*

The third basic result is a limit theorem saying that optimal partitions for statistical models P can be approximated by Monte Carlo simulation. This result is related to a consistency theorem by Pollard, [12], for the principle point problem which, however, requires considerably stronger regularity conditions.

(2.17) THEOREM *(Consistency theorem)*
Suppose that the assumptions (2.4) are satisfied and that hyperplanes in $\mathbb{R}^d$ *are of* P*-measure zero. Let* $(\hat{P}_n)_{n\in\mathbb{N}}$ *be any sequence of empirical distributions which converges weakly to* P*. If* $m \leq m^*$ *and if* $(\mathcal{B}_n) \subseteq \mathcal{S}_m$ *is a sequence of segmentations such that*

$$\lim_{n\to\infty} \left(I_f(\mathcal{B}_n, \hat{P}_n) - I_f^m(\hat{P}_n) \right) = 0, \tag{2.18}$$

then the sequence $(\mathcal{B}_n)$ *is asymptotically optimal for* P*, i.e.*

$$\lim_{n\to\infty} I_f(\mathcal{B}_n, P) = I_f^m(P). \tag{2.19}$$

2.4 The Fixpoint Method

Now we turn to the discussion of the fixpoint algorithm. The starting point is the equivalence theorem (2.12). We may restrict our attention to segmentations and and have the choice of maximizing either I_f or F_m.

The function $F_m : A^m \to \mathbb{R}$ is neither convex nor concave. In order to find its global maximum we could apply any stochastic search method like simulated annealing. In view of bounded computing time such methods can arrive only at approximate solutions of the problem. But if we are satisfied with an approximate solution then further methods are at our disposal. The fixpoint algorithm is such a method which improves a starting solution step by step until it stops at a fixpoint.

Before we go into the details of the fixpoint algorithm we have to discuss a technical point. We want to find partitions $\mathcal{B}$ of size $|\mathcal{B}| = m$ where $m \leq m^*$ is fixed. Up to this point we have admitted partitions where some of the sets are of measure zero. This will be excluded in the following.

(2.20) DEFINITION *A partition* $\mathcal{B}$ *is called to be degenerate if it contains sets of* P*-measure zero.*

From Corollary (2.15), (2), it follows, that in case of $m \leq m^*$ any optimal partition is non-degenerate. Therefore it makes sense for an algorithm to exclude degenerate partitions from the game. However, while an algorithm is in action, sometimes

degenerate partitions may come across. We have to discuss the question of how to handle such situations. There are several practical possibilities for that purpose. Our proposal is to split a suitable set of positive P-measure.

(2.21) LEMMA *Let $m \leq m^*$ and let $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ be a degenerate partition. Then at least one set with $P(B_k) > 0$ can be splitted in such a way that $B_k = C \cup D$, $P(C) > 0$, $P(D) > 0$, and*

$$P(B_k)f(E(X|B_k)) < P(C)f(E(X|C)) + P(D)f(E(X|D)). \qquad (2.22)$$

It follows that by splitting some set any degenerate partition can be improved in such a way that the f-information $I_f(\mathcal{B})$ is strictly increased.

The fixpoint algorithm is a generalization of the method of k-means clustering. This method has been proposed for the first time by Fix und Hodges, [5], and is a popular method of statistical cluster analysis. We refer to Bock, [1], Abschnitt 15d.

The method of k-means clustering can be described in a simple way: Starting with any point $\boldsymbol{a} \in A^m$ we apply an improvement procedure consisting of two steps. In the first step we choose a suitable segmentation $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$. In the second step we compute the means of the sets in that segmentation. The vector of those means is the improvement of $\boldsymbol{a}$ and takes the role of the starting value for the next iteration.

If we want to apply this procedure to our more general optimization problem then we are faced with a complication. The convex function f and its conjugate convex function f^c are typically different and have different domains. Only in the very special case of $f(x) = ||x||^2/2$ both function coincide which makes the announced complication disappear. In the general case the means of the sets in a partition are elements of the domain of f whereas the function F_m applies to points in the domain of f^c. Therefore, after having performed the second step of the improvement procedure we are not in a position of having obtained a new starting point for the next iteration. We have to insert a third step which turns our mean values into the domain of f^c in such a way that the objective function increases.

A thorough analysis of the proof of the equivalence theorem shows us what has to be done. As a result the derivative of the convex function f comes into the game. (For the case of $f(x) = ||x||^2/2$ the derivative is the identity function which explains its invisibility in the k-means clustering algorithm.) Let $D(f, b)$ be the subdifferential of f at b. From Corollary (4.4) it follows that $D(f, E(X|B)) \neq \emptyset$ whenever $P(B) > 0$.

The following definition contains a description of a single iteration of the fixpoint algorithm consisting of three steps.

(2.23) DEFINITION *Let $\boldsymbol{a} \in A^m$ be any approximate solution of the dual optimization problem. This solution $\boldsymbol{a}$ is improved to a solution $\boldsymbol{b} = T(\boldsymbol{a})$ by the following steps:*

1. *In the first step we choose a segmentation $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ for which $I_f(\mathcal{B})$ is maximal.*
2. *If $\mathcal{B}$ is non-degenerate, the we leave it unchanged: $\mathcal{B}^* := \mathcal{B}$. If $\mathcal{B}$ is degenerate, then we improve $\mathcal{B}$ according to Lemma (2.21) arriving at a non-degenerate partition $\mathcal{B}^*$ with $I_f(\mathcal{B}) < I_f(\mathcal{B}^*)$. We compute the means $E(X|B_k)$ of the sets in the partition $\mathcal{B}^*$.*
3. *In the third step we choose any $b_k \in D(f, E(X|B_k))$ and define $T(\boldsymbol{a}) := \boldsymbol{b} = (b_1, b_2, \ldots, b_m)$.*

The operator T is an improvement procedure. The fixpoint algorithm is the iterative application of this operator. Beginning with a starting value $\boldsymbol{a}_0 \in A^m$ we may apply the operator an arbitrary number of times getting a sequence $\boldsymbol{a}_n := T(\boldsymbol{a}_{n-1})$, $n = 1, 2, \ldots$ It is clear that computing stops if we arrive at a fixpoint $\boldsymbol{a} = T(\boldsymbol{a})$.

The following theorem contains the dynamical properties of the fixpoint algorithm.

(2.24) THEOREM *Suppose that the assumptions (2.4) are satisfied and let $(\boldsymbol{a}_n)$ be a sequence which is obtained by iteration of the operator T.*

1. *Then $F_m(\boldsymbol{a}_{n-1}) \leq F_m(\boldsymbol{a}_n)$ for all $n \in \mathbb{N}$.*
2. *If f is differentiable on $K^\circ(f)$ then we have $F_m(\boldsymbol{a}_{n-1}) < F_m(\boldsymbol{a}_n)$ whenever $\boldsymbol{a}_{n-1} \neq \boldsymbol{a}_n$.*
3. *If f is differentiable on $K^\circ(f)$ and if P is an empirical distribution then the iteration arrives at a fixpoint after finitely many steps.*

A remarkable fact is that the objective function is strictly increasing until the procedure stops. This excludes any periodicity of the algorithm and is due to our definition of Step 1 which is slightly different to the ordinary k-means algorithm where periodicity cannot be excluded (confer Bock, [1], pp. 172–173). The essential point of our formulation of Step 1 is our handling of the situation when $\mathcal{S}(\boldsymbol{a})$ contains more than only one segmentation. In that case we prescribe to choose the best segmentation.

Although choosing the best segmentation in the family $\mathcal{S}(\boldsymbol{a})$ is no problem from the theoretical point of view it might be a problem for practical application since this optimization may take a lot of computing time. It is therefore an important

point to ask how often the case may happen where $\mathcal{S}(\boldsymbol{a})$ contains more than one segmentation.

Let $\boldsymbol{a} \in A^m$ and for $x \in \mathbb{R}^d$ let

$$\alpha(x, \boldsymbol{a}) := \Big\{ j : a_j' x - f^c(a_j) = \max_{1 \le k \le m} \Big(a_k' x - f^c(a_k) \Big) \Big\}. \tag{2.25}$$

The family $\mathcal{S}(\boldsymbol{a})$ contains more than one segmentation iff $|\alpha(x, \boldsymbol{a})| > 1$ for some $x \in \mathbb{R}^d$. Our first result concerning that point is concerned with continuous statistical models and shows that for such models there is only one segmentation in $\mathcal{S}(\boldsymbol{a})$ up to P-negligibility.

(2.26) THEOREM *Suppose that the assumptions (2.4) are satisfied. If $P \ll \lambda_d$ and if $\boldsymbol{a} \in A^m$ consists of pairwise different components then $P\{x \in \mathbb{R}^d : |\alpha(x, \boldsymbol{a})| > 1\} = 0$ and $I_f(\mathcal{B})$ is constant on $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$.*

Our second result to non-uniqueness of $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ is concerned with empirical distributions and is therefore relevant for practical considerations. We will show that uniqueness is the typical situation.

(2.27) DEFINITION *Let P be the empirical distribution of $(x_i)_{1 \le i \le n}$. A vector $\boldsymbol{a} \in A^m$ is said to be in generic position (with respect to P), if $|\alpha(x_i, \boldsymbol{a})| = 1$ for all $i = 1, 2, \ldots, n$.*

It is clear that for $\boldsymbol{a} \in A^m$ in generic position there is exactly one segmentation $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$. The following theorem shows that points $\boldsymbol{a} \in A^m$ in generic position are typical, i.e. their complement is a topologically small set.

(2.28) THEOREM *Let P be an empirical distribution, whose support $(x_i)_{1 \le i \le n}$ is contained in $K^\circ(f)$. Then the following assertions are valid:*

1. *The set of all points $\boldsymbol{a} \in A^m$ in generic position is open in A^m.*
2. *If f is differentiable on $K^\circ(f)$ and if f^c is continuous on A, then the set of all points $\boldsymbol{a} \in A^m$ in generic position is the complement of a nowhere dense set.*

By the preceding results the basic dynamical properties of the fixpoint algorithm are settled. It remains to get an idea of the fixpoints of the algorithm.

Let $\boldsymbol{a} \in A^m$. Let us call $\boldsymbol{h} = (h_1, \ldots, h_m) \in (\mathbb{R}^d)^n$ an admissible direction for $\boldsymbol{a}$, if there is a positive number $\epsilon > 0$, such that $\boldsymbol{a} + \epsilon \boldsymbol{h} \in A^m$. If $\boldsymbol{a}$ is an inner point of A^m then clearly every direction is admissible.

Our final results consider the relation between fixpoints of the algorithm and directional derivatives of the function F_m. It will be shown that fixpoints cannot be improved by gradient methods.

(2.29) THEOREM *Suppose that the assumptions (2.4) are satisfied and let* $\boldsymbol{a} \in A^m$. *Then the following assertions are valid:*

1. *For every admissible direction* $\boldsymbol{h}$ *of* $\boldsymbol{a}$ *there exists*

$$D^+F_m(\boldsymbol{a},\boldsymbol{h}) := \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon}\Big(F_m(\boldsymbol{a}+\epsilon\boldsymbol{h}) - F_m(\boldsymbol{a})\Big).$$

2. *If* $\boldsymbol{a}$ *is a fixpoint then* $D^+(\boldsymbol{a},\boldsymbol{h}) \leq 0$ *for all admissible directions* $\boldsymbol{h}$.

3. *Assume that* f *is differentiable and* $\mathcal{S}(\boldsymbol{a})$ *contains only non-degenerate partitions. If* $D^+F_m(\boldsymbol{a},\boldsymbol{h}) \leq 0$ *for all admissible directions* $\boldsymbol{h}$, *then* $\boldsymbol{a}$ *is a fixpoint.*

For points in generic position the result can be improved considerably.

(2.30) COROLLARY *Suppose that the assumptions (2.4) are satisfied and that* P *is an empirical distribution. Let* $\boldsymbol{a} \in A^m$ *be in generic position. Then the following assertions are valid:*

1. *If* $\boldsymbol{a}$ *is a fixpoint then it is a local maximum of* F_m.

2. *Assume that* f *is differentiable. If* $\boldsymbol{a}$ *is a non-degenerate local maximum of* F_m *then it is a fixpoint.*

3 Examples

In this section we will consider some examples of convex functions and compute explicitely the computational form of the primal and dual optimization problem.
We will restrict our interest to convex functions of the form $f(x) = \Phi(||x||)$ where $\Phi : [0,\infty) \to [0,\infty)$ is a convex function satisfying $\Phi(0) = 0$. In such cases the derivative

$$\phi(x) = \lim_{\epsilon \downarrow 0} \frac{\Phi(x+\epsilon) - \Phi(x)}{\epsilon}$$

is well defined and we have

$$\Phi(x) = \int_0^x \phi(s)\,ds, \quad x \geq 0.$$

The properties of the derivative ϕ play a substantial role for the behaviour of the compression procedure define by f. We will define the various examples of convex function by fixing the derivative ϕ of Φ.

(3.1) EXAMPLE If $\phi(s) \equiv 1$, then $\Phi(x) = x$ and we get $f(x) = ||x||$. This is the convex function f of the Kohonen problem D*.

(3.2) EXAMPLE If $\phi(s) = s$ then $\Phi(x) = x^2/2$ and $f(x) = ||x||^2/2$. For this choice of Φ we obtain the situation which is typical for statistical cluster analysis and the MVP-problem.

(3.3) EXAMPLE Both examples (3.1) and (3.2) are special cases of a general type of convex function. Defining $\phi(s) = s^{p-1}$ with $1 \leq p \leq 2$ we arrive at $\Phi(x) = x^p/p$ and $f(x) = ||x||^p/p$.

In case of example (3.2) and of example (3.3) with $1 < p \leq 2$ the derivative ϕ is not bounded. However, for robustness of the compression procedure a bounded derivative is highly desirable, and this is fulfilled in example (3.1). Our next examples have bounded derivatives, too.

(3.4) EXAMPLE Let $\phi(s) = s/\sqrt{1+s^2}$. Then we have $\Phi(x) = \sqrt{1+x^2} - 1$ and hence $f(x) = \sqrt{1+||x||^2} - 1$. As mentioned before and shown below this convex function corresponds to the modified Kohonen method of Masters, [11]:

(3.5) EXAMPLE In the field of robust statistics Huber proposed influence functions for the construction of robust tests and estimates according to

$$\phi(s) = \begin{cases} s & s \leq 1, \\ 1 & s > 1. \end{cases}$$

Such functions viewed as derivatives of convex function lead us to

$$\Phi(s) = \begin{cases} x^2/2 & x \leq 1, \\ x - 1/2 & x > 1, \end{cases} \qquad f(x) = \begin{cases} ||x||^2/2 & ||x|| \leq 1, \\ ||x|| - 1/2 & ||x|| > 1. \end{cases}$$

(3.6) EXAMPLE Activating functions of neural networks are often defined as

$$\phi(s) = \frac{e^s - 1}{e^s + 1}.$$

These functions lead to

$$\Phi(x) = 2 \ln \cosh \frac{x}{2}, \quad f(x) = 2 \ln \cosh \frac{||x||}{2}.$$

The convex functions of examples (3.1) to (3.3) are homogeneous functions, i.e. $f(\lambda x) = \lambda^r f(x)$ for some $r > 0$. Therefore a scaling factor applied to the data does not affect the compression procedure. A completeley different situation arises with the non-homogeneous functions of examples (3.4) to (3.6). Here any scaling of the data changes the properties of the compression result drastically. It is obvious that compression is the more robust the greater the part of the data which corresponds to the saturation level of the derivative ϕ. Scaling the data with a factor $\lambda > 1$ makes the method more robust. Smaller scale factors move the procedure towards usual vector quantization (in the sense of Problem C).

(3.7) EXAMPLE Let $\phi_1 : \mathbb{R} \to \mathbb{R}$ be increasing and such that $\phi_1(0) = 0$. Defining $\phi_\lambda(s) = \phi_1(\lambda s)$, $\lambda > 0$, we obtain the convex functions

$$\Phi_\lambda(x) = \frac{1}{\lambda}\Phi_1(\lambda x), \qquad f_\lambda(x) = \frac{1}{\lambda} f_1(\lambda x).$$

In the following we will explicitely state the primal and dual optimization problems corresponding to our examples of convex functions, as well as the derivatives which are required for step 3 of the fixpoint iteration.

(3.8) DISCUSSION Let $\phi(s) = 1$, $\Phi(x) = x$ and $f(x) = ||x||$. The primal problem is the optimization problem

$$\int ||E(X|\mathcal{B})||\, dP = \text{Max !}$$

For the dual problem we require the conjugate convex function f^c. From

$$a'x - ||x|| \leq 0 \quad \text{if } ||a|| \leq 1,\ x \in \mathbb{R}^d,$$

and

$$\sup_x (a'x - ||x||) \geq a'(\lambda a) - \lambda||a|| = \lambda(||a||^2 - ||a||), \lambda > 0,$$

we obtain that

$$f^c(a) = \begin{cases} 0 & ||a|| \leq 1, \\ \infty & ||a|| > 1. \end{cases}$$

This implies that for the optimization of $a'x - f^c(a)$ we may restrict our attention to $||a|| \leq 1$. For these vectors a we have $f^c(a) = 0$ and the dual problem amounts to

$$F(a_1, \ldots, a_m) = \int \max_j a_j' X\, dP = \text{Max !} \quad \text{for } ||a_j|| \leq 1,\ j = 1, 2, \ldots, m.$$

For performing step 3 of the fixpoint iteration we require the derivative of f. This is

$$D(f,x) = \frac{x}{||x||} \quad \text{whenever } x \neq 0.$$

If $x = 0$ then we may choose any slope of a support function, e.g. $D(f,0) = 0$.

(3.9) DISCUSSION Let us consider the case of $\phi(s) = s^{p-1}$ for $1 < p \leq 2$. This case covers examples (3.2) and (3.3). The primal problem is

$$\int ||E(X|\mathcal{B})||^p \, dP = \text{Max !}$$

For the dual problem we have to compute the conjugate convex function. Let $q > 0$ be such that $1/p + 1/q = 1$. Young's inequality gives

$$st \leq \frac{s^p}{p} + \frac{t^q}{q},$$

where equality holds iff $t = s^{p-1}$ or $s = t^{q-1}$. This implies that $f^c(a) = ||a||^q/q$. Hence, the dual problem is given by

$$F(a_1, \ldots, a_m) = \int \max_j \left(a_j' X - ||a_j||^q/q \right) dP = \text{Max !}$$

The derivative of f which we need for step 3 of the fixpoint iteration is given by

$$D(f,x) = ||x||^{p-2} x \quad \text{whenever } x \neq 0.$$

If $x = 0$ and $p < 2$ then we may choose any slope of a support function, e.g. $D(f,0) = 0$.

(3.10) DISCUSSION Let $\phi(s) = s/\sqrt{1+s^2}$ and $\Phi(x) = \sqrt{1+x^2} - 1$. Then the primal problem is

$$\int \sqrt{1 + ||E(X|\mathcal{B})||^2} \, dP = \text{Max !}$$

For the dual problem we compute the conjugate convex function. The inverse function of $t = \phi(s)$ is $s = \psi(t) = t/\sqrt{1-t^2}$ if $|t| < 1$. The primitive of ψ is

$$\Psi(y) = \int_0^y \psi(t) \, dt = 1 - \sqrt{1-y^2} \quad \text{if } |y| \leq 1.$$

From Young's inequality

$$xy \leq \Phi(x) + \Psi(y)$$

we obtain the conjugate convex function as

$$f^c(a) = \begin{cases} 1 - \sqrt{1 - ||a||^2} & ||a|| \leq 1, \\ \infty & ||a|| > 1. \end{cases}$$

Therefore the dual problem is given by

$$F(a_1, \ldots, a_m) = \int \max_j \left(a_j' X + \sqrt{1 - ||a_j||^2} \right) dP = \text{Max !}$$

if $||a_j|| \leq 1$, $j = 1, 2, \ldots, m$.
The derivative of f is

$$D(f, x) = \phi(||x||) \frac{x}{||x||} = \frac{x}{\sqrt{1 + ||x||^2}}.$$

At this point we leave our presentation of examples for a moment and show that the modified Kohonen problem proposed by Masters, [11], is equivalent to the problem considered in example (3.4) and discussion (3.10).

(3.11) DISCUSSION Masters, [11], proposes a modification of the Kohonen method augmenting data $x \in \mathbb{R}^d$ to $x^* = (1, x) \in \mathbb{R}^{d+1}$ and applying Kohonen's method to the augmented data. Denoting $X^* = (1, X)$ the primal problem is

$$\int ||E(X^*|\mathcal{B})|| \, dP = \text{Max !}$$

which by $||(1, x)|| = \sqrt{1 + ||x||^2}$ is equivalent to the primal problem of discussion (3.10). However, this is not sufficient to prove the asserted equivalence. We rather have to show that Kohonen's method applied to X^* gives the fixpoint method for the dual problem of discussion (3.10). Kohonen's method consists of two steps:
1. For any given $a_j^* \in \mathbb{R}^{d+1}$ with $||a_j^*|| = 1$ define a partition according to

$$B_j^* = \{x^* = (1, x) \in \mathbb{R}^{d+1} : j = \text{argmax}_k a_k^* x^*\}, \quad j = 1, 2, \ldots, m.$$

2. For any given partition $\mathcal{B}^* = (B_1^*, \ldots, B_m^*)$ of $\mathbb{R}^{d+1}$ define vectors

$$a_j^* = \frac{\overline{x^*}_{B_j^*}}{||\overline{x^*}_{B_j^*}||}, \quad j = 1, 2, \ldots, m.$$

Translating step 1 into vectors in $\mathbb{R}^d$ we obtain:

1'. For any given $a_j \in \mathbb{R}^d$ with $||a_j^*|| \leq 1$ define a partition according to

$$B_j = \{x \in \mathbb{R}^d : j = \text{argmax}_k \Big(a_k' x + \sqrt{1 - ||a_k||^2} \}, \quad j = 1, 2, \ldots, m.$$

This is exactly the recipe of constructing a partition for the dual problem in discussion (3.10).

Translating step 2 into vectors in $\mathbb{R}^d$ gives:

2'. For any given partition $\mathcal{B} = (B_1, \ldots, B_m)$ of $\mathbb{R}^d$ define vectors

$$a_j = \frac{\overline{x}_{B_j}}{\sqrt{1 + ||\overline{x}_{B_j}||^2}}, \quad j = 1, 2, \ldots, m.$$

From the formula for the derivative in discussion (3.10) we see that this is step 3 of the fixpoint iteration in the case of discussion (3.10).

(3.12) DISCUSSION Let us turn to the case of example (3.5). The primal problem is

$$\int \Phi(||X||)\, dP = \text{Max !}$$

where Φ is the function defined in example (3.5). By similar methods as in discussion (3.10) it can be shown that $f^c(a) = \Psi(||a||)$ where

$$\Psi(y) = \begin{cases} y^2/2 & y \leq 1, \\ \infty & y > 1. \end{cases}$$

For the derivative we have

$$D(f, x) = \begin{cases} x & ||x|| \leq 1, \\ x/||x|| & ||x|| > 1. \end{cases}$$

(3.13) DISCUSSION Finally we discuss example (3.6). The primal problem is

$$\int 2 \ln \cosh \frac{||X||}{2}\, dP = \text{Max !}$$

By similar method as in discussion we may show that $f^c(a) = \Psi(||a||)$ where

$$\Psi(y) = \begin{cases} (1+y)\ln(1+y) + (1-y)\ln(1-y) & y \leq 1, \\ \infty & y > 1. \end{cases}$$

For the derivative we have

$$D(f,x) = \frac{e^{||x||}-1}{e^{||x||}+1}\frac{x}{||x||}.$$

Examples (3.1), (3.4) to (3.6), are of special interest. These convex functions have bounded derivatives and define robust compression procedures. The vectors a_j which are the basis of the dual problem and of the fixpoint iteration are for these convex functions bounded to the unit ball. However, there is a difference between cases (3.1) and (3.5) on one side and cases (3.4) and (3.6) on the other side. In case (3.1) (Kohonen method) the vectors a_j are even restricted to the boundary of the unit ball, i.e. the unit sphere. In case (3.5) (Huber's influence functions) at least some of the vectors a_j are elements of the boundary. This is not the case with examples (3.4) and (3.6). Since for these cases the penalty term Ψ of the dual problem has derivative ∞ at the boundary the vectors a_j are kept away from the boundary during the algorithm.

4 Proofs

Let us start with a few remarks on the assumptions (2.4).

(4.1) REMARK Conditions (2.4), (1)-(3), are not severe and do not restrict the applicability of our results. Things are slightly different with condition (2.4), (4). This condition is certainly satisfied if $K(f) = \mathbb{R}^d$. However, if $K(f) \neq \mathbb{R}^d$, then the properties of f on the boundary $\partial K(f)$ play a major role if P puts some mass on this boundary. Condition (4) avoids such complications. If we deal with situations where condition (4) is violated then we need additional information concerning the properties of f on the boundary $\partial K(f)$. In this paper such situations are not considered.

A basic consequence of the assumptions (2.4) is contained in Corollary (4.4) below. It is prepared by the following lemma on conditional expectations.

(4.2) LEMMA *Let $E(||X||) < \infty$ and let $C \subseteq \mathbb{R}^d$ be a convex set with $P(C^\circ) = 1$. If $B_n \subseteq C$, $n \in \mathbb{N}$ are measurable subsets then*

$$\lim_{n\to\infty} E(X|B_n) \to a \in \mathbb{R}^d \setminus C^\circ \quad \textit{implies} \quad \lim_{n\to\infty} P(B_n) = 0. \tag{4.3}$$

Proof: W.l.g. we may assume that $a_1 = 0$ and $C \subseteq \mathbb{R}_+ \times \mathbb{R}^{d-1}$. For every $\epsilon > 0$ let $H_\epsilon := [\epsilon, \infty) \times \mathbb{R}^{d-1}$. The following inequalities are obvious:

$$\begin{aligned} ||x - a|| &\geq |x_1 - a_1|, && \text{if } x \in C, \\ ||x - a|| &\geq |x_1 - a_1| \geq \epsilon, && \text{if } x \in C \cap H_\epsilon, \\ |x_1 - a_1| &< \epsilon, && \text{if } x \in B \setminus H_\epsilon. \end{aligned}$$

This implies

$$\begin{aligned} ||a - E(X|B_n)|| &\geq |a_1 - E(X|B_n)_1| \\ &= \left| a_1 - E(X|B_n \cap H_\epsilon)_1 \frac{P(B_n \cap H_\epsilon)}{P(B_n)} - E(X|B_n \setminus H_\epsilon)_1 \frac{P(B_n \setminus H_\epsilon)}{P(B_n)} \right| \\ &\geq \left| a_1 - E(X|B_n \cap H_\epsilon)_1 \right| \frac{P(B_n \cap H_\epsilon)}{P(B_n)} - \left| a_1 - E(X|B_n \setminus H_\epsilon)_1 \right| \frac{P(B_n \setminus H_\epsilon)}{P(B_n)} \\ &\geq \epsilon \Big(\frac{P(B_n \cap H_\epsilon)}{P(B_n)} - \frac{P(B_n \setminus H_\epsilon)}{P(B_n)} \Big) \\ &= \epsilon \Big(2 \frac{P(B_n \cap H_\epsilon)}{P(B_n)} - 1 \Big) \end{aligned}$$

and therefore

$$\limsup_{n \to \infty} \frac{P(B_n \cap H_\epsilon)}{P(B_n)} \leq \frac{1}{2}.$$

It follows that $\limsup_{n\to\infty} P(B_n) \leq 2P(H'_\epsilon)$ for all $\epsilon > 0$. From $P(C^\circ) = 1$ we obtain that $\lim_{n\to\infty} P(B_n) = 0$. □

(4.4) COROLLARY *Suppose that the assumptions (2.4) are satisfied. Then we have* $E(X|B) \in K^\circ(f)$ *for all Borel sets* B *with* $P(B) > 0$.

From Corollary (4.4) it follows that the convex function f is finite and continuous at every $x = E(X|B)$ with $P(B) > 0$, and that the set of support functions is not empty: $S(f, x) \neq \emptyset$. If $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ is an arbitrary partition then the f-information defined by (1.11) and (2.3) is real-valued.

Let us turn to the Equivalence theorem (2.12). The proof is based on two simple but basic inequalities which are isolated for subsequent quotation.

(4.5) LEMMA *Suppose that the assumptions (2.4) are satisfied. If* $\boldsymbol{a} \in A^m$ *then* $F(\boldsymbol{a}) \leq I_f(\mathcal{B})$ *for all* $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$.

Proof: Let $\boldsymbol{a} \in A^m$ and $\mathcal{B} = (B_1, B_2, \ldots, B_m) \in \mathcal{S}(\boldsymbol{a})$. Then we have

$$\begin{aligned} F(\boldsymbol{a}) &= \int \max_{1\le k\le m} \Big(a_k' X - f^c(a_k)\Big)\, dP \\ &= \sum_{k=1}^m \int_{B_k} \Big(a_k' X - f^c(a_k)\Big)\, dP = \sum_{k=1}^m P(B_k)\Big(a_k' E(X|B_k) - f^c(a_k)\Big) \\ &\le \sum_{k=1}^m P(B_k) f\Big(E(X|B_k)\Big) = I_f(\mathcal{B}). \end{aligned}$$

□

(4.6) LEMMA *Suppose that the assumptions (2.4) are satisfied. Let* $\mathcal{B} = (B_1, ..., B_m)$ *be any partition and assume that* $a_j \in D(f, E(X|B_j))$ *for every* $j = 1, 2, \ldots, m$. *Denoting* $\boldsymbol{a} := (a_1, a_2, \ldots, a_m)$ *we have* $I_f(\mathcal{B}) \le F(\boldsymbol{a})$.

Proof: Let $\ell_k(x) := a_k' x - f^c(a_k)$. Since $\ell_k(E(X|B_k)) = f(E(X|B_k))$ it follows that

$$\begin{aligned} I_f(\mathcal{B}) &= \sum_{k=1}^m P(B_k) f\Big(E(X|B_k)\Big) \\ &= \sum_{k=1}^m P(B_k) \ell_k\Big(E(X|B_k)\Big) = \sum_{k=1}^m \int_{B_k} \ell_k(X)\, dP \\ &\le \int \max_{1\le k\le m} \Big(a_k' X - f^c(a_k)\Big)\, dP = F(\boldsymbol{a}). \end{aligned}$$

□

Proof: (of the Equivalence theorem (2.12).) Theorem (2.12) follows immediately by application of Lemmas (4.5) and (4.6). □

The proofs of the Existence theorem (2.14) and of the Consistency theorem (2.17) require additional auxiliary lemmas.

(4.7) COROLLARY *Suppose that the assumptions (2.4) are satisfied. Then the condition* $\liminf_{n\to\infty} P(B_n) > 0$ *implies that the sequence of means* $(E(X|B_n))_{n\in\mathbb{N}}$ *is bounded and that all accumulation points of that sequence are contained in* $K(f)^\circ$.

Proof: Boundedness of the sequence is implied by

$$\|E(X|B)\| \le \frac{1}{P(B)} \int_B \|X\|\, dP \le \frac{1}{P(B)} \int \|X\|\, dP.$$

The second part of the assertion follows from (4.2). □

(4.8) LEMMA *Suppose that the assumptions (2.4) are satisfied. Let $m \in \mathbb{N}$ be such that $m \leq m^* := \min\{k : I_f^k = I_f^{k+1}\}$. If $(\mathcal{B}^n)$ is a sequence of partitions with $|\mathcal{B}_n| = m$, $n \in \mathbb{N}$ and such that $\lim_{n\in\mathbb{N}} I_f(\mathcal{B}_n) = I_f^m$ then*

$$\liminf_{n\to\infty} \min_{1\leq j\leq m} P(B_j^n) > 0. \tag{4.9}$$

Proof: Assume that there is a subsequence $\mathbb{N}_0 \subseteq \mathbb{N}$ and an index j_1 such that

$$\lim_{n\in\mathbb{N}_0} P(B_{j_1}^n) = 0.$$

For each $n \in \mathbb{N}$ there exists a number j with $P(B_j^n) \geq 1/m$. W.l.g. we may assume that

$$\lim_{n\in\mathbb{N}_0} P(B_1^n) = 0, \quad \liminf_{n\in\mathbb{N}_0} P(B_2^n) \geq \frac{1}{m}.$$

By Corollary (4.7) it follows that the sequence $(E(X|B_2^n))_{n\in\mathbb{N}_0}$ is bounded and all accumulation point are contained in $K(f)^\circ$. Choosing a further subsequence $\mathbb{N}_1 \subseteq \mathbb{N}_0$ we arrive at

$$\lim_{n\in\mathbb{N}_1} E(X|B_2^n) =: a \in K^\circ(f)$$

which implies

$$\lim_{n\in\mathbb{N}_1} f\Big(E(X|B_2^n)\Big) = \lim_{n\in\mathbb{N}_1} f\Big(E(X|B_1^n \cup B_2^n)\Big) = f(a).$$

Now we construct a new sequence of partitions $\mathcal{C}_n$ by

$$\mathcal{C}_n := (B_1^n \cup B_2^n, B_3^n, \ldots, B_m^n).$$

We will show that

$$\lim_{n\in\mathbb{N}_1} I_f(\mathcal{B}_n) - I_f(\mathcal{C}_n) = 0, \tag{4.10}$$

which is a contradiction since the inequality $m \leq m^*$ implies

$$\limsup_{n\to\infty} I_f(\mathcal{C}_n) \leq I_f^{m-1} < I_f^m = \lim_{n\to\infty} I_f(\mathcal{B}_n).$$

Let us prove (4.10). We have

$$\begin{aligned} &I_f(\mathcal{B}_n) - I_f(\mathcal{C}_n) \\ &= P(B_1^n) f\Big(E(X|B_1^n)\Big) + P(B_2^n) f\Big(E(X|B_2^n)\Big) \\ &\quad - P(B_1^n \cup B_2^n) f\Big(E(X|B_1^n \cup B_2^n)\Big). \end{aligned}$$

From $\lim_{n\in\mathbb{N}_1} P(B_1^n) = 0$ it follows

$$\lim_{n\in\mathbb{N}_1} \Big(P(B_2^n) - P(B_1^n \cup B_2^n)\Big) = 0$$

and

$$\limsup_{n\in\mathbb{N}_1} P(B_1^n)\, f\Big(E(X|B_1^n)\Big) \le \lim_{n\in\mathbb{N}_1} \int_{B_1^n} f(X)\, dP = 0.$$

This gives (4.10). □

(4.11) LEMMA *Let P be any distribution on $\mathbb{R}^d$. Then for P-almost all $x \in \mathbb{R}^d$ we have $P(B(x,r)) > 0$ whenever $r > 0$.*

Proof: Let $M_r := \{y \in \mathbb{R}^d : P(B(y,r)) = 0\}$ and $M := \bigcup_{r>0} M_r$. By the Lindelöf-property of $\mathbb{R}^d$ each M_r can be covered by a countable number of open balls with P-measure zero. This implies $P(M_r) = 0$ and hence $P(M) = 0$. □

(4.12) LEMMA *Suppose that X is P-integrable and let $g : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. If $g(E(X|B)) = 0$ for all Borel sets B with $P(B) > 0$ then $g(x) = 0$ at P-almost all continuity points of g.*

Proof: Let x be a continuity point of g satisfying $P(B(x,\epsilon)) > 0$ for every $\epsilon > 0$. This implies $\lim_{n\to\infty} E(X|B(x,1/n)) = x$ and hence $\lim_{n\to\infty} g(E(X|B(x,1/n)) = g(x)$. Thus we have $g(x) = 0$. By Lemma (4.11) P-almost all continuity points of g share this property. □

The next lemma contains another important step for the proof of the Existence theorem.

(4.13) LEMMA *Suppose that the assumptions (2.4) are satisfied. If for all Borel sets B with $P(B) > 0$ the equation*

$$f(E(X)) = P(B)\, f(E(X|B)) + P(B')\, f(E(X|B')), \tag{4.14}$$

is valid then f is P-almost surely affine linear.

Proof: Let $\ell \in S(f, E(X))$, i.e. $\ell(y) = f(E(X)) + a'(y - E(X))$. It is easy to see that for each Borel set B with $P(B) > 0$

$$P(B')\, a'\,(E(X|B') - E(X)) = -P(B)\, a'\,(E(X|B) - E(X)).$$

This implies

$$\begin{aligned} f(E(X)) &\geq P(B)\, f(E(X|B)) + P(B')\ell(E(X|B')) \\ &= P(B)\, f(E(X|B)) + P(B')\, f(E(X)) + P(B')\, a'\,(E(X|B') - E(X)) \\ &= P(B)\, f(E(X|B)) + P(B')\, f(E(X)) - P(B)\, a'\,(E(X|B) - E(X)) \\ &= P(B)\, f(E(X|B)) + f(E(X)) - P(B)\,\ell(E(X|B)). \end{aligned}$$

Thus, from $P(B) > 0$ it follows that $f(E(X|B)) \leq \ell(E(X|B))$ which implies $f(E(X|B)) = \ell(E(X|B))$. Applying Lemma (4.12) gives the assertion. □

Now we are in the position to prove the Existence theorem (2.14).

Proof: (of the Existence theorem (2.14) and of Corollary (2.15)) The proof is divided into two parts. In the first part we assume that $m \in \mathbb{N}$ satsfies the inequality

$$m \leq m^* := \min\{k \in \mathbb{N} : I_f^k = I_f^{k+1}\}.$$

If $(\mathcal{B}_n)$ is a sequence of partitions such that $|\mathcal{B}_n| = m$ and $I_f(\mathcal{B}_n) \to I_f^m$ then Lemma (4.8) and Corollary (4.7) imply that for each $k = 1, 2, \ldots, m$ the sequences of means $x_k^n := E(X|B_k^n)$ have accumulation points in $x_k \in K^\circ(f)$. W.l.g. we assume that these sequences of means converge. Let $a_k^n \in D(f, E(X|B_k^n))$ and define $\boldsymbol{a}^n := (a_1^n, a_2^n, \ldots, a_m^n)$. By Lemma (5.5) we may assume that $a_k^n \to a_k \in A$ for every $k = 1, 2, \ldots, m$. Let $\boldsymbol{a} := (a_1, a_2, \ldots, a_m)$ and $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$. It follows that

$$I_f^m = \lim_{n\to\infty} F_m(\boldsymbol{a}_n) = F_m(\boldsymbol{a}) \leq I_f(\mathcal{B}) \leq I_f^m.$$

Thus, $\mathcal{B}$ is an optimal partition and $\boldsymbol{a}$ is a maximum of F_m.

In the second part of the proof we will show that in case of $m^* < \infty$ the equation $I_f^m = I_f^{m+1}$ is valid for all $m \geq m^*$. Each optimal partition of size m^* is then also an optimal partition of size $m \geq m^*$.

Let $m^* < \infty$. If $|\mathcal{B}| = m^*$ and $I_f(\mathcal{B}) = I_f^{m^*}$, then in view of $I_f^{m^*} = I_f^{m^*+1}$ no further splitting of sets $\mathcal{B}$ leads to any increase of I_f. By Lemma (4.13) it follows that f is affine linear P-a.e. on each set of $\mathcal{B}$. Therefore there do not exist any other partitions whose f-Information is larger than $I_f^{m^*} = I_f(\mathcal{B})$. □

An important consequence is concerned with the convergence of approximately optimal partitions.

(4.15) COROLLARY *Suppose that the assumptions (2.4) are satisfied and let* $m \leq m^*$. *If* $(\mathcal{B}_n)$ *is an approximately optimal sequence of partitions of size* m, *i.e.*

$$\lim_{n\to\infty} I_f(\mathcal{B}_n) = I_f^m, \tag{4.16}$$

then all sequences $(a^n)_{n\in\mathbb{N}}$ *in* A^m *satisfying* $a_k^n \in D(f, E(X|B_k^n))$ *are bounded and all accumulation points of such sequences are maxima of* F_m.

The following lemma prepares the proof of the Consistency theorem (2.17).

(4.17) LEMMA *Suppose that the assumptions (2.4) are satisfied and assume that* $P \ll \lambda_d$. *If* $(X_n)_{n\in\mathbb{N}}$ *are independent and* P*-distributed random variables and* $(\hat{P}_n)_{n\in\mathbb{N}}$ *denotes the associated sequence of empirical distributions then*

$$\lim_{n\to\infty} \sup_{\mathcal{B}\in\mathcal{S}_m} \left| I_f(\mathcal{B}, \hat{P}_n) - I_f(\mathcal{B}, P) \right| = 0. \tag{4.18}$$

Proof: Let $\epsilon > 0$ be any positive number.

Let $\mathcal{C}$ the family of all convex polyhedra which are defined by at most m hyperplanes. This is a P-uniform class, and by the law of large numbers there is some $N(\epsilon) =: N \in \mathbb{N}$ such that for $n \geq N$ and uniformly for all $B \in \mathcal{C}$

$$\begin{aligned}
&|\hat{P}_n(B) - P(B)| < \epsilon, \\
&\left| \int_B ||X||\, \hat{P}_n - \int_B ||X||\, dP \right| < \epsilon, \\
&\left| \int_B f(X)\, \hat{P}_n - \int_B f(X)\, dP \right| < \epsilon.
\end{aligned}$$

Chosse $\delta(\epsilon) =: \delta > 0$ in such a way that

$$P(B) < \delta \quad \Rightarrow \quad \int_B f(X)\, dP < \epsilon.$$

Then $P(B) < \delta$ and $n \geq N$ imply

$$\begin{aligned}
P(B)\, f(E(X|B)) \;&<\; \int_B f(X)\, dP < \epsilon, \\
\hat{P}_n(B) f(\hat{E}_n(X|B)) \;&\leq\; \int_B f(X)\, d\hat{P}_n < \int_B f(X)\, dP + \epsilon < 2\epsilon.
\end{aligned}$$

Thus we obtain for any segmentation $\mathcal{B} \in \mathcal{S}_m$

$$\begin{aligned}
&|I_f(\mathcal{B}, \hat{P}_n) - I_f(\mathcal{B}, P)| \\
&\quad\leq\; m \sup_{B\in\mathcal{C}} \left| \hat{P}_n(B)\, f\Big(\hat{E}_n(X|B)\Big) - P(B) f\Big(E(X|B)\Big) \right| \\
&\quad\leq\; m \Big(\sup_{B\in\mathcal{C}:\, P(B)\geq\delta} \left| \hat{P}_n(B) f\Big(\hat{E}_n(X|B)\Big) - P(B)\, f\Big(E(X|B)\Big) \right| + 3\epsilon \Big).
\end{aligned}$$

In order to prove the assertion we have to show that

$$\lim_{n\to\infty} \sup_{B\in\mathcal{C}:\, P(B)\geq\delta} \left| \hat{P}_n(B) f\Big(\hat{E}_n(X|B)\Big) - P(B) f\Big(E(X|B)\Big) \right| = 0.$$

For this it is sufficient to show that f is uniformly continuous on the set of all means occurring in this equation. Let us define

$$\begin{aligned} C_\delta &:= \{E(X|B) : B \in \mathcal{C},\ P(B) \geq \delta\}, \\ C_{n,\delta} &:= \{\hat{E}_n(X|B) : B \in \mathcal{C},\ P(B) \geq \delta\}. \end{aligned}$$

From Lemma (4.2) it follows that $\overline{C_\delta}$ is a compact set which is contained in $K^\circ(f)$. Let M be another compact set such that $\overline{C}_\delta \subseteq M^\circ$ and $M \subseteq K^\circ(f)$. Then for sufficiently large $n \in \mathbb{N}$ we have $\overline{C}_{n,\delta} \subseteq M^\circ$. Since f is continuous on $K^\circ(f)$ it is even uniformly continuous on M. □

Proof: (of the Consistency theorem (2.17)) This is an immediate consequence of (4.17). □

Let us turn to the proofs concerning the fixpoint algorithm.

Proof: (of Lemma (2.21)) This lemma follows from Lemma (4.13) and Corollary (2.15). □

(4.19) LEMMA *Suppose that the assumptions (2.4) are satisfied and let* $\boldsymbol{a} \in A^m$.

1. *If* $\boldsymbol{a}$ *is a fixpoint, for all* $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ *the following is true:* $\mathcal{B}$ *is non-degenerate and* $a_k \in D(f, E(X|B_k))$ *for all* $B_k \in \mathcal{B}$.

2. *If* f *is differentiable on* $K^\circ(f)$ *then the converse of (1) is valid, too.*

Proof: Let $\boldsymbol{a}$ be a fixpoint. Then from

$$F_m(\boldsymbol{a}) \leq \sup_{\mathcal{B}\in\mathcal{S}(\boldsymbol{a})} I_f(\mathcal{B}) \leq I_f(\mathcal{B}^*) \leq F_m(T(\boldsymbol{a})) = F_m(\boldsymbol{a}),$$

it follows that in this chain equality holds everywhere. Thus, $\mathcal{S}(\boldsymbol{a})$ cannot contain degenerate partitions since otherwise an improvement were possible with step 2 of the iteration. Moreover, the equation

$$F_m(\boldsymbol{a}) = \sup_{\mathcal{B}\in\mathcal{S}(\boldsymbol{a})} I_f(\mathcal{B})$$

implies that in Lemma (4.5) equality holds which is only possible if

$$P(B_k)\left(a_k' E(X|B_k) - f^c(a_k)\right) = P(B_k)\, f\Big(E(X|B_k)\Big)$$

for all $B_k \in \mathcal{B} \in \mathcal{S}(\boldsymbol{a})$. Since $P(B_k) > 0$ part 1 of the assertion follows.

Let $\boldsymbol{a} \in A^m$ be such that no partition $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ is degenerate and that $a_k \in D(f, E(X|B_k))$ for all $B_k \in \mathcal{B} \in \mathcal{S}(\boldsymbol{a})$. If f is differentiable on $K^\circ(f)$ then $a_k \in D(f, E(X|B_k))$ is uniquely determined and the third step of the iteration procedure (2.23) leads to $\boldsymbol{a} = T(\boldsymbol{a})$. □

Proof: (of Theorem (2.24)) By Lemmas (4.5) and (4.6) it is clear that steps 1 and 3 of the procedure (2.23) the inequalitites $F_m(\boldsymbol{a}) \le I_f(\mathcal{B})$ and $I_f\mathcal{B}^*) \le F_m(\boldsymbol{b})$ are valid. This proves part 1 of the assertion.

Suppose that $\boldsymbol{a}$ ist not a fixpoint. Let us show that in this case either step 1 or step 2 yield a strict inequality in the chain

$$F_m(\boldsymbol{a}) \le \sup_{\mathcal{B}\in\mathcal{S}(\mathbf{a})} I_f(\mathcal{B}) \le I_f(\mathcal{B}^*) \le F_m(T(\boldsymbol{a})).$$

Indeed, if

$$F_m(\boldsymbol{a}) < \sup_{\mathcal{B}\in\mathcal{S}(\mathbf{a})} I_f(\mathcal{B}),$$

then we achieve a strict inequality already at step 1. On the other hand, if this is not the case then we have $a_k \in D(f, E(X|B_k))$ for all $B_k \in \mathcal{B} \in \mathcal{S}(\mathbf{a})$ and by Theorem (4.19), (2), at least one partition $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ must be degenerate. This gives a strict inequality in step 2. Thus, we have proved part 2 of the assertion.

Part 3 of the assertion follows from the trivial fact that there are only finitely many partitions of a finite set. From part 2 we obtain that no partition can be met twice during the iteration. □

Proof: (of Theorem (2.26)) Let $M = \{x \in \mathbb{R}^d : |\alpha(x, \boldsymbol{a})| > 1\}$. From

$$M \subseteq \bigcup_{j\neq k} \{x \in \mathbb{R}^d : a_k' x - f^c(a_k) = a_j' x - f^c(a_j)\}$$

it follows that $P(M) = 0$. Since the partitions $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ can differ from one another only on their trace in M they must have identical f-information $I_f(\mathcal{B})$. □

The following two lemmas prepare the proof of Theorem (2.26).

(4.20) LEMMA *For every $x \in \mathbb{R}^d$ and every $\boldsymbol{a} \in A^m$ there exists $\epsilon(x, \boldsymbol{a}) > 0$, such that for $\boldsymbol{b} \in A^m$*

$$\max_{1 \le j \le m} ||a_j - b_j|| < \epsilon(x, \boldsymbol{a}) \quad \Rightarrow \quad \alpha(x, \boldsymbol{b}) \subseteq \alpha(x, \boldsymbol{a}), \tag{4.21}$$

Proof: Let $x \in \mathbb{R}^d$, $\boldsymbol{a} \in A^m$ and define $\gamma := \max_{1 \le j \le m} \left(a_j' x - f^c(a_j)\right)$. Then

$$\delta := \gamma - \max_{j \notin \alpha(x, \mathbf{a})} (a_j' x - f^c(a_j)) > 0.$$

We will show that there exists an $\epsilon > 0$ such that

$$||a_j - b_j|| < \epsilon \quad \Rightarrow \quad \begin{cases} b_j' x < a_j' x + \delta/2 \\ f^c(b_j) > f^c(a_j) - \delta/2 \end{cases} \tag{4.22}$$

Provided this is true, we choose $\epsilon(x, \boldsymbol{a}) > 0$ such that (4.22) holds for $j = 1, 2, \ldots, m$. Then the assertion follows.

There is no problem of choosing $\epsilon > 0$ in such a way that the first of the inequalities of (4.22) is true. In order to obtain the second inequality recall that f^c is lower semicontinuous and $\{b \in A : f^c(b) > f^c(a_j) - \delta/2\}$ is therefore an open set. Hence we may choose $\epsilon > 0$ such that

$$||a_j - b|| < \epsilon \quad \Rightarrow \quad f^c(b) > f^c(a_j) - \delta/2.$$

□

(4.23) LEMMA *Suppose that f is differentiable on $K^\circ(f)$ and f^c is continuous on $A = K(f^c)$. Suppose further that $x \in K^\circ(f)$, $\boldsymbol{a} \in A^m$ has pairwise different components, and that*

$$a_j' x - f^c(a_j) = \max_{1 \le k \le m} \left(a_k' x - f^c(a_k)\right). \tag{4.24}$$

Then the following assertion is valid:

For every choice of neighbourhoods $U_k \in \mathcal{U}(a_k)$ there are points $b_k \in U_k \cap A$ and neighbourhoods $V_k \subseteq U_k$, $V_k \in \mathcal{U}(b_k)$, such that for any choice of $c_k \in V_k \cap A$

$$c_j' x - f^c(c_j) > c_k' x - f^c(c_k), \quad \text{whenever } k \neq j. \tag{4.25}$$

Proof: Since f is differentiable, for every x there is a uniquely determined $d \in A$ such that $d'x - f^c(d) = f(x)$. If $a_j = d$, then we define $b_j := a_j = d$. If $a_j \neq d$, then we define $b_j := (1-\epsilon)a_j + \epsilon d$ where $\epsilon > 0$ is such that $b_j \in U_j$. For $k \neq j$ we define $b_k = a_k$.

We have $a_k \neq d$ whenever $k \neq j$. Indeed, if $a_k = d$ then (4.24) would imply $a_j = a_k = d$ which contradicts the assumptions. Thus, we have $d'x - f^c(d) > a_k'x - f^c(a_k)$ whenever $k \neq j$, and we may choose $\delta > 0$ in such a way that

$$d'x - f^c(d) > a_k'x - f^c(a_k) + \delta, \quad \text{if } k \neq j.$$

Let further $\eta < \epsilon\delta/4$ and let V_k be neighbourhoods of b_k, such that

$$c \in V_k \cap A \quad \Rightarrow \quad \begin{cases} |c'x - b_k'x| < \eta, \\ |f^c(c) - f^c(b_k)| < \eta. \end{cases}$$

Now, for every choice of $c_j \in V_j \cap A$, $c_k \in V_k \cap A$, $k \neq j$, it follows that

$$\begin{aligned} c_j'x - f^c(c_j) &\geq b_j'x - f^c(b_j) - 2\eta \\ &\geq (1-\epsilon)(a_j'x - f^c(a_j)) + \epsilon(d'x - f^c(d)) - 2\eta \\ &\geq (1-\epsilon)(a_k'x - f^c(a_k)) + \epsilon(d'x - f^c(d)) - 2\eta \\ &> (1-\epsilon)(a_k'x - f^c(a_k)) + \epsilon(a_k'x - f^c(a_k)) + \epsilon\delta - 2\eta \\ &= a_k'x - f^c(a_k) + \epsilon\delta - 2\eta \\ &= b_k'x - f^c(b_k) + \epsilon\delta - 2\eta \\ &\geq c_k'x - f^c(c_k) + \epsilon\delta - 4\eta \\ &> c_k'x - f^c(c_k). \end{aligned}$$

which proves the assertion. □

Proof: (of Theorem (2.28) Let us begin with the proof of assertion 1. Let $\boldsymbol{a} \in A^m$ be a point in generic position, i.e. $|\alpha(x_i, \boldsymbol{a})| = 1$ for all $i = 1, 2, \ldots, n$. By Lemma (4.20) for every $i = 1, 2, \ldots, n$ there exists $\epsilon_i > 0$ such that for $\boldsymbol{b} \in A^m$

$$\max_{1 \leq j \leq m} ||a_j - b_j|| < \epsilon_i \quad \Rightarrow \quad \alpha(x_i, \boldsymbol{b}) \subseteq \alpha(x_i, \boldsymbol{a}).$$

Let $\epsilon := \min_i \epsilon_i$. Then it follows that

$$\max_{1 \leq j \leq m} ||a_j - b_j|| < \epsilon \quad \Rightarrow \quad |\alpha(x_i, \boldsymbol{b})| = 1 \quad \text{for all } i = 1, 2, \ldots, n.$$

Hence the set of points in generic position is open.
The assertion 2 is proved by Lemma (4.23). This lemma shows that a point $\boldsymbol{a}$ that is not in generic position can be varied within an arbitrary small neighbourhood in such a way that for any fixed x_i the condition $|\alpha(x_i, \boldsymbol{b})| = 1$ becomes valid. If we do this for every of the finitely many x_i we arrive at a point $\boldsymbol{b}$ in generic position. □

The following is another consequence of Lemma (4.20).

(4.26) COROLLARY *For each point $\boldsymbol{a} \in A^m$ in generic position there is a neighbourhood U such that the unique segmentation $\mathcal{B} \in \mathcal{S}(\boldsymbol{b})$, $b \in U$, is constant.*

Proof: This is an immediate consequence of Lemma (4.20). □

The final couple of proofs is concerned with directional derivatives and gradients.

(4.27) LEMMA *Suppose that the assumptions (2.4) are satisfied and let $\boldsymbol{a} \in A^m$. Then for every admissible direction $\boldsymbol{h}$ of $\boldsymbol{a}$ we have*

$$D^+ F_m(\boldsymbol{a}, \boldsymbol{h}) = \int \max_{j \in \alpha(\mathbf{a},x)} \left(h_j' x - D^+ f^c(a_j, h_j) \right) dP. \tag{4.28}$$

Proof: For notational convenience let us define

$$\begin{aligned} A_j(\epsilon, x) &:= (a_j + \epsilon h_j)' x - f^c(a_j + \epsilon h_j) \\ B_j(x) &:= a_j' x - f^c(a_j) \\ C_j(x) &:= h_j' x - D^+ f^c(a_j, h_j) \end{aligned}$$

and

$$\delta_j(\epsilon) := D^+ f^c(a_j, h_j) - \frac{f^c(a_j + \epsilon h_j) - f^c(a_j)}{\epsilon}.$$

Since f^c is a convex function it follows that $\delta_j(\epsilon) \uparrow 0$ if $\epsilon \downarrow 0$. We have

$$\frac{A_j(\epsilon, x) - B_j(x)}{\epsilon} = C_j(x) + \delta_j(\epsilon). \tag{4.29}$$

The assertion says that

$$\int \frac{\max_j A_j(\epsilon, x) - \max_j B_j(x)}{\epsilon} dP \to \int \max_{j \in \alpha(\mathbf{a},x)} C_j(x)\, dP.$$

We will prove the assertion by showing pointwise dominated convergence of the integrands.

Let us begin with pointwise convergence. Let x be fixed. We have

$$\max_j B_j(x) = \max_{j\in\alpha(\mathbf{a},x)} B_j(x) > B_k(x), \quad \text{if } k \notin \alpha(\boldsymbol{a}, x).$$

There is some $\eta > 0$ such that

$$\max_j B_j(x) > B_k(x) + \eta, \quad \text{if } k \notin \alpha(\boldsymbol{a}, x).$$

Let $\epsilon > 0$ be such that $\epsilon(C_j(x) + \delta_j(\epsilon)) < \eta$ for every j. Then we obtain (observing that $B_j(x)$ does not depend on $j \in \alpha(\boldsymbol{a}, x)$)

$$\begin{aligned}
\max_j A_j(\epsilon, x) &= \max_{j\in\alpha(\mathbf{a},x)} A_j(\epsilon, x) \\
&= \max_{j\in\alpha(\mathbf{a},x)} \Big(B_j(x) + \epsilon(C_j(x) + \delta_j(\epsilon))\Big) \\
&= \max_{j\in\alpha(\mathbf{a},x)} B_j(x) + \max_{j\in\alpha(\mathbf{a},x)} \epsilon(C_j(x) + \delta_j(\epsilon)) \\
&= \max_j B_j(x) + \max_{j\in\alpha(\mathbf{a},x)} \epsilon(C_j(x) + \delta_j(\epsilon))
\end{aligned}$$

whence the asserted pointwise convergence is obvious.
In order to prove dominated convergence we observe that

$$\frac{|\max_j A_j(\epsilon, x) - \max_j B_j(x)|}{\epsilon} \le \max_j |C_j(x) + \delta_j(\epsilon)|$$

$$\le \max_j ||h_j||\,||x|| + \max_j D^+ f^c(a_j, h_j) + \max_j |\delta_j(\epsilon)|.$$

□

Proof: (of Theorem (2.29)) The first part of the assertion follows from Lemma (4.27). To prove the second part let $\boldsymbol{a}$ be a fixpoint and let $\boldsymbol{h}$ be an admissible direction of $\boldsymbol{a}$. Every partition $\mathcal{B}$ which satisfies

$$x \in B_j \quad \Rightarrow \quad j \in \alpha(\boldsymbol{a}, x)$$

is contained in $\mathcal{S}(\boldsymbol{a})$. We choose a partition $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ such that

$$\begin{aligned}
&\int \max_{j\in\alpha(\mathbf{a},x)} \Big(h_j' x - D^+ f^c(a_j, h_j)\Big)\, dP \\
&= \sum_{j=1}^m \int_{B_j} \Big(h_j' x - D^+ f^c(a_j, h_j)\Big)\, dP \\
&= \sum_{j=1}^m P(B_j)\Big(E(X|B_j)' h_j - D^+ f^c(a_j, h_j)\Big).
\end{aligned}$$

Since $\boldsymbol{a}$ is a fixpoint we obtain from Lemma (4.19) that $a_j \in Df(E(X|B_j))$ and by Lemma (5.6) we get

$$D^+ f^c(a_j, h) \geq E(X|B_j)'h \quad \text{for all } h \in \mathbb{R}^d.$$

Hence the assertion.
Let us prove the third part of the assertion. Let $\mathcal{B} \in \mathcal{S}(\boldsymbol{a})$ be any segmentation. Then we have for every admissible direction $\boldsymbol{h}$

$$\sum_{j=1}^{m} \int_{B_j} \left(h_j' x - D^+ f^c(a_j, h_j) \right) dP$$
$$\leq \int \max_{j \in \alpha(\mathbf{a}, x)} \left(h_j' x - D^+ f^c(a_j, h_j) \right) dP \leq 0.$$

Choose j and keep it fixed. By Lemma (4.19) we have to show that $a_j \in D(f, E(X|B_j))$ which is equivalent to

$$f(E(X|B_j)) \leq a_j' E(X|B_j) - f^c(a_j). \tag{4.30}$$

Since $\mathcal{B}$ is non-degenerate we have $P(B_j) > 0$. This implies $E(X|B_j) \in K^\circ(f)$ by Corollary (4.4) and thus $D(f, E(X|B_j)) \neq \emptyset$. Let $c_j \in D(f, E(X|B_j))$ which means that

$$f(E(X|B_j)) = c_j' E(X|B_j) - f^c(c_j). \tag{4.31}$$

Defining an admissible direction by

$$h_k = \begin{cases} 0 & \text{if } k \neq j, \\ c_j - a_j & \text{if } k = j, \end{cases}$$

gives

$$\int_{B_j} \left(h_j' x - D^+ f^c(a_j, h_j) \right) dP \leq 0.$$

Since we have

$$f^c(c_j) - f^c(a_j) = \frac{f^c(a_j + h_j) - f^c(a_j)}{1} \geq D^+ f^c(a_j, h_j),$$

we obtain

$$\int_{B_j} \left(h_j' x - f^c(c_j) + f^c(a_j) \right) dP \leq 0.$$

This implies

$$c_j' E(X|B_j) - f^c(c_j) \leq a_j' E(X|B_j) - f^c(a_j),$$

which proves the assertion (4.30) in view of (4.31). □

Proof: (of Corollary (2.30)) If $\boldsymbol{a}$ is in generic position then it follows from Corollary (4.26) that

$$F_m(b) = \sum_{j=1}^{m} \int_{B_j} (b_j' X - f^c(b_j))\, dP$$

is a concave function on a neighbourhood of $\boldsymbol{a}$. Then the assertion follows from Theorem (2.29). □

5 Appendix: Some concepts from convex analysis

In this section we summarize some concepts and facts from convex analysis for the reader's convenience. Our basic reference is Stoer and Witzgall, 1970, [14].

Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a convex function. We will consider only lower semicontinuous convex functions, i.e. where the level sets $\{f \leq \beta\}$ are closed sets for all $\beta \in \mathbb{R}$.

The domain of a convex function f is defined by

$$K(f) = \{x \in \mathbb{R}^d : f(x) < \infty\} \tag{5.1}$$

On the interior of the domain $K(f)^\circ$ the convex function f is continuous (Stoer and Witzgall, [14], Theorem 4.1.5). If f is finite, then $K(f) = K(f)^\circ = \mathbb{R}^d$. Otherwise the convex function f can be discontinuous on the boundary $\partial K(f)$.

Convex functions are closely related to linear functions. Let

$$L(f) = \{\ell : \ell(x) = a'x - \beta \leq f(x),\ x \in \mathbb{R}^d\} \tag{5.2}$$

be the set of all affine linear functions which are dominated by f. If we ask for affine linear functions $\ell(x) = a'x - \beta$ which are dominated by f and which have a given derivative $a \in \mathbb{R}^d$ then we arrive at the concept of the conjugate convex function of Definition2.8 (Stoer and Witzgall, [14], Abschnitt 4.6). The conjugate convex function f^c is a convex function with values in $(-\infty, \infty]$ and it is lower semicontinuous. The relation of conjugate convex functions f^c to dominated affine linear functions $\ell \leq f$ is described in the following lemma.

(5.3) LEMMA *For the conjugate convex function f^c the following assertions are true:*

1. *We have $f^c(a) < \infty$ iff there is an affine linear function $\ell \in L(f)$ with derivative a.*

2. *If $f^c(a) < \infty$ then $\ell(x) = a'x - f^c(a)$ is the largest of all affine linear functions in $L(f)$ with derivative a.*

If $f^c(a) < \infty$, then the affine linear function $\ell(x) = a'x - f^c(a)$ is called a support function of f. Of special interest are support functions which have a common point with f.

(5.4) DEFINITION *Let $S(f,x)$ the set of all affine linear functions $\ell \in L(f)$ such that $\ell(x) = f(x)$. The subdifferential $D(f,x)$ is the set of all derivatives of functions in $S(f,x)$.*

Every function $\ell \in S(f,a)$ is a support function. For every point of $K(f)^\circ$ the subdifferential is not empty (Stoer und Witzgall, [14], Theorem 4.2.8). If the set $S(f,a)$ contains exactly one affine linear function then the convex function f is differentiable at a.

For technical reasons we need the following simple facts.

(5.5) LEMMA *Let $x_n \to x \in K^\circ(f)$. Then every sequence of support functions $\ell_n \in S(f,x_n)$ is bounded and all accumulation points are contained in $S(f,x)$.*

(5.6) LEMMA *If $a_0 \in D(f,b)$ then we have*

$$D^+ f^c(a_0,h) \geq b'h \quad \text{for all } h \in \mathbb{R}^d.$$

Proof: Let $\epsilon > 0$. Since we have $a'x - f^c(a) \leq f(x)$ for all x and a, we have in particular

$$(a_0 + \epsilon h)'b - f^c(a_0 + \epsilon h) \leq f(b) = a_0'b - f^c(a_0).$$

This implies

$$f^c(a_0 + \epsilon h) - f^c(a_0) \geq \epsilon b'h.$$

□

Acknowledgements

The authors wish to thank J. Mazanec for posing the problem and for stimulating discussions. The work was partly supported by the Spezialforschungsbereich about Adaptive Information Systems and Modelling in Economics and Management Science at the Vienna University of Economics and Business Administration.

References

[1] H.H. Bock. *Automatische Klassifikation.* Vandenhoeck und Ruprecht, 1974.

[2] Bouton C. and Pages G. About the multidimensional competitve learning vector quantization algorithm with constant gain. *Annals of Applied Probability*, 7:679–710, 1997.

[3] J. Conway and N. Sloane. *Sphere-packings, lattices and groups.* Springer, 1993.

[4] L. Devroye, L. Györfi, and L. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics.* Springer, 1996.

[5] E. Fix and J.L. Hodges. Discriminatory analysis:nonparametric discrimination. Technical Report 21–49–004, USAF School of Aviation Medicine, 1951.

[6] B. Flury, Tarpey T., and Li L. Principal points and self-consistent points of elliptical distributions. *Annals of Statistics*, 23:102–112, 1995.

[7] B.A. Flury. Principal points. *Biometrika*, 77:33–41, 1990.

[8] J.A. Hartigan. *Clustering algorithms.* Wiley, New York, 1975.

[9] Cuesta-Albertos J.A., Gordaliza A., and Matran C. Trimmed k-means: An attempt to robustify quantizers. *Annals of Statistics*, 25:553–576, 1997.

[10] T. Kohonen. *Self-organization and associative memory.* Springer, 1984.

[11] T. Masters. *Practical Neural Network Recipes in C++.* Academic Press, 1993.

[12] D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9:135–140, 1981.

[13] D. Pollard. A central limit theorem for k-means clustering. *Annals of Probability*, 10:919–926, 1982.

[14] J. Stoer and C. Witzgall. *Convexity and optimization in finite dimensions I*, volume 163 of *Die Grundlehren der mathematischen Wissenschaften.* Springer, 1970.

[15] H. Strasser. *Mathematical theory of statistics: Statistical experiments and asymptotic decision theory*, volume 7 of *De Gruyter Studies in Mathematics.* de Gruyter, 1985.

[16] E.N. Torgersen. *Comparison of statistical experiments.* Cambridge Univ. Press, 1991.