ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

Fitting parametric sets to probability distributions

MEELIS KÄÄRIK AND KALEV PÄRNA

ABSTRACT. We consider the problem of approximation of distributions by sets. Given a probability P on a metric space (S,d) and a class $\mathcal{A}$ of subsets of S, find an approximative set $A \in \mathcal{A}$ that minimizes the mean discrepancy of a random point $X \sim P$ from the set A:

$$W(A,P) = \int_S \varphi(d(x,A))P(dx) \to \min_{A \in \mathcal{A}}.$$

We are especially interested in the case of parametric approximative sets, $\mathcal{A} = \{A(\Theta) : \Theta \in T\}$, where the aim is to find a value of the parameter $\Theta \in T$ which minimizes

$$\overline{W}(\Theta, P) = W(A(\Theta), P) \to \min_{\Theta \in T}.$$

Current article is an extension of Käärik and Pärna (2003) to the case where approximating sets $A(\Theta)$ are unions of k parametric sets of the same type. We will prove the convergence of optimal values of loss-functions and the existence of optimal sets in some functional spaces.

1. Introduction

Approximation of distributions by sets and consistency of empirically optimal approximations is an old but still developing issue in probability theory and mathematical statistics.

Classical examples of one-point approximations (mean, median) are well known and consistency of empirical means is a fundamental fact called the Law of Large Numbers. Extension of the class of approximations to k-points sets leads us to k-centres (sometimes called k-means, principal points), the most studied type of approximative sets during last decades.

Received December 2, 2003.

2000 *Mathematics Subject Classification.* 60B05, 60B10.

Key words and phrases. Approximation of distributions, loss-function, discrepancy function, consistency, M-estimation, k-centres.

Supported by Estonian Science Foundation Grant 5277 and Target Financed Project 0181776s01.

The properties of k-centres are studied in various functional spaces, starting from the case $S = \mathbb{R}^d$ (Pollard (1981); Zhang and Zhu (1993); Cuesta-Albertos, Gordaliza and Matran (1997); Graf and Luschgy (2000)) to general Banach spaces (Cuesta and Matran (1988), (1989); Pärna (1990); Lember and Pärna (1999)). A systematic overview of the theory of k-centres in separable metric spaces is given in Lember (1999).

Besides k-centres, some more complex approximations have been investigated. For instance, approximation of distributions by spheres with fixed radius in finite-dimensional Euclidean spaces is considered in Käärik (2000). A closely related problem (called median ball problem) is studied in Averous and Meste (1997).

Mathematically, the problem of approximation of distributions by sets can be formulated as follows. Let P be a probability distribution on a separable metric space (S, d). Let $\mathcal{A} \subset 2^S$ be a class of subsets of S. Our aim is to find an approximative set $A \in \mathcal{A}$ that minimizes the mean discrepancy of a random point $X \sim P$ from the set A:

$$W(A, P) = \int_S \varphi(d(x, A))P(dx) \to \min_{A \in \mathcal{A}},$$

where $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ is a nondecreasing discrepancy function and the distance $d(x, A) = \inf\{d(x, a) : a \in A\}$.

Let $W(P) := \inf_{A \in \mathcal{A}} W(A, P)$ denote the infimum value of the loss function.

Definition 1.1. Any set A satisfying $W(A, P) = W(P)$ is called *optimal approximation* for P (shortly, *P-optimal*).

An approximation A is called *ε-optimal* for P (or, *P-ε-optimal*) if it verifies $W(A, P) < W(P) + \varepsilon$.

Two important questions are:
1) when does an optimal approximating set exist,
2) does a sequence of optimal approximations converge?

The latter question arises when optimal approximative sets have been found for measures from a sequence $\{P_n\}$ which converges weakly to the measure P. In this paper our final goal is to give answers to these questions for the special case of subsets A which are unions of k parametric sets from the same family. The paper is a follow-up to Käärik and Pärna (2003), where the case of a single approximative parametric set was studied.

The paper is organized in the following way. In Section 2 we reveal some general properties of our loss-functions for an arbitrary class $\mathcal{A}$ of approximative sets. Then, in Section 3, we cover the case where approximative sets are unions of k elements of a given class $\mathcal{A}$. In practice, this case is of interest if the class $\mathcal{A}$ itself is too "narrow" or "limited" to provide a good approximation for P. These general properties are exploited in Section 4, where the distribution P is being approximated by multiple parametric sets of the same type (k lines, k circles, etc). The restriction to have all sets of the same type is used to prevent some technical difficulties. However, we believe that the main idea will work in the case of sets of different type as well.

2. Arbitrary class of approximative sets

Here we deduce some properties of the loss-function $W(A, P)$ for an arbitrary class of approximative sets.

2.1. Assumptions. Let us introduce some requirements for the discrepancy function φ and for the measures P and $\{P_n\}$.

A1. The discrepancy function φ has following properties: $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ is continuous, nondecreasing, $\varphi(0) = 0$, and it has Δ_2-property, i.e. $\exists \lambda > 0 : \quad \varphi(2x) \leq \lambda\varphi(x), \quad \forall x > 0$. We also assume that the inequality $\lim_{s\to\infty} \varphi(s) =: \varphi(\infty) > 0$ holds, since the approximation problem has no meaning with $\varphi(x) \equiv 0$.

A2. For some $x_0 \in S$ we have $\int \varphi(d(x, x_0))P(dx) < \infty$.

A3. The weak convergence $P_n \Rightarrow P$ holds.

A4. For some $x_0 \in S$ the function $\varphi(d(x, x_0))$ is uniformly integrable with respect to $\{P_n\}$.

Comments. The requirements A1 are weak enough to be satisfied by any power function. Assumption A2 guarantees the finiteness of the loss-function. Conditions A3 and A4 are satisfied if $\{P_n\}$ is a sequence of *empirical measures* corresponding to P, for example, The weak convergence of empirical measures is shown to take place with probability 1 in Varadarajan (1958), and the property A4 for these measures is verified in Pärna (1988).

2.2. General properties of the loss-function. Throughout this section we assume that A1–A4 are fulfilled and the space S is a separable metric space. The class of approximative sets $\mathcal{A} \subset 2^S$ is assumed to be arbitrary class satisfying $W(A, P) < \varphi(\infty)$ for some $A \in \mathcal{A}$. This is a very natural assumption, since in the case $W(A, P) \equiv \varphi(\infty)$ for each $A \in \mathcal{A}$, any $A \in \mathcal{A}$ is optimal. We will derive some general properties of the loss-function, the main result being the convergence of infimum values of the loss-function, $W(P_n) \to W(P)$, provided that $P_n \Rightarrow P$.

Under the assumptions A1–A4 the following assertions are valid.

Lemma 2.1. *Let A be a non-empty subset of a separable metric space S. Then for each $R \geq 0$ we have*

a) $\int \varphi(d(x, A) + R)P(dx) < \infty$,

b) *the function* $\varphi(d(x,A)+R)$ *is uniformly integrable with respect to* $\{P_n\}$, *i.e.*

$$\lim_{a\to\infty}\sup_n \int_{\varphi(d(x,A)+R)\geq a} \varphi(d(x,A)+R)P_n(dx)=0,$$

c) *the convergence*

$$\int \varphi(d(x,A)+R)P_n(dx)\to\int \varphi(d(x,A)+R)P(dx)$$

takes place.

Proof. For the proof of Lemma 2.1 see, e.g. Käärik and Pärna (2003). □

By taking $R=0$ in the statement c), one immediately has that for each fixed subset $A\in\mathcal{A}$ the convergence $W(A,P_n)\to W(A,P)$ takes place.

Lemma 2.2. *The inequality* $\limsup_n W(P_n)\leq W(P)$ *holds.*

Proof. For every fixed $A\in\mathcal{A}$ we have $W(P_n)\leq W(A,P_n)$ and $W(A,P_n)\to W(A,P)$. Let $\varepsilon>0$ be arbitrary and let $A\in\mathcal{A}$ satisfy $W(A,P)\leq W(P)+\varepsilon$. Then $W(P_n)\leq W(A,P_n)\to W(A,P)\leq W(P)+\varepsilon$ which implies $\limsup_n W(P_n)\leq W(P)+\varepsilon$. Since $\varepsilon>0$ was arbitrary, the lemma is proved. □

Lemma 2.3. *The inequality* $W(P)<\varphi(\infty)$ *holds.*

Proof. By the assumption outlined in the beginning of the present subsection, there exists $A\in\mathcal{A}$ such that $W(A,P)<\varphi(\infty)$. Therefore $W(P)\leq W(A,P)<\varphi(\infty)$. □

Lemma 2.4. *For any class* $\mathcal{A}\subset 2^S$ *and any closed ball* $\bar{B}(x_0,R)\subset S$ *the uniform convergence*

$$\lim_n \sup_{\substack{A\in\mathcal{A}\\ A\cap\bar{B}(x_0,R)\neq\emptyset}} |W(A,P_n)-W(A,P)|=0 \tag{1}$$

takes place.

Proof. Denote $f_A(x)=\varphi(d(x,A))$ and $\Phi=\{f_A|A\in\mathcal{A},A\cap\bar{B}(x_0,R)\neq\emptyset\}$. Then the condition (1) is equivalent to

$$\lim_n \sup_{f_A\in\Phi}|\int f_A(x)P_n(dx)-\int f_A(x)P(dx)|=0.$$

Some sufficient conditions for this uniform convergence to hold are given by Ranga Rao (1962): there exists $g(x)$ such that $|f_A(x)|<g(x)$ for all $f_A(x)\in\Phi$ and $x\in S$, $P_n\Rightarrow P$, $\int gdP_n\to\int gdP<\infty$, and the class Φ is equicontinuous. We shall check that these conditions are fulfilled.

Define the function $g(x) := \varphi(d(x, x_0) + R)$, then for every A satisfying $A \cap \bar{B}(x_0, R) \neq 0$ the inequalities

$$f_A(x) = \varphi(d(x, A)) \leq \varphi(d(x, x_0) + d(x_0, A)) \leq \varphi(d(x, x_0) + R) = g(x)$$

hold. The convergence $P_n \Rightarrow P$ takes place by assumption A3 and Lemma 2.1 c) ensures $\int g dP_n \to \int g dP < \infty$.

To prove the equicontinuity of Φ, we follow the ideas used in Pärna (1986), Lemma 1.

Let $x_1 \in S$ and $\varepsilon > 0$ be arbitrary. Since the discrepancy function φ is continuous, it is uniformly continuous in the closed interval $I := [0, d(x_1, x_0) + R]$, i.e. for any $\varepsilon > 0$ there exists such $\delta = \delta(\varepsilon, x_1)$ that $|\varphi(s) - \varphi(t)| < \varepsilon$ holds for every $s \in I$ and $t > 0$ where $|s - t| < \delta$. Choose $s = d(x_1, A)$ and $t = d(x_2, A)$ and let $d(x_1, x_2) < \delta$. We can see that $s \in I$ (from one side $s = d(x_1, A) \leq d(x_1, x_0) + d(x_0, A)) \leq d(x_1, x_0) + R$ and from the other side $s \geq 0$).

Furthermore, by the triangle inequality $d(x_1, A) \leq d(x_1, x_2) + d(x_2, A)$, we have $|s - t| = |d(x_1, A) - d(x_2, A)| \leq d(x_1, x_2) < \delta$.

The fluctuation of any $f_A \in \Phi$ can now be estimated as follows:

$$|f_A(x_1) - f_A(x_2)| = |\varphi(d(x_1, A)) - \varphi(d(x_2, A))| = |\varphi(s) - \varphi(t)| < \varepsilon,$$

which completes the proof of equicontinuity of the class Φ. □

Lemma 2.5. *Let S be a separable metric space, and let ε satisfy $0 \leq \varepsilon < \varphi(\infty) - W(P)$. Then there exists a ball $\bar{B}(x_0, R) \subset S$ that eventually intersects*

a) *every ε-minimizing sequence $\{B_n\}$ for the loss-function $W(\cdot, P)$, i.e. every sequence $\{B_n\}$ satisfying $\limsup_n W(B_n, P) \leq W(P) + \varepsilon$;*
b) *every sequence of P_n-ε-optimal approximations.*

Proof. To prove a) assume, on contrary, that no such ball exists. Then, for any $x \in S$, there exists a subsequence $\{B_{n'}\}$ such that the distance $d(x, B_{n'})$ will tend to infinity, and by the continuity of the discrepancy function we get $\lim_{n'} \varphi(d(x, B_{n'}) = \varphi(\infty)$.

We now integrate this equality and use the Fatou lemma and Lemma 2.2:

$$\begin{aligned} \varphi(\infty) &= \int \liminf_{n'} \varphi(d(x, B_{n'}))P(dx) \leq \liminf_{n'} \int \varphi(d(x, B_{n'}))P(dx) \\ &= \liminf_{n'} W(B_{n'}, P) \leq \limsup_{n} W(B_n, P) \leq W(P) + \varepsilon < \varphi(\infty) \end{aligned}$$

– a contradiction.

The proof of b) is similar (cf., e.g. Proposition 3.3.1 in Lember (1999)). □

Proposition 2.1. *If the assumptions A1–A4 are fulfilled and S is a separable metric space, then the convergence $W(P_n) \to W(P)$ takes place.*

Proof. Choose an arbitrary sequence $\varepsilon_n \downarrow 0$ and consider a sequence of P_n-ε_n-optimal approximations $\{A_n^{\varepsilon_n}\}$. By Lemma 2.5 b), this sequence intersects with a ball $\bar{B}(x_0, R) \subset S$, and Lemma 2.4 implies

$$\lim_n |W(A_n^{\varepsilon_n}, P_n) - W(A_n^{\varepsilon_n}, P)| = 0. \tag{2}$$

Now the difference between optimal values can be estimated by

$$\begin{aligned} W(P) - W(P_n) &\leq W(A_n^{\varepsilon_n}, P) - W(P_n) \\ &\leq W(A_n^{\varepsilon_n}, P) - W(A_n^{\varepsilon_n}, P_n) + \varepsilon_n. \end{aligned} \tag{3}$$

If $n \to \infty$, then, due to (2), the difference between the two first terms in (3) tends to zero and we obtain that $\liminf_n W(P_n) \geq W(P)$. From the other side, Lemma 2.2 states that $\limsup_n W(P_n) \leq W(P)$. Hence we have $\lim_n W(P_n) = W(P_n)$. □

Lemma 2.6. *Assume that A1–A4 hold. Let S be a separable metric space and let ε satisfy $0 \leq \varepsilon < \varphi(\infty) - W(P)$. Then every sequence of P_n-ε-optimal approximations $\{A_n^\varepsilon\}$ is ε-minimizing for the loss-function $W(\cdot, P)$.*

Proof. Choose an arbitrary $0 \leq \varepsilon < \varphi(\infty) - W(P)$ and estimate

$$\begin{aligned} &|W(A_n^\varepsilon, P) - W(P)| \\ &\quad\leq |W(A_n^\varepsilon, P) - W(A_n^\varepsilon, P_n)| + |W(A_n^\varepsilon, P_n) - W(P)| \\ &\quad\leq |W(A_n^\varepsilon, P) - W(A_n^\varepsilon, P_n))| + |W(P_n) - W(P)| + \varepsilon. \end{aligned}$$

Lemmas 2.4 and 2.5 b) imply that the first term tends to zero, and by Proposition 2.1 also the second term tends to zero, which implies

$$\limsup_n |W(A_n^\varepsilon, P) - W(P)| \leq \varepsilon.$$

□

3. Approximation with multiple sets

As before, we assume that A1–A4 are fulfilled and the space S is a separable metric space. Let $\mathcal{A}$ be a class of subsets of S. In this section we consider approximation of distributions by multiple sets of type $\mathcal{A}$, i.e. approximative sets A are unions of k elements of $\mathcal{A}$, $A = A^1 \cup \ldots \cup A^k$ with all $A^i \in \mathcal{A}$ (the number k is fixed). Let $\mathcal{A}_k$ be the class of all such unions.

Since the results obtained in Section 2 were proved for an arbitrary class of approximative sets, they are valid for this special case as well.

Multiple approximative sets are of interest each time when the class $\mathcal{A}$ itself is too narrow to allow a reasonably good fit. Then finite unions of sets taken from $\mathcal{A}$ can offer better opportunities for approximation.

In what follows, we assume that the class $\mathcal{A}$ covers all the space S or at least the support of P: $P(\cup_{A\in\mathcal{A}}A) = 1$.

Let $W_k(P)$ denote the optimal value of the loss-function for the distribution P over the class $\mathcal{A}_k$. Then by Proposition 2.1 we have the convergence $W_k(P_n) \to W_k(P)$, provided that the assumptions A1–A4 are fulfilled and S is a separable metric space.

In order to analyze further properties based on the special structure of the approximative sets, and to rule out some trivial cases, we need to assume that the inequalities $W_k(P) < W_{k-1}(P) < \ldots < W_1(P)$ hold. Lemma 3.1 gives us an equivalent, but more convenient condition.

Lemma 3.1. *The inequalities $W_{k-1}(P) > 0$ and $W_k(P) < W_{k-1}(P) < \ldots < W_1(P)$ are equivalent.*

Proof. The same statement has been proved for the case of k-centre-approximation as Proposition 1.2.4 in Lember (1999). We can use main ideas of that proof in our case as well, the details are omitted. □

We will add a new assumption A5 to our special case of multiple sets:
A5. The inequality $W_{k-1}(P) > 0$ holds.

Notice that in the case $k = 1$ this assumption reduces to $W_0(P) = \varphi(\infty) > 0$, which is one of the general assumptions for the discrepancy function φ.

Lemma 3.2. *Suppose that $0 \le \varepsilon < W_{k-1}(P) - W_k(P)$. Then there exists a ball $\bar{B}(x_0, M)$, which, for every ε-minimizing sequence $\{A_n^\varepsilon(k)\}$ for $W(\cdot, P)$, eventually intersects each component set.*

Proof. Consider an arbitrary ε-minimizing sequence $\{A_n^\varepsilon(k)\}$ for $W(\cdot, P)$ and let $A_n^\varepsilon(k) = A_n^1 \cup \ldots \cup A_n^k$. Lemma 2.5 a) says that there exists a ball that intersects at least with one component-set, say A_n^1. We proceed by induction: assume that there exists a ball that intersects with l component-sets $(1 \le l < k)$, and show that then there exists a ball that intersects $l + 1$ component-sets as well.

Since all approximative sets are of the same type, we can always arrange indices into a suitable order, so we can say that by assumption a ball $\bar{B}(x_0, R_l)$ intersects the sets $A_n^1, \ldots, A_n^l$, $n > n_0$. Let us assume, on contrary, that for each radius R and index $n_1 > n_0$ there exists an index $s > n_1$ such that the ball $\bar{B}(x_0, R)$ does not intersect with any of the component-sets $A_s^{l+1}, \ldots, A_s^k$. Then there exists a subsequence $\{n'\}$, which satisfies $d(x_0, A_{n'}^{l+1}) \to \infty, \ldots, d(x_0, A_{n'}^k) \to \infty$.

Denote $B_n := A_n^1 \cup \ldots \cup A_n^l$. It is obvious that

$$\varphi(d(x, B_{n'})) - \varphi(d(x, A_{n'}^\varepsilon(k))) \to 0.$$

Now define the functions $h_n(x) := \varphi(d(x, A_n^\varepsilon(k))) - \varphi(d(x, B_n))$, $h(x) := 0$ and $g(x) := \varphi(d(x, x_0) + R)$, where $R \in \mathbb{R}^+$ is chosen to satisfy $R > R_l$ and $\varphi(d(x, x_0) + R) > 0$, for all $x \in S$.

We will now make use of the following result (Lember (1999), Corollary 3.2.2). *If a sequence $\{f_n\}$ satisfies $P\{x_n \to x \Rightarrow f_n(x_n) \to f(x)\} = 1$ and if the functions f_n and f are bounded by a continuous function g, satisfying $g(x) > 0$, for all $x \in S$, and the convergences $P_n \Rightarrow P$ and $\int g dP_n \to \int g dP < \infty$ hold, then $\int f_n dP_n \to \int f dP$.*

In our case, for every $x \in S$, we have that

$$h_n(x) \le \varphi(d(x, A_n^\varepsilon(k))) \le \varphi(d(x, x_0) + R_l) \le \varphi(d(x, x_0) + R) = g(x).$$

Then the assumption A3 and Lemma 2.1 c) imply

$$\int g(x) P_n(dx) = \int \varphi(d(x, x_0) + R) P_n(dx) \to$$
$$\int \varphi(d(x, x_0) + R) P(dx) = \int g(x) P(dx).$$

Therefore the assumptions for Corollary 3.2.2 in Lember (1999) are fulfilled and the convergence $\int h_n dP_n \to \int h dP$ or, equivalently $W(A_n^\varepsilon(k), P_n) - W(B_n, P_n) \to 0$, takes place.

Using now the fact that $A_n^\varepsilon(k)$ is an ε-minimizing sequence, Lemma 2.4 leads us to

$$\begin{aligned} W_k(P) + \varepsilon &\ge \limsup_n W(A_n^\varepsilon(k), P) = \limsup_n W(A_n^\varepsilon(k), P_n) \\ &= \limsup_n W(B_n, P_n) \ge \limsup_n W_l(P_n) = W_l(P) \\ &\ge W_{k-1}(P) > W_k(P) + \varepsilon \end{aligned}$$

– a contradiction. Lemma is proved. □

4. Multiple parametric approximation

In this section, the general results obtained above are applied to the particular case of approximative sets – the class of parametric sets.

4.1. Multiple parametric sets. Let P be a probability distribution on a separable metric space (S, d). Let $\mathcal{A} = \{A(\Theta) : \Theta \in T\} \subset 2^S$ be a parameterized class of subsets of S. The parameter space T is assumed to be a metric space with metrics ϱ_T. To approximate the distribution P we use unions of k parametric sets from the class $\mathcal{A}$. We denote the k-tuples of values of Θ by $\Theta(k) = (\Theta^1, \ldots, \Theta^k) \in T^k$. Each $\Theta(k)$ produces an approximative set $A(\Theta^1) \cup \ldots \cup A(\Theta^1) =: A(\Theta(k))$ and the class of all approximative sets will be denoted by $\mathcal{A}_k = \{A(\Theta(k)) : \Theta(k) \in T^k\}$. We assume that the product space T^k is endowed with a metrics ϱ. We are interested in finding such values of $\Theta(k)$ that minimize

$$\overline{W}(\Theta(k), P) = W(A(\Theta(k)), P) = \int_S \varphi(d(x, A(\Theta(k))))P(dx) \to \min_{\Theta(k)\in T^k}.$$

If $\Theta^*(k)$ is optimal, then corresponding set $A(\Theta^*(k))$ is called *optimal approximation* for P.

Let $W_k(P)$ denote the infimum value of the loss-function. A value of $\Theta(k)$ is called *ε-optimal* for P if it verifies $\overline{W}(\Theta(k), P) < W_k(P) + \varepsilon$. We denote the set of all optimal values of $\Theta(k)$ by $\mathcal{U}(P)$ and the set of all ε-optimal $\Theta(k)$ by $\mathcal{U}^\varepsilon(P)$. Similar notation is used when P is replaced by P_n.

We next recall and modify some properties of a "good" parameterization, first introduced in Käärik and Pärna (2003). Let $h(A, B)$ be the *Hausdorff distance* between sets A and B.

PROPERTY A. The mapping $A : T \to \mathcal{A}$ is locally uniformly continuous, i.e. for any $\varepsilon > 0$ and ball $B(x_0, M) \subset S$ there exists $\delta > 0$ such that $\varrho_T(\Theta_1, \Theta_2) < \delta$ implies $h(A_1, A_2) < \varepsilon$ provided that intersections $A_i := A(\Theta_i) \cap B(x_0, M) \neq \emptyset$, $i = 1, 2$.

PROPERTY B. Let us have a subset $U \subset T^k$. If there exist $x_0 \in S$ and $M > 0$ such that for all $\Theta(k) = (\Theta^1, \ldots, \Theta^k) \in U$ we have $A(\Theta^j) \cap B(x_0, M) \neq \emptyset$, $j = 1, \ldots, k$, then the subset U is bounded.

Remark. It is easy to show that the property B above for parameters in T^k is equivalent to the property C2 in Käärik and Pärna (2003) for parameters in T (describing one approximative component).

4.2. The existence and convergence of optimal parameters. Assume that A1–A5 are fulfilled, and the spaces S and T (and therefore T^k) are separable metric spaces.

Lemma 4.1. *Suppose that* $0 \leq \varepsilon < W_{k-1}(P) - W_k(P)$. *If the parameterization has Property B, then every ε-minimizing sequence* $\{\Theta_n(k)\} \subset T^k$ *for the loss-function* $\overline{W}(\cdot, P)$ *is bounded.*

Proof. Lemma 3.2 and Property B imply the result. □

Corollary 4.1. *If* $\varepsilon < W_{k-1}(P) - W_k(P)$ *and the parameterization has Property B, then every sequence of* P_n*-ε-optimal parameters is bounded.*

Proof. The result follows from lemmas 2.6 and 4.1. □

To prove the convergence of optimal parameters, we need to add more restrictions to spaces S and T, since it can be shown easily that in separable metric spaces the optimal approximations might not exist. Therefore we will assume that the spaces S and T (and therefore T^k) are finite-dimensional normed spaces. We will prove the existence of optimal parameters and convergence of P_n-optimal parameters by first proving the convergence of P_n-ε-optimal parameters.

Proposition 4.1. *Assume that the assumptions A1–A5 are fulfilled, the parameterization has properties A and B, and S and T are finite-dimensional normed spaces. If* $0 \le \varepsilon < W_{k-1}(P) - W_k(P)$, *then the distance between any sequence of* P_n-ε-*optimal parameters and the set of* P-ε-*optimal parameters converges to zero, i.e. the convergence* $\sup_{\Theta_n^\varepsilon \in \mathcal{U}^\varepsilon(P_n)} \varrho(\Theta_n^\varepsilon, \mathcal{U}^\varepsilon(P)) \to 0$ *takes place.*

Proof. We make use of the following property: the convergence $\varrho(\Theta_n^\varepsilon, \mathcal{U}^\varepsilon(P)) \to 0$ holds if every subsequence $\{\Theta_{n'}^\varepsilon\}$ has a (sub)subsequence $\{\Theta_{n''}^\varepsilon\}$ which satisfies $\varrho(\Theta_{n''}^\varepsilon, \mathcal{U}^\varepsilon(P)) \to 0$. Consider an arbitrary sequence $\{\Theta_n^\varepsilon\}$ and its subsequence $\{\Theta_{n'}^\varepsilon\}$. By Corollary 4.1, the sequence $\{\Theta_{n'}^\varepsilon\}$ is bounded. For S being finite-dimensional this sequence is relatively compact, which means there must exist a converging subsequence $\Theta_{n''}^\varepsilon \to \Theta \in T$. Denote $f_n(x) := \varphi(d(x, A(\Theta_n^\varepsilon)))$ and $f(x) := \varphi(d(x, A(\Theta)))$. To get the convergence of corresponding values of loss-functions, we will again use Corollary 3.2.2 from Lember (1999), given in the proof of Lemma 3.2 above. Let us check that the assumptions of this statement are fulfilled.

First notice that according to Lemma 2.5 b) and by the properties of distance d, for each $x \in S$ and P_n-ε-optimal sequence $\{\Theta_n^\varepsilon\}$ we can construct a ball $\bar{B}(x_0, M)$ such that the equalities $d(x, A(\Theta_n^\varepsilon)) = d(x, B_n)$ and $d(x, A(\Theta) = d(x, B)$ hold, eventually, for $B_n = A(\Theta_n^\varepsilon) \cap \bar{B}(x_0, M)$ and $B = A(\Theta) \cap \bar{B}(x_0, M)$.

Now consider a sequence $x_{n''} \to x$, and estimate:

$$
\begin{aligned}
&|d(x_{n''}, A(\Theta_{n''}^\varepsilon)) - d(x, A(\Theta)| \\
&\qquad \le |d(x_{n''}, A(\Theta_{n''}^\varepsilon)) - d(x, A(\Theta_{n''}^\varepsilon))| + |d(x, A(\Theta_{n''}^\varepsilon)) - d(x, A(\Theta)| \\
&\qquad \le d(x_{n''}, x) + |d(x, B_{n''}) - d(x, B)| \le d(x_{n''}, x) + h(B_{n''}, B).
\end{aligned}
$$

By the convergence $x_{n''} \to x$, the first term tends to zero, and using the fact that $\Theta_{n''}^\varepsilon \to \Theta$, and Property A, also the second term tends to zero. Therefore we get $d(x_{n''}, A(\Theta_{n''}^\varepsilon)) \to d(x, A(\Theta))$ and the continuity of φ implies $f_{n''}(x_{n''}) \to f(x)$. Since the elements of the sequence $\{\Theta_n^\varepsilon\}$ are bounded, they are contained in some ball $\bar{B}(\Theta_0, R) \subset T^k$. Choose R big enough to satisfy also $\Theta \subset \bar{B}(\Theta_0, R)$. The functions f_n and f are bounded by the continuous function g of the form $g(x) := \varphi(d(x, x_0) + R)$, and g is integrable by Lemma 2.1 a) – hence all the requirements of Corollary 3.2.2 of Lember (1999) are met and we have the convergence $\int \varphi(d(x, A(\Theta_{n''}^\varepsilon)))dP_{n''} \to \int \varphi(d(x, A(\Theta)))dP$.

Now the loss-function can be estimated:

$$
W(P_{n''}) + \varepsilon \;\ge\; \overline{W}(\Theta_{n''}^\varepsilon, P_{n''}) = \int \varphi(d(x, A(\Theta_{n''}^\varepsilon)))dP_{n''}
$$

$$\to \int \varphi(d(x, A(\Theta)))dP = \overline{W}(\Theta, P). \tag{4}$$

From the other side, Lemma 2.2 gives $\lim_{n''} W(P_{n''}) \leq \limsup_n W(P_n) \leq W(P)$, which together with inequality (4) implies $\overline{W}(\Theta, P) \leq W(P) + \varepsilon$ and hence $\Theta \in \mathcal{U}^\varepsilon(P)$.

Since the sequence and subsequence were arbitrary, the convergence $\sup_{\Theta_n^\varepsilon \in \mathcal{U}^\varepsilon(P_n)} \varrho(\Theta_n^\varepsilon, \mathcal{U}^\varepsilon(P)) \to 0$ takes place. The proof is complete. □

Proposition 4.2. *If the assumptions A1–A5 are fulfilled and the parameterization has Property B, then for finite-dimensional normed spaces S and T optimal approximations corresponding to distributions P and P_n do exist, i.e. the sets $\mathcal{U}(P)$ and $\mathcal{U}(P_n)$ are non-empty.*

Proof. We will only prove that $\mathcal{U}(P)$ is not empty, the remaining is similar. Let $\{\Theta_n\}$ be an arbitrary minimizing sequence for the loss-function $\overline{W}(\cdot, P)$. By Lemma 4.1 this sequence is bounded and therefore relatively compact. This means, there exists a converging subsequence $\Theta_{n'} \to \Theta \in T^k$. From the other side,

$$\overline{W}(\Theta, P) = \lim_{n'} \overline{W}(\Theta_{n'}, P) = W_k(P),$$

which implies $\Theta \in \mathcal{U}(P)$. □

Proposition 4.3. *If the assumptions A1–A5 are fulfilled and the parameterization has properties A and B, then for finite-dimensional normed spaces S and T the distance between any sequence of P_n-optimal parameters and the class of P-optimal parameters converges to zero, i.e. the convergence*

$$\sup_{\Theta_n \in \mathcal{U}(P_n)} \varrho(\Theta_n, \mathcal{U}(P)) \to 0$$

takes place.

Proof. By Proposition 4.2 a P_n-optimal Θ_n exists and $\mathcal{U}(P) \neq \emptyset$. Now the result follows by applying Proposition 4.1 with $\varepsilon = 0$. □

References

Averous, J. and Meste, M. (1997). Median balls: an extension of the interquantile intervals to multivariate distributions. *J. Multivariate Anal.* **63**, 222–241.

Cuesta, J. A. and Matran, C. (1988). The Strong Law of Large Numbers for k-means and best possible nets of Banach valued random variables. *Probab. Theory Related Fields* **78**, 523–534.

Cuesta, J. A. and Matran, C. (1989). Uniform consistency of r-means. *Statist. Probab. Lett.* **6**, 65–71.

Cuesta-Albertos, J. A., Gordaliza, A. and Matran, C. (1989). Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.* **25**, 553–576.

Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions.* Springer-Verlag.

Käärik, M. (2000). Approximation of distributions by sphere. In: *Multivariate Statistics. Proceedings of the 6th Tartu Conference*, Eds: T. Kollo et al., VSP/TEV, Vilnius-Utrecht, 61–66.

Käärik, M. and Pärna, K. (2003). Approximation of distributions by parametric sets. *Acta Appl. Math.* **78**, 175–183.

Lember, J. and Pärna, K. (1999). Strong consistency of k-centres in reflexive spaces. In: *Probability Theory and Mathematical Statistics. Proceedings of the Seventh Vilnius Conference (1998)*, Eds: B. Grigelionis et al., VSP/TEV, Vilnius-Utrecht, 441–452.

Lember, J. (1999). *Consistency of Empirical k-Centres.* PhD dissertation, Tartu University Press, Tartu.

Pärna, K. (1986). Strong consistency of k-means clustering criterion in separable metric spaces. *Tartu Riikl. Ül. Toimetised* **733**, 86–96.

Pärna, K. (1988). On the stability of k-means clustering in metric spaces. *Tartu Riikl. Ül. Toimetised* **798**, 19–36.

Pärna, K. (1990). On the existence and weak convergence of k-centres in Banach spaces. *Tartu Riikl. Ül. Toimetised* **893**, 17–28.

Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9**, 135–140.

Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.* **33**, 659–680.

Varadarajan, V. S. (1958). On the convergence of probability distributions. *Sankhyā* **19**, 23–26.

Zhang, D. and Zhu, L. (1993). PP K-mean clustering. *Systems Sci. Math. Sci.* **6**, 289–295.

Institute of Mathematical Statistics, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia

E-mail address: meelis.kaarik@ut.ee

E-mail address: kalev.parna@ut.ee