
DOI: https://doi.org/10.4213/tvp2411
QUANTIZATION FOR PROBABILITY MEASURES IN THE PROKHOROV METRIC1)

Для распределения вероятностей $P$ на $\mathbb{R}^d$ и $n \in \mathbb{N}$ рассматривается величина $e_n = \inf \pi(P,Q)$, где $\pi$ — метрика Прохорова и инфимум берется по всем дискретным распределением $Q$ таким, что $|\text{supp}(Q)| \leq n$. Изучаются решения $Q$ этой задачи минимизации, свойства устойчивости и состоятельности эмпирических оценок. Для некоторых классов распределений определяется точная скорость сходимости к нулю ошибки $n$-квантования $e_n$ при $n \to \infty$.

Ключевые слова и фразы: многомерное квантование, метрика Ки Фан, метрика Прохорова, оптимальные квантизаторы, эмпирические меры, асимптотическая ошибка квантования, энтропия, размерность квантования.

1. Introduction

Let $X$ be an $\mathbb{R}^d$-valued random variable with distribution $P$. For $n \in \mathbb{N}$, let $\mathcal{F}_n$ be the set of all Borel measurable maps $f: \mathbb{R}^d \to \mathbb{R}^d$ with $|f(\mathbb{R}^d)| \leq n$, where $|A|$ denotes the cardinality of $A \subset \mathbb{R}^d$. The elements of $\mathcal{F}_n$ are called $n$-quantizers. For each $f \in \mathcal{F}_n$, $f(X)$ gives a quantized version of $X$. The quantization problem is to find an $n$-optimal quantizer for $P$ or at least an asymptotically $n$-optimal sequence of quantizers as $n \to \infty$, where the optimality depends upon the measure for the quantization error. In electrical engineering this problem arises in the context of coding speech and visual signals effectively. For these applications in communication and information theory we refer to [10]. In statistics quantizers can be used as a model for the grouping of data. See, e.g., [21] and [4]. Quantization also seems to be a promising tool in recent developments in numerical probability. See, e.g., [18].

Much of the previous work is based on $L_r$-metrics as measure for the quantization error. In this case the minimal $n$th quantization error of $P$ is

---

1) Additional notes or references related to the topic.
given by
\[
 e_{n,r}(P) = \begin{cases} 
\inf \left\{ (\mathbb{E}\|X - f(X)\|)^{1/r} : f \in \mathcal{F}_n \right\}, & 0 < r < \infty, \\
\inf \left\{ \text{ess sup} \|X - f(X)\| : f \in \mathcal{F}_n \right\}, & r = \infty.
\end{cases}
\]

The idea of quantization is enlightened by the following result which shows how optimal quantizers approximate the original distribution \(P\). Let \(\rho_r\) denote the \(L_r\)-Wasserstein metric (\(L_\infty\)-minimal metric if \(r = \infty\)), and let \(f\) be an \(n\)-optimal quantizer. Then
\[
e_{n,r}(P) = \inf \{\rho_r(P, Q) : |\text{supp}(Q)| \leq n\} = \rho_r(P, P^f),
\]
where \(P^f\) denotes the image measure. For the mathematical aspects of the quantization problem for these metrics, one can consult [12].

In this paper we use the Ky Fan distance as measure of performance. Thus if \(\| \cdot \|\) denotes any norm on \(\mathbb{R}^d\) and \(\kappa(Y, Z)\) is the Ky Fan metric,
\[
\kappa(Y, Z) = \inf \{ \varepsilon > 0 : P\{|Y - Z| > \varepsilon\} \leq \varepsilon\},
\]
then the \(n\)th quantization error for \(P\) (or \(X\)) is defined by
\[
e_n(P) = \inf \{\kappa(X, f(X)) : f \in \mathcal{F}_n\}.
\]
A quantizer \(f \in \mathcal{F}_n\) is called \(n\)-optimal for \(P\) if
\[
e_n(P) = \kappa(X, f(X)).
\]
This amounts in the maximization of the probability for a certain (small) deviation between \(X\) and \(f(X)\) and yields seemingly an attractive feature of quantizers. Now optimal quantizers provide best approximations of \(P\) in the Prokhorov metric denoted by \(\pi\). If \(f_n\) is an \(n\)-optimal quantizer, then
\[
e_n(P) = \inf \{\pi(P, Q) : |\text{supp}(Q)| \leq n\} = \pi(P, P_\pi^f)
\]
(see Lemma 2.2). This shows that the quantization error rules the rate of convergence of \(P_{n}\) toward the original distribution \(P\) in the Prokhorov metric.

From a statistical point of view, quantization with respect to the Ky Fan metric can be seen as a particular type of trimming in \(L_\infty\)-based quantization (cf. [5]). While \(L_\infty\)-quantization requires \(P\) to have a compact support, Ky Fan quantization does not.

The paper is organized as follows. Section 2 presents the basic features of the quantization problem including equivalent formulations as optimal location problem for \(n\)-point sets and as an approximation problem for \(P\) with respect to the Prokhorov metric, existence of \(n\)-optimal quantizers, necessary conditions for \(n\)-optimality, and a uniqueness result for optimal quantizers in the one-dimensional case. Section 3 contains stability (or continuity) properties and applications to the consistency of empirical estimators. In
Sections 4 and 5 we investigate the asymptotic behavior of the nth quantization error $e_n(P)$ as $n \to \infty$. We derive the exact rate of convergence to zero of $e_n(P)$ for some classes of distributions $P$. Approximate rates are obtained in terms of the quantization dimension of $P$ which is shown to coincide with the epsilon entropy dimension of $P$. It turns out that the rate of convergence of $e_n(P)$ for absolutely continuous distributions with unbounded support is distribution dependent — in contrast with the behavior of the $L_r$-quantization error.

2. The quantization problem

2.1. Basic facts. The investigation of the quantization problem requires the concept of Voronoi partitions. If $x \in \mathbb{R}^d$ and $A$ is a nonempty subset of $\mathbb{R}^d$, then the distance from $x$ to $A$ is given by

$$d_A(x) = d(x, A) = \inf_{a \in A} ||x - a||.$$ 

In case $d = 1$, the underlying norm is throughout the absolute value. Consider a nonempty finite (or locally finite) subset $\alpha$ of $\mathbb{R}^d$. The Voronoi region generated by $a \in \alpha$ is defined by

$$W(a \mid \alpha) = \{x \in \mathbb{R}^d : \|x - a\| = d(x, \alpha)\}.$$ (2.1)

The Voronoi regions $W(a \mid \alpha)$ are closed and star-shaped relative to $a$ and the Voronoi diagram $\{W(a \mid \alpha) : a \in \alpha\}$ of $\alpha$ provides a covering of $\mathbb{R}^d$. A Borel measurable partition $\{A_a : a \in \alpha\}$ of $\mathbb{R}^d$ is called a Voronoi partition with respect to $\alpha$ if

$$A_a \subset W(a \mid \alpha) \text{ for every } a \in \alpha.$$ (2.2)

We start with two equivalent formulations of the quantization problem with respect to the Ky Fan error measure. For $g : \mathbb{R}^d \to \mathbb{R}$ Borel measurable, define

$$||g||_0 = ||g||_L(P) = \inf\{\varepsilon > 0 : P\{|g| > \varepsilon\} < \varepsilon\}.$$ (2.3)

Observe that

$$||g_1 + g_2||_0 \leq ||g_1||_0 + ||g_2||_0,$$

$$||cg||_0 \geq |c||g||_0 \quad \text{if } c \in \mathbb{R}, \quad |c| \leq 1,$$

$$||cg||_0 \leq |c||g||_0 \quad \text{if } |c| \geq 1.$$ (2.4)

For fixed $n \in \mathbb{N}$, searching for an $n$-optimal quantizer is equivalent to the $n$-centers problem with respect to $\| \cdot \|_0$.

Lemma 2.1. (a) Let $\alpha \subset \mathbb{R}^d$ with $1 \leq |\alpha| \leq n$, and let $f \in \mathcal{F}_n$ with $f(\mathbb{R}^d) = \alpha$. Then

$$\kappa(X, f(X)) \geq \|d_\alpha\|_0.$$
If the sets \( \{ f = a \}, a \in \alpha, \) form a Voronoi partition with respect to \( \alpha, \) then
\[
\kappa(X, f(X)) = \| d_\alpha \|_0.
\]

(b) \( e_n(P) = \inf \{ \| d_\alpha \|_0 : \alpha \subset \mathbb{R}^d, \ 1 \leq |\alpha| \leq n \}. \)

Proof. (a) For \( f \in \mathcal{F}_n \) with \( f(\mathbb{R}^d) = \alpha, \) let \( A_a = \{ f = a \}, a \in \alpha. \)

Then
\[
P\{\| X - f(X) \| > \varepsilon \} = \sum_{a \in \alpha} P\left( A_a \cap \{ x \in \mathbb{R}^d : \| x - a \| > \varepsilon \} \right)
\]
\[
\geq \sum_{a \in \alpha} P( A_a \cap \{ d_\alpha > \varepsilon \}) = P\{ d_\alpha > \varepsilon \}
\]
and hence \( \kappa(X, f(X)) \geq \| d_\alpha \|_0. \) If \( \{ A_a : a \in \alpha \} \) is a Voronoi partition of \( \mathbb{R}^d \) with respect to \( \alpha, \) we have
\[
P\{\| X - f(X) \| > \varepsilon \} = P\{ d_\alpha > \varepsilon \}
\]
and therefore \( \kappa(X, f(X)) = \| d_\alpha \|_0. \)

(b) is an immediate consequence of (a). Lemma 2.1 is proved.

The closed ball with center \( a \in \mathbb{R}^d \) and radius \( r \geq 0 \) is denoted by
\[
B(a, r) = \{ x \in \mathbb{R}^d : \| x - a \| \leq r \}.
\]

Since
\[
\| d_\alpha \|_0 = \inf \left\{ \varepsilon \geq 0 : P \left( \bigcup_{a \in \alpha} B(a, \varepsilon) \right) \geq 1 - \varepsilon \right\},
\]
the quantization problem with respect to \( \kappa \) may be seen as a stochastic covering problem (concerning «coverings» by balls of equal radius).

A set \( \alpha \subset \mathbb{R}^d \) with \( 1 \leq |\alpha| \leq n \) is called \( n \)-optimal set of centers for \( P \) if
\[
e_n(P) = \| d_\alpha \|_0.
\]

Let \( \mathcal{C}_n(P) \) denote the set of all \( n \)-optimal sets of centers for \( P. \) Notice that from \( \alpha \in \mathcal{C}_n(P) \) one can construct an \( n \)-optimal quantizer according to Lemma 2.1(a).

The quantization problem for \( P \) is further equivalent to the problem of approximating \( P \) by a discrete probability with at most \( n \) supporting points. Here the underlying distance is the Prokhorov metric which reflects the fact that this metric is minimal relative to the Ky Fan metric. For Borel probability measures \( P_1, P_2 \) on \( \mathbb{R}^d, \) the Prokhorov metric is defined by
\[
\pi(P_1, P_2) = \inf \{ \varepsilon \geq 0 : P_1(A) \leq P_2(\{ d_A \leq \varepsilon \}) + \varepsilon \text{ for all } A \in \mathcal{B}(\mathbb{R}^d) \}, \quad (2.5)
\]
where \( \mathcal{B}(\mathbb{R}^d) \) denotes the Borel \( \sigma \)-algebra on \( \mathbb{R}^d. \) The minimality of \( \pi \) relative to \( \kappa \) means
\[
\pi(P_1, P_2) = \inf \{ \kappa(Y, Z) : P^Y = P_1, \ P^Z = P_2 \}, \quad (2.6)
\]
where all random variables are defined on a nonatomic probability space. The representation (2.6) is a particular case of the Strassen-Dudley theorem (cf. [7] and [9, Corollary 11.6.4]). Let $\mathcal{P}_n$ denote the set of all discrete probabilities $Q$ on $\mathbb{R}^d$ with $|\text{supp}(Q)| \leq n$.

**Lemma 2.2.** (a) Let $\alpha \subset \mathbb{R}^d$ with $1 \leq |\alpha| \leq n$, $Q \in \mathcal{P}_n$ with $Q(\alpha) = 1$, and $f \in \mathcal{F}_n$ with $f(\mathbb{R}^d) = \alpha$. Then

$$\pi(P, Q) \geq \|d_{\alpha}\|_0 \quad \text{and} \quad \pi(P, P') \leq \kappa(X, f(X)).$$

If the sets $\{f = a\}$, $a \in \alpha$, form a Voronoi partition of $\mathbb{R}^d$ with respect to $\alpha$, then

$$\pi(P, Q) \geq \pi(P, P') = \kappa(X, f(X)).$$

(b) $e_n(P) = \inf\{\pi(P, Q) : Q \in \mathcal{P}_n\} = \inf\{\pi(P, P') : f \in \mathcal{F}_n\}$.

**Proof.** (a) Let $\varepsilon \geq \pi(P, Q)$. Then

$$Q(A) \leq P\{d_{\alpha} \leq \varepsilon\} + \varepsilon, \quad A \in \mathcal{B}(\mathbb{R}^d).$$

In particular,

$$1 = Q(\alpha) \leq P\{d_{\alpha} \leq \varepsilon\} + \varepsilon$$

which gives $\|d_{\alpha}\|_0 \leq \pi(P, Q)$. The second inequality follows from (2.6). Now assume that $\{\{f = a\} : a \in \alpha\}$ is a Voronoi partition with respect to $\alpha$. Since by the preceding part of the proof $\|d_{\alpha}\|_0 \leq \pi(P, P') \leq \kappa(X, f(X))$, it follows from Lemma 2.1(a) that $\|d_{\alpha}\|_0 = \pi(P, P') = \kappa(X, f(X))$ and hence $\pi(P, Q) \geq \pi(P, P')$.

(b) follows immediately from (a). Lemma 2.2 is proved.

**2.2. Optimal quantizers.** Next we deal with $n$-optimality. Under the assumption that $P$ vanishes on spheres some necessary conditions for $n$-optimality can be given.

**Theorem 2.1** (necessary conditions for optimality). Let $\alpha \in \mathcal{C}_n(P)$, $s = e_n(P)$, and let $\{A_a : a \in \alpha\}$ be a Voronoi partition of $\mathbb{R}^d$ with respect to $\alpha$. Assume

$$P(\partial B(b, r)) = 0 \quad \text{for all} \quad b \in \mathbb{R}^d, \ r > 0. \ (2.7)$$

Then

$$|\alpha| = n, \quad P(A_a \cap B(a, s)) > 0 \quad \text{for every} \quad a \in \alpha,$n

and

$$P\left(d_{\beta} \leq s \mid \bigcup_{a \in \beta} A_a\right) = \max \left\{P\left(d_{\gamma} \leq s \mid \bigcup_{a \in \beta} A_a\right) : \gamma \subset \mathbb{R}^d, \ 1 \leq |\gamma| \leq m\right\}$$

for every $\beta \subset \alpha$ with $|\beta| = m$. 


Proof. Note that $s < 1$. Let $\delta = \{a \in \alpha : P(A_a \cap B(a, s)) > 0\}$ and assume $|\delta| < n$. Since

$$1 - s \leq P\{d_\alpha \leq s\} = \sum_{a \in \alpha} P(A_a \cap \{d_\alpha \leq s\})$$

$$= \sum_{a \in \alpha} P(A_a \cap B(a, s)) = \sum_{a \in \delta} P(A_a \cap B(a, s))$$

$$= \sum_{a \in \delta} P(A_a \cap \{d_\delta \leq s\}) \leq P\{d_\delta \leq s\},$$

one gets $|\delta| \geq 1$ and $\|d_\delta\|_0 \leq s$, hence $\|d_\delta\|_0 = s$ and $\delta \in \mathcal{C}_n(P)$. Under (2.7), the distribution of $d_F$ under $P$ is continuous for any finite set $F \subset \mathbb{R}^d$ and thus $\|d_F\|_0$ is a unique solution in $(0, 1]$ of the equation $P\{d_F \leq \varepsilon\} = 1 - \varepsilon$. This gives

$$P\{d_\delta \leq s\} = 1 - s < 1.$$ 

Choose a compact set $K \subset \{d_\delta > s\}$ with $P(K) > 0$. Since $K \subset \bigcup_{a \in K} B(a, s)$, we can find a finite subset $F$ of $K$ such that $K \subset \bigcup_{a \in F} B(a, s)$. Therefore, there exists a point $b \in F$ such that $P(K \cap B(b, s)) > 0$. It follows

$$P\{d_{\delta \cup \{b\}} \leq s\} \geq P\{\{d_\delta \leq s\} \cup (K \cap B(b, s))\} = P\{d_\delta \leq s\} + P(K \cap B(b, s))$$

$$> P\{d_\delta \leq s\} = 1 - s.$$ 

By (2.7), this implies $\|d_{\delta \cup \{b\}}\|_0 < s$, a contradiction.

As for the assertion concerning $\beta$, assume that for some $\beta \subset \alpha$, $1 \leq |\beta| = m \leq n$ there exists $\gamma \subset \mathbb{R}^d$, $1 \leq |\gamma| \leq m$, such that

$$P\{d_\beta \leq s \mid \bigcup_{a \in \gamma} A_a\} < P\{d_\gamma \leq s \mid \bigcup_{a \in \beta} A_a\}.$$ 

Then

$$1 - s = P\{d_\alpha \leq s\} = \sum_{a \in \beta} P(A_a \cap \{d_\alpha \leq s\}) + \sum_{a \in \alpha \setminus \beta} P(A_a \cap \{d_\alpha \setminus \beta \leq s\})$$

$$< \sum_{a \in \beta} P(A_a \cap \{d_\gamma \leq s\}) + \sum_{a \in \alpha \setminus \beta} P(A_a \cap \{d_\alpha \setminus \beta \leq s\}) \leq P\{d_{\gamma \cup (\alpha \setminus \beta)} \leq s\}.$$ 

This yields $\|d_{\gamma \cup (\alpha \setminus \beta)}\|_0 < s$, a contradiction. Theorem 2.1 is proved.

Recall that absolutely continuous distributions (with respect to Lebesgue measure) satisfy (2.7). The same holds, for instance, for self-similar distributions under the $L_2$-norm (cf. [11, Proposition 4.1]).

Unlike $L_\infty$-quantization in the preceding theorem one cannot replace (2.7) by the condition $|\text{supp}(P)| \geq n$. This is exhibited by the following simple example.

Example 2.1. Let $d = 1$ and $P = \frac{1}{4} \sum_{i=1}^4 \delta_i$, where $\delta_x$ is the point mass at $x$. (Recall that $\|x\| = |x|$.) Then, for $n = 2$, we have $e_2(P) = \frac{1}{2}$ and the one-point set $\alpha = \{\frac{3}{2}\}$ is 2-optimal for $P$. 


To ensure the existence of $n$-optimal quantizers we will employ the continuity of the map $\alpha \mapsto \|d_\alpha\|_0$ with respect to the Hausdorff metric. For nonempty compact subsets $A, B$ of $\mathbb{R}^d$ the Hausdorff metric is given by

$$
H(A, B) = \max\left\{ \max_{a \in A} d(a, B), \max_{b \in B} d(b, A) \right\} = \sup_{x \in \mathbb{R}^d} |d(x, A) - d(x, B)|. \tag{2.8}
$$

For $g: \mathbb{R}^d \to \mathbb{R}$ Borel measurable, let

$$
\|g\|_r = \|g\|_{L_r(P)} = \left\{ \left( \int |g|^r dP \right)^{1/r} \right\} \quad \text{for } 0 < r < \infty, \quad \inf\{\varepsilon > 0: P\{|g| > \varepsilon\} = 0\}, \quad r = \infty.
$$

**Lemma 2.3.** If $A, B \subset \mathbb{R}^d$ are compact and nonempty, then

$$
\|d_A\|_0 - \|d_B\|_0 < H(A, B)^s \quad \text{for every } s \in [0, 1].
$$

**Proof.** Using the triangle inequality (see (2.4)) and the $L_r$-bounds for $\| \cdot \|_0$, one obtains for $0 < r < \infty$

$$
\|d_A\|_0 - \|d_B\|_0 < \|d_A - d_B\|_0 < \|d_A - d_B\|^{r/(r+1)} \leq H(A, B)^{r/(r+1)}.
$$

Lemma 2.3 is proved.

The compactness property appearing in the following theorem is important for stability results (see Section 3).

**Theorem 2.2** (existence, level compactness). (a) $\mathcal{C}_n(P) \neq \emptyset$.

(b) Assume that $P$ satisfies condition (2.7). Then $e_n(P) < e_{n-1}(P)$ ($e_0(P) := 1$) and the level set

$$
L(c) = \{\alpha \in \mathbb{R}^d: 1 \leq |\alpha| \leq n, \|d_\alpha\|_0 \leq c\}
$$

is $H$-compact for every $c$ with $e_n(P) \leq c < e_{n-1}(P)$. In particular, $\mathcal{C}_n(P)$ is $H$-compact.

**Proof.** **Step 1.** In the first step we assume $e_n(P) < e_{n-1}(P)$. Let $e_n(P) \leq c < e_{n-1}(P)$ and choose $0 < r < \infty$ (depending on $n$, $P$, and $c$) such that

$$
P(B(0, r - c)) > c \quad \text{and} \quad P(B(0, 2r)^c) < e_{n-1}(P) - c.
$$

Let $\alpha \in L(c)$. Then $|\alpha| = n$. Let $\alpha = \{a_1, \ldots, a_n\}$ and assume without loss of generality that $\|a_1\| \leq \cdots \leq \|a_n\|$. Then $\|a_1\| < r$. Otherwise $c > P\{d_a > c\} \geq P(B(0, r - c))$, a contradiction. Furthermore, $\|a_n\| \leq 5r$ (assuming now $n \geq 2$). Otherwise

$$
\|x - a_1\| \leq \|x - a_n\| 1_{B(0, 2r)}(x) + \|x\| 1_{B(0, 2r)^c}(x)
$$
for every $x \in \mathbb{R}^d$. If $\{A_1, \ldots, A_n\}$ with $A_i = A_\alpha$, denotes a Voronoi partition of $\mathbb{R}^d$ with respect to $\alpha$, $\beta = \{a_1, \ldots, a_{n-1}\}$, and $s = \|d_\beta\|_0$, then

$$s > \|d_\alpha\|_0,$$

$$P\{d_\alpha \geq s\} \leq P\{d_\alpha > \|d_\alpha\|_0\} \leq \|d_\alpha\|_0 < c,$$

$$P\{d_\beta \geq s\} \geq s,$$

$$P(A_n \cap \mathring{B}(a_1, s)^c) \leq P(A_n \cap \mathring{B}(a_n, s)^c) + P(B(0, 2r)^c),$$

where $\mathring{B}(a, s) = \{\|x - a\| < s\}$. Consequently,

$$e_{n-1}(P) < s = P\{d_\beta \geq s\} = \sum_{i=1}^{n} P(A_i \cap \{d_\beta \geq s\})$$

$$\leq \sum_{i=1}^{n-1} P(A_i \cap \mathring{B}(a_i, s)^c) + P(A_n \cap \mathring{B}(a_1, s)^c)$$

$$\leq \sum_{i=1}^{n} P(A_i \cap \mathring{B}(a_i, s)^c) + P(B(0, 2r)^c)$$

$$< P\{d_\alpha \geq s\} + e_{n-1}(P) - c \leq e_{n-1}(P),$$

a contradiction. We thus obtain

$$L(c) \subset \{\alpha \subset B(0, 5r): 1 \leq |\alpha| \leq n\}.$$

Since the latter set is $H$-compact, it follows from Lemma 2.3 that $L(c)$ is $H$-compact, too. This implies that $\mathcal{C}_n(P) = L(e_n(P))$ is not empty and $H$-compact.

**Step 2.** In the second step we prove (a) and (b). Clearly, we have $e_1(P) < 1 = e_0(P)$. By Step 1, this yields $\mathcal{C}_1(P) \neq \emptyset$. If $e_n(P) = e_1(P)$, then $\mathcal{C}_1(P) \subset \mathcal{C}_n(P)$ and hence $\mathcal{C}_n(P) \neq \emptyset$. If $e_n(P) < e_1(P)$, let

$$k = \max\{2 \leq j \leq n: e_n(P) < e_{j-1}(P)\}.$$

Then

$$e_n(P) = \cdots = e_k(P) < e_{k-1}(P)$$

and by Step 1, this implies $\emptyset \neq \mathcal{C}_k(P) \subset \mathcal{C}_n(P)$. Thus assertion (a) is proved. Now assume (2.7). Then by part (a) and Theorem 2.1, $e_n(P) < e_{n-1}(P)$. The remaining assertions of (b) follow from Step 1. Theorem 2.2 is proved.

We know from Example 2.1 that one cannot omit condition (2.7) in part (b) of the preceding theorem.

A bijective mapping $T: \mathbb{R}^d \to \mathbb{R}^d$ is called similarity transformation if there exists $c \in (0, \infty)$, the scaling number, such that $\|Tx - Ty\| = c\|x - y\|$ for every $x, y \in \mathbb{R}^d$. The effect of similarity transformations of $P$ is as follows.
Lemma 2.4. Let $T: \mathbb{R}^d \to \mathbb{R}^d$ be a similarity transformation with scaling number $c$, and let $\alpha \subset \mathbb{R}^d$ be a nonempty finite set.

(a) $\|d_{T(\alpha)}\|_{L_0(P_T)} = \|cd_{\alpha}\|_{L_0(P)}$, 

$e_n(P_T) = \inf\{\|cd_{\beta}\|_{L_0(P)}: \beta \subset \mathbb{R}^d, 1 \leq |\beta| \leq n\}$,

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \|d_{\alpha}\|_{L_0(P)} = \|d_{\alpha}\|_{L_\infty(P)}.$$ 

(b) If $c = 1$, then $e_n(P_T) = T_e_n(P)$, $e_n(P_T) = e_n(P)$.

Proof. (a) Since $\{d_{T(\alpha)} > \varepsilon\} = T\{cd_{\alpha} > \varepsilon\}$, we have $\|d_{T(\alpha)}\|_{L_0(P_T)} = \|cd_{\alpha}\|_{L_0(P)}$. This implies the second assertion, and the third one is obvious. (b) follows from (a).

Remark. Let $X_d$ denote the d-dimensional Lebesgue measure. For a compact set $A \subset \mathbb{R}^d$ with $X_d(A) > 0$, let $U(A)$ denote the uniform distribution on $A$. In some cases (small $d$, small $n$ or special norms), $e_n(P)$ for $P = U(A)$ is less than or equal to the «nth packing radius» of $A$. If this happens, one gets a sufficient condition for $n$-optimality as follows. Let $s = s_n$ be a unique solution in $(0, 1)$ of the equation

$$\frac{ne^d\lambda^d(B(0, 1))}{\lambda^d(A)} = 1 - \varepsilon.$$ (2.10)

For any $\alpha \subset \mathbb{R}^d$ with $1 \leq |\alpha| \leq n$, we have

$$\|d_{\alpha}\|_0 \geq \inf \left\{ \varepsilon \geq 0: \sum_{a \in \alpha} P(B(a, \varepsilon)) \geq 1 - \varepsilon \right\}$$

$$\geq \inf \left\{ \varepsilon \geq 0: \frac{ne^d\lambda^d(B(0, 1))}{\lambda^d(A)} \geq 1 - \varepsilon \right\} = s$$

and thus

$$e_n(P) \geq s.$$ 

Suppose there exists a subset $\alpha$ of $\mathbb{R}^d$ with $|\alpha| = n$ such that $\{B(a, s): a \in \alpha\}$ is a packing in $A$, i.e., $\bigcup_{a \in \alpha} B(a, s) \subset A$ and $\lambda^d(B(a, s) \cap B(b, s)) = 0$ for $a, b \in \alpha, a \neq b$, then $\|d_{\alpha}\| \leq s$. This implies $\alpha \in \mathcal{E}_n(P)$ and $e_n(P) = s$.

Example 2.2 (uniform distribution on a cube and cube quantizer). Let $P = U([0, b]^d)$ with $b > 0$ and consider a tessellation of $[0, b]^d$ consisting of $n = k^d$ translates $C_1, \ldots, C_n$ of the cube $C_1 = [0, b/k]^d$. Denote by $a_i$ the midpoint of $C_i$.

Then $\alpha = \{a_1, \ldots, a_n\} = \left\{\frac{(2i - 1)b}{2k}: i = 1, \ldots, k\right\}^d$.

Let $f = f_n \in \mathcal{F}_n$ with $f = \sum_{i=1}^n a_i 1_{C_i}$, $P$-a.s. Then

$$\kappa(X, f(X)) = \inf \left\{ \varepsilon \geq 0: \sum_{i=1}^n P(C_i \cap B(a_i, \varepsilon)) \geq 1 - \varepsilon \right\}$$
= \inf \left\{ \varepsilon \geq 0 : nb^{-d}\lambda^d(C_1 \cap B(a_1, \varepsilon)) \geq 1 - \varepsilon \right\}
= \inf \left\{ \varepsilon \geq 0 : U(C_1)(B(a_1, \varepsilon)) \geq 1 - \varepsilon \right\}
= \|d_{\{a_1\}}\|_{L_0(U(C_1))} = \|n^{-1/d}d_{\{u\}}\|_{L_0(P)},
where \(u = (b/2)(1, \ldots, 1).\) Here the last equality follows from Lemma 2.4. It shows that the error of the cube quantizer \(f\) is of optimal order \(n^{-1/d}\) (see Section 4).

Now let the underlying norm on \(\mathbb{R}^d\) be the \(l_p\)-norm for \(1 \leq p \leq \infty.\) Then \(C_i = [0, b]^d \cap W(a_i|\alpha)\) for every \(i\) and hence by Lemmas 2.1(a) and 2.2(a)
\[\kappa(X, f(X)) = \|d_{\alpha}\|_{L_0(P)} = \pi(P, P^f),\]
where \(P^f = (b^d/n)\sum_{i=1}^n \delta_{a_i}.\) If, for instance, \(p = 2, d = 2,\) and \(b = 1,\) then by Lemma 2.4,
\[\lim_{n \to \infty} n^{1/2}\pi(P, P^{f_{n}}) = \|d_{\{u\}}\|_{L_\infty(P)} = \frac{1}{\sqrt{2}} = 0.7071\ldots\]
This may be compared with the optimal constant given in Theorem 4.5.

Let \(s = s_n\) be a unique solution of (2.10) in \((0, 1)\) with \(A\) replaced by \([0, b]^d.\) Then \(\{B(a, s) : a \in \alpha\}\) is a packing in \([0, b]^d\) if and only if \(s \leq b/(2k) = b/(2n^{1/d})\) which is equivalent to
\[\frac{\lambda^d(B(0, 1))}{2^d} \geq 1 - \frac{b}{2n^{1/d}}.\] (2.11)

For the \(l_\infty\)-norm, (2.11) takes the form
\[1 \geq 1 - \frac{b}{2n^{1/d}}\]
which holds for every \(n.\) Therefore, \(f\) is an \(n\)-optimal quantizer, \(\alpha \in \mathcal{C}_n(P),\)
and \(e_n(P) = s\) for every \(n = \kappa^d.\) In particular, for \(d = 1\) one obtains
\[e_n(U([0, b])) = \frac{b}{2n + b}, \quad n \in \mathbb{N}.\] (2.12)
Notice that in case \(d = 1\) there are \(n\)-optimal sets of centers other than \(\alpha.\)
The sets \(\beta_x = \{x + bi/n : i = 0, 1, \ldots, n - 1\}\) with
\[\frac{b}{2n + b} \leq x \leq \frac{b}{n} - \frac{b}{2n + b}\]
also satisfy \(\beta_x \in \mathcal{C}_n(P)\) (cf. also Proposition 2.1).

2.3. Uniqueness in one dimension. In the one-dimensional case there is a reasonable criterion for the uniqueness of \(n\)-optimal sets of centers. Recall that a continuous probability \(P\) on \(\mathbb{R}\) is (strictly) unimodal about some mode \(m \in \mathbb{R}\) if \(P\) possesses a \(\lambda\)-density \(h\) such that \(h\) is (strictly) increasing on \((-\infty, m)((-\infty, m) \cap \text{supp}(P))\) and (strictly) decreasing on \((m, \infty)((m, \infty) \cap \text{supp}(P)).\) Introduce the concentration function
\[C_P(y) = \sup\{P([x, x + y]) : x \in \mathbb{R}\} \quad \text{for every} \quad y \geq 0.\]
Proposition 2.1. Let \( d = 1 \) and suppose \( P \) is continuous and unimodal. Let \( s = e_n(P) \). Then \( s \) is the unique solution in \((0, 1)\) of the equation
\[
C_P(2ns) = 1 - \varepsilon. \tag{2.13}
\]
If \( \alpha = \{x + (2i - 1)s : 1 \leq i \leq n\} \), where \( x \in \mathbb{R} \) satisfies
\[
C_P(2ns) = P([x, x + 2ns]),
\]
then \( \alpha \in \mathcal{C}_n(P) \). If \( P \) is strictly unimodal, then
\[
|\mathcal{C}_n(P)| = 1.
\]

Proof. Let \( \beta \in \mathcal{C}_n(P) \). By Theorem 2.1, we have \(|\beta| = n\) and \( P([b - s, b + s]) > 0 \) for every \( b \in \beta \). Note further that \( \text{supp}(P) \) is convex.

Let \( \beta = \{b_1, \ldots, b_n\} \) with \( b_1 < \cdots < b_n \). Shifting the \( b_i \)'s in direction of a mode we can find a set \( \gamma = \{c_1, \ldots, c_n\} \), \( c_1 < \cdots < c_n \), such that max \((c_{i+1} - c_i) \leq 2s\) and \( P\{d_\beta \leq s\} \leq P\{d_\gamma \leq s\} \). Then \( \gamma \in \mathcal{C}_n(P) \) and
\[
1 - s = P\{d_\gamma \leq s\} = P([c_1 - s, c_1 + s])
\leq C_P(c_1 + s - (c_1 - s)) \leq C_P(2ns) = P\{d_\alpha \leq s\}.
\]

This implies \( \alpha \in \mathcal{C}_n(P) \) and \( C_P(2ns) = 1 - s \). Moreover, since \( C_P \) is strictly increasing on \( \{0 < C_P < 1\} \), we get \( c_n + s - (c_1 - s) = 2ns \) and hence
\[
\gamma = \{c_1 - s + (2i - 1)s : 1 \leq i \leq n\}.
\]

Now assume that \( P \) is strictly unimodal. Then \( \max(b_{i+1} - b_i) \leq 2s \).

Otherwise, if \( b_i + s < b_{i+1} - s \) for some \( 1 \leq i \leq n - 1 \), set
\[
b'_i = \begin{cases} b_i, & m \leq b_i + s, \\
\quad m - s, & b_i + s < m < b_{i+1} - s, \\
b_{i+1} - 2s, & m \geq b_{i+1} - s,
\end{cases}
\]
and \( \beta' = \{b_1, \ldots, b_i - 1, b'_i, b'_i + 1, b_{i+2}, \ldots, b_n\} \), where \( m \) denotes a mode of \( P \). The strict unimodality yields
\[
P\{d_\beta \leq s\} < P\{d_\beta' \leq s\},
\]
a contradiction. As above, one obtains
\[
\beta = \{b_1 - s + (2i - 1)s : 1 \leq i \leq n\}
\]
and hence
\[
C_P(2ns) = P\{d_\beta \leq s\} = P([b_1 - s, b_1 - s + 2ns]).
\]
Since \( P([x, x + 2ns]) \) has a unique maximizer \( x \in \mathbb{R} \), we deduce \(|\mathcal{C}_n(P)| = 1\). Proposition 2.1 is proved.
3. Stability properties and empirical versions

A stability property for the nth quantization error in terms of the Prokhorov metric follows immediately from Lemma 2.2(b). In fact, if $P_1, P_2$ are Borel probability distributions on $\mathbb{R}^d$, then

$$|e_n(P_1) - e_n(P_2)| \leq \pi(P_1, P_2)$$

(3.1)

for every $n \in \mathbb{N}$. A stability result for $n$-optimal approximations from $\mathcal{P}_n$ can also be based on the Prokhorov metric and the Hausdorff metric is appropriate for formulating a stability property for $n$-optimal sets of centers. Let $\mathcal{D}_n(P)$ denote the set of all $n$-optimal approximations $Q$ to $P$ from $\mathcal{P}_n$, i.e., $Q \in \mathcal{P}_n$ and

$$e_n(P) = \pi(P, Q).$$

Recall that the Prokhorov metric metrizes the weak topology.

Theorem 3.1. Let $P_k \overset{D}{\rightarrow} P$ (weak convergence) for Borel probabilities $P_k, P$ on $\mathbb{R}^d$ and assume that $P$ satisfies (2.7).

(a) Let $Q_k \in \mathcal{D}_n(P_k)$, $k \in \mathbb{N}$. Then the set of $\pi$-cluster points of the sequence $(Q_k)_{k \geq 1}$ is a nonempty subset of $\mathcal{D}_n(P)$ and

$$\inf_{Q \in \mathcal{D}_n(P)} \pi(Q_k, Q) \to 0 \quad \text{as} \quad k \to \infty.$$

(b) Let $\alpha_k \in \mathcal{C}_n(P_k)$, $k \in \mathbb{N}$. Then the set of $H$-cluster points of the sequence $(\alpha_k)_{k \geq 1}$ is a nonempty subset of $\mathcal{C}_n(P)$ and

$$\inf_{\alpha \in \mathcal{C}_n(P)} H(\alpha_k, \alpha) \to 0 \quad \text{as} \quad k \to \infty.$$

The preceding theorem can be derived from the following simple lemma.

Lemma 3.1. Let $(M, d)$ be a metric space, and let $f: M \to \mathbb{R}_+$ be a lower semicontinuous function. Suppose the level set

$$L(c) = \{x \in M: f(x) \leq c\}$$

is compact for some $c > \inf_{y \in M} f(y)$. Let

$$D = \{x \in M: f(x) = \inf_{y \in M} f(y)\},$$

and let $(x_k)_{k \geq 1}$ be a minimizing sequence in $M$ for $f$, i.e., $f(x_k) \to \inf_{y \in M} f(y)$. Then the set of cluster points of $(x_k)_{k \geq 1}$ is a nonempty subset of $D$ and

$$\inf_{x \in D} d(x_k, x) \to 0, \quad k \to \infty.$$

Proof of Theorem 3.1. (a) We show that the assertion follows from Lemma 3.1 applied to the metric space $(\mathcal{P}_n, \pi)$ and $f = \pi(P, \cdot)$. By Theorem 2.2, we have $e_n(P) < e_{n-1}(P)$. Consider the level set

$$L(c) = \{Q \in \mathcal{P}_n: \pi(P, Q) \leq c\}$$
for $e_n(P) < c < e_{n-1}(P)$. For $Q \in L(c)$ and $\alpha = \text{supp}(Q)$, it follows from Lemma 2.2(a) that $\|d_\alpha\|_{L_0(P)} \leq \pi(P, Q) \leq c$. Hence, by Theorem 2.2
$L(c) \subset \{Q \in \mathcal{P}_n: Q(B) = 1\}$ for some compact subset $B$ of $\mathbb{R}^d$. We deduce
$\pi$-compactness of $L(c)$. Furthermore, $(Q_k)_{k \geq 1}$ is minimizing sequence in $\mathcal{P}_n$ for $\pi(P, \cdot)$. In fact, since
$$e_n(P) \leq \pi(P, Q_k) \leq \pi(P, P_k) + \pi(P_k, Q_k) = \pi(P, P_k) + e_n(P_k)$$
and $\pi(P, P_k) + e_n(P_k) \to e_n(P)$, one gets
$$\pi(P, Q_k) \to e_n(P), \quad k \to \infty.$$ 
Thus, all conditions of Lemma 3.1 are fulfilled, and the assertion follows.

(b) This part follows from an application of Lemma 3.1 to $M = \{\alpha \in \mathbb{R}^d: 1 \leq |\alpha| \leq n\}$ equipped with the Hausdorff metric $H$ and $f: M \to \mathbb{R}_+, f(\alpha) = \|d_\alpha\|_{L_0(P)}$. Note first that $f$ is $H$-continuous. This follows from Lemma 2.3. Next by Theorem 2.2, the level set
$$L(c) = \{\alpha \in M: f(\alpha) \leq c\}$$
is $H$-compact for $e_n(P) < c < e_{n-1}(P)$. Finally, we show that $(\alpha_k)_{k \geq 1}$ is a minimizing sequence in $M$ for $f$. For $k \in \mathbb{N}$, let $\{A_{k, a}: a \in \alpha_k\}$ be a Voronoi partition of $\mathbb{R}^d$ with respect to $\alpha_k$. Set $Q_k = \sum_{a \in \alpha_k} P_k(A_k, a) \delta_a$. Then by Lemmas 2.1(a) and 2.2(a), $Q_k \in \mathcal{P}_n(P_k)$ and
$$e_n(P) \leq f(\alpha_k) \leq \pi(P, Q_k).$$
Hence, as shown in the proof of (a), $\pi(P, Q_k) \to e_n(P)$ which implies
$$f(\alpha_k) \to e_n(P), \quad k \to \infty.$$ 
Thus, we see that all assumptions of Lemma 3.1 are fulfilled, and the assertion follows. Theorem 3.1 is proved.

The stability results can be applied to the empirical analysis of the quantization problem. Let $X_1, X_2, \ldots$ be i.i.d. $\mathbb{R}^d$-valued random variables with distribution $P$, and let $\tilde{P}_k = k^{-1} \sum_{i=1}^k \delta_{X_i}$ be the empirical measure of $X_1, \ldots, X_k$.

**Corollary 3.1** (consistency). (a) $e_n(\tilde{P}_k) \to e_n(P)$ a.s. as $k \to \infty$ uniformly in $n \in \mathbb{N}$.

(b) Let $\tilde{Q}_k = \tilde{Q}_k(X_1, \ldots, X_k) \in \mathcal{P}_n(\tilde{P}_k)$, $k \in \mathbb{N}$, and assume that $P$ satisfies (2.7). Then
$$\inf_{Q \in \mathcal{P}_n(P)} \pi(\tilde{Q}_k, Q) \to 0 \text{ a.s.,} \quad k \to \infty.$$ 

(c) Let $\tilde{\alpha}_k = \tilde{\alpha}_k(X_1, \ldots, X_k) \in \mathcal{P}_n(\tilde{P}_k)$, $k \in \mathbb{N}$, and assume that $P$ satisfies (2.7). Then
$$\inf_{\alpha \in \mathcal{P}_n(P)} H(\tilde{\alpha}_k, \alpha) \to 0 \text{ a.s.,} \quad k \to \infty.$$
Proof. Since \( \pi(\hat{P}_k, P) \to 0 \) a.s., the assertions follow from Theorem 3.1 and (3.1).

Remark. Let \( Y_k = \sup_{n \geq 1} |e_n(\hat{P}_k) - e_n(P)| \). Since \( e_n(\hat{P}_k) = 0 \) for \( n > k \) and thus \( Y_k = (\sup_{n \leq k-1} |e_n(\hat{P}_k) - e_n(P)|) \vee e_k(P) \), it follows from the subsequent Theorem 4.3 that

\[
\liminf_{k \to \infty} k^{1/d} Y_k \geq \liminf_{k \to \infty} k^{1/d} e_k(P) > 0
\]

provided the absolutely continuous part of \( P \) does not vanish.

For \( P = U([0, 1]^d) \), Yukich and Shor [22], [20] determined the exact a.s. order of convergence of \( \pi(\hat{P}_k, P) \). Using (3.1), one gets a.s. rates for \( Y_k \), namely,

\[
\limsup_{k \to \infty} \frac{k}{\ln \ln k} Y_k \leq c(1) \quad \text{a.s.,} \quad d = 1,
\]

\[
\limsup_{k \to \infty} \frac{k^{1/2}}{\ln k} Y_k \leq c(2) \quad \text{a.s.,} \quad d = 2,
\]

\[
\limsup_{k \to \infty} \frac{k^{1/4}}{\ln^2 k} Y_k \leq c(d) \quad \text{a.s.,} \quad d \geq 3,
\]

where \( c(d) \) are finite constants. (For \( d = 1 \), this is an immediate consequence of Smirnov's law of the iterated logarithm for \( \sup_{x \in \mathbb{R}^d} |\hat{F}_k(x) - F(x)| \).)

The asymptotic behavior of \( \pi(\hat{P}_k, P) \) in the one-dimensional, nonuniform case has been investigated in [15]. As a consequence one has, for instance, for symmetric unimodal distributions \( P \) on \( \mathbb{R} \) with unbounded support and exponentially small tails that \( k^{1/2}(\ln k)^{-\alpha} Y_k \) is stochastically bounded for a suitable \( \alpha > 0 \).

4. Asymptotic quantization error: distributions with compact support

The \( n \)th quantization errors of any Borel probability distribution \( P \) on \( \mathbb{R}^d \) obviously satisfy

\[
\lim_{n \to \infty} e_n(P) = 0. \tag{4.1}
\]

In this section we determine the rate of convergence to zero of \( e_n(P) \) and constants. In particular, we look for power laws

\[
e_n(P) \approx n^{-1/D}, \quad n \to \infty, \tag{4.2}
\]

for some \( D \in (0, \infty) \). It is convenient to use the symbols \( \approx \) and \( \sim \), where \( a_n \approx b_n \) means \( a_n = O(b_n) \) and \( b_n = O(a_n) \) and \( a_n \sim b_n \) means \( a_n/b_n \to 1 \). If (4.2) holds, the lower and upper quantization coefficient of \( P \) are defined by

\[
Q(P) = \liminf_{n \to \infty} n^{1/D} e_n(P), \quad \overline{Q}(P) = \limsup_{n \to \infty} n^{1/D} e_n(P), \tag{4.3}
\]
respectively. If $Q(P)$ and $\overline{Q}(P)$ agree, then their common value is denoted by $Q(P)$ and called the quantization coefficient of $P$. The question for which probabilities the quantization coefficient exists is more difficult to answer for the Ky Fan metric than for $L_r$-metrics mainly due to the lack of suitable scaling properties.

In the sequel the $L_\infty$-metric serves as upper bound for Borel probability measures with compact support. For a probability $P$ on $\mathbb{R}^d$ with compact support let

$$e_{n,\infty}(P) = \inf \left\{ \|d_\alpha\|_{L_\infty(P)} : \alpha \subset \mathbb{R}^d, 1 \leq |\alpha| \leq n \right\}. \quad (4.4)$$

Then by Lemma 2.1

$$e_n(P) \leq e_{n,\infty}(P). \quad (4.5)$$

Since

$$\|d_\alpha\|_\infty = \max_{x \in \text{supp}(P)} d_\alpha(x) = \min \left\{ \varepsilon \geq 0 : \bigcup_{a \in \alpha} B(a, \varepsilon) \supset \text{supp}(P) \right\},$$

$e_{n,\infty}(P)$ depends only on the topological support of $P$ and we also write $e_{n,\infty}(\text{supp}(P))$ instead of $e_{n,\infty}(P)$. Let

$$Q_\infty(d) = \inf_{n \geq 1} n^{1/d} e_{n,\infty}(U([0, 1]^d)). \quad (4.6)$$

Then the volume of the ball with radius $Q_\infty(d)$ coincides with the density of the thinnest covering of the whole space by translates of $B(0, 1)$ (cf. [13, p. 237, Definition 6] and [12, Remark 10.8]).

**Theorem 4.1.** Let $A \subset \mathbb{R}^d$ be a nonempty compact set. Then

$$\limsup_{n \to \infty} n^{1/d} e_{n,\infty}(A) \leq Q_\infty(d) \lambda^d(A)^{1/d}$$

and $Q_\infty(d) > 0$. If $\lambda^d(\partial A) = 0$, then

$$\lim_{n \to \infty} n^{1/d} e_{n,\infty}(A) = Q_\infty(d) \lambda^d(A)^{1/d}.$$

Proof. See [16, Theorem IX], [12, Theorem 10.7 and (10.8)].

The $L_1$-metric provides a lower bound. More generally, for $0 < r < \infty$, let

$$e_{n,r}(P) = \inf \left\{ \|d_\alpha\|_{L_r(P)} : \alpha \subset \mathbb{R}^d, 1 \leq |\alpha| \leq n \right\}. \quad (4.7)$$

and

$$Q_r(d) = \inf_{n \geq 1} n^{1/d} e_{n,r}(U([0, 1]^d)). \quad (4.8)$$

The main asymptotic result for $L_r$-quantization of nonsingular distributions is the following theorem, see [23], [24], and [3]. The first mathematically rigorous proof can be found in [12, Theorem 6.2]. Let $P_a$ denote the absolutely continuous part of $P$ (with respect to $\lambda^d$).
Theorem 4.2. Let \( r \in (0, \infty) \). We have
\[
\liminf_{n \to \infty} n^{1/d} \epsilon_{n,r}(P) \geq Q_r(d) \|h\|_{L_2(d + r)}^{1/r}
\]
and \( Q_r(d) > 0 \), where \( h \) is the \( \lambda^d \)-density of \( P_a \). If \( \int \|x\|^{r+\delta} dP(x) < \infty \) for some \( \delta > 0 \), then
\[
\lim_{n \to \infty} n^{1/d} \epsilon_{n,r}(P) = Q_r(d) \|h\|_{L_2(d + r)}^{1/r} < \infty.
\]

4.1. Nonsingular distributions. We will rely on the following lower bound.

Lemma 4.1. Let \( K \subset \mathbb{R}^d \) be a compact set with \( P(K) > 0 \). Then there exists \( k \in \mathbb{N} \) such that
\[
e_n(P) \geq \frac{P(K)}{2} e_{n+k,1}(P(\cdot | K)) \quad \text{for every} \quad n \in \mathbb{N}.
\]

Proof. Let \( \rho \) denote the Lipschitz metric,
\[
\rho(P_1, P_2) = \sup \left\{ \left| \int g dP_1 - \int g dP_2 \right| : g \in \text{Lip}^+(\mathbb{R}^d) \right\},
\]
where \( \text{Lip}^+(\mathbb{R}^d) \) stands for the class of function \( g: \mathbb{R}^d \to \mathbb{R}, 0 \leq g \leq 1 \), satisfying the Lipschitz condition \( |g(x) - g(y)| \leq \|x - y\| \), \( x, y \in \mathbb{R}^d \). Since \( \pi(P_1, P_2) \geq \rho(P_1, P_2)/2 \), we have
\[
e_n(P) \geq \frac{1}{2} \inf \{ \rho(P, Q) : Q \in \mathcal{P}_n \}.
\]
If \( Q \in \mathcal{P}_n \) and \( \alpha = \text{supp}(Q) \), then \( 1 \wedge d_\alpha \in \text{Lip}^+(\mathbb{R}^d) \) and hence
\[
\rho(P, Q) \geq \left| \int (1 \wedge d_\alpha) dP - \int (1 \wedge d_\alpha) dQ \right| = \int (1 \wedge d_\alpha) dP.
\]
This implies
\[
\inf \{ \rho(P, Q) : Q \in \mathcal{P}_n \} \geq \inf \left\{ \int (1 \wedge d_\alpha) dP : \alpha \subset \mathbb{R}^d, 1 \leq |\alpha| \leq n \right\}.
\]
Choose a finite set \( \beta \subset \mathbb{R}^d \) such that \( \max \{d_\beta(x) : x \in K\} \leq 1 \). Let \( k = |\beta| \). For \( \alpha \subset \mathbb{R}^d \) with \( 1 \leq |\alpha| \leq n \), one gets
\[
\int (1 \wedge d_\alpha) dP \geq \int (1 \wedge d_{\alpha \cup \beta}) dP \geq \int_K (1 \wedge d_{\alpha \cup \beta}) dP = \int_K d_{\alpha \cup \beta} dP
\]
and therefore
\[
\inf \left\{ \int (1 \wedge d_\alpha) dP : \alpha \subset \mathbb{R}^d, 1 \leq |\alpha| \leq n \right\} \geq P(K) e_{n+k,1}(P(\cdot | K)).
\]
The assertion follows from the above inequalities. Lemma 4.1 is proved.

Now we can employ the results known for \( e_{n,1} \) and \( e_{n,\infty} \).
Theorem 4.3. Assume that \( P_a \neq 0 \). Then
\[
\liminf_{n \to \infty} n^{1/d} e_n(P) \geq \frac{1}{2} Q_1(d) \|h\|_{L_1/(d+1)(\lambda^d)} > 0,
\]
where \( h \) is the \( \lambda^d \)-density of \( P_a \). If, additionally, \( \text{supp}(P) \) is compact, then
\[
e_n(P) \approx n^{-1/d}, \quad n \to \infty,
\]
and
\[
\frac{1}{2} Q_1(d) \|h\|_{L_1/(d+1)(\lambda^d)} \leq Q(P) < \overline{Q}(P) \leq Q_\infty(d) \lambda^d(\text{supp}(P))^{1/d}.
\]

Proof. Let \( K \subset \mathbb{R}^d \) be a compact set with \( P(K) > 0 \). It follows from Theorem 4.2 that
\[
\lim_{n \to \infty} n^{1/d} e_n,1(P(\cdot | K)) = P(K)^{-1} Q_1(d) \|h1_K\|_{L_1/(d+1)(\lambda^d)}.
\]
Using Lemma 4.1, we deduce
\[
\liminf_{n \to \infty} n^{1/d} e_n(P) \geq \frac{1}{2} Q_1(d) \|h\|_{L_1/(d+1)(\lambda^d)}.
\]
Since \( K \) is arbitrary, one obtains
\[
\liminf_{n \to \infty} n^{1/d} e_n(P) \geq \frac{1}{2} Q_1(d) \|h\|_{L_1/(d+1)(\lambda^d)}.
\]
Now suppose that \( \text{supp}(P) \) is compact. Then by Theorem 4.1
\[
\limsup_{n \to \infty} n^{1/d} e_n,\infty(P) \leq Q_\infty(d) \lambda^d(\text{supp}(P))^{1/d} < \infty.
\]
In view of (4.5) the assertions follows. Theorem 4.3 is proved.

The above upper bound for \( \overline{Q}(P) \) seems to be unbeatable. Except in dimension 1 or 2, the true values of the constants \( Q_1(d) \) and \( Q_\infty(d) \) are unknown. (Actually, \( Q_1(1) = \frac{1}{4} \), \( Q_\infty(1) = \frac{1}{2} \) and for the \( l_2 \)-norm on \( \mathbb{R}^2 \), \( Q_1(2) = (2 + 3 \ln \sqrt{3})/(3^{7/4} \sqrt{2}) = 0.3777 \ldots \) and \( Q_\infty(2) = (2(3 \sqrt{3})^{1/2} = 0.6204 \ldots \).)

However, some lower bounds on \( Q_1(d) \) and upper bounds on \( Q_\infty(d) \) can be obtained, using lattice coverings (see, e.g., [12]).

In the unimodal case the quantization coefficient exists.

Proposition 4.1. Let \( d = 1 \), and let \( P \) be continuous and unimodal. Then
\[
\lim_{n \to \infty} ne_n(P) = \sup_{n \geq 1} ne_n(P) = \frac{1}{2} \lambda(\text{supp}(P)).
\]
Thus, if \( \text{supp}(P) \) is compact, the quantization coefficient \( Q(P) \) exists (and equals the covering coefficient of \( \text{supp}(P) \)).

Proof. Easy consequence of Proposition 2.1 (and Theorem 4.1).
4.2. Regular distributions of dimension \( D \). An interesting class of probability measures is provided by Ahlfors–David-type regular measures. For \( D \in (0, \infty) \), \( P \) is called regular of dimension \( D \) if \( P \) has compact support and there exist \( c_1, c_2, r_0 \in (0, \infty) \) such that

\[
c_1 r^D \leq P(B(a, r)) \leq c_2 r^D \quad \text{for all} \quad a \in \text{supp}(P), \quad r \in (0, r_0). \tag{4.9}
\]

The class of regular distributions (with \( D < d \)) contains, for instance, the normalized surface measures on convex compact sets or on smooth compact manifolds and the normalized Hausdorff measure on self-similar sets (cf. [6], and [12, Section 12]).

Proposition 4.2. Assume that there is \( c \in (0, \infty) \) such that

\[
P(B(a, r)) \leq cr^D \quad \text{for all} \quad a \in \mathbb{R}^d, \quad r \in (0, 1). \tag{4.10}
\]

Then

\[
\liminf_{n \to \infty} n^{1/D} e_n(P) \geq c^{-1/D}
\]

(with the same constant \( c \)).

Proof. For \( \alpha \in \mathcal{C}_n(P) \), one obtains by (4.10)

\[
\|d_\alpha\|_\alpha \geq \inf \left\{ \varepsilon \geq 0 : \sum_{a \in \alpha} P(B(a, \varepsilon)) \geq 1 - \varepsilon \right\} \geq \inf \{ \varepsilon \geq 0 : n\varepsilon^D \geq 1 - \varepsilon \} =: s_n
\]

and thus \( e_n(P) \geq s_n \). The lower bound \( s_n \) is a unique solution in the interval \((0, 1)\) of the equation

\[
n\varepsilon^D = 1 - \varepsilon
\]

and therefore, \( n^{1/D} s_n \) is a unique solution in \((0, n^{1/D})\) of the equation \( c\varepsilon^D = 1 - n^{-1/D} \varepsilon \). We deduce that \( (n^{1/D} s_n)_{n \geq 1} \) is a strictly increasing sequence and

\[
\lim_{n \to \infty} n^{1/D} s_n = \sup_{n \geq 1} n^{1/D} s_n = c^{-1/D}.
\]

This yields \( \liminf_{n \to \infty} n^{1/D} e_n(P) \geq c^{-1/D} \). Proposition 4.2 is proved.

Notice that a Borel probability measure \( P \) on \( \mathbb{R}^d \) satisfies (4.10) with \( D = d \) if and only if \( P \) is absolutely continuous with a.s. bounded \( \lambda^d \)-density \( h \); the best constant \( c \) in (4.10) is then given by \( \lambda^d \)-ess \( \sup h \lambda^d(B(0,1)) \) (cf. [17, Theorem 2.12]).

One may reverse the inequality signs in Proposition 4.2 up to a modification for the constant.

Proposition 4.3. Assume that \( \text{supp}(P) \) is compact and there are constants \( c, r_0 \in (0, \infty) \) such that

\[
P(B(a, r)) \geq cr^D \quad \text{for all} \quad a \in \text{supp}(P), \quad r \in (0, r_0).
\]

Then

\[
\limsup_{n \to \infty} n^{1/D} e_n(P) \leq 2c^{-1/D}.
\]
P r o o f. The assertion follows from \( \limsup_{n \to \infty} n^{1/D} e_{n,\infty}(P) \leq 2c^{-1/D} \) (cf. [11, Proposition 12.17]) and (4.5).

We can deduce the exact rate of convergence of \( e_n(P) \) to zero for regular distributions.

**Theorem 4.4.** Let \( P \) be a regular probability measure of dimension \( D \). Then

\[
 e_n(P) \approx n^{-1/D} \quad \text{as} \quad n \to \infty.
\]

P r o o f. Regularity of dimension \( D \) implies condition (4.10) (cf. [11, Lemma 12.3]). Consequently, the assertion follows from Propositions 4.2 and 4.3.

As an example, for the classical Cantor distribution \( P \) one obtains

\[
e_n(P) \approx n^{-\ln 3/\ln 2}.
\]

4.3. Uniform distributions. For uniform distributions we obtain the following result.

**Corollary 4.1.** If \( A \subset \mathbb{R}^d \) is a compact set with \( \lambda^d(A) > 0 \), then \( e_n(U(A)) \approx n^{-1/d} \) and

\[
\left( \frac{\lambda^d(A)}{\lambda^d(B(0,1))} \right)^{1/d} \leq Q(U(A)) \leq \bar{Q}(U(A)) \leq Q_\infty(d) \lambda^d(A)^{1/d}.
\]

P r o o f. \( U(A) \) satisfies (4.10) with \( c = \lambda^d(B(0,1))/\lambda^d(A) \) and \( D = d \) which proves the assertion.

If the underlying norm on \( \mathbb{R}^d \) is the \( l_p \)-norm, \( 1 < p < \infty \), then the bounds in the preceding corollary for \( A = [0,1]^d \) are asymptotic to the same value as \( d \to \infty \). In fact, in view of the classical upper bound for \( Q_\infty(d) \),

\[
Q_\infty(d) \leq \frac{(d \ln d + d \ln \ln d + 5d)^{1/d}}{\lambda^d(B(0,1))^{1/d}}, \quad d \geq 3,
\]

which is due to [19] it follows from the formula

\[
\lambda^d(B(0,1)) = \frac{(2\Gamma(1+1/p))^d}{\Gamma(1+d/p)}
\]

and Stirling's formula for the \( \Gamma \)-function that

\[
\lim_{d \to \infty} d^{-1/p} Q(U([0,1]^d)) = \lim_{d \to \infty} d^{-1/p} \overline{Q}(U([0,1]^d)) = \lim_{d \to \infty} d^{-1/p} Q_\infty(d) = \frac{p}{2(ep)^{1/p} \Gamma(1/p)}. \quad (4.11)
\]

In a special case the bounds in Corollary 4.1 exactly agree, namely if the ball \( B(0,1) \) is space-filling (by translation). Then the quantization coefficient exists and is given by

\[
Q(U(A)) = Q_\infty(d) \lambda^d(A)^{1/d} = \left( \frac{\lambda^d(A)}{\lambda^d(B(0,1))} \right)^{1/d}.
\]
One obtains, for instance,

\[ Q(U(A)) = \frac{\lambda^d(A)^{1/d}}{2} \quad (l_\infty\text{-norm}), \]

\[ Q(U(A)) = \left( \frac{\lambda^d(A)}{2} \right)^{1/2}, \quad d = 2 \quad (l_1\text{-norm}). \]  

(4.12)

The following theorem concerns a nontrivial case, where the quantization coefficient exists.

**Theorem 4.5.** Let \( d = 2 \), and let the underlying norm be the \( l_2 \)-norm. Then

\[ \lim_{n \to \infty} n^{1/2}e_n(U([0,1]^2)) = Q_\infty(2) = \left( \frac{2}{3\sqrt{3}} \right)^{1/2} = 0.6204 \ldots \]

(and thus the quantization coefficient exists for \( U([0,1]^2) \) and coincides with the covering coefficient \( Q_\infty(2) \)).

For the proof we need the following estimate.

**Lemma 4.2.** If \( A, B \subset \mathbb{R}^d \) are compact with \( A \subset B \) and \( \lambda^d(A) > 0 \), then

\[ e_n(U(A)) \leq e_n(U(B)) \frac{\lambda^d(B)}{\lambda^d(A)}. \]

**Proof.** Let \( s = e_n(U(B)) \), \( \alpha \in \mathcal{C}_n(U(B)) \), and \( c = \lambda^d(B)/\lambda^d(A) \). Then

\[ s = U(B)(d_\alpha > s) \geq U(B)(d_\alpha > cs) \geq U(A)(d_\alpha > cs)c^{-1} \]

and hence \( cs \geq U(A)(d_\alpha > cs) \). This implies \( e_n(U(A)) \leq \|d_\alpha\|_{L_\infty(U(A))} \leq cs \).

Lemma 4.2 is proved.

**Proof of Theorem 4.5.** Let \( \Lambda = \mathbb{Z}(1,0) + \mathbb{Z}(\frac{1}{2}, \sqrt{3}/2) \) denote the hexagonal lattice. We approximate \([0,1]^2\) by packings in \([0,1]^2\) consisting of suitably scaled regular hexagons form the Voronoi diagram of \( \Lambda \). For \( n \in \mathbb{N} \), let

\[ r_n = \sup_{\alpha \in \Lambda, ||\alpha|| = n} \max \left\{ s \geq 0 : \bigcup_{a \in \alpha} sW(a | \Lambda) \subset [0,1]^2 \right\}. \]

Then the supremum is achieved by some \( \alpha_n \subset \Lambda \). Set

\[ B_n = \bigcup_{a \in \alpha_n} r_nW(a | \Lambda). \]

Since the regular hexagon \( A = W(0|\Lambda) \) is space-filling, one gets

\[ n^{1/2}r_n \to \lambda^2(A)^{-1/2}, \quad n \to \infty. \]  

(4.13)

According to an inequality obtained in [14] (see also [2]), we have for every \( \varepsilon > 0 \) and every \( \beta \subset \mathbb{R}^2 \) with \( ||\beta|| \leq n \)

\[ \int_{B_n} 1_{[0,\varepsilon]}(d_\beta(x)) \, dx \leq n \int_{r_nA} 1_{[0,\varepsilon]}(||x||) \, dx, \]
that is,
\[ U(B_n)(\beta \leq \varepsilon) \leq U(r_n A)(d_{[0]} \leq \varepsilon). \]
If \( s = e_n(U(B_n)) \) and \( \beta \in \mathcal{C}_n(U(B_n)) \), then
\[ s = U(B_n)(d_{\beta} > s) \geq U(r_n A)(d_{[0]} > s) \]
and hence \( \|d_{[0]}\|_{L_0(U(r_n A))} \leq s = e_n(U(B_n)) \). The set of midpoints \( r_n \alpha_n \)
satisfies \( \|d_{r_n \alpha_n}\|_{L_0(U(B_n))} = \|d_{[0]}\|_{L_0(U(r_n A))} \). Therefore,
\[ e_n(U(B_n)) = \|d_{[0]}\|_{L_0(U(r_n A))} \tag{4.14} \]
and \( r_n \alpha_n \in \mathcal{C}_n(U(B_n)) \). It follows from Lemma 2.4 and (4.13) that
\[ n^{1/2} \lambda^2(A)^{1/2} \|d_{[0]}\|_{L_0(U(r_n A))} = n^{1/2} \lambda^2(A)^{1/2} \|r_n d_{[0]}\|_{L_0(U(A))} \rightarrow \|d_{[0]}\|_{L_\infty(U(A))} = \max_{x \in A} \|x\|, \quad n \to \infty. \]
This implies
\[ \lim_{n \to \infty} n^{1/2} e_n(U(B_n)) = \max_{x \in A} \|x\| \lambda^2(A)^{-1/2} = Q_\infty(2), \]
where the last equality is well known. Since \( \lambda^2(B_n) \to 1 \), we deduce in view of Lemma 4.2 \( \lim_{n \to \infty} n^{1/2} e_n(U([0, 1]^2)) = Q_\infty(2) \). Theorem 4.5 is proved.

5. Asymptotic quantization error: distributions with unbounded support

5.1. The quantization dimension. Now we replace the compactness assumption for the support of \( P \) by a moment condition. However, we do not achieve an exact power law for probability measures \( P \) with \( P_a \neq 0 \) as in Theorem 4.2 but weaker dimension results. They are important, for instance, for our investigation of one-dimensional (unimodal) distributions in Subsection 5.2.

The lower and upper quantization dimensions of \( P \) (relative to the Prokhorov metric) are defined by
\[ D(P) = \liminf_{n \to \infty} \frac{\ln n}{-\ln e_n(P)}, \quad \overline{D}(P) = \limsup_{n \to \infty} \frac{\ln n}{-\ln e_n(P)}, \tag{5.1} \]
respectively. If these are equal, we refer to the common value as the quantization dimension of \( P \)
\[ D(P) = \lim_{n \to \infty} \frac{\ln n}{-\ln e_n(P)}. \tag{5.2} \]
Notice that the (lower, upper) quantization dimension has the same value for all norms on \( \mathbb{R}^d \).
Lemma 5.1. Let \( e_n = e_n(P) \).

(a) If \( 0 < t < \overline{D}(P) < s \), then

\[
\lim_{n \to \infty} n^{1/t} e_n = \infty \quad \text{and} \quad \liminf_{n \to \infty} n^{1/s} e_n = 0.
\]

(b) If \( 0 < t < \underline{D}(P) < s \), then

\[
\limsup_{n \to \infty} n^{1/t} e_n = \infty \quad \text{and} \quad \liminf_{n \to \infty} n^{1/s} e_n = 0.
\]

(c) If \( D \in (0, \infty) \) and \( \liminf_{n \to \infty} n^{1/D} e_n > 0 \), then \( D \leq \overline{D}(P) \).

(d) If \( D \in (0, \infty) \) and \( \limsup_{n \to \infty} n^{1/D} e_n < \infty \), then \( D \geq \underline{D}(P) \).

Proof. The proofs of (a) and (b) are omitted. Assertions (c) and (d) follow immediately from (a) and (b).

The quantization dimension coincides with the epsilon entropy dimension of \( P \). For \( \epsilon > 0 \), let

\[
N(\epsilon, P) = \min \left\{ n \geq 1 : \exists \alpha \subset \mathbb{R}^d, |\alpha| = n, P \left( \bigcup_{a \in \alpha} B(a, \epsilon) \right) \geq 1 - \epsilon \right\}.
\]

Proposition 5.1. We have

\[
\underline{D}(P) = \liminf_{\epsilon \to 0} \frac{\ln N(\epsilon, P)}{-\ln \epsilon}, \quad \overline{D}(P) = \limsup_{\epsilon \to 0} \frac{\ln N(\epsilon, P)}{-\ln \epsilon}.
\]

Proof. Set \( N(\epsilon) = N(\epsilon, P) \) and \( e_n = e_n(P) \) and notice that

\[
N(\epsilon) = \min \{ n \geq 1 : e_n \leq \epsilon \}.
\]

If \( e_n = 0 \) for some \( n \geq 1 \), then \( \underline{D}(P) = \overline{D}(P) = 0 \) and the assertions are obviously true. So assume \( e_n > 0 \) for every \( n \geq 1 \). Since \( N(e_n) \leq n \) for all \( n \geq 1 \), we deduce

\[
\underline{D}(P) = \liminf_{n \to \infty} \frac{\ln N(e_n)}{-\ln e_n} \geq \liminf_{\epsilon \to 0} \frac{\ln N(\epsilon)}{-\ln \epsilon}.
\]

The definition implies \( e_{N(\epsilon)} \leq \epsilon \) for all \( \epsilon > 0 \), hence

\[
\underline{D}(P) \leq \liminf_{\epsilon \to 0} \frac{\ln N(\epsilon)}{-\ln e_{N(\epsilon)}} \leq \liminf_{\epsilon \to 0} \frac{\ln N(\epsilon)}{-\ln \epsilon}.
\]

As for \( \overline{D}(P) \), we will prove first that there is a \( k \in \mathbb{N} \) such that

\[
e_{k(n+1)} < e_n \quad \text{for all} \quad n \geq 1.
\]

Let \( \beta \subset \mathbb{R}^d \) be of minimum cardinality with \( B(0, 1) \subset \bigcup_{b \in \beta} B(b, \frac{1}{2}) \). Let \( k = |\beta| \). (It is well known that \( k \leq 5^d \).) For \( x \in \mathbb{R}^d, \epsilon > 0 \), and \( \beta(x, \epsilon) = \epsilon \beta + x \), we have

\[
B(x, \epsilon) \subset \bigcup_{b \in \beta(x, \epsilon)} B(b, \frac{\epsilon}{2}).
\]
For $n \geq 1$, choose $\alpha \in \mathcal{C}_n(P)$. If $P\{d_\alpha \leq e_n\} = 1$, let $\gamma = \bigcup_{a \in \alpha} \beta(a,e_n)$. Then $|\gamma| \leq kn$ and $\{d_\alpha \leq e_n\} \subset \{d_\gamma \leq e_n/2\}$, hence

$$P\left\{d_\gamma \leq \frac{e_n}{2}\right\} \geq P\{d_\gamma \leq e_n\} = 1 \geq 1 - \frac{e_n}{2}.$$  

This implies $\|d_\gamma\|_0 \leq e_n/2$ and therefore

$$e_k(n+1) \leq e_{kn} \leq \|d_\gamma\|_0 \leq e_n/2 < e_n.$$

If $P\{d_\alpha \leq e_n\} < 1$, one can find $b \in \{d_\alpha > e_n\}$ such that $\alpha' = \alpha \cup \{b\}$ satisfies

$$P\{d_\alpha' \leq e_n\} > P\{d_\alpha \leq e_n\} \geq 1 - e_n$$

(see the proof of Theorem 2.1). Let $\gamma = \bigcup_{a \in \alpha'} \beta(a,e_n)$. Then $|\gamma| \leq k(n+1)$ and

$$P\left\{d_\gamma \leq \frac{e_n}{2}\right\} \geq P\{d_\alpha' \leq e_n\} > 1 - e_n,$$

hence $\|d_\gamma\|_0 < e_n$. This implies

$$e_k(n+1) \leq \|d_\gamma\|_0 < e_n.$$

It follows from (5.4) that $e_k(N(e_n)+1) < e_{N(e_n)} = e_n$ for all $n \geq 1$. This gives $k(N(e_n)+1) > n$ and therefore $N(e_n) \geq n/2k$ for all $n \geq 2k$. We thus obtain

$$\limsup_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{-\ln \varepsilon} \geq \limsup_{n \to \infty} \frac{\ln N(e_n)}{-\ln e_n} \geq \limsup_{n \to \infty} \left[ \frac{\ln n}{-\ln e_n} - \frac{\ln(2k)}{-\ln e_n} \right] = D(P).$$

To prove the converse inequality, observe that $e_{N(\varepsilon)-1} > \varepsilon$ for all $\varepsilon < \varepsilon_1$. This leads to

$$\limsup_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{-\ln \varepsilon} \leq \limsup_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{-\ln e_{N(\varepsilon)-1}} = \limsup_{\varepsilon \to 0} \frac{\ln (N(\varepsilon) - 1)}{-\ln e_{N(\varepsilon)-1}} \leq D(P).$$

Proposition 5.1 is proved.

The Hausdorff dimension of $P$ given by

$$\dim_H(P) = \inf \{\dim_H(A) : A \in \mathcal{B}(\mathbb{R}^d), P(A) = 1\} \quad (5.5)$$

provides a lower bound for the (lower) quantization dimension of $P$. Here $\dim_H(A)$ denotes the Hausdorff dimension of the set $A$.

**Proposition 5.2.** We have

$$\dim_H(P) \leq D(P).$$

**Proof.** If follows from Lemma 4.3 and [12, Theorem 11.6] that

$$\dim_H(P(\cdot | K)) \leq D(P)$$
for every compact set $K \subset \mathbb{R}^d$ with $P(K) > 0$. Choose a sequence $(K_n)_{n \geq 1}$ of compact sets $K_n \subset \mathbb{R}^d$ satisfying $P(K_n) > 0$ and $\lim P(K_n) = 1$. For every $n \in \mathbb{N}$, there exists a Borel set $A_n \subset \mathbb{R}^d$ such that $P(A_n \cap K_n)/P(K_n) = 1$ and

$$\dim_H(P(\cdot | K_n)) = \dim_H(A_n).$$

Set $A = \bigcup_{n=1}^{\infty} A_n$. One obtains $P(A) = 1$ and therefore

$$\dim_H(P) \leq \dim_H(A) = \sup_{n \geq 1} \dim_H(A_n) \leq D(P).$$

Proposition 5.2 is proved.

Since $\dim_H(P) \leq d$, the above inequality can be strict (see Example 5.1(b)).

An upper bound for the (upper) quantization dimension can be deduced under a moment condition.

**Proposition 5.3.** Let $r_0 = \sup \{r > 0 : E\|X\|^r < \infty \}$. Then

$$\overline{D}(P) \leq d + \frac{d}{r_0}.$$ 

**Proof.** We may assume $r_0 > 0$. Using the Markov inequality, one gets

$$e_n(P) \leq e_{n,r}(P)^{r/(r+1)}$$

for every $r \in (0, \infty)$. Let $r \in (0, r_0)$ and set $t = d + d/r$. Then there exists $\delta > 0$ such that $E\|X\|^{r+\delta} < \infty$. Consequently, by Theorem 4.2,

$$\limsup_{n \to \infty} n^{1/t} e_n(P) \leq \lim_{n \to \infty} n^{1/t} e_{n,r}(P)^{r/(r+1)} = \lim_{n \to \infty} \left( n^{1/d} e_{n,r}(P) \right)^{r/(r+1)} < \infty.$$ 

Using Lemma 5.1, this yields $\overline{D}(P) \leq t$. The assertion follows.

**Remarks.** (a) The above upper bound is sharp (at least for $d = 1$). See Theorem 5.2.

(b) The above bound for $\overline{D}(P)$ can also be deduced from a result of [8] and Proposition 5.1. In fact, (a slight modification of) Proposition 3.4 of [8] says that for $r \in (0, r_0)$ there is a constant $c \in (0, \infty)$ such that

$$N(\varepsilon, P) \leq ce^{-d(r+1)/r}$$

for all sufficiently small $\varepsilon$.

The main dimension result for probability distributions with all their moments finite is an immediate consequence.

**Theorem 5.1.** Assume that $P_a \neq 0$ and $E\|X\|^r < \infty$ for every $r > 0$. Then

$$D(P) = d.$$ 

**Proof.** By Theorem 4.4 and Lemma 5.1 (c), we have $d \leq D(P)$. Thus the assertion follows from Proposition 5.3.

Theorem 5.1 is not a final result concerning the rate of convergence of $e_n(P)$ to zero. It remains to be made more precise, perhaps with slowly varying (distribution dependent) factors.
5.2. One-dimensional unimodal distributions. Let \( d = 1 \), and let \( P \) be continuous and unimodal. Then, if \( \text{supp}(P) \) is unbounded, one obtains from Proposition 4.1 that

\[
\lim_{n \to \infty} n e_n(P) = \infty. \tag{5.6}
\]

If all moments are finite, we have \( D(P) = 1 \). (The moment condition holds if \( P \) is strongly unimodal.)

We determine the exact rates of \( e_n \) for two important classes of unimodal distributions with unbounded support, namely distributions having regularly varying tails or exponentially small tails.

**Theorem 5.2.** Assume that \( P \) is symmetric, continuous, unimodal with \( \text{supp}(P) = \mathbb{R} \). Let \( F \) denote the distribution function of \( P \).

(a) If \( 1 - F \) is regularly varying at infinity of index \(-\rho < 0\), that is, \( 1 - F(x) = x^{-\rho}g(x) \), where \( g \) is slowly varying at infinity, then \( D(P) = (1 + \rho)/\rho \) and

\[
\lim_{n \to \infty} n^{\rho/(1+\rho)} h(n)^{-1} e_n(P) = 2^{1/(1+\rho)}, \tag{5.7}
\]

where \( h \) is a slowly varying function at infinity such that

\[
h\left(\frac{x^{1+\rho}}{g(x)}\right) \sim g(x)^{1/(1+\rho)} \quad \text{as} \quad x \to \infty.
\]

(b) If \( -\ln(1 - F) \) is regularly varying at infinity of index \( \delta > 0 \), then \( D(P) = 1 \) and

\[
\lim_{n \to \infty} n c_n^{-1} e_n(P) = 1, \tag{5.8}
\]

where \( c_n = F^{-1}(1 - 1/n) \) and \((c_n)_{n \geq 1}\) is slowly varying.

Note that the restriction in part (a) on the index of \( 1 - F \) is natural, since otherwise \( 1 - F(x) \to \infty \) as \( x \to \infty \). If \( 1 - F(x) \sim cx^{-\rho} \) as \( x \to \infty \) with \( \rho > 0 \), then (5.7) shows that the quantization coefficient of \( P \) exists. In part (b), the restriction \( \delta > 0 \) on the index of \(-\ln(1 - F)\) is natural, since \( F(x) \to 0 \) as \( x \to \infty \) in case \( \delta < 0 \) and statement (5.8) is clearly not true in case \( \delta = 0 \).

**Proof.** Set \( e_n = e_n(P) \). The concentration function takes the form \( C_P(x) = 2F(x/2) - 1, x \geq 0 \). Therefore, it follows from Proposition 2.1 that \( e_n = 2(1 - F(n e_n))), n \in \mathbb{N} \). Furthermore, by (5.6), \( n e_n \to \infty \).

(a) We have \( e_n = 2(n e_n)^{-\rho} g(n e_n) \) and hence

\[
(n e_n)^{1+\rho} g(n e_n)^{-1} = 2n.
\]

Setting \( k(x) = x^{1+\rho}g(x)^{-1} \) and \( \tilde{k}(x) = x^{1/(1+\rho)}h(x) \) yields \( \tilde{k} \circ k(x) \sim x \) as \( x \to \infty \) which in turn implies

\[
e_n \sim \tilde{k}(2n) \quad \text{as} \quad n \to \infty.
\]

This gives (5.7). Since \( \ln h(n)/\ln n \to 0 \) (cf. [1, Proposition 1.3.6]), the dimension result follows from (5.7).
(b) Let \(-\ln(1 - F(x)) = x^\delta g(x)\), where \(g\) is slowly varying at infinity. Then we have

\[
- \ln e_n = - \ln 2 - \ln(1 - F(ne_n)) = - \ln 2 + (ne_n)^\delta g(ne_n).
\]

An application of Theorem 5.1 yields \(D(P) = 1\) and thus

\[
(ne_n)^\delta g(ne_n) \sim \ln n \quad \text{as} \quad n \to \infty.
\]

There exists a regularly varying function \(k\) of index \(1/\delta\) such that \(k(x^\delta g(x)) \sim x\) as \(x \to \infty\) (cf. [1, Theorem 1.5.12]). Consequently,

\[
ne_n \sim k(\ln n) \quad \text{and} \quad cn \sim k(\ln n) \quad \text{as} \quad n \to \infty.
\]

This implies (5.8). Since \(k \circ \ln\) is a slowly varying function at infinity, the sequence \((c_n)_{n \geq 1}\) is slowly varying.

Remark 5.1. A one-sided version of the preceding theorem reads as follows. Assume that \(P\) is continuous, unimodal with mode 0 and \(\text{supp}(P) = \mathbb{R}_+\). Then the concentration function takes the form \(C_P(x) = F(x), x > 0\). By Proposition 2.1,

\[
e_n = 1 - F(2ne_n)
\]

and (5.7) and (5.8) hold with the right-hand side replaced by \(2^{1/(1+\rho)}/2\) and \(\frac{1}{2}\), respectively.

Example 5.1. (a) (Normal distribution.) Let \(P = \mathcal{N}(0, \sigma^2)\).

Using which implies \(-\ln(1 - F(x)) \sim x^2/(2\sigma^2)\) as \(x \to \infty\) one gets

\[
F^{-1}\left(1 - \frac{1}{n}\right) \sim (2\sigma^2 \ln n)^{1/2}.
\]

Therefore, by Theorem 5.2(b),

\[
\lim_{n \to \infty} n^{(\alpha+1)/\alpha} e_n(P) = \sqrt{2}\sigma.
\]

(b) (Symmetric stable distributions.) The asymptotic behavior of the \(n\)th quantization error for the normal distribution is quite different from that of the other symmetric stable distributions. Let \(P\) be nondegenerate symmetric \(\alpha\)-stable with \(0 < \alpha < 2\). It is well known that \(P\) is continuous unimodal and \(1 - F(x) \sim cx^{-\alpha}\) as \(x \to \infty\) for some constant \(c > 0\). Therefore, by Theorem 5.2(a),

\[
\lim_{n \to \infty} n^{\alpha/(\alpha+1)} e_n(P) = (2c)^{1/(\alpha+1)}.
\]

In particular, \(D(P) = (\alpha+1)/\alpha\) and the quantization coefficient of \(P\) exists.
If, for instance, $P$ is a Cauchy distribution with scale parameter $a > 0$, then $D(P) = 2$,

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} \left( \frac{x}{a} \right), \quad 1 - F(x) \sim \frac{a}{\pi x} \quad \text{as} \quad x \to \infty,$$

hence

$$\lim_{n \to \infty} n^{1/2} e_n(P) = \left( \frac{2a}{\pi} \right)^{1/2}.$$

(c) (Logistic distribution.) Let $P = h\lambda$ with density

$$h(x) = \frac{\exp(x/a)}{a(1 + \exp(x/a))^2}, \quad x \in \mathbb{R}, \quad a > 0.$$ The distribution function is given by $F(x) = 1/(1 + \exp(-x/a))$ and hence $-\ln(1 - F(x)) \sim x/a$ as $x \to \infty$. One obtains $F^{-1}(1 - n^{-1}) \sim \ln n$ and therefore, by Theorem 5.2(b),

$$\lim_{n \to \infty} n(\ln n)^{-1} e_n(P) = a.$$

(d) (Weibull distribution.) Let $P = h\lambda$ with density

$$h(x) = \frac{b}{a^b} x^{b-1} \exp \left( - \left( \frac{x}{a} \right)^b \right) 1_{(0, \infty)}(x), \quad a > 0, \quad 0 < b \leq 1.$$ Then $F(x) = 1 - \exp(-(x/a)^b)$ for $x > 0$, $-\ln(1 - F(x)) = (x/a)^b$ for $x > 0$, and $F^{-1}(1 - n^{-1}) = a(\ln n)^{1/b}$. Therefore, by Remark 5.1,

$$\lim_{n \to \infty} n(\ln n)^{-1/b} e_n(P) = a/2.$$

5.3. Distributions with unimodal marginals. We rely on the following product quantizer upper bound for $e_n(P)$. Let $P_i$ denote the $i$th one-dimensional marginal distribution of $P$.

**Lemma 5.2.** Let $\| \cdot \| \leq c \| \cdot \|_{l_1}$, where $\| \cdot \|_{l_1}$ denotes the $l_1$-norm on $\mathbb{R}^d$. If $n_i \in \mathbb{N}$, $\prod_{i=1}^d n_i \leq n$, then

$$e_m(P) \leq (c \vee 1) \sum_{i=1}^d e_{n_i}(P_i).$$

**Proof.** First assume that $\| \cdot \| = \| \cdot \|_{l_1}$. For $i \in \{1, \ldots, d\}$, let $\alpha_i \in \mathcal{C}_{n_i}(P_i)$ and $s_i = e_{n_i}(P_i)$. Let $\alpha \subset \mathbb{R}^d$ denote the product quantizer $\bigotimes_{i=1}^d \alpha_i$ and $s = \sum_{i=1}^d s_i$. Then $|\alpha| \leq n$ and

$$P\{d_\alpha > s\} = P\left\{ x \in \mathbb{R}^d : \min_{a \in \alpha} \sum_{i=1}^d |x_i - a_i| > s \right\}$$

$$= P\left\{ x \in \mathbb{R}^d : \sum_{i=1}^d \min_{b \in \alpha_i} |x_i - b| > s \right\}$$

$$\leq \sum_{i=1}^d P_i\{d_{\alpha_i} > s_i\} \leq \sum_{i=1}^d s_i = s.$$
Consequently, $e_n(P) \leq \|d_a\|_{L_0(P)} \leq s$. The general case follows from $e_n(P) \leq (c \lor 1) e_n(P, \|\cdot\|_1)$. Lemma 5.2 is proved.

Combining the preceding lemma, Theorem 5.2, and Remark 5.1 provides rates for $e_n(P)$. In particular, using the uniform bit allocation $n_i = \cdots = n_d = [n^{1/d}]$, where $[x]$ denotes the integer part of $x \in \mathbb{R}$, we obtain the following result.

**Corollary 5.1.** Assume that $P_i$ is symmetric, continuous, unimodal with $\text{supp}(P_i) = \mathbb{R}$ for every $i \in \{1, \ldots, d\}$. Let $F_i$ denote the distribution function of $P_i$.

(a) If $1 - F_i(x) \sim c_i x^{-\rho_i}$ as $x \to \infty$ with $c_i, \rho_i > 0$, then
$$e_n(P) = O(n^{-\rho/d(1+\rho)}) \quad \text{as} \quad n \to \infty,$$
where $\rho = \min_{1 \leq i \leq d} \rho_i$.

(b) If $-\ln(1 - F_i) \sim c_i x^{\delta_i}$ as $x \to \infty$ with $c_i, \delta_i > 0$, then
$$e_n(P) = O(n^{-1/d}(\ln n)^{1/\delta}) \quad \text{as} \quad n \to \infty,$$
where $\delta = \min_{1 \leq i \leq d} \delta_i$.

For the $d$-dimensional normal distribution $P = N_d(0, \Sigma)$ with regular covariance matrix $\Sigma$ one obtains
$$e_n(P) = O(n^{-1/d}(\ln n)^{1/2}), \quad n \to \infty.$$

One may reasonably conjecture that this rate is the true one.

**СПИСОК ЛИТЕРАТУРЫ**

Quantization for probability measures


Поступила в редакцию
25.XI.2003

Исправленный вариант
22.V.2007