OPTIMAL FUNCTIONAL PRODUCT QUANTIZATION

TAO GUO, NIKITA KARAGODIN, AND EUGENE STEPANOV

Abstract. We introduce a mathematically new approach for quantization of vectorial signals which is already widely used by engineers, and discuss its basic properties. Consider a vectorial signal on the input to a computational device calculating a given function of the input. The main objective is to quantize the signal components separately in a way that optimizes the output quality. We study existence of optimal quantizers and estimate the optimal cost for several classes of functions.

Contents

1. Introduction 1
1.1. Functional product quantization problem 2
1.2. Comparison with classical quantization 3
2. Notation and preliminaries 3
3. A bridge between classical and functional product quantization 4
4. Random quantization and existence of optimal quantizers 4
4.1. Random quantization 5
4.2. Existence of nonrandom optimal quantizers 7
4.3. Properties of quantizing sets 8
5. Optimal quantizers for particular classes of functions 9
5.1. Characteristic functions of measurable rectangles and their finite sums 9
5.2. Characteristic functions of "nice" planar sets 12
5.3. Linear functions 16
5.4. Lower bounds for monotone functions 19
5.5. Quadratic cost 21
5.6. Further examples of functions 23
6. General upper estimate for Sobolev functions 25
Appendix A. Auxiliary statements 28
References 30

1. Introduction

Suppose that a d-dimensional vectorial signal $Z = (X_1, \ldots, X_d)$ with scalar components X_i is input to a computational device that produces the value $f(Z)$ of the given function f on the output. We want to quantize (substitute with a signal which might have only a discrete set of values) separately and independently the components of input, i.e. the scalar signals X_i, so as to maximize the quality of the output. That is, our goal is to minimize the expectation of the error between

1991 *Mathematics Subject Classification.* 49Q20, 49Q10; 49J55, 49Q22.

output on Z and output on its quantized version, once Z is a random vector with a given distribution law (common distribution law of $(X_1, \ldots, X_d)$). We will call this a functional product quantization problem. The general problem statement is described formally in the next section.

The term quantization is known as a process of mapping a large (probably continuous) set to a small (often finite) set and it has a long history. The idea is so natural, that it dates back hundreds of years, as it was used to find approximate values of integrals by discretization of input, which is, by modern standards, part of Numerical analysis. Later on, the process itself was defined formally as finding the best discrete measure with a given number of support points that approximates a given measure. This topic occurres in various fields, such as Information theory (compression), Stochastic processes (sampling), Machine learning (clustering), Numerical analysis. Naturally, there are many works dedicated to this subject, see [5] for an introduction to the field as well as [7] for a survey on classical results. Note that in the classical setting the space is quantized as a whole, which might lead to the "Curse of dimensionality" (various troubling effects that arise when the dimension of a quantized space increases). In addition, the quality of quantization is usually measured as the expectation of some power of the distance between signal and its quantized version. Product quantization has been introduced in Machine learning community by Jégou, Douze and Schmid [1] as a technique that allows to significantly reduce the dimension of the quantized space. It is most famous as it improves the nearest neighbor search algorithm. When considering quantization in Information theory, it appears that the idea of independent coding of joint sources dates back to the works of Slepian, Wolf [2] and Wyner, Ziv [3]. It gained attention recently with practical development of sensor networks, see Xiong, Liveris, Cheng [4].

As a step further, we suggest to use product quantization with an objective to improve quality of the output on the signal, not the signal itself. Note that this is exactly what happens when an integral of a function is being approximated with a discretization, one of the oldest appearances of quantization idea itself. Moreover, in recent studies related to quantization of neural networks, see for example [8], the most important part is also to improve quality of the output function. Thus, it seems natural to combine an efficient product quantization technique with an important goal of controlling quality of the output. This problem seems to be underdeveloped from a mathematical perspective, so in this work we lay its foundation and study general properties as well as asymptotic results for the most natural output functions.

1.1. Functional product quantization problem.

Let us fix $d = 2$ for simplicity. Assume that 2 sets $\mathcal{X}_1, \mathcal{X}_2$ with random elements $X_i \in \mathcal{X}_i$ and their common distribution law being given. Let $\mu = \mathrm{law}(X_1, X_2)$ be a Borel probability measure. Finally, assume that a Borel function $f\colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ is given. For $(n_1, n_2) \in \mathbb{N} \times \mathbb{N}$ one has to find the quantization maps

$$q_i\colon \mathcal{X}_i \to \mathcal{X}_i, \#q_i(\mathcal{X}_i) \leq n_i, i = 1, 2$$

such that for a given distance d on $\mathbb{R}$

$$L_f(q_1, q_2) := \mathbb{E}\, d\left(f\left(X_1, X_2\right), f\left(q_1(X_1), q_2(X_2)\right)\right) \to \min .$$

In this paper, we are mainly interested in the asymptotics of the *quantization cost*

$$C_f(n_1,n_2) := \inf\{L_f(q_1,q_2)\colon \#q_1(\mathcal{X}_1)\leq n_1, \#q_2(\mathcal{X}_2)\leq n_2\}.$$

Throughout the paper, we will assume, unless otherwise explicitly stated, the most common situation in applications, namely that $\mathcal{X}_i=\mathbb{R}^{k_i}$ is just the Euclidean space and $\mu\ll\mathcal{L}^{k_1}\otimes\mathcal{L}^{k_2}$ with compact support. Even more, in most cases, we will limit ourselves to the case $k_1=k_2=1$, i.e. $\mathcal{X}_1=\mathcal{X}_2=\mathbb{R}$, $\mu\ll\mathcal{L}^2$. We will see that this case already contains all the essential difficulties of the problem considered.

1.2. **Comparison with classical quantization.** The functional product quantization problem introduced above has to be compared with the following (relatively) well studied classical quantization problem, namely, that of finding the quantization map

$$q\colon \mathcal{Z}\ := \mathcal{X}\times\mathcal{Y}\to\mathcal{Z},\quad \#q(\mathcal{Z})\leq N,$$

so that

$$L(q) := \mathbb{E}\,c\left(Z,q(Z)\right)\to\min,$$

where c is the given cost on $\mathcal{Z}$. In other words, here, as opposed to the functional product quatization problem, one would like to quantize just the input vector minimizing the error on the input, i.e. the expectation of the norm of the difference between Z and its quantized version, without taking in consideration the function f to be calculated on the input. The cost of such classical quantization is given by

$$C(N) := \inf\{L_f(q)\colon \#q(\mathcal{Z})\leq N\}.$$

The case $\mathcal{X}=\mathcal{Y}=[0,1]\subset\mathbb{R}$, so that $\mathcal{Z}=[0,1]^2$ and $\mu=\mathcal{L}^2\llcorner[0,1]^2$ is the most well studied. In this case

$$C(N)\sim C/\sqrt{N},$$

with $C>0$ known.

2. Notation and preliminaries

For brevity we denote the whole space as $\mathcal{X}:=\mathcal{X}_1\times\ldots\times\mathcal{X}_d$, the signal as $X:=(X_1,\ldots,X_d)$ and the set of quantizers as $q:=(q_1,\ldots,q_d)$. Then, q_i has n_i values that we denote as $a_i^{s_i}, s_i=1,\ldots,n_i$. Define $A_i^{s_i}:=q_i^{-1}(a_i^{s_i}), s_i=1,\ldots,n_i, i=1,\ldots,d$.

Sometimes to emphasize the dependence of the costs on c and μ we write $L_{f,c,\mu}(q)$ and $C_{f,c,\mu}(n_1,\ldots,n_d)$ instead of $L_f(q)$ and $C_f(n_1,\ldots,n_d)$ respectively. Also for the classical quantization problem, to emphasize the dependence of the cost on c and μ we may write $L_{c,\mu}(q)$ and $C_{c,\mu}(n_1,\ldots,n_d)$ instead of $L(q)$ and $C(n_1,\ldots,n_d)$ respectively.

For a Borel measure μ on a metric space E and $D\subset E$ Borel, we let $\mu\llcorner D$ stand for the restriction of μ to D and by 1_D the characteristic function of D. If μ and ν are measures with μ absolutely continuous with respect to ν, we write $\mu\ll\nu$. By $\mathcal{L}^d$ we denote the Lebesgue measure over the Euclidean space $\mathbb{R}^d$. The notation $L^p(E,\mu)$ stands for the usual Lebesgue space of functions over a metric space E which are p-integrable with respect to μ, if $1\leq p<+\infty$, or μ-essentially bounded, if $p=+\infty$. The norm in this space is denoted by $\|\cdot\|_p$. The reference to the metric space E will be often omitted from the notation when not leading to a confusion, i.e. we will often write $L^p(\mu)$ instead of $L^p(E,\mu)$. Similarly, if $E=\mathbb{R}^d$ is a Euclidean space and $\mu=\mathcal{L}^d$ is the Lebesgue measure, then we will omit the reference to μ

writing just $L^p(\mathbb{R}^d)$ instead of $L^p(\mathbb{R}^d,\mu)$. The weak* convergence in $L^\infty(E,\mu)$ is denoted by $\overset{*}{\rightharpoonup}$. For a random variable Y we denote by $\mathbb{E}\,(Y)$ its expectation, by $\mathrm{Var}\,(Y)$ its variance and by $\mathrm{law}(Y)$ its law.

3. A bridge between classical and functional product quantization

The quantization of only one of the variables is a bridge between classical case and the one we are studying. In this case the following estimate is considered

$$L_f(q) = \mathbb{E}\, c(f(X,Y), f(q(X),Y))$$

and

$$C_f(N) = \inf\{L_f(q) : \# q(X) \le N\}.$$

On one hand, if c and f are continuous and the support of the measure is compact, by taking a uniform quantization over the second coordinate we get

$$C_f(N) \ge \lim_{n_2\to\infty} C_f(N, n_2).$$

Surprisingly, the reverse inequality is not true. Even the slightest quantization of the second coordinate may drastically decrease the total error, as the following example shows.

Example 3.1. Let $\mu := \mathcal{L}^2\llcorner[0,1]\times[0,1]$, $c(u,v) := |u-v|$ and let

$$f(x,y) := (1_{[1/3,2/3]\times[0,1/3]} + 1_{[0,1/3]\times[1/3,2/3]} + 1_{[2/3,1]\times[2/3,1]})(x,y).$$

Let $N := 1$. Then whatever q is, one has that $f(q(x),y)$ differs from $f(x,y)$ on the union of 4 squares of the total area $4/9$, so that $C_f(1) = 4/9$. On the other hand, if $q([0,1]) \in (0,1/3)$ and $q_2([0,1]) \in (0,1/3)$, then $f(q(x),q_2(y))$ differs from $f(x,y)$ on the union of 3 squares of the total area $3/9$, so that

$$4/9 = C_f(1) > 3/9 \ge C_f(1,1) \ge C_f(1,n_2)$$

for all $n_2 \in \mathbb{N}$. Note that this result does not change if we ask for f to be smooth, since one can just approximate a characteristic function with smooth functions.

4. Random quantization and existence of optimal quantizers

The goal of this section is to prove the existence of optimal quantizers. For a particular quantizing lattice $w := \{(x_1^{s_1},\dots,x_d^{s_d}), s_i = 1,\dots,n_i\}$ denote values of f at its points as $f(w) = (f(x_1^{s_1},\dots,x_d^{s_d}))_{s_i=1,\dots,n_i}$. Denote by W the set of all lattices with $x_i^{s_i} \in \mathcal{X}_i$ and by $f(\mathrm{W}) = \{f(w)\colon w \in \mathrm{W}\} \subset \mathbb{R}^{n_1\cdots n_d}$. Essentially, $f(\mathrm{W})$ describes all the potential quantizations of the output. In order to have the existence of optimal quantizers we request $f(\mathrm{W})$ to be compact. Note that this requirement is in particular satisfied in the following two important cases indicated in the statement below

Proposition 4.1. *The set f(W) is compact in $\mathbb{R}^{n_1\cdots n_d}$, in particular, when either*

(A) f has finite set of values

(B) or f is continuous and all $\mathcal{X}_i$ are compact.

Proof. In case (A) the set $f(\mathrm{W})$ is finite thus compact.

For the case (B) $f(\mathrm{W})$ is precompact as a subset of $f(\mathcal{X})^{n_1\cdots n_d}$. To show that it is closed consider a sequence of lattices w_k such that $f(w_k)$ converges. Then, since all $\mathcal{X}_i$ are compact metric spaces, we can pick a subsequence of lattices (not relabelled) such that each point $x_{i,k}^{s_i}$ converges to some $x_i^{s_i}$ for $i = 1, \dots, d, s_i = 1, \dots, n_i$. Then, for all $s_i = 1, \dots, n_i$ one has

$$f(x_{1,k}^{s_1}, \dots, x_{d,k}^{s,d}) \to f(x_1^{s_1}, \dots, x_d^{s_d}).$$

Thus $f(w_k) \to f(w)$ where $w = \{(x_1^{s_1}, \dots, x_d^{s_d}), s_i = 1, \dots, n_i\}$, proving the claim. □

We often face a situation of non-compact $\mathcal{X}_i$, for instance $\mathcal{X}_i = \mathbb{R}$. If $\mathcal{X}_i$ are not compact it is easy to construct an example with nice continuous functions such that the problem has no minimizers, see Example 4.2. However, for practical use in engineering applications the sets $\mathcal{X}_i$ may always assumed to be compact.

Example 4.2. Consider $f(x,y) := x + y$, $c(u,v) := e^{-|u-v|^2}$ and $\mu := \mathcal{L}^2 \llcorner [0,1]^2$. Take $n_1 = n_2 = 1$ and $q_{1,k}(x) = q_{2,k}(x) = k$. Then $\mathcal{L}_f(q_{1,k}, q_{2,k}) \to 0$, but there is no quantizers providing zero cost.

Theorem 4.3. *Assume that $\mu = w(x_1, \dots, x_d)\mu_1 \otimes \dots \otimes \mu_d$ for Borel probability measures μ_i on $\mathcal{X}_i$ and $w(x_1, \dots, x_d) \in L^1(\mathcal{X}, \mu_1 \otimes \dots \otimes \mu_d)$. Let $f(W)$ be compact $c(u,v) \geq 0$ and the map $v \mapsto c(u,v)$ be lower semicontinuous for all u. Then the best quantization error $\mathcal{C}_f(n_1, \dots, n_d)$ is achievable as $\mathcal{L}_f(q_1, \dots, q_d)$ for some quantizers $q_1, \dots, q_d$.*

To prove this result we will introduce the relaxed problem setting, that of random quantization, show that it has solution, and then show that the same quantization error can be achieved by usual (non random, or deterministic) quantizers.

4.1. **Random quantization.**

In a random quantization setting we are looking for sets of n_i quantization points $\{x_i^1, \dots, x_i^{n_i}\} \subset \mathcal{X}_i$ and weight functions $p_i^1, \dots, p_i^{n_i}$ such that for all $x \in \mathbb{R}$ one has

$$0 \leq p_i^{s_i}(x) \leq 1 \quad \text{for all } s_i = 1, \dots, n_i, \qquad \sum_{s_i=1}^{n_i} p_i^{s_i}(x) = 1$$

where $i = 1, \dots, d$. For brevity we denote

$$\bar{p}_i(\cdot) := (p_i^1(\cdot), \dots, p_i^{n_i}(\cdot)), \qquad \bar{x}_i := (x_i^1, \dots, x_i^{n_i}).$$

The best random quantization by definition minimizes the error

$$\begin{aligned} &\mathcal{L}_f(\bar{p}_1, \dots, \bar{p}_d, \bar{x}_1, \dots, \bar{x}_d) \\ &\quad := \sum_{s_1=1}^{n_1} \dots \sum_{s_d=1}^{n_d} \int_{\mathcal{X}} p_1^{s_1}(x_1) \dots p_d^{s_d}(x_d) c(f(x), f(x_1^{s_1}, \dots, x_d^{s_d}))\, d\mu(x). \end{aligned}$$

In other words, we pick n_i quantizing points in $\mathcal{X}_i$ and we quantize every point x_i in one of $x_i^1, \dots, x_i^{n_i}$ with probabilities $p_i^1(x_i), \dots, p_i^{n_i}(x_i)$ independently from everything else.

Nonrandom quantization problem that we are most interested in corresponds to the case of random quantization where all the weights except one are zero, i.e. $p_i^{s_i}(x_i) = \delta(x_i^{s_i}, q_i(x_i))$, where $\delta(a,b)$ stands for Kronecker symbol.

The following proposition shows that the best error for a random quantization problem is achievable.

Proposition 4.4. *Assume that $\mu = w(x_1,\ldots,x_d)\mu_1\otimes\ldots\otimes\mu_d$, for Borel probability measures μ_i on $\mathcal{X}_i$ and $w(x_1,\ldots,x_d)\in L^1(\mathcal{X},\mu_1\otimes\ldots\otimes\mu_d)$. Let $f(W)$ be compact, $c(u,v)\geq 0$ and the map $v\mapsto c(u,v)$ be lower semicontinuous for all u. Then random quantization functional $\mathcal{L}_f$ attains its minimum.*

Proof. The proof is divided in two steps.

Step 1. We will further prove that if $p_{i,k}^{s_i}\overset{*}{\rightharpoonup} p_i^{s_i}$ in $L^\infty(\mathcal{X}_i,\mu_i)$ (here $\overset{*}{\rightharpoonup}$ denotes weak* convergence) and $f(x_{1,k}^{s_1},\ldots,x_{d,k}^{s_d})\to a_{s_1,\ldots,s_d}$ as $k\to\infty$, then

$$\begin{aligned}\liminf_{k\to\infty}&\left(\int_{\prod_{j=1}^d\mathcal{X}_j} p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d)c(f(x),f(x_{1,k}^{s_1},\ldots,x_{d,k}^{s_d}))d\mu(x)\right)\\ &\geq\int_{\prod_{j=1}^d\mathcal{X}_j} p_1^{s_1}(x_1)\ldots p_d^{s_d}(x_d)c(f(x),a_{s_1,\ldots,s_d})d\mu(x).\end{aligned}\tag{4.1}$$

Taking for the moment (4.1) for granted, we deduce from it the lower semicontinuity of $\mathcal{L}_f$. Namely, we show that, denoting

$$\bar{p}_{i,k}(\cdot) := (p_{i,k}^1(\cdot),\ldots,p_{i,k}^{n_i}(\cdot)),\qquad \bar{x}_{i,k} := (x_{i,k}^1,\ldots,x_{i,k}^{n_i}),$$

one has

$$\begin{aligned}&\liminf_{k\to\infty}\mathcal{L}_f(\bar{p}_{1,k},\ldots,\bar{p}_{d,k},\bar{x}_{1,k},\ldots,\bar{x}_{d,k})\\ &=\liminf_{k\to\infty}\sum_{s_1=1}^{n_1}\cdots\sum_{s_d=1}^{n_d}\left(\int_{\prod_{j=1}^d\mathcal{X}_j} p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d)c(f(x),f(x_{1,k}^{s_1},\ldots,x_{d,k}^{s_d}))\,d\mu(x)\right)\\ &\geq\sum_{s_1=1}^{n_1}\cdots\sum_{s_d=1}^{n_d}\liminf_{k\to\infty}\left(\int_{\prod_{j=1}^d\mathcal{X}_j} p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d)c(f(x),f(x_{1,k}^{s_1},\ldots,x_{d,k}^{s_d}))\,d\mu(x)\right)\\ &\geq\sum_{s_1=1}^{n_1}\cdots\sum_{s_d=1}^{n_d}\int_{\prod_{j=1}^d\mathcal{X}_j} p_1^{s_1}(x_1)\ldots p_d^{s_d}(x_d)c(f(x),a_{s_1,\ldots,s_d})d\mu(x)\\ &=\mathcal{L}_f(\bar{p}_1,\ldots,\bar{p}_d,\bar{x}_1,\ldots,\bar{x}_d),\end{aligned}$$

where points $x_i^{s_i}$ are such that $a_{s_1,\ldots,s_d}=f(x_1^{s_1},\ldots,x_d^{s_d})$. Note that such points exist because $f(\mathrm{W})$ is closed, thus limit of values of f on a sequence of lattices is a value of f on some lattice. To finish the proof it remains to take a minimizing sequence of $\bar{p}_{1,k},\ldots,\bar{p}_{d,k},\bar{x}_{1,k},\ldots,\bar{x}_{d,k}$ for $\mathcal{L}_f$, extract convergent subsequences (not relabeled) such that $p_{i,k}^{s_i}\overset{*}{\rightharpoonup} p_i^{s_i}$ in $L^\infty(\mathcal{X}_i,\mu_i)$, $f(x_{1,k}^{s_1},\ldots,x_{d,k}^{s_d})\to a_{s_1,\ldots,s_d}$ as $k\to\infty$ for all $i=1,\ldots,d$, $s_i=1,\ldots,n_i$, and apply the inequality above. Note, that a convergent subsequence can be chosen because a unit ball in $L^\infty(\mathcal{X}_i,\mu_i)$ with weak* topology is compact and metrizable, while $f(\mathrm{W})$ is assumed to be compact.

Step 2. It remains thus to prove (4.1). To thos aim let us show that

$$p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d)\overset{*}{\rightharpoonup} p_1^{s_1}(x_1)\ldots p_d^{s_d}(x_d)\quad\text{ in } L^\infty(\mathcal{X},\mu).\tag{4.2}$$

It suffices in fact to check that for $\phi\in L^\infty(\mathcal{X},\mu)$ one has

$$\int_{\mathcal{X}} p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d)\phi(x)d\mu\to\int_{\mathcal{X}} p_1^{s_1}(x_1)\ldots p_d^{s_d}(x_d)\phi(x)d\mu.$$

The latter is true, because $\phi(x)w(x_1,\dots,x_d) \in L^1(\mathcal{X},\mu_1\otimes\ldots\otimes\mu_d)$ and

$$p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d) \overset{*}{\rightharpoonup} p_1^{s_1}(x_1)\ldots p_d^{s_d}(x_d) \text{ in } L^\infty(\mathcal{X},\mu_1\otimes\ldots\otimes\mu_d),$$

thus proving (4.2).

Now, from (4.2) one has that the sequence of measures $p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d)d\mu(x)$ converges setwise to the measure $p_1^{s_1}(x_1),\ldots p_d^{s_d}(x_d)d\mu(x)$, because for any Borel $A\subset\mathcal{X}$ one has $\mathbf{1}_A\in L^1(\mathcal{X},\mu)$, and thus

$$\int_A p_{1,k}^{s_1}(x_1)\ldots p_{d,k}^{s_d}(x_d)d\mu(x) \to \int_A p_1^{s_1}(x_1),\ldots p_d^{s_d}(x_d)d\mu(x).$$

Now, the statement (4.1) follows from the Fatou lemma with varying measures [9, section 11.4, proposition 17] □

4.2. **Existence of nonrandom optimal quantizers.** Now we are going to show that this minimum can be obtained by nonrandom quantizers, and therefore the best error in nonrandom quantization is also achievable.

Proof of Theorem 4.3: We are going to prove a stronger statement, namely that although nonrandom quantization is a particular case of random quantization, the best quantizers are actually nonrandom. For the proof there is no need in assumptions on $\mu, f(x), c(u,v)$, they only appear so that the best quantizers in random setting exist. Consider the optimum for a random quantization problem $p_i^{s_i}(x_i), x_i^{s_i}, s_i = 1,\dots,n_i, i = 1,\dots,d$. We will show that it is achievable by nonrandom quantizers. We disintegrate

$$\mu(x_1,\dots,x_d) = \mu_{x_i}(x_1,\dots,x_{i-1},x_{i+1},\dots,x_d)\otimes d\mu_{X_i}(x_i),$$

where μ_{x_i} are the rspective conditional measures. Among all optimal quantizers $p_i^{s_i}, x_i^{s_i}$ pick one with the least number of random quantizers (we name a quantizer $p_i^{s_i}, s_i = 1,\dots,n_i$ non-random, if one of the weights is one and the others are zero), and show that it is non-random (i.e. the number of random quantizers is zero). Suppose the contrary. Without loss of generality we may assume that p_1^s is not random. Define

$$\hat{s}_1(x_1) := \operatorname*{arg\,min}_{s_1=1,\dots,n_1} g_{x_1}(s_1), \quad \text{where}$$

$$g_{x_1}(s_1) :=$$

$$\int\limits_{\mathcal{X}_2\times\ldots\times\mathcal{X}_d} \sum_{s_2,\dots,s_d} p_2^{s_2}(x_2)\ldots p_d^{s_d}(x_d)c(f(x),f(x_1^{s_1},\dots,x_d^{s_d}))d\mu_{x_1}(x_2,\dots,x_d).$$

Here and below we abbreviate $\sum_{s_1=1}^{n_1}\dots\sum_{s_d=1}^{n_d}$ as $\sum_{s_1,\dots,s_d}$. Denoting $x^{\bar{s}} := (x_1^{s_1},\dots,x_d^{s_d})$ for brevity, one clearly has

$$\begin{aligned}
&\int\limits_{\mathcal{X}} \sum_{s_1,\dots,s_d} p_1^{s_1}(x_1)\dots p_d^{s_d}(x_d) c(f(x), f(x^{\bar{s}}))d\mu(x)\\
&\quad= \int\limits_{\mathcal{X}_1} \sum_{s_1=1}^{n_1} p_1^{s_1}(x_1) g_{x_1}(s_1)\, d\mu_{\mathcal{X}_1}(x_1)\\
&\quad\geq \int\limits_{\mathcal{X}_1} \sum_{s_1=1}^{n_1} p_1^{s_1}(x_1) g_{x_1}(\hat{s}_1(x_1))\, d\mu_{\mathcal{X}_1}(x_1) = \int\limits_{\mathcal{X}_1} g_{x_1}(\hat{s}_1(x_1))\, d\mu_{\mathcal{X}_1}(x_1)\\
&\quad= \int\limits_{\mathcal{X}} \sum_{s_2,\dots,s_d} p_2^{s_2}(x_2)\dots p_d^{s_d}(x_d) c(f(x), f(x_1^{\hat{s}_1(x_1)}, x_2^{s_2},\dots,x_d^{s_d}))\, d\mu(x).
\end{aligned}$$

In other words, we transformed random quantizer $p_1^s(x_1)$ into non-random one (corresponding to the choice of quantization function $q_1(x_1) = x_1^{\hat{s}_1(x_1)}$) without increasing the cost. Thus, this is an optimal quantizer with less random quantizers than before, contradicting the construction. Thus, there were no random quantizers to begin with, meaning that there is an optimal completely non-random quantization strategy. □

Remark 4.5. As a byproduct of the above proof we have that the best quantization error is equal to the best random quantization error.

4.3. **Properties of quantizing sets.**

We prove here a simple property of optimal quantizers

Lemma 4.6. *Let f be bounded, $c(u,v)\geq 0$ and $c(u,v)=0$ only if $u=v$, the map $v\mapsto c(u,v)$ be lower semicontinuous for all u, and $\mu(f^{-1}(\lambda))=0$ for all $\lambda\in\mathbb{R}$. Let q_i, $i=1,\dots,d$, be quantization maps. Denoting $\{a_i^{s_i}\}_{s_i=1}^{n_i} := q_i(\mathcal{X}_i)$, set*

$$A_i^{s+i}(n_i) := q_i^{-1}(a_i^{s_i}).$$

Assuming that $L_f(q_1,\dots,q_d)\to 0$ as $(n_1,\dots,n_d)\to\infty$, one has then

$$\max_{s_1,\dots,s_d} \mu(A_1^{s_1}(n_1)\times\ldots\times A_d^{s_d}(n_d))\to 0, \qquad \text{as } n_1,\dots,n_d\to\infty.$$

Proof. If not, there is an $\varepsilon>0$ and some $A_1^{s_1}(n_1),\dots,A_d^{s_d}(n_d)$ with

$$\mu(A_1^{s_1}(n_1)\times\ldots\times A_d^{s_d}(n_d))\geq\varepsilon \quad \text{with } s_i = s_i(n_i).$$

Note that

$$L_f(q_1,\dots,q_d)\geq \int_{A_1^{s_1}(n_1)\times\ldots\times A_d^{s_d}(n_d)} c(f(x), f(a_1^{s_1},\dots,a_d^{s_d}))\, d\mu(x).$$

Up to a subsequence (not relabeled) one has $\mathbf{1}_{A_1^{s_1}(n_1)\times\ldots\times A_d^{s_d}(n_d)}\to\varphi$ in the weak* sense of $L^\infty(\mu)$ and $f(a_1^{s_1},\dots,a_d^{s_d})\to\lambda$ as $(n_1,\dots,n_d)\to\infty$. Moreover,

$$\int_{\mathcal{X}}\varphi\, d\mu\geq\varepsilon$$

and $\varphi \geq 0$ μ-a.e. Therefore, again due to the Fatou lemma with varying measures [9, section 11.4, proposition 17], one has

$$\begin{aligned}
&\int_{\mathcal{X}} \varphi(x) c(f(x), \lambda)\, d\mu(x) \\
&\leq \liminf_{(n_1,\ldots,n_d)\to\infty} \int_{\mathcal{X}} \mathbf{1}_{A_1^{s_1}(n_1)\times\ldots\times A_d^{s_d}(n_d)}(x) c(f(x), f(a_1^{s_1},\ldots,a_d^{s_d}))\, d\mu(x) \\
&\leq \liminf_{(n_1,\ldots,n_d)\to\infty} L_f(q_1,\ldots,q_d) = 0.
\end{aligned}$$

Since $c \geq 0$ this gives

$$\int_{\mathcal{X}} \varphi(x) c(f(x), \lambda)\, d\mu(x) = 0,$$

which implies $f(x) = \lambda$ on the set $\{\varphi(x) > 0\}$ which has positive measure μ, contrary to the assumptions. □

5. Optimal quantizers for particular classes of functions

5.1. **Characteristic functions of measurable rectangles and their finite sums.**

We first consider the case when f is a characteristic function of a measurable rectangle, i.e. $f = \mathbf{1}_{A_1\times\ldots\times A_d}$ for $A_i \subset \mathcal{X}_i$ measurable sets.

Proposition 5.1. *If $f(x) = \mathbf{1}_{A_1\times\ldots\times A_d}(x)$, with measurable $A_i \subset \mathcal{X}_i$ then for $n_i \geq 2$ for all $i = 1,\ldots,d$, one has $C_f(n_1,\ldots,n_d) = 0$.*

Proof. Take $a_i^1 \in A_i, a_i^2 \in \mathcal{X}_i \setminus A_i$ and set

$$q_i(x_i) \quad := \quad \begin{cases} a_i^1, & x_i \in A_i, \\ a_i^2, & x_i \in \mathcal{X} \setminus A_i, \end{cases}$$

□

Now, it is easy to generalize this to the case of f being a finite sum of characteristic functions of measurable rectangles.

Proposition 5.2. *If*

$$f(x) = \sum_{j=1}^{N} c_j \mathbf{1}_{A_1^j}(x_1)\ldots\mathbf{1}_{A_d^j}(x_d),$$

where $A_i^j \subset \mathcal{X}_i$ whatever is $\mathcal{X}_i$, then there is an $\bar{N}$ such that for $n_i \geq \bar{N}$, one has $C_f(n_1,\ldots,n_d) = 0$.

Proof. Let us encode each point with the sets containing it. Denote

$$e_i(x_i) = \Big(\mathbf{1}_{A_i^j}(x_i)\Big)_{j=1}^{N}.$$

By definition the images of e_i are binary codes of size N. For every binary code w in the image $e_i(\mathcal{X})$ pick x_i^w such that $e_i(x_i^w) = w$. Consider the following quantization: $q_i(x_i) = x_i^{e_i(x)}$. Then for all $x \in \mathcal{X}$ $e_i(x_i) = e_i(q_i(x_i))$. Therefore from definition of e_i one has

$$f(x) = f(q_1(x_1),\ldots,q_d(x_d)).$$

Consequently, $L_f(q_1,\ldots,q_d) = 0$ for any cost function c. □

Remark 5.3. Note that in Proposition 5.2

(1) the measurable rectangles $A_1^j \times \ldots \times A_d^j$ may be intersecting.
(2) in general, one has $\bar N = O(2^N)$ as $N \to \infty$ because it is a total number of binary strings of length N. Nevertheless, when $\mathcal{X}_i = \mathbb{R}$ and all A_i^j are intervals one has $\bar N \leq 2N$.
(3) the statement is constructive, i.e. it provides an algorithm for quantization.

To prove (2) note that N intervals in $\mathbb{R}$ divide it into at most $2N$ parts. Moreover, all of them, except the union of two rays, are intervals. The encodings $e_i(\mathcal{X}_i)$ are constant on these intervals, therefore their images consist of at most $2N$ elements.

Finally, the reverse statement, that only the finite sum of characteristic functions of measurable rectangles has zero-quantization cost, is also true to some extent.

Proposition 5.4. *Let $c \geq 0$ be a Borel function such that $c(u,v) = 0$ only if $u = v$. If $C_f(n_1, \ldots, n_d) = 0$ and this error is achievable, then there are disjoint measurable sets $A_i^{s_i} \subset \mathcal{X}$, $s_i = 1, \ldots, n_i$, $i = 1, \ldots, d$ such that the union $\cup_{s_1,\ldots,s_d} A_1^{s_1} \times \ldots \times A_d^{s_d}$ covers $\mathcal{X}$ up to a μ-negligible set and*

$$f(x) = \sum_{s_1=1}^{n_1} \ldots \sum_{s_d=1}^{n_d} c_{s_1,\ldots,s_d} \mathbf{1}_{A_1^{s_1}}(x_1) \ldots \mathbf{1}_{A_d^{s_d}}(x_d) \tag{5.1}$$

for some $c_{s_1,\ldots,s_d} \in \mathbb{R}$, whatever are $\mathcal{X}_i$.

Proof. By definition there are $q_1, \ldots, q_d$ such that $\mathcal{L}_f(q_1, \ldots, q_d) = 0$. If $q_i(\mathcal{X}_i) = \{a_i^s\}_{s=1}^{n_i}$, set $A_i^s = q_i^{-1}(a_i^s)$. One has then

$$\begin{aligned} 0 = \mathcal{L}_f(q_1, \ldots, q_d) &= \int_{\mathcal{X}} c(f(x), f(q_1(x_1), \ldots, q_d(x_d)))\, d\mu(x) \\ &= \sum_{s_1=1}^{n_1} \ldots \sum_{s_d=1}^{n_d} \int_{A_1^{s_1} \times \ldots \times A_d^{s_d}} c(f(x), f(a_1^{s_1}, \ldots, a_d^{s_d}))\, d\mu(x) \end{aligned}$$

which means that $f(x) = f(a_1^{s_1}, \ldots, a_d^{s_d})$ for μ - a.e. $x \in A_1^{s_1} \times \ldots \times A_d^{s_d}$. Denote $c_{s_1,\ldots,s_d} = f(a_1^{s_1}, \ldots, a_d^{s_d})$ and get that (5.2) is true. □

We can now apply Theorem 4.3 to get the following statement.

Corollary 5.5. *Suppose that $\mu = w\mu_1 \otimes \ldots \otimes \mu_d$ with Borel probability measures μ_i on $\mathcal{X}_i$, $w \in L^1(\mathcal{X}, \mu_1 \otimes \ldots \otimes \mu_d)$ and $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is nonnegative Borel function such that the map $v \mapsto c(u,v)$ is lower semicontinuous for all u. If, moreover, f is bounded and $c(u,v) = 0$ iff $u = v$, then $C_f(n_1, \ldots, n_d) = 0$ implies that there are disjoint measurable sets $A_i^{s_i} \subset \mathcal{X}_i$, $s_i = 1, \ldots, n_i$, $i = 1, \ldots, d$ such that the union $\cup_{s_1,\ldots,s_d} A_1^{s_1} \times \ldots \times A_d^{s_d}$ covers $\mathcal{X}$ up to a μ-negligible set and for μ-a.e. x one has*

$$f(x) = \sum_{s_1=1}^{n_1} \ldots \sum_{s_d=1}^{n_d} c_{s_1,\ldots,s_d} \mathbf{1}_{A_1^{s_1}}(x_1) \ldots \mathbf{1}_{A_d^{s_d}}(x_d) \tag{5.2}$$

for some $c_{s_1,\ldots,s_d} \in \mathbb{R}$.

Proof. Under the assumptions of corollary being proven the zero cost is achievable by Theorem 4.3 and Proposition 4.1 once one shows that f has a finite number of values. This would allow us to use Proposition 5.4 to finish the proof. However, this property cannot be proven for f directly, and therefore we are going to construct a new function $\tilde f$ with a finite set of values, that equals f μ-a.e. and has zero quantization cost. To his aim, consider a sequence of quantizers $q_{1,k},\dots,q_{d,k}$ such that

$$\begin{aligned}0 = \lim_{k\to\infty}\mathcal{L}_f(q_{1,k},\dots,q_{d,k}) &= \int_{\mathcal{X}} c(f(x), f(q_{1,k}(x_1),\dots,q_{d,k}(x_d)))\,d\mu(x)\\ &= \sum_{s_1=1}^{n_1}\cdots\sum_{s_d=1}^{n_d}\int_{A^{s_1}_{1,k}\times\ldots\times A^{s_d}_{d,k}} c(f(x), f(a^{s_1}_{1,k},\dots,a^{s_d}_{d,k}))\,d\mu(x).\end{aligned}$$

Now, by taking a weak* converging subsequence (not relabelled) we obtain that $\mathbf{1}_{A^{s_1}_{1,k}\times\ldots\times A^{s_d}_{d,k}} \overset{*}{\rightharpoonup} \phi_{s_1,\dots,s_d}$ in $L^\infty(\mathcal{X},\mu)$ for all $s_i = 1,\dots,n_i$. Clearly $\phi_{s_1,\dots,s_d}(x_1,\dots,x_d)\in[0,1]$ μ-a.e. Note that since

$$\sum_{s_1,\dots,s_d}\mathbf{1}_{A^{s_1}_{1,k}\times\ldots\times A^{s_d}_{d,k}}(x_1,\dots,x_d) = 1$$

for all $x_i\in\mathcal{X}_i$, one has

$$\sum_{s_1,\dots,s_d}\phi_{s_1,\dots,s_d}(x_1,\dots,x_d) = 1$$

for μ-a.e. $(x_1,\dots,x_d)$. Moreover, consider a subsequence (not relabelled) such that $f(a^{s_1}_{1,k},\dots,a^{s_d}_{d,k})$ converges to some $c_{s_1,\dots,s_d}\in\mathbb{R}$. Now, from weak* convergence we get that the measure $\mathbf{1}_{A^{s_1}_{1,k}\times\ldots\times A^{s_d}_{d,k}}(x)d\mu(x)$ setwise converges to $\phi_{s_1,\dots,s_d}(x)d\mu(x)$. Thus, by Fatou lemma with varying measures [9, section 4, proposition 17], we get

$$\begin{aligned}0 &= \lim_{k\to\infty}\int_{A^{s_1}_{1,k}\times\ldots\times A^{s_d}_{d,k}} c(f(x), f(a^{s_1}_{1,k},\dots,a^{s_d}_{d,k}))\,d\mu(x)\\ &\geq \int_{\mathcal{X}}\phi_{s_1,\dots,s_d}(x_1,\dots,x_d)\liminf_{k\to\infty} c(f(x), f(a^{s_1}_{1,k},\dots,a^{s_d}_{d,k}))\,d\mu(x)\\ &\geq \int_{\mathcal{X}}\phi_{s_1,\dots,s_d}(x_1,\dots,x_d)c(f(x),c_{s_1,\dots,s_d})\,d\mu(x),\end{aligned}$$

where the last inequality follows from lower semicontinuity of $v\mapsto c(u,v)$. Since integrand of the r.h.s. is non-negative, then

$$\int_{\mathcal{X}}\phi_{s_1,\dots,s_d}(x_1,\dots,x_d)c(f(x),c_{s_1,\dots,s_d})\,d\mu(x) = 0. \tag{5.3}$$

Thus $f(x) = c_{s_1,\dots,s_d}$ μ-a.e. on a set $D_{s_1,\dots,s_d} = \{\phi_{s_1,\dots,s_d} > 0\}$. Consequently, f has a finite number of values μ-a.e. Now, let us construct $\tilde f$ with a finite set of values that has zero-cost and equals f μ-a.e. First of all, take $\tilde f := f$ on $D_{s_1,\dots,s_d}$ for all s_i and set it to 0 elsewhere. Secondly, take any lattice $w = (x_1^{s_1},\dots,x_d^{s_d}), s_i = 1,\dots,n_i$ and redefine

$$\tilde f(x_1^{s_1},\dots,x_d^{s_d}) := c_{s_1,\dots,s_d}.$$

We claim that $C_{\tilde f}(n_1,\dots,n_d) = 0$. Define $\tilde q_{i,k}:\mathcal{X}_i\to\mathcal{X}_i, i = 1\dots,d$ by setting

$$\tilde q_{i,k}(x) := x_i^{s_i}, \text{if } x\in A^{s_i}_{i,k}, \qquad s_i = 1,\dots,n_i.$$

In other words, we leave quantizing sets the same as for f, but instead of taking $f(a^{s_1}_{1,k},\dots,a^{s_d}_{d,k})$ as values, we take $c_{s_1,\dots,s_d}$. Clearly, from weak* convergence of

$\mathbf{1}_{A_{1,k}^{s_1}\times\ldots\times A_{d,k}^{s_d}} \rightharpoonup \phi_{s_1,\ldots,s_d}$ in $L^\infty(\mathcal{X},\mu)$, one has

$$\begin{aligned}
&\lim_{k\to\infty}\int_{A_{1,k}^{s_1}\times\ldots\times A_{d,k}^{s_d}} c(\tilde f(x),\tilde f(x_1^{s_1},\ldots,x_d^{s_d}))d\mu(x)\\
&\qquad=\lim_{k\to\infty}\int_{A_{1,k}^{s_1}\times\ldots\times A_{d,k}^{s_d}} c(\tilde f(x), c_{s_1,\ldots,s_d})d\mu(x)\\
&\qquad=\int_{\mathcal{X}}\phi_{s_1,\ldots,s_d}(x_1,\ldots,x_d)c(\tilde f(x),c_{s_1,\ldots,s_d})\,d\mu(x)\\
&\qquad=\int_{\mathcal{X}}\phi_{s_1,\ldots,s_d}(x_1,\ldots,x_d)c(f(x),c_{s_1,\ldots,s_d})\,d\mu(x)\quad\text{since } \tilde f = f\ \mu\text{-a.e.}\\
&\qquad=0,\quad\text{by (5.3)},
\end{aligned}$$

which proves $C_{\tilde f}(n_1,\ldots,n_d)=0$. Consequently, by Proposition 4.1 and Theorem 4.3 we get that the best quantization error is achievable for $\tilde f$. Thus, the claim follows from Proposition 5.4 for $\tilde f$, and thus also for f because $f=\tilde f$ μ-a.e. □

5.2. **Characteristic functions of "nice" planar sets.** In this subsection we estimate the quantization cost for f being a characteristic function of some sufficiently nice planar set K, i.e. $f = 1_K : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$. Without loss of generality we suppose $K\subset[0,1]^2$. Let μ be the standard Lebesgue measure $\mu=\mathcal{L}^2\llcorner[0,1]^2$ and wlog $c(1,0)=c(0,1)=1$.

Theorem 5.6. *Let f be a characteristic function $f(x,y)=1_K(x,y)$ for an open $K\subset[0,1]^2$, standard Lebesgue measure $\mu=\mathcal{L}^2\llcorner[0,1]^2$ and cost $c(1,0)=c(0,1)=1$. Then*

(i) if K has a piecewise smooth topological boundary, one has

$$\mathcal{C}_f(n_1,n_2)\le\frac{\sqrt{2}P(K)(1+o(1))}{\min(n_1,n_2)},\qquad \textit{as } n_1,n_2\to\infty,$$

the upper bound being achieved by uniform quantization.

(ii) if, moreover, K is convex different from a rectangle, one has

$$\mathcal{C}_f(n_1,n_2)\ge\frac{c(1+o(1))}{\min(n_1,n_2)},\qquad \textit{as } n_1,n_2\to\infty, \textit{where } c \textit{ depends only on } K.$$

Remark 5.7. For a fixed total number of points $N=n_1+n_2$ it is clear that

$$\frac{c_1}{N}\le C_f(N)\le\frac{c_2}{N},\quad\text{as } N\to\infty$$

for some positive constants c_1 and C_2.

Proof. Step 1. The upper bound holds for a uniform quantization, i.e.

$$q_i(x_i):=\frac{\lfloor n_ix_i\rfloor}{n_i}+\frac{1}{2n_i}.$$

This way we have a lattice with n_1n_2 small rectangles of area $n_1^{-1}n_2^{-1}$ with different quantizing points each. Clearly, only the ones that intersect ∂K add value to the error. All such rectangles belong to $(\partial K)_\varepsilon$ – the ε-neighbourhood of ∂K with $\varepsilon:=\sqrt{2}\max(n_1^{-1},n_2^{-1})$. But for a K with a piecewise smooth boundary

$$\lim_{\varepsilon\to0}\frac{1}{\varepsilon}\mathcal{L}^2((\partial K)_\varepsilon)=P(K).$$

Hence, the total area of such rectangles is bounded by

$$\mathcal{L}^2((\partial K)_\varepsilon) = \varepsilon P(K) + o(\varepsilon) = \frac{\sqrt{2}P(K)}{\min(n_1, n_2)} + o\left(\frac{1}{\min(n_1, n_2)}\right), \text{ as } \min(n_1, n_2) \to \infty.$$

Since the quantization cost is bounded by the total area of these rectangles, we get the claim (i).

Step 2. To prove the lower bound we reformulate the statement in the following way. Without loss of generality we assume that $n_1 \leq n_2$. Consider the quantizing sets of q_1 and q_2, $A_j, j = 1, \dots, n_1$ and $\tilde{B}_k, k = 1, \dots, n_2$ respectively. For each $j = 1, \dots, n_1$ we take $K_j = \{k \in 1, \dots, n_2 \colon f(q_1(A_j), q_2(\tilde{B}_k)) = 1\}$ and construct

$$B_j := \bigcup_{k \in K_j} \tilde{B}_k.$$

In other words, $f(q_1(x), q_2(y)) = 1$, if and only if $(x, y) \in \cup_{j=1}^{n_1}(A_j \times B_j)$. Our next step is to show that one has

$$\mathcal{L}^2\left(K \triangle \bigcup_{j=1}^{n_1}(A_j \times B_j)\right) \geq \frac{c(1 + o(1))}{n_1}, \qquad \text{as } n_1 \to \infty. \tag{5.4}$$

Note that this is exactly the lower bound we want, since the symmetric difference $\mathcal{L}^2(K \triangle \bigcup_{j=1}^{n_1}(A_j \times B_j))$ is the set where $f(x, y) \neq f(q_1(x), q_2(y))$, thus it contributes its measure to the total error.

Consider a smooth part of the ∂K where all the outward normal vectors have nonzero coordinates. Denote its natural parametrization as $\theta(t)$. Denote lengths of its x and y projections as $\tilde{P}_x$ and $\tilde{P}_y$. By choosing the directions of coordinate axes appropriately, we may assume that all the coordinates of the considered normal vectors are strictly positive, i.e. they look in the north-east direction. For some constant C that we specify later, consider a polygonal line of $k = Cn_1$ segments that are tangent to the chosen part of ∂K in its points of differentiability and have x-projections of the same length. Construct k right triangles with their vertices at the right angle inside K by using segments of this polygonal line as hypothenuses. Enumerate all the triangles such that their y-coordinate is increasing and x-coordinate is decreasing. Let X_i, Y_i be the projections of cathetes of the i-th triangle on x and y axes. Define

$$P_x = \sum_{i=1}^{k} |X_i| = k|X_1|, \qquad P_y = \sum_{i=1}^{k} |Y_i|.$$

Clearly, $P_x = (1 + o(1))\tilde{P}_x$, and $P_y = (1 + o(1))\tilde{P}_y$ as $n_1 \to \infty$. Denote $(\nu_i, \sqrt{1 - \nu_i^2})$ the unit outward normal vector to ∂K in the tangency point of ∂K and the hypothenuses of the i-th triangle. Then

$$|Y_i| = \frac{|X_i|\nu_i}{\sqrt{1 - \nu_i^2}},$$

consequently,

$$P_y = \sum_{i=1}^{k} |Y_i| = |X_1| \sum_{i=1}^{k} \frac{\nu_i}{\sqrt{1 - \nu_i^2}}.$$

Denote

$$\bar\rho_1 := \left(\frac{\sqrt{1-\nu_1^2}}{\nu_1 k}\sum_{i=1}^k \frac{\nu_i}{\sqrt{1-\nu_i^2}}\right)^{-1}, \qquad \bar\rho_2 = \frac{\sqrt{1-\nu_k^2}}{\nu_k k}\sum_{i=1}^k \frac{\nu_i}{\sqrt{1-\nu_i^2}}.$$

Note that

$$\frac{1}{k}\sum_{i=1}^k \frac{\nu_i}{\sqrt{1-\nu_i^2}} = \frac{1}{k|X|_1}\sum_{i=1}^k \nu_i\sqrt{|X_i|^2+|Y_i|^2},$$

and thus for ℓ denoting the length of θ one has

$$\bar\rho_1 \to \rho_1 := \left(\frac{1}{\tilde P_x}\frac{\dot\theta(0)_x}{\dot\theta(0)_y}\int_\theta \dot\theta_y\right)^{-1}, \qquad \bar\rho_2 \to \rho_2 := \frac{1}{\tilde P_x}\frac{\dot\theta(\ell)_x}{\dot\theta(\ell)_y}\int_\theta \dot\theta_y, \qquad \text{as } n_1\to\infty.$$

From definition of ρ_1 and ρ_2 we have

$$\bar\rho_1^{-1}\max_i |Y_i| \le \frac{P_y}{k} \le \bar\rho_2 \min_i |Y_i|,$$

hence

$$(1+o(1))\rho_1^{-1}\max_i |Y_i| \le \frac{P_y}{k} \le (1+o(1))\rho_2\min_i |Y_i|, \qquad \text{as } n_1\to\infty. \tag{5.5}$$

Now we can clarify the choice of C, namely we set $C := 4\rho_1$, i.e. $k = 4\rho_1 n_1$.

In what follows we prove that the inequality (5.4) holds with $c := \frac{\tilde P_x \tilde P_y}{16\rho_1(2\rho_1\rho_2+1)}$. In order to prove this, we will show that the following claim.

Claim 5.8. *For $c := \frac{\tilde P_x \tilde P_y}{16\rho_1(2\rho_1\rho_2+1)}$ the set $\cup_{j=1}^{n_1}(A_j\times B_j)$ either does not cover area of at least $(1+o(1))cn_1^{-1}$ inside considered triangles, or covers at least $(1+o(1))cn_1^{-1}$ outside of K, as $n_1\to\infty$.*

The inequality (5.4) follows from Claim 5.8 because the area of triangles outside of K is asymptotically smaller than total area of triangles, i.e. it is $o(\sum_{i=1}^k |X_i||Y_i|) = o(|X_1|\sum_{i=1}^k |Y_i|) = o(P_xP_y/k) = o(n_1^{-1})$, which is asymptotically negligible for (5.4). Thus Claim 5.8 concludes the proof.

Step 3. It remains to prove Claim 5.8. To this aim, denote $a_i^j := |A_j\cap X_i|/|X_i|$ and $b_i^j := |B_j\cap Y_i|/|Y_i|$. Clearly $a_i^j, b_i^j \in [0,1]$. We now make the following estimates.

(i) The area that $A_j\times B_j$ covers inside of the union of triangles is not greater than

$$\sum_{i=1}^k a_i^j b_i^j |X_i||Y_i| \le (1+o(1))k^{-2}\rho_1 P_xP_y\sum_{i=1}^k a_i^j b_i^j. \tag{5.6}$$

This is because $A_j\times B_j$ covers at most $(A_j\cap X_i)\times(B_j\cap Y_i)$ inside of the i-th triangle. Thus, it covers area of at most $a_i^j b_i^j|X_i||Y_i|$ inside i-th triangle. Now we sum up over all triangles. The estimate on the r.h.s. follows from the equality $|X_i| = P_x/k$ and the inequality (5.5).

(ii) The area that $A_j\times B_j$ covers outside of K is not smaller than

$$\sum_{i=1}^{k-1} a_i^j|X_i|(b_{i+1}^j|Y_{i+1}|+\ldots+b_k^j|Y_k|) \ge (1+o(1))k^{-2}\rho_2^{-1}P_xP_y\sum_{i=1}^{k-1} a_i^j(b_{i+1}^j+\ldots+b_k^j). \tag{5.7}$$

This is because the set $\cup_{h\geq 1}((A_j \cap X_i) \times (B_j \cap Y_{i+h}))$ lies outside of K (so does the union of rectangles $\cup_{h\geq 1} X_i \times Y_{i+h}$ due to the fact that considered curve θ is a graph of a monotone function $x_2 = x_2(x_1)$) and its area is the l.h.s.. The estimate on the r.h.s. follows from the equality $|X_i| = P_x/k$ and the inequality (5.5).

By Lemma A.1 one has

$$\sum_{i=1}^{k-1} a_i^j (b_{i+1}^j + \ldots + b_k^j) \geq \frac{1}{2} \sum_{i=1}^{k} a_i^j b_i^j - \frac{1}{2}. \tag{5.8}$$

The whole area of all the triangles is $\sum_{i=1}^k |X_i||Y_i|/2 = P_x P_y/(2k)$ since all the $|X_i|$ are equal. Let

$$\lambda := \frac{4\rho_1\rho_2 + 1}{4\rho_1\rho_2 + 2}.$$

If at least $(1-\lambda)$-portion of the total area of triangles is not covered by $\cup_{j=1}^{n_1}(A_j \times B_j)$, Claim 5.8 immediately follows since

$$(1-\lambda)P_x P_y/(2k) = \frac{P_x P_y}{(8\rho_1\rho_2 + 4)k} = \frac{P_x P_y}{16\rho_1(2\rho_1\rho_2+1)n_1} = \frac{(1+o(1))c}{n_1}.$$

Therefore, it remains to consider the case when at least λ portion of the total area of triangles is covered by $\cup_{j=1}^{n_1}(A_j \times B_j)$, that is the covered area is at least $\lambda P_x P_y/(2k)$. From claim (i) above and (5.6) we get

$$k^{-2}\rho_1 P_x P_y \sum_{j=1}^{n_1}\sum_{i=1}^{k} a_i^j b_i^j \geq (1+o(1))\lambda P_x P_y/(2k). \tag{5.9}$$

Thus, one has

$$\begin{aligned}
&\frac{P_x P_y}{k^2\rho_2} \sum_{j=1}^{n_1}\sum_{i=1}^{k-1} a_i^j (b_{i+1}^j + \ldots + b_k^j) \\
&\qquad \geq \frac{P_x P_y}{2k^2\rho_2} \sum_{j=1}^{n_1}\sum_{i=1}^{k} a_i^j b_i^j - \frac{n_1 P_x P_y}{2k^2\rho_2} \quad \text{by (5.8)} \\
&\qquad \geq (1+o(1))\frac{\lambda P_x P_y}{4\rho_1\rho_2 k} - \frac{n_1 P_x P_y}{2k^2\rho_2} \quad \text{by (5.9)} \\
&\qquad = \frac{(1+o(1))P_x P_y}{16\rho_1(2\rho_1\rho_2+1)n_1} \quad \text{by definitions of } \lambda \text{ and } k \\
&\qquad = \frac{(1+o(1))c}{n_1}.
\end{aligned} \tag{5.10}$$

But claim (ii) and (5.7) implies that $\cup_{j=1}^{n_!}(A_j \times B_j)$ covers outside of K the area at least

$$(1+o(1))\frac{P_x P_y}{k^2\rho_2} \sum_{j=1}^{n_1}\sum_{i=1}^{k-1} a_i^j (b_{i+1}^j + \ldots + b_k^j),$$

hence, by (5.10), at least $(1+o(1))c/n_1$, which concludes the proof of Claim 5.8. □

The careful inspection of Step 2 and Step 3 of the proof of the above Theorem 5.6 provides the following curious corollary for the case when $K \subset \mathbb{R}^2$ is a right-angled triangle with catheti parallel to the coordinate axes.

Corollary 5.9. *For a characteristic function of a right-angled triangle with sides P_x, P_y the quantizing error is bounded from below*

$$C_f(n_1, n_2) \geq \frac{(1+o(1))P_xP_y}{48\min(n_1, n_2)}, \qquad \text{as } \min(n_1, n_2) \to \infty.$$

Proof. In terms of the above proof of Theorem 5.6 one can explicitly calculate $\rho_1 = \rho_2 = 1$, and, therefore, $c = (16(2\rho_1\rho_2 + 1))^{-1} = 1/48$. □

5.3. **Linear functions.** For the case when f is a linear function we are able to calculate exactly the quantization cost for a fairly large class of cost functions c.

Theorem 5.10. *Let $f(x) := \sum_{i=1}^d w_ix_i$ with $w_i \neq 0$ for all $i = 1, \dots, d$, and $c(u, v) := p(|u - v|)$, where $t \mapsto p(t)$ is convex and strictly increasing for $t \geq 0$, while $\mu := \mathcal{L}^d \llcorner [0, 1]^d$. Then*

$$C_f(n) = \left| \frac{1}{\prod_{i=1}^d w_i} \int_{-w_1/2}^{w_1/2} \dots \int_{-w_d/2}^{w_d/2} p\left(\left|\sum_{i=1}^d x_i/n_i\right|\right) dx_d \dots dx_1 \right|.$$

Moreover, the best quantization functions are uniform, i.e. for $x \in [0, 1]^d$ take

$$q_i(x_i) = \frac{\lfloor n_ix_i \rfloor}{n_i} + \frac{1}{2n_i}.$$

Proof. The absolute value in the formula for C_f is to cover the case of negative coefficients, but in the proof it is convenient to consider all $w_i > 0, i = 1, \dots, d$. To see that this restriction does not lose generality, note that linearity of f allows us to shift the defining measure $\mathcal{L}^d \llcorner [0, 1]^d$ to $\mathcal{L}^d \llcorner [-1/2, 1/2]^d$. This translation changes f up to a constant, but an additive constant gets canceled in $f(x) - f(q(x))$. Now, when we work in a symmetrical region, for a negative w_i one can change $x_i \to -x_i$ and $w_i \to -w_i$. The function f and the measure μ do not change, i.e. the error remains the same. Therefore, we work with the case all $w_i > 0, i = 1, \dots, d$.

Let $\tilde{A}_i^{s_i}, s_i = 1, \dots, n_i$ denote the level sets of $q_i, i = 1, \dots, d$ with $\tilde{a}_{s_i}^i := q_i(\tilde{A}_i^{s_i})$. Denote for brevity $s = (s_1, \dots, s_d)$, $c_s := f(q_1(\tilde{a}_1^{s_1}), \dots, q_d(\tilde{a}_d^{s_d}))$. Then

$$\begin{aligned} C_f(n_1, \dots, n_d) &= \sum_{s_1,\dots,s_d} \int_{\tilde{A}_1^{s_1} \times \dots \times \tilde{A}_d^{s_d}} p\left(\left|\sum_{i=1}^d w_i\tilde{x}_i - c_s\right|\right) d\tilde{x} \\ &= \sum_{s_1,\dots,s_d} \frac{1}{\prod_{i=1}^d w_i} \int_{A_1^{s_1} \times \dots \times A_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c_s\right|\right) dx, \end{aligned} \tag{5.11}$$

where $A_i^{s_i} := w_i\tilde{A}_i^{s_i}$. Note that $\cup_{s_i=1}^{n_i} A_i^{s_i} = [0, w_i]$. Let us write a single error term in the above sum in the following way

$$\int_{A_1^{s_1} \times \dots \times A_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c_s\right|\right) dx = \int_{A_1^{s_1}} G(x_1)\, dx_1,$$

where

$$G(x_1) := \int_{A_2^{s_2} \times \dots \times A_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c_s\right|\right) dx_d \dots dx_2.$$

We consider G to be defined on the whole real line. Note, that all the functions $x_1 \mapsto p(|\sum_{i=1}^d x_i - c_s|)$ are convex, implying that the function G is also convex.

In addition, G extended to the whole real line is not monotone, since assuming an extension $p(+\infty) = +\infty$ we get $G(-\infty) = +\infty$ and $G(+\infty) = +\infty$. Therefore, for the extreme point α of G the function G decreases up to α and increases after.

Now, consider the following transformation of $A_1^{s_1}$ into an interval of the same measure. Denote $a_1^{s_1} := \mathcal{L}^1(A_1^{s_1})/2$. Take $t \in \mathbb{R}$ such that $\alpha - t = \mathcal{L}^1(A_1^{s_1} \cap (-\infty, \alpha))$. We will prove that

$$\int_{A_1^{s_1}} G(x_1)\, dx_1 \geq \int_t^{t+2a_1^{s_1}} G(x_1)\, dx_1. \tag{5.12}$$

To this aim we rewrite (5.12) as

$$\int_0^\infty \mathcal{L}^1(\{x_1 \in A_1^{s_1} \colon G(x_1) > r\})\, dr \geq \int_0^\infty \mathcal{L}^1(\{x_1 \in [t, t+2a_1^{s_1}] \colon G(x_1) > r\})\, dr. \tag{5.13}$$

To prove (5.13) it suffices to show that for all $r \geq 0$ one has

$$\mathcal{L}^1(\{x_1 \in A_1^{s_1} \colon G(x_1) > r\}) \geq \mathcal{L}^1(\{x_1 \in [t, t+2a_1^{s_1}] \colon G(x_1) > r\}).$$

Since $\mathcal{L}^1(A_1^{s_1}) = 2a_1^{s_1} = \mathcal{L}^1([t, t+2a_1^{s_1}])$ it is enough to prove the opposite, i.e. that

$$\mathcal{L}^1(\{x_1 \in A_1^{s_1} \colon G(x_1) \leq r\}) \leq \mathcal{L}^1(\{x_1 \in [t, t+2a_1^{s_1}] \colon G(x_1) \leq r\}). \tag{5.14}$$

Clearly, it is enough to consider $r \geq G(\alpha)$. Then the condition $G(x_1) \leq r$ can be reformulated as $x_1 \in [u, v]$ with $u \leq \alpha \leq v$, because G is convex (the endpoints of the interval might not be included, but it does not affect the measure anyway). Now (5.14) would follow once one shows that for any $u \leq \alpha \leq v$ one has

$$\begin{aligned} &\mathcal{L}^1(A_1^{s_1} \cap [u, \alpha]) \leq \mathcal{L}^1([\max(t, u), \alpha]) = \min(\alpha - t, \alpha - u), \\ &\mathcal{L}^1(A_1^{s_1} \cap [\alpha, v]) \leq \mathcal{L}^1([\alpha, \min(t+2a_1^{s_1}, v)]) = \min(t + 2a_1^{s_1} - \alpha, v - \alpha). \end{aligned} \tag{5.15}$$

By definition $\mathcal{L}^1(A_1^{s_1} \cap [-\infty, \alpha]) = \alpha - t$, which proves the first inequality. The second one follows from $\mathcal{L}^1(A_1^{s_1} \cap [\alpha, +\infty)) = t + 2a_1^{s_1} - \alpha$. This finishes the proof of (5.14) hence (5.13) hence (5.12).

After that, similarly, one by one we transform all the other sets $A_i^{s_i}$ into intervals in a way that decreases the error term. As a result, we get that for some $t_i \in \mathbb{R}$ one has

$$\int_{A_1^{s_1} \times \ldots \times A_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c_s\right|\right) dx \geq \int_{t_1}^{t_1+2a_1^{s_1}} \ldots \int_{t_d}^{t_d+2a_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c_s\right|\right) dx.$$

Performing a linear change of variables, we write the latter integral as

$$\int_{-a_1^{s_1}}^{a_1^{s_1}} \ldots \int_{-a_d^{s_d}}^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c\right|\right) dx, \tag{5.16}$$

with a new constant $c := c_s - \sum_{i=1}^d (t_i + a_i^{s_i})$. In order to get rid of c we use the following simple lemma.

Lemma 5.11. *Let Z be a centrally symmetric real random variable and $t \mapsto p(|t|)$ be a convex function with minimum at zero. Then*

$$\min_{c \in \mathbb{R}} \mathbb{E}\, p(|Z - c|) = \mathbb{E}\, p(|Z|).$$

Proof. The function $c \mapsto \mathbb{E}\, p(|Z-c|)$ is convex, because for a fixed z the function $c \mapsto p(|z-c|)$ is convex. Moreover it is centrally symmetric, because so is Z, i.e.

$$\mathbb{E}\, p(|Z-c|) = \mathbb{E}\, p(|-Z-c|) = \mathbb{E}\, p(|Z+c|).$$

Clearly, any centrally symmetric convex function has its minimum at zero. □

The distribution of $Z_1 + \ldots + Z_d$ for a vector $(Z_1, \ldots, Z_d)$ uniformly distributed on $[-a_1^{s_1}, a_1^{s_1}] \times \ldots \times [-a_d^{s_d}, a_d^{s_d}]$ is symmetric with respect to zero. Therefore, by Lemma 5.11 the integral (5.16) is minimal when c is zero. Note that $c = 0$ gives $c_s = \sum_{i=1}^d (t_i + a_i^{s_i})$. Putting all together, we obtain the inequality

$$\int_{A_1^{s_1} \times \ldots \times A_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c_s\right|\right) dx \geq \int_{-a_1^{s_1}}^{a_1^{s_1}} \cdots \int_{-a_d^{s_d}}^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i\right|\right) dx.$$

Then, using this estimate for all the terms in the initial formula (5.11) for a quantization error, we get following lower bound

$$C_f(n_1, \ldots, n_d) \geq \frac{1}{\prod_{i=1}^d w_i} \sum_{s_1, \ldots, s_d} \int_{-a_1^{s_1}}^{a_1^{s_1}} \cdots \int_{-a_d^{s_d}}^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i\right|\right) dx,$$

where for all $i = 1, \ldots, d$ one has $\sum_{s_i=1}^{n_i} a_i^{s_i} = w_i/2$, since $A_i^{s_i}, s_i = 1, \ldots, n_i$ cover $[0, w_i]$ and this sum is half the measure of their union. Now, to finish the proof, we have to find the minimum of the right hand side with respect to all $a_i^{s_i}$. This is provided by Lemma A.4, which implies that

$$C_f(n_1, \ldots, n_d) \geq \frac{\prod_{i=1}^d n_i}{\prod_{i=1}^d w_i} \int_{-\frac{w_1}{2n_1}}^{\frac{w_1}{2n_1}} \cdots \int_{-\frac{w_d}{2n_d}}^{\frac{w_d}{2n_d}} p\left(\left|\sum_{i=1}^d x_i\right|\right) dx.$$

The latter becomes the claimed lower bound after a linear change of variables $y_i := n_i x_i$.

To prove the second part of the statement, it remains to verify that this error is achieved for a uniform quantization, i.e. for

$$q_i(x_i) := \frac{\lfloor n_i x_i \rfloor}{n_i} + \frac{1}{2n_i}.$$

Note that linearity of the function implies that the error is the same on all the rectangles of the form $\prod_i [\frac{k_i}{n_i}, \frac{k_i+1}{n_i}]$ where $k_i = 0, \ldots, n_i - 1$. Therefore, it is sufficient to check that for one rectangle $\prod_i [0, \frac{1}{n_i}]$ the error term is equal to

$$\left| \frac{1}{\prod_{i=1}^d n_i w_i} \int_{-w_1/2}^{w_1/2} \cdots \int_{-w_d/2}^{w_d/2} p\left(\left|\sum_{i=1}^d x_i/n_i\right|\right) dx_d \ldots dx_1 \right|.$$

At the same time, by definition this term is

$$\int_0^{\frac{1}{n_1}} \cdots \int_0^{\frac{1}{n_d}} p\left(\left|\sum_{i=1}^d w_i x_i - \sum_{i=1}^d \frac{w_i}{2n_i}\right|\right) dx.$$

A linear change of variables $y_i := w_i(n_i x_i - 1/2)$ comcludes the proof. □

One might wonder what is the best quantizing error when the total number of points in the grid $n_1 n_2 \ldots n_d$ is fixed. The next remark answers this question, its proof is postponed to the Appendix A.

A standard example of a cost function is the power of the euclidean distance. In this case, the error can be calculated explicitly.

Remark 5.12. For a linear function $f(x) = \sum_{i=1}^d w_i x_i$, cost $c(u,v) = |u-v|^\gamma, \gamma \geq 1$ and Lebesgue measure $\mu(x) = \mathcal{L}^d \llcorner [0,1]^d$ Theorem 5.10 gives the exact error

$$C_f = \frac{\prod_{i=1}^d n_i w_i^{-1}}{2^{\gamma+d}\gamma(\gamma+1)\dots(\gamma+d-1)} \sum_{\varepsilon_1=\pm1} \dots \sum_{\varepsilon_d=\pm1} \prod_{i=1}^d \varepsilon_i \left| \sum_{i=1}^d \frac{\varepsilon_i w_i}{n_i} \right|^{\gamma+d}.$$

Remark 5.13. Under conditions of Remark 5.12, when $N = n_1 + n_2 + \ldots + n_d$ is fixed, one can show that the best possible quantizing error has the following order

$$\min_{n_1,\dots,n_d : \sum_i n_i = N} C_f \sim C/N^\gamma,$$

with $C = C(w_1, \dots, w_d) > 0$.

5.4. **Lower bounds for monotone functions.** The approach we used for a linear function works in a slightly more general case, but gives only a lower bound.

Theorem 5.14. *Let $f(x_1, \dots, x_d)$ be monotone in each coordinate and satisfy $|f(x_1, \dots, x_i + \Delta_i, \dots, x_d) - f(x_1, \dots, x_d)| \geq w_i \Delta_i$ for all $\Delta_i > 0$, $i = 1, \dots, d$ and some fixed positive w_i. In addition, $c(u,v) = p(|u-v|)$ for an increasing function $t \mapsto p(t), t \geq 0$ and $\mu = \mathcal{L}^d \llcorner [0,1]^d$. Then*

$$\mathcal{C}_f(n_1, \dots, n_d) \geq \frac{1}{\prod_{i=1}^d w_i} \int_0^{\frac{w_1}{2}} \dots \int_0^{\frac{w_d}{2}} p\left(\left| \sum_{i=1}^d x_i/n_i \right| \right) dx.$$

Proof. First of all, f is not required to be increasing in each coordinate, similarly to the linear case, where negativity of coefficients does not affect the result. To see this, one can use translation to work with $\mathcal{L}^d \llcorner [-1/2, 1/2]^d$ instead of $\mathcal{L}^d \llcorner [0,1]^d$ and then change sign of all coordinates along which f is decreasing, obtaining a new function that is increasing in each coordinate.

Let $A_i^{s_i}, s_i = 1 \dots, n_i$ denote the level sets of $q_i, i = 1, \dots, d$. Denote an output on one quantizing value as $c_s := f(q_1(A_1^{s_1}), \dots, q_d(A_d^{s_d}))$. Then

$$C_f(n_1, \dots, n_d) = \sum_{s_1,\dots,s_d} \int_{A_1^{s_1} \times \ldots \times A_d^{s_d}} p(|f(x) - c_{s_1,\dots,s_d}|)\, dx.$$

Denote $A^s := A_1^{s_1} \times \ldots \times A_d^{s_d}$ for brevity. Let us estimate one term of the sum as follows. Denote centers of mass of $A_i^{s_i}$ as α_i respectively. Consider the case $f(\alpha_1, \dots, \alpha_d) > c_s$, the opposite one is completely analogous. Since f is increasing in each coordinate, one has $f(x_1, \dots, x_d) > f(\alpha_1, \dots, \alpha_d) > c_s$ when all $x_i > \alpha_i$ (for the opposite case take all $x_i < \alpha_i$). Then, from monotonicity of $p(\cdot)$ we obtain

$$\int_{A^s} p(|f(x) - c_s|)\, dx \geq \int_{\alpha_1}^\infty \dots \int_{\alpha_d}^\infty \mathbf{1}_{A^s}(x) p(|f(x) - f(\alpha_1, \dots, \alpha_d)|)\, dx$$

From the assumptions on f the r.h.s is not less than

$$\int_{\alpha_1}^\infty \dots \int_{\alpha_d}^\infty \mathbf{1}_{A^s}(x) p\left(\left| \sum_{i=1}^d w_i(x_i - \alpha_i) \right| \right) dx.$$

For $a_i^{s_i} := |A_i^{s_i}|/2$, since α_i is a center of mass of $A_i^{s_i}$, this integral is not less than

$$\int_{\alpha_1}^{\alpha_1+a_1^{s_1}} \dots \int_{\alpha_d}^{\alpha_d+a_d^{s_d}} p\left(\left|\sum_{i=1}^d w_i(x_i-\alpha_i)\right|\right) dx = \int_0^{a_1^{s_1}} \dots \int_0^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d w_i x_i\right|\right) dx.$$

By definition, $A_i^{s_i}, s_i = 1, \dots, n_i$ cover $[0,1]$, thus $\sum_{s_i=1}^{n_i} a_i^{s_i} = 1/2$. Combining this for all terms in C_f we get a lower bound

$$\mathcal{C}_f(n_1,\dots,n_d) \geq \min_{a_i^{s_i}:\sum_{s_i=1}^{n_i} a_i^{s_i}=1/2} \sum_{s_1,\dots,s_d} \int_0^{a_1^{s_1}} \dots \int_0^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d w_i x_i\right|\right) dx.$$

It remains to show the the right hand side attains its minimum for $a_i^{s_i} = \frac{1}{2n_i}$. The proof of this bound is based on the same idea, as the proof of Lemma A.4, i.e. uses the Lagrange condition, but it is easier because all the variables are positive now. It remains to prove that

$$\sum_{s_1,\dots,s_d} \int_0^{a_1^{s_1}} \dots \int_0^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d w_i x_i\right|\right) dx \geq \prod_{i=1}^d n_i \int_0^{\frac{1}{2n_1}} \dots \int_0^{\frac{1}{2n_d}} p\left(\left|\sum_{i=1}^d w_i x_i\right|\right) dx,$$

because after a linear change of variables $y_i = w_i n_i x_i$ the latter integral becomes exactly what we need, namely

$$\frac{1}{\prod_{i=1}^d w_i} \int_0^{\frac{w_1}{2}} \dots \int_0^{\frac{w_d}{2}} p\left(\left|\sum_{i=1}^d y_i/n_i\right|\right) dy.$$

Clearly, the r.h.s. is decreasing in n_i. Now, we use a standard argument. Take $n_1, \dots, n_d$ with the smallest sum, such that for them there is a point contradicting the inequality. Since the condition $\sum_{s_i=1}^{n_i} a_i^{s_i} = 1/2, a_i^{s_i} \geq 0$ describes a compact and the difference between l.h.s. and r.h.s. is continuous w.r.t. $a_i^{s_i}$, this difference attains its minimum at some point, clearly that minimum being less than zero. At this point all $a_i^{s_i}$ are strictly positive, otherwise one could get rid of zero values, as this would only increase right hand side due to its monotonicity in n_i, but would not change the left hand side. Then we would obtain a contradictory configuration with smaller sum of n_i. When all the variables are strictly positive, one can apply Lagrange conditions and get that for any fixed $i = 1, \dots, d$ all the partial derivatives with respect to $a_i^{s_i}, s_i = 1, \dots, n_i$ are the same. The derivative with respect to $a_1^{s_1}$ is

$$\sum_{s_2=1}^{n_2} \dots \sum_{s_d=1}^{n_d} \int_0^{a_2^{s_2}} \dots \int_0^{a_d^{s_d}} p\left(\left|w_1 a_1^{s_1} + \sum_{i=2}^d w_i x_i\right|\right) dx_d \dots dx_2.$$

It is monotone in $a_1^{s_1}$, i.e. Lagrange condition implies $a_1^1 = \dots = a_1^{n_1}$. Similarly, we get $a_i^1 = \dots = a_i^{n_i}$ for all $i = 1, \dots, d$. Note that this is exactly the point of equality.

□

Remark 5.15. Using this lower bound for a linear function f we would get a result worse than the exact error in Theorem 5.10, but it loses only by a factor not greater than 2^d. On the other hand, the restrictions in Theorem 5.10 are stronger, because the function $t \mapsto p(|t|)$ is convex and f is linear.

The following easy statement is also worth mentioning.

Proposition 5.16. *For any function f and nonnegative cost c and two measures $\mu \le \nu$, in the sense that for any Borel set B one has $\mu(B) \le \nu(B)$, it is true that*

$$C_{f,c,\mu}(n_1,\dots,n_d) \le C_{f,c,\nu}(n_1,\dots,n_d).$$

Proof. For any quantization functions q_1, q_2 one has

$$\begin{aligned} L_{f,c,\mu}(q_1,\dots,q_d) &= \int c(f(x), f(q_1(x_1),\dots,q_d(x_d)))\, d\mu(x) \\ &\le \int c(f(x), f(q_1(x_1),\dots,q_d(x_d)))\, d\nu(x) = L_{f,c,\nu}(q_1,\dots,q_d). \end{aligned}$$

By passing to the infimum over all $q_1,\dots,q_d$ we finish the proof. □

This immediately implies the following corollary,

Corollary 5.17. *Let f and c be as in Theorem 5.10. If for some rectangle $R = [a_1, a_1 + r_1] \times \ldots \times [a_d, a_d + r_d]$ one has the inequality $\mu \le C\mathbf{1}_R\,\mathcal{L}^d$, it is true that*

$$\mathcal{C}_{f,c,\mu} \le \left| \frac{C}{\prod_i w_i r_i} \int_{-w_1 r_1/2}^{w_1 r_1/2} \cdots \int_{-w_d r_d/2}^{w_d r_d/2} p\left(\left| \sum_i x_i/n_i \right| \right) dx \right|.$$

If for some rectangle $R' = [a_1, a_1 + r_1'] \times \ldots \times [a_d, a_d + r_d']$ one has $\mu \ge c1_{R'}\,\mathcal{L}^d$, then

$$\mathcal{C}_{f,c,\mu} \ge \left| \frac{c}{\prod_i w_i r_i'} \int_{-w_1 r_1'/2}^{w_1 r_1'/2} \cdots \int_{-w_d r_d'/2}^{w_d r_d'/2} p\left(\left| \sum_i x_i/n_i \right| \right) dx \right|$$

In particular, for a cost function $c(u,v) = |u-v|^\gamma, \gamma \ge 1$, if $N = n_1 + \ldots + n_d$ is fixed and $\mu \ll \mathcal{L}^d$ with bounded l.s.c. density and compact support, then

$$\frac{c}{N^\gamma} \le \mathcal{C}_{f,c,\mu} \le \frac{C}{N^\gamma}$$

for some $c > 0$, $C > 0$ depending on the data.

Proof. Note that due to Proposition 5.16 for the upper estimate it is enough to prove the same upper bound for the measure $C\mathcal{L}^d \llcorner R$. Since f is linear we can change the variables $y_i = (x_i - a_i)/r_i$, where $\overline{y} \in [0,1]^d$. Then $f(x) := \sum_i w_i x_i = \sum w_i r_i y_i + const = \tilde{f}(\overline{y})$ for a linear function $\tilde{f}$. The cost $c(u,v)$ is translation invariant, thus the constant in $\tilde{f}$ can be omited. Finally, the loss $\mathcal{L}_{f,\mu}(q_1,\dots,q_d)$ is clearly linear in μ, therefore we can use Theorem 5.10 to obtain claimed estimate. The lower estimate is completely analogous and the last statement follows from the Remark 5.13. □

5.5. **Quadratic cost.** For the quadratic cost $c(u,v) := |u-v|^2$ we are able to say slightly more.

Theorem 5.18. *Let $f(x) = \sum_{i=1}^d \phi_i(x_i)$, where all ϕ_i have convex image and $c(u,v) := |u-v|^2$. Let X_i be independent with $\operatorname{law}(X_i) = \nu_i$, so that the joint law is $\mu = \otimes_i \nu_i$. Then one can choose the best quantization functions $q_i(x_i)$ independently from each other, minimizing $\mathbb{E}\,|\phi_i(X_i) - \phi_i(q_i(X_i))|^2$ respectively. The error is then the sum of separate errors, i.e.*

$$C_f(n_1,\dots,n_d) = \sum_{i=1}^d C_{\phi_i,c,\nu_i}(n_i)$$

Proof. Let $A_i^{s_i}, s_i = 1, \dots, n_i$ denote the level sets of q_i respectively and $a_i^{s_i} := q_i(A_i^{s_i})$. Denote $q := (q_1, \dots, q_d)$ for brevity. Then, for $c_s := f(q_1(a_1^{s_1}), \dots, q_d(a_d^{s_d})))$, by definition one has

$$L_f(q) = \sum_{s_1,\dots,s_d} \int_{A_d^{s_d}} \dots \int_{A_1^{s_1}} \left(\sum_{i=1}^d \phi_i(x_i) - c_s \right)^2 d\nu_1(x_1) \dots d\nu_d(x_d).$$

Consider one term of this sum. Define a random vector

$$(X_1^{s_1}, \dots, X_d^{s_d}) = (X | X \in A_1^{s_1} \times \dots \times A_d^{s_d}) \sim \otimes_i \left(\mathbf{1}_{A_i^{s_i}}(x_i) \frac{\nu_i(x_i)}{\nu_i(A_i^{s_i})} \right).$$

The integral can be expressed as

$$\int_{A_1^{s_1} \times \dots \times A_d^{s_d}} \left(\sum_{i=1}^d \phi_i(x_i) - c_s \right)^2 d\mu(x) = \prod_{i=1}^d \nu_i(A_i^{s_i}) \mathbb{E} \left[\left(\sum_{i=1}^d \phi_i(X_i^{s_i}) - c_s \right)^2 \right].$$

It is well-known (one can show it by taking the derivative with respect to c), that this expectation is at minimum for

$$c_s = \mathbb{E} \left[\sum_{i=1}^d \phi_i(X_i^{s_i}) \right] = \sum_{i=1}^d \mathbb{E} \left[\phi_i(X_i^{s_i}) \right]$$

and the minimum value is exactly

$$\min_{c_s \in \mathbb{R}} \mathbb{E} \left[\left(\sum_{i=1}^d \phi_i(X_i^{s_i}) - c_s \right)^2 \right] = \operatorname{Var} \left[\sum_{i=1}^d \phi_i(X_i^{s_i}) \right] = \sum_{i=1}^d \operatorname{Var} \left[\phi_i(X_i^{s_i}) \right],$$

because the variables $X_i^{s_i}$ are independent. Consequently, we obtain a lower bound

$$L_f(q) \geq \sum_{s_1,\dots,s_d} \left(\prod_{i=1}^d \nu_i(A_i^{s_i}) \sum_{i=1}^d \operatorname{Var} \left[\phi_i(X_i^{s_i}) \right] \right) = \sum_{i=1}^d \sum_{s_i=1}^{n_i} \nu_i(A_i^{s_i}) \operatorname{Var} \phi_i(X_i^{s_i}),$$

and the equality is achieved for the right choice of c_s, namely $c_s = \sum_{i=1} \mathbb{E}\, \phi_i(X_i^{s_i})$. Recall that by definiton $c_s = \sum_{i=1} \phi_i(a_i^{s_i})$. It is possible to pick $a_i^{s_i} \in \phi_i^{-1}(\mathbb{E}\, \phi(X_i^{s_i}))$, because all ϕ_i have convex image. Therefore, for fixed level sets $A_i^{s_i}$ and the best choice of $q_i(a_i^{s_i})$ for such $A_i^{s_i}$ we get

$$L_f(q) = \sum_{i=1}^d \sum_{s_i=1}^{n_i} \nu_i(A_i^{s_i}) \operatorname{Var} \phi_i(X_i^{s_i}).$$

What is convenient here, is that different quantizers are completely separated, reducing the problem to a classical quantization.

More precisely, one term of this sum is exactly a classical quantization error for the same choice of $q_i(x_i)$

$$\sum_{s_i=1}^{n_i} \nu_i(A_i^{s_i}) \operatorname{Var} \phi_i(X_i^{s_i}) = L_{\phi_i, c, \nu_i}(q_i).$$

This follows from exactly the same argument that we used to obtain this sum in the first place. Therefore, one can pick the best quantizers minimizing their own errors. □

5.6. **Further examples of functions.** The above theorem can be combined with the following statement (of immediate proof) to provide a lot of examples for the asymptotic behaviour of costs.

Lemma 5.19. *Let $g\colon \mathbb{R} \to \mathbb{R}$ satisfy the estimate*

$$\underline{c}(x,y) \le c(g(x),g(y)) \le \bar{c}(x,y)$$

for all $x, y \in f(\operatorname{supp}\mu)$. Then

$$C_{f,\underline{c}}(n_1,n_2) \le C_{g\circ f,c}(n_1,n_2) \le C_{f,\bar{c}}(n_1,n_2).$$

Corollary 5.20. *Let $c(u,v) = p(|u-v|)$ for an increasing function $p(t), t \ge 0$ and $\mu = \mathcal{L}^d \llcorner [0,1]^d$. Let $f(x) = g(\langle w, x\rangle)$. Assuming that for some function s the function $t \mapsto (p\circ s)(t), t \ge 0$ is convex increasing and $|g(a)-g(b)| \le s(|a-b|), a, b$ in the range of $x \mapsto \langle w, x\rangle$, one has*

$$\mathcal{C}_f(n_1,\dots,n_d) \le \left| \frac{1}{\prod_i w_i} \int_{-w_1/2}^{w_1/2} \dots \int_{-w_d/2}^{w_d/2} (p\circ s)\left(\left| \sum_i x_i/n_i \right| \right) dx \right|$$

Assuming that for some convex function r it is true that $(p\circ s)(t), t \ge 0$ is convex increasing and $|g(a)-g(b)| \ge r(|a-b|), a, b$ in the range of $x \mapsto \langle w, x\rangle$, one has

$$\mathcal{C}_f(n_1,\dots,n_d) \ge \left| \frac{1}{\prod_i w_i} \int_{-w_1/2}^{w_1/2} \dots \int_{-w_d/2}^{w_d/2} (p\circ r)\left(\left| \sum_i x_i/n_i \right| \right) dx \right| .$$

Proof. Both inequalities immediately follow from Lemma 5.19 and Theorem 5.10. □

Remark 5.21. Let $f(x) = g(\langle w, x\rangle)$, where g is α-Hölder with a constant C, $c(u,v) = |u-v|^\gamma, \gamma \ge 1/\alpha$, and $\mu := \mathcal{L}^d \llcorner [0,1]^d$. Then

$$\begin{aligned}&\mathcal{C}_f(n_1,\dots,n_d)\\ &\quad\le \frac{C^\gamma \prod_i n_i w_i^{-1}}{2^{\alpha\gamma+d}\alpha\gamma(\alpha\gamma+1)\dots(\alpha\gamma+d-1)} \sum_{\varepsilon_1=\pm1} \cdots \sum_{\varepsilon_d=\pm1} \prod_{i=1}^d \varepsilon_i \left| \sum \frac{\varepsilon_i w_i}{n_i} \right|^{\alpha\gamma+d} .\end{aligned}$$

If instead $|g(a)-g(b)| \ge c|a-b|^\alpha, \{a,b\}$ in the range of $x \mapsto \langle w, x\rangle$, then

$$\begin{aligned}&\mathcal{C}_f(n_1,\dots,n_d)\\ &\quad\ge \frac{c^\gamma \prod_i n_i w_i^{-1}}{2^{\alpha\gamma+d}\alpha\gamma(\alpha\gamma+1)\dots(\alpha\gamma+d-1)} \sum_{\varepsilon_1=\pm1} \cdots \sum_{\varepsilon_d=\pm1} \prod_{i=1}^d \varepsilon_i \left| \sum \frac{\varepsilon_i w_i}{n_i} \right|^{\alpha\gamma+d} .\end{aligned}$$

Proof. If g is α-Hölder with a constant C, then $c(g(x),g(y)) = |g(x)-g(y)|^\gamma \le C^\gamma |x-y|^{\alpha\gamma}$. Therefore, by using Lemma 5.19 and Remark 5.12 we obtain the upper bound inequality. Analogously, when $|g(a)-g(b)| \ge c|a-b|^\alpha, \{a,b\}$ in the range of $x \mapsto \langle w, x\rangle$, then $c(g(x),g(y)) = |g(x)-g(y)|^\gamma \ge c^\gamma|x-y|^{\alpha\gamma}$., and hence the lower bound inequality follows again by combining Lemma 5.19 and Remark 5.12. □

Corollary 5.22. *Let $f(x) = g(\sum_i \phi_i(x_i))$, while X_i are independent with the joint law $\otimes_i \nu_i$. If and $c(g(a),g(b)) \le |a-b|^2$, then*

$$\mathcal{C}_f(n_1,\dots,n_d) \le \sum_{i=1}^d C_{2,\phi_i,\nu_i}(n_i).$$

If $c(g(a),g(b)) \geq |a-b|^2$, *then*

$$\mathcal{C}_f(n_1,\ldots,n_d) \geq \sum_{i=1}^{d} C_{2,\phi_i,\nu_i}(n_i).$$

Remark 5.23. Let $f(x) = g(\sum_i \phi_i(x_i))$ and $c(u,v) = |u-v|^\gamma$, while the joint law of X_i is $\otimes_i \nu_i$. If g is $2/\gamma$-Hölder with a constant R, then

$$\mathcal{C}_f(n_1,\ldots,n_d) \leq R \cdot \sum_{i=1}^{d} C_{2,\phi_i,\nu_i}(n_i).$$

If $|g(a)-g(b)| \geq r|a-b|^{2/\gamma}$, then

$$\mathcal{C}_f(n_1,\ldots,n_d) \geq r \cdot \sum_{i=1}^{d} C_{2,\phi_i,\nu_i}(n_i).$$

The next statement demonstrates how one can estimate the error by using general results listed here. For simplicity of calculations, consider $d=2$.

Remark 5.24. Let $f(x,y) = \phi(x) + \psi(y)$ and consider the cost function $c(u,v) = |1-u/v|^2$ which arizes frequently in engineering practice. Assume the joint law of X and Y be $\mu\otimes\nu$ supported on $[a_1,a_2]\times[b_1,b_2]$, with $a_1>0$ and $b_1>0$. Assume that $f(x,y)>\delta>0$ on a support of $\mu\otimes\nu$ (so that our cost function does not tend to infinity inside the area we are working with). Then, as $n_1,n_2\to\infty$ one has

$$\mathcal{C}_f(n_1,n_2) \leq \frac{1+o(1)}{a_1+b_1}(C_{2,\phi_\#\mu}(n_1) + C_{2,\psi_\#\nu}(n_2))$$

and for some constant c

$$\mathcal{C}_f(n_1,n_2) \geq c(C_{2,\phi_\#\mu}(n_1) + C_{2,\psi_\#\nu}(n_2)).$$

Proof. Note that as $u/v\to 1$ one has $c(u,v) = |1-u/v|^2 \sim |\ln u - \ln v|^2$. Quantizing f with a cost function $|\ln u - \ln v|^2$ is the same as quantizing $\tilde{f}(x,y) = \ln(\phi(x)+\psi(y))$ with $\tilde{c}(u,v) = |u-v|^2$ while the joint law of X and Y is $\mu\otimes\nu$. Then the previous remarks provide us with inequalities

$$\mathcal{C}_{\tilde{f}}(n_1,n_2) \leq \frac{1}{a_1+b_1}(C_{2,\phi_\#\mu}(n_1) + C_{2,\psi_\#\nu}(n_2))$$

and

$$\mathcal{C}_{\tilde{f}}(n_1,n_2) \geq \frac{1}{a_2+b_2}(C_{2,\phi_\#\mu}(n_1) + C_{2,\psi_\#\nu}(n_2)).$$

It remains to check how good the approximation $|1-u/v|^2 \sim |\ln u - \ln v|^2$ is. First of all, for an upper bound we use a uniform quantization, therefore the ratio $f(x,y)/f(q_1(x),q_2(y))$ tends to 1 uniformly over all x,y in this case. That is why the approximation is good enough for an upper bound. Now let us assume that we can achieve a better quantizing error, i.e. there is a sequence of quantizers $q_1,q_2 = q_1(n_1,n_2), q_2(n_1,n_2)$ with an error $L_f(q_1,q_2)$ better that the one we claim. Lemma 4.6 implies that the maximum measure of level sets of quantizers tends to zero, as $n_1,n_2\to\infty$. The actual lower bound can be written in the following way. We divide all the points $(x,y)\in[a_1,a_2]\times[b_1,b_2]$ into two classes S_ε and B_ε, where $S_\varepsilon = \{(x,y) : |1-f(q_1(x),q_2(y))/f(x,y)| < \varepsilon\}$ and $B_\varepsilon = [a_1,a_2]\times[b_1,b_2]\setminus S_\varepsilon$.

To calculate the error divide the integral into 2 parts integrating over S_ε and B_ε respectively. The latter integral is trivially bounded from below, thus we get

$$\mathcal{L}_f(q_1,q_2) \geq \int_{S_\varepsilon} |1 - f(q_1(x),q_2(y))/f(x,y)|^2 \mu(dx)\otimes\nu(dy) + \varepsilon^2 \mu\otimes\nu(B_\varepsilon).$$

Now since our error is asymptotically better than $C_{2,\phi_\#\mu}(n_1)+C_{2,\psi_\#\nu}(n_2)$ one can pick $\varepsilon = \varepsilon(n_1,n_2) \to 0$ so slowly, as $n_1,n_2\to\infty$, such that inevitably $\mu\otimes\nu(B_\varepsilon) = o(C_{2,\phi_\#\mu}(n_1)+C_{2,\psi_\#\nu}(n_2)$, because $\varepsilon^2\nu\otimes\mu(B_\varepsilon) = O(L_f(q_1,q_2)) = o(C_{2,\phi_\#\mu}(n_1)+C_{2,\psi_\#\nu}(n_2))$. Thus almost the whole measure is concentrated in S_ε and in S_ε one has $f(q_1(x),q_2(y))/f(x,y)$ uniformly close to 1, i.e. the cost $|1-u/v|^2 \sim |\ln u - \ln v|^2$ there. Thereby, as $n_1,n_2\to\infty$

$$\begin{aligned}
&\int_{S_\varepsilon} |1 - f(q_1(x),q_2(y))/f(x,y)|^2 \mu(dx)\otimes\nu(dy) \\
&\geq \int_{S_\varepsilon} |\ln f(q_1(x),q_2(y)) - \ln f(x,y)|^2 (1-\varepsilon)\mu(dx)\otimes\nu(dy) \\
&\sim \int_{S_\varepsilon\cup B_\varepsilon} |\ln f(q_1(x),q_2(y)) - \ln f(x,y)|^2 \mu(dx)\otimes\nu(dy).
\end{aligned}$$

The last part is due to the fact that $\varepsilon\to 0$ and that since the integrable function is uniformly bounded and the for the measure we know that $\mu\otimes\nu(B_\varepsilon) = o(C_{2,\phi_\#\mu}(n_1)+C_{2,\psi_\#\nu}(n_2))$ we obtain

$$\int_{B_\varepsilon} |\ln f(q_1(x),q_2(y)) - \ln f(x,y)|^2 \mu(dx)\otimes\nu(dy) = o(C_{2,\phi_\#\mu}(n_1)+C_{2,\psi_\#\nu}(n_2)).$$

On the other hand

$$\int_{S_\varepsilon\cup B_\varepsilon} |\ln f(q_1(x),q_2(y)) - \ln f(x,y)|^2 \mu(dx)\otimes\nu(dy) \asymp C_{2,\phi_\#\mu}(n_1)+C_{2,\psi_\#\nu}(n_2).$$

Thus the equivalence for the cost is good enough for the lower bound too, i.e. there is no asymptotically better quantization possible for $|1-u/v|^2$ rather than one considered for $|\ln u - \ln v|^2$. □

Example 5.25. Let $f(x,y) = (x+y)^2$, $c(u,v) = |u-v|^2$, while the joint law of X and Y is $\mu\times\nu$ in a rectangle $[a_1,a_2]\times[b_1,b_2]$. Then

$$\mathcal{C}_f(n_1,n_2) \leq 2(\max(|a_1|,|a_2|)+\max(|b_1|,|b_2|))(C_{2,x_\#\mu}(n_1)+C_{2,y_\#\nu}(n_2))$$

and if $a_1\geq 0, b_1\geq 0$ and they are not 0 simultaneously, one has

$$\mathcal{C}_f(n_1,n_2) \geq 2(a_1+b_1)(C_{2,x_\#\mu}(n_1)+C_{2,y_\#\nu}(n_2))$$

Proof. This example immediately follows from Remark 5.23. Here $g(t) = t^2$, i.e. $g'(t) = 2t$, thereby g is a Lipschitz function with a constant $2(\max(|a_1|,|a_2|)+\max(|b_1|,b_2))$ and the first claim is true. Additionally, if $a_1\geq 0, b_1\geq 0$ it is true that $|g(t)-g(s)|\geq 2(a_1+b_1)|t-s|$ for all $t,s\in[a_1+b_1,a_2+b_2]$ and consequently the second claim is true. □

6. General upper estimate for Sobolev functions

Assume that X_i are random vectors in $\mathcal{X}_i = \mathbb{R}^{k_i}, i=1,\dots,d$. Set $k := \sum_i k_i$.

Lemma 6.1. *Let $A_i \subset \mathcal{X}_i$ be open rectangles and $f \in C^1(\bar{A}_1 \times \ldots \times \bar{A}_d)$. Then for $\gamma \geq 1$ it is true that*

$$\int_{A_1\times\ldots\times A_d} |f(x)-f(a)|^\gamma\, dx \leq C_k \mathrm{diam}\,(A_1\times\ldots\times A_d)^\gamma \mathcal{L}^k(A_1\times\ldots\times A_d) M^*|\nabla f|^\gamma(a),$$

where M^* stands for the uncentered maximal function.

Proof. We denote for brevity $\Omega = A_1 \times \ldots \times A_d$ and $D := \mathrm{diam}\,(\Omega)$ and write

$$f(x) - f(a) = \int_0^1 \frac{d}{dt} f(tx + (1-t)a)\, dt,$$

so that

$$\begin{aligned}
\int_{\overline{\Omega}} |f(x) - f(a)|^\gamma\, dx &\leq \int_{\overline{\Omega}} dx \left| \int_0^1 \frac{d}{dt} f(tx + (1-t)a)\, dt \right|^\gamma \\
&\leq \int_{\overline{\Omega}} dx \int_0^1 \left| \frac{d}{dt} f(tx + (1-t)a) \right|^\gamma dt \\
&\leq D^\gamma \int_{\overline{\Omega}} dx \int_0^1 |\nabla f|^\gamma\, (tx + (1-t)a)\, dt \\
&= D^\gamma \int_0^1 \frac{dt}{t^d} \int_{(1-t)a+t\overline{\Omega}} |\nabla f|^\gamma\, (w)\, dw \\
&= D^\gamma \int_0^1 \frac{dt}{t^d} t^d \mathcal{L}^c(\overline{\Omega}) \fint_{(1-t)a+t\overline{\Omega}} |\nabla f|^\gamma\, (w)\, dw \\
&\leq D^\gamma \mathcal{L}^c(\overline{\Omega}) M^* |\nabla f|^\gamma(a)
\end{aligned}$$

as claimed. □

Theorem 6.2. *Let $A_i \subset \mathcal{X}_i$ be open cubes of sidelength r_i, $\Omega := A_1 \times \ldots \times A_d$, $f \in W^{1,p}(\Omega)$, $p \geq \gamma$. If $\mu \ll dx$ with density $\varphi \in L^\infty(\mathbb{R}^k)$ has compact support* $\mathrm{supp}\,\varphi \subset \Omega$, *while $c(u,v) = |u-v|^\gamma$, then*

$$C_f(n_1, \ldots, n_d) \leq C_k \|\varphi\|_\infty \|M^*|\nabla f|^\gamma\|_1 \max_i (r_i n_i^{-1/k_i})^\gamma + o\left(\max_i (r_i n_i^{-1/k_i})^\gamma\right) \tag{6.1}$$

as $n_1, \ldots, n_d \to \infty$, where M^* stands for the uncentered maximal function.

Moreover, if $p > \gamma$, then

$$C_f(n_1, \ldots, n_d) \leq C_{k,p} \|\varphi\|_\infty \|\nabla f\|_p^\gamma \max_i (r_i n_i^{-1/k_i})^\gamma + o\left(\max_i (r_i n_i^{-1/k_i})^\gamma\right). \tag{6.2}$$

Proof. We approximate $f \in W^{1,p}(\Omega)$ by $f_k \in C^1(\bar{\Omega})$ converging in Sobolev norm, and in particular with $\lim_k f_k(y) = f(y)$ and $\lim_k M^*|\nabla f_k|^\gamma(y) = M^*|\nabla f|^\gamma(y)$ for a.e. $y \in \Omega$, i.e. for all $y \in \Omega \setminus N$ with $\mathcal{L}^c(N) = 0$.

It is enough to prove the statement for $n_i^{1/k_i} \in \mathbb{Z}$, $i = 1, \ldots, d$, otherwise one could take $m_i = \lfloor n_i^{1/k_i} \rfloor^{d_i}$ with $m_i^{1/k_i} \leq n_i^{1/k_i} \leq 2m_i^{1/k_i}$. Then the inequalities for m_i combined with

$$C_f(n_1, \ldots, n_d) \leq C_f(\overline{m}) \qquad \text{and} \qquad \max_i (r_i m_i^{-1/k_i}) \leq 2 \max_i (r_i n_i^{-1/k_i})$$

would imply the estimate for any n_i with a constant multiplied by 2^γ.

Divide each A_i into n_i rectangles $A_i^1, \ldots, A_i^{n_1}$ and take $a_1^{s_1} \in A_1^{s_1}, \ldots, a_d^{s_d} \in A_d^{s_d}$, such that $(a_1^{s_1}, \ldots, a_d^{s_d}) \notin N$ for all $s_i = 1, \ldots, n_i, i = 1, \ldots, d$. Define then q_i by setting

$$q_i(x) := a_i^{s_i} \text{ whenever } x \in A_i^{s_i}.$$

Denote $A^s := A_1^{s_1} \times \ldots \times A_d^{s_d}$ and $a^s := (a_1^{s_1}, \ldots, a_d^{s_d})$. Recalling that Lemma 6.1 implies

$$\int_{A_{\overline{s}}} |f_k(x) - f_k(a^s)|^\gamma \, dx \le C_k \mathrm{diam}\,(A^s)^\gamma \mathcal{L}^k(A^s) M^* |\nabla f_k|^\gamma (a^s).$$

Summing up these inequalities, we get

(6.3)
$$\int_\Omega |f_k(x) - f_k(q_1(x_1), \ldots, q_d(x_d))|^\gamma \, dx \le C_k \max_i \left(r_i n_i^{-1/k_i} \right)^\gamma \Delta(f_k, \Omega, n_1, \ldots, n_d),$$

where $\Delta(f_k, \Omega, n_1, \ldots, n_d) := \sum_{s_1, \ldots, s_d} \mathcal{L}^k(A^s) M^* |\nabla f_k|^\gamma (a^s)$.

Passing to the limit as $k \to \infty$ in (6.3), one arrives by Fatou's lemma at

(6.4)
$$\begin{aligned}\int_\Omega |f(x) - f(q_1(x_1), \ldots, q_d(x_d))|^\gamma \, dx &\le \liminf_k \int_\Omega |f_k(x) - f_k(q_1(x_1), \ldots, q_d(x_d))|^\gamma \, dx \\ &\le C_k \max \max_i \left(r_i n_i^{-1/k_i} \right)^\gamma \Delta(f, \Omega, n_1, \ldots, n_d).\end{aligned}$$

Since $M^* |\nabla f|^\gamma$ is continuous, one has

$$\Delta(f, \Omega, n_1, \ldots, n_d) \to \int_\Omega M^* |\nabla f|^\gamma (x) \, dx$$

as $(n_1, \ldots, n_d) \to \infty$, and hence (6.4) gives

(6.5)
$$\begin{aligned} C_f(n_1, \ldots, n_d) &\le \int_\Omega |f(x) - f(q_1(x_1), \ldots, q_d(x_d))|^\gamma \, d\mu(x) \\ &\le \|\varphi\|_\infty \int_\Omega |f(x) - f(q_1(x_1), \ldots, q_d(x_d))|^\gamma \, dx \\ &\le C_k \|\varphi\|_\infty \|M^* |\nabla f|^\gamma\|_1 \max_i \left(r_i n_i^{-1/k_i} \right)^\gamma + o \left(\max_i \left(r_i n_i^{-1/k_i} \right)^\gamma \right) \end{aligned}$$

as $(n_1, \ldots, n_d) \to \infty$, which is (6.1) In particular, if $p > \gamma$, then estimating $\|M^* |\nabla f|^\gamma\|_1$ by Hardy-Littlewood theorem, we get (6.2). □

Remark 6.3. When $N = n_1 + \ldots + n_d$ is fixed, the upper estimate is minimum at

$$n_i = \frac{N r_i^{k_i}}{\sum_i r_i^{k_i}},$$

hence providing the following estimates for $C_f(N) = \min_{\sum n_i = N} C_f(n_1, \ldots, n_d)$

$$C_f(N) \le C_k \|\phi\|_\infty \|M^* |\nabla f|^\gamma\|_1 \max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{\gamma/k_i} + o \left(\max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{\gamma/k_i} \right),$$

as $N \to \infty$. Moreover, for $p > \gamma$

$$C_f(N) \le C_{k,p} \|\phi\|_\infty \|\nabla f\|_p^\gamma \max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{\gamma/k_i} + o\left(\max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{\gamma/k_i} \right).$$

Appendix A. Auxiliary statements

Here we collect some auxiliary statements used in proofs of results in the main body of the paper.

Lemma A.1. *For any $k \ge 1$, $a_i, b_i \in [0,1]$ one has*

$$\sum_{i=1}^{k-1} a_i(b_{i+1} + \ldots + b_k) \ge \frac{1}{2} \sum_{i=1}^{k} a_i b_i - \frac{1}{2}.$$

Proof. This inequality is linear in all variables, therefore it is enough to prove it for $a_i, b_i \in \{0,1\}$. If $a_i = 0$, then there is no b_i in the right hand side but there is b_i with a nonnegative coefficient in the left hand side, thus it is enough to prove the statement for $b_i = 0$. Similarly, if $b_i = 0$, it is enough to prove the statement for $a_i = 0$. Therefore, we can omit all the pairs of zeros and check the same inequality where all the variables are equal to one. It remains to note that for any k' it is true that

$$\sum_{i=1}^{k'-1} (k' - i) = \frac{k'^2 - k'}{2} \ge \frac{1}{2} k' - \frac{1}{2},$$

implying the inequality, where k' is the number of pairs such that $a_i = b_i = 1$. □

Lemma A.2. *For a convex and strictly increasing on $[0, +\infty)$ function $p(\cdot)$ and a fixed t_0 the function $t \mapsto p(|t_0 + t|) + p(|t_0 - t|)$ is*

(i) *non-decreasing on $[0, +\infty)$,*
(ii) *and, in addition, strictly increasing on $[|t_0|, +\infty)$.*

Proof. First, without loss of generality, by symmetry, we might assume $t_0 \ge 0$. We want to show that for any $a > b \ge 0$ one has

$$p(|t_0 + a|) + p(|t_0 - a|) \ge p(|t_0 + b|) + p(|t_0 - b|).$$

By convexity of $t \mapsto p(|t|)$ one has

$$\begin{aligned} \frac{a+b}{2a} p(|t_0 + a|) + \frac{a-b}{2a} p(|t_0 - a|) &\ge p\left(\left| \frac{a+b}{2a}(t_0 + a) + \frac{a-b}{2a}(t_0 - a) \right| \right) \\ &= p(|t_0 + b|), \\ \frac{a-b}{2a} p(|t_0 + a|) + \frac{a+b}{2a} p(|t_0 - a|) &\ge p\left(\left| \frac{a-b}{2a}(t_0 + a) + \frac{a+b}{2a}(t_0 - a) \right| \right) \\ &= p(|t_0 - b|). \end{aligned}$$

It remains to sum these two inequalities to get the claim (i).

For $t \ge t_0$ the function becomes $t \mapsto p(t_0 + t) + p(t - t_0)$ and thus it is strictly increasing because so is $p(\cdot)$, proving the claim (ii). □

Lemma A.3. *For a convex and strictly increasing on* $[0,+\infty)$ *function* $p(\cdot)$ *and fixed* $x_2,\dots,x_d$ *the function*

$$x_1 \mapsto (Tp)(x_1,\dots,x_d) := \sum_{\varepsilon_1=\pm1}\dots\sum_{\varepsilon_d=\pm1} p\left(\left|\sum_{i=1}^d \varepsilon_i x_i\right|\right)$$

is

(i) non-decreasing on $[0,+\infty)$

(ii) and, moreover, strictly increasing on $[|x_2|+\dots+|x_d|,+\infty)$.

Proof. By definition one has

$$(Tp)(x_1,\dots,x_d) = \sum_{\varepsilon_2=\pm1}\dots\sum_{\varepsilon_d=\pm1}\left(p\left(\left|\sum_{i=2}^d \varepsilon_i x_i + x_1\right|\right) + p\left(\left|\sum_{i=2}^d \varepsilon_i x_i - x_1\right|\right)\right).$$

Then, Lemma A.2 implies that each term of this sum is non-decreasing as a function of x_1 on $[0,+\infty)$ and strictly increasing as a function of x_1 on $[|\sum_{i=2}^d \varepsilon_i x_i|,+\infty)$. Then both claims immediately follow, since $\sum_{i=2}^d |x_i| \geq |\sum_{i=2}^d \varepsilon_i x_i|$. □

Lemma A.4. *For any* $n_1,\dots,n_d \in \mathbb{N}$ *and* $a_i^{s_i} \geq 0, s_i = 1,\dots,n_i, i=1,\dots,d$ *such that* $\sum_{s_i=1}^{n_i} a_i^{s_i} = w_i/2$ *for any* $i=1,\dots,d$, *one has*

(A.1)

$$\sum_{s_1,\dots,s_d}\int_{-a_1^{s_1}}^{a_1^{s_1}}\dots\int_{-a_d^{s_d}}^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i\right|\right)dx \geq \prod_{i=1}^d n_i \int_{-\frac{w_1}{2n_1}}^{\frac{w_1}{2n_1}}\dots\int_{-\frac{w_d}{2n_d}}^{\frac{w_d}{2n_d}} p\left(\left|\sum_{i=1}^d x_i\right|\right)dx.$$

Proof. We divide the proof in two steps.

Step 1. We first show that the right hand side of (A.1) is non-increasing with respect to n_i. Set

$$(Tp)(x_1,\dots,x_d) := \sum_{\varepsilon_1=\pm1}\sum_{\varepsilon_2=\pm1}\dots\sum_{\varepsilon_d=\pm1} p\left(\left|\sum_{i=1}^d \varepsilon_i x_i\right|\right) \tag{A.2}$$

Note, that the integral in the right-hand side of (A.1) can be rewritten in the following form

$$\int_0^{\omega_1/2}\dots\int_0^{\omega_d/2}(Tp)(x_1/n_1,\dots,x_d/n_d)\,dx.$$

The inner function is non-increasing in n_i due to Lemma A.3. Therefore, the integral is also non-increasing in n_i.

Step 2. We now prove the claim of the lemma. Assuming that there is a set of numbers $((a_i^{s_i})), s_i=1,\dots,n_i, i=1,\dots,d$ for which inequality (A.1) fails, take the one with minimal $n_1+\dots+n_d$. We will show that one can change $a_1^1,\dots,a_1^{n_1}$ to be equal and inequality (A.1) would still fail. By doing similar change for all $i=1,\dots,d$, we would then obtain that inequality (A.1) must fail when for all $i=1,\dots,d$ one has $a_i^1=\dots=a_i^{n_i}$.

To show that $a_1^1,\dots,a_{n_1}^1$ can be set equal, consider the left hand side as a function F of $(a_1^1,\dots,a_1^{n_1})$ on a compact set $\{(a_1^1,\dots,a_1^{n_1})\colon a_1^{s_1}\geq 0, G(a_1^1,\dots,a_1^{n_1})=0\}$, where $G(a_1^1,\dots,a_1^{n_1}) := \sum_{s_1=1}^{n_1} a_1^{s_1} - w_1/2$. Since F is continuous in $a_1^{s_1}, s_1 = 1,\dots,n_1$ it attains its minimum at some point $(\tilde a_1^{s_1}), s_1=1,\dots,n_1$ for which also (A.1) fails. If some of the $\tilde a_1^{s_1}$ were 0, we could remove it from the set $(\tilde a_1^{s_1}), s_1 = 1,\dots,n_1$, obtaining a set of variables not satisfying (A.1) with a smaller sum $n_1+\dots+n_d$, because the right hand side is decreasing with respect to n_1. Therefore,

$(\tilde{a}_1^{s_1}), s_1 = 1, \ldots, n_1$ belongs to a relative interior point of a compact set we are working with. Thus, method of Lagrange multipliers provides us with the following equations on $(\tilde{a}_1^{s_1})$: for some scalar λ and σ that are not 0 at the same time

$$\lambda \cdot \nabla F(\tilde{a}_1^1, \ldots, \tilde{a}_1^{n_1}) = \sigma \cdot \nabla G(\tilde{a}_1^1, \ldots, \tilde{a}_1^{n_1}) = \sigma \cdot (1, 1, \ldots, 1).$$

Note that $\lambda \neq 0$, otherwise we would get $\sigma = 0$ too. Thus, for all $s_1 = 1, \ldots, n_1$ all the derivatives

$$\frac{\partial F}{\partial a_1^{s_1}}(\tilde{a}_1^1, \ldots, \tilde{a}_1^{n_1})$$

are equal. Note that the function F can be written as

$$F(\tilde{a}_1^1, \ldots, \tilde{a}_1^{n_1}) = \sum_{s_1, \ldots, s_d} \int_0^{\tilde{a}_1^{s_1}} \int_0^{a_2^{s_2}} \ldots \int_0^{a_d^{s_d}} (Tp)(y_1, \ldots, y_d)\, dy_d \ldots dy_1.$$

Therefore,

$$\frac{\partial F}{\partial a_1^{s_1}}(\tilde{a}_1^1, \ldots, \tilde{a}_1^{n_1}) = \sum_{s_2=1}^{n_2} \ldots \sum_{s_d=1}^{n_d} \int_0^{a_2^{s_2}} \ldots \int_0^{a_d^{s_d}} (Tp)(\tilde{a}_1^{s_1}, y_2, \ldots, y_d)\, dy_d \ldots dy_2.$$

Let us show that the integral is strictly increasing as a function of $\tilde{a}_1^{s_1} > 0$. First of all, due to Lemma A.3 an integrand is non-decreasing. In addition, when $y_2 + \ldots + y_d < \tilde{a}_1^{s_1}$ the integrand is strictly increasing again by Lemma A.3. Therefore, the whole integral is also strictly increasing.

Now, equality of partial derivatives implies that for any $s_1 = 1, \ldots, n_1$ one has

$$\sum_{s_2=1}^{n_2} \ldots \sum_{s_d=1}^{n_d} \int_0^{a_2^{s_2}} \ldots \int_0^{a_d^{s_d}} (Tp)(\tilde{a}_1^{s_1}, y_2, \ldots, y_d)\, dy_d \ldots dy_2$$
$$= \sum_{s_2=1}^{n_2} \ldots \sum_{s_d=1}^{n_d} \int_0^{a_2^{s_2}} \ldots \int_0^{a_d^{s_d}} (Tp)(\tilde{a}_1^1, y_2, \ldots, y_d)\, dy_d \ldots dy_2.$$

Hence $\tilde{a}_1^1 = \tilde{a}_1^{s_1}$, i.e. $\tilde{a}_1^1 = \ldots = \tilde{a}_1^{n_1}$. Now, applying the same argument to all $(a_i^1, \ldots, a_i^{n_i}), i = 1, \ldots, d$ one by one we get that the inequality (A.1) has to be false for the point where $a_i^1 = \ldots = a_i^{n_i} = w_i/(2n_i), i = 1, \ldots, d$ (the latter equality is due to the fact that $\sum_{s_i} a_i^{s_i} = w_i/2$). But this is exactly the point where equality holds in (A.1) . □

References

[1] H. Jégou, M. Douze and C. Schmid, ”Product Quantization for Nearest Neighbor Search” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117-128, 2011, 10.1109/TPAMI.2010.57.

[2] D. Slepian and J.K. Wolf, “Noiseless coding of correlated information sources,” IEEE Trans. Inform. Theory, vol. 19, no. 4, pp. 471–480, 1973.

[3] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” IEEE Trans. Inform. Theory, vol. 22, no. 1, pp. 1–10, 1976.

[4] Zixiang Xiong, A. D. Liveris and S. Cheng, ”Distributed source coding for sensor networks,” in IEEE Signal Processing Magazine, vol. 21, no. 5, pp. 80-94, Sept. 2004, 10.1109/MSP.2004.1328091.

[5] S. Graf, H. Luschgy, ”Foundations of Quantization for Probability Distributions”, Lecture Notes in Mathematics 1730, Springer., 2007, 10.1007/BFb0103945

[6] A. Suzuki, Z. Drezner, "The p-center location", Location science, 4(1-2):69–82, 1996. 10.1016/S0966-8349(96)00012-5

[7] R. M. Gray and D. L. Neuhoff, "Quantization," in IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2325-2383, Oct. 1998, 10.1109/18.720541.

[8] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, Kurt Keutzer, "A Survey of Quantization Methods for Efficient Neural Network Inference", Book Chapter: Low-Power Computer Vision: Improving the Efficiency of Artificial Intelligence, 10.48550/arXiv.2103.13630

[9] Royden H. L. "Real Analysis", New York: Macmillan, 2nd ed., 1968

School of Cyber Science and Engineering, Southeast University, Nanjing, China
Email address: taoguo@seu.edu.cn

St.Petersburg Branch of the Steklov Mathematical Institute of the Russian Academy of Sciences, Fontanka 27, 191023 St.Petersburg, Russia
Email address: nikitus20@gmail.com

Scuola Normale Superiore, Pisa, Piazza dei Cavalieri 7, 56126 Pisa, Italy and St.Petersburg Branch of the Steklov Mathematical Institute of the Russian Academy of Sciences, Fontanka 27, 191023 St.Petersburg, Russia and HSE University, Moscow
Email address: stepanov.eugene@gmail.com