OPTIMAL FUNCTIONAL PRODUCT QUANTIZATION

TAO GUO, NIKITA KARAGODIN, AND EUGENE STEPANOV

ABSTRACT. We will discuss and compare two approaches for quantization of vectorial signals on the input to a computational device: quantizing the whole signal and optimizing the input error, or quantizing separately the components but optimizing the output error.

CONTENTS

1. INTRODUCTION

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

The research has been financed by Huawei.

with a given distribution law (i.e. the 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 below.

1.1. Functional product quantization problem.

Assume d sets $\mathcal{X}_1,\ldots,\mathcal{X}_d$, with random elements $X_i\in\mathcal{X}_i$, $Y\in\mathcal{Y}$, their common distribution law being the given Borel probability measure μ, i.e.

$$\mathrm{law}(X,Y)=\mu.$$

Finally assume that a Borel function $f\colon\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ be given. For $(n_1,n_2)\in\mathbb{N}\times\mathbb{N}$ one has to find the quantization maps

$$q_1\colon\mathcal{X}\to\mathcal{X},\#q_1(\mathcal{X})\leq n_1,\quad q_2\colon\mathcal{Y}\to\mathcal{Y},\#q_2(\mathcal{Y})\leq n_2$$

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

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

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

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

Throughout the paper we will assume, unless otherwise explicitly stated, the most common situation in applications, namely, that $\mathcal{X}=\mathbb{R}^{d_1}$, $\mathcal{Y}=\mathbb{R}^{d_2}$ be just the Euclidean spaces and $\mu\ll\mathcal{L}^{d_1}\otimes\mathcal{L}^{d_2}$ with compact support. Even more, in most cases we will limit ourselves to the case $d_1=d_2=1$, i.e. $\mathcal{X}=\mathcal{Y}=\mathbb{R}$, $\mu\ll\mathcal{L}^2$. We will see that already this case contains all the essential difficulties of the problem considered.

Literature overview neded!

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

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

where d is the given distance 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

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

with $C>0$ known.

2. Notation and preliminaries

Sometimes to emphasize the dependence of the costs on d and μ we write $L_{f,d,\mu}(q_1, q_2)$ and $C_{f,d,\mu}(n_1, n_2)$ instead of $L_f(q_1, q_2)$ and $C_f(n_1, n_2)$ respectively. Also for the classical quantization problem, to emphasize the dependence of the cost on d and μ we may write $L_{d,\mu}(q)$ and $C_{d,\mu}(N)$ instead of $L(q)$ and $C(N)$ respectively. For $d(u, v) := |u - v|^q$ we write $L_{q,\mu}(q)$, $C_{q,\mu}(N)$ 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. 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 $1 \le p < +\infty$, 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 $L^p(\mathbb{R}^d, \mu)$.

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

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

and

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

On one hand, if d 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 $\mu := \mathcal{L}^2 \llcorner [0,1] \times [0,1]$, $d(u, v) := |u - v|$ 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 our indicator function with smooth functions.

4. Random quantization and existence of optimal quantizers

4.1. **Random quantization.** In a random quantization setting we are looking for n_i quantization points $\{x_i^1,\dots,x_i^{n_i}\}$ and weight functions $c_i^1,\dots,c_i^{n_i}$ such that for all $x\in\mathbb{R}$ one has

$$0\le c_i^s(x)\le 1 \quad \text{for all } s=1,\dots,n_i, \qquad \sum_{s=1}^{n_i} c_i^s(x)=1,$$

The best random quantization by definition minimizes the error

$$\mathcal{L}_f(c_i^s,x_i^s) := \sum_{s_1=1}^{n_1}\dots\sum_{s_d=1}^{n_d}\int_{\mathcal{X}_1\times\ldots\times\mathcal{X}_d} c_1^{s_1}(x_1)\dots c_d^{s_d}(x_d)d(f(\overline{x}),f(x_1^{s_1},\dots,x_d^{s_d}))d\mu(\overline{x}).$$

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 $c_i^1(x_i),\dots,c_i^{n_i}(x_i)$ 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. $c_i^s(x_i)=\delta_{x_i^s,q_i(x_i)}$, where $\delta_{a,b}$ stands for Kronecker symbol.

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

Proposition 4.1. *Assume that f has a compact range, $d(u,v)\ge 0$ and the map $v\mapsto d(u,v)$ is lower semicontinuous for all u, while μ has compact support. Then random quantization functional $\mathcal{L}_f(c_i^s,x_i^s)$ attains its minimum for some quantizers and weights.*

Proof. For any $m\in\{1,\dots,d\}$ and any measure $\nu(x_m,\dots,x_d)$ denote its marginal distribution on $\mathcal{X}_m$ as $\nu_{X_m}(x_m)$, so that it can be disintegrated as

$$\nu(x_m,\dots,x_d)=\nu_{x_m}(x_{m+1},\dots,x_d)\otimes d\nu_{\mathcal{X}_m}(x_m),$$

for $\nu_{\mathcal{X}_m}$-a.e. $x_m\in\mathcal{X}_m$, where ν_{x_m} is the conditional probability.

Denote $x_k^s=(x_{1,k}^{s_1},\dots,x_{d,k}^{s_d})$ for simplicity. Consider $c_i^s(\cdot)$ in the positive part of a unit sphere in $L^\infty(\mu_{\mathcal{X}_i})$ with *-weak topology, value $f(x_1^{s_1},\dots,x_d^{s_d})$ in the compact range of f. Then $\mathcal{L}_f(c_i^s,x_i^s)$, being a function of $c_i^s, s\in\{1,\dots,n_i\}, i\in\{1,\dots,n_i\}, f(x_1^{s_1},\dots,x_d^{s_d}), s_i\in\{1,\dots,n_i\}\,\forall i\in\{1,\dots,d\}$ is a function in product of all described spaces, and this product is compact. Therefore, in order to shot that the functional L_f attains its minimum it is sufficient to prove that it is lower semicontinuous. Let us assume that as $k\to\infty$ one has $c_{i,k}^s\overset{*}{\rightharpoonup}c_i^s$, $f(x_{1,k}^{s_1},\dots,x_{d,k}^{s_d})\to a_{s_1,\dots,s_d}$. We prove the following inequality by induction.

For any $1\le m\le d+1$, fixed $x_1,\dots,x_{m-1}$ and measure $\nu(x_{m+1},\dots,x_d)$ one has

$$\liminf_{k\to\infty}\left(\int_{\mathcal{X}_m\times\ldots\times\mathcal{X}_d} c_{m,k}^{s_m}(x_m)\dots c_{d,k}^{s_d}(x_d)d(f(\overline{x}),f(x_k^{\overline{s}})d\nu(x_m,\dots,x_d)\right)$$
$$\ge\int_{\mathcal{X}_m\times\ldots\times\mathcal{X}_d} c_m^{s_m}(x_m)\dots c_d^{s_d}(x_d)d(f(\overline{x}),a_{\overline{s}})d\nu(x_m,\dots,x_d).$$

Induction base. Clearly, for $m=d+1$ it is simply

$$\liminf d(f(\overline{x}),f(x_k^{\overline{s}}))\ge d(f(\overline{x}),a_{\overline{s}}).$$

This inequality follows from the assumption for a function $v\to d(u,v)$ to be lower semicontinuous for any u.

Induction step. We go from $m+1$ to m. First of all, disintegrate the integral into

$$\int_{\mathcal{X}_m} c_{m,k}^{s_m}(x_m) d\nu_{\mathcal{X}_m} \left(\int_{\mathcal{X}_{m+1}\times\ldots\times\mathcal{X}_d} c_{m+1,k}^{s_{m+1}}(x_{m+1}) \ldots c_{d,k}^{s_d}(x_d) d(f(\overline{x}), f(x_k^{\overline{s}})) d\nu_{x_m} \right).$$

From induction hypothesis one has

$$\begin{aligned}\liminf_{k\to\infty} \int_{\mathcal{X}_{m+1}\times\ldots\times\mathcal{X}_d} & c_{m+1,k}^{s_{m+1}}(x_{m+1}) \ldots c_{d,k}^{s_d}(x_d) d(f(\overline{x}), f(x_k^{\overline{s}})) d\nu_{x_m} \\ & \geq \int_{\mathcal{X}_{m+1}\times\ldots\times\mathcal{X}_d} c_{m+1}^{s_{m+1}}(x_{m+1}) \ldots c_d^{s_d}(x_d) d(f(\overline{x}), a_{\overline{s}})) d\nu_{x_m}\end{aligned}$$

Now, to estimate the whole term let us use Fatou's Lemma with varying measures. It is applicable if the measure $c_{m,k}^{s_m} d\nu_{X_m}(x_m)$ converges setwise to $c_m^{s_m}(x_m) d\nu_{X_m}(x_m)$. This convergence immediately follows from *-weak convergence of $c_{m,k}^{s_m}(x) \to c_m^{s_m}(x)$. Therefore, one has

$$\begin{aligned}&\liminf_{k\to\infty} \int_{\mathcal{X}_m} c_{m,k}^{s_m}(x_m) d\nu_{\mathcal{X}_m} \left(\int_{\mathcal{X}_{m+1}\times\ldots\times\mathcal{X}_d} c_{m+1,k}^{s_{m+1}}(x_{m+1}) \ldots c_{d,k}^{s_d}(x_d) d(f(\overline{x}), f(x_k^{\overline{s}})) d\nu_{x_m} \right) \\ &\geq \int_{\mathcal{X}_m} c_{m,k}^{s_m}(x_m) d\nu_{\mathcal{X}_m} \liminf_{k\to\infty} \left(\int_{\mathcal{X}_{m+1}\times\ldots\times\mathcal{X}_d} c_{m+1,k}^{s_{m+1}}(x_{m+1}) \ldots c_{d,k}^{s_d}(x_d) d(f(\overline{x}), f(x_k^{\overline{s}})) d\nu_{x_m} \right) \\ &\geq \int_{\mathcal{X}_m} c_m^{s_m}(x_m) d\nu_{\mathcal{X}_m} \left(\int_{\mathcal{X}_{m+1}\times\ldots\times\mathcal{X}_d} c_{m+1}^{s_{m+1}}(x_{m+1}) \ldots c_d^{s_d}(x_d) d(f(\overline{x}), a^{\overline{s}}) d\nu_{x_m} \right) \\ &= \int_{\mathcal{X}_m\times\ldots\times\mathcal{X}_d} c_m^{s_m}(x_m) \ldots c_d^{s_d}(x_d) d(f(\overline{x})), a^{\overline{s}}) d\nu(x_m, \ldots, x_d)\end{aligned}$$

In result, we have shown that one term of the sum defining $\mathcal{L}_f(c_i^s, x_i^s)$ is lower semicontinous in considered space. It remains to use a standard observation that $\liminf \sum \int \ldots \geq \sum \liminf \int \ldots$ and get lower semicontinuity of the whole sum. $\square$

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.

Theorem 4.2. *Assume that f has a compact range and $d(u,v) \geq 0$ is lower semicontinuous in second coordinate, measure $\mu(x,y)$ has a compact support. Then the best quantization error $\mathcal{C}_f(\overline{n})$ is achievable as $\mathcal{L}_f(\overline{q})$ for some quantizers $q_1, \ldots, q_d$. Moreover, the best nonrandom quantization error and the best random quantization error are equal.*

Proof. Since nonrandom quantization is a particular case of random quantization, it is sufficient to prove that the optimum for a random quantization problem from 4.1 is achievable by nonrandom quantizers. Consider the optimum for a random quantization problem $c_i^{s_i}(x_i), x_i^{s_i}, s_i \in \{1, \ldots, n_i\}, i \in \{1, \ldots, d\}$. Denote $x^{\overline{s}} = (x_1^{s_1}, \ldots, x_d^{s_d})$ for convenience. Disintegration measure denote as

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

Essentially, we use linearity of $\mathcal{L}_f(c_i^s, x_i^s)$ in every weight function c_i^s. It allows us to improve the total error by transforming one random quantizer into a non-random one. Now, out of all optimal quantizers c_i^s, x_i^s pick one with the least number of random quantizers (we name a quantizer $c_i^s, s \in \{1, \ldots, n_i\}$ non-random, if one of the weights is one and the others are zero). Without loss of generality assume that c_1^s is not random. Define

$$s_{1,b}(x_1) = \operatorname*{arg\,min}_{s_1\in\{1,\dots,n_1\}} \int_{\mathcal{X}_2\times\ldots\times\mathcal{X}_d} \sum_{s_2=1}^{n_d}\cdots\sum_{s_d=1}^{n_d} c_2^{s_2}(x_2)\ldots c_d^{s_d}(x_d)d(f(\overline{x}), f(x^{\overline{s}}))d\mu_{x_1}.$$

Clearly,

$$\int_{\mathcal{X}_1\times\ldots\mathcal{X}_d}\sum_{s_1,\dots,s_d} c_1^{s_1}(x_1)\ldots c_d^{s_d}(x_d)d(f(\overline{x}),f(x^{\overline{s}}))d\mu(\overline{x})$$
$$= \int_{\mathcal{X}_1}\sum_{s_1=1}^{n_1} c_1^{s_1}(x_1)d\mu_{\mathcal{X}_1}(x_1)\left(\int_{\mathcal{X}_2\times\ldots\times\mathcal{X}_d}\sum_{s_2,\dots,s_d} c_2^{s_2}(x_2)\ldots c_d^{s_d}(x_d)d(f(\overline{x}),f(x^{\overline{s}}))d\mu_{x_1}\right)$$
$$\geq \int_{\mathcal{X}_1} d\mu_{\mathcal{X}_1}(x_1)\left(\int_{\mathcal{X}_2\times\ldots\times\mathcal{X}_d}\sum_{s_2,\dots,s_d} c_2^{s_2}(x_2)\ldots c_d^{s_d}(x_d)d(f(\overline{x}),f(x_1^{s_{1,b}(x_1)},x_2^{s_2},\dots,x_d^{s_d}))d\mu_{x_1}\right)$$
$$= \int_{\mathcal{X}_1\times\ldots\times\mathcal{X}_d}\sum_{s_2,\dots,s_d} c_2^{s_2}(x_2)\ldots c_d^{s_d}(x_d)d(f(\overline{x}),f(x_1^{s_{1,b}(x_1)},x_2^{s_2},\dots,x_d^{s_d}))d\mu$$

In other words, we transformed random quantizer $c_1^s(x_1)$ into non-random one $q_1(x_1) = x_1^{s_{1,b}(x_1)}$ and the error improved. But the error was optimal already, and the number of random quantizers was minimal by our choice. Therefore, all the quantizers were non-random. □

4.3. **Properties of quantizing sets.**

Lemma 4.3. *Assume that f has a compact range, $d(u,v)\geq 0$ is lower semicontinuous in v, $d(u,v)=0$ iff $u=v$. In addition, $\mu(\overline{x})$ has a compact support and $\mu(f^{-1}(c))=0$ for all $c\in\mathbb{R}$. If $q_i(\mathcal{X}_i)=\{a_i^s\}_{s=1}^{n_i}$, set $A_i^s=q_i^{-1}(a_i^s)$. Assuming that $L_f(\overline{q})\to 0$ as $n_1,\dots,n_d\to\infty$, one has*

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

Proof. If not, there is an $\varepsilon>0$ and some $(A_1^{s_1}\times\ldots\times A_d^{s_d})(\overline{n})$ with

$$\mu((A_1^{s_1}\times\ldots\times A_d^{s_d})(\overline{n}))\geq\varepsilon \quad \text{with } (s_1,\dots,s_d)=(s_1,\dots,s_d)(\overline{n}).$$

Note that

$$L_f(\overline{q})\geq\int_{A_1^{s_1}\times\ldots\times A_d^{s_d}} d(f(\overline{x}),f(a_1^{s_1},\dots,a_d^{s_d}))\,d\mu(\overline{x}).$$

Up to a subsequence (not relabeled) one has $\mathbf{1}_{A_1^{s_1}\times\ldots\times A_d^{s_d}(\overline{n})}\to\varphi$ in the *weak sense of $L^\infty(\mu)$ and $f(a_1^{s_1},\dots,a_d^{s_d})\to c$ as $n_1,\dots,n_d\to\infty$. Moreover, $\int\varphi\,d\mu\geq\varepsilon$ and $\varphi\geq 0$ μ-a.e.. Therefore, due to Ioffe's theorem [1], one has

$$\begin{aligned}&\int_{\mathcal{X}_1\times\ldots\times\mathcal{X}_d}\varphi(\overline{x})d(f(\overline{x}),c)\,d\mu(\overline{x})\\ &\leq \liminf_{n_1,\dots,n_d\to\infty}\int_{\mathcal{X}_1\times\ldots\times\mathcal{X}_d}\mathbf{1}_{A_1^{s_1}\times\ldots\times A_d^{s_d}}(\overline{x})d(f(\overline{x}),f(a_1^{s_1},\dots,a_d^{s_d}))\,d\mu(\overline{x})\\ &\leq \liminf_{n_1,\dots,n_d\to\infty} L_f(\overline{q})=0.\end{aligned}$$

Since $d(u,v)$ is nonnegative, it is true that

$$\int_{\mathcal{X}_1\times\ldots\times\mathcal{X}_d}\varphi(\overline{x})d(f(\overline{x}),c)\,d\mu(\overline{x})=0,$$

which implies $f(\overline{x}) = c$ for a set of positive measure μ of $\overline{x}$ (area where $\varphi > 0$), 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 $f(\overline{x}) = \mathbf{1}_{A_1\times\ldots\times A_d}$, with measurable $A_i \subset \mathcal{X}_i$ then for $\forall i\, n_i \geq 2$ one has $C_f(\overline{n}) = 0$.*

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

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

□

We are now able to consider the case when f is a finite sum of characteristic functions of measurable rectangles.

Proposition 5.2. *If*

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

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 $C_f(\overline{n}) = 0$.

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 $\overline{x} \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ $e_i(x_i) = e_i(q_i(x_i))$. Therefore from definition of e_i one has

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

Consequently, $L_f(\overline{q}) = 0$ for any distance function d. □

- Note that the measurable rectangles $A_1^j \times \ldots \times A_d^j$ may be intersecting.
- 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 $\bar{N} \leq 2N$.
- The statement is constructive, i.e. it provides an algorithm for quantization.

Proof. The statement follows from the notion 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.

In order to algorithmically construct $q_i(\cdot)$ order all the edges of $A_i^j, j \in \{1, \ldots, N\}$ and consider the intervals between adjacent ones after ordering – parts of the division. Take any part of the division and define q_i there as a middle point of this part. The coding $e_i(x_i)$ is constant inside each part, i.e. $e_i(x_i) = e_i(q_i(x_i))$. Consequently, $f(\overline{x}) = f(q_1(x_1), \ldots, q_d(x_d))$ and the error is zero. □

Proposition 5.3. *If $C_f(\overline{n}) = 0$ and this error is achievable, then there are disjoint measurable sets $A_i^{s_i} \subset \mathcal{X}$, $1 \leq s_i \leq n_i$, $1 \leq i \leq d$ such that the union $\cup_{s_1,\ldots,s_d} A_1^{s_1} \times \ldots \times A_d^{s_d}$ covers $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ up to a μ-negligible set and*

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

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 $\mathcal{L}_f(\overline{q}) = 0$. 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

$$\begin{aligned} 0 = \mathcal{L}_f(\overline{q}) &= \int_{\mathcal{X}_1 \times \ldots \times \mathcal{X}_d} d(f(\overline{x}), f(q_1(x_1), \ldots, q_d(x_d)))\, d\mu(\overline{x}) \\ &= \sum_{s_1=1}^{n_1} \ldots \sum_{s_d=1}^{n_d} \int_{A_1^{s_1} \times \ldots \times A_d^{s_d}} d(f(\overline{x}), f(a_1^{s_1}, \ldots, a_d^{s_d}))\, d\mu(\overline{x}) \end{aligned}$$

which means that $f(\overline{x}) = f(a_1^{s_1}, \ldots, a_d^{s_d})$ for μ - a.e. $\overline{x} \in A_1^{s_1} \times \ldots \times A_d^{s_d}$. Denote $c_{s_1,\ldots,s_d} = f(a_1^{s_1}, \ldots, a_d^{s_d})$ and get that (5.1) is true. □

Remark 5.4. Here d may be any positive measurable function (not necessarily distance) such that $d(u, v) = 0$ implies $u = v$.

Remark 5.5. Due to Theorem 4.2 the best error is achievable for f with a compact range and d lower semicontinuous in the second coordinate. Thus, under these conditions, the above Proposition 5.3 gives us the representation of functions with zero quantization cost.

5.2. Characteristic functions of "nice" sets in $\mathbb{R} \times \mathbb{R}$.

Theorem 5.6. *Consider a characteristic function $f(x, y) = 1_K(x, y)$, standard Lebesgue measure $\mu = \mathcal{L}^2 \llcorner [0, 1]^2$ and distance $d(1, 0) = d(0, 1) = 1$. Then*

- *For a convex body K with a piecewise smooth boundary that is not a rectangle one has*
$$\mathcal{C}_f(n_1, n_2) \geq \frac{c(1 + o(1))}{\min(n_1, n_2)}, \text{ as } n_1, n_2 \to \infty.$$
- *For any body K with a piecewise smooth boundary one has*
$$\mathcal{C}_f(n_1, n_2) \leq \frac{P(K)(1 + o(1))}{\min(n_1, n_2)}, \text{ as } n_1, n_2 \to \infty.$$

Remark 5.7. The constant c in this relation depends on K. Clearly, the distance function is not important, because f has only 2 values. The upper bound error is achieved for a uniform quantization.

Remark 5.8. For a fixed total number of points $N = n_1 + n_2$ it is clear that

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

for some positive constants c_1 and C_2.

Proof. One can easily show that for a uniform quantization the upper bound holds. Divide $\mathcal{X} = [0,1]$ and $\mathcal{Y} = [0,1]$ into n_1 and n_2 equal intervals, then put a quantizing point into each interval. This way we have a lattice with $n_1 n_2$ small rectangles of size $n_1^{-1} n_2^{-1}$ with different quantizing points each. It is clear that only rectangles that intersect boundary of K add value to the error. Since for a convex K there are at most $4\max(n_1, n_2)$ small rectangles intersecting the boundary of K, we get the upper bound of $4\min(n_1, n_2)^{-1}$.

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 level sets of q_1 and q_2, $A_j, 1 \leq j \leq n_1$ and $\tilde{B}_k, 1 \leq k \leq n_2$ respectively. For each $1 \leq j \leq n_1$ we take $K_j = \{k \in \{1, \ldots, n_2\} | f(q_1(A_j), q_2(\tilde{B}_k)) = 1\}$ and construct

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

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

$$|K \triangle \bigcup_{j=1}^{n} A_j \times B_j| \succ n^{-1}.$$

Note that this is exactly the lower bound we want, since the symmetric difference $|K \triangle \bigcup_{j=1}^{n} (A_j \times B_j)|$ 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 piecewise smooth part of the boundary of K where all the outward normal vectors have strictly positive coordinates. Denote lengths of its x and y projections as P_x and P_y. For some constant C that we specify later, consider a polygonal chain of $k = Cn$ segments that are tangent to the chosen part of δK in its points of differentiability and have x-projections of the same length. Construct k right triangles with their main vertices inside K by using segments of this chain as hypothenuses. Enumerate all the triangles such that their y-coordinate is increasing and x-coordinate is decreasing. Let X_i, Y_i be projections of legs of the i-th triangle on x and y axes. Since the coordinates of the normal vectors are separated from 0, one can define ρ_1, ρ_2 depending only on K such that

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

Note that since all the $|X_i|$ are equal, we compare the maximum/minimum and average of the fraction of coordinates of normal vectors. Now we can clarify the choice of C – pick $C = 4\rho_1$. In the next section we prove that $A_j \times B_j$ either do not cover area of at least $(16\rho_1(2\rho_1\rho_2 + 1)n)^{-1} P_x P_y$ inside triangles, or cover at least $(16\rho_1(2\rho_1\rho_2 + 1)n)^{-1} P_x P_y$ outside of K. Then, to obtain similar estimate for K it remains to notice that a negligible part of triangles lies outside of K, since they use tangents as hypothenuses.

For the sets A_j and B_j we denote $a_i^j = |A_j \cap X_i| / |X_i|$ and $b_i^j = |B_j \cap Y_i| / |Y_i|$. Of course $a_i^j, b_i^j \in [0,1]$.
On one hand, the contribution of one set $A_j \times B_j$ to the covering of the triangles is not greater than

$$\sum_{i=1}^{k} a_i^j b_i^j |X_i||Y_i| \leq k^{-2} \rho_1 P_x P_y \sum_{i=1}^{k} a_i^j b_i^j.$$

On the other hand, $A_j \times B_j$ covers the part outside K with an area at least

$$\sum_{i=1}^{k-1} a_i^j |X_i| (b_{i+1}^j |Y_{i+1}| + \ldots + b_k^j |Y_k|) \geq k^{-2} \rho_2^{-1} P_x P_y \sum_{i=1}^{k-1} a_i^j (b_{i+1}^j + \ldots + b_k^j).$$

Now we need the following lemma.

Lemma 5.9. *For any $k \geq 1$, $a_i^j, b_i^j \in [0,1]$ one has*

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

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

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

□

We proceed with the original statement. The whole area of all the triangles is $\sum_i |X_i||Y_i|/2 = P_x P_y/(2k)$ since all the $|X_i|$ are equal. Thus, for

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

at least $\lambda P_x P_y/(2k)$ of it is covered, otherwise the uncovered part would be at least

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

Therefore,

$$k^{-2} \rho_1 P_x P_y \sum_{j=1}^{n} \sum_{i=1}^{k} a_i^j b_i^j \geq \lambda P_x P_y/(2k)$$

Then, lemma implies

$$\frac{P_x P_y}{k^2 \rho_2} \sum_{j=1}^{n} \sum_{i=1}^{k-1} a_i^j (b_{i+1}^j + \ldots + b_k^j) \geq \frac{P_x P_y}{2k^2 \rho_2} \sum_{j=1}^{n} \sum_{i=1}^{k} a_i^j b_i^j - \frac{n P_x P_y}{2k^2 \rho_2} \geq \frac{\lambda P_x P_y}{4\rho_1\rho_2 k} - \frac{n P_x P_y}{2k^2 \rho_2}.$$

Due to definitions of λ and k one has

$$\frac{\lambda P_x P_y}{4\rho_1\rho_2 k} - \frac{n P_x P_y}{2k^2 \rho_2} = \frac{P_x P_y}{16\rho_1(2\rho_1\rho_2 + 1)n}.$$

□

Corollary 5.10. *For a characteristic function of a right-angled triangle with sides P_x, P_y the quantizing error is bounded from below*

$$C_f(n_1, n_2) \geq \frac{P_x P_y}{48 \min(n_1, n_2)}.$$

5.3. **Linear functions over $\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$.** For the case when f is a linear function we are able to calculate exactly the quantization cost for a fairly large class of distance functions d.

Theorem 5.11. *Let $f(x) := \sum_{i=1}^{d} w_i x_i$ and $d(u,v) := p(|u-v|)$, where $t \to p(t)$ is convex and strictly increasing for $t \geq 0$, while $\mu := \mathcal{L}^d \llcorner [0,1]^d$. Then*

$$C_f(n) = \left| \frac{1}{\prod_i w_i} \int_{-w_1/2}^{w_1/2} \cdots \int_{-w_d/2}^{w_d/2} p\left(\left| \sum_i x_i/n_i \right| \right) dx_d \ldots dx_1 \right|.$$

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

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

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 $w_i > 0$. 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 a constant additive gets canceled in $f(x) - f(q(x))$. Now, when we work in the 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 does not change, i.e. the error remains the same. Therefore, we work with the case $w_i > 0$.

Let $\tilde{A}^i_k, k \in \{1, \ldots, n_i\}$ denote the level sets of $q_i, i \in \{1, \ldots, d\}$. Then

$$\begin{aligned} C_f(n_1, \ldots, n_d) &= \sum_{k_1=1}^{n_1} \cdots \sum_{k_d=1}^{n_d} \int_{\tilde{A}^1_{k_1} \times \ldots \times \tilde{A}^d_{k_d}} p(|\sum_i w_i \tilde{x}_i - c_{k_1,\ldots,k_d}|)\, d\tilde{x} \\ &= \sum_{k_1=1}^{n_1} \cdots \sum_{k_d=1}^{n_d} \frac{1}{\prod_i w_i} \int_{A^1_{k_1} \times \ldots \times A^d_{k_d}} p(|\sum_i x_i - c_{k_1,\ldots,k_d}|)\, dx, \end{aligned}$$

where $A^i_k = w_i \tilde{A}^i_k$. Note that $A^i_k, k = 1, \ldots, n_i$ cover $[0, w_i]$. Let us write one error term in the following way

$$\int_{A^1_{k_1} \times \ldots \times A^d_{k_d}} p(|\sum_i x_i - c_{k_1,\ldots,k_d}|)\, dx = \int_{A^1_{k_1}} G(x_1)\, dx_1,$$

where

$$G(x_1) = \int_{A^2_{k_2} \times \ldots \times A^d_{k_d}} p(|\sum_i x_i - c_{k_1,\ldots,k_d}|)\, dx_d \ldots dx_2.$$

Note, that all the functions $x_1 \to p(|\sum_i x_i - c_{k_1,\ldots,k_d}|)$ are convex, implying that the function $G(x_1)$ is also convex. In addition, $G(x_1)$ is not monotone, since $G(-\infty) = \infty$ and $G(\infty) = \infty$. Therefore, for some point α the function $G(x_1)$ decreases up to α and increases after.

Now, consider the following transformation of $A^1_{k_1}$ into an interval of the same measure. Denote $2a^1_{k_1} = |A^1_{k_1}|$. Take $t \in \mathbb{R}$ such that $\alpha - t = |A^1_{k_1} \cap (-\infty, \alpha)|$. Let us show that

$$\int_{A^1_{k_1}} G(x_1)\, dx_1 \geq \int_t^{t+2a^1_{k_1}} G(x_1)\, dx_1.$$

We divide this inequality into two separate ones – integrating up to α and after α. They are similar, so let us prove the first one, i.e. that

$$\int_{-\infty}^{\alpha} \mathbf{1}_{A^1_{k_1}}(x_1)G(x_1)\,dx_1 \geq \int_{-\infty}^{\alpha} \mathbf{1}_{[t,\alpha]}(x_1)G(x_1)\,dx_1.$$

After integrating both sides by parts, it remains to prove that

$$|A^1_{k_1}\cap(-\infty,\alpha)|G(\alpha)-\int_{-\infty}^{\alpha}|A^1_{k_1}\cap(-\infty,x_1)|dG(x_1) \geq |\alpha-t|G(x_1)-\int_{t}^{\alpha}|x_1-t|dG(x_1).$$

The first parts are equal due to the definition of t. To compare the integrals, taking into the account that $G(x_1)$ is decreasing for $x_1 < \alpha$, it is enough to show that $|A^1_{k_1} \cap (-\infty, x_1)| \geq |x_1 - t|$ for $x_1 \in (t, \alpha)$. It follows from an obvious observation that $|A^1_{k_1} \cap (x_1, \alpha)| \leq |\alpha - x_1|$ for $x_1 < \alpha$, combined with the definition of t.

After that, similarly, one by one we transform all the other sets $A^i_{k_i}$ into intervals in a way that decreases the error term. As a result, we get

$$\int_{A^1_{k_1}\times\ldots\times A^d_{k_d}} p(|\sum_i x_i-c_{k_1,\ldots,k_d}|)\,dx \geq \int_{t^1}^{t^1+2a^1_{k_1}}\ldots\int_{t^d}^{t^d+2a^d_{k_d}} p(|\sum_i x_i-c_{k_1,\ldots,k_d}|)\,dx$$

By doing linear change of variables we write the latter integral as

$$\int_{-a^1_{k_1}}^{a^1_{k_1}}\ldots\int_{-a^d_{k_d}}^{a^d_{k_d}} p(|\sum_i x_i - c|)\,dx. \tag{5.2}$$

Now, in order to get rid of c we use the following simple lemma.

Lemma 5.12. *Let X be a centrally symmetric real random variable and $t \to p(|t|)$ be a convex function with minimum at zero. Then*

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

Proof. The function $c \to \mathbb{E}\, p(|X - c|)$ is convex, because for a fixed x the function $c \to p(|x - c|)$ is convex. Moreover it is centrally symmetric, because X is centrally symmetric

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

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

The distribution of $X_1 + X_2 + \ldots + X_d$ for a vector $(X_1, \ldots, X_d)$ uniformly distributed on $[-a^1_{k_1}, a^1_{k_1}] \times \ldots \times [-a^d_{k_d}, a^d_{k_d}]$ is symmetrical about zero. Therefore, by lemma 5.12 the integral (5.2) is minimal when c is zero. Note that $c = 0$ is equivalent to $c_{k_1,\ldots,k_d} = t^1 + a^1_{k_1} + \ldots + t^d + a^d_{k_d}$. Putting all together, we obtain the estimate

$$\int_{A^1_{k_1}\times\ldots\times A^d_{k_d}} p(|\sum_i x_i - c_{k_1,\ldots,k_d}|)\,dx \geq \int_{-a^1_{k_1}}^{a^1_{k_1}}\ldots\int_{-a^d_{k_d}}^{a^d_{k_d}} p(|\sum_i x_i|)\,dx.$$

Then, using this for all the terms in the initial formula for a quantization error, we get the inequality

$$C_f(n_1, n_2, \ldots, n_d) \geq \frac{1}{\prod w_i}\sum_{k_1=1}^{n_1}\ldots\sum_{k_d=1}^{n_d}\int_{-a^1_{k_1}}^{a^1_{k_1}}\ldots\int_{-a^d_{k_d}}^{a^d_{k_d}} p(|\sum_i x_i|)\,dx,$$

where for all $i \in \{1, \ldots, d\}$ one has $\sum_{k_i=1}^{n_i} a^i_{k_i} = w_i/2$, since $A^i_{k_i}, k_i = 1, \ldots, n_i$ cover $[0, w_i]$ and their measure is defined as $2a^i_{k_i}$. Now, to finish the proof, we have to find the minimum of the right hand side with respect to all $a^i_{k_i}$. This part is purely technical and stated here as a lemma, its proof can be found in the section 7.

Lemma 5.13. *For any* $n_1, \dots, n_d \in \mathbb{N}$ *and nonnegative numbers* $a^i_{k_i}, k_i \in \{1, \dots, n_i\}, i \in \{1, \dots, d\}$ *such that for any* i *it is true that* $\sum_{k_i} a^i_{k_i} = w_i/2$*, one has*
(5.3)
$$\sum_{k_1=1}^{n_1} \dots \sum_{k_d=1}^{n_d} \int_{-a^1_{k_1}}^{a^1_{k_1}} \dots \int_{-a^d_{k_d}}^{a^d_{k_d}} p(|\sum_i x_i|)\, dx \geq n_1 \dots n_d \int_{-\frac{w_1}{2n_1}}^{\frac{w_1}{2n_1}} \dots \int_{-\frac{w_d}{2n_d}}^{\frac{w_d}{2n_d}} p(|\sum_i x_i|)\, dx.$$

This lemma implies that
$$C_f(n_1, \dots, n_d) \geq \frac{\prod n_i}{\prod w_i} \int_{-\frac{w_1}{2n_1}}^{\frac{w_1}{2n_1}} \dots \int_{-\frac{w_d}{2n_d}}^{\frac{w_d}{2n_d}} p(|\sum_i x_i|)\, dx,$$
which is exactly the lower bound we claimed, subjected to a simple linear change of variables.

To prove the second part of the statement one has to check that this error is achieved for a uniform quantization, i.e. for
$$q_i(x_i) = \frac{\lfloor n_i x_i \rfloor}{n_i} + \frac{1}{2n_i}.$$
Clearly, this statement can be checked via simple calculations, but to avoid those we can verify that all the inequalities in the proof of the lower bound become equalities.

- All $A^i_{k_i}$ are already intervals.
- Every $c_{k_1,\dots,k_d}$ is exactly sum of the centers of $A^i_{k_i}$.
- All $a^i_{k_i}, k_i \in \{1, \dots, n_i\}$ are the same for any fixed i.

□

One might wonder what is the best quantizing error when the total number of points in the grid $n_1 n_2 \dots n_d$ is fixed. The next remark answers this question, its proof is postponed to the section 7.

Remark 5.14. If $n_1 n_2 \dots n_d$ is fixed, the minimum of C_f in the theorem 5.11 is at $n_1/w_1 = n_2/w_2 = \dots = n_d/w_d$.

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

Remark 5.15. For a linear function $f(\overline{x}) = \sum_{i=1}^d w_i x_i$, Minkowski distance $d(u,v) = |u-v|^q, q \geq 1$ and Lebesgue measure $\mu(\overline{x}) = \mathcal{L}^d \llcorner [0,1]^d$ Theorem 5.11 gives the exact error
$$C_f = \frac{n_1 n_2}{2^{q+1}(q+1)(q+2)w_1 w_2} \left(\left(\frac{w_1}{n_1} + \frac{w_2}{n_2} \right)^{q+2} - \left| \frac{w_1}{n_1} - \frac{w_2}{n_2} \right|^{q+2} \right).$$
$$C_f = \frac{\prod_i n_i w_i^{-1}}{2^{q+d} q(q+1)\dots(q+d-1)} \sum_{\varepsilon_1 \in \{-1,1\}} \dots \sum_{\varepsilon_d \in \{-1,1\}} \prod_i \varepsilon_i \left| \sum_i \frac{\varepsilon_i w_i}{n_i} \right|^{q+d}.$$

Remark 5.16. Under conditions of remark 5.15, when $w_1/n_1 \to 0$ we get

$$C_f \to \frac{w_2^q}{2^q(q+1)n_2^q}.$$

Remark 5.17. Under conditions of remark 5.15, when $N = n_1 + n_2 + \ldots + n_d$ is fixed, one can show that the best possible quantizing error has the following order

$$\min_{\overline{n}:\sum_i n_i = N} C_f \sim C/N^q,$$

with $C = C(\overline{w}) > 0$.

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

$$\mathcal{C}_f(n_1, \ldots, n_d) \geq \frac{1}{\prod w_i} \int_0^{\frac{w_1}{2}} \ldots \int_0^{\frac{w_d}{2}} p(|\sum_i x_i/n_i|)\, dx.$$

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 $\mathcal{L}^d \llcorner [-1/2, 1/2]^d$ and then change sign of all coordinates along which f is decreasing, obataining a new function that is increasing in each coordinate.

Let $A^i_k, k \in \{1, \ldots, n_i\}$ denote the level sets of $q_i, i \in \{1, \ldots, d\}$. Then

$$C_f(n_1, \ldots, n_d) = \sum_{k_1=1}^{n_1} \ldots \sum_{k_d=1}^{n_d} \int_{A^1_{k_1} \times \ldots \times A^d_{k_d}} p(|f(x) - c_{k_1,\ldots,k_d}|)\, dx.$$

Denote $A_{k_1,\ldots,k_d} = A^1_{k_1} \times \ldots \times A^d_{k_d}$. Let us estimate one term of this sum as follows. Denote centers of mass of $A^i_{k_i}$ as α_i respectively. Consider the case $f(\alpha_1, \ldots, \alpha_d) > c_{k_1,\ldots,k_d}$, the opposite one is completely analogous. Since f is increasing in each coordinate, one has $f(x_1, \ldots, x_d) > f(\alpha_1, \ldots, \alpha_d) > c_{k_1,\ldots,k_d}$ when all $x_i > \alpha_i$ (for the opposite case take all $x_i < \alpha_i$). Then, from monotonicity of $p(\cdot)$ we obtain

$$\int_{A_{k_1,\ldots,k_d}} p(|f(x) - c_{k_1,\ldots,k_d}|)\, dx \geq \int_{\alpha_1}^{\infty} \ldots \int_{\alpha_d}^{\infty} \mathbf{1}_{A_{k_1,\ldots,k_d}}(x) p(|f(x) - f(\alpha)|)\, dx$$

From the condition of f we get a lower bound

$$\int_{\alpha_1}^{\infty} \ldots \int_{\alpha_d}^{\infty} \mathbf{1}_{A_{k_1,\ldots,k_d}}(x) p(|\sum_i w_i(x_i - \alpha_i)|)\, dx.$$

For $2a^i_{k_i} = |A^i_{k_i}|$, since α_i is a center of mass of $A^i_{k_i}$, this integral is not less than

$$\int_{\alpha_1}^{\alpha_1 + a^1_{k_1}} \ldots \int_{\alpha_d}^{\alpha_d + a^d_{k_d}} p(|\sum_i w_i(x_i - \alpha_i)|) dx = \int_0^{a^1_{k_1}} \ldots \int_0^{a^d_{k_d}} p(|\sum_i w_i x_i|)\, dx.$$

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

$$\mathcal{C}_f(n_1, \ldots, n_d) \geq \min_{a^i_{k_i} : \sum_{k_i=1}^{n_i} a^i_{k_i} = 1/2} \sum_{k_1=1}^{n_1} \ldots \sum_{k_d=1}^{n_d} \int_0^{a^1_{k_1}} \ldots \int_0^{a^d_{k_d}} p(|\sum_i w_i x_i|)\, dx.$$

It remains to show the the right hand side attains its minimum for $a^i_{k_i} = \frac{1}{2n_i}$. The proof of this bound is based on the same idea, as the proof of lemma 5.13, i.e. uses the Lagrange condition, but it is easier because all the variables are positive now. It remains to prove that

$$\sum_{k_1=1}^{n_1} \ldots \sum_{k_d=1}^{n_d} \int_0^{a^1_{k_1}} \ldots \int_0^{a^d_{k_d}} p(|\sum_i w_i x_i|)\, dx \geq \prod_i n_i \int_0^{\frac{1}{2n_1}} \ldots \int_0^{\frac{1}{2n_d}} p(|\sum_i w_i x_i|)\, dx,$$

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

$$\frac{1}{\prod_i w_i} \int_0^{\frac{w_1}{2}} \ldots \int_0^{\frac{w_d}{2}} p(|\sum_i y_i/n_i|) dy.$$

Clearly, this expression is decreasing in n_i. Now, we use a standard argument. Take $n_1, \ldots, n_d$ contradicting the inequality with the smallest sum. Since the condition $\sum_{k_i=1}^{n_i} a^i_{k_i} = 1/2, a^i_{k_i} \geq 0$ describes a compact space and the difference between l.h.s. and r.h.s. is continuous in this space, it attains its minimum at some point, clearly that minimum being less than zero. At this point all $a^i_{k_i}$ 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. In other words, we would obtain a contradictory configuration with smaller sum of n_i. Finally, when all the variables are strictly positive, one can apply Lagrange conditions and get that all the partial derivatives with respect to $a^i_{k_i}$ are the same for any fixed i. Consider the derivative with respect to $a^1_{k_1}$

$$\sum_{k_2=1}^{n_2} \ldots \sum_{k_d=1}^{n_d} \int_0^{a^2_{k_2}} \ldots \int_0^{a^d_{k_d}} p(|w_1 a^1_{k_1} + \sum_{i=2}^{d} w_i x_i|)\, dx_d \ldots dx_2.$$

It is monotone in $a^1_{k_1}$, i.e. Lagrange condition implies $a^1_1 = \ldots = a^1_{n_1}$. Similarly we get $a^i_1 = \ldots = a^i_{n_i}$, but this is exactly the point of equality.

□

Remark 5.19. Using this lower bound for a linear function f we would get a result worse than the exact error in Theorem 5.11, but it loses only by a factor not greater than 2^d. On the other hand, the restrictions in Theorem 5.11 are stronger, because the function $t \to p(|t|)$ is convex and f is linear.

The following easy statement is also worth mentioning.

Proposition 5.20. *For any function f and nonnegative distance d and two measures $\mu \leq \nu$, in the sense that for any Borel set B one has $\mu(B) \leq \nu(B)$, it is true that*

$$C_{f,d,\mu}(\overline{n}) \leq C_{f,d,\nu}(\overline{n}).$$

Proof. For any quantization functions q_1, q_2 one has

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

By passing to the infimum over all $\overline{q}$ we finish the proof. □

This immediately implies the following corollary,

Corollary 5.21. *Let f and d be as in Theorem 5.11. 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*

$$\mathcal{C}_{f,d,\mu} \le \left| \frac{C}{\prod_i w_i r_i} \int_{-w_1r_1/2}^{w_1r_1/2} \ldots \int_{-w_dr_d/2}^{w_dr_d/2} p\left(\left| \sum_i x_i/n_i \right| \right) d\overline{x} \right|.$$

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

$$\mathcal{C}_{f,d,\mu} \ge \left| \frac{c}{\prod_i w_i r'_i} \int_{-w_1r'_1/2}^{w_1r'_1/2} \ldots \int_{-w_dr'_d/2}^{w_dr'_d/2} p\left(\left| \sum_i x_i/n_i \right| \right) d\overline{x} \right|$$

In particular, for a distance function $d(u, v) = |u - v|^q, q \ge 1$, 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

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

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

Proof. Note that due to Proposition 5.20 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 $f(\overline{x}) := \sum_i w_i x_i = \sum w_i r_i y_i + const = \tilde{f}(\overline{y})$ for a linear function $\tilde{f}$. The distance $d(u, v)$ is translation invariant, thus the constant in $\tilde{f}$ can be omited. Finally, the loss $\mathcal{L}_{f,\mu}(\overline{q})$ is clearly linear in μ, therefore we can use Theorem 5.11 to obtain claimed estimate. The lower estimate is completely analogous and the last statement follows from the Remark 5.17. □

5.5. **Further examples of functions over $\mathbb{R}\times\mathbb{R}$.** For the quadratic cost $d(u, v) := |u - v|^2$ we are able to say slightly more.

Theorem 5.22. *Let $f(\overline{x}) = \sum_{i=1}^d \phi_i(x_i)$, where all ϕ_i have convex image and $d(u, v) := |u - v|^2$. Let X_i be independent with* 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 errors, i.e.*

$$C_f(\overline{n}) = \sum_{i=1}^d C_{\phi_i,d,\nu_i}(n_i)$$

Proof. Let $A^i_k, k \in \{1, \ldots, n_i\}$ denote the level sets of q_i respectively. Then

$$L_f(\overline{q}) = \sum_{k_1=1}^{n_1} \ldots \sum_{k_d=1}^{n_d} \int_{A^d_{k_d}} \ldots \int_{A^1_{k_1}} \left(\sum_i \phi_i(x_i) - c_{k_1,\ldots,k_d} \right)^2 d\nu_1(x_1) \ldots d\nu_d(x_d).$$

Consider one term of this sum. Define a random vector

$$(X^1_{k_1}, \ldots, X^d_{k_d}) = (\overline{X} | \overline{X} \in A^1_{k_1} \times \ldots \times A^d_{k_d}) \sim \otimes_i \mathbf{1}_{A^i_{k_i}} \frac{\nu_i}{\nu_i(A^i_{k_i})}.$$

The integral can be expressed as

$$\int_{A^1_{k_1} \times \ldots \times A^d_{k_d}} \left(\sum_i \phi_i(x_i) - c_{\overline{k}} \right)^2 d\mu(\overline{x}) = \prod_i \nu_i(A^i_{k_i}) \mathbb{E} \left[\left(\sum_i \phi_i(X^i_{k_i}) - c_{\overline{k}} \right)^2 \right].$$

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

$$c_{\overline{k}} = \mathbb{E} \left[\sum_i \phi_i(X^i_{k_i}) \right] = \sum_i \mathbb{E} \left[\phi_i(X^i_{k_i}) \right]$$

and for such a choice of $c_{\overline{k}}$ we get

$$\operatorname{Var} \left[\sum_i \phi_i(X^i_{k_i}) \right] = \sum_i \operatorname{Var} \left[\phi_i(X^i_{k_i}) \right],$$

because the variables X_{k_i} are independent. Consequently, we obtain a lower bound

$$L_f(\overline{q}) \geq \sum_{k_1=1}^{n_1} \ldots \sum_{k_d=1}^{n_d} \left(\prod_{i=1}^{d} \nu_i(A^i_{k_i}) \sum_{i=1}^{n} \operatorname{Var} \left[\phi_i(X^i_{k_i}) \right] \right) = \sum_{i=1}^{d} \sum_{k_i=1}^{n_i} \nu_i(A^i_{k_i}) \operatorname{Var} \phi_i(X^i_{k_i}),$$

and the equality is achieved for the right choice of $c_{\overline{k}}$, namely $c_{\overline{k}} = \sum_i \mathbb{E}\, \phi_i(X^i_{k_i})$. Recall that $c_{\overline{k}} = \sum_i \phi_i(x^i_{k_i})$, where $x^i_{k_i}$ are quantization points for $A^i_{k_i}$ respectively. Pick $x^i_{k_i} \in \phi_i^{-1}(\mathbb{E}\, \phi(X^i_{k_i}))$, it is possible because all ϕ_i have convex image. Therefore, for fixed level sets of q_i and the best choice of their values we get

$$L_f(\overline{q}) = \sum_{i=1}^{d} \sum_{k_i=1}^{n_i} \nu_i(A^i_{k_i}) \operatorname{Var} \phi_i(X^i_{k_i}).$$

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

$$\sum_{k=1}^{n_i} \nu_i(A^i_k) \operatorname{Var} \phi_i(X^i_k) = L_{\phi_i, d, \nu_i}(q_i).$$

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. □

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

$$\underline{d}(x, y) \leq d(g(x), g(y)) \leq \bar{d}(x, y)$$

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

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

Corollary 5.24. *Let $d(u,v) = p(|u-v|)$ for an increasing function $p(t), t \geq 0$ and $\mu = \mathcal{L}^d \llcorner [0,1]^d$. Let $f(\overline{x}) = g(\langle w, x\rangle)$.*
Assuming that for some function s it is true that $(p \circ s)(t), t \geq 0$ is convex increasing and $|g(a) - g(b)| \leq s(|a-b|), a, b$ in the range of $x \to \langle w, x\rangle$, one has

$$\mathcal{C}_f(\overline{n}) \leq \left| \frac{1}{\prod_i w_i} \int_{-w_1/2}^{w_1/2} \dots \int_{-w_d/2}^{w_d/2} (p \circ s)\left(\left| \sum_i x_i/n_i \right| \right) d\overline{x} \right|$$

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

$$\mathcal{C}_f(\overline{n}) \geq \left| \frac{1}{\prod_i w_i} \int_{-w_1/2}^{w_1/2} \dots \int_{-w_d/2}^{w_d/2} (p \circ r)\left(\left| \sum_i x_i/n_i \right| \right) d\overline{x} \right|.$$

Proof. Both inequalities immediately follow from Lemma 5.23 and Theorem 5.11. □

Remark 5.25. Let $f(x,y) = g(\langle w, x\rangle)$, where g is α-Hölder with a constant C, $d(u,v) = |u-v|^q, q \geq 1/\alpha$, and $\mu := \mathcal{L}^d \llcorner [0,1]^d$. Then

$$\mathcal{C}_f(\overline{n}) \leq \frac{C^q \prod_i n_i w_i^{-1}}{2^{\alpha q + d} \alpha q (\alpha q + 1) \dots (\alpha q + d - 1)} \sum_{\varepsilon_1 \in \{-1,1\}} \cdots \sum_{\varepsilon_d \in \{-1,1\}} \prod_i \varepsilon_i \left| \sum \frac{\varepsilon_i w_i}{n_i} \right|^{\alpha q + d}.$$

Proof. Note that $d(g(x), g(y)) = |g(x) - g(y)|^q \leq C^q |x-y|^{\alpha q}$. Therefore, by using Lemma 5.23 and Remark 5.15 we obtain the inequality. □

Remark 5.26. Let $f(x,y) = g(\langle w, x\rangle)$, where $|g(a) - g(b)| \geq c|a-b|^\alpha, \{a, b\}$ in the range of $x \to \langle w, x\rangle$, $d(u,v) = |u-v|^q, q \geq 1/\alpha$, and $\mu := \mathcal{L}^d \llcorner [0,1]^d$. Then

$$\mathcal{C}_f(\overline{n}) \geq \frac{c^q \prod_i n_i w_i^{-1}}{2^{\alpha q + d} \alpha q (\alpha q + 1) \dots (\alpha q + d - 1)} \sum_{\varepsilon_1 \in \{-1,1\}} \cdots \sum_{\varepsilon_d \in \{-1,1\}} \prod_i \varepsilon_i \left| \sum \frac{\varepsilon_i w_i}{n_i} \right|^{\alpha q + d}.$$

Proof. Note that $d(g(x), g(y)) = |g(x) - g(y)|^q \geq c^q |x-y|^{\alpha q}$. Consequently, Lemma 5.23 and Remark 5.15 combined prove the inequality. □

Corollary 5.27. *Let $f(\overline{x}) = g(\sum_i \phi_i(x_i))$ and $d(g(a), g(b)) \leq |a-b|^2$, while X_i are independent with the joint law $\otimes_i \nu_i$. Then*

$$\mathcal{C}_f(\overline{n}) \leq \sum_{i=1}^d C_{2,\phi_i,\nu_i}(n_i).$$

Corollary 5.28. *Let $f(\overline{x}) = g(\sum_i \phi_i(x_i))$ and $d(g(a), g(b)) \geq |a-b|^2$, while X_i are independent with the joint law $\otimes_i \nu_i$. Then*

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

Remark 5.29. Let $f(\overline{x}) = g(\sum_i \phi_i(x_i))$ and $d(u,v) = |u-v|^q$, where g is $2/q$ -Hölder with a constant R, while the joint law of X_i is $\otimes_i \nu_i$. Then

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

Remark 5.30. Let $f(\overline{x}) = g(\sum_i \phi_i(x_i))$ and $d(u,v) = |u-v|^q$, where $|g(a)-g(b)| \geq r|a-b|^{2/q}$, while the joint law of X_i is $\otimes_i \nu_i$. Then

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

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

Remark 5.31. Let $f(x,y) = \phi(x) + \psi(y)$ and $d(u,v) = |1 - u/v|^2$, while the joint law of X and Y is $\mu \otimes \nu$ 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 distance 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 $d(u,v) = |1-u/v|^2 \sim |\ln u - \ln v|^2$. Quantizing f with a distance function $|\ln u - \ln v|^2$ is the same as quantizing $\tilde{f}(x,y) = \ln(\phi(x)+\psi(y))$ with $\tilde{d}(u,v) = |u-v|^2$ 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 4.3 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 distance $|1-u/v|^2 \sim |\ln u - \ln v|^2$ 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 distance 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 5.32. Let $f(x,y) = (x+y)^2$, $d(u,v) = |u-v|^2$, 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 Remarks 5.30 and 5.29. 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 $\mathcal{X}_i = \mathbb{R}^{k_i}$, $i \in \{1,\ldots,d\}$. 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*

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

where M^* stands for the uncentered maximal function.

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

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

so that

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

as claimed. □

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

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

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

Moreover, if $p > 1$, then

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

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 $\lim_k M^*|\nabla f_k|(y) = M^*|\nabla f|(y)$ for a.e. $y \in \Omega$, i.e. for all $y \in \Omega \setminus N$ with $\mathcal{L}^d(N) = 0$.

It is enough to prove the statement for $n_i^{1/k_i} \in \mathbb{Z} \forall i \in \{1, \ldots, d\}$, 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

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

would imply the estimate for any n_i with a twice bigger constant.

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 $s_i \in \{1, \ldots, n_i\}, i \in \{1, \ldots, d\}$. Define then q_i by setting

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

Denote $A_{\overline{s}} = A_1^{s_1} \times \ldots A_d^{s_d}$ and $a_{\overline{s}} = a_1^{s_1}, \ldots, a_d^{s_d}$. Recalling that Lemma 6.1 implies

$$\int_{A_{\overline{s}}} |f_k(\overline{x}) - f_k(a_{\overline{s}})|\, d\overline{x} \leq C_k \operatorname{diam}(A_{\overline{s}}) \mathcal{L}^k(A_{\overline{s}}) M^* |\nabla f_k|(a_{\overline{s}}).$$

Summing up these inequalities, we get

$$\int_\Omega |f_k(\overline{x}) - f_k(q_1(x_1), \ldots, q_d(x_d))|\, d\overline{x} \leq C_k \max_i \left(r_i n_i^{-1/k_i}\right) \Delta(f_k, \Omega, \overline{n}), \tag{6.3}$$

$$\text{where } \Delta(f_k, \Omega, \overline{n}) := \sum_{s_1=1}^{n_1} \ldots \sum_{s_d=1}^{n_d} \mathcal{L}^k(A_1^{s_1} \times \ldots \times A_d^{s_d}) M^* |\nabla f_k|(a_1^{s_1}, \ldots, a_d^{s_d}).$$

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

$$\begin{aligned}\int_\Omega |f(\overline{x}) - f(q_1(x_1), \ldots, q_d(x_d))|\, d\overline{x} &\leq \liminf_k \int_\Omega |f_k(\overline{x}) - f_k(q_1(x_1), \ldots, q_d(x_d))|\, d\overline{x} \\ &\leq C_k \max \max_i \left(r_i n_i^{-1/k_i} \right) \Delta(f, \Omega, \overline{n}).\end{aligned} \tag{6.4}$$

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

$$\Delta(f, \Omega, \overline{n}) \to \int_\Omega M^*|\nabla f|(\overline{x})\, d\overline{x}$$

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

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

as $n_1, \ldots, n_d \to \infty$, which is (6.1) In particular, if $p > 1$, then estimating $\|M^*|\nabla f|\|_1$ 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 $C_f(N) = \min_{\sum n_i = N} C_f(\overline{n})$

$$C_f(N) \leq C_k \|\phi\|_\infty \|M^*|\nabla f|\|_1 \max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{1/k_i} + o\left(\max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{1/k_i} \right),$$

as $N \to \infty$. Moreover, for $p > 1$

$$C_f(N) \leq C_{k,p} \|\phi\|_\infty \|\nabla f\|_p \max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{1/k_i} + o\left(\max_i \left(\frac{\sum_i r_i^{k_i}}{N} \right)^{1/k_i} \right).$$

7. Technical proofs

7.1. Proof of lemma 5.13.

Proof. Step 1.

We first show that the righthand side of (5.3) is non-increasing with respect to n_i. Consider the following functional operator, that transforms p into

$$Dp(x_1, \ldots, x_d) = \sum_{\varepsilon_1 \in \{-1,1\}} \sum_{\varepsilon_2 \in \{-1,1\}} \ldots \sum_{\varepsilon_d \in \{-1,1\}} p(|\sum_i \varepsilon_i x_i|) \tag{7.1}$$

Note, that the integral can be rewritten in the following form

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

The inner function is decreasing in n_1 and n_2 due to lemma 7.2. Therefore, the integral is also decreasing in n_1 and n_2.

Lemma 7.1. *For a convex and strictly increasing on* $[0, \infty)$ *function* $p(\cdot)$ *and a fixed* t_0 *the function* $t \to p(|t + t_0|) + p(|t - t_0|)$ *is increasing on* $t \geq 0$.

Proof. First, without loss of generality we might assume $t_0 \geq 0$. The derivative of this function is $p'(|t + t_0|) + p'(|t - t_0|)\mathrm{sgn}(t - t_0)$. Since $p'(\cdot)$ is increasing and $|t + t_0| > |t - t_0|$ for $t_0, t > 0$, the derivative is positive. □

Lemma 7.2. *For a convex and strictly increasing on* $[0, \infty)$ *function* $p(\cdot)$ *and fixed* $x_2, \dots, x_d$ *the function* $x_1 \to Dp(x_1, \dots, x_d)$ *is increasing on* $x_1 \geq 0$.

Proof. Recall that

$$Dp(x_1, \dots, x_d) = \sum_{\varepsilon_1 \in \{-1,1\}} \dots \sum_{\varepsilon_d \in \{-1,1\}} p(|\sum_i \varepsilon_i x_i|).$$

Since $|\sum_i \varepsilon_i x_i| = |\sum_i -\varepsilon_i x_i|$, we can fix $\varepsilon_1 = 1$ and rewrite

$$Dp(x_1, \dots, x_d) = 2 \sum_{\varepsilon_2 \in \{-1,1\}} \dots \sum_{\varepsilon_d \in \{-1,1\}} p(|x_1 + \sum_{i=2}^d \varepsilon_i x_i|).$$

Now, from lemma 7.1 we obtain that for any particular choice of ε_i the sum

$$p(|x_1 + \sum_{i=2}^d \varepsilon_i x_i|) + p(|x_1 + \sum_{i=2}^d -\varepsilon_i x_i|)$$

is increasing on $x_1 \geq 0$. Now, we sum this over all choices of ε_i and get exactly $Dp(x_1, \dots, x_d)$. Consequently, Dp is increasing in x_1 for $x_1 \geq 0$. □

Step 2. We now prove the claim. Assuming that there is a set of numbers $((a_i), (b_j))$ for which inequality (5.3) fails, take the one with minimal $n_1 + n_2$. For convenience (a_i) and (b_j) are nondecreasing sequences. Consider the left hand side as a function $F(a_1, \dots, a_n)$ on a compact set $a_i \geq 0$, $\sum_i a_i = a/2$. By the extreme value theorem F attains its minimum at some point $(\tilde{a}_i)$ also being contrary to the inequality. If some of the $\tilde{a}_i$ was 0, we could remove it from the set $((\tilde{a}_i), (b_j))$, obtaining a smaller set of variables not satisfying the inequality, because the right hand side is decreasing with respect to n_1 and n_2. Therefore, $(\tilde{a}_i)$ is an interior point of a compact set we are working with. Thus, method of Lagrange multipliers provides us with the following equations on $(\tilde{a}_i)$: for some scalar λ and σ that are not 0 at the same time

$$\lambda \cdot \nabla F(\tilde{a}_1, \dots, \tilde{a}_n) = \sigma \cdot \nabla G(\tilde{a}_1, \dots, \tilde{a}_n) = \sigma \cdot (1, 1, \dots, 1).$$

Note that $\lambda \neq 0$, otherwise we would get $\sigma = 0$ too. Then, in coordinate form we get that for all i

$$\frac{\partial F}{\partial a_i}(\tilde{a}_1, \dots, \tilde{a}_n) = \frac{\partial F}{\partial a_j}(\tilde{a}_1, \dots, \tilde{a}_n).$$

Note that

$$\frac{\partial F}{\partial a_i}(\tilde{a}_1, \dots, \tilde{a}_n) = \sum_{j=1}^{n_2} \int_0^{b_j} p(|\tilde{a}_i + y|) + p(|\tilde{a}_i - y|)\, dy.$$

Due to the lemma 7.2, the derivative is strictly growing with respect to $\tilde{a}_i$.

Therefore, under the Lagrange condition we obtained, one has $\tilde{a}_1 = \ldots = \tilde{a}_{n_1}$. Now apply similar argument to $(b_1, \dots, b_{n_2})$ and get that the minimum with respect to (b_j) has to be at $b_1 = \ldots = b_{n_2}$. Consequently the inequality should not be true for some n_1, n_2 and $a_i = \frac{a}{2n_1}, b_j = \frac{b}{2n_2}$. But this is exactly the point of equality. □

7.2. **Proof of the remark 5.14.**

Proof. We want to prove that the minimum is attained when all n_i/w_i are equal. Let us show, that if we fix $n_1 n_2$ and shift n_1 and n_2 towards the point $n_1/w_1 = n_2/w_2$ then the functional decreases. After that, from symmetry we obtain that this is true for any pair n_i, n_j. Finally, we use a standard argument. Denote $\alpha = (\prod n_i/w_i)^{1/d}$. If not all n_i/w_i are equal to α, pick $n_j/w_j < \alpha$ and $n_k/w_k > \alpha$. Then we can shift n_j/w_j and n_k/w_k towards each other with a fixed product in a way that one of them becomes α. This operation does not change $n_1 \ldots n_d$, decreases total number of n_i such that $n_i \neq \alpha w_i$ and decreases the functional. Therefore, any choice of $n_1, \ldots, n_d$ can be transformed into $\alpha w_1, \ldots, \alpha w_d$ and the functional decreases along the way.

In order to prove a simplified statement we denote $t = \frac{w_1}{2n_1}$ and $\frac{c}{t} = \frac{w_2}{2n_2}$, where c is fixed and w.l.o.g. $w_1/n_1 > w_2/n_2$. Then shift towards equilibrium is equivalent to a reduction of t, therefore our goal is to show that the derivative of the error with respect to t is positive for $t > c/t$. Let us remind that the error is of the form

$$\frac{\prod_i n_i}{\prod_i w_i} \int_{-\frac{w_1}{2n_1}}^{\frac{w_1}{2n_1}} \cdots \int_{-\frac{w_d}{2n_d}}^{\frac{w_d}{2n_d}} p\left(\left|\sum_i x_i\right|\right) dx_d \ldots dx_1.$$

Using symmetry and an operator D defined in (7.1) we can rewrite this integral as follows

$$\frac{\prod_i n_i}{\prod_i w_i} \int_0^{\frac{w_1}{2n_1}} \int_0^{\frac{w_2}{2n_2}} \cdots \int_0^{\frac{w_d}{2n_d}} (Dp)(x_1, x_2, \ldots, x_d)\, dx_d \ldots dx_1.$$

For simplicity denote $W_+ = [0, \frac{w_3}{2n_3}] \times \ldots \times [0, \frac{w_d}{2n_d}]$. Let us also omit the constant factor. Finally, substitute t and c/t into the formula and get

$$\int_{W_+} \int_0^t \int_0^{c/t} (Dp)(x_1, \ldots, x_d)\, dx_d \ldots dx_1.$$

Its derivative with respect to t is

$$\int_{W_+} dx_d \ldots dx_3 \left(\int_0^{c/t} (Dp)(t, x_2, \ldots, x_d)\, dx_2 - ct^{-2} \int_0^t (Dp)(x_1, c/t, x_3, \ldots, x_d)\, dx_1 \right).$$

We aim to prove its positivity, thus it is enough to show that the inner part is greater than zero. Now, change the variables $y_2 = tx_2$ and $y_1 = c/tx_1$ in order to make these integrals more similar. The expression becomes

$$t^{-1} \int_0^c (Dp)(t, y_2/t, x_3, \ldots, x_d) dy_2 - t^{-1} \int_0^c (Dp)(ty_1/c, c/t, x_3, \ldots, x_d) dy_1.$$

From the definition, (Dp) is symmetric, thus we can rewrite this integral with a unified variable $z = y_2 = y_1$

$$t^{-1} \int_0^c (Dp)(t, z/t, x_3, \ldots, x_d) - (Dp)(c/t, tz/c, x_3, \ldots, x_d) dz.$$

It suffices to prove that when $t > c/t > 0$ and $0 < z < c$ one has

$$(Dp)(t, z/t, x_3, \ldots, x_d) > (Dp)(c/t, tz/c, x_3, \ldots, x_d).$$

Now, a simple symmetry argument allows to write the following formula

$$(Dp)(x_1, x_2, \dots, x_d) = 2(Dp)(x_1 + x_2, x_3, \dots, x_d) + 2(Dp)(|x_1 - x_2|, x_3, \dots, x_d).$$

This representation reduces our statement to two easier ones

$$\begin{cases} Dp(t + z/t, x_3, \dots, x_d) > Dp(c/t + tz/c, x_3, \dots, x_d), \\ Dp(|t - z/t|, x_3, \dots, x_d) > Dp(|c/t - tz/c|, x_3, \dots, x_d). \end{cases}$$

Lemma 7.2 simplifies even further to the form

$$t + z/t > c/t + tz/c, \qquad \text{and} \qquad |t - z/t| > |c/t - tz/c|.$$

The first one is equivalent to $(t - c/t)(1 - z/c) > 0$, which is true under our assumptions. To prove the second one it is enough to check that

$$\begin{cases} t - z/t > c/t - tz/c \\ t - z/t > tz/c - c/t. \end{cases}$$

These inequalities are equivalent to the following trivial ones

$$\begin{cases} (t - c/t)(1 + z/c) > 0 \\ (t + c/t)(1 - z/c) > 0. \end{cases}$$

□

References

[1] A. D. Ioffe, *On Lower Semicontinuity of Integral Functionals. I*, SIAM Journal on Control and Optimization., Vol. 15, Iss. 4 (1977) https://doi.org/10.1137/0315035

Southeast China University, China

St.Petersburg Branch of the Steklov Mathematical Institute of the Russian Academy of Sciences, Fontanka 27, 191023 St.Petersburg, Russia

St.Petersburg Branch of the Steklov Mathematical Institute of the Russian Academy of Sciences, Fontanka 27, 191023 St.Petersburg, Russia and Department of Mathematical Physics, Faculty of Mathematics and Mechanics, St. Petersburg State University, and HSE University, Moscow

Email address: stepanov.eugene@gmail.com