OPTIMAL FUNCTIONAL PRODUCT QUANTIZATION

TAO GUO, NIKITA KARAGODIN, AND EUGENE STEPANOV

ABSTRACT. We introduce a mathematically new approach for quantization of vectorial signals. Namely, we are aimed at quantizing vectorial signals on input to a computational device that calculates some given function of the input. As opposed to the classical approach which optimizes the quality of the quantized signal, without taking into account any further transformation of the latter by means of the computational device, here we optimize the quality of quantization on the output of this device. Moreover, we quantize components of the signal separately. This leads to a quantization problem qualitatively different from the classical one. We study existence of optimal quantizers (which is not at all obvious in this setting) and estimate the optimal cost for several classes of functions.

CONTENTS

2020 *Mathematics Subject Classification.* 49Q20, 49Q10; 49J55, 49Q22.

The first author was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62301144, and in part by the Zhishan Young Scholar Fund No. 2242025RCB0032. The third author is a member of GNAMPA and acknowledges also the support of the MIUR Excellence Department Project awarded to the Department of Mathematics, University of Pisa, CUP I57G22000700001. The results of section 3 are an output of a research project implemented as part of the Basic Research Program at HSE University.

1. Introduction

Consider a d-dimensional vectorial signal modelled by a random vector $X = (X_1, \ldots, X_d)$ with components X_i taking values in some set $\mathcal{X}_i$ (usually the components are scalar, i.e $\mathcal{X}_i$ are just subsets of reals), $i = 1, \ldots, d$. Suppose X has to be *quantized*, i.e. transformed into a signal which might assume only a discrete set of at most N values. Usually one is interested in performing this in an optimal way according to some chosen optimization criterion. The components are assumed to be random with their joint law (i.e. the law of the vectorial signal X) to be a known Borel probability measure μ. We describe now and compare between each other the two possible approaches to optimal quantization, the classical one, and the one proposed in this paper (called *functional product quantization* in the sequel).

1.1. Classical quantization.

Assume for simplicity that $d := 2$, each of the components X_i of the signal take values in some set $\mathcal{X}_i$, and denote $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ (usually $\mathcal{X}_i$ are just intervals of the real line, and hence $\mathcal{X}$ is a rectangle in $\mathbb{R}^2$, possibly even the whole of $\mathbb{R}^2$). The relatively well studied classical quantization approach is that of finding the *quantization map*

$$q \colon \mathcal{X} \to \mathcal{X}, \quad \# q(\mathcal{X}) \le N,$$

so as to minimize the mean quantization error

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

where c is the given cost function on $\mathcal{X}$, and $\mathbb{E}$ stands as usual for expectation. For instance, when one chooses $c(x_1, x_2) := |x_1 - x_2|^p$, which is the most frequent choice in applications, this amounts to minimizing the expectation of the power p (usually $p = 2$ or $p = 1$) of the absolute value of the difference between X and its quantized version. The cost of such a classical optimal quantization which measures the best possible quality of the latter, is given by

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

The case $\mathcal{X}_1 = \mathcal{X}_2 = [0,1] \subset \mathbb{R}$, so that $\mathcal{X} = [0,1]^2$ and $\mu = \mathcal{L}^2 \llcorner [0,1]^2$ is the usual Lebesgue measure (i.e. uniform distribution) over $\mathcal{X}$, is the most well studied. In this case $C(N) > 0$ for all N and

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

asymptotically as $N \to \infty$, with $C > 0$ a known constant (at least for $d = 2$; for general space dimension the explicit value of this constant by now is still unknown).

This quantization problem setting is very well-known and is in fact used in many different branches of mathematics under different names, see [5] for an introduction to the field as well as [7] for a survey on classical results. For instance, in the language of measure theory, this is a problem of finding the best possible approximation of the given measure μ by a discrete measure (i.e. finite sum of Dirac point masses) supported on a set of at most N points; in the case of the power cost $c(x_1, x_2) := |x_1 - x_2|^p$ this means finding the respective discrete measure which is nearest to μ in the sense of the Kantorovich (also called Wasserstein[1]) p-distance $W_p(\mu,\cdot)$. In such a formulation it is present in statistics and data science where it is used, e.g., for clustering. On the other hand the same problem can be formulated as finding the optimal location of an N-point set $\Sigma\subset\mathcal{X}$ minimizing the functional

$$\Sigma\mapsto\int_{\mathcal{X}}\mathrm{dist}_c(x,\Sigma)\,d\mu(x),$$

where $\mathrm{dist}_c(x,\Sigma):=\inf\{c(x,y)\colon y\in\Sigma\}$. In such a setting this problem is known in urban planning as *facility location* problem (with Σ interpreted as the set of facilities to locate, μ as the density of population in the given geographic area, and $c(x,y)$ the individual cost of getting from point x ro point y), or N-point problem (also used in many applications, e.g. in information theory for data compression). If $\mathcal{X}$ is, say, $\mathbb{R}^2$ (or a sufficiently large rectangle), c is the usual power cost as above and μ has compact support, then the optimal set Σ clearly exists, and once one knows Σ then one can easily find the optimal quantization map q as a nearest point projection onto Σ. The optimal discrete approximation of the measure μ is given just by the push-forward $q_{\#}\mu$. In different words, at least when $\mu\ll\mathcal{L}^2$, this is the measure $\sum_{j=1}^k w_j\delta_{x_j}$ where δ_x is the Dirac point mass located in x, $k\le N$, $\{x_j\}_{j=1}^k=\Sigma$ and the weights $w_j>0$ are given by

$$w_j:=\mu(V_j).$$

where $V_j:=\{x\in\mathcal{X}\colon |x-x_j|\le|x-x_i|\}$ for all $i\ne j$, is the Voronoi cell corresponding to the point x_j.

1.2. **Functional product quantization.** It is however sometimes important to know that the signal might be transmitted in order to be sent on *input* to some device computing a given function $f\colon\mathcal{X}_1\times\mathcal{X}_2\to\mathbb{R}$ (which will always in the sequel be considered Borel). If the signal has to be quantized, it is then more natural to measure the quality of quantization on the output of this device rather than on the input. Further, for engineers it is preferable to quantize each signal component separately and independently of other components. This leads to the following problem. Recalling that we have assumed $d:=2$ for simplicity, for $(n_1,n_2)\in\mathbb{N}\times\mathbb{N}$ one has to find the quantization maps

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

so as to minimize for a given cost function c on $\mathbb{R}$ the mean quantization error

$$L_f(q_1,q_2):=\mathbb{E}\,c\left(f\left(X_1,X_2\right),f\left(q_1(X_1),q_2(X_2)\right)\right)$$

on *output* of the computational device.

Similar quantization approach has been introduced in machine learning community in [1]. In information theory, it appears that the idea of independent coding of joint sources dates back to [2] and [3]. It gained attention recently with practical development of sensor networks, see [4]. The idea of combining product quantization with an objective to improve the quality of the output of the computational device rather than the signal itself. is also quite natural. Note that this is exactly what happens when an integral of a function is being approximated by a discretization technique. Moreover, in recent studies related to quantization of neural networks,

[1] The name Wasserstein distance is historically incorrect, and we prefer to attribute the name of Kantorovich to the latter

see for example [8], the most important part is also to improve the quality of the output function.

In the present paper we pursue exactly the above described approach of functional product quantization. We will see that it leads to the quantization problem strikingly different from the classical one. In fact, here already the existence of optimal quantization maps q_1, q_2 even for very nice data, is a priori quite unclear. In this paper, we are mainly interested in the asymptotical behaviour of the *quantization cost*

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

or alternatively, given a single number $N > 0$, of the minimum of the above costs $C_f(n_1, n_2)$ over all couples (n_1, n_2) such that $n_1 + n_2 \leq N$ (with an obvious abuse of notation we still denote this cost $C(N)$). We will see that, as opposed to the classical quantization, for certain functions f one can achieve zero cost even for finite N. We will further give upper and lower estimates on the quantization cost for particular classes of functions f (in particular, for linear f we are able to calculate the cost explicitly). These estimates are easily observed to be qualitatively different from those of the quantization cost for the classical setting.

2. Notation and preliminaries

2.1. **General notation.** Throughout the paper, we will assume, unless otherwise explicitly stated, that $\mathcal{X}_i$ be Polish spaces and μ_i Borel probability measures. Measurable sets in each $\mathcal{X}_i$ are those belonging to the completion of the Borel σ-algebra with respect to μ_i. However, big part of our results is related to the most common situation in applications when $\mathcal{X}_i \subset \mathbb{R}^{k_i}$ is just the subset of a Euclidean space and μ absolutely continuous with respect to the Lebesgue measure and with compact support. Even more, we will sometimes limit ourselves to the case $d = 2$ and $k_1 = k_2 = 1$, i.e. $\mathcal{X}_1 = \mathcal{X}_2 \subset \mathbb{R}$, $\mu \ll \mathcal{L}^2$ (with $\mathcal{L}^d$ standing for the d-dimensional Lebesgue measure). In fact, this case already contains all the essential difficulties of the problem considered.

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 $\stackrel{*}{\rightharpoonup}$.

Finally, the Euclidean norm in $\mathbb{R}^k$ will be denoted by $|\cdot|$, and the scalar product by $\langle \cdot, \cdot \rangle$.

2.2. **Signals.** For brevity we denote $\mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ and the signal by $X := (X_1, \ldots, X_d)$. The signals and there components are seen as random elements of the respective spaces. For the law μ of the random element Y of a Polish space $\mathcal{Y}$ we write $law(Y) = \mu$ or just $Y \sim \mu$. Note that the same symbol $\sim$ is used also to denote asymptotic equivalence of sequences. If Y is a random variable (i.e. $\mathcal{Y} = \mathbb{R}$), we denote by $\mathbb{E}(Y)$ its expectation and by $\mathrm{Var}(Y)$ its variance.

2.3. **Quantization.** The maps $q_i : \mathcal{X}_i \to \mathcal{X}_i$ will be usually referred to as *quantizers* or *quantization maps*, and we denote $q := (q_1, \dots, q_d)$. Each q_i takes n_i values that we denote as $a_i^{s_i}, s_i = 1, \dots, n_i$. Define the *quantizing sets* to be the level sets of quantizers, i.e. $A_i^{s_i} := q_i^{-1}(a_i^{s_i}), s_i = 1, \dots, n_i, i = 1, \dots, d$. Clearly the role of Voronoi cell in the classical quantization problem is played here by products of quantizing sets.

2.4. **Costs.** Sometimes to emphasize the dependence of the costs on c and μ we write $L_{f,c,\mu}(q)$ and $C_{f,c,\mu}(n_1, \dots, n_d)$ instead of $L_f(q)$ and $C_f(n_1, \dots, n_d)$ 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 $C_{c,\mu}(n_1, \dots, n_d)$ instead of $L(q)$ and $C(n_1, \dots, n_d)$ respectively.

3. Random quantization and existence of optimal quantizers

The goal of this section is to prove the existence of optimal quantizers. We only assume the spaces $\mathcal{X}_i$ to be Polish.

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 $f(\mathrm{W}) = \{f(w) : w \in \mathrm{W}\} \subset \mathbb{R}^{n_1 \cdots n_d}$. 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 3.1. *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 3.2. However, for practical use in engineering applications the sets $\mathcal{X}_i$ may always assumed to be compact.

Example 3.2. 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 quantizer providing zero cost.

Theorem 3.3. *Assume that $\mu = w(x_1,\dots,x_d)\mu_1\otimes\ldots\otimes\mu_d$ for Borel probability measures μ_i on $\mathcal{X}_i$ and $w(x_1,\dots,x_d)\in L^1(\mathcal{X},\mu_1\otimes\ldots\otimes\mu_d)$. 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 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.

3.1. **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

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

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.

Non-random 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 3.4. *Assume that $\mu=w(x_1,\dots,x_d)\mu_1\otimes\ldots\otimes\mu_d$, for Borel probability measures μ_i on $\mathcal{X}_i$ and $w(x_1,\dots,x_d)\in L^1(\mathcal{X},\mu_1\otimes\ldots\otimes\mu_d)$. 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 the 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 $L^\infty(\mathcal{X}_i,\mu_i)$ (recall that $\overset{*}{\rightharpoonup}$ stands for the weak* convergence) and $f(x_{1,k}^{s_1},\dots,x_{d,k}^{s_d})\to a_{s_1,\dots,s_d}$ as $k\to\infty$, then

$$\begin{aligned}&\liminf_{k\to\infty}\left(\int_{\prod_{j=1}^d\mathcal{X}_j} p_{1,k}^{s_1}(x_1)\dots p_{d,k}^{s_d}(x_d)c(f(x),f(x_{1,k}^{s_1},\dots,x_{d,k}^{s_d}))d\mu(x)\right)\\ &\qquad\geq\int_{\prod_{j=1}^d\mathcal{X}_j} p_1^{s_1}(x_1)\dots p_d^{s_d}(x_d)c(f(x),a_{s_1,\dots,s_d})d\mu(x).\end{aligned}\tag{3.1}$$

Taking for the moment (3.1) for granted, we deduce from it the lower semicontinuity of $\mathcal{L}_f$. Namely, we show that, denoting

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

one has

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

where points $x^{s_i}_i$ are such that $a_{s_1,\dots,s_d} = f(x^{s_1}_1, \dots, x^{s_d}_d)$. 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 $\bar{p}_{1,k}, \dots, \bar{p}_{d,k}, \bar{x}_{1,k}, \dots, \bar{x}_{d,k}$ for $\mathcal{L}_f$, extract convergent subsequences (not relabelled) such that $p^{s_i}_{i,k} \overset{*}{\rightharpoonup} p^{s_i}_i$ in $L^\infty(\mathcal{X}_i, \mu_i)$, $f(x^{s_1}_{1,k}, \dots, x^{s_d}_{d,k}) \to a_{s_1,\dots,s_d}$ as $k \to \infty$ for all $i = 1, \dots, d$, $s_i = 1, \dots, n_i$, 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 (3.1). To this aim let us show that

$$p^{s_1}_{1,k}(x_1) \dots p^{s_d}_{d,k}(x_d) \overset{*}{\rightharpoonup} p^{s_1}_1(x_1) \dots p^{s_d}_d(x_d) \quad \text{in } L^\infty(\mathcal{X}, \mu). \tag{3.2}$$

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

$$\int_{\mathcal{X}} p^{s_1}_{1,k}(x_1) \dots p^{s_d}_{d,k}(x_d) \phi(x) d\mu \to \int_{\mathcal{X}} p^{s_1}_1(x_1) \dots p^{s_d}_d(x_d) \phi(x) d\mu.$$

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

$$p^{s_1}_{1,k}(x_1) \dots p^{s_d}_{d,k}(x_d) \overset{*}{\rightharpoonup} p^{s_1}_1(x_1) \dots p^{s_d}_d(x_d) \text{ in } L^\infty(\mathcal{X}, \mu_1 \otimes \dots \otimes \mu_d),$$

thus proving (3.2).

Now, from (3.2) one has that the sequence of measures $p^{s_1}_{1,k}(x_1) \dots p^{s_d}_{d,k}(x_d) d\mu(x)$ converges setwise to the measure $p^{s_1}_1(x_1), \dots p^{s_d}_d(x_d) d\mu(x)$, because for any Borel $A \subset \mathcal{X}$ one has $\mathbf{1}_A \in L^1(\mathcal{X}, \mu)$, and thus

$$\int_A p^{s_1}_{1,k}(x_1) \dots p^{s_d}_{d,k}(x_d) d\mu(x) \to \int_A p^{s_1}_1(x_1), \dots p^{s_d}_d(x_d) d\mu(x).$$

Now, the statement (3.1) follows from the Fatou's lemma with varying measures [9, section 11.4, proposition 17] □

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

Proof of Theorem 3.3: We are going to prove a stronger statement, namely that although non-random quantization is a particular case of random quantization, the best quantizers are actually non-random. For the proof we only need the assumptions on the data (i.e. μ, f, c) ensuring the existence of optimal random quantizers. Consider the optimum for a random quantization problem $p_i^{s_i}(x_i), x_i^{s_i}, s_i = 1, \ldots, n_i, i = 1, \ldots, d$. We will show that it is achievable by non-random quantizers. We disintegrate

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

where μ_{x_i} are the respective 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 $p_i^{s_i}, s_i = 1, \ldots, n_i$ 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

$$\begin{aligned} &\hat{s}_1(x_1) := \operatorname*{arg\,min}_{s_1=1,\ldots,n_1} g_{x_1}(s_1), \quad \text{where} \\ &g_{x_1}(s_1) := \\ &\qquad \int\limits_{\mathcal{X}_2\times\ldots\times\mathcal{X}_d} \sum_{s_2,\ldots,s_d} p_2^{s_2}(x_2)\ldots p_d^{s_d}(x_d) c(f(x), f(x_1^{s_1}, \ldots, x_d^{s_d})) d\mu_{x_1}(x_2, \ldots, x_d). \end{aligned}$$

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

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

In other words, we transformed random quantizer $p_1^s(x_1)$ into non-random one (corresponding to the choice of quantization map $q_1(x_1) = x_1^{\hat{s}_1(x_1)}$) 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 3.5. As a by-product of the above proof we have that the best quantization error is equal to the best random quantization error.

3.3. **Properties of quantizing sets.** We prove here a simple property of optimal quantizers.

Lemma 3.6. *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 relabelled) one has $\mathbf{1}_{A_1^{s_1}(n_1) \times \ldots \times A_d^{s_d}(n_d)} \overset{*}{\rightharpoonup} \varphi$ 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, due to the Fatou's lemma with varying measures [9, section 11.4, proposition 17], one has

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

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. ☐

3.4. **A bridge between classical and functional product quantization.** The quantization of only one of the components of the input signal is a bridge between the classical approach and the functional product quantization. In this case the following estimate is considered

$$L_f(q_1, \mathrm{id}) := \mathbb{E}\, c(f(X_1, X_2), f(q_1(X_1), X_2))$$

and

$$C_f(n_1) := \inf\{L_f(q_1, \mathrm{id}) \colon q_1 \colon \mathcal{X}_1 \to \mathcal{X}_1, \#q(\mathcal{X}_1) \leq n_1\}$$

(of course there is an obvious abuse of notation here).

It is easy to observe that if c and f, say, are continuous over $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, and $\mathcal{X}_i$ are compact, then

$$C_f(n_1) \geq \lim_{n_2 \to \infty} C_f(n_1, n_2). \tag{3.3}$$

In fact, let $q_1^\varepsilon\colon \mathcal{X}_1 \to \mathcal{X}_1$ be such that $\# q_1^\varepsilon(\mathcal{X}_1) \le n_1$ and

$$L_f(q_1^\varepsilon, \mathrm{id}) \le C_f(n_1) + \varepsilon,$$

and $q_2^{n_2}\colon \mathcal{X}_2 \to \mathcal{X}_2$ be projections onto finite ε_{n_2}-nets in $\mathcal{X}_2$ of cardinality n_2 such that $\varepsilon_{n_2} \to 0$ as $n_2 \to \infty$. Then

$$\lim_{n_2\to\infty} C_f(n_1, n_2) \le \lim_{n_2\to\infty} L_f(q_1^\varepsilon, q_2^{n_2}) = L_f(q_1^\varepsilon, \mathrm{id}) \le C_f(n_1) + \varepsilon,$$

showing (3.3) because $\varepsilon > 0$ can be taken arbitrary. Of course already this proof suggests that (3.3) is valid under even less restrictive conditions on c, f and $\mathcal{X}_i$ (e.g. when only $\mathcal{X}_2$ is compact, f and c are Borel functions continuous only in the second variable, but c is bounded over $f(\mathcal{X}) \times f(\mathcal{X})$).

Surprisingly, the reverse inequality to (3.3) is not true: even the slightest quantization of the second component may drastically decrease the total error, as the following example shows.

Example 3.7. Let $\mathcal{X}_1 = \mathcal{X}_2 := [0,1]$, $\mu := \mathcal{L}^2 \llcorner [0,1]\times[0,1]$, $c(u,v) := |u-v|$ and let

$$f(x_1,x_2) := (1_{[1/3,2/3]\times[0,1/3]} + 1_{[0,1/3]\times[1/3,2/3]} + 1_{[2/3,1]\times[2/3,1]})(x_1,x_2).$$

Let $n_1 := 1$. Then whatever q_1 is, one has that $f(q_1(x_1), x_2)$ differs from $f(x_1,x_2)$ on the union of 4 squares of the total area $4/9$, so that $C_f(1) = 4/9$. On the other hand, if $q_1([0,1]) \in (0,1/3)$ and $q_2([0,1]) \in (0,1/3)$, then $f(q_1(x_1), q_2(x_2))$ differs from $f(x_1,x_2)$ 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 request f to be continuous or even smooth, since one can make the same example just by approximating characteristic functions involved with smooth ones.

4. Optimal quantizers for particular classes of functions

4.1. **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 (in this section, if not explicitly stated otherwise, $\mathcal{X}_i$ are generic Polish spaces).

Proposition 4.1. *If $f(x) = \mathbf{1}_{A_1\times\ldots\times A_d}(x)$, with measurable $A_i \subset \mathcal{X}_i$ then for $n_i \ge 2$ 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

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

concluding the proof. □

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

Proposition 4.2. *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 $n_i \ge \bar N$, one has $C_f(n_1,\ldots,n_d) = 0$.

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

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

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 quantization maps $q_i(x_i) := x_i^{e_i(x)}$. 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 4.3. Note that in Proposition 4.2

(1) 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 $\bar{N} \leq 2N$.

(2) the statement is constructive, i.e. it provides an algorithm for quantization.

To prove (1) 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 4.4. *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 $A_i^{s_i} \subset \mathcal{X}_i$, $s_i = 1, \ldots, n_i$, $i = 1, \ldots, d$ 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*

$$f(x) = \sum_{s_1=1}^{n_1} \ldots \sum_{s_d=1}^{n_d} c_{s_1,\ldots,s_d} \mathbf{1}_{A_1^{s_1}}(x_1) \ldots \mathbf{1}_{A_d^{s_d}}(x_d) \tag{4.1}$$

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 (4.1) is true. □

We can now apply Theorem 3.3 to get the following statement.

Corollary 4.5. *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 non-negative Borel function such that the map $v \mapsto c(u, v)$ is lower semicontinuous for all u. If, moreover, f is bounded and $c(u, v) = 0$, if and only if $u = v$, 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*

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

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 3.3 and Proposition 3.1 once one shows that f has a finite number of values. This would allow us to use Proposition 4.4 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 $q_{1,k}, \ldots, q_{d,k}$ such that

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

Now, by taking a weak* converging subsequence (not relabelled) we obtain that $\mathbf{1}_{A_{1,k}^{s_1} \times \ldots \times A_{d,k}^{s_d}} \stackrel{*}{\rightharpoonup} \phi_{s_1,\ldots,s_d}$ in $L^\infty(\mathcal{X}, \mu)$ for all $s_i = 1, \ldots, n_i$. Clearly,

$$\phi_{s_1,\ldots,s_d}(x_1, \ldots, x_d) \in [0,1]$$

for μ-a.e $(x_1, \ldots, x_d)$. Note that since

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

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

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

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

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

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

$$\int_{\mathcal{X}} \phi_{s_1,\ldots,s_d}(x_1, \ldots, x_d) c(f(x), c_{s_1,\ldots,s_d})\, d\mu(x) = 0. \tag{4.2}$$

Thus $f(x) = c_{s_1,\ldots,s_d}$ μ-a.e. on a set $D_{s_1,\ldots,s_d} = \{\phi_{s_1,\ldots,s_d} > 0\}$. 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

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

In other words, we leave quantizing sets the same as for f, but instead of taking $f(a_{1,k}^{s_1}, \dots, a_{d,k}^{s_d})$ as values, we take $c_{s_1,\dots,s_d}$. Clearly, from weak* convergence of $\mathbf{1}_{A_{1,k}^{s_1} \times \dots \times A_{d,k}^{s_d}} \rightharpoonup \phi_{s_1,\dots,s_d}$ in $L^\infty(\mathcal{X}, \mu)$, one has

$$\begin{aligned}
&\lim_{k\to\infty} \int_{A_{1,k}^{s_1} \times \dots \times A_{d,k}^{s_d}} c(\tilde{f}(x), \tilde{f}(x_1^{s_1}, \dots, x_d^{s_d})) d\mu(x) \\
&\quad = \lim_{k\to\infty} \int_{A_{1,k}^{s_1} \times \dots \times A_{d,k}^{s_d}} c(\tilde{f}(x), c_{s_1,\dots,s_d}) d\mu(x) \\
&\quad = \int_{\mathcal{X}} \phi_{s_1,\dots,s_d}(x_1, \dots, x_d) c(\tilde{f}(x), c_{s_1,\dots,s_d})\, d\mu(x) \\
&\quad = \int_{\mathcal{X}} \phi_{s_1,\dots,s_d}(x_1, \dots, x_d) c(f(x), c_{s_1,\dots,s_d})\, d\mu(x) \quad \text{since } \tilde{f} = f \ \mu\text{-a.e.} \\
&\quad = 0, \quad \text{by (4.2)},
\end{aligned}$$

which proves $C_{\tilde{f}}(n_1, \dots, n_d) = 0$. Consequently, by Proposition 3.1 and Theorem 3.3 we get that the best quantization error is achievable for $\tilde{f}$. Thus, the claim follows from Proposition 4.4 for $\tilde{f}$, and thus also for f because $f = \tilde{f}$ μ-a.e. □

4.2. **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. $f = 1_K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Here $d = 2$ and $\mathcal{X}_1 = \mathcal{X}_2 = \mathbb{R}$. Without loss of generality we suppose $K \subset [0,1]^2$ and $c(1,0) = c(0,1) = 1$. Let μ be the standard Lebesgue measure $\mu := \mathcal{L}^2 \llcorner [0,1]^2$.

Theorem 4.6. *Let $d = 2$ and $\mathcal{X}_1 = \mathcal{X}_2 = \mathbb{R}$, 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

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

the upper bound being achieved by uniform quantization.

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

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

Remark 4.7. 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.

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

This way we have a lattice with $n_1 n_2$ 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

$$\lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\mathcal{L}^2((\partial K)_\varepsilon) = P(K).$$

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 $n_1 \le n_2$. Consider the quantizing sets of q_1 and q_2, $A_j, j = 1, \ldots, n_1$ and $\tilde{B}_k, k = 1, \ldots, n_2$ respectively. For each $j = 1, \ldots, n_1$ we take $K_j = \{k \in 1, \ldots, n_2 \colon f(q_1(A_j), q_2(\tilde{B}_k)) = 1\}$ 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

$$\mathcal{L}^2\left(K \triangle \bigcup_{j=1}^{n_1}(A_j \times B_j)\right) \ge \frac{c(1 + o(1))}{n_1}, \qquad \text{as } n_1 \to \infty. \tag{4.3}$$

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 $f(x, y) \ne f(q_1(x), q_2(y))$, thus it contributes its measure to the total error.

Consider a smooth part of the ∂K where all the outward normal vectors have non-zero 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 hypotenuses. 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 catheti 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 hypotenuses 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

$$(1+o(1))\rho_1^{-1} \max_i |Y_i| \le \frac{P_y}{k} \le (1+o(1))\rho_2 \min_i |Y_i|, \qquad \text{as } n_1 \to \infty. \tag{4.4}$$

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 (4.3) 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 4.8. *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 (4.3) follows from Claim 4.8 because the area of triangles outside of K is asymptotically smaller than total area of triangles, i.e. it is

$$o(\sum_{i=1}^{k} |X_i||Y_i|) = o(|X_1| \sum_{i=1}^{k} |Y_i|) = o(P_x P_y / k) = o(n_1^{-1}),$$

which is asymptotically negligible for (4.3). Thus Claim 4.8 concludes the proof.

Step 3. It remains to prove Claim 4.8. 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

$$\sum_{i=1}^{k} a_i^j b_i^j |X_i||Y_i| \le (1+o(1))k^{-2}\rho_1 P_x P_y \sum_{i=1}^{k} a_i^j b_i^j. \tag{4.5}$$

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 (4.4).

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

$$\sum_{i=1}^{k-1} a_i^j |X_i|(b_{i+1}^j |Y_{i+1}| + \ldots + b_k^j |Y_k|) \ge (1+o(1))k^{-2}\rho_2^{-1} P_x P_y \sum_{i=1}^{k-1} a_i^j (b_{i+1}^j + \ldots + b_k^j). \tag{4.6}$$

This is because the set $\cup_{h\ge 1}((A_j \cap X_i) \times (B_j \cap Y_{i+h}))$ lies outside of K (so does the union of rectangles $\cup_{h\ge 1} X_i \times Y_{i+h}$ 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 (4.4).

By Lemma A.1 one has

$$\sum_{i=1}^{k-1} a_i^j (b_{i+1}^j + \ldots + b_k^j) \ge \frac{1}{2}\sum_{i=1}^{k} a_i^j b_i^j - \frac{1}{2}. \tag{4.7}$$

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 4.8 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 (4.5) we get

$$k^{-2}\rho_1 P_x P_y \sum_{j=1}^{n_1}\sum_{i=1}^{k} a_i^j b_i^j \ge (1+o(1))\lambda P_x P_y/(2k). \tag{4.8}$$

Thus, one has

$$
\begin{aligned}
&\frac{P_xP_y}{k^2\rho_2}\sum_{j=1}^{n_1}\sum_{i=1}^{k-1}a_i^j(b_{i+1}^j+\ldots+b_k^j)\\
&\qquad\geq \frac{P_xP_y}{2k^2\rho_2}\sum_{j=1}^{n_1}\sum_{i=1}^{k}a_i^jb_i^j-\frac{n_1P_xP_y}{2k^2\rho_2}\quad\text{by (4.7)}\\
&\qquad\geq (1+o(1))\frac{\lambda P_xP_y}{4\rho_1\rho_2 k}-\frac{n_1P_xP_y}{2k^2\rho_2}\quad\text{by (4.8)}\\
&\qquad= \frac{(1+o(1))P_xP_y}{16\rho_1(2\rho_1\rho_2+1)n_1}\quad\text{by definitions of } \lambda \text{ and } k\\
&\qquad= \frac{(1+o(1))c}{n_1}.
\end{aligned}
\tag{4.9}
$$

But claim (ii) and (4.6) implies that $\cup_{j=1}^{n_1}(A_j\times B_j)$ covers outside of K the area at least

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

hence, by (4.9), at least $(1+o(1))c/n_1$, which concludes the proof of Claim 4.8. □

The careful inspection of Step 2 and Step 3 of the proof of the above Theorem 4.6 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 4.9. *For a characteristic function of a right-angled triangle with sides P_x, P_y the quantizing error is bounded from below*

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

Proof. In terms of the above proof of Theorem 4.6 one can explicitly calculate $\rho_1=\rho_2=1$, and, therefore, $c=(16(2\rho_1\rho_2+1))^{-1}=1/48$. □

4.3. **Linear functions.** For the case when all $\mathcal{X}_i=\mathbb{R}$ and f is a linear function we are able to calculate exactly the quantization cost for a fairly large class of cost functions c.

Theorem 4.10. *Let all $\mathcal{X}_i=\mathbb{R}$, $i=1,\ldots,d$, $f(x):=\sum_{i=1}^d w_ix_i$ with $w_i\neq 0$ for all $i=1,\ldots,d$, and $c(u,v):=p(|u-v|)$, where $p\colon[0,+\infty)\to\mathbb{R}$ is convex and strictly increasing, while $\mu:=\mathcal{L}^d\llcorner[0,1]^d$. Then*

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

Moreover, the best quantization maps 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 $w_i>0, i=1,\ldots,d$. To see that this restriction does not lose generality, note that linearity of f allows us to shift the 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 cancelled 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 $\tilde{a}^i_{s_i}:=q_i(\tilde{A}_i^{s_i})$. 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

$$C_f(n_1,\dots,n_d)=\sum_{s_1,\dots,s_d}\int_{\tilde{A}_1^{s_1}\times\ldots\times\tilde{A}_d^{s_d}} p\left(\left|\sum_{i=1}^d w_i\tilde{x}_i-c_s\right|\right)d\tilde{x} \tag{4.10}$$
$$=\sum_{s_1,\dots,s_d}\frac{1}{\prod_{i=1}^d w_i}\int_{A_1^{s_1}\times\ldots\times A_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i-c_s\right|\right)dx,$$

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

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

where

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

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, from $\lim_{t\to+\infty}p(t)=+\infty$ we get $\lim_{x_1\to\pm\infty}G(x_1)=+\infty$. Therefore, there is the unique minimizer α of G.

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

$$\int_{A_1^{s_1}}G(x_1)\,dx_1\geq\int_t^{t+2a_1^{s_1}}G(x_1)\,dx_1. \tag{4.11}$$

To this aim we rewrite (4.11) as

$$\int_0^\infty\mathcal{L}^1(\{x_1\in A_1^{s_1}\colon G(x_1)>r\})\,dr\geq\int_0^\infty\mathcal{L}^1(\{x_1\in[t,t+2a_1^{s_1}]\colon G(x_1)>r\})\,dr. \tag{4.12}$$

To prove (4.12) 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

$$\mathcal{L}^1(\{x_1\in A_1^{s_1}\colon G(x_1)\leq r\})\leq\mathcal{L}^1(\{x_1\in[t,t+2a_1^{s_1}]\colon G(x_1)\leq r\}). \tag{4.13}$$

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 (4.13) would follow once one shows that for any $u\leq\alpha\leq v$ one has

$$\begin{aligned}&\mathcal{L}^1(A_1^{s_1}\cap[u,\alpha])\leq\mathcal{L}^1([\max(t,u),\alpha])=\min(\alpha-t,\alpha-u),\\&\mathcal{L}^1(A_1^{s_1}\cap[\alpha,v])\leq\mathcal{L}^1([\alpha,\min(t+2a_1^{s_1},v)])=\min(t+2a_1^{s_1}-\alpha,v-\alpha).\end{aligned} \tag{4.14}$$

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 (4.13) hence (4.12) hence (4.11).

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

$$\int_{-a_1^{s_1}}^{a_1^{s_1}} \ldots \int_{-a_d^{s_d}}^{a_d^{s_d}} p\left(\left|\sum_{i=1}^d x_i - c\right|\right) dx, \tag{4.15}$$

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 4.11. *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 4.11 the integral (4.15) 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

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

Then, using this estimate for all the terms in the initial formula (4.10) for a quantization error, we get following lower bound

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

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

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

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 $y_i := w_i(n_i x_i - 1/2)$ concludes 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 4.12. For a linear function $f(x) := \sum_{i=1}^d w_i x_i$, the cost

$$c(u,v) := |u - v|^\gamma, \quad \gamma \geq 1,$$

and Lebesgue measure $\mu := \mathcal{L}^d \llcorner [0,1]^d$ the Theorem 4.10 gives the exact error

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

In case $d = 2$ one gets

$$C_f = \frac{n_1 n_2}{2^{\gamma+1}(\gamma+1)(\gamma+2) w_1 w_2} \left(\left(\frac{w_1}{n_1} + \frac{w_2}{n_2} \right)^{\gamma+2} - \left| \frac{w_1}{n_1} - \frac{w_2}{n_2} \right|^{\gamma+2} \right).$$

Remark 4.13. Under conditions of Remark 4.12, when $N = n_1 + n_2 + \ldots + n_d$ is fixed, one can show that the best possible quantizing error has the order

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

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

4.4. **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. Here again all $\mathcal{X}_i = \mathbb{R}$.

Theorem 4.14. *Let all $\mathcal{X}_i = \mathbb{R}$, $i = 1, \ldots, d$, f 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 $\Delta_i > 0$, $i = 1, \ldots, d$ 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

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

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

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

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 centres of mass of $A_i^{s_i}$ as α_i respectively. Consider the case $f(\alpha_1, \ldots, \alpha_d) > c_s$, the opposite one is completely analogous. Since f is increasing in each coordinate, one has $f(x_1, \ldots, x_d) > f(\alpha_1, \ldots, \alpha_d) > c_s$ when all $x_i > \alpha_i$ (for the opposite case take all $x_i < \alpha_i$). Then, from monotonicity of $p(\cdot)$ we obtain

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

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

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

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

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

By definition, $A_i^{s_i}, s_i = 1, \ldots, n_i$ 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

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

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

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

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 4.15. Using this lower bound for a linear function f we would get a result worse than the exact error in Theorem 4.10, but it loses only by a factor not greater than 2^d. On the other hand, the restrictions in Theorem 4.10 are stronger, because the function $t \mapsto p(|t|)$ is convex and f is linear.

The following easy statement is also worth mentioning.

Proposition 4.16. *For any function f and non-negative cost c 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*

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

Proof. For any quantization maps q_1, q_2 one has

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

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

This immediately implies the following corollary,

Corollary 4.17. *Let f and c be as in Theorem 4.10. If for some rectangle $R = [a_1, a_1 + r_1] \times \dots \times [a_d, a_d + r_d]$ one has the inequality $\mu \leq C \mathbf{1}_R \mathcal{L}^d$, it is true that*

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

If for some rectangle $R' = [a_1, a_1 + r'_1] \times \ldots \times [a_d, a_d + r'_d]$ one has $\mu \geq c 1_{R'} \mathcal{L}^d$, then

$$\mathcal{C}_{f,c,\mu} \geq \left| \frac{c}{\prod_i w_i r'_i} \int_{-w_1 r'_1/2}^{w_1 r'_1/2} \cdots \int_{-w_d r'_d/2}^{w_d r'_d/2} p \left(\left| \sum_i x_i/n_i \right| \right) dx \right|$$

In particular, for a cost function $c(u,v) = |u-v|^\gamma, \gamma \geq 1$, 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

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

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

Proof. Note that due to Proposition 4.16 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 omitted. Finally, the loss $\mathcal{L}_{f,\mu}(q_1, \ldots, q_d)$ is clearly linear in μ, therefore we can use Theorem 4.10 to obtain claimed estimate. The lower estimate is completely analogous and the last statement follows from the Remark 4.13. □

4.5. **Quadratic cost.** For the quadratic cost $c(u,v) := |u-v|^2$ we are able to say slightly more. The respective result is valid for generic (Polish) spaces $\mathcal{X}_i$.

Theorem 4.18. *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 random elements of Polish spaces $\mathcal{X}_i$, with $\mathrm{law}(X_i) = \mu_i$, so that their joint law is $\mu = \otimes_i \mu_i$. Then one can choose the best quantization maps q_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.*

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

Proof. For $c_s := f(q_1(a_1^{s_1}), \ldots, q_d(a_d^{s_d})))$ by definition one has

$$L_f(q) = \sum_{s_1, \ldots, s_d} \int_{A_d^{s_d}} \cdots \int_{A_1^{s_1}} \left(\sum_{i=1}^d \phi_i(x_i) - c_s \right)^2 d\mu_1(x_1) \ldots d\mu_d(x_d).$$

Consider one term of this sum. Define a random vector

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

The integral can be expressed as

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

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

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

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

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

and the equality is achieved for the right choice of c_s, namely $c_s = \sum_{i=1}^d \mathbb{E}\,\phi_i(X_i^{s_i})$. Recall that by definition

$$c_s = \sum_{i=1}^d \phi_i(a_i^{s_i})..$$

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

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

Note that each quantizer here appears in a separate additive term, which allows to reduce the problem to the separate quantization in each variable. Indeed, using this equality for $d=1$ we get that

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

In other words, $L_f(q)$ is the sum of errors of classical quantization problems for the same choice of quantizers q_i. Then, they can be optimised separately to obtain the optimal error

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

hence concluding the proof. □

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

Lemma 4.19. *Let $g\colon \mathbb{R}\to\mathbb{R}$ satisfy the estimate*

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

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

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

Corollary 4.20. *Let all $\mathcal{X}_i=\mathbb{R}$, $c(u,v)=p(|u-v|)$ for an increasing function $p(t), t\geq 0$ and $\mu = \mathcal{L}^d\llcorner[0,1]^d$. Let $f(x) = g(\langle w,x\rangle)$. Assuming that for some function s the function $t\mapsto (p\circ s)(t), t\geq 0$ is convex increasing and $|g(a)-g(b)|\leq s(|a-b|), a,b$ in the range of $x\mapsto\langle w,x\rangle$, one has*

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

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

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

Proof. Both inequalities immediately follow from Lemma 4.19 and Theorem 4.10. ☐

Remark 4.21. Let $f(x) = g(\langle w, x \rangle)$, where g is α-Hölder with a Hölder constant C, $c(u, v) = |u - v|^\gamma, \gamma \geq 1/\alpha$, and $\mu := \mathcal{L}^d \llcorner [0,1]^d$. Then

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

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

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

Proof. If g is α-Hölder with a Hölder constant C, then $c(g(x), g(y)) = |g(x) - g(y)|^\gamma \leq C^\gamma |x-y|^{\alpha\gamma}$. Therefore, by using Lemma 4.19 and Remark 4.12 we obtain the upper bound inequality. Analogously, when $|g(a) - g(b)| \geq c|a-b|^\alpha, \{a, b\}$ in the range of $x \mapsto \langle w, x \rangle$, then $c(g(x), g(y)) = |g(x) - g(y)|^\gamma \geq c^\gamma |x-y|^{\alpha\gamma}$., and hence the lower bound inequality follows again by combining Lemma 4.19 and Remark 4.12. ☐

Corollary 4.22. *Let $f(x) = g(\sum_i \phi_i(x_i))$, while X_i are independent random elements of Polish spaces $\mathcal{X}_i$ with the law μ_i (i.e. their joint law being $\otimes_i \mu_i$). If $c(g(a), g(b)) \leq |a-b|^2$, then*

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

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

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

Remark 4.23. Let $f(x) = g(\sum_i \phi_i(x_i))$ and $c(u, v) = |u - v|^\gamma$, while the law of each X_i is μ_i, and their joint law is $\otimes_i \mu_i$. If g is $2/\gamma$-Hölder with a Hölder constant R, then

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

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

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

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

Remark 4.24. Let $f(x, y) = \phi(x) + \psi(y)$ and consider the cost function $c(u, v) = |1 - u/v|^2$ which arises 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 3.6 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

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

Let us assume that the error asymptotically better than $C_{2,\phi_\#\mu}(n_1) + C_{2,\psi_\#\nu}(n_2)$ can be achieved. In this case, one can pick $\varepsilon = \varepsilon(n_1, n_2) \to 0$ so slowly, as $n_1, n_2 \to \infty$, 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

$n_1, n_2 \to \infty$ one has

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

asymptotically. The last asymptotic relationship is due to the fact that $\varepsilon \to 0$ and that since the integrable function is uniformly bounded and $\mu \otimes \nu(B_\varepsilon) = o(C_{2,\phi_\# \mu}(n_1) + C_{2,\psi_\# \nu}(n_2))$, 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 4.25. 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 4.23. Here $g(t) = t^2$, i.e. $g'(t) = 2t$, so that g is a Lipschitz function with Lipschitz constant

$$\operatorname{Lip} g \leq 2(\max(|a_1|, |a_2|) + \max(|b_1|, b_2)),$$

and the first claim is true. Additionally, if $a_1 \geq 0, b_1 \geq 0$ it is true that

$$|g(t) - g(s)| \geq 2(a_1 + b_1)|t - s| \text{ for all } t, s \in [a_1 + b_1, a_2 + b_2]$$

and hence the second claim is true. ∎

5. General upper estimate for Sobolev functions

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

Lemma 5.1. *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*

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

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 5.2. *Let $\mathcal{X}_i := A_i \subset \mathbb{R}^{k_i}$ be open cubes of side length 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*

(5.1)
$$C_f(n_1, \ldots, n_d) \leq C_k \|\varphi\|_\infty \|M^* |\nabla f|^\gamma\|_1 \max_i (r_i n_i^{-1/k_i})^\gamma + o\left(\max_i (r_i n_i^{-1/k_i})^\gamma\right)$$

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

Moreover, if $p > \gamma$, then

(5.2)
$$C_f(n_1, \ldots, n_d) \leq C_{k,p} \|\varphi\|_\infty \|\nabla f\|_p^\gamma \max_i (r_i n_i^{-1/k_i})^\gamma + o\left(\max_i (r_i n_i^{-1/k_i})^\gamma\right).$$

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 $m_i^{1/k_i} \leq n_i^{1/k_i} \leq 2 m_i^{1/k_i}$. 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 5.1 implies

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

Summing up these inequalities, we get

(5.3)

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

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 (5.3), one arrives by Fatou's lemma at

(5.4)

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

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 (5.4) gives

(5.5)

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

as $(n_1, \ldots, n_d) \to \infty$, which is (5.1) In particular, if $p > \gamma$, then estimating $\|M^*|\nabla f|^\gamma\|_1$ by Hardy-Littlewood theorem, we get (5.2). ☐

Remark 5.3. 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 $p > \gamma$

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

Remark 5.4. In the formulation of the above Theorem 5.2 one could have assumed that $\mathcal{X}_i = \mathbb{R}^{k_i}$ and f be defined over the whole of $\mathcal{X} = \mathbb{R}^k$ by zero outside of Ω. The upper estimate provided by this theorem is clearly still valid, since increasing the sets $\mathcal{X}_i$ may only decrease the optimal cost.

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 $k \geq 1$, $a_i, b_i \in [0,1]$ one has*
$$\sum_{i=1}^{k-1} a_i(b_{i+1} + \ldots + b_k) \geq \frac{1}{2}\sum_{i=1}^{k} a_i b_i - \frac{1}{2}.$$

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 non-negative 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
$$\sum_{i=1}^{k'-1}(k'-i) = \frac{k'^2 - k'}{2} \geq \frac{1}{2}k' - \frac{1}{2},$$
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 $t_0 \geq 0$. We want to show that for any $a > b \geq 0$ one has
$$p(|t_0+a|) + p(|t_0-a|) \geq p(|t_0+b|) + p(|t_0-b|).$$
By convexity of $t \mapsto p(|t|)$ one has
$$\begin{aligned}\frac{a+b}{2a}p(|t_0+a|) + \frac{a-b}{2a}p(|t_0-a|) &\geq p\left(\left|\frac{a+b}{2a}(t_0+a) + \frac{a-b}{2a}(t_0-a)\right|\right)\\ &= p(|t_0+b|),\\ \frac{a-b}{2a}p(|t_0+a|) + \frac{a+b}{2a}p(|t_0-a|) &\geq p\left(\left|\frac{a-b}{2a}(t_0+a) + \frac{a+b}{2a}(t_0-a)\right|\right)\\ &= p(|t_0-b|).\end{aligned}$$
It remains to sum these two inequalities to get the claim (i).

For $t \geq t_0$ 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 $x_2, \ldots, x_d$ the function*
$$x_1 \mapsto (Tp)(x_1, \ldots, x_d) := \sum_{\varepsilon_1=\pm1} \cdots \sum_{\varepsilon_d=\pm1} p\left(\left|\sum_{i=1}^{d} \varepsilon_i x_i\right|\right)$$
is
- *(i) non-decreasing on $[0,+\infty)$*
- *(ii) and, moreover, strictly increasing on $[|x_2| + \ldots + |x_d|,+\infty)$.*

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 $n_1+\ldots+n_d$. 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 $a_i^1=\ldots=a_i^{n_i}$.

To show that $a_1^1,\dots,a_{n_1}^1$ can be set equal, consider the left hand side of (A.1) as a function $F=F(a_1^1,\dots,a_1^{n_1})$ 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 with a smaller sum $n_1+\ldots+n_d$ still not satisfying (A.1), because the right hand side is non-increasing with respect to n_1 (by Step 1). Therefore, $(\tilde{a}_1^{s_1}), s_1=1,\dots,n_1$ belongs to a relative interior point of a compact set we are working with. Thus, the method of Lagrange multipliers provides us with the following equations on $(\tilde{a}_1^{s_1})$: for some scalar λ and σ that do not vanish simultaneously:

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

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,

(A.3)
$$\frac{\partial F}{\partial a_1^{s_1}}(\tilde{a}_1^1, \ldots, \tilde{a}_1^{n_1}) = \sum_{s_2=1}^{n_2} \ldots \sum_{s_d=1}^{n_d} \int_0^{a_2^{s_2}} \ldots \int_0^{a_d^{s_d}} (Tp)(\tilde{a}_1^{s_1}, y_2, \ldots, y_d)\, dy_d \ldots dy_2.$$

Let us show that the integral in (A.3) is strictly increasing as a function of $\tilde{a}_1^{s_1} > 0$. First of all, due to Lemma A.3 its integrand is non-decreasing. In addition, when $y_2 + \ldots + y_d < \tilde{a}_1^{s_1}$ this 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) . □

Data Availability Statement

No new data were generated or analysed in support of this paper.

References

[1] H. Jégou, M. Douze and C. Schmid, "Product Quantization for Nearest Neighbor Search"', 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] Z. Xiong, A. D. Liveris and S. Cheng, "Distributed source coding for sensor networks", 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", IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2325-2383, Oct. 1998, 10.1109/18.720541.

[8] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, K. 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] H. L. Royden, "Real Analysis", New York: Macmillan, 2nd ed., 1968

School of Cyber Science and Engineering, Southeast University, Nanjing 210096, 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`

Dipartimento di Matematica, Università di Pisa, Largo Bruno Pontecorvo 5, 56127 Pisa, Italy and St.Petersburg Branch of the Steklov Mathematical Institute of the Russian Academy of Sciences, St.Petersburg, Russia and HSE University, Moscow
Email address: `stepanov.eugene@gmail.com`