Robust k-means Clustering for Distributions with Two Moments

Yegor Klochkov[1], Alexey Kroshnin[2], and Nikita Zhivotovskiy[3]

[1] *Cambridge-INET, Faculty of Economics, University of Cambridge, yk376@cam.ac.uk*
[2] *HSE University and Institute for Information Transmission Problems, RAS akroshnin@hse.ru*
[3] *Google Research, Brain Team, zhivotovskiy@google.com*

Abstract

We consider the robust algorithms for the k-means clustering problem where a quantizer is constructed based on N independent observations. Our main results are median of means based non-asymptotic excess distortion bounds that hold under the two bounded moments assumption in a general separable Hilbert space. In particular, our results extend the renowned asymptotic result of Pollard (1981) who showed that the existence of two moments is sufficient for strong consistency of an empirically optimal quantizer in $\mathbb{R}^d$. In a special case of clustering in $\mathbb{R}^d$, under two bounded moments, we prove matching (up to constant factors) non-asymptotic upper and lower bounds on the excess distortion, which depend on the probability mass of the lightest cluster of an optimal quantizer. Our bounds have the sub-Gaussian form, and the proofs are based on the versions of uniform bounds for robust mean estimators.

1 Introduction

Statistical (sample-based) k-means clustering is the classical form of quantization for probability measures. In this framework, given a distribution P defined on a normed space $(E, \|\cdot\|)$ and an integer $k \geq 1$, one wants to find $A^* \subset E$ such that the *distortion*

$$D(A) = \mathbb{E} \min_{a \in A} \|X - a\|^2 \quad \text{is minimized among all} \quad A \subset E, |A| = k.$$

It is well known that if $(E, \|\cdot\|)$ is $\mathbb{R}^d$ with the Euclidean norm and if $\mathbb{E}\,\|X\|^2 < \infty$ then this *optimal quantizer* A^* exists (see e.g., Theorem 1 in (Linder, 2002)), although it is not necessarily unique for $k \geq 2$. The value of the optimal distortion can be written as $D(A^*)$. In the statistical setup, the access to P is achieved via N independent observations $X_1, \ldots, X_N$ sampled according to P. Consider again the case of $\mathbb{R}^d$ and the Euclidean norm. The following renowned result due to Pollard (1981) states strong consistency of (any) *empirically optimal quantizer*, which is defined by

$$\hat{A} \in \operatorname*{arg\,min}_{A \subset \mathbb{R}^d, |A| = k} \frac{1}{N} \sum_{i=1}^{N} \min_{a \in A} \|X_i - a\|^2. \tag{1}$$

Theorem 1.1 (Strong consistency of k-means (Pollard, 1981))**.** *For any distribution P such that $\mathbb{E}\,\|X\|^2 < \infty$ and any integer $k \geq 1$, it holds that*

$$D(\hat{A}) - D(A^*) \overset{a.s.}{\to} 0, \quad \textit{as} \quad N \to \infty.$$

This consistency result was extended to the case where the space $(E, \|\cdot\|)$ is a general separable Hilbert space (Biau, Devroye and Lugosi, 2008; Levrard, 2015). Clearly, the consistency alone does not provide any information on how many training samples are needed to ensure that the excess distortion is below a certain level. Moreover, it does not allow the underlying distribution to be different for each N. Over the last three decades a lot of efforts were made in order to prove non-asymptotic results for the *excess distortion* $D(\hat{A}) - D(A^*)$ where the space is $\mathbb{R}^d$ or a general separable Hilbert space. We refer to various bounds in (Bartlett, Linder and Lugosi, 1998; Linder, 2002; Biau, Devroye and Lugosi, 2008; Graf and Luschgy, 2007; Maurer and Pontil, 2010; Narayanan and Mitter, 2010; Levrard, 2013, 2015; Fefferman, Mitter and Narayanan, 2016; Maurer, 2016) and the references therein. However, almost all the results were provided under the strong assumption that the domain is bounded. That is, it is usually assumed that $\|X\| \le T$ almost surely where X is distributed according to P and $T > 0$ is a constant. This simple setup allows one to use the tools of Empirical Process Theory in order to prove results of the form (see e.g., Theorem 2.1 by Biau, Devroye and Lugosi (2008), where the space (E, d) is assumed to be a separable Hilbert space)

$$D(\hat{A}) - D(A^*) \lesssim T^2\sqrt{\frac{k^2 + \log\frac{1}{\delta}}{N}}, \tag{2}$$

holding, with probability at least $1-\delta$, for $\delta \in (0,1)$. The question of general unbounded distributions is more challenging and has been studied less. The case where the vectors X_i have well behaved exponential moments was analyzed in (Cadre and Paris, 2012). Results under less restrictive assumptions include: the uniform deviation bounds in (Telgarsky and Dasgupta, 2013; Bachem, Lucic, Hassani and Krause, 2017); a *sub-Gaussian* excess distortion bound in (Brownlees, Joly and Lugosi, 2015) for the so-called k-medians problem; and the results for trimmed quantizers in (Brécheteau, Fischer and Levrard, 2018). We will discuss some of these results in more detail in what follows. However, we emphasize that in our particular setup the results we are aware of require the existence of at least *four moments* (that is, $\mathbb{E}\|X\|^4 < \infty$) compared to the minimal assumption under which the problem makes sense — $\mathbb{E}\|X\|^2 < \infty$ — which we are aiming for in this paper (this assumption is required to define the distortion $D(A^*)$). The question whether non-asymptotic results of the form (2) are possible under the minimal assumption $\mathbb{E}\|X\|^2 < \infty$ (as in (Pollard, 1981)) appeared naturally in several papers (see e.g., (Levrard, 2013)) but has not yet been addressed.

Our motivating example is the sub-Gaussian mean estimator introduced in (Lugosi and Mendelson, 2019c). Consider the situation where $E = \mathbb{R}^d$ with the Euclidean norm, set $\mu = \mathbb{E}\,X$, and assume that the covariance matrix $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^\top$ exists. If $k = 1$, we obviously have that the optimal quantizer A^* is actually the mean μ. In this case, our problem boils down to the estimation of the mean of a random vector. It was shown by Lugosi and Mendelson that there is an estimator such that, with probability at least $1-\delta$,

$$\mathbb{E}\|X - \hat{a}\|^2 - \mathbb{E}\|X - \mu\|^2 \lesssim \frac{\mathbb{E}\|X-\mu\|^2 + \lambda_{\max}(\Sigma)\log\frac{1}{\delta}}{N}, \tag{3}$$

where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of the covariance matrix Σ, the expectation is taken only with respect to X, and $\hat{a} = \hat{a}(X_1, \ldots, X_N)$ is random. It is known that this bound is valid for the sample mean in the case where the underlying distribution is multivariate Gaussian. The bound (3) has some remarkable properties:

- The dependence on N is $O\left(\frac{1}{N}\right)$.
- It only requires the existence of two moments, that is $\mathbb{E}\|X\|^2 < \infty$. We note that in $\mathbb{R}^d$, $\lambda_{\max}(\Sigma) \le \mathrm{Tr}(\Sigma) = \mathbb{E}\|X-\mu\|^2$.
- It has the logarithmic dependence on the confidence, which is $\log\frac{1}{\delta}$ and corresponds to the *sub-Gaussian tails* (see e.g., (Vershynin, 2016) for various equivalent definitions of sub-Gaussian distributions).
- Finally, even in the favorable bounded case where $\|X - \mu\| \le T$ almost surely the bound (3) does not scale as T^2 (compare it with the typical k-means bound (2)) but as $\mathbb{E}\|X-\mu\|^2$ which can be much smaller than T^2.

Therefore, extending the original question of whether the non-asymptotic excess distortion bounds are possible under $\mathbb{E}\|X\|^2 < \infty$, it is natural to ask if one can prove a result of the form (3) for $k \ge 2$. Unfortunately, fully general picture is much more subtle. In particular, even in the favorable bounded case for $k \geq 2$ the lower bounds of order $\Omega\left(\frac{1}{\sqrt{N}}\right)$ are known (see e.g., (Antos, 2005)) making the simple bound (2) sharp with respect to N.

Further, if $k = 1$ we observe that the right-hand side of (3) converges to zero as N goes to infinity even if *the underlying distribution $P = P(N)$ is different for each N*. Our only condition is that $\mathbb{E}\,\|X - \mu\|^2 = \operatorname{Tr}(\Sigma)$ does not grow too fast as N goes to infinity. Surprisingly, Example 1.2 below will show that the same is not true for general $k \geq 2$.

Risk bounds having the sub-Gaussian form for heavy-tailed distributions have attracted a lot of attention recently. Among these advances are (almost) optimal results on mean estimation in various norms (Minsker, 2018; Lugosi and Mendelson, 2019a); in robust regression (Minsker and Mathieu, 2019; Lugosi and Mendelson, 2019b; Lecué and Lerasle, 2017); in covariance estimation (Mendelson and Zhivotovskiy, 2018); in classification (Lecué, Lerasle and Mathieu, 2018). All the technical results in this area are based on different versions of the so-called *median of means estimator*, which was first introduced and analyzed by Nemirovsky and Yudin (1983) and independently in (Alon, Matias and Szegedy, 1999). For the sake of completeness, let us recall this basic result.

Assume that $Y_1, \ldots, Y_N$ are independent random variables with the same mean μ and the same variance σ^2. Fix the confidence level δ and assume that $\ell = \lceil 8 \log \frac{1}{\delta} \rceil$ is such that $N = m\ell$, where m is integer. Split the set $\{1, \ldots, N\}$ into ℓ blocks $I_1, \ldots, I_\ell$ of equal size such that $I_j = \{1 + m(j-1), \ldots, mj\}$. Denote the *median of means* (MOM for short) estimator by

$$\hat{\mu} = \operatorname{Median}\left(\frac{\ell}{N}\sum_{i \in I_1} Y_i, \ldots, \frac{\ell}{N}\sum_{i \in I_\ell} Y_i\right).$$

For this estimator we have the following sub-Gaussian behaviour. It holds, with probability at least $1 - \delta$, that

$$|\hat{\mu} - \mu| \leq \sigma\sqrt{\frac{32 \log \frac{1}{\delta}}{N}}.$$

Returning to the question of k-means clustering and the inequalities of the form (3) for general $k \geq 2$, the following simple example inspired by Bachem, Lucic, Hassani and Krause highlights some of the obstacles we will have to handle.

Example 1.2. *Let N be the sample size. Consider the real line $\mathbb{R}$, $k = 2$ and the distribution P supported on $\{0, \sqrt{N}\}$ such that $P(\{0\}) = 1 - \frac{1}{N}$ and $P(\{\sqrt{N}\}) = \frac{1}{N}$. In this case we have $D^*(A) = 0$.*

One may easily see that, with constant probability, the value $\sqrt{N}$ is not among $X_1, \ldots, X_N$. That will obviously force $\hat{A} = \{0\}$ and

$$D(\hat{A}) - D(A^*) = \mathbb{E}\,X^2 = 1,$$

which is not converging to zero as N goes to infinity.

In Section 3.1 we will significantly extend this construction. Of course, Example 1.2 does not contradict the strong consistency result of Theorem 1. Although $\mathbb{E}\,X^2 = 1$, the distribution $P = P(N)$ changes with N, which is, of course, not allowed in Theorem 1. However, in the statistical learning theory literature, the underlying distribution P is usually allowed to change with N, and this provides an additional motivation for our study. Surprisingly, our general bounds will provide consistency even for some sequences of distributions changing with the sample size N.

On Voronoi cells and clustering. From now on we assume that $(E, \|\cdot\|)$ is a separable Hilbert space with the inner product denoted by $\langle\cdot,\cdot\rangle$. Any quantizer $A = \{a_1, \ldots, a_k\}$ induces a partition of E into the so-called *Voronoi cells*, which for $a \in A$ consists of the points that have a as the closest point from A. To avoid the uncertainty at the boundaries, we assume that the elements of each quantizer $A = \{a_1, \ldots, a_k\}$ are ordered, and define the cells for $j = 1, \dots, k$,

$$\begin{aligned} V_j(A) = \{x \in E : & \|x - a_j\| < \|x - a_{j'}\|, \ j' = 1, \dots, j-1, \\ & \|x - a_j\| \le \|x - a_{j'}\|, \ j' = j+1, \dots, k\}. \end{aligned}$$

This way, we ensure that the cells are non-intersecting, and each of them is an intersection of $k-1$ open or closed half-spaces. Slightly abusing the notation, we sometimes write V_{a_j} instead of V_j.

We recall some basic properties of an *optimal* quantizer under the assumption $\mathbb{E}\,\|X\|^2 < \infty$.

1. For any distribution P with $\mathbb{E}\|X\|^2 < \infty$ and any k there exists an optimal k-elements quantizer A^* (see (Fischer, 2010), Corollary 3.1.) Note that an optimal quantizer is not necessarily unique.
2. For any optimal A^* and $i \neq j$,
$$P\left(\|X - a_j\| = \|X - a_i\|\right) = 0,$$
which means that the measure of intersection of any two cells is zero, thus it does not matter to which cell the boundary points are assigned (see (Graf and Luschgy, 2007), Theorem 4.2.)
3. The *centroid condition* (Graf and Luschgy, 2007): for $j = 1, \dots, k$,
$$\mathbb{E}\|X - a_j\|^2 \,\mathbb{1}[X \in V_j] = \inf_{a \in E} \mathbb{E}\|X - a\|^2 \,\mathbb{1}[X \in V_j] \quad \text{and} \quad a_j = \frac{\mathbb{E}\, X\,\mathbb{1}[X \in V_j]}{P(V_j)}.$$
4. Once the support of P consists of at least k elements, there is a well-defined real number $M = M(P, k)$ such that for any optimal A^*,
$$\|a\| \le M \quad \text{for all} \quad a \in A^*. \tag{4}$$
We refer to the original proof of Pollard or to Lemma 5.1 in (Fischer, 2010).
5. Due to Theorem 4.1 in (Graf and Luschgy, 2007) provided that the support of P consists of at least k elements, there exists $p_{\min} > 0$ s.t. for any optimal A^*,
$$\min_j P(V_j(A^*)) \ge p_{\min}.$$

Observe that the same conclusions work if we replace P by its empirical counterpart P_N. In particular, a version of centroid condition is also valid for the empirically optimal quantizer defined by (1). However, it is not true for a MOM based estimator in general.

Structure of the paper

- Section 2 is devoted to a high probability excess distortion bounds that hold in the case where a good guess on the localization radius of the optimal quantizer A^* is available. The result generalizes naturally several known bounds for the empirically optimal quantizer in separable Hilbert spaces.
- Section 3 contains our main results. We show that there is a consistent median-of-means based estimator that gives the sub-Gaussian performance under our minimal moment assumption provided that a good guess on $p_{\min}$ is given and $p_{\min}N \to \infty$. We also prove a minimax lower bound showing that our dependence on $p_{\min}$ and N is sharp up to constant factors in the special case of $\mathbb{R}^d$.
- Finally, Section 4 is devoted to the generalization of our main results. We show that it is possible to prove a slightly weaker bound using the procedure that does not require the knowledge of the parameters of P.
- Section 5 is devoted to discussions and some final remarks.

Notation. For $a, b \in \mathbb{R}$, we set $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$ and for two real valued functions f, g, we write $f \lesssim g$ iff there is an absolute constant $c > 0$ such that $f \le cg$. We set $f \simeq g$ if $f \lesssim g$ and $g \lesssim f$. Given a probability measure P let $P^{\otimes N}$ denote the measure which is N-times product of P. For the sake of simplicity, we always assume that $\log x$ is equal to $\log x \, \vee \, 1$. The indicator of an event A is denoted by $\mathbb{1}[A]$. We also use the standard $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$ notation as well as $\mathrm{KL}(P, Q)$ and $\mathrm{TV}(P, Q)$ for Kullback–Leibler divergence and Total Variation distance between two measures P and Q (see e.g., (Boucheron et al., 2013)). The support of a measure P is denoted by $\mathrm{supp}(P)$. For a normed space $(E, \|\cdot\|)$ let B_R denote the closed ball of radius R centered at the origin. To avoid the measurability issues we use the standard convention for the supremum of stochastic processes (see Paragraph 2 in (Talagrand, 2014)). Given the sample $X_1, \ldots, X_N$ sampled i.i.d. from P and a function $f : E \to \mathbb{R}$ we denote $P_N f = P_N f(X) = \frac{1}{N}\sum\limits_{i=1}^{N} f(X_i)$. In general, the symbol P_N will denote the empirical measure.

We are interested in $L_2(P_N)$ space, and the corresponding covering number of a functional class $\mathcal{G}$ will be denoted by $\mathcal{N}_2(\mathcal{G}, x, P_N)$, where x is the corresponding radius (see e.g., (Vershynin, 2016) for more details on covering numbers).

2 Simple Case: Known Magnitude of an Optimal Quantizer

In this section we provide our simplest result which can serve as a good illustration of the underlying techniques. In Sections 3 and 4 we will be focusing on sharpening our basic bound as well as weakening some of the assumptions.

We first show a simple bound which holds in the situations where a good guess on M is available (recall the property 4 and (4)). The result of Theorem 2.2 below can be seen as a strengthening of Theorem 11 in (Brownlees et al., 2015).

First, we observe that $M = M(P, k)$ defined above is not translation invariant. This means that if the distribution P of X is changed in a way such that we replace X by $X + c$, where $c \in E$ is a constant vector, the value of M may increase, while the clustering problem will remain the same. Therefore, we slightly redefine the quantity. Let $M = M(P, k)$ be a number such that

$$\|a - \mu\| \le M \quad \text{for all} \quad a \in A^*, \tag{5}$$

where $\mu = \mathbb{E}\, X$ and A^* is any particular optimal quantizer. Fortunately, due to e.g., (3) there is a very efficient way to estimate μ. One may split the sample of size N in two almost equal parts and estimate μ based on the first part. Therefore, for the sake of presentation, we assume that $\mu = 0$ in this section.

Remark 2.1. *It is important to note that the boundedness of the vectors in the finite set A^* has nothing in common with the boundedness of the observations $X_1, \ldots, X_N$. The latter can still be unbounded and the distribution P can be heavy-tailed.*

We proceed with the main result of this section.

Theorem 2.2. *Fix $\delta \in (0, 1)$. Let some M satisfying (5) be known and assume that $\mu = \mathbb{E}\, X = 0$. There is an estimator $\hat{A}_{M,\delta}$ that depends on M and δ such that, with probability at least $1 - \delta$,*

$$D(\hat{A}_{M,\delta}) - D(A^*) \lesssim M(M + \sqrt{\mathbb{E}\,\|X\|^2}) \left(\frac{k}{\sqrt{N}} + \sqrt{\frac{\log \frac{1}{\delta}}{N}} \right).$$

Let us now define the estimator that we use in Theorem 2.2. Notice that minimizing $\min_{a \in A} \|a - X\|^2$ with respect to A is equivalent to

$$l_A(X) \to \min, \qquad l_A(x) = \min_{a \in A} -2\langle x, a\rangle + \|a\|^2 .$$

Fix $1 \leq \ell \leq N$ and assume without loss of generality that $m = N/\ell$ is integer. Split the set $\{1, \ldots, N\}$ into ℓ blocks $I_1, \ldots, I_\ell$ of equal size such that $I_j = \{1 + (j-1)m, \ldots, jm\}$. For any real-valued function f and random variables $X_1, \ldots, X_N$ define

$$\mathrm{MOM}(f) = \mathrm{Median}\left(\frac{\ell}{N}\sum_{i \in I_1} f(X_i), \ldots, \frac{\ell}{N}\sum_{i \in I_\ell} f(X_i)\right).$$

Slightly abusing the notation we set

$$\mathcal{A}^k = \{A \subset E : |A| \leq k\} \quad \text{and} \quad \mathcal{A}^k_M = \{A \in \mathcal{A}^k : \max_{a \in A}\|a\| \leq M\}. \tag{6}$$

The estimator of Theorem 2.2. Define

$$\hat{A}_{M,\delta} = \arg\min_{A \in \mathcal{A}^k_M} \mathrm{MOM}(l_A),$$

with the number of blocks $\ell = 8\lceil \log \frac{2}{\delta} \rceil + 1$. If there are many minimizers, we may choose any of them.

The proof of Theorem 2.2 relies on the uniform bound for the median of means estimator. However, instead of restricting our attention to the medians only, we consider the *quantiles of means* (QOM). That is, for a given level $\alpha \in (0, 1)$,

$$\mathrm{QOM}_\alpha(f) = \mathrm{Quant}_\alpha\left(\frac{\ell}{N}\sum_{i \in I_1} f(X_i), \ldots, \frac{\ell}{N}\sum_{i \in I_\ell} f(X_i)\right),$$

where $\mathrm{Quant}_\alpha(x_1, \ldots, x_\ell) = x^{(\lceil \alpha\ell \rceil)}$, for $x^{(1)}, \ldots, x^{(\ell)}$ being a non-decreasing rearrangement of the original sequence. For the sake of simplicity, we always assume that $\ell\alpha$ is non-integer, such that the quantile is uniquely defined, and, in particular $\mathrm{Quant}_\alpha(x_1, \ldots, x_\ell) = -\mathrm{Quant}_{1-\alpha}(-x_1, \ldots, -x_\ell)$. It will be usually enough to assume that ℓ is not even which can be always achieved by adding at most one extra block. Obviously, $\mathrm{QOM}_{\frac{1}{2}}$ corresponds to the median of means.

Lemma 2.3. *Fix $\alpha \in (0, 1)$ and consider a separable class $\mathcal{F}$ of square integrable real-valued functions. Suppose, we have ℓ blocks and $\ell\alpha$ is a non-integer. It holds that, with probability at least $1 - e^{-\alpha^2\ell/2}$,*

$$\sup_{f \in \mathcal{F}}(\mathbb{E} f - \mathrm{QOM}_\alpha(f)) \lesssim \frac{1}{\alpha}\,\mathbb{E}\sup_{f \in \mathcal{F}}\left(\frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i)\right) + \frac{1}{\alpha^{1/2}}\sqrt{\sup_{f \in \mathcal{F}} \mathrm{Var}(f(X))\ \frac{\ell}{N}}, \tag{7}$$

as well as with probability at least $1 - e^{-(1-\alpha)^2\ell/2}$,

$$\sup_{f \in \mathcal{F}}(\mathrm{QOM}_\alpha(f) - \mathbb{E} f) \lesssim \frac{1}{1-\alpha}\,\mathbb{E}\sup_{f \in \mathcal{F}}\left(\frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i)\right) + \frac{1}{(1-\alpha)^{1/2}}\sqrt{\sup_{f \in \mathcal{F}} \mathrm{Var}(f(X))\ \frac{\ell}{N}}, \tag{8}$$

where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. Rademacher signs.

Remark 2.4. *In the case where α is fixed, we can take $\ell \simeq \log\frac{1}{\delta}$, so that with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}}|\mathbb{E} f - \mathrm{QOM}_\alpha(f)| \lesssim \mathbb{E}\sup_{f \in \mathcal{F}}\left(\frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i)\right) + \sqrt{\sup_{f \in \mathcal{F}} \mathrm{Var}(f(X))\ \frac{\log\frac{1}{\delta}}{N}},$$

where the first term represents the expectation of the empirical process, whereas the second term corresponds to a tail with the sub-Gaussian behavior. Compare this inequality with Talagrand's inequality for empirical processes, where the assumption $\sup_{f \in \mathcal{F}}|f(X)| \leq C$ almost surely is needed (see Chapter 12 in Boucheron et al. (2013)).

As noticed by Minsker (2018) (see equation (2.7)) an inequality similar to (7) of Lemma 2.3 for $\alpha = \frac{1}{2}$ follows from the proof of Theorem 2 in (Lecué, Lerasle and Mathieu, 2018). However, to the best of our knowledge, Lemma 2.3 in this form is not presented explicitly in the literature. We provide its proof in the appendix for the sake of completeness.

Proof of Theorem 2.2. **Step 1.** First, we provide the high probability part of the analysis. Observe that

$$\begin{aligned} D(\hat{A}_{M,\delta}) - D(A^*) &= \mathbb{E}\, l_{\hat{A}_{M,\delta}} - \mathbb{E}\, l_{A^*} \\ &\le \mathbb{E}\, l_{\hat{A}_{M,\delta}} - \mathrm{MOM}(l_{\hat{A}_{M,\delta}}) + \mathrm{MOM}(l_{A^*}) - \mathbb{E}\, l_{A^*} \\ &\le 2 \sup_{A \in \mathcal{A}_M^k} \left|\mathbb{E}\, l_A - \mathrm{MOM}(l_A)\right|, \end{aligned}$$

where we used $\mathrm{MOM}(l_{A^*}) \ge \mathrm{MOM}(l_{\hat{A}_{M,\delta}})$ since $A^* \in \mathcal{A}_M^k$. We have by Lemma 2.3 that, with probability at least $1-\delta$,

$$\sup_{A \in \mathcal{A}_M^k} \left|\mathbb{E}\, l_A(X) - \mathrm{MOM}(l_A)\right| \lesssim \mathbb{E} \sup_{A \in \mathcal{A}_M^k} \frac{1}{N}\sum_{i=1}^N \epsilon_i l_A(X_i) + \sqrt{\sup_{A \in \mathcal{A}_M^k} \mathrm{Var}(l_A(X)) \frac{\log \frac{1}{\delta}}{N}},$$

where $\epsilon_1, \dots, \epsilon_N$ are independent Rademacher signs. Here, we have for each $A \in \mathcal{A}_M^k$,

$$\begin{aligned} l_A(x)^2 &= (\|x - a_x\|^2 - \|x\|^2)^2 = (\|x - a_x\| - \|x\|)^2 (\|x - a_x\| + \|x\|)^2 \\ &\le \|a_x\|^2 (2\|x\| + \|a_x\|)^2, \end{aligned}$$

where $a_x \in \operatorname{Arg\,min}_{a \in A} \|x - a\|$. Then, since $\|a_x\| \le M$ for any x, we have

$$\mathrm{Var}(l_A(X)) \le \mathbb{E}\, l_A(X)^2 \lesssim M^2(M^2 + \mathbb{E}\|X\|^2).$$

Step 2. Note that $\hat{A}_{M,\delta}$ can consist of less than k points. However, in this case we can always add the copies of some of them and identify $\hat{A}_{M,\delta}$ with $(a_1, \dots, a_k)$. This does not change $l_{\hat{A}_{M,\delta}}$ and preserves the Voronoi partition of the space since the cells corresponding to the newly added points are empty. Finally, we estimate

$$\mathbb{E} \sup_{A \in \mathcal{A}_M^k} \frac{1}{N} \sum_{i=1}^N \epsilon_i l_A(X_i). \tag{9}$$

Consider the set $\mathcal{F}_{\mathcal{A}} = \{f_A : A \in \mathcal{A}_M^k\}$ of $\mathbb{R}^k$-valued functions such that for $A = \{a_1, \dots, a_k\},\ A \in \mathcal{A}_M^k$ we set

$$f_A(x) = \left(-2\langle x, a_1\rangle + \|a_1\|^2, \dots, -2\langle x, a_k\rangle + \|a_k\|^2\right).$$

For $\mathbf{c} \in \mathbb{R}^k$ let $\phi(\mathbf{c}) = \min\limits_{i \le k} c_i$. We obviously have $l_A(X) = \phi(f_A(X))$. Following the analysis of Section 3.2 in (Maurer, 2016) we have for any two $A = \{a_1, \dots, a_k\}$ and $B = \{b_1, \dots, b_k\}$ from $\mathcal{A}_M^k$,

$$|\phi(f_A(X_i)) - \phi(f_B(X_i))| \le \left\|\left(\|X_i - a_1\|^2 - \|X_i - b_1\|^2, \dots, \|X_i - a_k\|^2 - \|X_i - b_k\|^2\right)\right\|_2.$$

This allows us to use the ℓ_2-contraction to upper bound (9) with the quantity scaling linearly in k. To do so, we observe that Maurer's vector contraction inequality (Theorem 3 in (Maurer, 2016)) implies

$$\mathbb{E} \sup_{A \in \mathcal{A}_M^k} \frac{1}{N} \sum_{i=1}^N \epsilon_i l_A(X_i) \le \frac{\sqrt{2}}{N} \left(2\,\mathbb{E} \sup_{A \in \mathcal{A}_M^k} \sum_{i,j=1}^{N,k} \epsilon_{i,j} \langle X_i, a_j\rangle + \mathbb{E} \sup_{A \in \mathcal{A}_M^k} \sum_{i,j=1}^{N,k} \epsilon_{i,j} \|a_j\|^2 \right),$$

where $\epsilon_{i,j}$, $i = 1, \dots, N$, $j = 1, \dots, k$ are independent Rademacher signs, and $A = \{a_1, \dots, a_k\}$. We further have by Khintchine's inequality,

$$
\begin{aligned}
\mathbb{E} \sup_{A \in \mathcal{A}_M^k} \sum_{i,j=1}^{N,k} \epsilon_{i,j} \langle X_i, a_j \rangle &\le \sum_{j=1}^{k} \mathbb{E} \sup_{A \in \mathcal{A}_M^k} \left\langle \sum_{i=1}^{N} \epsilon_{i,j} X_i, a_j \right\rangle \le \sum_{j=1}^{k} \mathbb{E} \sup_{A \in \mathcal{A}_M^k} \left\| \sum_{i=1}^{k} \epsilon_{i,j} X_i \right\| \|a_j\| \\
&\le kM \max_{j \le k} \mathbb{E} \left\| \sum_{i=1}^{N} \epsilon_{i,j} X_i \right\| \le kM \max_{j \le k} \sqrt{\mathbb{E} \left\| \sum_{i=1}^{N} \epsilon_{i,j} X_i \right\|^2} \\
&\le kM \sqrt{\sum_{i=1}^{N} \mathbb{E} \|X_i\|^2},
\end{aligned}
$$

and also

$$
\mathbb{E} \sup_{A \in \mathcal{A}_M^k} \sum_{i,j=1}^{N,k} \epsilon_{i,j} \|a_j\|^2 \le \sum_{j=1}^{k} \mathbb{E} \sup_{A \in \mathcal{A}_M^k} \left| \sum_{i=1}^{N} \epsilon_{i,j} \right| \|a_j\|^2 \lesssim kM^2 \sqrt{N}.
$$

Finally, taking the expectation with respect to $X_1, \dots, X_N$ and using Jensen's inequality we obtain an analog of (2). That is,

$$
\mathbb{E} \sup_{A \in \mathcal{A}_M^k} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i l_A(X_i) \lesssim \frac{kM \left(\sqrt{\mathbb{E} \|X\|^2} + M \right)}{\sqrt{N}}.
$$

Combining the above bounds we prove the claim. □

It is by now well known that in our setup in the bounded case (e.g., when $\|X\| \le T$ almost surely) the right dependence of the excess distortion on the number of clusters is $\sqrt{k}$ up to logarithmic factors (Fefferman et al., 2016; Narayanan and Mitter, 2010). It is natural to ask if the same dependence is possible in our Theorem 2.2. First, observe that in the unbounded case, there are some complications. In particular, our parameter $M = M(P, k)$ can also depend on k. This means that the overall dependence of the excess distortion on k can be more complicated. Nevertheless, in the next section we show, among other things, that these improvements are possible and, in particular, the k-term will be replaced by the $\sqrt{k}$-term.

3 Towards Better Bounds Based on $p_{\min}$

This section is devoted to our main results. We prove almost optimal non-asymptotic bounds for k-means. Recall that if $\mathbb{E}\|X\|^2 < \infty$ we have for any optimal quantizer

$$
p_{\min} = \min_{a \in A^*} P(V_a) > 0,
$$

unless the support of P has less than k points. Notice that $p_{\min}$ controls the magnitude of the largest vector in A^*. Indeed, using the centroid condition, Jensen's inequality and the Cauchy–Schwarz inequality we have for any $a \in A^*$,

$$
\|a\| = \|\mathbb{E}[X | \, X \in V_a]\| = \frac{\|\mathbb{E}\, X\, \mathbb{1}[X \in V_a]\|}{P(V_a)} \le \frac{\mathbb{E}^{1/2}\|X\|^2}{\sqrt{P(V_a)}} \le \frac{\mathbb{E}^{1/2}\|X\|^2}{\sqrt{p_{\min}}}. \tag{10}
$$

This confirms that the mass of the lightest cluster of an optimal quantizer should affect the quality of any empirical quantizer.

Let us return to Example 1.2. In this case we have $k = 2$, $M \le \sqrt{N}$, $p_{\min} = \frac{1}{N}$, $\mathbb{E}\|X\|^2 \le 1$ and the bound (10) is tight. However, the bound of Theorem 2.2 is not tight anymore as it scales as $O(\sqrt{N})$. Indeed, Theorem 2.2 implies the bound $O\left(\frac{k(M^2 + M\sqrt{\mathbb{E}\|X\|^2})}{\sqrt{N}}\right)$ which is $O\left(\frac{k\,\mathbb{E}\|X\|^2}{p_{\min}\sqrt{N}}\right)$ whenever (10) is tight.

The challenging part is to get the optimal dependence on $p_{\min}$ and N in the excess distortion bound. In what follows, we show that the dependence $\Theta\left(\frac{1}{\sqrt{Np_{\min}}}\right)$ is achievable with respect to these parameters. The result of this form guarantees the consistency for sequences of distributions depending on N as long as $Np_{\min} \to \infty$ and the second moment is uniformly bounded. This extends the original asymptotic result of Pollard (1981) to the case where the distribution is allowed to change with N.

Suppose that we know the value of $p_{\min} > 0$ for at least one optimal quantizer. Denote for short, $P_N(V) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[X_i \in V]$. Naturally, we want to find a solution $\hat{A}$ such that the corresponding Voronoi cells are of measure at least $p_{\min} > 0$ which translates into $P_N(V_j) \geq p_{\min}/2$ due to concentration. It implies that each cell corresponding to $\hat{A}$ contains enough sample points, which corresponds to the so-called *constrained* k-means clustering. In $\mathbb{R}^d$ the algorithmic side of constrained clustering is well studied in the context of optimal transport and has numerous practical applications, see (Ng, 2000; Cuturi and Doucet, 2014; Genevay, Dulac-Arnold and Vert, 2019) and references therein. We have additional motivation to introduce $p_{\min}$ since this quantity appears naturally in the condition implying the so-called *fast rates* of the excess distortion in the bounded case (Levrard, 2015). Finally, recalling Example 1.2 we know that in any reasonable clustering problem $p_{\min} \gg \frac{1}{N}$ which means that the optimal solution A^* has enough observations in each cell. At the same time, we do not have such a natural preliminary guess on M.

As before, the number of blocks depends solely on the desired confidence level. The estimator of this section is translation invariant and the assumption $\mathbb{E}\,X = 0$ is not needed anymore. Our main result is the following theorem.

Theorem 3.1. *Fix $\delta \in (0,1)$. Suppose, $\min_{a\in A^*} P(V_a) \geq p_{\min}$ for some optimal quantizer A^*. There is an estimator $\hat{A}_{\delta,p_{\min}}$ that depends on $p_{\min}$ and δ such that, with probability at least $1-\delta$,*

$$D(\hat{A}_{\delta,p_{\min}}) - D(A^*) \lesssim \mathbb{E}\|X-\mu\|^2 \left((\log N)^2\sqrt{\frac{k}{Np_{\min}}} + \sqrt{\frac{\log\frac{1}{\delta}}{Np_{\min}}}\right).$$

Let us now present our estimator.

The estimator of Theorem 3.1. We set

$$\hat{A}_{p_{\min},\delta} = \operatorname*{arg\,min}_{\substack{A\in\mathcal{A}^k \\ \min\limits_{a\in A} P_N(V_a)\geq p_{\min}/2}} \mathrm{MOM}(l_A), \tag{11}$$

with the number of blocks $\ell = 12\lceil\log\frac{6}{\delta}\rceil + 1$.

The idea behind this estimator is quite natural: we guarantee the robustness by using the MOM principle and by restricting our attention only on the cells containing enough points. As already mentioned, this is essentially a robust version of the constrained k-means quantizer introduced in Ng (2000).

We introduce several technical results that together will lead us to the proof of Theorem 3.1 at the end of this section. As before, since the estimator we consider is translation invariant, we can assume that $\mathbb{E}\,X = 0$ in the proof without loss of generality. As previously, our main tool is the concentration of MOM for a suitably chosen subset of $\{l_A : A \in \mathcal{A}^k\}$. We show that the restriction $P_N(V_j) \geq p_{\min}/2$ in (11) implies a convenient bound for the vectors in the resulting empirical quantizer. Let us define the following class of quantizers:

$$\mathcal{A}^k_{M,m} = \left\{A \in \mathcal{A}^k : \min_{a\in A}\|a\| \leq m,\ \max_{a\in A}\|a\| \leq M\right\}, \quad 0 < m \leq M.$$

The following lemma says that with high probability all the solutions corresponding to $\hat{A}_{\delta,p_{\min}}$ are bounded which is, of course, natural in view of the proof of Theorem 2.2. However, the key technical observation is that we also need to control the smallest norm by saying that there is at least one center in $\hat{A}_{\delta,p_{\min}}$ which is relatively close to the actual expectation. Surprisingly, in order to show this we do not have to use any uniform results that hold simultaneously for the entire class $\mathcal{A}^k$. Therefore, we have the following property.

Lemma 3.2. *With probability at least $1 - e^{-\ell/12}$, it holds that simultaneously for all $A \in \mathcal{A}^k$ such that $MOM(l_A) \le 0$,*

$$\min_{a \in A} \|a\| \le 4\sqrt{2\,\mathbb{E}\|X\|^2}.$$

Moreover, with probability at least $1 - \left(e^{-\ell/12} + e^{-Np_{\min}/12}\right)$, it holds that

$$\hat{A}_{\delta,p_{\min}} \in \mathcal{A}^k_{M,m} \quad \textit{for} \quad m = 4\sqrt{2\,\mathbb{E}\|X\|^2} \quad \textit{and} \quad M = 10\sqrt{\frac{\mathbb{E}\|X\|^2}{p_{\min}}}.$$

Remark 3.3. *Note that the first statement of the above lemma gives us a prior bound on the excess distortion $D(\hat{A}_{\delta,p_{\min}}) - D(A^*)$. Indeed, since $\ell \ge 12 \log \frac{1}{\delta}$, with probability at least $1 - \delta$, we have $\min_{a \in \hat{A}_{\delta,p_{\min}}} \|a\| \le m$, thus*

$$D(\hat{A}_{\delta,p_{\min}}) - D(A^*) \le D(\hat{A}_{\delta,p_{\min}}) = \mathbb{E} \min_{a \in \hat{A}_{\delta,p_{\min}}} \|a - X\|^2 \le \mathbb{E}(m + \|X\|)^2 \lesssim \mathbb{E}\|X\|^2. \tag{12}$$

Before going to the proof of Lemma 3.2, let us state the following trivial result on empirical quantiles. We postpone its proof to Appendix.

Lemma 3.4. *Let $\xi_1, \dots, \xi_\ell$ be i.i.d. random values such that $\mathbb{E}\,\xi < \infty$ and $\xi \ge 0$ almost surely. Then for any $0 < \alpha < 1$ we have*

$$\mathbb{P}\left(\mathrm{Quant}_{1-\alpha}(\xi_1, \dots, \xi_\ell) \ge \frac{2}{\alpha}\,\mathbb{E}\,\xi\right) \le \exp\left(-\frac{\alpha\ell}{6}\right).$$

Proof of Lemma 3.2. **Step 1.** First, let us prove the bound on the minimal norm. Consider $A \in \mathcal{A}^k$ such that $\min_{a \in A}\|a\| \ge m$. Then for any $x \in B_{m/2}$ (recall that $B_{m/2}$ is a ball of radius $m/2$ centered at the origin) it holds that $\min_{a \in A}\|a - x\| \ge \frac{m}{2}$, thus for all $x \in E$,

$$l_A(x) = \min_{a \in A}\|a - x\|^2 - \|x\|^2 \ge \frac{m^2}{4}\,\mathbb{1}\left[\|x\| \le \frac{m}{2}\right] - \|x\|^2 = \frac{m^2}{4} - \left(\frac{m^2}{4}\,\mathbb{1}\left[\|x\| > \frac{m}{2}\right] + \|x\|^2\right),$$

and hence

$$\mathrm{MOM}(l_A) \ge \frac{m^2}{4} - \mathrm{MOM}\left(\frac{m^2}{4}\,\mathbb{1}\left[\|X\| > \frac{m}{2}\right] + \|X\|^2\right).$$

According to Lemma 3.4, with probability at least $1 - e^{-\ell/12}$, it holds that

$$\mathrm{MOM}\left(\frac{m^2}{4}\,\mathbb{1}\left[\|X\| > \frac{m}{2}\right] + \|x\|^2\right) \le 4\,\mathbb{E}\left(\frac{m^2}{4}\,\mathbb{1}\left[\|X\| > \frac{m}{2}\right] + \|X\|^2\right) < 8\,\mathbb{E}\|X\|^2,$$

thus, simultaneously for all $A \in \mathcal{A}^k$ satisfying $\min_{a \in A}\|a\| \ge m$ we have

$$\mathrm{MOM}(l_A) > \frac{m^2}{4} - 8\,\mathbb{E}\|X\|^2 = 0.$$

In particular, since $\{0\}$ is always one of the potential candidates for $\hat{A}_{\delta,p_{\min}}$ and $\mathrm{MOM}(l_{\{0\}}) = 0$, we have

$$\min_{a \in \hat{A}_{\delta,p_{\min}}} \|a\| < m.$$

Step 2. Now consider $A \in \mathcal{A}^k$ such that there is $b \in A$ with $\|b\| \le m$. It is easy to see that for any $a \in A$, $x \in V_a$ implies $\|a - x\| \le \|b - x\|$, thus

$$\|x\| \ge \frac{\|a\| - \|b\|}{2}. \tag{13}$$

Assume $\|a\| > M$ for some $a \in A$, then $\|x\| > \frac{M-m}{2}$ for any $x \in V_a$. Hence,

$$P_N(V_a) \leq P_N\left(\mathbb{1}\left[\|X\| > (M-m)/2\right]\right).$$

At the same time,

$$P\left(\|X\| > (M-m)/2\right) < \frac{4\,\mathbb{E}\|X\|^2}{(M-m)^2} \leq \frac{p_{\min}}{4}.$$

Now Chernoff's bound yields that, with probability at least $1 - e^{-Np_{\min}/12}$,

$$P_N\left(\|X\| > (M-m)/2\right) \leq P\left(\|X\| > (M-m)/2\right) + \frac{p_{\min}}{4} < \frac{p_{\min}}{2}.$$

This implies $P_N(V_a) < \frac{p_{\min}}{2}$, what means that none of such A can be chosen by our estimator. By the union bound, we finally get that $\hat{A}_{\delta, p_{\min}} \in \mathcal{A}^k_{M,m}$ with probability at least $1 - \left(e^{-\ell/12} + e^{-Np_{\min}/12}\right)$. □

The next step is to provide a uniform concentration of MOM over a class of quantizers $\mathcal{A}^k_{M,m}$. First, we estimate the L_2-diameter and covering numbers of the functional class corresponding to $\mathcal{A}^k_{M,m}$:

$$\mathcal{F}^k_{M,m} = \left\{l_A : A \in \mathcal{A}^k_{M,m}\right\}.$$

Lemma 3.5. *For any distribution P and any finite set $A \subset \mathcal{A}^k_{M,m}$ it holds that*

$$\sum_{a \in A} \|a\|^2 P(V_a) \leq 2m^2 + 8\,\mathbb{E}\|X\|^2, \tag{14}$$

and

$$\mathbb{E}\, l_A^2(X) \leq 4M^2\left(m^2 + 6\,\mathbb{E}\|X\|^2\right). \tag{15}$$

Proof. Fix $A \in \mathcal{A}^k$ and let $b \in A$ be such that $\|b\| \leq m$. Then for any $a \in A$ and $x \in V_a$ it holds from (13) that $\|a\| \leq \|b\| + 2\|x\| \leq m + 2\|x\|$. Therefore,

$$\sum_{a \in A} P(V_a)\|a\|^2 = \mathbb{E}\sum_{a \in A} \mathbb{1}[X \in V_a]\|a\|^2 \leq \mathbb{E}\left(m + 2\|X\|\right)^2 \leq 2m^2 + 8\,\mathbb{E}\|X\|^2.$$

Further, we easily have using (14)

$$\begin{aligned}\mathbb{E}\, l_A^2(X) &\leq \sum_{a \in A} P(V_a)\,\mathbb{E}\left[\left(\|a\|^2 + 2\|a\| \cdot \|X\|\right)^2 | X \in V_a\right] \\ &\leq 2M^2\left(\sum_{a \in A} P(V_a)\|a\|^2 + 4\,\mathbb{E}\|X\|^2\right) \\ &\leq 4M^2\left(m^2 + 6\,\mathbb{E}\|X\|^2\right).\end{aligned}$$

□

The next technical lemma is one of our main contributions which can be of independent interest. It states the upper bounds on $\log \mathcal{N}_2\left(\mathcal{F}^k_{M,m}, t, P_N\right)$ for general separable Hilbert spaces as well as for $\mathbb{R}^d$. The question on sharp bounds on covering numbers for the classes of functions indexed by $\mathcal{A}^k$ appeared naturally in the analysis of k-means clustering in the uniformly bounded case. The way to do it is to estimate the so-called fat-shattering dimension (Narayanan and Mitter, 2010; Fefferman, Mitter and Narayanan, 2016) or to decompose the covering numbers as a product of k covering numbers of some simpler classes indexed by $\mathcal{A}$ as in

(Brownlees, Joly and Lugosi, 2015; Foster and Rakhlin, 2019). Furthermore, in the special case of $\mathbb{R}^d$, the analysis can be done via the computation of Pollard's pseudodimension (Bachem et al., 2017). Unfortunately, it seems that these approaches are better tuned to the analysis of uniformly bounded distributions or to the finite dimensional case. Our approach will be based on direct computations of these covering numbers via the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), and in $\mathbb{R}^d$ our analysis will remove the unnecessary logarithmic factors appearing in some previous works in the bounded case.

Lemma 3.6. *For any $0 < t < \operatorname{diam}_2\left(\mathcal{F}^k_{M,m}, P_N\right)$ it holds that*

$$\log \mathcal{N}_2\left(\mathcal{F}^k_{M,m}, t, P_N\right) \lesssim \frac{kM^2\left(m^2 + P_N\|X\|^2\right)\log N}{t^2} \log \frac{M\left(m + P_N^{1/2}\|X\|^2\right)}{t}.$$

Moreover, if $E = \mathbb{R}^d$, then

$$\log \mathcal{N}_2\left(\mathcal{F}^k_{M,m}, t, P_N\right) \lesssim kd \log \frac{M\left(m + P_N^{1/2}\|X\|^2\right)}{t}.$$

The proof of this fact is differed to Appendix. With this result in mind we are ready to show the following uniform bound.

Lemma 3.7. *Fix the number of blocks ℓ and assume that ℓ divides N. Then for any fixed $\alpha \in (0,1)$ with $\ell\alpha$ being non-integer, with probability at least $1 - e^{-\alpha^2\ell/2}$,*

$$\sup_{A \in \mathcal{A}^k_{M,m}} \left(\mathbb{E}\, l_A(X) - \mathrm{QOM}_\alpha(l_A)\right) \lesssim M\left(m + \mathbb{E}^{1/2}\|X\|^2\right)\left(\frac{(\log N)^2}{\alpha}\sqrt{\frac{k}{N}} + \sqrt{\frac{\ell}{\alpha N}}\right), \tag{16}$$

as well as, with probability at least $1 - e^{-(1-\alpha)^2\ell/2}$,

$$\sup_{A \in \mathcal{A}^k_{M,m}} \left(\mathrm{QOM}_\alpha(l_A) - \mathbb{E}\, l_A(X)\right) \lesssim M\left(m + \mathbb{E}^{1/2}\|X\|^2\right)\left(\frac{(\log N)^2}{1-\alpha}\sqrt{\frac{k}{N}} + \sqrt{\frac{\ell}{(1-\alpha)N}}\right). \tag{17}$$

Remark 3.8. *It is interesting to compare the bound (16) with the result one can obtain via the vector contraction inequalities which were used in the proof of Theorem 2.2. First, we notice that the leading term $M^2 + M\,\mathbb{E}^{1/2}\|X\|^2$ is replaced in (16) by a much better term $Mm + M\,\mathbb{E}^{1/2}\|X\|^2$.*

Proof. By Lemma 2.3, with probability at least $1 - e^{-\alpha^2\ell/2}$, it holds that

$$\sup_{A \in \mathcal{A}^k_{M,m}} \left(\mathbb{E}\, l_A(X) - \mathrm{QOM}_\alpha(l_A)\right) \lesssim \mathbb{E} \sup_{A \in \mathcal{A}^k_{M,m}} \left(\frac{1}{\alpha N}\sum_{i=1}^N \epsilon_i l_A(X_i)\right) + \sqrt{\sup_{A \in \mathcal{A}^k_{M,m}} \mathrm{Var}(l_A(X))\frac{\ell}{\alpha N}}.$$

Now we are going to bound the first term of the right-hand side for a fixed sample $X_1, \ldots, X_N$ using the Dudley integral argument. It follows from (15) applied to the empirical distribution P_N that

$$\operatorname{diam}_2\left(\mathcal{F}^k_{M,m}, P_N\right) \le 10M\sigma_N, \quad \text{where} \quad \sigma_N = m + P_N^{1/2}\|X\|^2,$$

thus, the standard Dudley integral argument (e.g., Lemma A.3 in (Srebro et al., 2010)) together with Lemma 3.6 ensure the following bound on the Rademacher averages of l_A,

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{A \in \mathcal{A}^k_{M,m}} \left(\frac{1}{N}\sum_{i=1}^N \epsilon_i l_A(X_i)\right) &\lesssim \beta + \frac{1}{\sqrt{N}}\int_\beta^{\operatorname{diam}_2\left(\mathcal{F}^k_{M,m}, P_N\right)} \sqrt{\log \mathcal{N}_2\left(\mathcal{F}^k_{M,m}, t, P_N\right)}dt \\
&\lesssim \beta + \frac{1}{\sqrt{N}}\int_\beta^{10M\sigma_N} \frac{M\sigma_N}{t}\sqrt{k\log N \log\frac{M\sigma_N}{t}}dt \\
&\lesssim \beta + M\sigma_N\sqrt{\frac{k\log N}{N}}\log^{3/2}\left(\frac{M\sigma_N}{\beta}\right)
\end{aligned}$$

Further, choosing $\beta = M\sigma_N\sqrt{\frac{k}{N}}$ we have

$$\mathbb{E}_\epsilon \sup_{A\in\mathcal{A}^k_{M,m}} \left(\frac{1}{N}\sum_{i=1}^N \epsilon_i l_A(X_i)\right) \lesssim M\sigma_N\sqrt{\frac{k}{N}\log N \log^3\frac{N}{k}} \le M\sigma_N(\log N)^2\sqrt{\frac{k}{N}}.$$

Finally, (15) implies

$$\mathrm{Var}(l_A(X)) \le \mathbb{E}\, l_A^2(X) \lesssim M^2\left(m^2 + \mathbb{E}\|X\|^2\right),$$

thus we conclude that, with probability at least $1 - e^{-\alpha^2\ell/2}$,

$$\begin{aligned}\sup_{A\in\mathcal{A}^k_{M,m}} \left(\mathbb{E}\, l_A(X) - \mathrm{QOM}_\alpha(l_A)\right) &\lesssim \frac{\mathbb{E}\, M\sigma_N(\log N)^2}{\alpha}\sqrt{\frac{k}{N}} + \sqrt{\sup_{A\in\mathcal{A}^k_{M,m}} \mathrm{Var}(l_A(X))\frac{\ell}{\alpha N}} \\ &\lesssim M\left(m + \mathbb{E}^{1/2}\|X\|^2\right)\left(\frac{(\log N)^2}{\alpha}\sqrt{\frac{k}{N}} + \sqrt{\frac{\ell}{\alpha N}}\right).\end{aligned}$$

Inequality (17) can be similarly derived from (8). □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. In order to finish the proof we need to combine several results. Let us fix some optimal quantizer A^*, satisfying $P(V_a) \ge p_{\min}$ for all $a \in A^*$. We derive the bound on the union of the events below:

- by Chernoff's and the union bounds, it holds with probability at least $1 - ke^{-Np_{\min}/8}$ that for any $a \in A^*$

$$P_N(V_a) \ge \frac{P(V_a)}{2} \ge \frac{p_{\min}}{2},$$

  hence A^* is in the set of possible solutions;
- by Lemma 3.2, with probability at least $1 - e^{-\ell/12} - e^{-Np_{\min}/12}$, we have $\hat{A}_{\delta,p_{\min}} \in \mathcal{A}^k_{M,m}$ with $M = 10\sqrt{\frac{\mathbb{E}\|X\|^2}{p_{\min}}}$ and $m = 4\sqrt{2\,\mathbb{E}\|X\|^2}$. In addition, a similar property can be derived for A^*. Indeed, if $\min_{a\in A^*}\|a\| > m$, then

$$\mathbb{E}\, l_{A^*}(X) \ge \frac{m^2}{4}P\left(\|X\| \le \frac{m}{2}\right) - \mathbb{E}\|X\|^2 \ge 8\,\mathbb{E}\|X\|^2\left(1 - \frac{1}{8}\right) - \mathbb{E}\|X\|^2 > 0 = \mathbb{E}\, l_{\{0\}}(X).$$

  This contradicts the optimality of A^*. Now assume $\min_{a\in A^*}\|a\| \le m$, but $\|a\| > M$ for some $a \in A^*$. Arguing as in the proof of Lemma 3.2, we obtain

$$P(V_a) \le P\left(\|X\| > \frac{M-m}{2}\right) < \frac{4\,\mathbb{E}\|X\|^2}{(M-m)^2} \le \frac{p_{\min}}{4}.$$

  This contradicts the lower bound $P(V_a) \ge p_{\min}$;
- by Lemma 3.7, taking M and m as above, we have that, with probability at least $1 - 2e^{-\ell/8}$,

$$\sup_{A\in A_{M,m}} |\mathbb{E}\, l_A - \mathrm{MOM}(l_A)| \lesssim \mathbb{E}\|X\|^2\left((\log N)^2\sqrt{\frac{k}{Np_{\min}}} + \sqrt{\frac{\ell}{Np_{\min}}}\right).$$

All three assertions take place with probability at least $1 - 3e^{-\ell/12} - (k+1)e^{-Np_{\min}/12}$. Suppose for a moment that $Np_{\min} \ge 12\log\frac{2(k+1)}{\delta}$. Then, additionally, due to the choice $\ell = 12\lceil\log\frac{6}{\delta}\rceil + 1$, we have that the total probability is at least $1-\delta$. Since we know that on this event $\hat{A}_{\delta,p_{\min}}, A^* \in A_{M,m}$ and $\mathrm{MOM}(l_{\hat{A}_{\delta,p_{\min}}}) \le \mathrm{MOM}(l_{A^*})$, we have

$$\begin{aligned} D(\hat{A}_{\delta,p_{\min}}) - D(A^*) &= \mathbb{E}\, l_{\hat{A}_{\delta,p_{\min}}} - \mathbb{E}\, l_{A^*} \\ &\le \mathbb{E}\, l_{\hat{A}_{\delta,p_{\min}}} - \mathrm{MOM}(l_{\hat{A}_{\delta,p_{\min}}}) - \mathbb{E}\, l_{A^*} + \mathrm{MOM}(l_{A^*}) \\ &\lesssim \mathbb{E}\|X\|^2 \left((\log N)^2 \sqrt{\frac{k}{Np_{\min}}} + \sqrt{\frac{\log \frac{1}{\delta}}{Np_{\min}}} \right). \end{aligned}$$

Finally, consider the case $Np_{\min} < 12 \log \frac{2(k+1)}{\delta}$. According to (12) one has, with probability at least $1-\delta$,

$$D(\hat{A}_{\delta,p_{\min}}) - D(A^*) \lesssim \mathbb{E}\|X\|^2 \lesssim \mathbb{E}\|X\|^2 \left((\log N)^2 \sqrt{\frac{k}{Np_{\min}}} + \sqrt{\frac{\log \frac{1}{\delta}}{Np_{\min}}} \right).$$

The claim follows. □

Finally, we present an analog of Theorem 3.1 in $\mathbb{R}^d$. We are able to completely remove the $(\log N)$-factor by making d appear in the bound. First, we need the following simple result.

Corollary 3.9. *For any $\alpha \in (0,1)$ with $\ell\alpha$ being non-integer, we have, with probability at least $1 - e^{-\alpha^2 \ell/2}$,*

$$\sup_{A \in \mathcal{A}^k_{M,m}} \left(\mathbb{E}\, l_A(X) - \mathrm{QOM}_\alpha(l_A) \right) \lesssim M \left(m + \mathbb{E}^{1/2}\|X\|^2 \right) \left(\frac{1}{\alpha} \sqrt{\frac{kd}{N}} + \sqrt{\frac{\ell}{\alpha N}} \right).$$

as well as, with probability at least $1 - e^{-(1-\alpha)^2 \ell/2}$,

$$\sup_{A \in \mathcal{A}^k_{M,m}} \left(\mathrm{QOM}_\alpha(l_A) - \mathbb{E}\, l_A(X) \right) \lesssim M \left(m + \mathbb{E}^{1/2}\|X\|^2 \right) \left(\frac{1}{1-\alpha} \sqrt{\frac{kd}{N}} + \sqrt{\frac{\ell}{(1-\alpha) N}} \right).$$

Proof. Using the Dudley integral argument again we have

$$\begin{aligned} \mathbb{E}_\epsilon \sup_{A \in \mathcal{A}^k_{M,m}} \left(\frac{1}{N} \sum_{i=1}^N \epsilon_i l_A(X_i) \right) &\lesssim \frac{1}{\sqrt{N}} \int_0^{\mathrm{diam}_2\left(\mathcal{F}^k_{M,m}, P_N\right)} \sqrt{\log \mathcal{N}_2 \left(\mathcal{F}^k_{M,m}, t, P_N \right)} dt \\ &\lesssim \frac{1}{\sqrt{N}} \int_0^{10 M \sigma_N} \sqrt{kd \log \left(\frac{M \sigma_N}{t} \right)} dt \\ &\lesssim M \sigma_N \sqrt{\frac{kd}{N}}. \end{aligned}$$

The rest of the proof is exactly the same as for Lemma 3.7. □

With this result in mind, we can immediately prove our second main result.

Theorem 3.10. *Consider the case of $\mathbb{R}^d$ with the Euclidean distance. Fix $\delta \in (0,1)$. Suppose, $\min_{a \in A^*} P(V_a) \ge p_{\min}$ for some optimal quantizer A^*. The same estimator $\hat{A}_{\delta,p_{\min}}$ satisfies, with probability at least $1-\delta$,*

$$D(\hat{A}_{\delta,p_{\min}}) - D(A^*) \lesssim \mathbb{E}\|X - \mu\|^2 \sqrt{\frac{kd + \log \frac{1}{\delta}}{Np_{\min}}}.$$

Proof. The proof repeats the same lines of the proof of Theorem 3.1. The only difference is that the bound of Lemma 3.7 is replaced by the bound of Corollary 3.9. □

3.1 A lower bound with $p_{\min}$

Here we study the question of minimax lower bounds for the problem considered above. The lower bounds for the excess distortion appeared first in (Bartlett et al., 1998) for the bounded case ($\|X\| \leq 1$ almost surely), where they show a minimax lower bound of order $\Omega\left(\sqrt{\frac{k^{1-4/d}}{N}}\right)$. Furthermore, Linder (2002) recovers this bound for constant d and $k \geq 3$, while Antos (2005) shows the same bound for $k = 2$ but only for empirically optimal quantizers. Below, we focus on how the mass of the lightest cluster affects the excess distortion in the unbounded case. We extend the construction of Linder (2002) to derive a bound that confirms that the dependence on $p_{\min}$ and N in Theorem 3.10 is sharp.

Fix $k = 4$ and $d = 1$. Consider a class of probability measures on $\mathbb{R}$,

$$\mathcal{P}(p_{\min}, \sigma) = \left\{P : \mathbb{E}\, X^2 \leq \sigma^2 \text{ and there is } A^* \in \arg\min D(A, P) \text{ such that } \min_{a \in A^*} P(V_a) \geq p_{\min}\right\},$$

i.e., the probability measures that have an optimal quantizer based on k points such that the probability of X falling into each Voronoi cell under P is at least $p_{\min}$. Theorem 3.10 implies that there is an estimator $\hat{A}_N$ based on the i.i.d. sample $X_1, \ldots, X_N$, such that for any $P \in \mathcal{P}(p_{\min}, \sigma)$, we have with probability at least (say) 0.99,

$$D(\hat{A}_N, P) - D(A^*, P) \lesssim \sigma^2 \sqrt{\frac{1}{N p_{\min}}},$$

where the probability of the event is measured with respect to the joint distribution $\mathbb{P} = P^{\otimes N}$. The following result shows that when d and k are constants, the result is sharp up to a constant factor.

Theorem 3.11. *Under the notation introduced above let $\sigma > 0$, $p_{\min} \leq 1/10$. Then, for any empirically designed quantizer $\hat{A}_N$ there is a distribution $P \in \mathcal{P}(p_{\min}, \sigma)$, such that, with probability at least $\frac{1}{4}$,*

$$D(\hat{A}_N, P) - D(A^*, P) \geq \frac{\sigma^2}{80} \sqrt{\frac{1}{N p_{\min}}}.$$

Let us first present a heuristic argument showing the validity of Theorem 3.11. For $p \in (0, 1/2)$, $\delta \in (-1/2, 1/2)$ consider a distribution $P_{p,\delta}$ supported on five points

$$P_{p,\delta}(X = -\tfrac{1}{2}p^{-1/2}) = P_{p,\delta}(X = -p^{-1/2}) = \frac{p(1-\delta)}{4}, \qquad P_{p,\delta}(X = 0) = 1 - p,$$
$$P_{p,\delta}(X = \tfrac{1}{2}p^{-1/2}) = P_{p,\delta}(X = p^{-1/2}) = \frac{p(1+\delta)}{4}.$$

We have $\mathbb{E}\, X^2 = 5/8$. Obviously, we can rescale these values, so it is enough to consider the case $\sigma^2 = 5/8$. It is easy to see that for $\delta > 0$ the optimal quantizer is $A^* = (0, \frac{1}{2}p^{-1/2}, p^{-1/2}, -\frac{3}{4}p^{-1/2})$ with the distortion

$$D(A^*, P_{p,\delta}) = \frac{p(1-\delta)}{2}\left(\frac{p^{-1/2}}{4}\right)^2 = \frac{1-\delta}{32}.$$

For $\delta = \frac{1}{\sqrt{Np}}$, the number of points on the negative side will be greater with constant probability (see p. 27 in (Linder, 2002)). In such a case, the empirically optimal quantizer must be $\hat{A} = (0, -p^{-1/2}, -\frac{1}{2}p^{-1/2}, (\frac{1}{2}+a)p^{-1/2})$, where a is some value between 0 and $\frac{1}{2}$. Thus, the distortion of any empirically optimal quantizer will be at least

$$D(\hat{A}, P_{p,\delta}) \geq \frac{p(1+\delta)}{2}\left(\frac{p^{-1/2}}{4}\right)^2 = \frac{1+\delta}{32},$$

which implies

$$D(\hat{A}, P_{p,\delta}) - D(A^*, P_{p,\delta}) \geq \frac{\delta}{16} = \frac{1}{64}\frac{1}{\sqrt{Np}}.$$

However, this only touches the empirically optimal quantizer. The proof of the minimax bound relies on a standard reduction to hypothesis testing.

Proof of Theorem 3.11. As pointed above, we can fix $\sigma^2 = 5/8$ without loss of generality. Set $p = 4p_{\min} \leq 2/5$. Then we have that $P_{p,\delta}, P_{p,-\delta} \in \mathcal{P}(p_{\min}, 5/8)$. Denote, $P_1 = P_{p,\delta}$ and $P_2 = P_{p,-\delta}$. The Kullback-Leibler divergence between the two satisfies

$$\mathrm{KL}(P_1, P_2) = p\delta \log \frac{1+\delta}{1-\delta} \leq p\delta^2 \,.$$

Using Pinsker's inequality and additivity of the KL-divergence for product measures (see e.g., (Boucheron et al., 2013)) we have

$$\mathrm{TV}(P_1^{\otimes N}, P_2^{\otimes N}) \leq \sqrt{\frac{N}{2}\mathrm{KL}(P_1, P_2)} \leq \sqrt{Np\delta^2} = 1/2,$$

where we choose $\delta = 1/(4\sqrt{Np})$. Below, we only consider the distributions $P \in \{P_1, P_2\}$ instead of the whole class $\mathcal{P}(p_{\min}, \sqrt{5/8})$. Consider an empirical quantizer $\hat{A}_N = \hat{A}_N(X_1, \ldots, X_N)$ that takes only the values $\{A_1, A_2\}$, where

$$A_1 = (-\tfrac{3}{4}p^{-1/2}, 0, \tfrac{1}{2}p^{-1/2}, p^{-1/2}) \quad \text{and} \quad A_2 = (-p^{-1/2}, -\tfrac{1}{2}p^{-1/2}, 0, \tfrac{3}{4}p^{-1/2}).$$

Let $\Omega_1 \subset \mathbb{R}^N$ is the set where $\hat{A}_N = A_1$, and $\hat{A}_N = A_2$ elsewhere. Since $\mathrm{TV}(P_1^{\otimes N}, P_2^{\otimes N}) \leq 1/2$, we have

$$\max_{j=1,2} P_j^{\otimes N}\{\hat{A}_N \neq A_j\} = \max(1 - P_1^{\otimes N}(\Omega_1), P_2^{\otimes N}(\Omega_1)) \geq \max(1/2 - P_2^{\otimes N}(\Omega_1), P_2^{\otimes N}(\Omega_1)) \geq 1/4.$$

Notice that under $\mathbb{P}_j$ the event $\hat{A}_N \neq A_j$ corresponds to the distortion

$$D(A_1, P_2) = D(A_2, P_1) = \frac{p(1+\delta)}{2}\left(\frac{p^{-1/2}}{4}\right)^2 = \frac{1+\delta}{32},$$

whereas the minimal distortion is

$$D(A_j, P_j) = \frac{p(1-\delta)}{2}\left(\frac{p^{-1/2}}{4}\right)^2 = \frac{1-\delta}{32}\,.$$

Since $\delta = 1/(4\sqrt{Np})$, $p = 4p_{\min}$, and $\sigma^2 = 5/8$ the result follows.

It remains to show why only $\hat{A}_N \in \{A_1, A_2\}$ matters. For an arbitrary $\hat{A}_N$, the corresponding Voronoi cells could be one of the following:

1. $\big\{\{-p^{-1/2}, -\frac{1}{2}p^{-1/2}\}, \{0\}, \{\frac{1}{2}p^{-1/2}\}, \{p^{-1/2}\}\big\}$,
2. $\big\{\{-p^{-1/2}\}, \{-\frac{1}{2}p^{-1/2}, 0\}, \{\frac{1}{2}p^{-1/2}\}, \{p^{-1/2}\}\big\}$,
3. $\big\{\{-p^{-1/2}\}, \{-\frac{1}{2}p^{-1/2}\}, \{0, \frac{1}{2}p^{-1/2}\}, \{p^{-1/2}\}\big\}$,
4. $\big\{\{-p^{-1/2}\}, \{-\frac{1}{2}p^{-1/2}\}, \{0\}, \{\frac{1}{2}p^{-1/2}, p^{-1/2}\}\big\}$.

Denote by $\tilde{A}_N$ an empirical quantizer such that it equals to A_1 in the cases 1. and 2., and equals to A_2 in the cases 3. and 4. Let us show case by case, that the distortion of $\tilde{A}_N$ is smaller under either measure.

1. This case is trivial: by the centroid condition under both measures the optimal center for the cluster $\{-p^{-1/2}, -\frac{1}{2}p^{-1/2}\}$ is $-\frac{3}{4}p^{-1/2}$, which corresponds to A_1.
2. It is easy to calculate that the minimal distortion of a cluster on two points a, b with probabilities q, r, respectively, is $(a-b)^2\frac{qr}{q+r}$. Therefore, using only the distortion on $\{-\frac{1}{2}p^{-1/2}, 0\}$,

$$D(\hat{A}_N, P_1) \geq \frac{1}{4p}\frac{p(1-\delta)(1-2p)}{p(1-\delta) + 4(1-2p)} > \frac{1-\delta}{32} = D(A_1, P_1),$$

   where the second inequality follows from $p \leq 2/5$. Using additionally $\delta < 1$, we have as well

$$D(\hat{A}_N, P_2) \geq \frac{1}{4p}\frac{p(1+\delta)(1-2p)}{p(1+\delta) + 4(1-2p)} > \frac{1+\delta}{32} = D(A_1, P_2)\,.$$

Due to the symmetry, case 3. is similar to case 2. and case 4. is similar to case 1. We conclude that we always have $D(\hat{A}_N, P_j) \geq D(\tilde{A}_N, P_j)$ for both $j = 1, 2$. □

4 Unknown Parameters of Distributions

In this section we show that a convergence rate similar to one in Theorem 2.2 and Theorem 3.1 holds without any prior knowledge on M or $p_{\min}$. Our motivation is that in practice we may not have any information about the underlying distribution P. We show that even in this case the sub-Gaussian excess distortion bounds are possible. However, as a result, our bounds become more sensitive to some specific properties of P. The following theorem is the main result of this section.

Theorem 4.1. *Fix $\delta \in (0,1)$. There is an estimator $\hat{A}_\delta$ depending on δ such that, with probability at least $1-\delta$,*

$$D(\hat{A}_\delta) - D(A^*) \lesssim R\ \mathbb{E}^{1/2}\|X-\mu\|^2 \left((\log N)^2 \sqrt{\frac{k}{N}} + \sqrt{\frac{\log \frac{1}{\delta}}{N}} \right),$$

where R is such that

$$\mathbb{E}\|X-\mu\|^2\, \mathbb{1}[\|X-\mu\| > R] \le \frac{\Delta}{64},$$

and

$$\Delta = \inf_{A \in \mathcal{A}^{k-1}} D(A) - \inf_{A\in\mathcal{A}^k} D(A).$$

Remark 4.2. *Observe that both R and Δ played an import role in the original proof of the strong consistency by Pollard.*

Let us first define our estimator. As before, in this section we use the notation (6).

The estimator of Theorem 4.1. We set

$$\hat{A}_\delta = \operatorname*{arg\,min}_{A\in\mathcal{A}^k} \mathrm{MOM}(l_A),$$

with the number of blocks $\ell = 32\lceil \log \frac{4}{\delta} \rceil + 1$.

As before, our estimator $\hat{A}_\delta$ is an analog of an empirically optimal quantizer (1) with the only difference that instead of the sample mean we minimize the MOM criterion. Note that the estimator is translation invariant, so we can once again assume that $\mathbb{E}\,X = 0$ without loss of generality.

Proof of Theorem 4.1. We are going to compare $\hat{A}_\delta$ with $\hat{A}_\delta \cap B_M$ for some $M \gtrsim R$ and show that with high probability either $\mathbb{E}\, l_{\hat{A}_\delta \cap B_M}$ is close to $\mathbb{E}\, l_{\hat{A}_\delta}$ (for small N) or $\hat{A}_\delta \subset B_M$ (for large N), where B_M is a ball of radius M centred at the origin. Moreover, $\min_{a\in\hat{A}_\delta}\|a\| \lesssim \mathbb{E}^{1/2}\|X\|^2$ with high probability, thus in both cases we can apply Lemma 3.7 to obtain the convergence rate of the form

$$D(\hat{A}_\delta) - D(A^*) \lesssim M\,\mathbb{E}^{1/2}\|X\|^2 \sqrt{\frac{k\log^4 N + \log\frac{1}{\delta}}{N}}.$$

First, according to the first part of Lemma 3.2, with probability at least $1 - e^{-\ell/12} \ge 1 - \delta/4$,

$$\min_{a\in\hat{A}_\delta} \|a\| \le m = 4\sqrt{2\,\mathbb{E}\|X\|^2}.$$

Let us define $M = m + 2(R \vee m)$. Note that

$$\mathbb{E}\|X\|^2 \le R^2 + \mathbb{E}\|X\|^2\,\mathbb{1}[\|X\| > R] \le R^2 + \frac{\Delta}{64} \le R^2 + \frac{\mathbb{E}\|X\|^2}{64},$$

thus $R \geq 0.99\, \mathbb{E}^{1/2}\|X\|^2$, which implies $M \simeq R$.

Now fix $A \in \mathcal{A}^k$ such that $\min_{a\in A}\|a\| \leq m$. Then by (13) for any $a \in A$ (if it exists) such that $\|a\| > M$ and any $x \in V_a$, one has $\|x\| > \frac{M-m}{2} = R \vee m$, thus $l_A \equiv l_{A\cap B_M}$ on $B_{R\vee m}$. Moreover, recall that for all $x \in E$,

$$\min_{a\in A}\|a - x\| \leq m + \|x\|, \qquad \min_{a\in A\cap B_M}\|a - x\| \leq m + \|x\|,$$

and thus

$$l_{A\cap B_M}(x) - l_A(x) = \min_{a\in A\cap B_M}\|a - x\|^2 - \min_{a\in A}\|a - x\|^2 \leq (m + \|x\|)^2\, \mathbb{1}[\|x\| > R \vee m] \leq 4\|x\|^2\, \mathbb{1}[\|x\| > R],$$

Therefore,

$$\begin{aligned}\mathrm{MOM}(l_A) &\geq \mathrm{MOM}\left(l_{A\cap B_M}(X) - 4\|X\|^2\, \mathbb{1}[\|X\| > R]\right)\\ &\geq \mathrm{QOM}_{1/4}(l_{A\cap B_M}) - 4\,\mathrm{QOM}_{3/4}\left(\|X\|^2\, \mathbb{1}[\|X\| > R]\right).\end{aligned}$$

The last term can be bounded by Lemma 3.4: with probability at least $1 - \delta/4$,

$$\mathrm{QOM}_{3/4}\left(\|X\|^2\, \mathbb{1}[\|X\| > R]\right) \leq 8\, \mathbb{E}\|X\|^2\, \mathbb{1}[\|X\| > R] \leq \frac{\Delta}{8},$$

thus

$$\mathrm{QOM}_{1/4}(l_{A\cap B_M}) \leq \mathrm{MOM}(l_A) + \frac{\Delta}{2}.$$

Further, we can assume without loss of generality that A^* belongs to $\mathcal{A}^k_{M,m}$. Indeed, $\min_{a\in A^*}\|a\| \leq m$ according to the proof of Theorem 3.1, and if $A^* \not\subset B_M$, then $|A^* \cap B_M| < k$ and hence

$$\mathbb{E}\, l_{A^*} + \Delta \leq \mathbb{E}\, l_{A^*\cap B_M} \leq \mathbb{E}\, l_{A^*} + 4\, \mathbb{E}\|X\|^2\, \mathbb{1}[\|X\| > R] \leq \mathbb{E}\, l_{A^*} + \frac{\Delta}{16},$$

which is possible only if $\Delta = 0$. But in this case $\|X\| \leq R \leq M$ almost surely and $|\mathrm{supp}(\mathrm{P})| \leq k - 1$, thus one can choose $A^* = \mathrm{supp}(\mathrm{P}) \in \mathcal{A}^k_{M,m}$. Lemma 3.7 ensures that, with probability at least $1 - \delta/2$, for any $\alpha \in \left\{\frac{1}{4}, \frac{1}{2}\right\}$ it holds that

$$\sup_{A\in\mathcal{A}^k_{M,m}} |\mathbb{E}\, l_A(X) - \mathrm{QOM}_\alpha(l_A)| \leq CR\, \mathbb{E}^{1/2}\|X\|^2 \left((\log N)^2\sqrt{\frac{k}{N}} + \sqrt{\frac{\log\frac{1}{\delta}}{N}}\right), \tag{18}$$

where $C > 0$ is an absolute constant. Finally, we get the following lines of inequalities, which hold with probability at least $1 - \delta$,

$$\begin{aligned}\mathbb{E}\, l_{\hat{A}_\delta\cap B_M} &\leq \mathrm{QOM}_{1/4}(l_{\hat{A}_\delta\cap B_M}) + CR\, \mathbb{E}^{1/2}\|X\|^2 \left((\log N)^2\sqrt{\frac{k}{N}} + \sqrt{\frac{\log\frac{1}{\delta}}{N}}\right)\\ &\leq \mathrm{MOM}(l_{\hat{A}_\delta}) + \frac{\Delta}{2} + CR\, \mathbb{E}^{1/2}\|X\|^2 \left((\log N)^2\sqrt{\frac{k}{N}} + \sqrt{\frac{\log\frac{1}{\delta}}{N}}\right)\\ &\leq \mathrm{MOM}(l_{A^*}) + \frac{\Delta}{2} + CR\, \mathbb{E}^{1/2}\|X\|^2 \left((\log N)^2\sqrt{\frac{k}{N}} + \sqrt{\frac{\log\frac{1}{\delta}}{N}}\right)\\ &\leq \mathbb{E}\, l_{A^*} + \frac{\Delta}{2} + 2CR\, \mathbb{E}^{1/2}\|X\|^2 \left((\log N)^2\sqrt{\frac{k}{N}} + \sqrt{\frac{\log\frac{1}{\delta}}{N}}\right).\end{aligned}$$

Now there are two possible cases. If

$$CR\,\mathbb{E}^{1/2}\|X\|^2\left((\log N)^2\sqrt{\frac{k}{N}}+\sqrt{\frac{\log\frac{1}{\delta}}{N}}\right)\geq\frac{\Delta}{4},$$

then

$$\mathbb{E}\,l_{\hat{A}_\delta}\leq\mathbb{E}\,l_{\hat{A}_\delta\cap B_M}\leq\mathbb{E}\,l_{A^*}+4CR\,\mathbb{E}^{1/2}\|X\|^2\left((\log N)^2\sqrt{\frac{k}{N}}+\sqrt{\frac{\log\frac{1}{\delta}}{N}}\right).$$

Otherwise, we have

$$CR\,\mathbb{E}^{1/2}\|X\|^2\left((\log N)^2\sqrt{\frac{k}{N}}+\sqrt{\frac{\log\frac{1}{\delta}}{N}}\right)<\frac{\Delta}{4},$$

then $\hat{A}_\delta\subset B_M$: indeed, $\mathbb{E}\,l_{\hat{A}_\delta\cap B_M}<\mathbb{E}\,l_{A^*}+\Delta$, now assume $|\hat{A}_\delta\cap B_M|<k$, then

$$\mathbb{E}\,l_{\hat{A}_\delta\cap B_M}\geq\inf_{A\in\mathcal{A}^{k-1}}\mathbb{E}\,l_A=\mathbb{E}\,l_{A^*}+\Delta,$$

and we obtain a contradiction. Thus, $\hat{A}_\delta\in\mathcal{A}^k_{M,m}$, and (18) again yields

$$\begin{aligned}D(\hat{A}_\delta)-D(A^*)&=\mathbb{E}\,l_{\hat{A}_\delta}-\mathbb{E}\,l_{A^*}\\&\leq\mathbb{E}\,l_{\hat{A}_\delta}-\mathrm{MOM}(l_{\hat{A}_\delta})-\mathbb{E}\,l_{A^*}+\mathrm{MOM}(l_{A^*})\\&\lesssim R\,\mathbb{E}^{1/2}\|X\|^2\left((\log N)^2\sqrt{\frac{k}{N}}+\sqrt{\frac{\log\frac{1}{\delta}}{N}}\right).\end{aligned}$$

□

We conclude by comparing our Theorem 4.1 to Theorem 2.2 presented in (Biau, Devroye and Lugosi, 2008). The form of the latter result is somewhat similar to our excess distortion bound. However, the proof of Theorem 2.2 contains an inaccuracy which, to the best of our understanding, can not be immediately fixed. The problem in the proof comes from the application of Corollary 2.1 in (Biau et al., 2008) which requires that the centres belong to the set $\mathcal{A}^k_M$ (which is called $\mathcal{F}^k_M$ there) and also that the observations $X_1,\ldots,X_N$ are in a bounded domain with probability one. The last fact does not hold for the unbounded distributions considered there (recall our Remark 2.1). Fortunately, with additional technical efforts and by replacing the empirically optimal quantizer with our MOM minimizer, we achieve the result even stronger in a manner than one claimed in Theorem 2.2 by Biau et al.

5 Discussions

Finally, we discuss several previous results related to clustering for heavy-tailed distributions as well as directions for future work.

The results of Brownlees, Joly and Lugosi (2015) are only presented for k-medians (where the distortion is defined as $D(A)=\mathbb{E}\min_{a\in A}\|X-a\|$ instead of $D(A)=\mathbb{E}\min_{a\in A}\|X-a\|^2$). However, we believe that their techniques, at least if applied straightforwardly, would require $\mathbb{E}\,\|X\|^4<\infty$. Our Theorem 2.2 only requires $\mathbb{E}\,\|X\|^2<\infty$ and is valid for any separable Hilbert space, whereas Theorem 11 in (Brownlees et al., 2015) depends explicitly on the dimension and has a worse dependence on the $\log\frac{1}{\delta}$-term. The uniform bounds in (Telgarsky and Dasgupta, 2013; Bachem et al., 2017) provide uniform convergence bounds under $\mathbb{E}\,\|X\|^4<\infty$ in $\mathbb{R}^d$ that can not be immediately converted into the excess distortion bounds similar to ours. Since these uniform bounds are tuned to the analysis of empirically optimal quantizers, they obviously have a suboptimal dependence on the confidence parameter δ.

A natural course of further research is to introduce some favorable assumptions on the distribution P leading to the so-called *fast rates* for the excess distortion. These are the excess distortion bounds scaling as $O(\frac{1}{N})$ instead of $O(\frac{1}{\sqrt{N}})$ which, of course, can not be obtained for free (Antos, 2005). By now, these assumptions and their analysis are well-understood in the bounded case (see (Levrard, 2015) and references therein). Another interesting direction is to sharpen our bounds and make our robust algorithms more practical. As already mentioned, we believe that making some assumptions on $p_{\min}$ and thus restricting the sizes of clusters is somewhat more natural than assuming that M is known in advance.

Acknowledgements. We would like to thank Olivier Bachem for stimulating discussions, Gabor Lugosi for a valuable feedback and Marco Cuturi and Nikita Puchkin for providing several important references.

The work of Alexey Kroshnin was conducted within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'.

References

Alon, N., Matias, Y. and Szegedy, M. (1999). The space complexity of approximating the frequency moments, *Journal of Computer and system sciences* **58**(1): 137–147.

Antos, A. (2005). Improved minimax bounds on the test and training distortion of empirically designed vector quantizers, *IEEE Transactions on Information Theory* **51**(11): 4022–4032.

Bachem, O., Lucic, M., Hassani, S. H. and Krause, A. (2017). Uniform deviation bounds for k-means clustering, *International Conference on Machine Learning*, pp. 283–291.

Bartlett, P. L., Linder, T. and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design, *IEEE Transactions on Information theory* **44**(5): 1802–1813.

Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces, *IEEE Transactions on Information Theory* **54**(2): 781–790.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press.

Brécheteau, C., Fischer, A. and Levrard, C. (2018). Robust Bregman clustering, *arXiv:1812.04356* .

Brownlees, C., Joly, E. and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses, *The Annals of Statistics* **43**(6): 2507–2536.

Cadre, B. and Paris, Q. (2012). On Hölder fields clustering, *Test* **21**(2): 301–316.

Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters, *International Conference on Machine Learning*, pp. 685–693.

Fefferman, C., Mitter, S. and Narayanan, H. (2016). Testing the manifold hypothesis, *Journal of the American Mathematical Society* **29**(4): 983–1049.

Fischer, A. (2010). Quantization and clustering with Bregman divergences, *Journal of Multivariate Analysis* **101**(9): 2207–2221.

Foster, D. J. and Rakhlin, A. (2019). ℓ_∞-Vector contraction for Rademacher complexity, *arXiv preprint arXiv:1911.06468* .

Genevay, A., Dulac-Arnold, G. and Vert, J.-P. (2019). Differentiable deep clustering with cluster size constraints, *arXiv preprint arXiv:1910.09036* .

Graf, S. and Luschgy, H. (2007). *Foundations of quantization for probability distributions*, Springer.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space, *Contemporary Mathematics* **26**(1): 189–206.

Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice, *arXiv preprint arXiv:1711.10306, Annals of Statistics (forthcoming)* .

Lecué, G., Lerasle, M. and Mathieu, T. (2018). Robust classification via MOM minimization, *arXiv preprint arXiv:1808.03106* .

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: Isoperimetry and Processes*, Springer-Verlag Berlin Heidelberg.

Levrard, C. (2013). Fast rates for empirical vector quantization, *Electronic Journal of Statistics* **7**: 1716–1746.

Levrard, C. (2015). Nonasymptotic bounds for vector quantization in Hilbert spaces, *The Annals of Statistics* **43**(2): 592–619.

Linder, T. (2002). Learning-theoretic methods in vector quantization, *Principles of nonparametric learning*, Springer, pp. 163–210.

Lugosi, G. and Mendelson, S. (2019a). Near-optimal mean estimators with respect to general norms, *Probability Theory and Related Fields* **175**: 957–973.

Lugosi, G. and Mendelson, S. (2019b). Regularization, sparse recovery, and median-of-means tournaments, *Bernoulli* **25**(3): 2075–2106.

Lugosi, G. and Mendelson, S. (2019c). Sub-gaussian estimators of the mean of a random vector, *The Annals of Statistics* **47**(2): 783–794.

Maurer, A. (2016). A vector-contraction inequality for Rademacher complexities, *International Conference on Algorithmic Learning Theory*, Springer, pp. 3–17.

Maurer, A. and Pontil, M. (2010). k-dimensional coding schemes in Hilbert spaces, *IEEE Transactions on Information Theory* **56**(11): 5839–5846.

Mendelson, S. (2015). Learning without concentration, *Journal of the ACM* **62**(3).

Mendelson, S. and Zhivotovskiy, N. (2018). Robust covariance estimation under $L_4 - L_2$ norm equivalence, *arXiv preprint arXiv:1809.10462, Annals of Statistics (forthcoming)* .

Minsker, S. (2018). Uniform bounds for robust mean estimators, *arXiv preprint arXiv:1812.03523* .

Minsker, S. and Mathieu, T. (2019). Excess risk bounds in robust empirical risk minimization, *arXiv preprint arXiv:1910.07485* .

Narayanan, H. and Mitter, S. (2010). Sample complexity of testing the manifold hypothesis, *Advances in Neural Information Processing Systems*, pp. 1786–1794.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*, Wiley, New York.

Ng, M. K. (2000). A note on constrained k-means algorithms, *Pattern Recognition* **33**(3): 515–519.

Pollard, D. (1981). Strong consistency of k-means clustering, *The Annals of Statistics* **9**(1): 135–140.

Srebro, N., Sridharan, K. and Tewari, A. (2010). Smoothness, low noise and fast rates, *Advances in Neural Information Processing Systems*, pp. 2199–2207.

Talagrand, M. (2014). *Upper and lower bounds for stochastic processes: modern methods and classical problems*, Vol. 60, Springer Science & Business Media.

Telgarsky, M. J. and Dasgupta, S. (2013). Moment-based uniform deviation bounds for k-means and friends, *Advances in Neural Information Processing Systems*, pp. 2940–2948.

Vershynin, R. (2016). *High-Dimensional Probability: An Introduction with Applications*, Vol. 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.

Appendix

Proof of Lemma 2.3. First, notice that $\mathbb{E}\, f - \mathrm{QOM}_\alpha(f) = \mathrm{QOM}_{1-\alpha}(\mathbb{E}\, f - f)$. Therefore, $\sup_{f \in \mathcal{F}}(\mathbb{E}\, f - \mathrm{QOM}_\alpha(f)) > x$ is equivalent to

$$\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \mathbb{1}[\mathbb{E}\, f - \tilde{f}_t > x] \geq \alpha,$$

where $\tilde{f}_t = \frac{\ell}{N} \sum_{i \in I_t} f(X_i)$. Using the idea of Mendelson (2015), denote the function $\phi(u) = (u-1)\,\mathbb{1}[1 \leq u \leq 2] + \mathbb{1}[u \geq 2]$, so that ϕ is 1-Lipschitz, and $\phi(u) \geq \mathbb{1}[u \geq 2]$. Then, the above event is included in the following event

$$\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) \geq \alpha\,.$$

Next, we write the bounded difference inequality (see (Boucheron et al., 2013)) since the summands in the above are independent and bounded by one. We have that, with probability at least $1 - e^{-2\ell y^2}$,

$$\begin{aligned}\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) \leq & \sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \mathbb{E}\, \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) \\ & + \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \left\{ \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) - \mathbb{E}\, \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) \right\} \\ & + y.\end{aligned}$$

For the first part, since $\phi(u) \leq \mathbb{1}[u \geq 1]$ and using Chebyshev's inequality, we write

$$\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \mathbb{E}\, \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) \leq \sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \mathbb{P}\left(\mathbb{E}\, f - \tilde{f}_t \leq x/2\right) \leq \sup_{f \in \mathcal{F}} \frac{\mathrm{Var}(\tilde{f}_t)}{(x/2)^2} = \sup_{f \in \mathcal{F}} \mathrm{Var}(f) \frac{4\ell}{N x^2}\,.$$

For the second part we, use the symmetrization and contraction arguments of Ledoux and Talagrand (2013), so that together

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \left\{ \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) - \mathbb{E}\, \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) \right\} & \leq 2\, \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \epsilon_t \phi\left(\frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x}\right) \\ & \leq 2\, \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \epsilon_t \frac{2(\mathbb{E}\, f - \tilde{f}_t)}{x},\end{aligned}$$

where $\epsilon_1, \dots, \epsilon_t$ are i.i.d. Rademacher signs. Using the symmetrization argument again, we have

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \epsilon_t \frac{2(\mathbb{E} f - \tilde{f}_t)}{x} \le \frac{4}{x} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i) \,.$$

Collecting the three terms together we have that, with probability at least $1 - e^{-2\ell y^2}$,

$$\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \mathbb{1} \left[\mathbb{E} f - \tilde{f}_t > x \right] \le y + \frac{8}{x} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i) + \frac{4}{x^2} \sup_{f \in \mathcal{F}} \mathrm{Var}(f) \frac{\ell}{N}.$$

We need the right-hand side of the last display to be smaller than α. Let us take $y = \alpha/2$ and

$$x = \frac{16}{\alpha} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i) + \frac{2}{\alpha^{1/2}} \sqrt{2 \sup_{f \in \mathcal{F}} \mathrm{Var}(f) \frac{\ell}{N}}.$$

Then, with probability at least $1 - e^{-\alpha^2 \ell/2}$, it holds that

$$\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{t=1}^{\ell} \mathbb{1} \left[\mathbb{E} f - \tilde{f}_t > x \right] \le \alpha.$$

To derive the other tail, we can use the symmetry $\mathrm{QOM}_\alpha(f) = -\mathrm{QOM}_{1-\alpha}(-f)$. □

Proof of Lemma 3.4. By Markov's inequality and Chernoff's bound for the binomial distribution, we have

$$\mathbb{P}\left(\mathrm{Quant}_{1-\alpha}(\xi_1, \dots, \xi_\ell) \ge 2\,\mathbb{E}\,\xi/\alpha\right) = \mathbb{P}\left(\sum_{i=1}^{\ell} \mathbb{1}[\xi_i \ge 2\,\mathbb{E}\,\xi/\alpha] \ge \alpha\ell\right) \le \exp\left(-\frac{1}{3}\left(\alpha - \frac{\alpha}{2}\right)\ell\right) = \exp\left(-\frac{\alpha\ell}{6}\right).$$

□

Proof of Lemma 3.6. **Step 1.** We start with $E = \mathbb{R}^d$. In order to prove the bound we observe that $L_2(P_N)$ distance between l_A, $A = (a_1, \dots, a_k) \in \mathcal{A}^k_{M,m}$, and l_B, $B = (b_1, \dots, b_k) \in \mathcal{A}^k_{M,m}$ (we can multiply some points if $|A| < k$ or $|B| < k$), is controlled by the maximum of the Euclidean distances between the corresponding vectors a_j and b_j. Indeed, let $x \in V_{a_j} \cap V_{b_s}$, then the following assertions hold:

$$\begin{aligned} l_B(x) - l_A(x) &\le \|b_j\|^2 - \|a_j\|^2 - 2\langle x, b_j - a_j \rangle \le 2\left(\|a_j\| + \|x\|\right)\|a_j - b_j\| + \|a_j - b_j\|^2, \\ l_A(x) - l_B(x) &\le \|a_s\|^2 - \|b_s\|^2 - 2\langle x, a_s - b_s \rangle \le 2\left(\|b_s\| + \|x\|\right)\|a_s - b_s\| + \|a_s - b_s\|^2. \end{aligned}$$

Therefore,

$$|l_A(x) - l_B(x)| \lesssim \left(\|a_j\| + \|b_s\| + \|x\|\right) \max_r \|a_r - b_r\| + \max_r \|a_r - b_r\|^2.$$

On the other hand,

$$\|a_j - x\| = \min_{a \in A} \|a - x\| \le \|x\| + \min_{a \in A} \|a\| \le \|x\| + m,$$

as well as $\|b_s - x\| \le \|x\| + m$, thus

$$|l_A(x) - l_B(x)| = \left|\|a_j - x\|^2 - \|b_s - x\|^2\right| \le \left(\|x\| + m\right)^2.$$

Combining the above bounds and using the inequality $u^2 \wedge v^2 \le uv$ for $u, v \ge 0$, we conclude that

$$\begin{aligned} |l_A(x) - l_B(x)| &\lesssim \left(\|a_j\| + \|b_s\| + \|x\|\right) \max_r \|a_r - b_r\| + \left(\|x\| + m\right)^2 \wedge \max_r \|a_r - b_r\|^2 \\ &\lesssim \left(\|a_j\| + \|b_s\| + \|x\|\right) \max_r \|a_r - b_r\| + \left(\|x\| + m\right) \max_r \|a_r - b_r\| \\ &\lesssim \left(\|a_j\| + \|b_s\| + \|x\| + m\right) \max_r \|a_r - b_r\|. \end{aligned}$$

Note that (14) applied to the empirical measure P_N ensures

$$\sum_{a\in A}\|a\|^2 P_N(V_a) \le \sigma_N^2, \quad \sum_{b\in B}\|b\|^2 P_N(V_b) \le \sigma_N^2, \quad \text{where} \quad \sigma_N^2 = 2m^2 + 8P_N\|X\|^2. \tag{19}$$

Therefore, we have

$$\begin{aligned}\|l_A - l_B\|_{L_2(P_N)} &\lesssim \left(\sqrt{\sum_{a\in A}\|a\|^2 P_N(V_a)} + \sqrt{\sum_{b\in B}\|b\|^2 P_N(V_b)} + \sqrt{\frac{1}{N}\sum_i \|X_i\|^2} + m\right)\max_r \|a_r - b_r\| \\ &\le \left(2\sigma_N + m + P_N^{1/2}\|X\|^2\right)\max_r \|a_r - b_r\| \\ &\lesssim \sigma_N \max_r \|a_r - b_r\|.\end{aligned}$$

Finally, we use that in $(\mathbb{R}^d)^k$ it holds that

$$\log \mathcal{N}_\infty\left((B_M)^k, t\right) \le k \log \mathcal{N}\left(B_M, t\right) \lesssim kd \log \frac{M}{t},$$

(see e.g., (Vershynin, 2016)).

Step 2. Now we are ready to prove the bound in its full generality. First, note that (15) implies (which holds for P_N as well)

$$\|l_A - l_{\{0\}}\|_{L_2(P_N)} \le 2M\sqrt{6P_N\|X\|^2 + m^2} \le 2M\sigma_N \quad \text{for all} \quad A \in \mathcal{A}_{M,m}^k,$$

where σ_N comes from (19), so it is enough to consider $t \le 2M\sigma_N$.

We are going to apply the Johnson–Lindenstrauss lemma, and to do this, we first show that it is enough to consider quantizers A from some finite-dimensional subspace of E, depending on the sample $X_1, \dots, X_N$. Let us fix an arbitrary vector $u \notin \operatorname{Span}(\{X_1, \dots, X_N\})$. It is easy to see that for any $a \in E$ there exists $\tilde{a} \in S = \operatorname{Span}(\{u, X_1, \dots, X_N\})$ such that $\|\tilde{a}\| = \|a\|$ and $\langle \tilde{a}, X_i\rangle = \langle a, X_i\rangle$ for all $1 \le i \le N$. Therefore, without loss of generality, one can restrict $\mathcal{A}_{M,m}^k$ to the sets from the $(N+1)$-dimensional subspace S. Using the last observation, by the version of Johnson–Lindenstrauss lemma for products (p.998 in (Fefferman, Mitter and Narayanan, 2016)) for any $0 < \epsilon \le 1/2$ and any fixed set $Q \subset S$ with $|Q| \le N + k$, it holds that

$$\mathbb{P}_L\left(E_L^Q\right) = \mathbb{P}_L\left(|\langle R_L x, R_L y\rangle - \langle x, y\rangle| \le \epsilon\|x\| \cdot \|y\| \quad \text{for all} \quad x, y \in Q\right) \ge \frac{1}{2}.$$

Here L is a random uniformly distributed d-dimensional subspace of S with $d = \left\lceil \frac{c\log(N+k)}{\epsilon^2}\right\rceil$ (for a rigorous mathematical definition see (Johnson and Lindenstrauss, 1984)), and $R_L = \sqrt{\frac{N+k}{d}}\Pi_L$, where Π_L is the orthogonal projector on L. Let $\mathcal{P}(t) \subset \mathcal{A}_{M,m}^k$ be such that $\{l_A : A \in \mathcal{P}(t)\}$ is a t-packing set of $\mathcal{F}_{M,m}^k$ (i.e., a maximal t-separated set), so that by the standard relation $\mathcal{N}_2(\mathcal{F}_{M,m}^k, t, P_N) \le |\mathcal{P}(t)|$. Now notice that for A chosen uniformly at random from $\mathcal{P}(t)$ and random L with joint probability at least $\frac{1}{2}$ the above condition holds for the set $Q_A = \{X_1, \dots, X_N\} \cup A$:

$$\mathbb{P}\left(E_L^{Q_A}\right) = \mathbb{E}_A\, \mathbb{P}_L(E_L^{Q_A}) \ge \frac{1}{2}.$$

Note that we consider the union of the sample and a quantizer since we will have to bound both the norms of projections and products of the form $\langle R_L X_i, R_L a\rangle$. On the other hand, by Fubini's theorem

$$\mathbb{P}\left(E_L^{Q_A}\right) = \mathbb{E}_L\, \mathbb{P}_A(E_L^{Q_A}) \ge \frac{1}{2},$$

therefore, there exists subspace L such that $\mathbb{P}_A(E_L^{Q_A}) \ge \frac{1}{2}$, that is, the event $E_L^{Q_A}$ holds for at least half of quantizers from $\mathcal{P}(t)$ — let us denote this set by $\mathcal{P}_L(t)$.

Consider an arbitrary quantizer $A \in \mathcal{P}_L(t)$. In what follows we use for brevity the following simple notation: $x' = R_L x$ for any $x \in E$ and, respectively, $A' = \{a' : a \in A\}$ and $P'_N = \frac{1}{N}\sum_i \delta_{X'_i}$. Clearly, by Johnson-Lindenstrauss lemma for any $x \in Q_A$,

$$\frac{1}{2}\|x\|^2 \leq (1-\epsilon)\|x\|^2 \leq \|x'\|^2 = \langle R_L x, R_L x\rangle \leq (1+\epsilon)\|x\|^2 \leq \frac{3}{2}\|x\|^2,$$

thus

$$\max_{a' \in A'}\|a'\|^2 \leq \frac{3}{2}M^2, \quad \min_{a' \in A'}\|a'\|^2 \leq \frac{3}{2}m^2, \quad P_N\|X'\|^2 \leq \frac{3}{2}P_N\|X\|^2.$$

In particular, this implies that

$$A' \in \mathcal{A}^k_{3M/2,3m/2} \quad \text{and} \quad \sum_{a' \in A'}\|a'\|^2 P(V_{a'}) \leq \frac{3}{2}\sigma_N,$$

with σ_N defined by (19). Now for any $X_i \in V_a(A)$, where $V_a(A)$ is the Voronoi cell from partition induced by the set A, corresponding to the point a, we have

$$l_{A'}(X'_i) \leq \|a'\|^2 - 2\langle X'_i, a'\rangle \leq \|a\|^2 - 2\langle X_i, a\rangle + \epsilon\left(\|a\|^2 + 2\|X_i\|\cdot\|a\|\right) \leq l_A(X_i) + \epsilon M\left(\|a\| + 2\|X_i\|\right),$$

and in the same way we obtain that for $X'_i \in V_{a'}(A')$,

$$l_A(X_i) \leq l_{A'}(X'_i) + \epsilon\left(\|a\|^2 + 2\|X_i\|\cdot\|a\|\right) \leq l_{A'}(X'_i) + \epsilon M\left(\|a\| + 2\|X_i\|\right) \leq l_{A'}(X'_i) + \epsilon M\left(\sqrt{2}\|a'\| + 2\|X_i\|\right).$$

Therefore, recalling the definition of σ_N (19), we have

$$\|l_A(X) - l_{A'}(X')\|_{L_2(P_N)} \leq \epsilon M\left(\sqrt{3}\sigma_N + 2P_N^{1/2}\|X\|^2\right) \leq 3\epsilon M\sigma_N.$$

Setting $\epsilon = \frac{t}{12M\sigma_N} \leq \frac{1}{6}$ we get $\|l_A(X) - l_{A'}(X')\|_{L_2(P_N)} \leq \frac{t}{4}$, thus

$$\|l_{A'}(X) - l_{B'}(X)\|_{L_2(P'_N)} \geq \|l_A(X) - l_B(X)\|_{L_2(P_N)} - \frac{t}{2} > \frac{t}{2} \quad \text{for any} \quad A \neq B \in \mathcal{P}_L(t).$$

This implies by the standard relation between the covering and packing numbers that $|\mathcal{P}_L(t)| \leq \mathcal{N}_2\left(\mathcal{F}', t/4, P'_N\right)$, where $\mathcal{F}' = \left\{l_{A'} : A \in \mathcal{A}^k_{M,m}\right\}$. As was shown above $\mathcal{F}' \subset \mathcal{F}^k_{3M/2,3m/2}$, and since the corresponding quantizers belong to the d-dimensional subspace L, we have by **Step 1** that

$$\log\mathcal{N}_2\left(\mathcal{F}', t/4, P'_N\right) \lesssim kd\log\frac{M\sigma_N}{t} \lesssim \frac{kM^2\sigma_N^2\log(N+k)}{t^2}\log\frac{M\sigma_N}{t}.$$

Combining our bounds, we conclude that

$$\log\mathcal{N}_2(\mathcal{F}^k_{M,m}, t, P_N) \leq \log|\mathcal{P}(t)| \leq \log\left(2|\mathcal{P}_L(t)|\right) \lesssim \log\mathcal{N}_2\left(\mathcal{F}', t/4, P'_N\right) \lesssim \frac{kM^2\sigma_N^2\log(N+k)}{t^2}\log\frac{M\sigma_N}{t}.$$

To obtain the claimed bound, it remains to notice that for $N < k$ it is enough to consider the class $\mathcal{A}^N_{M,m}$ instead of $\mathcal{A}^k_{M,m}$, thus we can always assume $k \leq N$. Hence,

$$\log\mathcal{N}_2(\mathcal{F}^k_{M,m}, t, P_N) \lesssim \frac{kM^2\sigma_N^2\log(2N)}{t^2}\log\frac{M\sigma_N}{t}.$$

The claim follows. □