JMLR: Workshop and Conference Proceedings vol 35:1–17, 2014

Density-preserving quantization with application to graph downsampling

Morteza Alamgir ALAMGIR@INFORMATIK.UNI-HAMBURG.DE
Department of Computer Science, University of Hamburg

Gábor Lugosi GABOR.LUGOSI@UPF.EDU
ICREA and Department of Economics, Universitat Pompeu Fabra

Ulrike von Luxburg LUXBURG@INFORMATIK.UNI-HAMBURG.DE
Department of Computer Science, University of Hamburg

Abstract

We consider the problem of vector quantization of i.i.d. samples drawn from a density p on $\mathbb{R}^d$. It is desirable that the representatives selected by the quantization algorithm have the same distribution p as the original sample points. However, quantization algorithms based on Euclidean distance, such as k-means, do not have this property. We provide a solution to this problem that takes the unweighted k-nearest neighbor graph on the sample as input. In particular, it does not need to have access to the data points themselves. Our solution generates quantization centers that are "evenly spaced". We exploit this property to downsample geometric graphs and show that our method produces sparse downsampled graphs. Our algorithm is easy to implement, and we provide theoretical guarantees on the performance of the proposed algorithm.

Keywords: Vector quantization, kNN graph, sampling

1. Introduction

Vector quantization is the task of compressing a large set of data points into a set of representatives called centroids or centers. Its applications are abundant, just consider the examples of visual word models in computer vision (Leung and Malik, 2001; Csurka et al., 2004), color quantization in image processing (Heckbert, 1982), or selecting landmark points (de Silva and Tenenbaum, 2004). For data points in $\mathbb{R}^d$, the standard approach is to minimize a quantization error measured in terms of the Euclidean distance. The best-known algorithm is k-means (where k is relatively large, as opposed to clustering problems, where k is usually chosen much smaller). Because the selected centers are supposed to be a "faithful representation" of the original data points, it is a highly desirable property that the centroids have the same distribution as the original data points. However, it is known that this is not the case if we minimize the quantization error with respect to the Euclidean distance (Graf and Luschgy, 2000). The optimal centers that minimize the quantization error with respect to the Euclidean distance are distributed as a power of the original density instead of the density itself.
How can we find a set of centroids that matches the underlying density of our data, without even observing the location of the points? This is the key problem we study in this paper. Previous work assumes that the coordinates of the input points are observed. Delp and Mitchell (1991) looked for a weaker property, which is to match the moments of the underlying density. Hegde et al. (2004) explicitly minimize the KL-divergence between a kernel density estimate of the original data and the estimated density of the centroids by gradient descent. Hulle (1999) and Meinicke and Ritter

(2001) use a similar approach to build a compact density estimator. Instead of relying on all samples to estimate the density, they select few points that best describe the density. Recently, Li et al. (2011) considered using the Rosenblatt transformation which transforms an arbitrary distribution to a uniform distribution. After quantizing the transformed data, one can use the inverse transformation to get back to the original density.

In our paper, we introduce a completely new approach to the problem. The algorithm we suggest is conceptually simple: we construct an unweighted k-nearest neighbor (kNN) graph with respect to the Euclidean distance on the sample points and use the k-medoids algorithm with respect to the shortest path distance on this graph for selecting the representatives. The k-medoids algorithm is similar to k-means, but the centroids of the clusters have to be chosen from the datapoints. Quite surprisingly, we do not even need to have access to the data points themselves or to their Euclidean distances. Our algorithm works as soon as we know which are the k-nearest neighbors of each data point. It also works when the distribution of the data is concentrated on a manifold, and we do not even need to know the intrinsic dimension of the data.

Computing the k-medoids with respect to the graph shortest path distance is an extension of k-means for graphs, and have been used by Kim et al. (2007) and Feil and Janos (2007) for data clustering. However, the particular choice of the unweighted kNN graph and the intriguing properties of the resulting centroids have not been studied in the literature.

Our optimal quantization centers enjoy another interesting property: They are "evenly spaced" on the domain. They distribute well in the domain and do not bunch together. This can be formally stated using the dispersion of the centers. The dispersion of a set of samples is the radius of the largest uncovered ball in the domain. We show that the set of optimal quantization centers has a dispersion lower than the dispersion of a random subsample of the datapoints. We exploit this property and apply our algorithm to downsample random geometric graphs. A large geometric graph G is given, where its vertices are n i.i.d. samples from an unknown density p. We do not even need to have access to coordinates of the vertices. Our task is to downsample G to n' vertices: to build a graph G' which "looks like" a geometric graph built on n' samples from p. Our algorithm provides a sparse solution to this problem. It builds a geometric graph whose number of edges depends linearly on n'. Compare this to the naive approach of taking a random subsample of the vertices. The selected vertices clearly have the desired distribution, but a connected neighborhood graph built from these vertices is less sparse (number of edges of the order $O(n \log n)$).

A major part of this paper is devoted to a thorough statistical analysis of our algorithm. We first introduce a new distance function D_{PD} on $\mathbb{R}^d$ which depends on the data density p. We then show that this distance function plays the role of a "uniformizing transformation" of the space: optimally quantizing p-distributed data with respect to the distance function D_{PD} behaves as quantizing uniform data with respect to the Euclidean norm. As a second step, we exploit that the D_{PD}-distance on $\mathbb{R}^d$ can be approximated by the shortest path distance in unweighted kNN graphs.

2. Definitions and formal setup

Consider a connected and compact subset $\mathcal{X} \subset \mathbb{R}^d$ endowed with a probability measure P that has a Lipschitz continuous density p with Lipschitz constant L. Assume that for all $x \in \mathcal{X}$ we have $0 < p_{\min} \leq p(x) \leq p_{\max} < \infty$. The set $V_n = \{X_1, ..., X_n\} \subset \mathcal{X}$ has been drawn i.i.d. according to p. Denote by P_n the empirical measure of the sample. The unweighted and undirected k-nearest neighbor graph G_n is the graph with vertex set V_n where we connect X_j to X_i by an

unweighted undirected edge if X_j is among the k nearest neighbors of X_i or vice versa (according to the Euclidean metric). In the following, we use the letter k to denote the parameter of the kNN graph and the letter $\mathfrak{K}$ to denote the number of representatives for $\mathfrak{K}$-means and $\mathfrak{K}$-medoids.
For two vertices $x, y \in V_n$, the *shortest path distance* $D_{sp}^{G_n}(x, y)$ is the length of the shortest path connecting x to y in G_n. When the underlying graph is clear from the context, we drop the index G_n and simply use the notation $D_{sp}(x, y)$.

Definition 1 (Following / resembling a density) *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of sets with $A_n \subset \mathcal{X}$, $|A_n| = n$. We say that A_n follows (or resembles) the density function p if for any measurable set $S \subset \mathcal{X}$, the fraction of points of A_n that lie inside S converges to the probability mass of S: $\frac{1}{n}|S \cap A_n| \to \int_S p(x)dx$ as $n \to \infty$.*

Let f be a positive, continuous, real-valued function defined on $\mathcal{X}$. We define the f-weighted length or *f-length* of a differentiable curve $\gamma : [0, 1] \to \mathcal{X}$ as

$$D_f(\gamma) = \int_0^1 f(\gamma(t))|\gamma'(t)|dt.$$

The f-distance between x and y is defined as $D_f(x, y) := \inf_\gamma D_f(\gamma)$ where the infimum is over all rectifiable paths γ with finite length that connect x to y. We introduce a shorthand notation for the f-distance with $f(x) = p(x)^{1/d}$ and call it PD-*distance*. This distance is a metric (Alamgir and von Luxburg, 2012) and belongs to the family of density-based distances (Sajama and Orlitsky, 2005). In Lemma 3 we show that the PD-distance induces a uniform structure on non-uniform densities.

3. Vector quantization

Let $\mathfrak{K}$ be a positive integer. A *level vector quantizer* maps vectors in $\mathbb{R}^d$ to a set $A = \{a_1, \cdots, a_\mathfrak{K}\} \subset \mathbb{R}^d$. Each element a_i is called a *centroid*, a *center* or a *representative*. The set $\mathcal{V}_i \subset \mathbb{R}^d$ of vectors that are mapped to the centroid a_i is called a *cell*. A common assignment is based on Euclidean distance: every point x is assigned to its nearest centroid with respect to the Euclidean distance. A widely used example of Euclidean distance quantizers is least squares quantization, also known as $\mathfrak{K}$-means (Lloyd, 1982). As we will work with several metrics, we use an explicit notation for the *centroid assignment function* $\mathcal{C}_A : \mathbb{R}^d \to A$. This is the function that determines the cells $\mathcal{V}_i = \{x \in \mathbb{R}^d | \ \mathcal{C}_A(x) = a_i\}$. In particular, we consider the functions $\mathcal{C}_{A,\|\cdot\|}$, $\mathcal{C}_{A,\text{PD}}$ and $\mathcal{C}_{A,sp}$ that assign points to the closest center according to the Euclidean distance, the PD-distance and the graph shortest path distance, respectively.
The *representation error* $g(x, y)$ quantifies the error of representing y by x. In $\mathfrak{K}$-means we have $g(x, y) = \|x - y\|^2$. The quality Φ of a set of centroids is measured by the expected representation error of the centroid mapping function

$$\Phi(g, \mathcal{C}_A, P) := \int g(\mathcal{C}_A(x), x)P(dx).$$

The set of *optimal centroids* of size $\mathfrak{K}$ with respect to the error function $\Phi(g, \mathcal{C}_A, P)$ is defined as the set A that minimizes the expected error

$$A^* = \operatorname*{argmin}_{A,|A|=\mathfrak{K}} \Phi(g, \mathcal{C}_A, P).$$

The set of empirical optimal centroids A_n is defined analogously with respect to $\Phi(g, \mathcal{C}_A, P_n)$. Finding the set of empirical optimal centroids A_n is an NP-hard problem in general. However, there exist EM-type algorithms and several heuristics to find a good local optimum of $\Phi(g, \mathcal{C}_A, P_n)$. In this work, we ignore the issue of algorithmic complexity and assume that an empirically optimal quantizer can be computed.
The topology of optimal centroids and the shape of their corresponding cells have been studied in several papers, see Graf and Luschgy (2000) and Gruber (2004) for references. As a brief summary of their results, the optimal centroids in the case $g(x, y) = \|x - y\|^\alpha$ are distributed asymptotically with density $p^{d/(d+\alpha)}$ when the number of centroids $\mathfrak{K}$ goes to infinity.

4. Quantization with the PD-distance

In this section, we introduce a new technique to transform a space with a non-uniform density to a space with uniform density. The PD-distance plays a key role in this construction. The transformation is defined for all smooth compact Riemannian manifolds, but later we only use it for full-dimensional subsets of $\mathbb{R}^d$. We use this transformation to construct a density-preserving quantization procedure. In Section 4.3 we then show how to approximate the quantization procedure and why this approximation works.

4.1. Uniformizing metric

Let us start with some intuition. Consider a uniform elastic rubber band with printed lines graduated in centimeters. Stretch the band in different places to get a non-uniform band. The printed lines will also displace and their Euclidean distances will change. However, these displaced lines show a uniformity property: The mass of the elastic band between successive lines is the same all over the stretched band. We show that the PD-distance corresponds to such a stretched distance when the density function p corresponds to the density of the rubber at each point on the band.

Definition 2 (Uniformizing metric) *Let (M, h) be a differentiable compact Riemannian manifold with intrinsic dimension d, where h is the standard Riemannian metric. Let $p(x)$ be a continuous and bounded density defined on M. Consider the differentiable Riemannian manifold (M, h^p) where $h^p_x(u, v) := h_x(p(x)^{1/d}u, p(x)^{1/d}v)$ for tangent vectors u and v at x. We call the metric h^p a metric uniformizing the density p.*

By assumption, the density p is strictly positive and the uniformizing metric is a conformal change of metric with $h^p_x = p(x)^{2/d}h_x$. Let ϱ (resp. ϱ_p) and w (resp. w_p) denote the geodesic distance and the volume element on (M, h) (resp. on (M, h^p)). We denote the volume of a set $\mathcal{A} \subset \mathcal{X}$ in (M, h^p) as $\mathrm{Vol}_p(\mathcal{A})$.

Lemma 3 (Density becomes uniform and PD-distance becomes Euclidean) *Consider a compact differentiable Riemannian manifold (M, h) and its uniformizing transformation (M, h^p).*

1. *The length of a continuously differentiable curve $\gamma : [0, 1] \to M$ in (M, h^p) is the* PD*-length of γ in (M, h).*
2. *The geodesic distance between two points in (M, h^p) corresponds to their* PD*-distance in (M, h).*

3. *The uniformizing transformation induces a uniform density on M: all sets $\mathcal{A}_1, \mathcal{A}_2 \subset M$ with $P(\mathcal{A}_1) = P(\mathcal{A}_2)$ also have the same volume in (M, h^p), that is $Vol_p(\mathcal{A}_1) = Vol_p(\mathcal{A}_2)$.*

Proof
Part 1. The length of a continuously differentiable curve γ in (M, h^p) is defined as $\int_0^1 \|\gamma'(t)\|_{h^p} dt$, where $\|\cdot\|_{h^p}$ is the norm induced by the inner product on the tangent space $T^pM(\gamma(t))$ at point $\gamma(t)$. The result follows from $\|x\|_{h^p} = p(x)^{1/d}\|x\|_h$ and the definition of the PD-length.
Part 2. The geodesic distance is the infimum over the lengths of paths connecting the two points. Exploiting Part 1 gives $\varrho_p(c, x) = D_{\text{PD}}(c, x)$.
Part 3. From properties of the conformal change of a metric (see Theorem 1.159 in Besse, 1987) we get $dw_p(x) = p(x)dw(x)$. Then

$$\frac{\text{Vol}_p(\mathcal{A}_1)}{\text{Vol}_p(\mathcal{A}_2)} = \frac{\int_{\mathcal{A}_1} dw_p(x)}{\int_{\mathcal{A}_2} dw_p(x)} = \frac{\int_{\mathcal{A}_1} p(x)dw(x)}{\int_{\mathcal{A}_2} p(x)dw(x)} = \frac{P(\mathcal{A}_1)}{P(\mathcal{A}_2)} = 1.$$

■

4.2. Quantization with the exact PD-distance achieves the correct distribution

From here on we restrict ourselves to Euclidean spaces for the sake of simplicity. However, all theorems can be generalized to smooth manifolds. We now study the behavior of optimal centroids with respect to the PD-distance, and later the behavior of the empirical centroids with respect to the shortest path distance. In the following we always assume that the set of optimal centers is unique. Our results can also be extended in a straightforward manner to the non-unique case by studying the set of unique centers, but this introduces heavier notation. Although all results in this paper hold for $g_f(x, y) = D_f(x, y)^t$ with arbitrary $t \geq 1$, we restrict ourselves to the more common case $t = 2$. Specifically, we mainly deal with the representation error functions

$$g_{\text{PD}}(x, y) = D_{\text{PD}}(x, y)^2 \quad \text{and} \quad g_{sp}(x, y) = D_{sp}(x, y)^2.$$

The next theorem specifies the distribution of the centroids with respect to the PD-distance. The intuition behind the theorem is that the density of $\mathfrak{K}$-means centroids matches the density of the data points when the latter is uniform. So we use a distance measure that induces a uniform density on the underlying space.

Theorem 4 (Quantization with the exact PD-distance leads to correct distribution of centers) *Let $\mathcal{X}$ be a connected and compact subset of $\mathbb{R}^d$ endowed with a Lipschitz continuous density p. Assume that for all $x \in \mathcal{X}$ we have $0 < p_{\min} \leq p(x) \leq p_{\max} < \infty$. Let $A^{*\mathfrak{K}}$ be the optimal $\mathfrak{K}$-means centroid set with respect to the* PD*-distance that attains the minimum*

$$A^{*\mathfrak{K}} = \underset{A, |A|=\mathfrak{K}}{\text{argmin}} \left\{ \int_{\mathcal{X}} \min_{c \in A} D_{\text{PD}}(c, x)^2 p(x) dx \right\}. \tag{1}$$

Then $A^{*\mathfrak{K}}$ follows the density p as $\mathfrak{K} \to \infty$.

Proof Using Lemma 3,

$$\int \min_{c \in A} \varrho_p(c, x)^2 dw_p(x) = \int \min_{c \in A} D_{\text{PD}}(c, x)^2 p(x) dw(x),$$

where $w(x)$ is the Lebesgue measure on $\mathcal{X}$. Now Part 2 from Theorem 1 in Gruber (2004) shows that the centers minimizing

$$\int \min_{c \in A} \varrho_p(c,x)^2 dw_p(x)$$

follow a uniform density with respect to the area measure w_p as $\mathfrak{K} \to \infty$. This means that they follow density p with respect to the Lebesgue measure. ■

Theorem 4 shows that the optimal $\mathfrak{K}$-means centers with respect to the PD-distance follow the underlying density of our data. However, it is neither easy to compute the PD-distance from p, nor from a set of i.i.d. sample points from p. In the next section we present a simple algorithm to approximate the $\mathfrak{K}$-means centers with respect to the PD-distance.

4.3. Quantization with approximate PD-distance using unweighted kNN graphs

We now provide a simple and effective way to approximate $\mathfrak{K}$-means centers with respect to the PD-distance. The procedure is the following: First build an unweighted kNN graph G_n based on samples $X_1, ..., X_n \in \mathcal{X}$ from the density p with a properly chosen k (see Theorem 5 for the exact condition on k). Then we quantize with respect to the graph shortest path distance in G_n. For this quantization step we can either use the $\mathfrak{K}$-means or the $\mathfrak{K}$-medoids algorithms. The latter can be used when we do not have access to the location of the vertices. This is the case on which we focus.We show that the centers constructed by this procedure converge to optimal $\mathfrak{K}$-means centers with respect to the PD-distance. This is proved in Section 4.3.1 for a fixed $\mathfrak{K}$ and $k, n \to \infty$. In Section 4.3.2, we show that if $\mathfrak{K}$ also goes to infinity "slowly" enough, the $\mathfrak{K}$-medoids centers will resemble the original density.

4.3.1. CONVERGENCE FOR A FIXED $\mathfrak{K}$

Set $\beta = \eta_d p_{\min}^{d+1}/(4L)^d$ where η_d is the volume of a Euclidean unit ball in $\mathbb{R}^d$ and L is the Lipschitz constant of the density p.

Theorem 5 (Convergence of $\mathfrak{K}$-medoids centers to $\mathfrak{K}$-means centers) *Consider the unweighted kNN graph G_n based on the i.i.d. sample $X_1, ..., X_n \in \mathcal{X}$ from the density p. We choose k such that for a fixed $\alpha > 0$,*

$$24 \log(n)^{1+\alpha} < k < \beta \frac{n}{(\log n)^\alpha}.$$

Fix $\mathfrak{K}$. Let A_n denote the set of optimal empirical $\mathfrak{K}$-medoids in graph G_n with respect to the shortest path distance. We assume that the set of optimal $\mathfrak{K}$-means centers for $\mathcal{X}$ with respect to the PD-distance is unique and denote this set by A^* . As n tends to infinity, the set of empirical $\mathfrak{K}$-medoids centers A_n converges almost surely to A^* with respect to the Hausdorff distance.

To prove this theorem, we need to know the behavior of the shortest path distance in unweighted kNN graphs. Alamgir and von Luxburg (2012) study this problem and show convergence of the shortest path distance to PD-distance in probability. In Lemma 6 we present a simplified version of their theorem. Then in Lemma 7 we use the Borel-Cantelli lemma to show almost sure convergence of $|\Phi(g_{sp}, \mathcal{C}_A, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n)|$ to zero. This means that the empirical quantization error with respect to the PD-distance approximates the normalized empirical quantization error with respect to

the shortest path distance. The price we pay to reach almost sure convergence instead of convergence in probability is to have a slightly stronger condition on k, the connectivity parameter of the kNN graph. To keep our proofs readable, we ignore boundary effects. One can show that the boundary effects are asymptotically negligible.

Lemma 6 (Convergence of D_{sp} to D_{PD})
Consider a sequence of i.i.d. samples $V_n = \{X_1, ..., X_n\} \subset \mathcal{X}$ drawn from the density p. Build an unweighted kNN graph G_n based on $X_1, ..., X_n$ and denote the shortest path distance on this graph by D^n_{sp}. Set $c_n = (\frac{k}{n\eta_d})^{1/d}$. Assume that for a fixed α, $24\log(n)^{1+\alpha} < k < \beta\frac{n}{(\log n)^\alpha}$. Set $\lambda = (\log n)^{-\frac{\alpha}{4d}}$ and let $n \geq exp(4^{(d+1)/\alpha^2})$. Then, with probability at least $1 - \frac{1}{n^2}$, for all pairs $x, y \in V_n$

$$\frac{1}{(1-\lambda)^{1/d}} D_{\text{PD}}(x,y) \leq c_n D^n_{sp}(x,y) \leq \frac{1}{\frac{1}{(1+\lambda)^{1/d}} - 2\frac{(1+\lambda)^{1/d}}{k^{\alpha^2}}} D_{\text{PD}}(x,y) + c_n.$$

Proof In Theorem 1 of Alamgir and von Luxburg (2012), set $a = 1 - \alpha^2$. For the selected n, we have

$$a < 1 - \log_k \left(4^d(1+\lambda)^2\right),$$

which is the condition needed for a. The probability that the statement in the theorem holds is $1 - P_{err}$ where $P_{err} = \frac{2^d}{(1-\lambda)^2} 3ne^{-k^a/6}$. Set k and λ as mentioned in the theorem to get

$$P_{err} \leq \frac{2^d}{(1-\lambda)^2} ne^{-k^a/6} \leq 3.2^{d+1} ne^{-4\log(n)} \leq \frac{1}{n^2}.$$

Therefore, with probability at least $1 - \frac{1}{n^2}$, for all pairs $x, y \in V_n$:

$$\frac{1}{(1-\lambda)^{1/d}} D_{\text{PD}}(x,y) \leq c_n D^n_{sp}(x,y) \leq \frac{1}{\frac{1}{(1+\lambda)^{1/d}} - 2\frac{(1+\lambda)^{1/d}}{k^{\alpha^2}}} D_{\text{PD}}(x,y) + c_n.$$

■

The condition $n \geq exp(4^{(d+1)/\alpha^2})$ can be relaxed by choosing k larger than the threshold $\log(n)$. For example if we choose $k \sim \log(n)^d$, we get the more realistic condition $n \geq e^4$.
The next lemma shows convergence of the normalized quantization error with respect to the shortest path distance $c_n^2\Phi(g_{sp}, \mathcal{C}_A, P_n)$ to the quantization error with respect to the PD-distance $\Phi(g_{\text{PD}}, \mathcal{C}_A, P_n)$.

Lemma 7 (Convergence of quantization errors, fixed $\mathfrak{K}$) *Consider the setting in Lemma 6. Let $n \to \infty$ and fix $\mathfrak{K}$. Then*

$$\sup_{A \subset \mathcal{X}, |A| = \mathfrak{K}, \forall \mathcal{C}_A} |c_n^2\Phi(g_{sp}, \mathcal{C}_A, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n)| \to 0 \ \ a.s.$$

Proof By the definition of $\Phi(g_{sp}, \mathcal{C}_A, P_n)$ and Lemma 6, the following inequalities hold with probability at least $1 - n^{-2}$ for all $A \subset \mathcal{X}$ and $|A| = \mathfrak{K}$:

$$\begin{aligned} c_n^2\Phi(g_{sp}, \mathcal{C}_A, P_n) &= \frac{c_n^2}{n} \sum_{v \in V_n} D_{sp}(\mathcal{C}_A(v), v)^2 \geq \frac{1}{n(1-\lambda)^{2/d}} \sum_{v \in V_n} D_{\text{PD}}(\mathcal{C}_A(v), v)^2 \\ &= \frac{1}{(1-\lambda)^{2/d}} \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n). \end{aligned} \tag{2}$$

For the upper bound, set $q = 1/(\frac{1}{(1+\lambda)^{1/d}} - 2\frac{(1+\lambda)^{1/d}}{k^{\alpha^2}})$. Then we have

$$c_n^2 \Phi(g_{sp}, \mathcal{C}_A, P_n) \leq q^2 \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n) + c_n^2 + 2c_n q \Phi(D_{\text{PD}}, \mathcal{C}_A, P_n). \tag{3}$$

Note that in Lemma 6, the property holds for all pairs of vertices simultaneously. It is easy to show that for a fixed $\mathfrak{K}$, $\Phi(g_{\text{PD}}, \mathcal{C}_A, P)$ and consequently $\Phi(g_{\text{PD}}, \mathcal{C}_A, P_n)$ and $\Phi(D_{\text{PD}}, \mathcal{C}_A, P_n)$ are bounded. Using Inequalities 2 and 3, the boundedness of $\Phi(g_{\text{PD}}, \mathcal{C}_A, P_n)$ and $\Phi(D_{\text{PD}}, \mathcal{C}_A, P_n)$, the Borel-Cantelli lemma and the convergence of hyperharmonic series $\sum i^{-2}$ leads to almost sure convergence of $\sup_{|A|=\mathfrak{K}} |c_n^2 \Phi(g_{sp}, \mathcal{C}_A, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n)|$ to zero.

■

Proof [Theorem 5]. *Overview:* The standard technique is to bound the difference between representation errors with A_n and A^* and show that it converges almost surely to 0. This is done by using the triangle inequality to bring the empirical representation errors into the play. Then the uniform strong law of large numbers for the representation error shows the convergence of the empirical quantization centers to the optimal quantization centers (see, e.g., Pollard, 1981). In the proof of this theorem, we deal with extra intermediate terms that need to be bounded using properties of the shortest path distance (see below).

Define $\mathcal{E}_{\mathfrak{K}} = \{A \subset \mathcal{X} \mid |A| = \mathfrak{K}\}$, the set of all possible sets of $\mathfrak{K}$-centers. The set $\mathcal{E}_{\mathfrak{K}}$ is compact under the topology induced by the Hausdorff metric. When we work with the Euclidean distance, the map $A \to \Phi(g_{\|.\|}, \mathcal{C}_{A,\|.\|}, P)$ is continuous on $\mathcal{E}_{\mathfrak{K}}$ (Pollard, 1981). This result can be extended to the continuity of $A \to \Phi(g_{\text{PD}}, \mathcal{C}_{A,\text{PD}}, P)$ when we use the PD-distance as our metric. We assumed that A^* is unique. Hence the almost sure convergence of $\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P)$ to $\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P)$ implies the almost sure convergence of A_n to the optimal $\mathfrak{K}$-means centers A^* with respect to the Hausdorff distance.

Thus, we only need to show that $\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P)$ converges to $\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P)$ almost surely. We prove this by using several intermediate steps. In these steps we alternate between P and P_n, between g_{PD} and g_{sp}, and between $\mathcal{C}_{A,\text{PD}}$ and $\mathcal{C}_{A,sp}$. Also we need to reach A^* from A_n. This is done using an intermediate step that goes from vertices A_n to a set of vertices near to A^*. To this end we define $\tilde{A^*}$ as the nearest subset of samples to A^*:

$$\tilde{A^*} = \{\tilde{v}_j | \tilde{v}_j = \operatorname*{argmin}_{v \in V_n} \| v - A^*(j) \|; j = 1, ..., \mathfrak{K}\}.$$

For $v_j \in A^*$, we also use the notation $\tilde{v}_j = \tilde{A^*}(v_j)$ to denote the corresponding vertex from V_n. Set $c_n = (k/\eta_d n)^{1/d}$ as before. We bound $\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P) - \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P)$ as follows:

$$\begin{aligned}
\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P) - \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P) \leq\ & |\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P) - \Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P_n)| && \text{(4a)}\\
& + \Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_{A_n,sp}, P_n) && \text{(4b)}\\
& + |\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,sp}, P_n) - c_n^2 \Phi(g_{sp}, \mathcal{C}_{A_n,sp}, P_n)| && \text{(4c)}\\
& + c_n^2 \big(\Phi(g_{sp}, \mathcal{C}_{A_n,sp}, P_n) - \Phi(g_{sp}, \mathcal{C}_{\tilde{A^*},sp}, P_n)\big) && \text{(4d)}\\
& + c_n^2 \big(\Phi(g_{sp}, \mathcal{C}_{\tilde{A^*},sp}, P_n) - \Phi(g_{sp}, \mathcal{C}_{\tilde{A^*},\text{PD}}, P_n)\big) && \text{(4e)}\\
& + |c_n^2 \Phi(g_{sp}, \mathcal{C}_{\tilde{A^*},\text{PD}}, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_{\tilde{A^*},\text{PD}}, P_n)| && \text{(4f)}\\
& + |\Phi(g_{\text{PD}}, \mathcal{C}_{\tilde{A^*},\text{PD}}, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n)| && \text{(4g)}\\
& + |\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P)|. && \text{(4h)}
\end{aligned}$$

Now we show that every term on the right hand side is either non-positive or almost surely converges to zero.
(4a) The proof of the uniform strong law of large numbers (SLLN) for $\Phi(g_{\|.\|}, \mathcal{C}_{.,\|.\|}, P)$ (see e.g. Section 4 in Pollard 1981) holds for all metric spaces with compact closed balls. This results in the uniform SLLN for $\Phi(g_{\text{PD}}, \mathcal{C}_{.,\text{PD}}, P)$ (uniform over all set of $\mathfrak{K}$-centers) and almost sure convergence of $\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P_n)$ to $\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P)$. So

$$|\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P) - \Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P_n)| \xrightarrow{a.s.} 0.$$

We present another proof of this statement in Appendix A.
(4b) By definition, $\mathcal{C}_{A,\text{PD}}(x)$ is the centroid closest to x with respect to the PD-distance, and therefore

$$\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P_n) \leq \Phi(g_{\text{PD}}, \mathcal{C}_{A_n,sp}, P_n).$$

(4c) By Lemma 7,

$$\sup_{A_n} |\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,sp}, P_n) - c_n^2 \Phi(g_{sp}, \mathcal{C}_{A_n,sp}, P_n)| \xrightarrow{a.s.} 0.$$

(4d) The centers A_n are the optimal $\mathfrak{K}$-medoids centers with respect to the shortest path distance, so

$$\Phi(g_{sp}, \mathcal{C}_{A_n,sp}, P_n) - \Phi(g_{sp}, \mathcal{C}_{\tilde{A}^*,sp}, P_n) \leq 0.$$

(4e) $\mathcal{C}_{A,sp}(x)$ is defined as the centroid closest to x with respect to the PD-distance, so

$$\Phi(g_{sp}, \mathcal{C}_{\tilde{A}^*,sp}, P_n) \leq \Phi(g_{sp}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n).$$

(4f) From Lemma 7,

$$\sup_{\tilde{A}^*} |c_n^2 \Phi(g_{sp}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n)| \xrightarrow{a.s.} 0.$$

(4g) By the construction of $\tilde{A}^*$, the distance between $\tilde{A}^*$ and A^* decreases as the sample size n increases. Define $\delta = \max_{v \in A^*} D_{\text{PD}}(\tilde{A}^*(v), v)$. It is easy to check that δ almost surely converges to zero as $n \to \infty$. By the squared triangle inequality,

$$\begin{aligned}
\Phi(g_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n) &= \frac{1}{n}\sum D_{\text{PD}}(\mathcal{C}_{\tilde{A}^*,\text{PD}}(X_i), X_i)^2 \leq \frac{1}{n}\sum D_{\text{PD}}(\tilde{A}^*(\mathcal{C}_{A^*,\text{PD}}(X_i)), X_i)^2 \\
&\leq \frac{1}{n}\sum D_{\text{PD}}(\mathcal{C}_{A^*,\text{PD}}(X_i), X_i)^2 + \frac{1}{n}\sum D_{\text{PD}}(\mathcal{C}_{A^*,\text{PD}}(X_i), \tilde{A}^*(\mathcal{C}_{A^*,\text{PD}}(X_i)))^2 \\
&\quad + \frac{2}{n}\sum D_{\text{PD}}(\mathcal{C}_{A^*,\text{PD}}(X_i), X_i) D_{\text{PD}}(\mathcal{C}_{A^*,\text{PD}}(X_i), \tilde{A}^*(\mathcal{C}_{A^*,\text{PD}}(X_i))) \\
&\leq \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n) + \delta^2 + 2\delta\Phi(D_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n).
\end{aligned}$$

Similarly, we get

$$\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n) \leq \Phi(g_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n) + \delta^2 + 2\delta\Phi(D_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n).$$

Note that $\Phi(D_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n)$ and $\Phi(D_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n)$ almost surely converge to $\Phi(D_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P)$ and $\Phi(D_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P)$, which are bounded. Therefore,

$$|\Phi(g_f, \mathcal{C}_{\tilde{A}^*,f}, P_n) - \Phi(g_f, \mathcal{C}_{A^*,f}, P_n)| \xrightarrow{a.s.} 0.$$

(4h) By the strong law of large numbers for $\Phi(g_{\text{PD}}, \mathcal{C}_{\cdot,\text{PD}}, P)$, we have

$$|\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P)| \xrightarrow{a.s.} 0.$$

So all in all we get $|\Phi(g_{\text{PD}}, \mathcal{C}_{A_n,\text{PD}}, P) - \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P)| \xrightarrow{a.s.} 0$, which finishes the proof. ■

4.3.2. CONVERGENCE FOR $\mathfrak{K} \to \infty$

So far we only proved the convergence results for fixed $\mathfrak{K}$. However, Theorem 4 holds for $\mathfrak{K} \to \infty$. As we work with empirical centers, it is necessary that $\mathfrak{K}$, k and n go to infinity together. The next theorem specifies the interplay between these parameters. It shows that if all parameters go to infinity at an appropriate speed, the empirical $\mathfrak{K}$-medoids centers converge to the optimal $\mathfrak{K}$-means centers and will resemble the original density.

Theorem 8 ($\mathfrak{K}$-medoids centers resemble the original density)
Consider the unweighted kNN graph G_n based on the i.i.d. sample $X_1, ..., X_n \in \mathcal{X}$ drawn from the density p. Assume that for all $x \in \mathcal{X}$ we have $0 < p_{\min} \leq p(x) \leq p_{\max} < \infty$, and let $\alpha > 0$ be a constant. We choose $\mathfrak{K}, k$ and n such that $k \geq \log(n)^{1+\alpha}$, $k \log(n)/n \to 0$ and $\mathfrak{K}k/n \to 0$. Let A_n denote the optimal empirical $\mathfrak{K}$-medoids in graph G_n with respect to the shortest path distance. Assume that the set of optimal $\mathfrak{K}$-means centers with respect to the PD*-distance is unique. Then as $\mathfrak{K}, k$ and n tend to infinity, the empirical $\mathfrak{K}$-medoids centers A_n follow the density p.*

The conditions on k and $\mathfrak{K}$ have intuitive interpretations. The condition on k is a bit stronger than the usual condition $k \geq C \log(n)$ that guarantees connectivity in random kNN graphs. The condition on $\mathfrak{K}$ also has an intuitive meaning. If we choose a large $\mathfrak{K}$, say $\mathfrak{K} \approx n/\log(n)$, each center would only have around $\log(n)$ points in its own cell. Thus each center is connected to almost all points inside the cell with a path of length 1. Therefore, the shortest paths inside the cells are not good approximations for PD-distances. The condition $\mathfrak{K}k/n \to 0$ ensures that for each center, many of the points inside the cell has shortest path distance $\omega(1)$. As a rule of thumb, we can set $k \approx \log(n)$ and choose $\mathfrak{K}$ smaller than $n/\log(n)^2$.
The proof of this theorem needs a more careful investigation of Equation (4). We have to show that each term on the right-hand side converges to zero if $\mathfrak{K}, k$ and n go to infinity at the specified speed. For terms (4a) and (4h), this can be done using standard results from the literature (Pollard, 1982; Bartlett et al., 1998). However, the terms (4c) and (4f) only appear in our algorithm and need a separate analysis. Details can be found in Appendix A.

5. Dispersion of optimal centroids

In this section, we show that the optimal centroids spread well in the domain and do not leave a large part of the domain uncovered. Formally, we bound the dispersion of the optimal centroids: the radius of the largest ball in the domain that does not contain any centroid.
We motivate our result by an example. Let u denote the uniform density on the unit square. We want to represent u by a subset of m^2 points from $[0, 1]^2$. Consider two solutions for this problem: select m^2 random samples from the density u or select m^2 points on a grid of width $\Theta(1/m)$. One

can show that the grid solution is a better representation of the density u in the following sense: For the first solution, there exist a circle inside $[0,1]^2$ with radius $\Theta(\sqrt{\log(m)}/m)$ that does not contain any sample point, with high probability (see, e.g., Theorem 1.1 in Penrose, 1999). For the grid solution, the largest radius of such a circle is $\Theta(1/m)$. This shows that points on a grid have a smaller dispersion than a set of random samples.
The next theorem shows that the optimal centers of $\mathfrak{K}$-means with respect to the PD-distance behave similar to grid points: they also have dispersion $\Theta(\mathfrak{K}^{-d})$, where d is the dimensionality of the underlying space. The proof is discussed in Appendix B.

Theorem 9 (Dispersion of optimal centers) *Let $\mathcal{X}$ be a connected and compact subset of $\mathbb{R}^d$ endowed with a Lipschitz continuous density p. Assume that for all $x \in \mathcal{X}$ we have $0 < p_{\min} \leq p(x) \leq p_{\max} < \infty$. Let $A^{*\mathfrak{K}}$ be the set of optimal $\mathfrak{K}$-means centroids with respect to the* PD*-distance. Then there exists a constant c such that the dispersion of $A^{*\mathfrak{K}}$ in $\mathcal{X}$ is $c\mathfrak{K}^{-d}$. The constant c is related to the doubling constant of the underlying space $\mathcal{X}$, which depends on d but not on $\mathfrak{K}$.*

5.1. Application to downsampling geometric graphs

In this section, we use the density preserving quantization algorithm to downsample random geometric graphs. The algorithm is applicable on a general random geometric graph, but we discuss it only for unweighted kNN graphs. Assume that we are given a massive unweighted kNN graph G with n vertices. The vertices of the graph are sampled from a probability density p, but we neither have access to the point locations of vertices, nor to the underlying density. Our task is to "downsample" G to a much smaller graph G' with $n' \ll n$ vertices, i.e. to build a graph G' that "looks like" a kNN graph built on n' samples from the density p. A general downsampling procedure consists of two steps: choosing vertices, and assigning edges. Our main focus in this section is on the first step: How to choose the vertices?
A simple idea is to randomly select n' vertices and connect each vertex to its k' nearest neighbors, where distances are measured by the shortest path distances in the original graph G. If we choose k' in the right range ($k' \geq c\log(n')$ for a sufficiently large c), we will end up with a connected kNN graph [1]. Can we find a sparser solution, that is, with fewer edges?
In Theorem 9, we proved that our density-preserving downsampling algorithm results in samples that are more evenly spaced than a random sample. Here, we show that if we connect each center to a constant number of its nearest neighbors, the graph will be connected. The constant depends on the geometry and the dimension of our underlying space, but not on the number of centers. The next theorem states this for optimal $\mathfrak{K}$-means centers in the continuous case.

Theorem 10 (Sparse neighborhood graphs on optimal centers) *Let $\mathcal{X}$ be a connected and compact subset of $\mathbb{R}^d$ endowed with a Lipschitz continuous density p. Assume that for all $x \in \mathcal{X}$ we have $0 < p_{\min} \leq p(x) \leq p_{\max} < \infty$. Let $A^{*\mathfrak{K}}$ be the set of optimal $\mathfrak{K}$-means centroids with respect to the* PD*-distance. Then there exists a constant c such that the c-nearest neighbor graph (both with respect to the Euclidean distance and the* PD*-distance) built on centroids is connected. The constant c is related to the doubling constant of the underlying space $\mathcal{X}$, which depends on d but not on $\mathfrak{K}$.*

1. One can prove that connecting the sampled points based on their shortest path distance in G produces a graph "similar" to a neighborhood graph built on the points using their actual Euclidean distances. Here, "similar" means that the number of common edges between two graphs dominates the number of edges that do not appear in one of the graphs. The proof is beyond the scope of this paper.

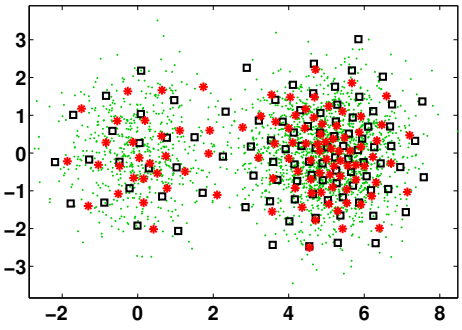
This theorem is in sharp contrast with the fact that for connectivity of a kNN graph, it is necessary that $k \geq C \log(n)$ (see, e.g., Penrose, 1999). The proof of the theorem uses a geometric argument based on the *Delone property* of Voronoi cells corresponding to optimal $\mathfrak{K}$-means centers. It also shows how we can find the constant c: Look at the Voronoi diagram of the optimal centers and set c as the maximum number of facets of a Voronoi cell in that diagram. Details of the proof are discussed in Appendix C.
Now we can put Theorems 8 and 10 together to get our desired result. Namely, if we apply the density-preserving downsampling on unweighted kNN-graphs, the centers will follow the underlying density of the original graph. Moreover, we can build a sparse neighbourhood graph on these $\mathfrak{K}$ downsampled vertices. The total number of edges in this graph will depend linearly on $\mathfrak{K}$.

6. Implementation

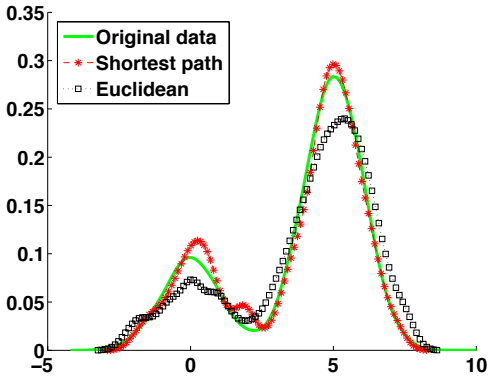
We first suggest a heuristic to simplify the implementation of the quantization procedure when our sample points are embedded in an Euclidean space. Then we discuss the case where we do not have access to the embedding of our sample points. We are just given the adjacency matrix of the kNN graph. Finally, we illustrate an example of our density preserving quantization on a synthetic dataset as a proof of concept.
The classical implementation of $\mathfrak{K}$-means is an EM-type algorithm by Lloyd (1982). After initializing the centers, it iterates over two steps:

1. The assignment or the expectation step: In this step, data points are assigned to the centers.
2. The update or the maximization step: Update the centers given the assignments in Step 1.

It is easy to adapt the assignment step for the shortest path distance. The centers are not necessarily vertices from our graph, so we connect each center to its nearest sample point. Then we assign every point to a nearest center according to the shortest path distance. For the update step, instead of finding the exact center of each Voronoi cell with respect to the shortest path distance, we use a simple heuristic: assume that the density inside each cell is constant. Therefore, the update would be the same as the standard $\mathfrak{K}$-means with respect to Euclidean distance. If there are c_i points in the Voronoi cell i, this reduces the computational complexity of finding the center from $O(c_i^2)$ to $O(c_i)$. This heuristic is useful when c_i is relatively large and applicable when the density of points inside each cell is close to uniform.
As a proof of concept, we apply this algorithm on a synthetic dataset. We draw 12,000 points from an unbalanced mixture of two Gaussian distributions in $\mathbb{R}^2$ and build an unweighted kNN graph with $k = 15$. Then we quantize to $\mathfrak{K} = 100$ centers using the shortest path distance and the Euclidean distance. The results are depicted in Figure 1(*a*). One can see that quantization based on the Euclidean distance tends to put more centers in the low density regions compared to the quantization with respect to the shortest path distance. In the former, centers resemble the density $p(x)^{1/2}$ (Graf and Luschgy, 2000), hence the low density regions are amplified. To compare the density of the centers of standard $\mathfrak{K}$-means and our algorithm, we plot their estimated marginal density in the direction of x-axis in Figure 1(*b*). The density bias of standard $\mathfrak{K}$-means centers is apparent in this figure, whereas the marginal distribution of the centers with respect to the shortest path distance is close to the marginal distribution of the underlying density.
Alternatively, Asgharbeygi and Maleki (2008) propose a relatively simple algorithm for finding centers when there is no access to the embedding of the data. They first show that for proper distance



(*a*) Original data (green), centroids based on the shortest path distance (red) and the Euclidean distance (black).



(*b*) The marginal distribution of samples and quantization centroids in the direction of x-axis, where quantization is with respect to D_{sp} or Euclidean distance.

matrices, one can apply multidimensional scaling (MDS) to find a distance preserving embedding of the points in a higher dimensional Euclidean space and use the Lloyd algorithm in the embedded space. At the end, they show how to do all the computations in the original space without paying the cost of the embedding step.

Acknowledgments

We thank Tamás Linder for pointing out relevant references. This research is partly supported by the German Research Foundation via the Research Unit 1735 "Structural Inference in Statistics: Adaptation and Efficiency" and grant LU1718/1-1. G. Lugosi acknowledges support by the Spanish Ministry of Science and Technology grant MTM2012-37195.

References

M. Alamgir and U. von Luxburg. Shortest path distance in random k-nearest neighbor graphs. In *ICML*, 2012.

N. Asgharbeygi and A. Maleki. Geodesic k-means clustering. In *ICPR*, 2008.

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

A.L. Besse. *Einstein Manifolds*. Classics in Mathematics. Springer, 1987.

G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

V. de Silva and J.B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004.

E.J. Delp and O.R. Mitchell. Moment preserving quantization. *IEEE Transactions on Communications*, 39(11):1549–1558, 1991.

B. Feil and A. Janos. Geodesic distance based fuzzy clustering. *Lecture Notes in Computer Science, Soft Computing in Industrial Applications*, pages 50–59, 2007.

S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Lecture notes in mathematics. Springer-Verlag New York, Inc., 2000.

P. M. Gruber. Optimal configurations of finite sets in Riemannian 2-manifolds. *Geometriae Dedicata*, 84(1-3):271–320, 2001.

P. M. Gruber. Optimum quantization and its applications. *Advances in Mathematics*, 186(2):456 – 497, 2004.

P. Heckbert. Color image quantization for frame buffer display. In *SIGGRAPH*, 1982.

A. Hegde, D. Erdogmus, T. Lehn-Schioler, Y. Rao, and J. Principe. Vector-Quantization by density matching in the minimum Kullback-Leibler divergence sense. In *IEEE International Conference on Neural Networks*, volume 1, pages 105–109, 2004.

M.M. Van Hulle. Faithful representations with topographic maps. *Neural Networks*, 12(6):803 – 823, 1999.

J. Kim, K. Shim, and S. Choi. Soft geodesic kernel k-means. In *ICASSP*, pages 429–432, 2007.

T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Computer Vision*, 43(1):29–44, 2001.

M. Li, J. Klejsa, and W. Bastiaan Kleijn. On distribution preserving quantization. *CoRR*, abs/1108.3728, 2011.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982.

P. Meinicke and H. Ritter. Quantizing density estimators. In *Neural Information Processing Systems (NIPS)*, 2001.

H. Niederreiter. *Random Number Generation and quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

M. Penrose. A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability*, (1), 1999.

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

D. Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, (4), 1982.

Sajama and A. Orlitsky. Estimating and computing density based distance metrics. In *ICML*, 2005.

S. Yakowitz, P. L'Ecuyer, and F. Vázquez-Abad. Global stochastic optimization with low-dispersion point sets. *Operational Research*, 48:939–950, 2000.

Appendix A. Proof of Theorem 8

The first step of the proof is similar to the proof of Theorem 5, up to Equation (4). We show that each term on the right-hand side of that equation converges to zero if $k, \mathfrak{K}$ and n go to infinity at specified rates.

(4a) Let $\mathcal{F}$ denote the class of functions

$$\mathcal{F}_{\mathfrak{K}} = \{f_A(x) | f_A(x) = D_{\text{PD}}(x, \mathcal{C}_A(x))^2, A \subset \mathcal{X}, |A| = \mathfrak{K}\},$$

and $\mathcal{N}(t, \mathcal{F}_{\mathfrak{K}})$ the covering number of $\mathcal{F}_{\mathfrak{K}}$ with respect to the supremum norm. To cover $\mathcal{X}$ with PD-balls of radius t, we need $\Theta(t^{-d})$ such balls (see Lemma 7 in Alamgir and von Luxburg, 2012). This means that $\mathcal{N}(t, \mathcal{F}_{\mathfrak{K}}) = \Theta(t^{-\mathfrak{K}d})$. We also have

$$P(\sup_A \left|\Phi(g_{\text{PD}}, \mathcal{C}_{A,\text{PD}}, P) - \Phi(g_{\text{PD}}, \mathcal{C}_{A,\text{PD}}, P_n)\right| > t) \leq \mathcal{N}(t, \mathcal{F}_{\mathfrak{K}}) e^{-nt^2/B},$$

where B is a constant depending on the diameter of $\mathcal{X}$. This shows that the difference converges to zero with high probability (probability converging to 1) if $\mathfrak{K} \log(n)/n \to 0$.

(4c) We revisit Lemma 7 for $\mathfrak{K} \to \infty$. We rewrite the proof of the lemma for a general $\mathcal{C}_A$:

$$\frac{1}{(1-\lambda)^{2/d}} \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n) \leq c_n^2 \Phi(g_{sp}, \mathcal{C}_A, P_n) \leq e^2 \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n) + c_n^2 + 2c_n e \Phi(D_{\text{PD}}, \mathcal{C}_A, P_n).$$

Set $\mathcal{C}_A = \mathcal{C}_{A,sp}$. We need to show that as $\mathfrak{K}, k$ and n go to infinity,

$$\frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A,sp}, P_n)}{c_n^2} \to \infty \quad \text{and} \quad \frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A,sp}, P_n)}{c_n e \Phi(D_{\text{PD}}, \mathcal{C}_{A,sp}, P_n)} \to \infty.$$

From Theorem 1 in Gruber (2004), we know that

$$\Phi(g_{\text{PD}}, \mathcal{C}_{A,sp}, P_n) = \Theta(\frac{1}{\mathfrak{K}^{2/d}}) \quad \text{and} \quad \Phi(D_{\text{PD}}, \mathcal{C}_{A,sp}, P_n) = \Theta(\frac{1}{\mathfrak{K}^{1/d}}).$$

Recalling $c_n = (\frac{k}{n\eta_d})^{1/d}$, we get

$$\frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A,sp}, P_n)}{c_n^2} = \Theta\big((\frac{\eta_d n}{\mathfrak{K}k})^{2/d}\big) \quad \text{and} \quad \frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A,sp}, P_n)}{c_n e \Phi(D_{\text{PD}}, \mathcal{C}_{A,sp}, P_n)} = \Theta\big((\frac{\eta_d n}{\mathfrak{K}k})^{1/d}\big).$$

This shows that $|c_n^2 \Phi(g_{sp}, \mathcal{C}_A, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_A, P_n)|$ almost surely converges to zero if $\mathfrak{K}k/n \to 0$.

(4f) Similar to Part (4c).

(4g) Consider the proof of the corresponding part in Theorem 5. We need to show that

$$\frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n)}{\delta^2} \to \infty \quad \text{and} \quad \frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n)}{\delta \Phi(D_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n)} \to \infty.$$

Instead of bounding δ, it is easier to bound $\delta' = \sup_{x \in \mathcal{X}, v \in V_n} \|x - v\|$. Using a standard sphere packing lemma, we know that $\delta' \leq (k/n)^{1/d}$ with probability converging to 1. For large enough n, we have

$$\delta = \max_{v \in A^*} D_{\text{PD}}(\tilde{A}^*(v), v) \leq 2p_{\max}^{1/d} \max_{v \in A^*} \|\tilde{A}^*(v) - v\| \leq 2p_{\max}^{1/d} \delta'.$$

This means that, with high probability, $\delta \leq 2p_{\max}^{1/d}(k/n)^{1/d}$. Similar to Part (4c), we have

$$\frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n)}{\delta^2} = \Theta\big((\frac{n}{k\mathfrak{K}})^{2/d}\big) \quad \text{and} \quad \frac{\Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n)}{\delta\Phi(D_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n)} = \Theta\big((\frac{n}{k\mathfrak{K}})^{1/d}\big).$$

This shows that $|\Phi(g_{\text{PD}}, \mathcal{C}_{\tilde{A}^*,\text{PD}}, P_n) - \Phi(g_{\text{PD}}, \mathcal{C}_{A^*,\text{PD}}, P_n)|$ almost surely converges to zero if $\mathfrak{K}k/n \to 0$.
(4h) Similar to Equation (4a), this term converges to zero if $\mathfrak{K}\log(n)/n \to 0$.

Appendix B. Proof of Theorem 9

We recall a proposition from Gruber (2004) on the shape of optimal Voronoi cells.

Proposition 11 (Optimal Voronoi cells are Delone) *Let $\mathcal{X}$ be a compact and differentiable Riemannian d-manifold. Let $A^{*\mathfrak{K}}$ be the optimal quantization centroids with respect to the Riemannian metric ϱ. Then there exist constants $a, b > 0$ such that $A^{*\mathfrak{K}}$ is $(a\mathfrak{K}^{-1/d}, b\mathfrak{K}^{-1/d})$-Delone. This means that*

- *Every two distinct centers of $A^{*\mathfrak{K}}$ have distance at least $a\mathfrak{K}^{-1/d}$.*
- *For each point of $\mathcal{X}$, there exists a center in $A^{*\mathfrak{K}}$ at distance at most $b\mathfrak{K}^{-1/d}$.*

Constants a and b depend on d and the geometry of $\mathcal{X}$, but not on $\mathfrak{K}$.

Note that this proposition is not asymptotic and holds for every $\mathfrak{K}$. The proof of Theorem 9 only needs the second part of this proposition.
Proof Let $B^*(x, r)$ be the largest ball inside $\mathcal{X}$ that does not contain a centroid. Here x denotes the center of the ball and r is its radius. From the second part of Proposition 11, there exists an optimal centroid in $A^{*\mathfrak{K}}$ at distance at most $b\mathfrak{K}^{-1/d}$ from x. This shows that $r \leq b\mathfrak{K}^{-1/d}$ which finishes the proof. ■

Appendix C. Proof of Theorem 10

Consider the Voronoi diagram induced by the optimal centers. Connect each center to the centers of all neighbor cells to attain a connected graph. Let c be the maximum degree in this graph. We extend this graph to a nearest neighbor graph with neighborhood size c. Consider a vertex v_i with degree d_i. Extend the neighborhood of v_i by connecting it to its next $c - d_i$-nearest neighbors. This extension can be done with respect to the Euclidean distance (when the coordinates of vertices are available) or with respect to the PD-distance. We show that c is a constant independent of $\mathfrak{K}$, which proves the theorem.
The Voronoi cells corresponding to optimal centers $A^{*\mathfrak{K}}$ cannot be very thin or very long (Proposition 11). This property and a sphere packing lemma are used to bound the number of neighbors of each Voronoi cell.
Let $B_\varrho(x, r) = \{y | \varrho(x, y) \leq r\}$ denote the closed ϱ-ball with radius r and center x. Also denote the Voronoi cell of an optimal center $a_i \in A^{*\mathfrak{K}}$ by D_i. The next lemma provides the necessary tools for our proof.

Lemma 12 *Consider the setting in Proposition 11 and let D_i and D_j be two neighboring Voronoi cells.*

- *The neighbor centers are not far from each other: $\varrho(a_i, a_j) \leq 2b\mathfrak{K}^{-1/d}$.*
- *The cell D_j is inside the ϱ-ball around a_i with radius $3b\mathfrak{K}^{-1/d}$: $D_j \subset B_\varrho(a_j, 3b\mathfrak{K}^{-1/d})$.*
- *Voronoi cells are fat: The ϱ-ball with radius $0.5a\mathfrak{K}^{-1/d}$ around a_i is completely inside D_i*

$$B_\varrho(a_i, 0.5a\mathfrak{K}^{-1/d}) \subset D_i.$$

Proof *Part 1*. Consider the shortest path between a_i and a_j that passes through a boundary point between D_i and D_j. Let m be the intersection of this path with the boundary between D_i and D_j. We show that $\varrho(a_i, m) \leq b\mathfrak{K}^{-1/d}$. If not, from Part 2 of Proposition 11, there exist a center a_k such that $\varrho(a_k, m) \leq b\mathfrak{K}^{-1/d}$. This means that $\varrho(a_k, m) < \varrho(a_i, m)$, which contradicts the fact that m is in D_i. Similarly we have $\varrho(a_j, m) \leq b\mathfrak{K}^{-1/d}$. Use the triangle inequality to get

$$\varrho(a_i, a_j) \leq \varrho(a_i, m) + \varrho(m, a_j) \leq 2b\mathfrak{K}^{-1/d}.$$

Part 2. Consider a point $x \in D_j$. Similar to Part 1, we can show that $\varrho(a_j, x) \leq b\mathfrak{K}^{-1/d}$. Using the triangle inequality, we have

$$\varrho(a_i, x) \leq \varrho(a_i, a_j) + \varrho(a_j, x) \leq 3b\mathfrak{K}^{-1/d}.$$

Part 3. If not, there exists a point x such that $\varrho(a_i, x) \leq 0.5a\mathfrak{K}^{-1/d}$ but $x \notin D_i$. Therefore, the point x is inside a cell D_l with center a_l such that $\varrho(a_l, x) < \varrho(a_i, x)$. This means that

$$\varrho(a_i, a_l) \leq \varrho(a_i, x) + \varrho(x, a_l) < a\mathfrak{K}^{-1/d},$$

which is in contradiction with Part 1 of Proposition 11. ■

Consider an optimal center a_i and denote the set of centers of all neighboring cells by A_i. The PD-balls with radius $0.5a\mathfrak{K}^{-1/d}$ around centers in A_i are all disjoint. These balls are also completely inside $B_{\text{PD}}(a_i, 3b\mathfrak{K}^{-1/d})$, thus

$$p\big(B_{\text{PD}}(a_i, 3b\mathfrak{K}^{-1/d})\big) \geq \sum_{v \in A_i} p\big(B_{\text{PD}}(v, 0.5a\mathfrak{K}^{-1/d})\big) \geq |A_i| p_{\min} e_1 \mathfrak{K},$$

where e_1 is a constant (depending on d). Also

$$p\big(B_{\text{PD}}(a_i, 3b\mathfrak{K}^{-1/d})\big) \leq e_2 p_{\max} \mathfrak{K},$$

for a constant e_2. All in all, we have $|A_i| \leq \frac{p_{\max} e_2}{p_{\min} e_1}$ which is a constant independent of $\mathfrak{K}$.