Graph Quantization

Brijnesh J. Jain[1] and Klaus Obermayer[1]

[1]Berlin Institute of Technology, Germany
{jbj|oby}@cs.tu-berlin.de

Abstract. Vector quantization(VQ) is a lossy data compression technique from signal processing, which is restricted to feature vectors and therefore inapplicable for combinatorial structures. This contribution presents a theoretical foundation of graph quantization (GQ) that extends VQ to the domain of attributed graphs. We present the necessary Lloyd-Max conditions for optimality of a graph quantizer and consistency results for optimal GQ design based on empirical distortion measures and stochastic optimization. These results statistically justify existing clustering algorithms in the domain of graphs. The proposed approach provides a template of how to link structural pattern recognition methods other than GQ to statistical pattern recognition.

1 Introduction

Vector quantization is a classical technique from signal processing suitable for lossy data compression, density estimation, and prototype-based clustering [7, 14, 30]. The problem of optimal vector quantizer design is to find a codebook consisting of a finite set of prototypes such that an expected distortion with respect to some (differentiable) distortion measure is minimized.

Since the probability distribution of the input patterns is usually unknown, vector quantizer design techniques use empirical data. Extensively studied design techniques are, for example, k-means and simple competitive learning. The k-means algorithm is also commonly referred to as the Linde-Buzo-Gray (LBG) algorithm [24] the generalized Lloyd algorithm [25]. This algorithm is a local optimizer of the empirical sum-of-squared-error distortion without any global optimal or consistency guarantees. In contrast to k-means, competitive learning directly minimizes the expected distortion and is a consistent learner under very general conditions in the sense that it almost surely converges to a local optimal solution of the expected distortion.

One limitation of VQ is its restriction to patterns that are represented by vectors. For patterns that are more naturally represented by finite combinatorial structures, the theoretical framework of VQ as well its design techniques are no longer applicable. Examples of such structures include, for example, point patterns, strings, trees, and graphs arising from diverse application areas like proteomics, chemoinformatics, and computer vision.

To overcome this limitation, we generalize vector quantization to quantization of graphs. A number of graph quantizer design techniques for the purpose of prototype-based clustering have already been proposed. Examples include competitive learning algorithms in the domain of graphs [16–20, 22] and k-means as well as k-medoids algorithms [12, 13, 19, 20, 23, 28, 29]. Related clustering method are presented in [3, 26, 31]. Due to a lack of an appropriate theoretical framework, all these graph quantizer design techniques (or clustering methods) have been developed in order to minimize an empirical distortion function without justifying whether the solutions found are statistically consistent estimators of the true but unknown solutions. In addition, it is unclear whether the nearest neighbor and centroid condition, which are also referred to as the Lloyd-Max conditions, are necessary conditions for optimality.

In this contribution, we propose graph quantization in a mathematically principled way as an extension of vector quantization, where we consider the graph edit distance as an underlying graph distortion measure. The key results of this contribution are consistency statements for estimators based on empirical distortion measures and estimators based on stochastic optimization. Furthermore, we prove that the Llyod-Max conditions are also necessary condition for optimal graph quantizers. In order to achieve the consistency results and the Lloyd-Max conditions, we isometrically embed – without loss of structural information – graphs as points into some Riemannian orbifold. An orbifold is the quotient of a manifold by a finite group action and therefore generalizes the notion of manifold. Using orbifolds we can define geometric and analytic concept such as length, angle, derivative, gradient, and integral locally to a Euclidean space. This construction forms the basis for extending consistency results from Euclidean vector spaces to the domain of graphs.

The proposed approach has the following properties: First, it can be applied to finite combinatorial structures other than graphs like, for example, point patterns, sequences, trees, and hypergraphs. For the sake of concreteness, we restrict our attention exclusively to the domain of graphs. Second, for graphs consisting of a single vertex with feature vectors as attributes, graph quantization coincides with vector quantization. Third, the proposed consistency results justify some of the above referenced graph clustering methods as statistically consistent learners. Fourth, the underlying mathematical framework can be applied in order to link other structural pattern recognition methods that directly operate in the domain of graphs to methods from statistical pattern recognition.

The paper is organizes as follows. Section 2 describes the problem of graph quantizer design. Section 3 introduces Riemannian orbifolds. In Section 4, we extend VQ to GQ and present consistency result for GQ design techniques. Section 5 briefly discusses the case of general graph edit distance functions. Finally, Section 6 concludes.

2 The Problem of Graph Quantizer Design

This section aims at outlining the problem of extending VQ to the quantization of graphs.

2.1 Attributed Graphs

To begin with, we first describe the structures we want to quantize.

Let $\mathcal{A}$ be a set of *attributes* and let $\varepsilon \in \mathcal{A}$ be a distinguished element denoting the *null* or *void* element. An *attributed graph* is a tuple $X = (V, \alpha)$ consisting of a finite nonempty set V of *vertices* and an *attribute function* $\alpha : V \times V \to \mathcal{A}$. Elements of the set

$$E = \{(i,j) \in V \times V \,:\, i \neq j \text{ and } \alpha(i,j) \neq \varepsilon\}$$

are the *edges* of X. By $\mathcal{G}_{\mathcal{A}}$ we denote the set of all attributed graphs with attributes from $\mathcal{A}$. The vertex set of an attributed graph X is often referred to as V_X and its attribute function as α_X.

An *alignment* of a graph X is a graph X' with $V_X \subseteq V_{X'}$ and

$$\alpha_{X'}(i,j) = \begin{cases} \alpha_X(i,j) & : \quad (i,j) \in V_X \times V_X \\ \varepsilon & : \quad \text{otherwise} \end{cases}$$

for all $i, j \in V_{X'}$. Thus, we obtain an alignment of X by adding isolated vertices with null-attribute. The set $V_{X'} \setminus V_X$ is the set of *aligned vertices.* By $\mathcal{A}(X)$ we denote the (infinite) set of all alignments of X.

A *pairwise alignment* of graphs X and Y is a triple (ϕ, X', Y') consisting of alignments $X' \in \mathcal{A}(X)$ and $Y' \in \mathcal{A}(Y)$ together with a bijective mapping

$$\phi : V_{X'} \to V_{Y'}, \quad i \mapsto i^{\phi}.$$

By $\mathcal{A}(X,Y)$ we denote the set of all pairwise alignments between X and Y. Sometimes we briefly write ϕ instead of (ϕ, X', Y').

2.2 The Graph Edit Distance

Fundamental for quantizing data is the notion of distortion. This section briefly introduces the graph edit distance functions as our choice of distortion measure. For a more detailed definition of the graph edit distance, we refer to [2]. In addition, we present an important graph metric based on a generalization of the concept of maximum common subgraph, which arises in various different guises as a common choice of proximity measure [1, 5, 6, 15, 32, 33]. For sake of convenience, we assume that all distances are metrics.

Each pairwise alignment $(\phi, X', Y') \in \mathcal{A}(X,Y)$ can be regarded as an edit path with cost

$$d_{\phi}\,(X,Y) = \sum_{i,j \in V_{X'}} d_{\mathcal{A}}\left(\alpha_{X'}(i,j), \alpha_{Y'}(i^{\phi}, j^{\phi})\right),$$

where $d_{\mathcal{A}} : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$ is a distance function defined on the set $\mathcal{A}$ of attributes. Observe that deletion (insertion) of vertices also deletes (inserts) all edges the respective vertices are incident to.

The *graph edit distance* of X and Y is then defined as the edit path with minimal cost

$$d(X,Y) = \min\left\{d_\phi\left(X,Y\right) \,:\, \phi \in \mathcal{A}(X,Y)\right\}.$$

Note that the set $\mathcal{A}(X,Y)$ of pairwise alignments is of infinite cardinality. But since $d_\mathcal{A}(\varepsilon,\varepsilon) = 0$, we actually take the minimum over a finite subset by ignoring all pairwise alignments that map aligned vertices with null-attributes onto each other.

Next, we consider an important example of the graph edit distance based on a generalization of the concept of maximum common subgraph. We derive this graph metric from a similarity measure in the same way the Euclidean distance is derived from an inner product.

Suppose that $k_\mathcal{A} : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$ with $k_\mathcal{A}(\cdot,\varepsilon) = 0$ is a positive definite kernel. We measure the quality of a pairwise alignment $\phi \in \mathcal{A}(X,Y)$ by

$$k_\phi(X,Y) = \sum_{i,j \in V_X} k_\mathcal{A}\left(\alpha_X(i,j), \alpha_Y(i^\phi, j^\phi)\right).$$

An *optimal alignment kernel* is a graph similarity measure of the form

$$k(X,Y) = \max\left\{k_\phi(X,Y) \,:\, \phi \in \mathcal{A}(X,Y)\right\}. \tag{1}$$

Note that $k\left(\cdot|\cdot\right)$ is symmetric but indefinite as a pointwise maximizer of a set of positive definite kernels.

The distance metric on $\mathcal{G}_\mathcal{A}$ induced by an optimal alignment kernel $k\left(\cdot|\cdot\right)$ is defined by

$$d(X,Y) = \sqrt{l(X)^2 - 2k(X,Y) + l(Y)^2}, \tag{2}$$

where $l(X) = \sqrt{k(X,X)}$ denotes the *length* of an attributed graph X. As shown in [23], d is indeed a metric and can be expressed as a graph edit distance.

2.3 The Problem of Graph Quantizer Design

Let $(\mathcal{G}_\mathcal{A}, d)$ be a graph distance space, where $d\left(\cdot|\cdot\right)$ is a graph edit distance. Optimal graph quantization design aims at minimizing the expected distortion

$$D(\mathcal{C}) = \int_{\mathcal{G}_\mathcal{A}} d(X, Q(X))\, dP(X),$$

where $Q : \mathcal{G}_\mathcal{A} \rightarrow \mathcal{C}$ is a graph quantizer, $\mathcal{C} = \{Y_1, \ldots, Y_k\}$ a codebook consisting of k code graphs, and $P = P_{\mathcal{G}_\mathcal{A}}$ is a probability measure defined on some appropriate measurable space $(\mathcal{G}_\mathcal{A}, \Sigma_{\mathcal{G}_\mathcal{A}})$.

As opposed to vector quantization, the following factors complicate designing an optimal graph quantizer in a statistically consistent way:

1. The graph distance $d(X,Y)$ is in general non-convex and non-differentiable.
2. Neither a well-defined addition on graphs nor the notion of derivative for functions on graphs is known.

To overcome these difficulties, we isometrically embed graphs as points into a Riemannian orbifold in order to apply methods that generalize gradient descent techniques and methods from stochastic optimization for non-convex and non-differentiable distortion functions.

3 Riemannian Orbifolds

Orbifolds generalize the notion of manifold as locally being a quotient of $\mathbb{R}^n$ by finite group actions. Consequently, learning on orbifolds generalizes learning on Euclidean spaces and Riemannian manifolds. This section introduces Riemannian orbifolds and their intrinsic metric structure. Proofs for new results are delegated to Section B.1. For all other proofs we refer to [4, 21].

3.1 Riemannian Orbifolds

To keep the treatment simple, we assume that $\mathcal{X} = \mathbb{R}^n$ is the n-dimensional Euclidean vector space, and $\mathit{\Gamma}$ is a permutation group acting on $\mathcal{X}$. In a more general setting, however, we can assume that $\mathcal{X}$ is a Riemannian manifold, and $\mathit{\Gamma}$ is a finite group of isometries acting effectively on $\mathcal{X}$.

The binary operation

$$\cdot : \mathit{\Gamma} \times \mathcal{X} \to \mathcal{X}, \quad (\gamma, \boldsymbol{x}) \mapsto \gamma(\boldsymbol{x})$$

is a group action of $\mathit{\Gamma}$ on $\mathcal{X}$. For $\boldsymbol{x} \in \mathcal{X}$, the *orbit* of $\boldsymbol{x}$ is the set defined by

$$[\boldsymbol{x}] = \{\gamma(\boldsymbol{x}) \,:\, \gamma \in \mathit{\Gamma}\}\,.$$

The quotient set

$$\mathcal{X}_{\mathit{\Gamma}} = \mathcal{X}/\mathit{\Gamma} = \{[\boldsymbol{x}] \,:\, \boldsymbol{x} \in \mathcal{X}\}$$

consisting of all all orbits carries the structure of a *Riemannian orbifold.* Its *orbifold chart* is the surjective continuous mapping

$$\pi : \mathcal{X} \to \mathcal{X}_{\mathit{\Gamma}}, \quad \boldsymbol{x} \mapsto [\boldsymbol{x}]$$

that projects each point $\boldsymbol{x}$ to its orbit $[\boldsymbol{x}]$.

In the following, an orbifold is a triple $\mathcal{Q} = (\mathcal{X}, \mathit{\Gamma}, \pi)$ consisting of an Euclidean space $\mathcal{X}$, a permutation group $\mathit{\Gamma}$ acting on $\mathcal{X}$ and its orbifold chart π. With $\mathit{\Gamma} = \{\text{id}\}$ being the trivial permutation group consisting of the identity only, a manifold $\mathcal{X}$ is also an orbifold. In general, however, the underlying space $\mathcal{X}_{\mathit{\Gamma}}$ of an orbifold is not a manifold. Thus, orbifolds generalize the notion of manifold. The points at which an orbifold $\mathcal{X}_{\mathit{\Gamma}}$ is locally not homeomorphic to a manifold are its *singular points.* We call the elements of $\mathcal{X}_{\mathit{\Gamma}}$ *structures*, since they represent combinatorial structures like attributed graphs. We use capital letters $X, Y, Z, \ldots$ to denote structures from $\mathcal{X}_{\mathit{\Gamma}}$ and write, by abuse of notation, $\boldsymbol{x} \in X$ if $\pi(\boldsymbol{x}) = X$. Each vector $\boldsymbol{x} \in X$ is a *vector representation* of structure X and the set $\mathcal{X}$ of all vector representation is the *representation space* of $\mathcal{X}_{\mathit{\Gamma}}$.

Example 1. Let $\mathcal{X} = \mathbb{R}^2$ and let Γ be the group generated by reflections across the main-diagonal of the x-y-plane. Then $\mathcal{Q} = (\mathcal{X}_\Gamma, \Gamma, \pi)$ is a Riemannian orbifold with

$$\pi : \mathcal{X} \to \mathcal{X}_\Gamma, \quad \boldsymbol{x} = (x_1, x_2) \mapsto [\boldsymbol{x}] = \{(x_1, x_2), (x_2, x_1)\}\,.$$

The singular points of $\mathcal{X}_\Gamma$ are all structures X represented by vectors $\boldsymbol{x} = (x_1, x_2)$ with $x_1 = x_2$.

3.2 The Riemannian Orbifold of Attributed Graphs

In this section, we show that attributes graphs can be identified with points in some Riemannian orbifold.

Riemannian orbifolds of attributed graphs arise by considering equivalence classes of matrices representing the same graph. To identify graphs with points in a Riemannian orbifold without loss of structural information, some technical assumptions and restrictions to simplify the mathematical treatment are necessary. For this, let $(\mathcal{G}_\mathcal{A}, d)$ be a graph distance space with graph edit distance $d(\cdot|\cdot)$. Then we make the following assumptions:

P1 There is a feature map $\Phi : \mathcal{A} \to \mathcal{H}$ of the attributes into some finite dimensional Euclidean feature space $\mathcal{H}$ and a distance function $d_\mathcal{H} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$ such that $\Phi(\varepsilon) = \mathbf{0} \in \mathcal{H}$ and

$$d_\mathcal{A}(a, a') = d_\mathcal{H}(\Phi(a), \Phi(a'))$$

for all attributes $a, a' \in \mathcal{A}$.

P2 All graphs are finite of bounded order n, where n is a sufficiently large number. Graphs X of order less than n, say $m < n$, are aligned to graphs X' of order n by inserting $p = n - m$ isolated vertices with null attribute ε.

Before discussing the impact of both assumptions for practical application, we first restate our first assumptions for graph metrics induced by optimal alignment kernels. By definition $k_\mathcal{A} : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is a positive definite kernel corresponding to an inner product $k_\mathcal{A}(x, y) = \langle \Phi(x), \Phi(y) \rangle$ in some feature space $\mathcal{H}$. Our first assumption requires that $\mathcal{H}$ is a finite dimensional Euclidean space and $\Phi(\varepsilon) = \mathbf{0}$.

Now let us consider the above assumptions in more detail. Both conditions do not effect the graph edit distance, provided an appropriate feature map for the attributes can be found. Restricting to finite dimensional Euclidean feature spaces $\mathcal{H}$ is necessary for deriving consistency results and for applying methods from stochastic optimization. Limiting the maximum size of the graphs to some arbitrarily large number n and aligning smaller graphs to graphs of oder n are purely technical assumptions to simplify mathematics. For machine learning problems, this limitation should have no practical impact, because neither the bound n needs to be specified explicitly nor an extension of all graphs to an identical order needs to be performed. When applying the theory, all we actually require is that the order of the graphs is bounded.

With both assumptions in mind, we construct the Riemannian orbifold of attributed graphs. Let $\mathcal{X} = \mathcal{H}^{n \times n}$ be the set of all $(n \times n)$-matrices with elements from feature space $\mathcal{H}$. A graph X is completely specified by a *representation matrix* $\boldsymbol{X} = (\boldsymbol{x}_{ij})$ from $\mathcal{X}$ with elements

$$\boldsymbol{x}_{ij} = \begin{cases} \phi\left(\mu_X(i)\right) & : \quad i = j \\ \phi\left(\nu_X(i,j)\right) & : \quad (i,j) \in E \\ \mathbf{0} & : \quad \text{otherwise} \end{cases}$$

for all $i, j \in V_X$. The form of a representation matrix $\boldsymbol{X}$ of X is generally not unique and depends on how the vertices are arranged in the diagonal of $\boldsymbol{X}$.

Now suppose that Π^n be the set of all $(n \times n)$-permutation matrices. For each $\boldsymbol{P} \in \Pi^n$ we define a mapping

$$\gamma_{\boldsymbol{P}} : \mathcal{X} \to \mathcal{X}, \quad \boldsymbol{X} \mapsto \boldsymbol{P}^{\mathsf{T}} \boldsymbol{X} \boldsymbol{P}.$$

Then $\Gamma = \{\gamma_{\boldsymbol{P}} : \boldsymbol{P} \in \Pi^n\}$ is a permutation group acting on $\mathcal{X}$. Regarding an arbitrary matrix $\boldsymbol{X}$ as a representation of some graph X, then the orbit $[\boldsymbol{X}]$ consists of all possible matrices that can represent X. By identifying the orbits of $\mathcal{X}_\Gamma$ with attributed graphs, the set $\mathcal{G}_{\mathcal{A}}$ of attributed graphs of bounded order n is a Riemannian orbifold.

3.3 Metric Structures

Let $\mathcal{Q} = (\mathcal{X}, \Gamma, \pi)$ be an orbifold. We derive an intrinsic metric that enables us to do Riemannian geometry. In the case of a Riemannian orbifold of attributed graphs the intrinsic metric coincides with the graph metric of (2) induced by an optimal alignment kernel.

Any inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{X}$ gives rise to a maximizer of the form

$$k : \mathcal{X}_\Gamma \times \mathcal{X}_\Gamma \to \mathbb{R}, \quad (X, Y) \mapsto \max\left\{\langle \boldsymbol{x}, \boldsymbol{y} \rangle \,:\, \boldsymbol{x} \in X, \boldsymbol{y} \in Y\right\}.$$

We call the kernel function $k(\cdot|\cdot)$ *optimal alignment kernel*, induced by the inner product $\langle \cdot, \cdot \rangle$. Note that the maximizer of a set of positive definite kernels is an indefinite kernel in general. Since Γ is a group, we find that

$$k(X, Y) = \max\left\{\langle \boldsymbol{x}, \boldsymbol{y} \rangle \,:\, \boldsymbol{x} \in X\right\}.$$

where $\boldsymbol{y}$ is an arbitrary but fixed vector representation of Y. In general, we have

$$k(X, Y) \geq \langle \boldsymbol{x}, \boldsymbol{y} \rangle$$

for all $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$.

Example 2. Consider the Riemannian orbifold $(\mathcal{X}, \Gamma, \pi)$ of Example 1, where $\mathcal{X} = \mathbb{R}^2$ and $\Gamma = \{\mathrm{id}, \gamma\}$ is the group generated by reflections across the x-y-plane. Suppose that $\boldsymbol{x} = (1, 2)$ is a vector representation of X and $\boldsymbol{y} = (3, 2)$ is a vector representation of Y. Then the optimal alignment kernel $k\,(X, Y)$ induced by the standard inner product of $\mathcal{X}$ is given by

$$k(X, Y) = \max\left\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle, \langle \gamma(\boldsymbol{x}), \boldsymbol{y}\rangle, \langle \boldsymbol{x}, \gamma(\boldsymbol{y})\rangle, \langle \gamma(\boldsymbol{x}), \gamma(\boldsymbol{y})\rangle\right\}$$

Evaluating the inner products yields

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{y}\rangle &= \langle (1,2), (3,2)\rangle = 7 \\ \langle \gamma(\boldsymbol{x}), \boldsymbol{y}\rangle &= \langle (2,1), (3,2)\rangle = 8 \\ \langle \boldsymbol{x}, \gamma(\boldsymbol{y})\rangle &= \langle (1,2), (2,3)\rangle = 8 \\ \langle \gamma(\boldsymbol{x}), \gamma(\boldsymbol{y})\rangle &= \langle (2,1), (2,3)\rangle = 7. \end{aligned}$$

Thus, we have $k(X, Y) = 8$.

Example 3. Suppose that X and Y are attributed graphs where edges have attribute 1 and vertices have attribute 0. The optimal alignment kernel $k\,(X, Y)$ induced by the standard inner product of $\mathcal{X}$ is the number of edges of a maximum common subgraph of X and Y.

Example 4. More generally, if property P1 is satisfied, then any optimal alignment kernel on a bounded set of attributed graphs as defined in (1) is also an optimal assignment kernel of some Riemannian orbifold.

Suppose that $X \in \mathcal{X}_\Gamma$. Since $k(X, X) = \langle \boldsymbol{x}, \boldsymbol{x}\rangle$ for all $\boldsymbol{x} \in X$, we can define the *length* of X by

$$l(X) = \sqrt{k(X, X)}.$$

The optimal alignment kernel together with the length satisfies the Cauchy-Schwarz inequality

$$|k(X, Y)| \leq l(X) \cdot l(Y).$$

Since the Cauchy-Schwarz inequality is valid, the geometric interpretation of $k(\cdot|\cdot)$ is that it computes the cosine of a well-defined angle between X and X' provided they are normalized to length 1.

Likewise, $k(\cdot|\cdot)$ gives rise to a distance function defined by

$$d(X, Y) = \sqrt{l(X)^2 - 2k(X, Y) + l(Y)}.$$

From the definition of $k(\cdot|\cdot)$ follows that d is a metric. In addition, we have

$$d(X, Y) = \min\left\{\|\boldsymbol{x} - \boldsymbol{y}\| \,:\, \boldsymbol{x} \in X, \boldsymbol{y} \in Y\right\}, \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm induced by the inner product $\langle \cdot, \cdot\rangle$ of the Euclidean space $\mathcal{X}$.

Example 5. Consider the Riemannian orbifold $(\mathcal{X}, \Gamma, \pi)$ of Example 1 and 2. Suppose that $\boldsymbol{x} = (1, 2)$ is a vector representation of X and $\boldsymbol{y} = (3, 2)$ is a vector representation of Y. Then the squared lengths of X and Y are $l(X)^2 = 5$ and $l(Y)^2 = 13$. Since $k(X, Y) = 8$ according to Example 2, the distance is $d(X, Y) = \sqrt{5 - 16 + 13} = \sqrt{2}$.

Example 6. If properties P1 and P2 are satisfied, then the graph metric (2) coincides with the intrinsic orbifold metric (3).

Equation (3) states that $d\,(\cdot|\cdot)$ is the length of a minimizing geodesic of X and Y and therefore an intrinsic metric, because it coincides with the infimum of the length of all admissible curves from X to Y. In addition, we find that the topology of $\mathcal{X}_\Gamma$ induced by the metric d coincides with the quotient topology induced by the topology of the Euclidean space $\mathcal{X}$.

3.4 Orbifold Functions

Suppose that $\mathcal{Q} = (\mathcal{X}, \Gamma, \pi)$ is an orbifold. An *orbifold function* is a mapping

$$f : \mathcal{X}_\Gamma \to \mathbb{R}.$$

The *lift* of f is a function

$$\tilde{f} : \mathcal{X} \to \mathbb{R}$$

satisfying $\tilde{f} = f \circ \pi$. The lift $\tilde{f}$ is invariant under group actions of Γ, that is $\tilde{f}(\boldsymbol{x}) = \tilde{f}\,(\gamma(\boldsymbol{x}))$ for all $\gamma \in \Gamma$.

We say, an orbifold function $f : \mathcal{X}_\Gamma \to \mathbb{R}$ is continuous (locally Lipschitz, differentiable, generalized differentiable) at $X \in \mathcal{X}_\Gamma$ if its lift $\tilde{f}$ is continuous (locally Lipschitz, differentiable, generalized differentiable) at some vector representation $\boldsymbol{x} \in X$. The definition is independent of the choice of the vector representation that projects to X (see Section B.1, Prop. 1 – Prop. 4). For a definition of generalized differentiable functions and their basic properties we refer to Section A.

Example 7. Consider the Riemannian orbifold $(\mathcal{X}, \Gamma, \pi)$ of Example 1-5. The function

$$f_Y : \mathcal{X}_\Gamma \to \mathbb{R}, \quad X \mapsto k(X, Y)$$

for some $Y \in \mathcal{X}_\Gamma$ is an orbifold function with lift

$$\tilde{f}_Y : \mathcal{X} \to \mathbb{R}, \quad \boldsymbol{x} \mapsto \max\left\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle, \langle \boldsymbol{x}, \gamma(\boldsymbol{y})\rangle\right\},$$

where $\boldsymbol{y} \in Y$. Analytical properties of f such as continuity and differentiability can be investigated using the lift $\tilde{f}$ of f. For example, if $\tilde{f}$ is differentiable at $\boldsymbol{x} \in X$ then it is also differentiable at $\gamma(\boldsymbol{x})$ according to Prop. 3. Hence, differentiability of the orbifold function f is well-defined at X.

3.5 Gradients and Generalized Gradients of Orbifold Functions

We extend the notion of gradient and generalized gradient to differentiable and generalized differentiable orbifold functions.

Gradient of Differentiable Orbifold Functions. Suppose that $f : \mathcal{X}_\Gamma \rightarrow \mathbb{R}$ is differentiable at $X \in \mathcal{X}_\Gamma$. Then its lift $\tilde{f} : \mathcal{X} \rightarrow \mathbb{R}$ is differentiable at all vector representations that project to X. The *gradient* $\nabla f(X)$ of f at X is defined by the projection

$$\nabla f(X) = \pi\left(\nabla \tilde{f}(\boldsymbol{x})\right)$$

of the gradient $\nabla \tilde{f}(\boldsymbol{x})$ of $\tilde{f}$ at a vector representation $\boldsymbol{x} \in X$. This definition is independent of the choice of the vector representation. We have

$$\nabla \tilde{f}(\gamma(\boldsymbol{x})) = \gamma\left(\nabla \tilde{f}(\boldsymbol{x})\right)$$

for all $\gamma \in \Gamma$. This implies that the gradients of $\tilde{f}$ at $\boldsymbol{x}$ and $\gamma(\boldsymbol{x})$ are vector representations of the same structure, namely the gradient $\nabla f(X)$ of the orbifold function f at X. Thus, the gradient of f at X is a well-defined structure pointing to the direction of steepest ascent (see Section B.1, Prop. 3).

Subdifferential of Generalized Differentiable Orbifold Functions. Suppose that $f : \mathcal{X}_\Gamma \rightarrow \mathbb{R}$ is generalized differentiable at $X \in \mathcal{X}_\Gamma$. Then its lift $\tilde{f} : \mathcal{X} \rightarrow \mathbb{R}$ is generalized differentiable at all vector representations that project to X. The *subdifferential* $\partial f(X)$ of f at X is defined by the projection

$$\partial f(X) = \pi\left(\partial \tilde{f}(\boldsymbol{x})\right)$$

of the subdifferential $\partial \tilde{f}(\boldsymbol{x})$ of $\tilde{f}$ at a vector representation $\boldsymbol{x} \in X$. This definition is independent of the choice of the vector representation. We have

$$\partial \tilde{f}(\gamma(\boldsymbol{x})) = \gamma\left(\partial \tilde{f}(\boldsymbol{x})\right)$$

for all $\gamma \in \Gamma$. This implies that the subdifferentials $\partial \tilde{f}(\boldsymbol{x}) \subseteq \mathcal{X}$ and $\partial \tilde{f}(\gamma(\boldsymbol{x})) \subseteq \mathcal{X}$ are subsets that project to the same subset of $\mathcal{X}_\Gamma$, namely the subdifferential $\partial f(X)$ (see Section B.1, Prop. 4).

The properties of generalized differentiable function as listed in Section A carry over to generalized differentiable orbifold functions via their lifts. For example, a generalized differentiable orbifold function is locally Lipschitz and therefore differentiable almost everywhere.

Example 8. Let $(\mathcal{G}_\mathcal{A}, d)$ be a graph space, where

$$d(X, Y) = \min_{\phi \in \mathcal{A}(X,Y)} d_\phi(X, Y)$$

is a graph edit distance. We can identify $\mathcal{G}_\mathcal{A}$ with a Riemannian orbifold $\mathcal{Q} = (\mathcal{X}, \Gamma, \pi)$ and the graph edit distance $d(\cdot|\cdot)$ with a distance function defined on $\mathcal{X}_\Gamma$. Suppose that the cost functions $d_\phi(\cdot|\cdot)$ of the edit paths are continuously differentiable (generalized differentiable). Then the distance $d(\cdot|\cdot)$ is generalized differentiable.

Example 9. Let $\mathcal{Q}$ be a Riemannian orbifold of attributed graphs. Then (i) an optimal assignment kernel $k(\cdot|\cdot)$, (ii) the intrinsic metric $d(\cdot|\cdot)$ induced by $k(\cdot|\cdot)$, and (iii) the squared metric $d(\cdot|\cdot)^2$ are generalized differentiable.

3.6 Integration on Orbifolds

Suppose that $\mathcal{Q} = (\mathcal{X}, \Gamma, \pi)$ is a Riemannian orbifold with singular set $\mathcal{S}_{\mathcal{Q}}$. In order to integrate orbifold functions $f : \mathcal{X}_\Gamma \to \mathbb{R}$ by the Lebesgue integral, we need to construct an appropriate measurable space together with an orbifold measure. The measurable space is defined by the Borel set $\mathcal{B}(\mathcal{X}_\Gamma)$ generated by the open sets of $\mathcal{X}_\Gamma$. From the orbifold measure we expect that it is compatible with the local Riemannian measures. In addition, we demand that the singular set $\mathcal{S}_{\mathcal{Q}}$ has measure 0. This is motivated by the following fact: The singular set is covered locally by the finite union of totally geodesic submanifolds, which has measure 0 relative to the local canonical Riemannian measure. Since the projection to the orbifold is distance decreasing, it is reasonable to ask for an orbifold measure that assigns measure 0 to the singular set $\mathcal{S}_{\mathcal{Q}}$.

Let $\mathcal{B}\left(\mathcal{X}_\Gamma \setminus \mathcal{S}_{\mathcal{Q}}\right)$ denote the Borel set generated by the open sets of $\mathcal{X}_\Gamma \setminus \mathcal{S}_{\mathcal{Q}}$. Then there exists a complete canonical measure μ on the the Borel set $\mathcal{B}\left(\mathcal{X}_\Gamma \setminus \mathcal{S}_{\mathcal{Q}}\right)$ given by a unique volume form on $\mathcal{X}_\Gamma \setminus \mathcal{S}_{\mathcal{Q}}$. The measure μ can be extended to a complete measure ν on the Borel set $\mathcal{B}(\mathcal{X}_\Gamma)$ such that

$$\nu\left(\mathcal{A}\right) = \mu\left(\mathcal{A} \setminus \mathcal{S}_{\mathcal{Q}}\right) = \int_{\mathcal{A} \setminus \mathcal{S}_{\mathcal{Q}}} d\mu.$$

In particular, we have $\nu(\mathcal{A}) = 0$ for any subset $\mathcal{A} \subseteq \mathcal{S}_{\mathcal{Q}}$. For proofs we refer to [4].

In the following we write

$$\int_{\mathcal{U}_\Gamma} f(X) dX = \int_{\mathcal{U}_\Gamma} f d\nu$$

for the integral of an orbifold function $f : \mathcal{U}_\Gamma \to \mathbb{R}$ defined on a measurable subset $\mathcal{U}_\Gamma \subseteq \mathcal{X}_\Gamma$. We tacitly assume that all integrals occurring in the following sections exist.

4 Graph Quantization

This section extends vector quantization to quantization of graphs.

4.1 The Basics

Suppose that $\mathcal{Q} = (\mathcal{X}, \Gamma, \pi)$ is a Riemannian orbifold. A *graph quantizer* of size k is a mapping of the form

$$Q : \mathcal{X}_\Gamma \to \mathcal{C}$$

where $\mathcal{C} = \{Y_1, \dots, Y_k\} \subseteq \mathcal{X}_\Gamma$ is a finite set, called *codebook*. The elements $Y_j \in \mathcal{C}$ are the *code graphs*. The graph quantizer Q partitions the input space $\mathcal{X}_\Gamma$ into k disjoint *regions*

$$\mathcal{R}_j = \{X \in \mathcal{X}_\Gamma \,:\, Q(X) = Y_j\}$$

such that their union covers $\mathcal{X}_\Gamma$. By $\mathcal{P}_Q$ we denote the partition of Q consisting of all k regions $\mathcal{R}_j$.

Suppose that $\mathcal{J} = \{1, \dots, k\}$. The basic operation of a vector quantizer Q can be written as a composition $Q = d_Q \circ e_Q$ of an *encoder* $e_Q : \mathcal{X}_\Gamma \rightarrow \mathcal{J}$ and a *decoder* $d_Q : \mathcal{J} \rightarrow \mathcal{C}$. The encoder assigns each input graph to a region via the index set $\mathcal{J}$. The decoder maps indices of $\mathcal{J}$ referring to regions to code graphs.

4.2 Graph Quantizer Performance

We measure the performance of a graph quantizer Q by the expected distortion

$$D(Q) = \mathbb{E}_X \left[d\left(X, Q(X)\right) \right] = \int_{\mathcal{X}_\Gamma} d(X, Q(X)) dP(X),$$

where $X \in \mathcal{X}_\Gamma$ is a random variable with probability measure $P = P_{\mathcal{X}_\Gamma}$ representing the observable graphs to be quantized. The expectation $\mathbb{E}_X$ is taken with respect to some probability space $(\mathcal{X}_\Gamma, \Sigma_{\mathcal{X}_\Gamma}, P_{\mathcal{X}_\Gamma})$. The quantity $d(X, Y)$ measures the *distortion* of the random input graph X and code graph Y. Here we consider graph distortion measures that are graph edit distances. An example is the squared metric induced by an optimal alignment kernel

$$d\left(X, Y\right) = \min_{\boldsymbol{x} \in X, \boldsymbol{y} \in Y} \left\| \boldsymbol{x} - \boldsymbol{y} \right\|^2$$

Using the codebook and partition for the given quantizer Q, we can rewrite the expected distortion by

$$D(\mathcal{C}) = \sum_{j=1}^{k} \int_{\mathcal{R}_j} d(X, Y) dP(X).$$

4.3 The Problem of Optimal Graph Quantizer Design

The problem of optimal graph quantizer design is stated as follows: Find a codebook $\mathcal{C}$ specifying the decoder d_Q and a partition $\mathcal{P}_Q$ specifying the encoder e_Q such that the expected distortion $D(Q)$ is minimized. The composite mapping $Q = d_Q \circ e_Q$ of the resulting encoder and decoder is then an *optimal graph quantizer.*

An optimal graph quantizer satisfies the following necessary conditions, also known as the *Lloyd-Max conditions*:

1. *Nearest Neighbor Condition.* Given a fixed codebook $\mathcal{C}$, a graph quantizer Q is optimal, if the code vector $Q(X)$ of an input pattern X satisfies the nearest neighbor rule

$$Q(X) = \arg \min_{Y \in \mathcal{C}} d\left(X, Y\right)$$

   for all $X \in \mathcal{X}_\Gamma$, where ties are resolved according to some rule. A proof is given in Section B.2, Theorem 3.

2. *Centroid Condition.* Given a fixed partition $\mathcal{P}_Q$, a vector quantizer Q is optimal, if each code vector Y_j is the centroid of region $\mathcal{R}_j$, that is

$$Y_j = \arg \min_{Y \in \mathcal{X}_\Gamma} \mathbb{E}\left[d\left(X, Y\right) \mid X \in \mathcal{R}_j\right]$$

for all $Y \in \mathcal{X}_\Gamma$ and all $j \in \mathcal{J}$. A proof is given in Section B.2, Theorem 4.

Note that Y_j with

$$Y_j = \arg \min_{Y \in \mathcal{X}_\Gamma} \mathbb{E}\left[d\left(X, Y\right) \mid X \in \mathcal{R}_j\right]$$

is called a *centroid* of region $\mathcal{R}_j$. The centroids may not be unique. This also holds for squared metrics induced by some optimal assignment kernel, which are the counterparts of squared Euclidean distances.

4.4 Graph Quantizer Design

Since the distribution $P = P_{\mathcal{X}_\Gamma}$ of the observable graphs is usually unknown, the expected distortion $D(\mathcal{C})$ can neither be computed nor be minimized directly. Instead, we design (estimate) an optimal quantizer from empirical data. For vectors, prominent methods for designing an optimal quantizer are k-means and simple competitive learning. Both methods, k-means and simple competitive learning have been extended for designing graph quantizers in the context of prototype based clustering. To derive consistency results for k-means and simple competitive learning in the domain of graphs, we consider estimators based on empirical distortions and on stochastic approximation.

Estimators based on Empirical Distortion Measures. In order to derive consistency results, we restrict the set of feasible codebooks to a compact subspace

$$\mathcal{W} \subset \mathcal{X}_\Gamma^k = \underbrace{\mathcal{X}_\Gamma \times \cdots \times \mathcal{X}_\Gamma}_{\text{k-times}}$$

of the topological space $\mathcal{X}_\Gamma^k$. The problem of designing an optimal quantizer for graphs is then of the form

$$\min_{\mathcal{C} \in \mathcal{W}} \quad D(\mathcal{C}) = \sum_{j=1}^{k} \int_{\mathcal{R}_j} d(X, Y) dP(X).$$

where the minimum is taken over the compact set $\mathcal{W}$ rather than $\mathcal{X}_\Gamma^k$. Let

1. D^* be the set of minimal values of the expected distortion $D(\mathcal{C})$,
2. $\mathcal{W}^* = \{\mathcal{C} \in \mathcal{W} \,:\, D(\mathcal{C}) = D^*\}$ be the set of true (optimal) codebooks, and
3. $\mathcal{W}_\varepsilon^* = \{\mathcal{C} \in \mathcal{W} \,:\, D(\mathcal{C}) \leq D^* + \varepsilon\}$ be the set of approximate solutions.

To design an optimal graph quantizer, we minimize the *empirical distortion*

$$\hat{D}_N(\mathcal{C}) = \frac{1}{N}\sum_{i=1}^{N} \min_{j \in \mathcal{J}} d\left(X_i, Y_j\right),$$

where $\mathcal{C} \in \mathcal{W}$ and $\mathcal{S} = \{X_1, \ldots, X_N\}$ is a training set consisting of N independent graphs X_i drawn from $\mathcal{X}_\Gamma$. Let

1. $\hat{D}_N^*$ be the set of minimal values of the empirical distortion $\hat{D}_N(\mathcal{C})$,
2. $\mathcal{W}_N^* = \{\mathcal{C} \in \mathcal{W} \,:\, \hat{D}_N(\mathcal{C}) = \hat{D}_N^*\}$ be the set of empirical codebooks, and
3. $\mathcal{W}_{N\varepsilon}^* = \{\mathcal{C} \in \mathcal{W} \,:\, \hat{D}_N(\mathcal{C}) \leq \hat{D}_N^* + \varepsilon\}$ be the set of approximate solutions.

The next result shows that estimators based on empirical distortions are consistent estimators.

Theorem 1. *Suppose that $\mathcal{Q} = (\mathcal{X}, \Gamma, \pi)$ is a Riemannian orbifold, $d(X, Y)$ is a locally Lipschitz metric on $\mathcal{X}_\Gamma$ with integrable Lipschitz constant, and $\mathcal{W} \subseteq \mathcal{X}_\Gamma^k$ is compact. Then we have*

$$\lim_{N \to \infty} \hat{D}_N^*(\omega) = D^*$$
$$\lim_{N \to \infty} \mathcal{W}_N^*(\omega) = \mathcal{W}^*$$
$$\lim_{N \to \infty} \mathcal{W}_{\epsilon N}^*(\omega) = \mathcal{W}_\epsilon^*$$

almost surely.

The proof follows from [8] applied to the lift $\tilde{d}$ of distortion d. Examples of locally Lipschitz distance metrics on $\mathcal{X}_\Gamma$ with integrable Lipschitz constants are metrics induced by an optimal alignment kernel

$$d(X, Y) = \min_{\boldsymbol{x} \in X, \boldsymbol{y} \in Y} \|\boldsymbol{x} - \boldsymbol{y}\|$$

as well as $d(X, Y)^2$.

K-Means. In order to extend the standard k-means method to graphs for constructing an empirical codebook, we use the following update rule

$$\boldsymbol{y}_j^{t+1} = \frac{1}{N_j^t}\sum_{i=1}^{N} q_{ij}^t \boldsymbol{x}_i,$$

where $t > 0$ is the iteration, $\boldsymbol{x}_i \in X_i$ and $\boldsymbol{y}_j^t \in Y_j^t$ are vector representations that are optimally aligned,[1] and $\boldsymbol{Q}^t = \left(q_{ij}^t\right)$ is the matrix representation of the nearest neighbor quantizer Q^t restricted to the training set $\mathcal{S}$. The elements of $\boldsymbol{Q}^t$ are of the form

$$q_{ij}^t = \begin{cases} 1 & : \quad Q^t(X_i) = Y_j^t \\ 0 & : \quad \text{otherwise} \end{cases}.$$

[1] Recall that two vector representations $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$ are optimally aligned if $\|\boldsymbol{x} - \boldsymbol{y}\| = d(X, Y)$

The quantity N_j^t denotes the number of elements from the training sets that are quantized by code graph Y_j^t.

As for vectors, a drawback of k-means for graphs is that it is a local optimization technique for which existing consistency theorems are inapplicable, because Theorem 1 assumes global instead of local minimizers of the empirical distortion as estimators.

Estimators based on Stochastic Optimization. Suppose that $\mathcal{W} = \mathcal{X}_\Gamma^k$. Stochastic optimization methods directly minimize the expected distortion

$$\begin{aligned} D\left(\mathcal{C}\right) &= \sum_{j=1}^{k} \int_{\mathcal{R}_j} d\left(X, Y_j\right) d\, P(X) \\ &= \sum_{j=1}^{k} \int_{\mathcal{X}_\Gamma} \min_{1 \leq j \leq k} d\left(X, Y_j\right) d\, P(X), \end{aligned}$$

using a training set $\mathcal{S} = \{X_1, \ldots, X_N\}$ of N independent graphs X_i drawn from $\mathcal{X}_\Gamma$. We assume that the loss function

$$L(X, \mathcal{C}) = \min_{1 \leq j \leq k} d\left(X, Y_j\right)$$

is generalized-differentiable, hence $L(X, \mathcal{C})$ is differentiable almost everywhere.

Example 10. If he graph distortion $d(\cdot|\cdot)$ is generalized differentiable, then the loss function $L(X, \mathcal{C})$ is also generalized differentiable by calculus of generalized differentiable functions. This holds for graph distortions of Example 8 and 9.

Since the interchange of integral and generalized gradient remains valid for generalized differentiable loss functions, that is

$$\partial D(\mathcal{C}) = \mathbb{E}_X\left[\partial L(X, \mathcal{C})\right]$$

under mild assumptions (see [11, 27]), we can minimize the expected distortion $D(\mathcal{C})$ according to the following *stochastic generalized gradient* (SGG) method:

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t\left(\boldsymbol{x}_t - \boldsymbol{y}_t\right), \tag{4}$$

where $\boldsymbol{x}_t$ is a vector representation of input pattern $X_t \in \mathcal{S}$, which is optimally aligned to vector representation $\boldsymbol{y}_t$ of a code graph Y_t closest to X_t. The random elements $\boldsymbol{s}_t = \boldsymbol{x}_t - \boldsymbol{y}_t \in S_t$ are vector representations of *stochastic generalized gradients* S_t, i.e. random variables defined on the probability space $\left(\mathcal{X}_\Gamma, \Sigma_{\mathcal{X}_\Gamma}, P_{\mathcal{X}_\Gamma}\right)^\infty$ such that

$$\mathbb{E}\left[S_t \,|\, \mathcal{C}_0, \ldots, \mathcal{C}_t\right] \in \partial D\left(\mathcal{C}\right). \tag{5}$$

We consider the following conditions for almost sure convergence of stochastic optimization:

A1 The sequence $(\eta_t)_{t \geq 0}$ of step sizes satisfies

$$\eta_t > 0, \quad \lim_{t \to \infty} \eta_t = 0, \quad \sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty.$$

A2 The stochastic generalized gradients $(S_t)_{t \geq 0}$ satisfy (5).

A3 The expected squared norm of stochastic generalized gradients $(S_t)_{t \geq 0}$ is bounded by

$$\mathbb{E}\left[\|S_t\|^2\right] < +\infty.$$

The next result shows that the SGG method is a consistent estimator.

Theorem 2. *Let $\mathcal{Q} = (\mathcal{X}, \mathit{\Gamma}, \pi)$ be a Riemannian orbifold and let $d(X, Y)$ be a generalized differentiable metric on $\mathcal{X}_{\mathit{\Gamma}}$. Suppose that assumptions $(A1) - (A3)$ hold. Then the sequence $(\mathcal{C}_t)_{t \geq 0}$ generated by the SGG method converges almost surely to graphs satisfying necessary extremum conditions*

$$\mathcal{W}^* = \{\mathcal{C} \in \mathcal{W} \, : \, 0 \in \partial D(\mathcal{C})\}\,.$$

Besides the sequence $(D(\mathcal{C}_t))_{t \geq 0}$ converges almost surely and we have

$$\lim_{t \to \infty} D(\mathcal{C}_t) \in D(\mathcal{W}^*).$$

The proof is a direct consequence of Ermoliev and Norkin's Theorem [11] applied on the lift $\tilde{d}\,(\cdot|\cdot)$ of $d\,(\cdot|\cdot)$.

5 Remarks to GQ using the Graph Edit Distance

In many applications, the graph edit distance is discontinuous. Examples include edit distances with constant non-zero deletion and/or insertion cost. A necessary (but not sufficient) condition for the consistency results stated in Theorem 1 and 2 is that the underlying graph distortion is locally Lipschitz. Hence, both consistency results are inapplicable for discontinuous graph distortions. Let us consider both cases separately.

Estimators based on Empirical Distortion Measures. Estimators based on empirical distortion measures aim at approximating the expected distortion $D(\mathcal{C})$ by its empirical mean

$$\min_{\mathcal{C} \in \mathcal{W}} \quad \hat{D}_N(\mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} \min_{j \in \mathcal{J}} d\,(X_i, Y_j)\,.$$

As shown in [10], minimizing the empirical distortion is often meaningless, if the underlying graph edit distance function $d\,(\cdot|\cdot)$ and thus $\hat{D}_N(\mathcal{C})$ is discontinuous, even if the expectation $D(\mathcal{C})$ may be continuously differentiable. Since the local solutions of $\hat{D}_N(\mathcal{C})$ may have nothing in common with the local solutions of the original problem, estimators based on the empirical distortion $\hat{D}_N(\mathcal{C})$ can be statistically inconsistent. Hence, minimizing $\hat{D}_N(\mathcal{C})$ with underlying discontinuous graph edit distance using global or local optimization techniques like, for example, k-means lacks theoretical support.

Estimators based on Stochastic Optimization. The situation is better for estimators based on methods from stochastic optimization. For discontinuous graph edit distances $d\left(\cdot|\cdot\right)$ the expected distortion can be minimized in a statistically consistent way, for example, by methods based on approximations of $d\left(\cdot|\cdot\right)$ via averaged functions obtained by convolution with so-called mollifiers. For details, we refer to [9].

6 Conclusion

This contribution proposes a theoretical sound foundation of graph quantization generalizing the ideas of vector quantizations to the domain of attributed graph. We presented consistency results for graph quantizer design, where the underlying graph edit distances is generalized differentiable. As for vectors, estimators based on empirical distortion and stochastic optimization are statistically consistent. If the underlying distortion measure is a discontinuous graph edit distance, estimators based on empirical distortion measures lack theoretical justification. Thus, the proposed consistency results justify existing research on prototype-based clustering in the domain of graphs. In addition, we showed that the Lloyd-Max conditions are necessary conditions for optimality of GQ.

The mathematical framework that enables us to derive consistency results are Riemannian orbifolds. Identifying graphs with points in a Riemannian orbifold provides us locally access to a Euclidean space. This in turn allows us to introduce geometrical and analytical concepts for extending vector quantization to the domain of graphs. The implication of this approach is that it provides us a template for consistently linking methods from structural pattern recognition other than GQ to statistical pattern recognition methods.

Acknowledgments. The first author is very grateful to Vladimir Norkin for his kind support and valuable comments.

A Generalized Differentiable Functions

Let $\mathcal{X} = \mathbb{R}^n$ be a finite-dimensional Euclidean space. A function $f : \mathcal{X} \to \mathbb{R}$ is *generalized differentiable* at $\boldsymbol{x} \in \mathcal{X}$ in the sense of Norkin [27] if there is a multi-valued map $\partial f : \mathcal{X} \to 2^{\mathcal{X}}$ in a neighborhood of $\boldsymbol{x}$ such that

1. $\partial f(\boldsymbol{x})$ is a convex and compact set;
2. $\partial f(\boldsymbol{x})$ is upper semicontinuous at $\boldsymbol{x}$, that is, if $\boldsymbol{y}_i \to \boldsymbol{x}$ and $\boldsymbol{g}_i \in \partial f(\boldsymbol{y}_i)$ for each $i \in \mathbb{N}$, then each accumulation point $\boldsymbol{g}$ of $(\boldsymbol{g}_i)$ is in $\partial f(\boldsymbol{x})$;
3. for each $\boldsymbol{y} \in \mathcal{X}$ there is a $\boldsymbol{g} \in \partial f(\boldsymbol{y})$ with $f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle + o\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{g}\right)$, where
$$\lim_{i \to \infty} \frac{\left|o\left(\boldsymbol{x}, \boldsymbol{y}_i, \boldsymbol{g}_i\right)\right|}{\|\boldsymbol{y}_i - \boldsymbol{x}\|} = 0$$
for all sequences $\boldsymbol{y}_i \to \boldsymbol{y}$ and $\boldsymbol{g}_i \to \boldsymbol{g}$ with $\boldsymbol{g}_i \in \partial f\left(\boldsymbol{y}_i\right)$.

We call f *generalized differentiable* if it is generalized differentiable at each point $\boldsymbol{x} \in \mathcal{X}$. The set $\partial f(\boldsymbol{x})$ is the *subdifferential* of f at $\boldsymbol{x}$ and its elements are called *generalized gradients.*

Generalized differentiable functions have the following properties [27]:

(GD1) Generalized differentiable functions are locally Lipschitz and therefore continuous and differentiable almost everywhere.

(GD2) Continuously differentiable, convex, and concave functions are generalized differentiable.

(GD3) Suppose that $f_1, \ldots, f_n : \mathcal{X} \to \mathbb{R}$ are generalized differentiable at $\boldsymbol{x} \in \mathcal{X}$. Then

$$\begin{aligned} f_*(\boldsymbol{x}) &= \min(f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})) \\ f^*(\boldsymbol{x}) &= \max(f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})) \end{aligned}$$

are generalized differentiable at $\boldsymbol{x} \in \mathcal{X}$.

(GD4) Suppose that $f_1, \ldots, f_m : \mathcal{X} \to \mathbb{R}$ are generalized differentiable at $\boldsymbol{x} \in \mathcal{X}$ and $f_0 : \mathbb{R}^m \to \mathbb{R}$ is generalized differentiable at $\boldsymbol{y} = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})) \in \mathbb{R}^m$. Then $f(\boldsymbol{x}) = f_0(f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$ is generalized differentiable at $\boldsymbol{x} \in \mathcal{X}$. The subdifferential of f at $\boldsymbol{x}$ is of the form

$$\begin{aligned} \partial f(\boldsymbol{x}) = \text{con} \Big\{ \boldsymbol{g} \in \mathcal{X} : \; & \boldsymbol{g} = [\boldsymbol{g}_1 \boldsymbol{g}_2 \ldots \boldsymbol{g}_m] \boldsymbol{g}_0, \\ & \boldsymbol{g}_0 \in \partial f_0(\boldsymbol{y}), \\ & \boldsymbol{g}_i \in \partial f_i(\boldsymbol{x}), 1 \leq i \leq m \Big\}. \end{aligned}$$

where $[\boldsymbol{g}_1 \boldsymbol{g}_2 \ldots \boldsymbol{g}_m]$ is a $(N \times m)$-matrix.

(GD5) Suppose that $F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}} [f(\boldsymbol{x}, \boldsymbol{z})]$, where $f(\cdot, \boldsymbol{z})$ is generalized differentiable. Then F is generalized differentiable and its subdifferential at $\boldsymbol{x} \in \mathcal{X}$ is of the form $\partial F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}} [\partial f(\boldsymbol{x}, \boldsymbol{z})]$.

B Proofs

Suppose that $\mathcal{Q} = (\mathcal{X}, \Gamma, \pi)$ is a Riemannian orbifold. By $\mathcal{U}_\delta(\boldsymbol{x}) = \{\boldsymbol{x}' : \|\boldsymbol{x}'\| < \delta\}$ we denote the open ball with center $\boldsymbol{x}$ and radius $\delta > 0$. Note that $\mathcal{U}_\delta(\gamma(\boldsymbol{x})) = \gamma(\mathcal{U}_\delta(\boldsymbol{x}))$ for all $\gamma \in \Gamma$.

B.1 Orbifold Functions

Continuous Orbifold Functions

Proposition 1. *Let $f : \mathcal{X}_\Gamma \to \mathbb{R}$ be an orbifold function. Suppose that its lift $\tilde{f} : \mathcal{X} \to \mathbb{R}$ is continuous at a vector representation $\boldsymbol{x}$ that projects to $X \in \mathcal{X}_\Gamma$. Then $\tilde{f}$ is continuous at $\gamma(\boldsymbol{x})$ for all $\gamma \in \Gamma$.*

Proof. Let $\gamma \in \Gamma$ be a permutation and $\boldsymbol{x}' = \gamma(\boldsymbol{x})$. Suppose that $(\boldsymbol{y}'_i)_{i \in \mathbb{N}}$ is a sequence with $\boldsymbol{y}'_i \to \boldsymbol{x}'$. Then there is a sequence $(\boldsymbol{y}_i)_{i \in \mathbb{N}}$ with $\gamma(\boldsymbol{y}_i) = \boldsymbol{y}'_i$ for each $i \in \mathbb{N}$. Since permutations are homeomorphisms, we find that

$$\lim_{i \to \infty} \boldsymbol{y}_i = \lim_{i \to \infty} \gamma^{-1}(\boldsymbol{y}'_i) = \gamma^{-1}(\boldsymbol{x}') = \boldsymbol{x}.$$

From continuity of $\tilde{f}$ at $\boldsymbol{x}$ follows that $\tilde{f}(\boldsymbol{y}_i) \to \tilde{f}(\boldsymbol{x})$. Since $\tilde{f}$ is invariant under group actions from Γ, we have $\tilde{f}(\boldsymbol{x}) = \tilde{f}(\boldsymbol{x}')$ and $\tilde{f}(\boldsymbol{y}_i) = \tilde{f}(\boldsymbol{y}'_i)$ for each $i \in \mathbb{N}$. We obtain

$$\lim_{i \to \infty} \tilde{f}\left(\boldsymbol{y}'_i\right) = \lim_{i \to \infty} \tilde{f}\left(\boldsymbol{y}_i\right) = \tilde{f}(\boldsymbol{x}) = \tilde{f}(\boldsymbol{x}').$$

This proves that $\tilde{f}$ is continuous at each vector representation that projects to X. □

Locally Lipschitz Orbifold Functions

Proposition 2. *Let $f : \mathcal{X}_\Gamma \to \mathbb{R}$ be an orbifold function. Suppose that its lift $\tilde{f} : \mathcal{X} \to \mathbb{R}$ is locally Lipschitz at a vector representation $\boldsymbol{x}$ that projects to $X \in \mathcal{X}_\Gamma$. Then $\tilde{f}$ is locally Lipschitz at $\gamma(\boldsymbol{x})$ for all $\gamma \in \Gamma$.*

Proof. Since $\tilde{f}$ is locally Lipschitz at $\boldsymbol{x}$ there is a $L \geq 0$ and $\delta > 0$ such that

$$\left| \tilde{f}(\boldsymbol{y}) - \tilde{f}(\boldsymbol{z}) \right| \leq L \left\| \boldsymbol{y} - \boldsymbol{z} \right\|$$

for all $\boldsymbol{y}, \boldsymbol{z} \in \mathcal{U}_\delta(\boldsymbol{x})$. Let $\gamma \in \Gamma$ be a permutation and $\boldsymbol{x}' = \gamma(\boldsymbol{x})$. Since γ is an isometric homeomorphism, we have $\mathcal{U}_\delta(\boldsymbol{x}') = \gamma\left(\mathcal{U}_\delta(\boldsymbol{x})\right)$. From Γ-invariance of $\tilde{f}$ and the isometric property of γ follows

$$\left| \tilde{f}(\boldsymbol{y}') - \tilde{f}(\boldsymbol{z}') \right| = \left| \tilde{f}(\boldsymbol{y}) - \tilde{f}(\boldsymbol{z}) \right| \leq L \left\| \boldsymbol{y} - \boldsymbol{z} \right\| = L \left\| \boldsymbol{y}' - \boldsymbol{z}' \right\|$$

for all $\boldsymbol{y}', \boldsymbol{z}' \in \mathcal{U}_\delta(\boldsymbol{x}')$, where $\boldsymbol{y} = \gamma^{-1}(\boldsymbol{y}') \in \mathcal{U}_\delta(\boldsymbol{x})$ and $\boldsymbol{z} = \gamma^{-1}(\boldsymbol{z}) \in \mathcal{U}_\delta(\boldsymbol{x})$. This proves that $\tilde{f}$ is locally Lipschitz at each vector representation that projects to X. □

Differentiable Orbifold Functions

Proposition 3. *Let $f : \mathcal{X}_\Gamma \to \mathbb{R}$ be an orbifold function. Suppose that its lift $\tilde{f} : \mathcal{X} \to \mathbb{R}$ is differentiable at a vector representation $\boldsymbol{x}$ that projects to $X \in \mathcal{X}_\Gamma$. Then $\tilde{f}$ is differentiable at $\gamma(\boldsymbol{x})$ for all $\gamma \in \Gamma$. The gradient of $\tilde{f}$ at $\gamma(\boldsymbol{x})$ is of the form*

$$\nabla \tilde{f}(\gamma(\boldsymbol{x})) = \gamma\left(\nabla \tilde{f}(\boldsymbol{x})\right).$$

Proof. Since the lift $\tilde{f}$ of f is differentiable at $\boldsymbol{x}$, there is a $\delta > 0$ such that

$$\tilde{f}(\boldsymbol{x} + \boldsymbol{h}) = \tilde{f}(\boldsymbol{x}) + \left\langle \nabla \tilde{f}\left(\boldsymbol{x}\right), \boldsymbol{h} \right\rangle + o(\boldsymbol{h})$$

for all $\boldsymbol{h} \in \mathcal{U}_\delta(\mathbf{0})$. Let $\boldsymbol{x}'$ be an arbitrary vector representation that projects to X. Then there is a $\gamma \in \Gamma$ with $\boldsymbol{x}' = \gamma(\boldsymbol{x})$. Since $\tilde{f}$ is invariant under the group actions of Γ, we have $\tilde{f}(\boldsymbol{x}') = \tilde{f}(\boldsymbol{x})$. Then for each $\boldsymbol{h}' \in \mathcal{U}_\delta(\mathbf{0})$, we find that

$$\tilde{f}(\boldsymbol{x}' + \boldsymbol{h}') - \tilde{f}(\boldsymbol{x}') = \tilde{f}(\boldsymbol{x} + \boldsymbol{h}) - \tilde{f}(\boldsymbol{x}) = \left\langle \nabla \tilde{f}\,(\boldsymbol{x}), \boldsymbol{h} \right\rangle + o(\boldsymbol{h}),$$

where $\boldsymbol{h} \in \mathcal{X}$ with $\gamma(\boldsymbol{h}) = \boldsymbol{h}'$. Since the elements of Γ are isometries, we have $\|\boldsymbol{h}\| = \|\boldsymbol{h}'\|$ giving $\boldsymbol{h} \in \mathcal{U}_\delta(\mathbf{0})$. In addition, from isometry of γ follows

$$\langle f_{\boldsymbol{x}}, \boldsymbol{h} \rangle = \left\langle \gamma \left(\nabla \tilde{f}\,(\boldsymbol{x}) \right), \gamma(\boldsymbol{h}) \right\rangle = \left\langle \gamma \left(\nabla \tilde{f}\,(\boldsymbol{x}) \right), \boldsymbol{h}' \right\rangle.$$

We obtain

$$\tilde{f}(\boldsymbol{x}' + \boldsymbol{h}') - \tilde{f}(\boldsymbol{x}') = \left\langle \gamma \left(\nabla \tilde{f}\,(\boldsymbol{x}) \right), \boldsymbol{h}' \right\rangle + o'(\boldsymbol{h}'),$$

where $o'(\boldsymbol{h}') = o \circ \gamma^{-1}(\boldsymbol{h}')$ satisfies

$$\lim_{\boldsymbol{h}' \to 0} \frac{o'(\boldsymbol{h}')}{\|\boldsymbol{h}'\|} = \lim_{\boldsymbol{h}' \to 0} \frac{o(\gamma^{-1}(\boldsymbol{h}'))}{\|\boldsymbol{h}'\|} = \lim_{\boldsymbol{h}' \to 0} \frac{o(\gamma^{-1}(\boldsymbol{h}'))}{\|\gamma^{-1}(\boldsymbol{h}')\|} = 0.$$

This proves that $\tilde{f}$ is differentiable at each vector representation that projects to X. In addition, from the proof follows that the gradient of $\tilde{f}$ at $\boldsymbol{x}' = \gamma(\boldsymbol{x})$ is of the form

$$\nabla \tilde{f}\,(\boldsymbol{x}') = \gamma \left(\nabla \tilde{f}\,(\boldsymbol{x}) \right).$$

□

Generalized Differentiable Orbifold Functions

Proposition 4. *Let $f : \mathcal{X}_\Gamma \to \mathbb{R}$ be an orbifold function. Suppose that its lift $\tilde{f} : \mathcal{X} \to \mathbb{R}$ is generalized differentiable at a vector representation $\boldsymbol{x}$ that projects to $X \in \mathcal{X}_\Gamma$. Then $\tilde{f}$ is generalized differentiable at $\gamma(\boldsymbol{x})$ for all $\gamma \in \Gamma$ and*

$$\partial \tilde{f}(\gamma(\boldsymbol{x})) = \gamma \left(\partial \tilde{f}(\boldsymbol{x}) \right).$$

is a subdifferential of $\tilde{f}$ at $\gamma(\boldsymbol{x})$ for all $\gamma \in \Gamma$.

Proof. Since $\tilde{f}$ is generalized differentiable at $\boldsymbol{x}$, there is a multi-valued mapping $\partial \tilde{f} : \mathcal{U}_\delta(\boldsymbol{x}) \to 2^{\mathcal{X}}$ defined on some neighborhood $\mathcal{U}_\delta(\boldsymbol{x})$. Let $\gamma \in \Gamma$ be an arbitrary permutation and $\boldsymbol{x}' = \gamma(\boldsymbol{x})$. Then

$$\partial \tilde{f} : \mathcal{U}_\delta(\boldsymbol{x}') \to 2^{\mathcal{X}}, \quad \boldsymbol{y}' = \gamma(\boldsymbol{y}) \mapsto \gamma \left(\partial \tilde{f}(\boldsymbol{y}) \right)$$

is a multi-valued mapping in a neighborhood of $\boldsymbol{x}'$.

Since γ is a homeomorphic linear map, we find that $\gamma(\partial \tilde{f}(\boldsymbol{x})) = \partial \tilde{f}(\boldsymbol{x}')$ is a convex and compact set.

Next we show that $\tilde{f}$ is upper semicontinuous at $\boldsymbol{x}'$. Suppose that $\boldsymbol{y}_i' \to \boldsymbol{x}'$, $\boldsymbol{g}_i' \in \tilde{f}_c(\boldsymbol{y}_i')$ for each $i \in \mathbb{N}$, and $\boldsymbol{g}'$ is an accumulation point of $(\boldsymbol{g}_i')_{i \in \mathbb{N}}$. Then there is a $i_0 \in \mathbb{N}$ such that $\boldsymbol{y}_i' \in \mathcal{U}_\delta(\boldsymbol{x}')$ for all $i \geq i_0$. From

$$\mathcal{U}_\delta(\boldsymbol{x}') = \mathcal{U}_\delta(\gamma(\boldsymbol{x})) = \gamma\left(\mathcal{U}_\delta(\boldsymbol{x})\right)$$

follows that there are vector representations $\boldsymbol{y}_i \in \mathcal{U}_\delta(\boldsymbol{x})$ with $\gamma(\boldsymbol{y}_i) = \boldsymbol{y}_i'$ for each $i \geq i_0$. From continuity of γ^{-1} follows that $\boldsymbol{y}_i \to \boldsymbol{x}$. By construction of $\partial\tilde{f}$ follows that

$$\boldsymbol{g}_i' \in \partial\tilde{f}\left(\boldsymbol{y}_i'\right) = \partial\tilde{f}\left(\gamma\left(\boldsymbol{y}_i\right)\right) = \gamma\left(\partial\tilde{f}\left(\boldsymbol{y}_i\right)\right)$$

for each $i \geq i_0$. Hence, there are vector representations $\boldsymbol{g}_i \in \partial\tilde{f}(\boldsymbol{y}_i)$ with $\gamma(\boldsymbol{g}_i) = \boldsymbol{g}_i'$ for each $i \geq i_0$. Since $\tilde{f}$ is upper semicontinuous at $\boldsymbol{x}$, we find that $\boldsymbol{g} \in \partial\tilde{f}(\boldsymbol{x})$. Again by construction of $\partial\tilde{f}$ follows that

$$\boldsymbol{g}' = \gamma(\boldsymbol{g}) \in \gamma\left(\partial\tilde{f}(\boldsymbol{x})\right) = \partial\tilde{f}\left(\gamma(\boldsymbol{x})\right) = \partial\tilde{f}(\boldsymbol{x}').$$

This proves upper semicontinuity of $\partial\tilde{f}$ at all vector representations projecting to $X = \pi(\boldsymbol{x})$.

Finally, we prove that $\tilde{f}$ satisfies the subderivative property at $\boldsymbol{x}'$. Suppose that $\boldsymbol{y}', \boldsymbol{y} \in \mathcal{X}$ with $\boldsymbol{y}' = \gamma(\boldsymbol{y})$. By $\mathit{\Gamma}$-invariance of $\tilde{f}$, we have $\tilde{f}(\boldsymbol{y}') = \tilde{f}(\boldsymbol{y})$. Since $\tilde{f}$ is generalized differentiable at $\boldsymbol{x}$, we find a $\boldsymbol{g} \in \partial\tilde{f}(\boldsymbol{y})$ such that

$$\tilde{f}(\boldsymbol{y}') = \tilde{f}(\boldsymbol{y}) = \tilde{f}(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle + o(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{g})$$

with $o(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{g})$ tending faster to zero than $\|\boldsymbol{y} - \boldsymbol{x}\|$. Let $\boldsymbol{g}' = \gamma(\boldsymbol{g})$. Exploiting $\mathit{\Gamma}$-invariance of $\tilde{f}$ as well as isometry and linearity of γ yields

$$\begin{aligned} \tilde{f}(\boldsymbol{y}') &= \tilde{f}(\gamma(\boldsymbol{x})) + \langle \gamma(\boldsymbol{g}), \gamma(\boldsymbol{y} - \boldsymbol{x}) \rangle + o(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{g}) \\ &= \tilde{f}(\boldsymbol{x}') + \langle \boldsymbol{g}', \boldsymbol{y}' - \boldsymbol{x}' \rangle + o(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{g}). \end{aligned}$$

We define $o'(\boldsymbol{x}', \boldsymbol{y}', \boldsymbol{g}') = o \circ \gamma^{-1}(\boldsymbol{x}', \boldsymbol{y}', \boldsymbol{g}') = o(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{g})$ showing that o' tends faster to zero than $norm\boldsymbol{y}' - \boldsymbol{x}$. This proves the subderivative property of $\tilde{f}$ at all vector representations projecting to $X = \pi(\boldsymbol{x})$.

Putting all results together yields that $\tilde{f}$ is generalized differentiable at $\gamma(\boldsymbol{x})$ for all $\gamma \in \mathit{\Gamma}$. □

B.2 Lloyd-Max Necessary Conditions for Optimality

Due to the comparable nice analytical properties of Riemannian orbifolds, the proofs for the nearest neighbor and centroid condition of optimal graph quantizers are similar to their respective counterparts in vector quantization.

Theorem 3 (Nearest Neighbor Condition). *Suppose that $\mathcal{C}$ is a fixed codebook. Any graph quantizer $Q : \mathcal{X}_\mathit{\Gamma} \to \mathcal{C}$ with*

$$Q(X) = \arg\min_{Y \in \mathcal{C}} d\left(X, Y\right)$$

for all $X \in \mathcal{X}_\mathit{\Gamma}$, where ties are resolved according to some rule, has minimal expected distortion.

Proof. Suppose that $Q' : \mathcal{X}_\Gamma \to \mathcal{C}$ is a graph quantizer with arbitrary regions. Then we have

$$d(X, Q'(X)) \geq \min_{Y \in \mathcal{Y}} d(X, Y) = d(X, Q(X))$$

for all $X \in \mathcal{X}_\Gamma$. This implies

$$D(Q') = \mathbb{E}_X \left[d\left(X, Q'(X)\right)\right] \geq \mathbb{E}_X \left[d\left(X, Q(X)\right)\right] = D(Q).$$

□

Theorem 4 (Nearest Neighbor Condition). *Suppose that $\mathcal{P}_Q$ is a fixed partition and $Q : \mathcal{X}_\Gamma \to \mathcal{C}$ a graph quantizer with codebook $\mathcal{C}$ satisfying*

$$Y_j = \arg \min_{Y \in \mathcal{X}_\Gamma} \mathbb{E}\left[d\left(X, Y\right) \,|\, X \in \mathcal{R}_j\right]$$

for all $Y \in \mathcal{X}_\Gamma$ and all $j \in \mathcal{J}$. Then Q has minimal expected distortion.

Proof. Let $P_j = P(X \in \mathcal{R}_j)$. Suppose that Q' is a quantizer with partition $\{\mathcal{R}_1, \ldots, \mathcal{R}_k\}$ and arbitrary codebook $\mathcal{C} = \{Y'_1, \ldots, Y'_k\}$. Then we have

$$\begin{aligned}
\mathbb{E}\left[d(X, Q'(X))\right] &= \sum_{j=1}^{k} P_j \mathbb{E}\left[d(X, Q'(X)) \,|\, X \in \mathcal{R}_j\right] \\
&= \sum_{j=1}^{k} P_j \mathbb{E}\left[d(X, Y'_j) \,|\, X \in \mathcal{R}_j\right] \\
&\geq \sum_{j=1}^{k} P_j \min_{Y \in \mathcal{X}_\Gamma} \mathbb{E}\left[d(X, Y) \,|\, X \in \mathcal{R}_j\right] \\
&= \sum_{j=1}^{k} P_j \mathbb{E}\left[d(X, Y_j) \,|\, X \in \mathcal{R}_j\right] = \mathbb{E}\left[d(X, Q(X))\right]
\end{aligned}$$

□

References

1. H. Almohamad and S. Duffuaa, "A linear programming approach for the weighted graph matching problem", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5)522–525, 1993.
2. H. Bunke and B.T. Messmer, "Similarity measures for structured representations", *Lecture Notes in Computer Science*, 837.106–118, 1994.
3. H. Bunke, P. Foggia, C. Guidobaldi, and M. Vento, "Graph clustering using the weighted minimum common supergraph" *Graph Based Representations in Pattern Recognition*, Lecture Notes in Computer Science, 2726:235–246, 2003
4. J.E. Borzellino, *Riemannian geometry of orbifolds*, PhD thesis, University of California, Los Angelos, 1992.
5. T.S. Caetano, L. Cheng, Q.V. Le, and A.J. Smola, "Learning graph matching" *International Conference on Computer Vision*, p. 1–8, 2007.

6. T. Cour, P. Srinivasan, and J. Shi, "Balanced graph matching", *NIPS 2006 Conference Proceedings*, 2006.
7. R.O. Duda, P.E. Hart, and D.G. Stork *Pattern Classification*, Wiley & Sons, 2000.
8. Y.M. Ermoliev and V.I. Norkin, "Normalized convergence in stochastic optimization", *Annals of Operations Research*, 30:187–198, 1991,
9. Y.M. Ermoliev, V.I. Norkin, and R. Wets, "The minimization of discontinuous functions: mollifier subgradients", *SIAM Journal on Control and Optimization*, 33:149–167, 1995.
10. Y.M. Ermoliev and V.I. Norkin, "On nonsmooth and discontinuous problems of stochastic systems optimization", *European Journal of Operational Research*, 101:230–244, 1997.
11. Y. M. Ermoliev and V.I. Norkin, "Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization", *Cybernetics and Systems Analysis*, 34(2), 196–215, 1998.
12. M. Ferrer, *Theory and algorithms on the median graph. application to graph-based classification and clustering*, PhD Thesis, Univ. Aut'onoma de Barcelona, 2007.
13. M. Ferrer, E. Valveny, F. Serratosa, I. Bardají, and H. Bunke, "Graph-Based k-Means Clustering: A Comparison of the Set Median versus the Generalized Median Graph" *CAIP 2009 Conference Proceedings*, 2009.
14. A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.
15. S. Gold and A. Rangarajan, "Graduated Assignment Algorithm for Graph Matching", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18:377–388, 1996.
16. S. Gold, A. Rangarajan, and E. Mjolsness, "Learning with preknowledge: clustering with point and graph matching distance measures" *Neural Computation*, 8(4):787–804, 1996.
17. S. Günter and H. Bunke, "Self-organizing map for clustering in the graph domain", *Pattern Recognition Letters*, 23(4):405–417, 2002.
18. M. Hagenbuchner, A. Sperduti, and A.C. Tsoi, ÒA Self-Organizing Map for Adaptive Processing of Structured Data,Ó *IEEE Transaction on Neural Networks*, 14:491–505, 2003.
19. B. Jain and F. Wysotzki, "Central Clustering of Attributed Graphs", *Machine Learning*, 56, 169–207, 2004.
20. B. Jain and K. Obermayer, "On the sample mean of graphs", *IJCNN 2008 Conference Proceedings*, p. 993–1000, 2008.
21. B. Jain and K. Obermayer, "Structure Spaces", *Journal of Machine Learning Research*, 10:2667–2714, 2009.
22. B. Jain and K. Obermayer, "Accelerating Competitive Learning Graph Quantization", *Computer Vision and Image Understanding*, 2009 (submitted).
23. B. Jain and K. Obermayer, "Elkan's k-Means for Graphs", `arXiv:0912.4598v1 [cs.AI]`, 2009.
24. Y. Linde, A. Buzo, and R. M. Gray, ÒAn algorithm for vector quantizer design,Ó *IEEE Transactions on Communications*, 28:84–95, 1980.
25. S.P. Lloyd, ÒLeast squares quantization in PCMÓ, *IEEE Transactions on Information Theory*, 28:129–137, 1982, reprint of 1957.
26. M.A. Lozano and F. Escolano, "ACM attributed graph clustering for learning classes of images", *Graph Based Representations in Pattern Recognition*, Lecture Notes in Computer Science, 2726:247–258, 2003
27. V.I. Norkin, "Stochastic generalized-differentiable functions in the problem of nonconvex nonsmooth stochastic optimization", *Cybernetics*, 22(6), 804–809, 1986.

28. A. Schenker, M. Last, H. Bunke, and A. Kandel, "Clustering of web documents using a graph model", *Web Document Analysis: Challenges and Opportunities*, p. 1–16, 2003.
29. A. Schenker, M. Last, H. Bunke, and A. Kandel, *Graph-Theoretic Techniques for Web Content Mining*, World Scientific Publishing, 2005.
30. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Elsevier, 2009.
31. A. Torsello and E.R. Hancock, "Learning shape-classes using a mixture of tree-unions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):954-967, 2006.
32. S. Umeyama, "An eigendecomposition approach to weighted graph matching problems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):695–703, 1988.
33. M. Van Wyk, M. Durrani, and B. Van Wyk, "A RKHS interpolator-based graph matching algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):988–995, 2002.