On the Sample Mean of Graphs

Conference Paper · July 2008

DOI: 10.1109/IJCNN.2008.4633920 · Source: IEEE Xplore

CITATIONS
16

READS
58

2 authors:

Brijnesh Jain
Technische Universität Berlin
109 PUBLICATIONS **670** CITATIONS
SEE PROFILE

Klaus Obermayer
Technische Universität Berlin
459 PUBLICATIONS **8,276** CITATIONS
SEE PROFILE

On the Sample Mean of Graphs

Brijnesh Jain and Klaus Obermayer

***Abstract*— We present an analytic and geometric view of the sample mean of graphs. The theoretical framework yields efficient subgradient methods for approximating a structural mean and a simple plug-in mechanism to extend existing central clustering algorithms to graphs. Experiments in clustering protein structures show the benefits of the proposed theory.**

I. Introduction

Graphs often occur as "natural" representations of structured objects in different application areas of machine learning. To adopt methods like central clustering or principal component analysis for graphs, an understanding of the structural version of the sample mean is imperative. But the concept of sample mean of graphs is hardly investigated, although a number of central clustering algorithms for graphs have been devised [1]-[3].

The focus of this paper is on extending the concept of sample mean to graphs. The proposed formulation is based on the median graph [4]. A median graph is a graph with minimal sum of distances from the given sample graphs. This formulation corresponds to the formulation of the standard sample mean as an optimization problem. The median graph is a general and widely applicable measure of central tendency, because it makes no assumptions on the vertex and edge attributes and the underlying graph-edit distance measure. But this generality makes a theoretical analysis and a deeper insight into the nature of the concept more difficult. What we want is a measure of central tendency that summarizes a sample of graphs by recording the relative frequencies of common structural overlaps.

This paper aims at establishing a theoretical basis of the sample mean of graphs. In doing so we try to span a bridge from structural to statistical pattern recognition. The chosen approach is geometrically inspired and yields properties and characterizations that find — to a certain extent — their counterpart in the standard sample mean. We propose two subgradient methods to approximate the sample mean. The proposed framework yields a conceptually simple plug-in mechanism to extend central clustering algorithms like k-means or competitive learning to graphs. We apply both clustering algorithms to the problem of categorizing proteins structures. Although the proposed approach is general enough to apply to finite structures other than graphs, we will for sake of concreteness restrict our attention exclusively to this domain.

II. The basic approach

This section aims at providing a nontechnical overview of the proposed approach.

A weighted graph is a triple $X = (V, E, w)$ consisting of a finite set $V \neq \emptyset$ of *vertices*, a set $E \subseteq V \times V$ of *edges*, and a *weight function* $w : V \times V \to \mathbb{R}$ such that each edge has nonzero weight and each non-edge has weight zero. The weight $w(i, i)$ of vertex i can be any value from $\mathbb{R}$. A weighted graph X of order $|V| = n$ is completely specified by its *weight matrix* $\boldsymbol{X} = (x_{ij})$ with elements $x_{ij} = w(i, j)$ for all $1 \leq i, j \leq n$. Let $\mathcal{G}$ denote the set of weighted graphs.

Suppose that $\mathcal{D}_{\mathcal{T}} = (X_1, \ldots, X_k)$ is a sample of k not necessarily distinct graphs from $\mathcal{G}$. The standard method $M = (X_1 + \cdots X_k)/k$ for determining the sample mean fails, because a well-defined addition of graphs that meets our requirements is unknown. Following [4], we adopt the optimization formulation of the standard sample mean. For vectors, the sample mean minimizes the sum of squared Euclidean distances from the data points. In line with this formulation, we define a sample mean of $\mathcal{D}_{\mathcal{T}}$ as a global minimum of the cost function

$$F : \mathcal{G} \to \mathbb{R}, \quad F(X) = \sum_{i=1}^{k} D(X, X_i)^2, \tag{1}$$

where D is some suitable distance function on $\mathcal{G}$.

In principle, we can use any distance function D. Here, we focus on geometric distance functions that are related to the Euclidean metric, because the Euclidean metric is the underlying metric of the vectorial mean. The vectorial mean in turn provides a link to deep results in probability theory and is the foundation for a rich repository of analytical tools in pattern recognition. To access at least parts of these results, it seems to be reasonable to relate the distance function D to the Euclidean metric. This restriction is acceptable from an application point of view, because geometric distance functions on graphs and their related similarity functions are a common choice of proximity measure in a number of different applications [5-6].

Geometric distance functions D are typically defined as the maximum of a set of Euclidean distances. This definition implies that (1) the cost function F is neither differentiable nor convex; (2) the sample mean of graphs is not unique; and (3) determining a sample mean of graphs is NP-complete, because evaluation of D is NP-complete. Thus, we are faced with an intractable combinatorial optimization problem, where, at a first glance, a solution has to be found from an uncountable infinite set. In addition, multiple global minima of the cost function F complicates a characterization of a structural mean.

To cope with these difficulties, we view graphs as equivalence classes of vectors via their weight matrices, where the elements of the same equivalence class are different vector representations of the same graph. The resulting quotient set

(the set of equivalence classes) leads to the more abstract notion of $\mathcal{T}$-space. Formally, a $\mathcal{T}$-space $\mathcal{X}_\mathcal{T}$ over a vector space $\mathcal{X}$ is a quotient set of $\mathcal{X}$, where the equivalence classes are the orbits of the group action of a transformation group $\mathcal{T}$ on $\mathcal{X}$. The theory of $\mathcal{T}$-spaces generalizes the vector space concept to cope with combinatorial structures and aims at retaining the geometrical and algebraic properties of a vector space to a certain extent.

The $\mathcal{T}$-space concept clears the way to approach the structural version of the sample mean in a principled way. We present a geometric characterization of a structural mean that is closely related to the standard formulation of the vectorial mean. This characterization has important implications: (i) it shows that the solutions can come from a finite discrete set; and (ii) it shows that any sample mean is a well-defined weighted graph. The second important result we show is that the cost function F is locally Lipschitz and therefore differentiable almost everywhere. Hence, we can exploit generalized gradient information to minimize F. Both proposed subgradient methods emerge from this result.

III. $\mathcal{T}$-Spaces

In this section, we develop the theory of $\mathcal{T}$-spaces that allows us to formally adopt geometrical and analytical concepts for graphs. This framework provides the theoretical foundation for deriving results on the sample mean of graphs. All proofs are delegated to the appendix.

The first step to relate graphs to the Euclidean space is to align them to a fixed dimension. Let $\mathcal{G}[n]$ be the set of weighted graphs of order n. We can regard any weighted graph X of order $m < n$ as a graph of order n by adding $p = n - m$ isolated vertices. The weighted adjacency matrix of the aligned graph X' is then of the form

$$\boldsymbol{X'} = \begin{pmatrix} \boldsymbol{X} & \mathbf{0}_{m,p} \\ \mathbf{0}_{p,m} & \mathbf{0}_{p,p} \end{pmatrix},$$

where $\boldsymbol{X}$ is the weight matrix of X, and $\mathbf{0}_{m,p}$, $\mathbf{0}_{p,m}$, $\mathbf{0}_{p,p}$ are padding zero matrices. Using alignment, we can regard $\mathcal{G}[n]$ as the set of weighted graphs of bounded order n. Note that specifying an order n and aligning smaller graphs to graphs of order n are purely technical assumptions to simplify mathematics, which can be safely ignored in a practical setting. We only demand that the graphs are bounded.

The positions of the diagonal elements of $\boldsymbol{X}$ determine an ordering of the vertices. Conversely, different orderings of the vertices may result in different matrices. Since we are interested in the structure of a graph, the ordering of its vertices does not really matter. Therefore, we consider two matrices $\boldsymbol{X}$ and $\boldsymbol{X'}$ as being equivalent, denoted by $\boldsymbol{X} \sim \boldsymbol{X'}$, if they can be obtained from one another by reordering the vertices. Mathematically, the equivalence relation can be written as

$$\boldsymbol{X} \sim \boldsymbol{X'} \Leftrightarrow \exists \boldsymbol{P} \in \mathcal{P} \,:\, \boldsymbol{P}^\top \boldsymbol{X} \boldsymbol{P} = \boldsymbol{X'}, \tag{2}$$

where $\mathcal{P}$ denotes the set of all $(n \times n)$-permutation matrices. The equivalence class $[\boldsymbol{X}]$ of $\boldsymbol{X}$ describes the *abstract weighted graph* that is independent of the particular numbering of the vertices. By $\mathcal{G}[n]/\mathcal{P}$ we denote the set of all abstract weighted graphs of bounded order n. The equivalence relation $\sim$ gives rise to a more abstract and convenient approach of $\mathcal{G}[n]$ and $\mathcal{G}[n]/\mathcal{P}$ that allows to formally adopt statistical, analytical and geometrical concepts. The suggested approach are *$\mathcal{T}$-spaces* that will be introduced in the sequel.

Let $\mathcal{X} = \mathbb{R}^N$ be the N-dimensional Euclidean vector space, and let $\mathcal{T}$ be a subgroup of the group $\mathcal{P}$ of all $(N \times N)$-permutation matrices. Then the binary operation

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

is a group action of $\mathcal{T}$ on $\mathcal{X}$. For $\boldsymbol{x} \in \mathcal{X}$, the *orbit* of $\boldsymbol{x}$ is denoted by $[\boldsymbol{x}]_\mathcal{T} = \{T\boldsymbol{x} \,:\, T \in \mathcal{T}\}$. If no misunderstanding can occur, we write $[\boldsymbol{x}]$ instead of $[\boldsymbol{x}]_\mathcal{T}$.

A *$\mathcal{T}$-space* over $\mathcal{X}$ is the orbit space $\mathcal{X}_\mathcal{T} = \mathcal{X}/\mathcal{T}$ of all orbits of $\boldsymbol{x} \in \mathcal{X}$ under the action of $\mathcal{T}$. We call $\mathcal{X}$ the *representation space* of $\mathcal{X}_\mathcal{T}$. By $\mu : \mathcal{X} \to \mathcal{X}_\mathcal{T}$ we denote the *membership function* that sends vector representations to the structure they describe. We use capital letters $X, Y, Z, \ldots$ to denote the elements of $\mathcal{X}_\mathcal{T}$. Suppose that $X = \mu(\boldsymbol{x})$ for some $\boldsymbol{x} \in \mathcal{X}$. By abuse of notation, we sometimes identify X with $[\boldsymbol{x}]$ and write $\boldsymbol{x} \in X$ instead of $\boldsymbol{x} \in [\boldsymbol{x}]$.

Suppose that $N = n^2$. Then any matrix $\boldsymbol{X}$ representing a weighted graph $X \in \mathcal{G}[n]$ can be regarded as a vector $\boldsymbol{x}$ by stacking the columns of $\boldsymbol{X}$. Let $\mathcal{T}$ be the subgroup of all $(N \times N)$-permutation matrices that corresponds to the set of all $(n \times n)$-permutations matrices for renumbering the vertices of a graph of order n. Obviously, we have a relaxation in the sense that $\mathcal{G}[n] \subseteq \mathcal{X}$ and $\mathcal{G}[n]/\mathcal{P} \subseteq \mathcal{X}_\mathcal{T}$ such that μ restricted on $\mathcal{G}[n]$ sends vector representations to the graphs they represent. Note that there are structures in $\mathcal{X}_\mathcal{T}$ that are not well-defined graphs.

To transfer geometrical properties of the Euclidean space $\mathcal{X}$ to the $\mathcal{T}$-space $\mathcal{X}_\mathcal{T}$, we consider similarity and distance functions on $\mathcal{X}_\mathcal{T}$ of the following type. Let $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function satisfying $f(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. Then f induces symmetric functions

$$F^* : \mathcal{X}_\mathcal{T} \times \mathcal{X}_\mathcal{T} \to \mathbb{R}, \quad (X, Y) \mapsto \max_{\boldsymbol{x} \in X, \boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y}),$$
$$F_* : \mathcal{X}_\mathcal{T} \times \mathcal{X}_\mathcal{T} \to \mathbb{R}, \quad (X, Y) \mapsto \min_{\boldsymbol{x} \in X, \boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y}).$$

Since $\mathcal{T}$ is finite, the orbits $[\boldsymbol{x}]$ of $\boldsymbol{x}$ are finite. Hence, F^* and F_* assume an extremal value. We call F^* *maximizer* and F_* *minimizer* of f on $\mathcal{X}_\mathcal{T} \times \mathcal{X}_\mathcal{T}$. Let F be either a maximizer or minimizer of f. The *support* of F at $(X, Y) \in \mathcal{X}_\mathcal{T} \times \mathcal{X}_\mathcal{T}$ is the set defined by

$$\operatorname{supp} F(X, Y) = \{(\boldsymbol{x}, \boldsymbol{y}) \in X \times Y \,:\, F(X, Y) = f(\boldsymbol{x}, \boldsymbol{y})\}.$$

The class of similarity functions on $\mathcal{X}_\mathcal{T}$ we consider are maximizers of inner products $\langle \cdot, \cdot \rangle$ on $\mathcal{X}$. The *inner $\mathcal{T}$-product* induced by $\langle \cdot, \cdot \rangle$ is defined by

$$\langle \cdot, \cdot \rangle^* : \mathcal{X}_\mathcal{T} \times \mathcal{X}_\mathcal{T} \to \mathbb{R}, \quad (X, Y) \mapsto \max_{\boldsymbol{x} \in X, \boldsymbol{y} \in Y} \langle \boldsymbol{x}, \boldsymbol{y} \rangle.$$

The inner $\mathcal{T}$-product is *not* an inner product, because the maximum-operator in the definition of $\langle \cdot, \cdot \rangle^*$ does not preserve the bilinearity property of an inner product. But as

we will see shortly, the inner $\mathcal{T}$-product has the same geometrical properties as the standard inner product.

Any inner product space $\mathcal{X}$ is a normed space with norm $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ and a metric space with metric $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$. The norm $\|\cdot\|$ and the metric d on $\mathcal{X}$ give rise to minimizers $\|\cdot\|_*$ of $\|\cdot\|$ and D_* of d on $\mathcal{X}_\mathcal{T}$. Since elements from $\mathcal{T}$ preserve lengths and angles, we have $\|T\boldsymbol{x}\| = \|\boldsymbol{x}\|$ for all $T \in \mathcal{T}$. Hence, a $\mathcal{T}$-norm $\|X\|_*$ is independent from the choice of vector representation. We show that a $\mathcal{T}$-norm is related to an inner $\mathcal{T}$-product in the same way as a norm to an inner product.

Proposition 1: Let $\mathcal{X}_\mathcal{T}$ be a $\mathcal{T}$-space over $(\mathcal{X}, \langle \cdot, \cdot \rangle)$, and let $X \in \mathcal{X}_\mathcal{T}$. Then

1) $\langle X, X \rangle^* = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$ for all $\boldsymbol{x} \in X$.
2) $\|X\|_* = \sqrt{\langle X, X \rangle^*}$.

The inner $\mathcal{T}$-product together with its $\mathcal{T}$-norm satisfies a structural version of the Cauchy-Schwarz inequality.

Theorem 1: Let $\mathcal{X}_\mathcal{T}$ be a $\mathcal{T}$-space over an Euclidean space $(\mathcal{X}, \langle \cdot, \cdot \rangle)$, and let $X, Y \in \mathcal{X}_\mathcal{T}$. Then

$$\left| \langle X, Y \rangle^* \right| \leq \|X\|_* \|Y\|_* .$$

Using Theorem 1, we can introduce the notion of angle between structures in the usual way. This shows that the inner $\mathcal{T}$-product has the same geometrical properties as the standard inner product.

The next result states that the minimizer D_* of the Euclidean metric $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ is also a metric.

Theorem 2: Let $\mathcal{X}$ be an Euclidean space with Euclidean metric d. Then the minimizer D_* of d is a metric on the $\mathcal{T}$-space $\mathcal{X}_\mathcal{T}$.

We call the minimizer D_* of a Euclidean metric d on $\mathcal{X}$ the $\mathcal{T}$-metric induced by d. The $\mathcal{T}$-metric D_* can also be expressed in terms of $\langle \cdot, \cdot \rangle^*$.

Proposition 2: We have

$$D_*(X, Y)^2 = \|X\|_*^2 - 2 \langle X, Y \rangle^* + \|Y\|_*^2 .$$

A *$\mathcal{T}$-function* is a function of the form $F : \mathcal{X}_\mathcal{T} \to \mathbb{R}$, where $\mathcal{X}_\mathcal{T}$ is a T-space over $\mathcal{X}$. Instead of considering the $\mathcal{T}$-function F, it is often more convenient to consider its *representation function*

$$f : \mathcal{X} \to \mathbb{R}, \quad \boldsymbol{x} \mapsto F \circ \mu(\boldsymbol{x}),$$

which is invariant under transformations from $\mathcal{T}$.

IV. The sample mean of graphs

Throughout this section we assume that $\mathcal{X}$ is an Euclidean space with inner product $\langle \cdot, \cdot \rangle$, and $\mathcal{X}_\mathcal{T}$ is a $\mathcal{T}$-space over $\mathcal{X}$. Let $\mathcal{D}_\mathcal{T} = (X_1, \ldots, X_k)$ be a data sample of k (not necessarily distinct) elements from $\mathcal{X}_\mathcal{T}$. A *sample mean* of $\mathcal{D}_\mathcal{T}$ is any solution of the following optimization problem

$$(P_\mathcal{T}) \qquad \begin{array}{ll} \text{minimize} & F(X) = \sum_{i=1}^k D_*(X, X_i)^2 \\ \text{subject to} & X \in \mathcal{X}_\mathcal{T} \end{array} ,$$

where D_* is the $\mathcal{T}$-metric induced by the Euclidean metric d on $\mathcal{X}$. The *representation sample* of the sample $\mathcal{D}_\mathcal{T}$ is defined by the set $\mathcal{D} = X_1 \times \cdots \times X_k = \mu^{-1}(X_1) \times \cdots \times \mu^{-1}(X_k)$. Each element of $\mathcal{D}$ is a k-tuple $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$, where the components $\boldsymbol{x}_i$ are vector representations of X_i. We identify the elements of the representation sample $\mathcal{D}$ with matrices. A *matrix representation* of $\mathcal{D}_\mathcal{T}$ is a $(n \times k)$-matrix $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_k)$, where the columns of $\boldsymbol{X}$ form a k-tuple of $\mathcal{D}$.

Suppose that $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_k) \in \mathcal{D}$ is a matrix representation of $\mathcal{D}_\mathcal{T}$. We can rewrite $(P_\mathcal{T})$ to the equivalent form

$$(P) \qquad \begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) = \sum_{i=1}^k \min_{\boldsymbol{x_i} \in X_i} \|\boldsymbol{x} - \boldsymbol{x}_i\|^2 \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X} \end{array}$$

where $f = F \circ \mu$ is a representation function of F. Note that by Appendix A, Prop. 4, the solutions of problem (P) are independent from the particular choice of $\boldsymbol{x} \in X$. The following result shows that problem $(P_\mathcal{T})$ has a solution, which is in general not unique.

Proposition 3: Problem $(P_\mathcal{T})$ has a solution.

A. Characterization of the sample mean

Theorem 3 shows that a sample mean of $\mathcal{D}_\mathcal{T}$ is of similar form as the sample mean of a set of vectors.

Theorem 3: Let $\mathcal{D}_\mathcal{T} = (X_1, \ldots, X_k)$ be a sample of k structures from $\mathcal{X}_\mathcal{T}$. Then any vector representation $\boldsymbol{m}$ of a sample mean $M \in \mathcal{X}_\mathcal{T}$ of $\mathcal{D}_\mathcal{T}$ is of the form

$$\boldsymbol{m} = \frac{1}{k} \sum_{i=1}^k \boldsymbol{x}_i,$$

where $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_k)$ is a matrix representation of $\mathcal{D}_\mathcal{T}$ satisfying $(\boldsymbol{m}, \boldsymbol{x}_i) \in \operatorname{supp} D_*(M, X_i)$ for all $i \in \{1, \ldots .k\}$.

We call a matrix representation $\boldsymbol{X}$ of $\mathcal{D}_\mathcal{T}$ *conform* if the sample mean of the columns of $\boldsymbol{X}$ is a solution to problem (P), i.e. a vector representation of a sample mean of $\mathcal{D}_\mathcal{T}$.

Theorem 3 has three implications. First, any sample mean of well-defined graphs from the subset $\mathcal{G}[n]/\mathcal{P}$ of $\mathcal{X}_\mathcal{T}$ is a well-defined weighted graph from $\mathcal{G}[n]/\mathcal{P}$. Second, problem $(P_\mathcal{T})$ is a discrete combinatorial optimization problem. Any solution $\boldsymbol{m}$ of the equivalent problem (P) is a sample mean of the columns of a conform matrix representation $\boldsymbol{X} \in \mathcal{D} = X_1 \times \cdots \times X_k$. Since the group $\mathcal{T}$ is finite, the set $\mathcal{D}$ is also finite. Hence, solving problem $(P_\mathcal{T})$ reduces to finding a conform matrix representation $\boldsymbol{X}$ from a finite set $\mathcal{D}$. Third, Theorem 3 includes equivalent geometric characterizations of a sample mean. The remainder of this subsection is devoted to this issue.

The common tool to derive geoemtric characterizations of the sample mean is the Gram matrix. Let $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_k)$ be a matrix representation of a sample $\mathcal{D}_\mathcal{T} = (X_1, \ldots, X_k)$ of structures. The *Gram matrix* of $\boldsymbol{X}$ is defined by $\boldsymbol{G} = \boldsymbol{G}_{\boldsymbol{X}} = \boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}$, where $\boldsymbol{A}$ denotes the symmetric matrix representing the inner product $\langle \cdot, \cdot \rangle$ of $\mathcal{X}$. Thus, the elements of $\boldsymbol{G} = (g_{ij})$ are of the form $g_{ij} = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$. The *Gram sum* of $\boldsymbol{X}$ is defined by $\Gamma(\boldsymbol{X}) = \sum_{i=1}^n \sum_{j=1}^n g_{ij}$.

Theorem 4: Let $\mathcal{D}_\mathcal{T} = (X_1, \ldots, X_k)$ be a sample of structures from $\mathcal{X}_\mathcal{T}$. Then $\boldsymbol{X}$ is a conform matrix representation of $\mathcal{D}_\mathcal{T}$ if, and only if, $\Gamma(\boldsymbol{X}) \geq \Gamma(\boldsymbol{X'})$ for all $\boldsymbol{X'} \in \mathcal{D}$.

Theorem 4 states that a structure M is a sample mean of $\mathcal{D}_\mathcal{T} = (X_1, \ldots, X_k)$ if, and only if, it is represented by the sample mean of vector representations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \in \mathcal{D}$

with maximal Gram sum. Since the Gram sum contains information about pairwise similarities, distances, correlations, and angles, we can derive the following equivalent geometrical statements: A sample mean of graphs $X_1, \ldots, X_k$ is represented by the sample mean of those vector representations of X_i that have (1) maximal average pairwise Gram similarity, (2) minimal average pairwise Euclidean distance, (3) maximal average correlation coefficient, and (4) minimal average pairwise angle.

B. Subgradient methods for computing a sample mean

The cost function f of problem (P) is non-differentiable. As a consequence of [7], the function f is locally Lipschitz and therefore non-differentiable on a set of Lebesgue measure zero by Rademacher's Theorem. To minimize locally Lipschitz functions, the field of nonsmooth optimization offers a number of techniques [7]. The simplest and probably also the most used method for non-differentiable optimization are subgradient methods. The basic idea is to replace gradients by subgradients at non-differentiable points.

Algorithm 1 (Subgradient Method (SGM))

01 choose starting point $\boldsymbol{x}^1 \in \mathcal{X}$ and set $t := 0$
02 **repeat**
03 set $\tilde{\boldsymbol{x}}^{t,1} := \boldsymbol{x}^1$
04 **for** $i = 1, \ldots, k$ **do**
05 • choose $\boldsymbol{x}_i \in X_i$ with
06 $\left(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}^{t,1}\right) \in \operatorname{supp}\left(D_*\left(X_i, \tilde{X}^{t,1}\right)\right)$
07 • determine step size $\eta^{t,i} > 0$
08 • set $\tilde{\boldsymbol{x}}^{t,i+1} := \tilde{\boldsymbol{x}}^{t,i} + \eta^{t,i}\boldsymbol{x}_i / \|\boldsymbol{x}_i\|$
09 **if** $f\left(\boldsymbol{x}^t\right) > f\left(\tilde{\boldsymbol{x}}^{t,k+1}\right)$ **then** set $\boldsymbol{x}^{t+1} := \tilde{\boldsymbol{x}}^{t,k+1}$
10 Set $t := t + 1$
11 **until** some termination criterion is satisfied

For vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, the sample mean can be determined incrementally by setting $\boldsymbol{m}_1 = \boldsymbol{x}_1$ if $k = 1$ and $\boldsymbol{m}_k = (k-1)\boldsymbol{m}_{k-1}/k + \boldsymbol{x}_k/k$ if $k > 1$. In a similar way, we can approximate the sample mean of structures. In contrast to feature vectors, the incremental version of a sample mean for graphs is in general not a globally optimal and can be used as an efficient approximation, because it requires exactly k evaluations of D_*.

Algorithm 2 (Iterative Mean Algorithm (IMA))

01 set $\mathcal{I} := \{1, \ldots, k\}$ and set $t := 1$
02 randomly select $i \in \mathcal{I}$ and $\boldsymbol{x}_i \in X_i$. Set $\mathcal{I} := \mathcal{I} \setminus \{i\}$
03 set $\boldsymbol{m} := \boldsymbol{x}_i$
04 **while** $\mathcal{I} \neq \emptyset$ **do**
05 • randomly select $i \in \mathcal{I}$
06 • set $\mathcal{I} := \mathcal{I} \setminus \{i\}$ and set $t := t + 1$
07 • choose $\boldsymbol{x}_i \in X_i$ with
08 $(\boldsymbol{x}_i, \boldsymbol{m}) \in \operatorname{supp}\left(D_*\left(X_i, M\right)\right)$
09 • set $\boldsymbol{m} := t \cdot \boldsymbol{m}/(t-1) + \boldsymbol{x}_i/t$

C. Applying the sample mean to central clustering

Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ be a training sample of m graphs X_i drawn from $\mathcal{X}_\mathcal{T}$. The aim of central clustering is to find k cluster centers $\mathcal{Y} = \{Y_1, \ldots, Y_k\} \subseteq \mathcal{X}_\mathcal{T}$ such that

$$F(\boldsymbol{M}, \mathcal{Y}, \mathcal{X}) = \frac{1}{m} \sum_{j=1}^{k} \sum_{i=1}^{m} m_{ij} D_*(X_i, Y_j),$$

is minimized with respect to a given distortion measure D_*. The matrix $\boldsymbol{M} = (m_{ij})$ is a $(m \times k)$-membership matrix with elements $m_{ij} \in [0, 1]$ with $\sum_j m_{ij} = 1$ for all $i = 1, \ldots, m$. It can be shown that F as a function of the cluster centers Y_j is locally Lipschitz and therefore can be minimized by exploiting local subgradient information.

The last results provides new insight why existing central clustering methods [1-3] converge to local optima, although neither the cluster centers nor the update rule is well-defined. Intuitively, one might expect that central clustering could be prone to oscillations halfway between different vector representations of a cluster center. This phenomenon, however, is unlikely if the cluster criterion F is locally Lipschitz, because the gradient and therefore the update rule is well-defined at almost all points. By using a decreasing step size, the aforementioned oscillations can be avoided.

It is straightforward to extend central clustering methods like k-means and competitive learning to graphs. Structural k-means operates as the EM algorithm of standard k-means, where the chosen distortion measure in the E-step is D_* and either SGM or IMA is applied in the M-step to recompute the means. Structural competitive learning operates as simple competitive learning, where the standard inner product is replaced by $\langle \cdot, \cdot \rangle^*$. The update rule is a subgradient step of the form $\boldsymbol{w} := \boldsymbol{w} + \eta\boldsymbol{x}$, where $\boldsymbol{w}$ and $\boldsymbol{x}$ are vector representations of the weight graph W and the input graph X, resp., such that $(\boldsymbol{w}, \boldsymbol{x}) \in \operatorname{supp}\left(\langle W, X \rangle^*\right)$.

V. Experiments

A. Random graphs

To assess the performance of SGM and IMA, we used synthetic data that provide — at least to a certain extent — ground truth information. To execute step 5 in SGM and step 7 in IMA, we applied the graduated assignment algorithm proposed by [5].

The test data was set up as follows: A sample of 20 weighted graphs was drawn by first generating a random model graph M of order 10 and with edge probability 0.5. Next we generated the sample graphs as noisy versions of M by converting each edge (non-edges) to a non-edge (edge) with 5% probability. We added Gaussian noise with zero mean and deviation σ to the vertex and edge weights. Finally, we permuted the sample graphs. For each value of $\sigma \in \{0.01, 0.03, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0\}$, we generated 100 samples of 20 graphs. To assess an estimate of the solution quality, we used the arithmetic means M_A as our ground truth information. The arithmetic mean M_A of each sample is obtained by taking the average of the 20 weighted adjacency matrices after corruption and before permutation.

Figure 1 shows that the solution quality of IMA is comparable with SGM but clearly superior with respect to computation time. Although a fair comparison is intricate,

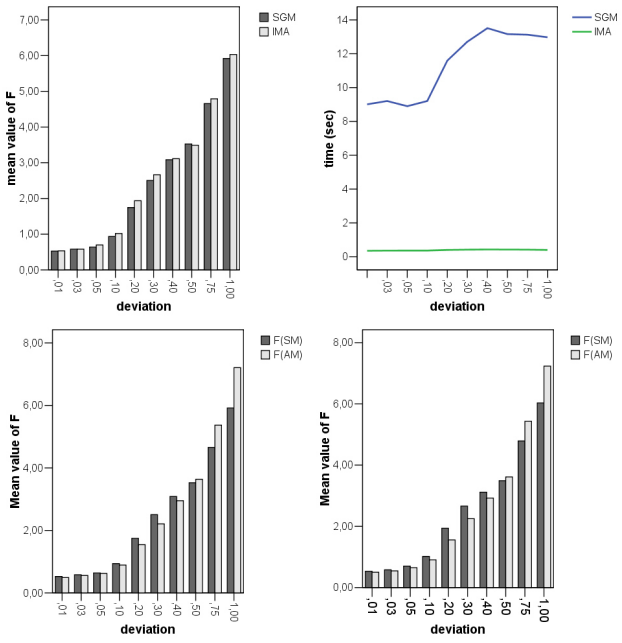

Fig. 1. Shown are the mean values $F(M_S)$ of SGM and IMA (**left**), the mean computation time of SGM and IMA in seconds (**2nd left**), the mean values $F(M_S)$ and $F(M_A)$ of SGM (**2nd right**) and IMA (**right**).

IMA is likely to be faster than the genetic algorithm proposed by [4]. For 20 graphs of order about 6 the genetic algorithm required in average about 40 seconds and IMA on a slightly larger problem about 0.4 seconds. Figure 1 shows that for deviations $\sigma < 0.5$ both algorithms approximate the ground truths M_A satisfactorily. For deviation $\sigma \geq 0.5$ nothing can be said about the solution quality, because the graphs are disrupted such that the sample means have lower error than the ground truths M_A.

B. Skolnick clustering test set

The 3D structure of a protein can provide important clues about how the protein performs it function. In this context, central clustering can be applied as a tool for categorizing similar protein structures and providing a template structure for homology modeling. One common model of the protein structure is the contact map. A contact map is a graph on an ordered set of vertices. The vertices represent residues. Edges connect two vertices if the corresponding residues are spatially close. Comparing two contact maps aims at finding an order preserving alignment of the vertices such that the number of common contacts is maximized. This optimization problem is referred to as the contact map overlap (CMO) problem, which is known to be NP-hard.

To demonstrate the utility of the proposed framework, we applied structural k-means and structural competitive learning to categorization and template generation of protein structures. We used a combination of softassign and dynamic programming [8] for solving the CMO problem. We applied the clustering algorithms to 40 proteins from the Skolnick test set. The contact maps of the domains were provided by [9]. The protein domains of the Skolnick test data are shown in Table I. Identifier ID refers to the index assigned to the domains, identifier PDB to the PDB code for the protein containing the domain, and identifier CID to the chain index of a protein. If a protein consists of a single chain, the corresponding entry in the CID column is left empty. Note that the IDs differ from those used in [9-10].

TABLE I

PROTEIN DOMAINS OF THE SKOLNICK TEST SET.

ID	PDB	CID	ID	PDB	CID	ID	PDB	CID	ID	PDB	CID
1	1b00	A	11	4tmy	B	21	2b3i	A	31	1tri	
2	1dbw	A	12	1rn1	A	22	2pcy		32	3ypi	A
3	1nat		13	1rn1	B	23	2plt		33	8tim	A
4	1ntr		14	1rn1	C	24	1amk		34	1ydv	A
5	1qmp	A	15	1baw	A	25	1aw2	A	35	1b71	A
6	1qmp	B	16	1byo	A	26	1b9b	A	36	1bcf	A
7	1qmp	C	17	1byo	B	27	1btm	A	37	1dps	A
8	1qmp	D	18	1kdi		28	1hti	A	38	1fha	
9	3chy		19	1nin		29	1tmh	A	39	1ier	
10	4tmy	A	20	1pla		30	1tre	A	40	1rcd	

Table II describes the protein domains and their families. Shown are the characteristics of the four families, the mean number of residues, the range of similarity obtained by sequence alignment and the identifiers of the protein domains.

The task is to classify the proteins into families according to their cluster membership. The characteristic feature of the Skolnick data is that sequence similarity fails for correct categorization of the proteins as indicated by the fourth column (*Seq. Sim.*) of Table II. This motivates structural alignment for solving the Skolnick clustering test. Previous approaches used pairwise clustering to categorize the proteins [9-10]. Competitive learning and the approaches from [9-10] correctly categorized the 40 proteins in 5 clusters as shown in Table III. Fold, family, and superfamily are according to the SCOP categories. K-Means found the same clusters but misclassified one protein.

Competitive learning is the fastest method with 120 and k-means the slowest with 960 pairwise structural alignments. Pairwise clustering methods require 780 structural alignments. The novel feature is that we are able to determine a sample mean of each cluster serving as a template of a category. Figure 2 shows approximations of sample means of the two largest clusters computed by competitive learning.

TABLE II

PROTEIN DOMAINS OF THE SKOLNICK TEST SET AND THEIR CATEGORIES AS TAKEN FROM [10].

Family	Style	Residues	Seq. Sim.	Proteins
1	alpha-beta	124	15-30%	1-14
2	beta	99	35-90%	15-23
3	alpha-beta	250	30-90%	24-34
4		170	7-70%	35-40

TABLE III

CLUSTERS OF PROTEINS FROM THE THE SKOLNICK TEST.

C	Dom.	Fold	Superfamily	Family
1	1-11	Flavodin-like	Che Y-like	Che Y-related
2	12-14	Microbial ribonucl.	Microbial ribonucl.	Fungi ribonucl.
3	15-23	Cuperdoxin-like	Cuperdoxins	Plastocyanim-like Plastoazurin-like
4	24-34	TIM-beta alpha-barrel	Triosephosphate isomerase (TIM)	Triosephosphate isomerase (TIM)
5	35-40	Ferritin-like	Ferritin-like	Ferritin

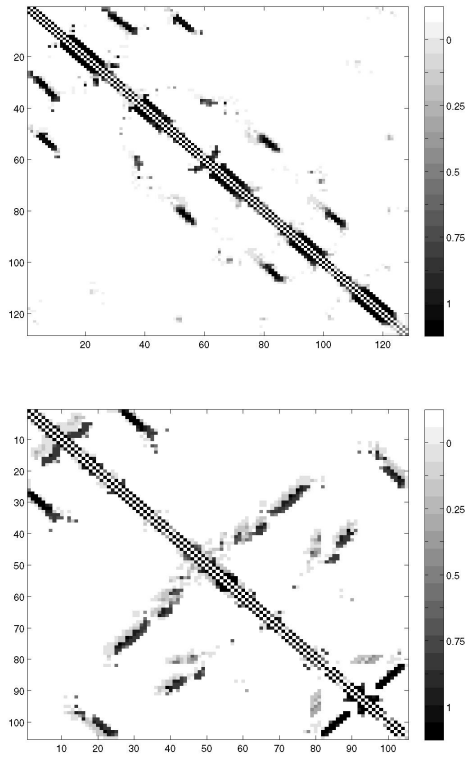

Fig. 2. Shown are approximated sample means of the Che Y-like superfamily (**left**) and the Cuperdoxins superfamily family (**right**). Diagonal elements show the residues and off-diagonal elements the contacts. Darker shading refers to a higher relative frequency of occurrence of residues/contacts over all cluster members.

VI. CONCLUSION

Based on the theory of $\mathcal{T}$-spaces, we have presented a principled approach to characterize the sample mean of graphs. We proposed a subgradient method and an incremental mean algorithm to approximate a structural mean and plugged the concepts into k-means and simple competitive learning. As shown in the experiments, this theoretical framework provides useful tools to generate templates and to provide a summary of a sample of structures, that captures the relative frequency of common structural overlaps.

REFERENCES

[1] S. Gold, A. Rangarajan, and S. Mjolsness, "Learning with preknowledge: Clustering with point and graph matching distance measures," *Neural Computation*, 8(4):787–804, 1996.

[2] S. Günter and H. Bunke, "Self-organizing map for clustering in the graph domain," *Pattern Recognition Letters*, 23:401–417, 2002.

[3] M.A. Lozano and F. Escolano, "Protein classification by matching and clustering surface graphs," *Pattern Recognition*, 39(4):539–551, 2006.

[4] X. Jiang, X., A. Munger, and H. Bunke, "On Median Graphs: Properties, Algorithms, and Applications," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(10):1144–1151, 2001.

[5] S. Gold and A. Rangarajan, "Graduated Assignment Algorithm for Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18:377–388, 1996.

[6] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of Molecular Biology*, 233(1):123–38, 1993.

[7] M. Mäkelä and P. Neittaanmäki, *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control.* World Scientific, 1992.

[8] B.J. Jain and M. Lappe, "Joining Softassign and Dynamic Programming for the Contact Map Overlap Problem," In *1st Int. Conf. Bioinformatics Research and Development (BIRD'07)*, 410–423, 2007.

[9] W. Xie and N.V. Sahinidis, "A Branch-and-Reduce Algorithm for the Contact Map Overlap Problem," In *10th Ann. Int. Conf. Research in Computational Molecular Biology (RECOMB'06)*, 516–529, 2006.

[10] A. Caprara and G. Lancia, "Structural alignment of large–size proteins via Lagrangian relaxation," *6th Ann. Int. Conf. Research in Computational Molecular Biology (RECOMB'02)*, 100–108, 2002.

APPENDIX

A. $\mathcal{T}$-Spaces

This section is a more detailed treatment of Section 3 including the results its results and their proofs.

Proposition 4: Let F be the minimizer or maximizer of a function $f : \mathcal{X}^n \to \mathbb{R}$. Let $X_1, \ldots, X_n \in \mathcal{X}_\mathcal{T}$. Then for each $\boldsymbol{x}_i \in X_i$ there are a $\boldsymbol{x}_j \in X_j$ for all $j \in \{1, \ldots, n\} \setminus \{i\}$ such that $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \operatorname{supp}(F|X_1, \ldots, X_n)$.

Proof: Let $\boldsymbol{x}_i \in X_i$. Suppose that $(\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*) \in \operatorname{supp}(F|X_1, \ldots, X_n)$. Then there is a transformation $T \in \mathcal{T}$ with $T\boldsymbol{x}_i = \boldsymbol{x}_i^*$. Since $\mathcal{T}$ is a group, the inverse T^{-1} exists. Hence, $\boldsymbol{x}_j = T^{-1}\boldsymbol{x}_j^*$ is an element of X_j for all $j \neq i$. Since T is an isometry, we obtain $F(X_1, \ldots, X_n) = f\left(\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*\right) = f\left(T\boldsymbol{x}_1, \ldots, T\boldsymbol{x}_n\right) = f\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)$. This shows the assertion. ■

1) Metric $\mathcal{T}$-Spaces:

Theorem 5: Let $\mathcal{X}_\mathcal{T}$ be a $\mathcal{T}$-space over the metric space $(\mathcal{X}, d)$. Then the minimizer

$$D_* : \mathcal{X}_\mathcal{T} \times \mathcal{X}_\mathcal{T} \to \mathbb{R}, \quad (X, Y) \mapsto \min_{\boldsymbol{x} \in X, \boldsymbol{y} \in Y} d(\boldsymbol{x}, \boldsymbol{y}).$$

is a metric on $\mathcal{X}_\mathcal{T}$.

Proof: Let X, Y, $Z \in \mathcal{X}_\mathcal{T}$.

1) We show $D_*(X, Y) = 0 \Leftrightarrow X = Y$. Let $\boldsymbol{x} \in X$ be a representation vector of X. According to Prop. 4 there is a $\boldsymbol{y} \in Y$ such that $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{supp}(D_*|\boldsymbol{x}, \boldsymbol{y})$. We have

$$\begin{aligned} D_*(X, Y) = 0 &\Leftrightarrow \forall \boldsymbol{x} \in X \, \exists \boldsymbol{y} \in Y \; d(\boldsymbol{x}, \boldsymbol{y}) = 0 \\ &\Leftrightarrow \forall \boldsymbol{x} \in X \, \exists \boldsymbol{y} \in Y \; \boldsymbol{x} = \boldsymbol{y} \\ &\Leftrightarrow X = Y. \end{aligned}$$

2) Symmetry of D_* follows from symmetry of d.
3) We show $D_*(X, Z) \leq D_*(X, Y) + D_*(Y, Z)$. Let $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{supp}(D_*|X, Y)$. There is a $\boldsymbol{z} \in Z$ such that $(\boldsymbol{y}, \boldsymbol{z}) \in \operatorname{supp}(D_*|Y, Z)$. Then

$$\begin{aligned} D_*(X, Y) + D_*(Y, Z) &= d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}) \\ &\geq d(\boldsymbol{x}, \boldsymbol{z}) \\ &\geq \min_{\boldsymbol{x} \in X, \boldsymbol{z} \in Z} d(\boldsymbol{x}, \boldsymbol{z}) \\ &= D_*(X, Z). \end{aligned}$$

■

Given the assumptions of Theorem 5, we call $(\mathcal{X}_\mathcal{T}, D_*)$ *metric $\mathcal{T}$-space* over $(\mathcal{X}, d)$.

Theorem 6: Any $\mathcal{T}$-space over a complete metric vector space is a complete metric space.

Proof: Let $\mathcal{X}_\mathcal{T}$ be a $\mathcal{T}$-space over the complete metric space $(\mathcal{X}, d)$. According to Theorem 5, $\mathcal{X}_\mathcal{T}$ is a metric space with metric D_*. To show that $\mathcal{X}_\mathcal{T}$ is complete, consider an arbitrary Cauchy sequence $(X_i)_{i\in\mathbb{N}}$ in $\mathcal{X}_\mathcal{T}$. We construct a Cauchy sequence $(\boldsymbol{x}_k)$ such that $(\mu(\boldsymbol{x}_k))$ is a subsequence of (X_i). For any $k > 0$ there is a n_k such that $D_*(X_i, X_j) < 1/2^k$ for all $i, j > n_k$. For each k, there are $\boldsymbol{x}_k \in X_{n_k}$ and $\boldsymbol{x}_{k+1} \in X_{n_{k+1}}$ with $d(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) \leq 1/2^k$. By the triangle inequality, we have

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \sum_{k=i}^{j-1} d(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) \leq \frac{1}{2^{i-1}}$$

for any i, j with $i < j$. Hence, $(\boldsymbol{x}_k)$ is a Cauchy sequence in $\mathcal{X}$ and $(\mu(\boldsymbol{x}_k))$ a subsequence of (X_i). Since $\mathcal{X}$ is complete, $(\boldsymbol{x}_k)$ converges to a limit point $\boldsymbol{x} \in \mathcal{X}$. By continuity of μ, we have $\lim_{k\to\infty} \mu(\boldsymbol{x}_k) = \mu(\boldsymbol{x})$, where $\mu(\boldsymbol{x}) \in \mathcal{X}_\mathcal{T}$. Thus, the whole sequence (X_i) converges to $\mu(\boldsymbol{x})$. This shows that $\mathcal{X}_\mathcal{T}$ is complete. ∎

2) $\mathcal{T}$-Spaces over Normed Vector Spaces: Let $\mathcal{X}_\mathcal{T}$ be a $\mathcal{T}$-space over the normed vector space $(\mathcal{X}, \|\cdot\|)$. As a normed vector space, $\mathcal{X}$ is a metric space with metric $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. For any homogeneous isometry T of $\mathcal{X}$, we have

$$\|T\boldsymbol{x}\| = \|T\boldsymbol{x} - \boldsymbol{0}\| = \|T\boldsymbol{x} - T\boldsymbol{0}\| = \|\boldsymbol{x} - \boldsymbol{0}\| = \|\boldsymbol{x}\|.$$

Hence, the minimizer $\|\cdot\|_*$ and maximizer $\|\cdot\|^*$ of $\|\cdot\|$ coincide, that is

$$\|X\|_* = \|X\|^* = \|\boldsymbol{x}\| \tag{3}$$

for all $X \in \mathcal{X}_\mathcal{T}$ and for all $\boldsymbol{x} \in X$.

We call the minimizer $\|\cdot\|_*$ the *$\mathcal{T}$-norm* induced by the norm $\|\cdot\|$. Note that a $\mathcal{T}$-norm is *not* a norm, because a $\mathcal{T}$-space has no well defined addition. But we can show that a $\mathcal{T}$-norm has norm-like properties. We use the notations λX for $\mu(\lambda X)$ and $X_{\boldsymbol{x}} + Y_{\boldsymbol{y}}$ for $\mu(\boldsymbol{x} + \boldsymbol{y})$.

Proposition 5: Let $(\mathcal{X}_\mathcal{T}, \|\cdot\|_*)$ be a $\mathcal{T}$-space over the normed space $(\mathcal{X}, \|\cdot\|)$. For all $X, Y \in \mathcal{X}_\mathcal{T}$, we have

1) $\|X\|_* = 0$ if, and only if, $X = 0_\mathcal{T}$.
2) $\|\lambda X\|_* = |\lambda| \, \|X\|_*$ for all $\lambda \in \mathbb{R}$.
3) $\|X_{\boldsymbol{x}} + Y_{\boldsymbol{y}}\|_* \leq \|X\|_* + \|Y\|_*$ for all $\boldsymbol{x} \in X$, $\boldsymbol{y} \in Y$.

Proof: Follows directly from first applying Equation (3) and then using the properties of the norm $\|\cdot\|$ defined on $\mathcal{X}$. ∎

3) $\mathcal{T}$-Spaces over Inner Product Spaces: Let $\mathcal{X}_\mathcal{T}$ be a $\mathcal{T}$-space over the inner product space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$, and let

$$\langle\cdot,\cdot\rangle^* : \mathcal{X}_\mathcal{T} \times \mathcal{X}_\mathcal{T} \to \mathbb{R}, \quad (X, Y) \mapsto \max\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle \,:\, \boldsymbol{x} \in X, \boldsymbol{y} \in Y\}$$

be the maximizer of the inner product $\langle\cdot,\cdot\rangle$.

We call $\langle\cdot,\cdot\rangle^*$ *inner $\mathcal{T}$-product* induced by the inner product $\langle\cdot,\cdot\rangle$. The inner $\mathcal{T}$-product is *not* an inner product, because the maximum-operator in the definition of $\langle\cdot,\cdot\rangle^*$ does not preserve the bilinearity property of an inner product. But we shall show later that an inner $\mathcal{T}$-product satisfies some weaker properties of an inner product.

Any inner product space $\mathcal{X}$ is a normed space with norm $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}$. The norm $\|\cdot\|$ on $\mathcal{X}$ in turn gives rise to the $\mathcal{T}$-norm $\|\cdot\|_*$ on $\mathcal{X}_\mathcal{T}$.

Proposition 6: Let $(\mathcal{X}_\mathcal{T}, \langle\cdot,\cdot\rangle^*)$ be a $\mathcal{T}$-space over the inner product space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$, and let $X \in \mathcal{X}_\mathcal{T}$. Then

1) $\langle X, X\rangle^* = \langle \boldsymbol{x}, \boldsymbol{x}\rangle$ for all $\boldsymbol{x} \in X$.
2) $\|X\|_* = \sqrt{\langle X, X\rangle^*}$.

Proof:

1) X is the orbit of $\boldsymbol{x}$ under the group action $\mathcal{T}$. The assertion follows from the fact that each transformation T of $\mathcal{T}$ satisfies
$$\langle T\boldsymbol{x}, T\boldsymbol{x}\rangle = \|T\boldsymbol{x}\|^2 = \|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x}\rangle$$
for all $\boldsymbol{x} \in \mathcal{X}$.
2) Follows from the first part by taking the square root.

∎

Next we show that an inner $\mathcal{T}$-product satisfies some weaker properties related to an inner product.

Proposition 7: Let $X, Y, Z \in \mathcal{X}_\mathcal{T}$, and let $\boldsymbol{x} \in X$, $\boldsymbol{y} \in Y$. Then

1) $\langle X, X\rangle^* \geq 0$ with $\langle X, X\rangle^* = 0 \Leftrightarrow X = 0_\mathcal{T}$ *(pos. definite)*
2) $\langle X, Y\rangle^* = \langle Y, X\rangle^*$ *(symmetric)*
3) $\langle \lambda X, Y\rangle^* = \lambda \langle X, Y\rangle^*$ for $\lambda \geq 0$ *(pos. homogeneous)*
4) $\langle X_{\boldsymbol{x}} + Y_{\boldsymbol{y}}, Z\rangle^* \leq \langle X, Z\rangle^* + \langle Y, Z\rangle^*$ *(sublinear)*

Proof:

1) Follows from Prop. 6 and the positive definiteness of $\langle\cdot,\cdot\rangle$.
2) Follows from the symmetry of $\langle\cdot,\cdot\rangle$.
3) Let $\lambda \geq 0$. Then
$$\begin{aligned}\langle \lambda X, Y\rangle^* &= \max_{\boldsymbol{x}\in X, \boldsymbol{y}\in Y} \langle \lambda\boldsymbol{x}, \boldsymbol{y}\rangle = \lambda \cdot \max_{\boldsymbol{x}\in X, \boldsymbol{y}\in Y} \langle \boldsymbol{x}, \boldsymbol{y}\rangle \\ &= \lambda \langle X, Y\rangle^*.\end{aligned}$$
4) Let $W = X_{\boldsymbol{x}} + Y_{\boldsymbol{y}}$, and let $(\boldsymbol{w}, \boldsymbol{z}) \in \operatorname{supp}(\langle\cdot,\cdot\rangle^* | W, Z)$. We have $W \subseteq X \oplus Y$. Hence, there are $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$ such that $\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{y}$. Thus,
$$\langle W, Z\rangle = \langle \boldsymbol{w}, \boldsymbol{z}\rangle = \langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z}\rangle = \langle \boldsymbol{x}, \boldsymbol{z}\rangle + \langle \boldsymbol{y}, \boldsymbol{z}\rangle.$$
From $\langle \boldsymbol{x}, \boldsymbol{z}\rangle \leq \langle X, Z\rangle^*$ and $\langle \boldsymbol{y}, \boldsymbol{z}\rangle \leq \langle Y, Z\rangle^*$ follows the assertion.

∎

Any inner product space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$ is a metric space with metric $d(x, y) = \|\boldsymbol{x} - \boldsymbol{y}\|$. The metric d induces a metric D_* on $\mathcal{X}_\mathcal{T}$.

Proposition 8: Let $(\mathcal{X}_\mathcal{T}, \langle\cdot,\cdot\rangle^*)$ be a $\mathcal{T}$-space over the inner product space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$. Then for all X, $Y \in \mathcal{X}_\mathcal{T}$, we have

$$D_*(X, Y)^2 = \|X\|_*^2 - 2\langle X, Y\rangle^* + \|Y\|_*^2.$$

Proof: We have

$$\begin{aligned}D_*(X, Y)^2 &= \min\left\{\|\boldsymbol{x} - \boldsymbol{y}\|^2 \,:\, \boldsymbol{x} \in X, \boldsymbol{y} \in Y\right\} \\ &= \min\left\{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y}\rangle \,:\, \boldsymbol{x} \in X, \boldsymbol{y} \in Y\right\} \\ &= \min\left\{\|\boldsymbol{x}\|^2 - 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \|\boldsymbol{y}\|^2 \,:\, \boldsymbol{x} \in X, \boldsymbol{y} \in Y\right\} \\ &= \|\boldsymbol{x}\|^2 - 2\max\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle \,:\, \boldsymbol{x} \in X, \boldsymbol{y} \in Y\} + \|\boldsymbol{y}\|^2 \\ &= \|X\|_*^2 - 2\langle X, Y\rangle^* + \|Y\|_*^2.\end{aligned}$$

∎

Theorem 7: Let $(\mathcal{X}_\mathcal{T}, \langle\cdot,\cdot\rangle^*)$ be a $\mathcal{T}$-space over the inner product space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$. Then

$$|\langle X, Y\rangle^*| \leq \|X\|_* \, \|Y\|_*.$$

for all $X, Y \in \mathcal{X}_\mathcal{T}$

Proof: Let $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{supp}(\langle\cdot,\cdot\rangle^* | X, Y)$. Applying the conventional Cauchy-Schwarz inequality for vectors and using Eq. (3) yields

$$|\langle X, Y\rangle^*| = |\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\| \, \|\boldsymbol{y}\| = \|X\|_* \, \|Y\|_*.$$

∎

Using Theorem 7, we can show that the angle of structures has a geometrical meaning. For two nonzero structures X and Y, the angle $\theta \in [0, \pi]$ between X and Y is defined (indirectly in terms of its cosine) by

$$\cos\theta = \frac{\langle X, Y\rangle^*}{\|X\|_* \, \|Y\|_*}. \tag{4}$$

Theorem 7 implies that

$$-1 \leq \frac{\langle X, Y\rangle^*}{\|X\|_* \, \|Y\|_*} \leq 1$$

and thus assures that this angle is well-defined. This shows that $\langle X, Y\rangle^*$ has the same geometrical properties as an inner product, although it does not satisfy the algebraic properties of an inner product.

B. Proofs of Results from Section 4

Proof of Proposition 3: Consider the representation mapping $f(\boldsymbol{x})$ of $F(X)$. Let $c = f(\boldsymbol{x}_0)$ for some arbitrary $\boldsymbol{x}_0 \in \mathcal{X}$, and let $\mathcal{U} = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \leq c\}$. Since, F is continuous, the representation mapping f is continuous. Hence, $\mathcal{U}$ is closed. In addition, $\mathcal{U}$ is also bounded. To see this, assume that $\mathcal{U}$ is unbounded. Then there is a sequence $(\boldsymbol{y}_i)_{i\in\mathbb{N}}$ in $\mathcal{U}$ with $\lim_{i\to\infty} \|\boldsymbol{y}_i\| = \infty$. From this follows $\lim_{i\to\infty} f(\boldsymbol{y}_i) = \infty$ as shown below in Lemma 1. But this is a contradiction to $f(\boldsymbol{y}_i) \leq c$ for all $i \in \mathbb{N}$. Hence, $\mathcal{U}$ is closed and bounded. By the Heine-Borel Theorem, $\mathcal{U}$ is compact. Since μ is a continuous mapping and since the continuous image of a compact set is compact, the set $\mathcal{U}_\mathcal{T} = \mu(\mathcal{U})$ is compact. The compact set $\mathcal{U}_\mathcal{T}$ is of the form $\mathcal{U}_\mathcal{T} = \{X \in \mathcal{X}_\mathcal{T} : F(X) \leq c\}$. Since a continuous function attains its minimum on a compact set, F has a minimum on $\mathcal{C}$, which is also a minimum on $\mathcal{X}_\mathcal{T}$ by construction of $\mathcal{U}_\mathcal{T}$.

To complete the proof it remains to show Lemma 1.

Lemma 1: We have $\lim_{\|X\|_*\to\infty} F(X) = \infty$.

Proof: Suppose that $(Y_i)_{i\in\mathbb{N}}$ is a sequence in $\mathcal{X}_\mathcal{T}$ with $\lim_{i\to\infty}\|Y_i\|_* = \infty$. Let

$$R = \max\left\{\|X_i\|_* \,:\, X_i \in \mathcal{D}_\mathcal{T}\right\} \quad \text{and} \quad r = \min\left\{\|X_i\|_* \,:\, X_i \in \mathcal{D}_\mathcal{T}\right\}.$$

From Prop. 8 follows

$$F(X) = \sum_{i=1}^{k} D_*(X, X_i)^2 = \sum_{i=1}^{k} \|X\|_*^2 - 2\langle X, X_i\rangle^* + \|X_i\|_*^2 \tag{5}$$

Exploiting Theorem 7 yields

$$\langle X, X_i\rangle^* = \|X\|_* \|X_i\|_* \cos\alpha_i, \tag{6}$$

where α_i denotes the well-defined angle between X and X_i. Substituting the inner $\mathcal{T}$-products in Equation (5) by the right hand side of Equation (6) yields

$$\begin{aligned} F(X) &= \sum_{i=1}^{k} \|X\|_*^2 - 2\|X\|_* \|X_i\|_* \cos\alpha_i + \|X_i\|_*^2 \\ &\geq \sum_{i=1}^{k} \|X\|_*^2 - 2\|X\|_* R + r^2 \\ &= \sum_{i=1}^{k} \left(\|X\|_* - R\right)^2 + \left(r^2 - R^2\right) \\ &= k\left(\left(\|X\|_* - R\right)^2 + \left(r^2 - R^2\right)\right). \end{aligned}$$

From $\lim_{i\to\infty}\|Y_i\|_* = \infty$ follows

$$\lim_{i\to\infty} F(Y_i) \geq \lim_{i\to\infty} k\left(\left(\|Y_i\|_* - R\right)^2 + \left(r^2 - R^2\right)\right) = \infty.$$

■

Proof of Theorem 3: Let $\mathcal{I} = \{1, \ldots, k\}$ denote the set of indexes. Since M is a sample mean, it is a global minimum of problem (P). Let $\boldsymbol{m}$ be an arbitrarily chosen vector representation of M, and let $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_k) \in \mathcal{D}$ be a matrix representation of $\mathcal{D}_\mathcal{T}$ with $(\boldsymbol{m}, \boldsymbol{x}_i) \in \operatorname{supp} D_*(M, X_i)$ for all $i \in \mathcal{I}$. Consider the function

$$f(\boldsymbol{x}|\boldsymbol{X}) = \sum_{i\in\mathcal{I}} \|\boldsymbol{x} - \boldsymbol{x}_i\|^2,$$

which is smooth, convex and has a unique global minimum of the form

$$\boldsymbol{x}_* = \frac{1}{k}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i.$$

Suppose that $\boldsymbol{m} \neq \boldsymbol{x}_*$. Then we have

$$f(\boldsymbol{x}_*|\boldsymbol{X}) < f(\boldsymbol{m}|\boldsymbol{X}) = F(M), \tag{7}$$

where strict inequality follows from the fact that the global minimum of $f(\boldsymbol{x}|\boldsymbol{X})$ is unique. Let $\mu(\boldsymbol{x}_*) = X_*$. Since M is a global minimum of F by assumption, we find that

$$F(X_*) \geq F(M) > f(\boldsymbol{x}_*|\boldsymbol{X}). \tag{8}$$

From $F(X_*) \neq f(\boldsymbol{x}_*|\boldsymbol{X})$ follows that there is a nonempty subset $\mathcal{J} \subseteq \mathcal{I}$ with $(\boldsymbol{x}_*, \boldsymbol{x}_i) \notin \operatorname{supp} D_*(X_*, X_i)$ for all $i \in \mathcal{J}$. Let $\tilde{\boldsymbol{x}}_i \in X_i$ be vector representations of X_i with $(\boldsymbol{x}_*, \tilde{\boldsymbol{x}}_i) \in \operatorname{supp} D_*(X_*, X_i)$ for all $i \in \mathcal{J}$. From

$$D_*(X_*, X_i) = \|\boldsymbol{x}_* - \tilde{\boldsymbol{x}}_i\| < \|\boldsymbol{x}_* - \boldsymbol{x}_i\|$$

for all $i \in \mathcal{J}$ follows

$$\begin{aligned} F(X_*) &= \sum_{i\in\mathcal{J}} \|\boldsymbol{x}_* - \tilde{\boldsymbol{x}}_i\|^2 + \sum_{i\in\mathcal{I}\setminus\mathcal{J}} \|\boldsymbol{x}_* - \boldsymbol{x}_i\|^2 \\ &< \sum_{i\in\mathcal{I}} \|\boldsymbol{x}_* - \boldsymbol{x}_i\|^2 + \sum_{i\in\mathcal{I}\setminus\mathcal{J}} \|\boldsymbol{x}_* - \boldsymbol{x}_i\|^2 \\ &= f(\boldsymbol{x}_*|\boldsymbol{X}) \end{aligned}$$

Thus, we obtain the inequality

$$F(X_*) < f(\boldsymbol{x}_*|\boldsymbol{X}) \tag{9}$$

Equation (9) is a contradiction to $F(X_*) > f(\boldsymbol{x}_*|\boldsymbol{X})$ in Equation (8). At the same time, combining Equation (9) with Equation (7) shows a second contradiction to our assumption that $F(M)$ is a global minimum. Hence, we find that $\boldsymbol{m} = \boldsymbol{x}_*$. ■

Proof of Theorem 4: We first show $1 \Rightarrow 2$. Let $\boldsymbol{X}$ be a conform matrix representation of $\mathcal{D}_\mathcal{T}$ and let $\boldsymbol{m}$ be the sample mean of the columns of $\boldsymbol{X}$. Since $\boldsymbol{X}$ is conform, the structure $M = \mu(\boldsymbol{m})$ is a sample mean of $\mathcal{D}_\mathcal{T}$ by Theorem 3. According to Lemma 2 shown below, we have

$$F(M) = \sum_{i=1}^{k} \|\boldsymbol{x}_i\|^2 - \frac{1}{k}\Gamma(\boldsymbol{X}).$$

Now suppose that there is a matrix representation $\boldsymbol{X'} = \left(\boldsymbol{x}'_1 \cdots \boldsymbol{x}'_k\right) \in \mathcal{D}$ such that $\Gamma(\boldsymbol{X}) < \Gamma(\boldsymbol{X'})$. Then $f(\boldsymbol{m'}) < f(\boldsymbol{m})$, where $\boldsymbol{m'}$ is the sample mean of the columns of $\boldsymbol{X'}$. With a similar argumentation as in the proof of Theorem 3, we can construct a contradiction to our assumption that M is a global minimum of F.

Next, we show $2 \Rightarrow 1$. Let $\boldsymbol{X} \in \mathcal{D}$ be a matrix representation of $\mathcal{D}_\mathcal{T}$ with maximal Gram sum. To show that $\boldsymbol{X}$ is conform is equivalent to show that the sample mean $\boldsymbol{m}$ of the columns of $\boldsymbol{X}$ represents a sample mean $M = \mu(\boldsymbol{m})$ of $\mathcal{D}_\mathcal{T}$. Suppose that M is not a sample mean. According to Prop. 3 there is a solution M' with $F(M') < F(M)$. From Theorem 3 follows that there is a conform matrix representation $\boldsymbol{X'} = \left(\boldsymbol{x}'_1 \cdots \boldsymbol{x}'_k\right) \in \mathcal{D}$ such that

$$M' = \mu\left(\frac{1}{k}\sum_{i=1}^{k} \boldsymbol{x}'_i\right).$$

From Lemma 2 follows that

$$F(M') = \sum_{i=1}^{k} \left\|\boldsymbol{x}'_i\right\|^2 - \frac{1}{k}\Gamma(\boldsymbol{X'}).$$

Since the length of a structure is independent of its vector representation, we have

$$\sum_{i=1}^{k} \left\|\boldsymbol{x}'_i\right\|^2 = \sum_{i=1}^{k} \|\boldsymbol{x}_i\|^2.$$

Thus, $F(M') < F(M)$ implies $\Gamma(\boldsymbol{X'}) > \Gamma(\boldsymbol{X})$. This is a contradiction to our assumption $\Gamma(\boldsymbol{X}) \geq \Gamma(\boldsymbol{X'})$. Hence, M is a sample mean and therefore $\boldsymbol{X}$ is conform.

To complete the proof of Theorem 4 we show the following Lemma:

Lemma 2: Let $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_k) \in \mathcal{D}$ be a conform matrix representation of a sample $\mathcal{D}_\mathcal{T} = \{X_1, \ldots, X_k\}$. Then

$$F(M) = \sum_{i=1}^{k} \|\boldsymbol{x}_i\|^2 - \frac{1}{k}\Gamma(\boldsymbol{X})$$

where

$$M = \mu\left(\frac{1}{k}\sum_{i=1}^{k} \boldsymbol{x}_i\right),$$

is the sample mean of $\mathcal{D}_\mathcal{T}$ determined by $\boldsymbol{X}$.

Proof: Let $\boldsymbol{m}$ be the sample mean of the columns of $\boldsymbol{X}$. Since $\boldsymbol{X}$ is conform, the structure $M = \mu(\boldsymbol{m})$ is a sample mean of $\mathcal{D}_\mathcal{T}$. We have

$$\begin{aligned} F(M) &= \sum_{i=1}^{k} \|\boldsymbol{m} - \boldsymbol{x}_i\|^2 = \sum_{i=1}^{k} \|\boldsymbol{m}\|^2 - 2\langle \boldsymbol{m}, \boldsymbol{x}_i\rangle + \|\boldsymbol{x}_i\|^2 \\ &= \sum_{i=1}^{k}\left\{\left\|\frac{1}{k}\sum_{j=1}^{k}\boldsymbol{x}_j\right\|^2 - 2\left\langle\frac{1}{k}\sum_{j=1}^{k}\boldsymbol{x}_j, \boldsymbol{x}_i\right\rangle + \|\boldsymbol{x}_i\|^2\right\} \\ &= \sum_{i=1}^{k}\left\{\frac{1}{k^2}\sum_{r=1}^{k}\sum_{s=1}^{k}\langle\boldsymbol{x}_r, \boldsymbol{x}_s\rangle - \frac{2}{k}\sum_{j=1}^{k}\langle\boldsymbol{x}_j, \boldsymbol{x}_i\rangle + \|\boldsymbol{x}_i\|^2\right\} \\ &= \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle - \frac{2}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle + \sum_{i=1}^{k}\|\boldsymbol{x}_i\|^2 \\ &= \sum_{i=1}^{k}\|\boldsymbol{x}_i\|^2 - \underbrace{\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle}_{=\Gamma(\boldsymbol{X})}. \end{aligned}$$

■