The natural pseudo-distance in topological data analysis

Patrizio Frosini

Department of Mathematics and ARCES, University of Bologna
patrizio.frosini@unibo.it

International Workshop and Conference on Topology & Applications, Rajagiri School of Engineering & Technology, Kochi, Kerala, INDIA, December 5-11, 2018

Outline

The definition of d_G

Theoretical results about d_G

Optimal homeomorphisms

A link between d_G and persistent homology

Group equivariant non-expansive operators

The definition of d_G

Theoretical results about d_G

Optimal homeomorphisms

A link between d_G and persistent homology

Group equivariant non-expansive operators

The definition of d_G

Let X and G be a topological space and a subgroup of the group $\mathrm{Homeo}(X)$ of all homeomorphisms from X to X, respectively. If φ_1, φ_2 are two continuous and bounded functions from X to $\mathbb{R}$ we can consider the value $\inf_{g \in G} \|\varphi_1 - \varphi_2 \circ g\|_\infty$. This value is called the *natural pseudo-distance* $d_G(\varphi_1, \varphi_2)$ between φ_1 and φ_2 with respect to the group G.

We endow both $C^0(X, \mathbb{R})$ and G with the topology of uniform convergence, so that G becomes a topological group acting continuously on $C^0(X, \mathbb{R})$ by composition on the right. We observe that the action of G on $C^0(X, \mathbb{R})$ is continuous.

The definition of d_G

If G is the trivial group Id, then d_G is the max-norm distance $\|\varphi_1 - \varphi_2\|_\infty$. Moreover, if G_1 and G_2 are subgroups of $\mathrm{Homeo}(X)$ and $G_1 \subseteq G_2$, then

$$d_{\mathrm{Homeo}(X)}(\varphi_1, \varphi_2) \leq d_{G_2}(\varphi_1, \varphi_2) \leq d_{G_1}(\varphi_1, \varphi_2) \leq \|\varphi_1 - \varphi_2\|_\infty$$

for every $\varphi_1, \varphi_2 \in C^0(X, \mathbb{R})$.

We usually restrict d_G to $\Phi \times \Phi$, where Φ is a bounded subset of $C^0(X, \mathbb{R})$.

Our ground truth: the natural pseudo-distance d_G

The natural pseudo-distance d_G is our ground truth: it describes the differences that the observer can perceive between the measurements in Φ with respect to the equivalence expressed by the group G.

A possible objection: *"The use of the concept of homeomorphism makes the natural pseudo-distance d_G difficult to apply. For example, in shape comparison two similar objects can be non-homeomorphic, hence this pseudo-metric cannot be applied to real problems."*

A possible objection

Answer: **the homeomorphisms do not concern the "objects" but the space X where the measurements are made**.

- For example, if we are interested in grey level images, the domain of our measurements can be modelled as the real plane and each image can be represented as a function from $\mathbb{R}^2$ to $\mathbb{R}$. Therefore, the space X is not given by the (possibly non-homeomorphic) objects displayed in the pictures, but by the topological space $\mathbb{R}^2$.
- If we make two CAT scans, the topological space X is always given by an helix turning many times around a body, and no requirement is made about the topology of such a body.

In other words, it is usually legitimate to assume that the topological space X is determined only by the measuring instrument we are using to get our measurements.

The definition of d_G

Theoretical results about d_G

Optimal homeomorphisms

A link between d_G and persistent homology

Group equivariant non-expansive operators

d_G and critical values: manifolds

When the filtering functions are defined on a regular closed manifold, some results restrict the range of values that can be taken by the natural pseudo-distance d_G.

Theorem

Assume that $\mathscr{M}$ is a closed manifold of class C^1 and that $\varphi_1, \varphi_2 : \mathscr{M} \to \mathbb{R}$ are two functions of class C^1. Set $d := d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$. Then a positive integer k exists for which one of the following properties holds:

(i) *k is odd and kd is the distance between a critical value of φ_1 and a critical value of φ_2;*

(ii) *k is even and kd is either the distance between two critical values of φ_1 or the distance between two critical values of φ_2.*

d_G and critical values: surfaces

Theorem

Assume that $\mathscr{S}$ is a closed surface of class C^1 and that $\varphi_1, \varphi_2 : \mathscr{S} \to \mathbb{R}$ are two functions of class C^1. Set $d := d_{\mathrm{Homeo}(\mathscr{S})}(\varphi_1, \varphi_2)$. Then a positive integer k exists for which at least one of the following properties holds:

(i) *d is the distance between a critical value of φ_1 and a critical value of φ_2;*

(ii) *d is half the distance between two critical values of φ_1.*

(iii) *d is half the distance between two critical values of φ_2.*

(iv) *d is one third of the distance between a critical value of φ_1 and a critical value of φ_2.*

d_G and critical values: curves

Theorem

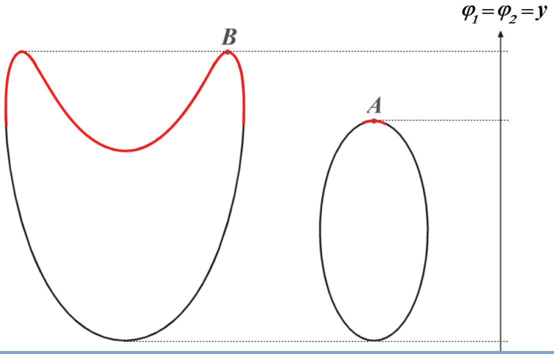
Assume that $\mathscr{C}$ is a closed curve of class C^1 and that $\varphi_1, \varphi_2 : \mathscr{C} \to \mathbb{R}$ are two functions of class C^1. Set $d := d_{\mathrm{Homeo}(\mathscr{C})}(\varphi_1, \varphi_2)$. Then a positive integer k exists for which at least one of the following properties holds:

a) *d is the distance between a critical value of φ_1 and a critical value of φ_2;*
b) *d is half the distance between two critical values of φ_1.*
c) *d is half the distance between two critical values of φ_2.*

The last theorem is sharp, as shown by the following examples.

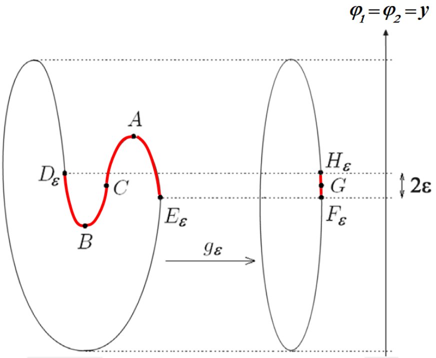
d_G and critical values: curves

Let us consider the two embeddings of S^1 in $\mathbb{R}^2$ represented in the following figure. The ordinate y defines two filtering functions φ_1, φ_2 on S^1. In this case $d_{\mathrm{Homeo}(S^1)}(\varphi_1, \varphi_2) = |\varphi_1(A) - \varphi(B)|$ is the distance between a critical value of φ_1 and a critical value of φ_2.



d_G and critical values: curves

Let us consider the two embeddings of S^1 in $\mathbb{R}^2$ represented in the following figure. The ordinate y defines two filtering functions φ_1, φ_2 on S^1. In this case $d_{\mathrm{Homeo}(S^1)}(\varphi_1, \varphi_2) = \frac{1}{2}|\varphi_1(A) - \varphi(B)|$ is half the distance between two critical values of φ_1.



A result concerning $d_{S^1}(\varphi_1, \varphi_2)$

The research concerning the case that G is a proper subgroup of $\mathrm{Homeo}(\mathscr{M})$ is still at its very beginning. As an example of the results concerning this line of research we cite the following theorem, concerning the Lie group S^1.

Theorem

Let φ_1, φ_2 be Morse functions from the Lie group S^1 to $\mathbb{R}$ and set $d = d_{S^1}(\varphi_1, \varphi_2)$. At least one of the following statements holds:

1) *There exist a critical point θ_1 for φ_1 and a critical point θ_2 for φ_2 such that $d = |\varphi_1(\theta_1) - \varphi_2(\theta_2)|$;*

($\longrightarrow$...)

A result concerning $d_{S^1}(\varphi_1, \varphi_2)$

Theorem (continued)

2) *There exist* θ_1, θ_2, $\tilde{\theta}_1$ *and* $\tilde{\theta}_2$ *such that* $d = |\varphi_1(\theta_1) - \varphi_2(\theta_2)| = |\varphi_1(\tilde{\theta}_1) - \varphi_2(\tilde{\theta}_2)|$ *with*

$$\begin{cases} \frac{d\varphi_1}{d\theta}(\theta_1) = \frac{d\varphi_2}{d\theta}(\theta_2) \text{ and } \frac{d\varphi_1}{d\theta}(\tilde{\theta}_1) = \frac{d\varphi_2}{d\theta}(\tilde{\theta}_2) \\ \theta_1 - \theta_2 = \tilde{\theta}_1 - \tilde{\theta}_2 \\ \frac{d\varphi_1}{d\theta}(\theta_1)\frac{d\varphi_1}{d\theta}(\tilde{\theta}_1) < 0 \end{cases}$$

if $\quad (\varphi_1(\theta_1) - \varphi_2(\theta_2)) \cdot (\varphi_1(\tilde{\theta}_1) - \varphi_2(\tilde{\theta}_2)) > 0,$

or

$(\longrightarrow \ldots)$

A result concerning $d_{S^1}(\varphi_1, \varphi_2)$

Theorem (continued)

$$\begin{cases} \frac{d\varphi_1}{d\theta}(\theta_1) = \frac{d\varphi_2}{d\theta}(\theta_2) \text{ and } \frac{d\varphi_1}{d\theta}(\tilde{\theta}_1) = \frac{d\varphi_2}{d\theta}(\tilde{\theta}_2) \\ \theta_1 - \theta_2 = \tilde{\theta}_1 - \tilde{\theta}_2 \\ \frac{d\varphi_1}{d\theta}(\theta_1)\frac{d\varphi_1}{d\theta}(\tilde{\theta}_1) > 0 \end{cases}$$

if $\quad (\varphi_1(\theta_1) - \varphi_2(\theta_2)) \cdot (\varphi_1(\tilde{\theta}_1) - \varphi_2(\tilde{\theta}_2)) < 0.$

Optimal homeomorphisms

Assume that X is a compact topological space and $\varphi_1, \varphi_2 : X \to \mathbb{R}$ are continuous functions. Let G be a subgroup of $\mathrm{Homeo}(X)$. We say that a homeomorphism $g \in G$ is *optimal* in G for (φ_1, φ_2) if $\|\varphi_1 - \varphi_2 \circ g\|_\infty = d_G(\varphi_1, \varphi_2)$. The following results hold for optimal homeomorphisms.

Theorem

Assume that $\mathscr{M}$ is a C^1 closed manifold and that $\varphi_1, \varphi_2 : \mathscr{M} \to \mathbb{R}$ are of class C^1. If an optimal homeomorphism $g \in \mathrm{Homeo}(\mathscr{M})$ for (φ_1, φ_2) exists, then $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ is the distance between a critical value of φ_1 and a critical value of φ_2.

Optimal homeomorphisms

Theorem

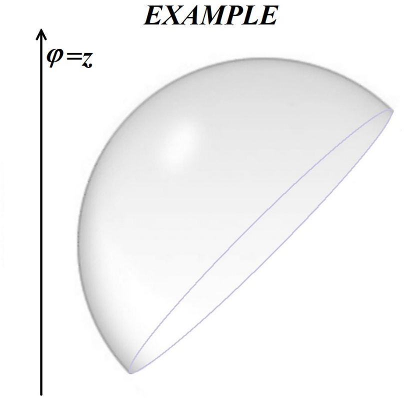
If $\varphi_1, \varphi_2 : S^1 \to \mathbb{R}$ are Morse functions and $d_{\mathrm{Homeo}(S^1)}(\varphi_1, \varphi_2)$ vanishes, then an optimal C^2-diffeomorphism exists in $\mathrm{Homeo}(S^1)$ for (φ_1, φ_2).

Theorem

The number of optimal homeomorphisms in the Lie group S^1 for a pair (φ_1, φ_2) of Morse functions from S^1 to $\mathbb{R}$ is finite.

What is persistent homology?

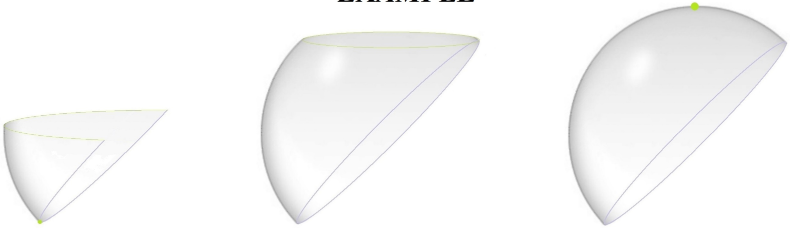
If $\varphi : X \to \mathbb{R}$ is a continuous function, we can consider the sublevel sets $X_t := \{x \in X : \varphi(x) \leq t\}$. When t varies we see the birth and death of k-dimensional holes.



What is persistent homology?

If $\varphi : X \to \mathbb{R}$ is a continuous function, we can consider the sublevel sets $X_t := \{x \in X : \varphi(x) \leq t\}$. When t varies we see the birth and death of k-dimensional holes.

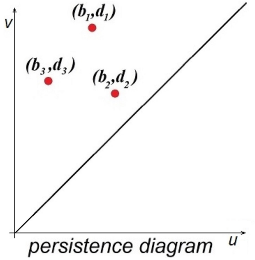
EXAMPLE



No 1-dimensional hole — *Birth of a 1-dimensional hole* — *Death of the 1-dimensional hole*

What is persistent homology?

In plain words, the persistence diagram in degree k of φ is the collection of the pairs (b_i, d_i) where b_i and d_i are the times of birth and death of the i-th hole of dimension k.



The points of the persistence diagram are endowed with multiplicity. Each point of the diagonal $u = v$ is assumed to be a point of the persistence diagram, endowed with infinite multiplicity.

What are persistent Betti numbers functions?

Persistence diagrams are not quite suitable for statistical purposes, because no good definition of average of persistence diagrams exists.

Persistent Betti numbers functions are more suitable for statistics.

Definition

The k-th persistent Betti numbers function $\beta_k(u, v)$ is the number of holes of dimension k whose time of birth is smaller than u and whose time of death is greater than v.

What are persistent Betti numbers functions?

Formally:

Definition

Let $\varphi : X \to \mathbb{R}$ be a continuous function. If $u, v \in \mathbb{R}$ and $u < v$, we can consider the inclusion i of X_u into X_v. Such an inclusion induces a homomorphism $i^* : H_k(X_u) \to H_k(X_v)$ between the homology groups of X_u and X_v in degree k. The group $PH_k^{\varphi}(u, v) := i^*(H_k(X_u))$ is called the *k-th persistent homology group with respect to the function $\varphi : X \to \mathbb{R}$, computed at the point* (u, v). The rank $r_k(\varphi)(u, v)$ of this group is said *the k-th persistent Betti numbers function with respect to the function $\varphi : X \to \mathbb{R}$, computed at the point* (u, v).

The average of persistent Betti numbers functions can be trivially defined as the usual average of real-valued functions.

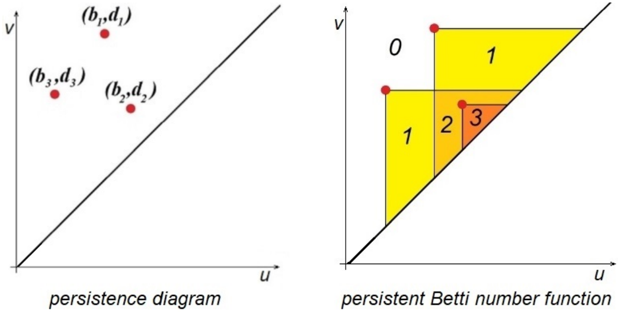
What are persistent Betti numbers functions?

The use of averages of persistent Betti numbers functions in degree 0 firstly appeared in the papers

- Pietro Donatini, Patrizio Frosini, Alberto Lovato, *Size functions for signature recognition*, Proceedings of SPIE, Vision Geometry VII, vol. 3454 (1998), 178183.
- Massimo Ferri, Patrizio Frosini, Alberto Lovato, Chiara Zambelli, *Point selection: A new comparison scheme for size functions (With an application to monogram recognition)*, Proceedings Third Asian Conference on Computer Vision, Lecture Notes in Computer Science 1351, vol. I, R. Chin, T. Pong (editors) Springer-Verlag, Berlin Heidelberg (1998), 329337.

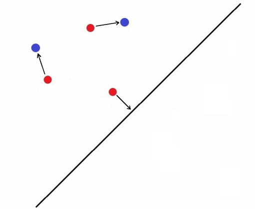
In these papers each point of the considered persistence diagram is replaced with a suitable function (usually a Gaussian function centered at that point).

What are persistent Betti numbers functions?



If we use Čech homology, persistence diagrams are equivalent to persistent Betti numbers functions.

Comparison of persistent Betti numbers functions



A way to change D_1 into D_2.

Persistence diagrams (and hence persistent Betti numbers functions) can be compared by means of the bottleneck distance. The bottleneck distance between two persistence diagrams D_1, D_2 is the minimum cost of changing the points of D_1 into the points of D_2, where the cost of moving each point is given by the max-norm distance in $\mathbb{R}^2$. Moving a point to the diagonal is equivalent to delete it.

Comparison of persistent Betti numbers functions

An important property of the metric d_{match} is its stability, as stated in the following result.

Theorem

If k is a natural number and $\varphi_1, \varphi_2 \in C^0(X, \mathbb{R})$, then

$$d_{\mathrm{match}}(r_k(\varphi_1), r_k(\varphi_2)) \leq d_{\mathrm{Homeo}(X)}(\varphi_1, \varphi_2) \leq \|\varphi_1 - \varphi_2\|_\infty .$$

Group Equivariant Non-Expansive Operators

Let X and G be a topological space and a subgroup of the group $\mathrm{Homeo}(X)$ of all homeomorphisms from X to X, respectively. Let $\Phi \subseteq C^0(X,\mathbb{R})$. We now consider the set $\mathscr{F}(\Phi, G)$ of all maps from Φ to Φ that verify the following two properties:

1. $F(\varphi \circ g) = F(\varphi) \circ g$ for every $\varphi \in \Phi$ and every $g \in G$ (i.e. F is equivariant with respect to G);
2. $\|F(\varphi_1) - F(\varphi_2)\|_\infty \leq \|\varphi_1 - \varphi_2\|_\infty$ for every $\varphi_1, \varphi_2 \in \Phi$ (i.e. F is non-expansive).

Obviously, $\mathscr{F}(\Phi, G)$ is not empty, since it contains at least the identity map. The maps in $\mathscr{F}(\Phi, G)$ will be called *Group Equivariant Non-Expansive Operators* (GENEOs).
In my next talk I will give an extension of this concept to operators from Φ to $\Psi \neq \Phi$.

Lower bounds for d_G via persistent homology

For every fixed k, we can consider the following pseudo-metric $D_{\mathrm{match}}^{\mathscr{F},k}$ on Φ:

$$D_{\mathrm{match}}^{\mathscr{F},k}(\varphi_1,\varphi_2) := \sup_{F\in\mathscr{F}} d_{\mathrm{match}}(r_k(F(\varphi_1)), r_k(F(\varphi_2)))$$

for every $\varphi_1, \varphi_2 \in \Phi$, where $r_k(\varphi)$ denotes the k-th persistent Betti numbers function with respect to the function $\varphi : X \to \mathbb{R}$. We will usually omit the index k, when its value is clear from the context or not influential.

We observe that
$D_{\mathrm{match}}^{\mathscr{F}}(\varphi_1, \varphi_2 \circ g) = D_{\mathrm{match}}^{\mathscr{F}}(\varphi_1 \circ g, \varphi_2) = D_{\mathrm{match}}^{\mathscr{F}}(\varphi_1, \varphi_2)$ for every $\varphi_1, \varphi_2 \in \Phi$ and every $g \in \mathrm{Homeo}(X)$.

Lower bounds for d_G via persistent homology

The importance of $D^{\mathscr{F}}_{\mathrm{match}}$ lies in the following two results, showing that it can be used to get information about the natural pseudo-distance d_G.

Theorem

If $\emptyset \neq \bar{\mathscr{F}} \subseteq \mathscr{F}(\Phi, G)$, *then* $D^{\bar{\mathscr{F}}}_{\mathrm{match}} \leq d_G$.

Theorem

$D^{\mathscr{F}(\Phi,G)}_{\mathrm{match}} = d_G$.

As a consequence, the topological and geometrical study of $\mathscr{F}(\Phi, G)$ is important in the research concerning the natural pseudo-distance.

Two relevant properties of $\mathscr{F}(\Phi, G)$

Two relevant properties of $\mathscr{F}(\Phi, G)$ are expressed by the following result.

Theorem

If Φ is compact, then $\mathscr{F}(\Phi, G)$ is compact.
If Φ is convex, then $\mathscr{F}(\Phi, G)$ is convex.

The compactness and convexity of $\mathscr{F}(\Phi, G)$ are important from the computational point of view.

An open problem

Let us consider a closed C^1 surface $\mathscr{M}$ and two C^1 filtering functions $\varphi_1, \varphi_2 : \mathscr{M} \to \mathbb{R}$. Let $\mathrm{Homeo}(\mathscr{M})$ be the group of all self-homeomorphisms of $\mathscr{M}$. It has been proved that at least one of the following statements holds:

1. $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ is the distance between a critical value of φ_1 and a critical value of φ_2;
2. $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ is half the distance between two critical values of φ_1;
3. $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ is half the distance between two critical values of φ_2;
4. $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ is one third of the distance between a critical value of φ_1 and a critical value of φ_2.

An open problem

Interestingly, no example of two functions $\varphi_1, \varphi_2 : \mathscr{M} \to \mathbb{R}$ is known, such that (4) holds but (1), (2), (3) do not hold. A natural question arises: Can we find an example of two such functions or prove that such an example cannot exist (so improving our result)?

An open problem

We recall that the usual technique to compute the natural pseudo-distance consists in

- finding a lower bound for $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ by computing the matching distance $d_{\mathrm{match}}(\mathrm{Dgm}_k(\varphi_1), \mathrm{Dgm}_k(\varphi_2))$ between the persistence diagrams in degree k of the functions φ_1 and φ_2;
- looking for a sequence (g_i) in $\mathrm{Homeo}(\mathscr{M})$, such that $\lim_{i\to\infty} \|\varphi_1 - \varphi_2 \circ g_i\|_\infty = d_{\mathrm{match}}(\mathrm{Dgm}_k(\varphi_1), \mathrm{Dgm}_k(\varphi_2))$.

If such a sequence (g_i) exists, then the value $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ is equal to $d_{\mathrm{match}}(\mathrm{Dgm}_k(\varphi_1), \mathrm{Dgm}_k(\varphi_2))$.

An open problem

Unfortunately, at least one of the following statements holds:

a) $d_{\mathrm{match}}(\mathrm{Dgm}_k(\varphi_1), \mathrm{Dgm}_k(\varphi_2))$ is the distance between a critical value of φ_1 and a critical value of φ_2;

b) $d_{\mathrm{match}}(\mathrm{Dgm}_k(\varphi_1), \mathrm{Dgm}_k(\varphi_2))$ is half the distance between two critical values of φ_1;

c) $d_{\mathrm{match}}(\mathrm{Dgm}_k(\varphi_1), \mathrm{Dgm}_k(\varphi_2))$ is half the distance between two critical values of φ_2.

Therefore, if (1),(2),(3) do not hold for $\varphi_1, \varphi_2 : \mathscr{M} \to \mathbb{R}$, then $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ cannot be equal to $d_{\mathrm{match}}(\mathrm{Dgm}_k(\varphi_1), \mathrm{Dgm}_k(\varphi_2))$. This means that if there exist two C^1 functions $\varphi_1, \varphi_2 : \mathscr{M} \to \mathbb{R}$ verifying (4) but not (1),(2),(3), then we need new methods to compute $d_{\mathrm{Homeo}(\mathscr{M})}(\varphi_1, \varphi_2)$ and to recognize the pair (φ_1, φ_2) as the right example. As a consequence, the answer to the question asked in this section is still unknown.


Thanks for your attention!
COCHIN INTERNATIONAL AIRPORT