**CGL**
Computational Geometric Learning

## Final Research Workshop - Third year

# Booklet of Abstracts

September 30 to October 2, 2013

National and Kapodistrian
UNIVERSITY OF ATHENS

# 1 General

The final CGLearning research workshop took place from Monday, September 30 (morning) to Wednesday, October 2 (afternoon), at Mare Nostrum Hotel, Vravrona (http://www.mare-nostrum.gr/), on the east of Athens, Greece. The Workshop was open to external participants. It featured tutorials from 3 tutorial speakers (the talks are on the project's webpage), and contributed talks by members of all CGL teams.

Local organizers: Ioannis Emiris and Vissarion Fisikopoulos. We acknowledge the useful help of Anna Karasoulou and Ioannis Psarros.

# 2 Schedule

- Sunday 29.09.2013: arrival.
- Monday 30.09.2013 (morning) – Wednesday 02.10.2013 (early afternoon): Main event, room "Artemis" (see detailed program below).
- Tuesday evening: Social dinner at the restaurant of the Acropolis Museum.
- Sunday 29.09.2013 (7 pm) and Wednesday 02.10.2013 (3 pm): Informal meetings of the CGL board.

| MONDAY | TUESDAY | WEDNESDAY |
|---|---|---|
| 9.30 Cecilia Clementi - Understanding the configurational space of macromolecules by means of Locally Scaled Diffusion Map | 9.30 Cecilia Clementi - Exploring high dimensional configurational spaces by Diffusion Map driven Molecular Dynamics | 9.30 Dimitrios Gunopulos - Similarity in Temporal, Spatio-Temporal and High-Dimensional Databases |
| 11.00 COFFEE | 11.00 COFFEE | 11.00 COFFEE |
| 11.30 Guillaume Obozinski - Machine learning and geometry: selected topics (tutorial 1) | 11.30 Guillaume Obozinski - Machine learning and geometry: selected topics (tutorial 2) | 11.30 Rien van de Weijgaert (RUG) - Gaussian field homology |
| | | 11.50 Sebastian Stich (ETH) - Optimization and Learning with Random Pursuit |
| | | 12.10 Vissarion Fisikopoulos (NKUA) - Vertex enumeration for polytopes defined by oracles |
| | | 12.30 Ebrahim Ehsanfar (TUD) - Fast Clustering Techniques for Range Queries and Probabilistic Data |
| | | 12.50 Clément Maria (INRIA) - The Compressed Annotation Matrix: an Efficient Data Structure for Computing Persistent Cohomology |
| 13.00 LUNCH | 13.00 LUNCH | 13.15 LUNCH |
| 15.00 Dimitrios Gunopoulos - An Introduction to Data Mining Techniques, focusing on Unsupervised and Semi-supervised learning | 15.00 Steve Oudot (INRIA) - Zigzag Zoology: Rips Zigzags for Homology Inference | CGL Board |
| | 15.20 Marc Glisse (INRIA) - Homological Reconstruction and simplification in $\mathbb{R}^3$ | |
| | 15.40 Dror Atariah (FU Berlin) - Improving Optimal Triangulation of Saddles | |
| | 16.00 Oren Salzman - Doron Sharahabani (TAU) - Sparsification of Motion-Planning Roadmaps by Edge Contraction | |
| 16.30 COFFEE | | |
| 17.00 David Cohen-Steiner (INRIA) - Improved bounds on higher eigenvalues of graphs | 17:00 Bus Trip to the Acropolis | |
| 17.20 Lars Kühne (FSU) - Sclow Plots: visualizing empty space | | |
| 17.40 Mathijs Wintraecken - Ramsay Dyer (RUG) - Intrinsic simplices on Riemannian manifold | 18:00 Dinner at the Acropolis Museum | |
| 18.00 Alix Lheritier (INRIA) - A High-dimensional Non-parametric Two-sample Test based on Bayesian Mixtures over Spatial Partitions | | |
| | 20:30 Free time | |
| | 21.30 Return | |

## 3 List of participants

| Name | Affiliation |
|---|---|
| Dror Atariah | FU Berlin |
| Jean-Daniel Boissonnat | INRIA |
| Frédéric Cazals | INRIA |
| Frédéric Chazal | INRIA |
| Cecilia Clementi | Rice U. |
| David Cohen-Steiner | INRIA |
| Anne Driemel | TUD |
| Ramsay Dyer | RUG |
| Ebrahim Ehsanfar | TUD |
| Ioannis Emiris | NKUA |
| Vissarion Fisikopoulos | NKUA |
| Bernd Gaertner | ETH |
| Panos Giannopoulos | FU Berlin |
| Joachim Giesen | FSU |
| Marc Glisse | INRIA |
| Dimitrios Gunopulos | NKUA |
| Dan Halperin | TAU |
| Anna Karasoulou | NKUA |
| Christos Konaxis | NKUA |
| Lars Kühne | FSU |
| Konstantinos Lentzos | NKUA |
| Alix Lheritier | INRIA |
| Clément Maria | INRIA |
| Guillaume Obozinski | Ecole des Ponts |
| Steve Oudot | INRIA |
| Ioannis Psarros | NKUA |
| Günter Rote | FU Berlin |
| Oren Salzman | TAU |
| Doron Sharahabani | TAU |
| Christian Sohler | TUD |
| Sebastian Stich | ETH |
| Raimundas Vidunas | NKUA |
| Rien van de Weijgaert | RUG |
| Mathijs Wintraecken | RUG |

## 4 Map

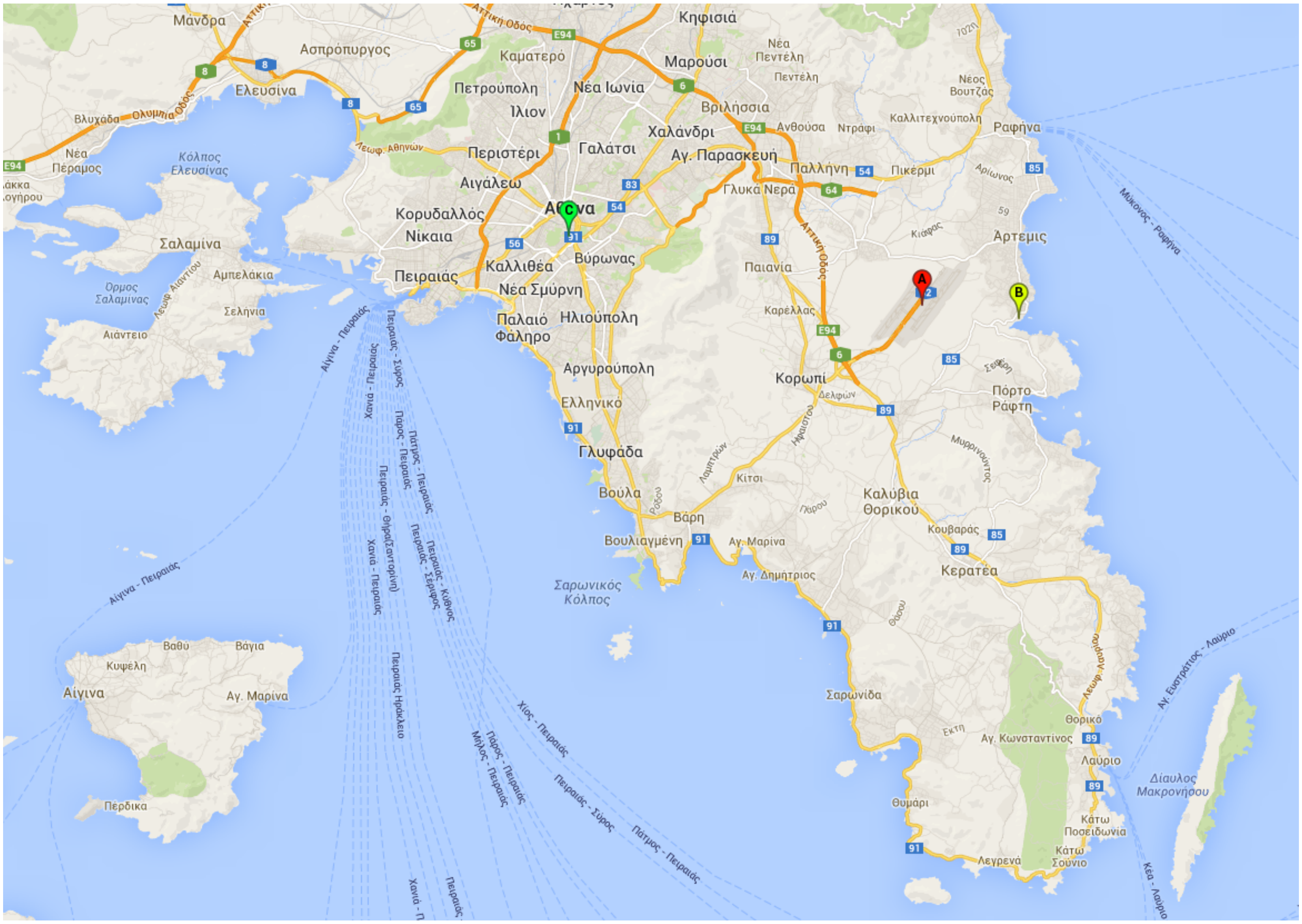

Figure 1: (A) Athens International Airport ”Eleftherios Venizelos”, (B) Mare Nostrum Hotel Club, (C) Acropolis Museum (Social Dinner)

# 5 Abstracts of tutorials

## Diffusion Map as a tool to characterize complex configurational spaces.

Cecilia Clementi
Rice University, Houston, Texas, USA

### 1. Understanding the configurational space of macromolecules by means of Locally Scaled Diffusion Map

The understanding of emerging collective behaviors in biomolecular complexes represents a major challenge in modern biophysics. Several groups have recently begun to adapt machine learning methods to analyze biomolecular systems. Many of these methods can be seen as dimensionality reduction algorithms. That is, they take as input data from a high-dimensional space (such as molecular configuration space) and output set of coordinates in a much lower dimensional space (such as a few collective variables). These methods rely on the assumption that physically relevant molecular configurations exist on a low-dimensional manifold that is embedded in a much higher dimensional space. Most of these methods preserve a metric of some sort.

In this tutorial I will first introduce the challenges in the analysis of the configurational space of macromolecules, then the machine learning approaches that have been proposed to address this problem. I will present both linear and non-linear methods, discussing the advantages and limitations of each approach.

The last part of the tutorial will discuss Diffusion Map and Locally Scaled Diffusion Map (LSDMap), a new development that combines multi-resolution nonlinear dimensionality reduction and diffusion analysis. The application of LSDMap produces reliable low-dimensional representations and models for the dynamics of apparently high-dimensional complex systems such as proteins in a biological environment.

### 2. Exploring high dimensional configurational spaces by Diffusion Map driven Molecular Dynamics.

In the last two decades, a number of methods have been developed to enhance the sampling of rare events in high dimensional dynamical systems.

One class of such techniques biases the dynamics according to one or a few collective variables and then unbiases the results to obtain the equilibrium distribution as a function of the collective variables used. However, oftentimes the definition of collective variables in complex macromolecular systems is *per se* a non trivial problem (as discussed in the first tutorial). The results obtained with these techniques depend on the collective variables used as input and the results are as meaningful as the collective variable chosen.

We will first review the most popular approaches that have been proposed to accelerate the sampling of the relevant part of the configurational space in high dimensional systems. We will then discuss a new approach for the sampling of rare events: Diffusion Map-directed Molecular Dynamics (DM-d-MD). This method builds on a previously developed dimensionality reduction technique, LSDMap. The collective variables from LSDMap, the diffusion coordinates (DC), capture the slowest collective motions of the system. The DM-d-MD algorithm uses local DCs to direct the dynamics. By periodically calculating DCs on the fly and restarting the dynamics from the boundary along the first DC, the system is more likely to visit new regions of the configuration space instead of being trapped in local energy minima.

# Data mining, and similarity in high-dimensional Databases.

Dimitrios Gunopulos
University of Athens, Greece

## 1. An Introduction to Data Mining Techniques, focusing on Unsupervised and Semi-supervised learning.

We give an overview of Data Mining Techniques from Clustering, Semi-supervised learning, Classification, and Pattern Mining, focusing on Clustering (Unsupervised Learning) and Semi-Supervised Learning. The goal of clustering is to provide useful information by organizing data into groups (clusters) such that data in a cluster are more similar to each other than are data belonging to different clusters. The study of clustering is only unified at this very general description level; there are many solutions proposed, based on diverse underlying principles and assumptions often leading to different results. In this tutorial we also focus on approaches that address the clustering problem by exploiting the users input while the clustering results are evaluated in terms of the users requirement satisfaction.

## 2. Similarity in Temporal, Spatio-Temporal and High-Dimensional Databases.

The aim of this tutorial is to give an overview of the various challenges encountered when assessing the similarity of Temporal, Spatio-Temporal and generally High-Dimensional data and to present the corresponding databases solution techniques. The tutorial presents an overview of similarity-related problems and solutions in temporal databases, followed by a thorough exposition of the various aspects of similarity in spatio-temporal settings. Motivation for considering similarity queries in such settings comes from the many novel applications that deal with mobile objects or participants, and we describe several such applications in the last part of the tutorial.

# Machine learning and geometry: selected topics.

GUILLAUME OBOZINSKI
Ecole des Ponts

## Tutorial 1.

Sparse methods, in statistics and machine learning, and compressed sensing, in signal processing, form today a well established body of techniques and theoretical ideas that have many concrete applications in different fields. Among others one can cite: more efficient acquisition of MRI signals, automated segmentation of hyperspectral images, denoising, inpainting, deblurring as well as other automated processing of natural images.

This tutorial will first introduce the problem of estimation of a sparse signals or sparse models. We will review the main formulations and algorithms used in the field. After discussing the specificities of compressed sensing, I will talk about theoretical garanties both for sparse methods in general and for compressed sensing.

The last part of the tutorial will focus on an elegant geometric perspective on compressed sensing, which allows among others to extend the existing theory to other more exotic forms of sparsity.

## Tutorial 2.

The goal of supervised learning is to learn a function $f$ mapping an input $x$ to an output $y$ from a *training set* consisting of pairs $(x_i, y_i)$ of input data $x_i$ together with a *label* $y_i$ to predict. In unsupervised learning, the training set consists of only input data with no labels; in semi-supervised learning only a fraction of the data is labelled.

In theory the unlabeled data should provide some information about the density of the input data. Unfortunately, in the typical problems considered in machine learning and in high-dimensional statistics, density estimation under general hypotheses is impossible, due to the curse of dimensionality. The geometric structure of the data encoded as a graph Laplacian is however very useful, because the support of the data distribution is very often essentially low-dimensional, and reveals some structure of the density.

This presentation will cover formulations for unsupervised and supervised learning that are based on the Laplacian, namely spectral clustering and other methods based on graph cuts, and for the semi-supervised setting Laplacian regularization.

# 6 Abstracts of contributed talks

## Improving Optimal Triangulation of Saddles

Dror Atariah
Freie Universität Berlin, Germany

Last year we presented an optimal (local) interpolating triangulation of saddle surfaces. That is, the triangulation comprised of triangles with vertices lying on the surface. It turns out, that by allowing a non interpolating triangles the approximation can be improve and attain an even better error. In this talk we shall discuss the some aspects of this improvement.

## Improved bounds on higher eigenvalues of graphs

David Cohen-Steiner
INRIA Sophia-Antipolis Méditerranée

The eigenvalues of a graph are intimately related to the existence of good partitions of the graph. We will show that recent eigenvalue bounds for Riemannian surfaces lead to improved bounds for graphs as a function of their genus. Our approach is based on a variant of Burger's comparison principle which is of independant interest.

joint work with Omid Amini (ENS).

## Fast Clustering Techniques for Range Queries and Probabilistic Data

Ebrahim Ehsanfar
Technische Universität Dortmund, Germany

This talk is about three different works on clustering problem. I will explain the general ideas that lead the fast orthogonal range clustering (k-median and k-means problem) using coreset constructions and proper data-structures. The second result is on implementation of a new heuristic for k-mediean problem for probabilistic data. I also investigate 1-center problem (i.e. the minimum enclosing ball problem) in a distributed setting as well as in a centralized setting for uncertain data.

## Vertex enumeration for polytopes defined by oracles

Vissarion Fisikopoulos
University of Athens, Greece

In general dimension, there is no known total polynomial algorithm for either convex hull or vertex enumeration, i.e. whose complexity depends polynomially in the input and output sizes. It is thus important to identify polytope constructions for which total polynomial-time algorithms can be obtained. We study classes of polytopes which are given implicitly by an LP (or separation) oracle. For the special case where we are also given a superset of the polytope's edge directions we present a total polynomial-time algorithm for computing the edge-skeleton (including vertex enumeration) of the polytope. All complexity bounds refer to the oracle Turing machine model. We consider two main applications, where we obtain (weakly) total polynomial-time algorithms: Signed Minkowski sums of convex polytopes, and computation of secondary, resultant, and discriminant polytopes. Further applications include convex combinatorial optimization and convex integer programming.
(Joint work with Ioannis Emiris and Bernd Gaertner)

## Homological Reconstruction and simplification in $\mathbb{R}^3$

Marc, Glisse
INRIA Saclay – Île de France

We consider the problem of deciding whether the persistent homology group of a simplicial pair $(K, L)$ can be realized as the homology $H_*(X)$ of some complex $X$ with $L \subset X \subset K$. We show that this problem is NP-complete even if $K$ is embedded in $\mathbb{R}^3$.
As a consequence, we show that it is NP-hard to simplify level and sublevel sets of scalar functions on $\mathbb{S}^3$ within a given tolerance constraint. This problem has relevance to the visualization of medical images by isosurfaces. We also show an implication to the theory of well groups of scalar functions: not every well group can be realized by some level set, and deciding whether a well group can be realized is NP-hard.

## Sclow Plots: Visualising Empty Space

Lars Kühne
Friedrich-Schiller-University Jena, Germany

Scatter plots are mostly used for correlation analysis, but are also a useful tool for understanding the distribution of high-dimensional point cloud data. An important characteristic of such distributions are clusters, and scatter plots have been used successfully to identify clusters in data. Another characteristic of point cloud data that has received less attention are regions that contain no or only very few data points. I will present Sclow Plots, a novel technique that augments scatter plots by projections of flow lines along the gradient vector field of the distance function to the point cloud. These augmented scatter plots enable a much better understanding of the geometry underlying the point cloud by revealing such empty regions or voids.

## A High-dimensional Non-parametric Two-sample Test based on Bayesian Mixtures over Spatial Partitions

Alix Lhéritier
Inria Sophia-Antipolis, France

Given two sets of samples in $\mathbb{R}^d$, a two-sample test aims at deciding whether they come from the same density or not. A common approach for designing these tests consists of estimating, in a non-parametric way, some function revealing differences between the two densities. In high-dimensional cases, two issues faced by this estimation are over-fitting and the curse-of-dimensionality, the latter even though real life data often lie close to a mixture of low dimensional manifolds. We propose a novel powerful two-sample test using spatial partitions that adapt to the intrinsic dimension of the data, as well as state-of-the-art statistical learning methods preventing over-fitting.

## The Compressed Annotation Matrix: an Efficient Data Structure for Computing Persistent Cohomology

Clément Maria
INRIA Sophia Antipolis-Méditerranée, France

Persistent homology with coefficients in a field $\mathbb{F}$ coincides with the same for cohomology because of duality. We propose an implementation of a recently introduced algorithm for persistent cohomology that attaches annotation vectors with the simplices. We separate the representation of the simplicial complex from the representation of the cohomology groups, and introduce a new data structure for maintaining the annotation matrix, which is more compact and reduces substancially the amount of matrix operations. In addition, we propose a heuristic to simplify further the representation of the cohomology groups and improve both time and space complexities. Join work with Jean-Daniel Boissonnat and Tamal K. Dey.

## Zigzag Zoology: Rips Zigzags for Homology Inference

Steve Oudot
Inria, France

For points sampled near a compact set $X$, the persistence barcode of the Rips filtration built from the sample contains information about the homology of $X$ as long as $X$ satisfies some geometric assumptions. The Rips filtration is prohibitively large, however zigzag persistence can be used to keep the size linear. In this talk I will present several species of Rips-like zigzags and compare them with respect to the signal-to-noise ratio, a measure of how well the underlying homology is represented in the persistence barcode relative to the noise in the barcode at the relevant scales. Some of these Rips-like zigzags have been available as part of the Dionysus library for several years while others are new. Interestingly, we showed that some species of Rips zigzags will exhibit less noise than the (non-zigzag) Rips filtration itself. Thus, Rips zigzags can offer improvements in both size complexity and signal-to-noise ratio.

Along the way, we developed new techniques for manipulating and comparing persistence barcodes from zigzag modules. In particular, we gave methods for reversing arrows and removing spaces from a zigzag while controlling the changes occurring in its barcode. These techniques were developed to provide our theoretical analysis of the signal-to-noise ratio of Rips-like zigzags, but they are of independent interest as they apply to zigzag modules generally.

This is joint work with Donald Sheehy.

## Sparsification of Motion-Planning Roadmaps by Edge Contraction

Oren Salzman
Tel-Aviv University, Israel

Asymptotically optimal motion-planning roadmaps constructed by the recently introduced PRM$^*$ algorithm provide high-quality, dense roadmaps. We consider the problem of *sparsifying* the roadmap, or reducing its size, while minimizing the effect on the quality of paths that can be extracted from the resulting roadmap. We present Roadmap Sparsification by Edge Contraction (RSEC), a simple and effective sparsifying algorithm. The primitive operation used by RSEC is *edge contraction*—the contraction of a roadmap edge $(v', v'')$ to a new vertex $v$ and the connection of the new vertex $v$ to the neighboring vertices of the contracted edge's vertices (i.e. to all neighbors of $v'$ and $v''$). For certain scenarios, we compress more than 98% of the edges and vertices of a given roadmap at the cost of degradation of average shortest path length by at most 2%.

*Shared talk with Doron Sharahabani.*

# Optimization and Learning with Random Pursuit

Sebastian Stich
ETH Zürich, Switzerland

We consider unconstrained randomized optimization of smooth convex functions in the gradient-free setting with Random Pursuit (RP). This algorithm only uses zeroth-order information of the objective function and computes an approximate solution by repeated optimization along a randomly chosen direction.

State-of-the-art (derivative-free) optimization algorithms often build and iteratively adapt a quadratic model of the objective function based only on the queried zeroth-order information. This allows to significantly increase the convergence rate of the optimization algorithm if the model describes the local geometry accurately enough.

We demonstrate how RP can not only be used for optimization, but the same algorithm can be used to build and iteratively update a quadratic model of a convex objective function. This model allows RP to achieve optimal convergence rates on quadratic functions. If time allows, we will also present some illustrative numerical experiments.

# Gaussian Field Homology

Rien van de Weygaert
Kapteyn Astronomical Institute, University of Groningen, the Netherlands

According to the *gravitational instability scenario*, structure in the Universe emerged from tiny primordial density and velocity perturbations which evolved under the force of gravity. Theories of the early universe predict that these primordial perturbations have the character of a homogeneous and isotropic spatial Gaussian process, as such fluctuations are produced by quantum processes during the inflationary epoch. There is overwhelming observational evidence in terms of the cosmic microwave background, most recently by the Planck satellite, that the structure of the primordial Universe is indeed Gaussian in nature.

In this presentation, we will describe our investigation of the homology of Gaussian random fields. We will study its hierarchical topological structure by means of persistence diagrams and assess the sensitivity to the primordial power spectrum of fluctuations. We will also discuss the Betti number curves for level set filtrations of Gaussian fields, and relate this to the predominance of islands, voids and tunnels at various excursion thresholds. We will also develop an analytical model for these descriptions, based on a graph representation of the Morse-Smale complex of a Gaussian field.

# Intrinsic Simplices on Riemannian Manifolds

Mathijs Wintraecken (joint work with Ramsay Dyer)
Rijksuniversiteit Groningen, the Netherlands

Given $m+1$ points on a Riemannian $m$-manifold, we present an intrinsic method to smoothly map a Euclidean simplex to the manifold, with the given points as vertices. We shall then argue that the image is diffeomorphic to the Euclidean simplex if the given vertices are nicely distributed. Here nice can be expressed in terms of Euclidean simplex quality measures. This work motivated by questions arising in the context of sampling manifolds for triangulation.