Adaptive Sampling with Topological Scores

Dan Maljovec[1], Bei Wang[1], Ana Kupresanin[2], Gardar Johannesson[2], Valerio Pascucci[1], and Peer-Timo Bremer[2]
[1]Scientific Computing and Imaging Institute, University of Utah
[2]Lawrence Livermore National Laboratory

Motivation

In the field of uncertainty quantification, the goal is to gain knowledge and insight about a phenomenon from a simulation or model. It is the hope that the phenomenon has some interesting relationships between the input parameters and the response values, thus by querying the simulation or model we can understand relationships and features of the phenomenon with respect to the input. Unfortunately given a complex simulation, querying data can be expensive, time-consuming, or otherwise difficult. Therefore, it is important to choose the “best” places to query the simulation. This is accomplished by adaptive sampling of the model or simulation.

Proposed Solution

We can approximate the expensive simulation with a simpler model given some initial information, and then locate the important areas by assigning scores to potential sample points.

Current methods for scoring candidates use geometric or statistical information about the dataset.

We seek to investigate novel methods for scoring candidates based on the topology of the dataset.

Definitions

Ground truth (GT) - The underlying “expensive” simulation we seek to avoid querying excessively

Training data (TD) - Initial data points sampled off the GT.

Predicting model (PM) - A statistical model constructed from the TD used to approximate the GT and to assign function values to candidate data.

Candidate data (CD) - Large selection of points to be scored by predicting their values from the PM.

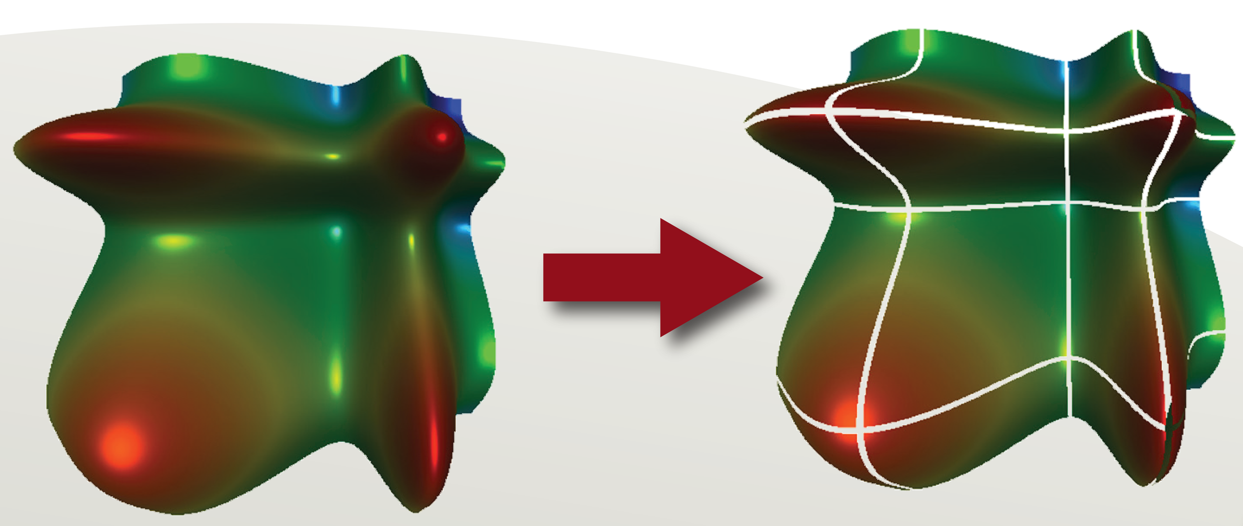
Metrics

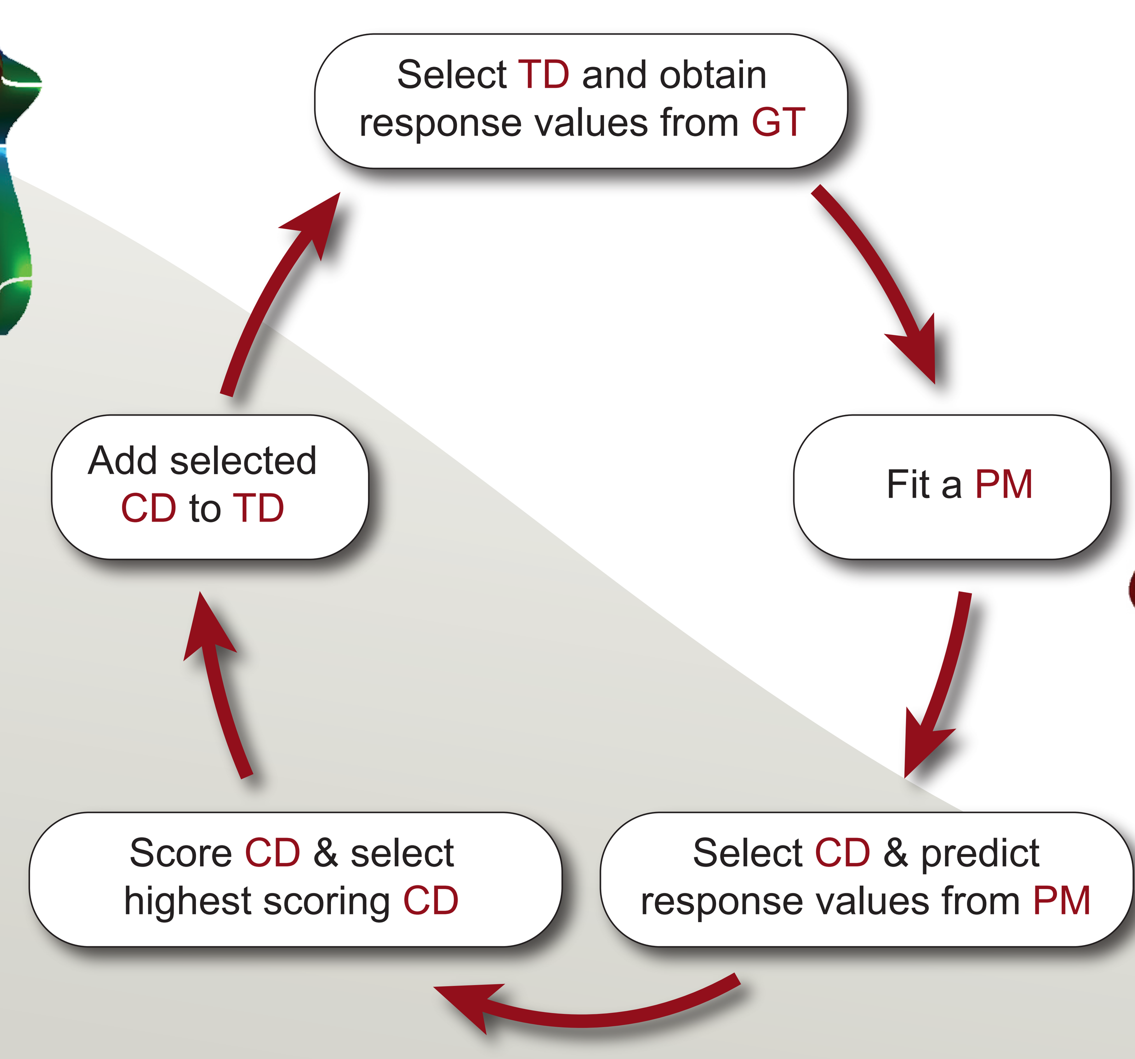
Classic Functions

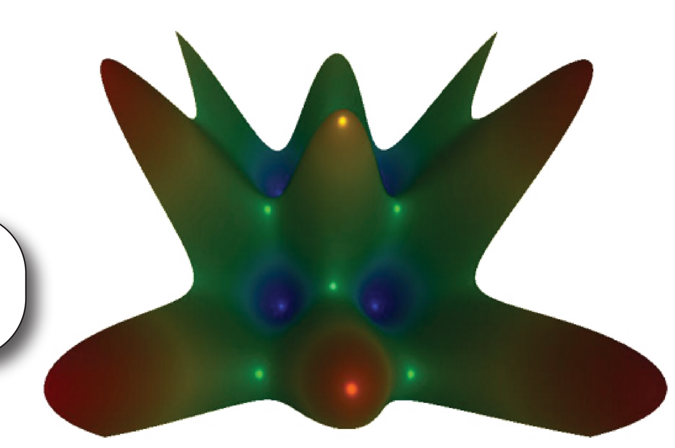
- ALM - Active Learning MacKay.
- Delta - Difference between predicted response at the CD point and the response value of the nearest TD point.
- EI - Expected Improvement.

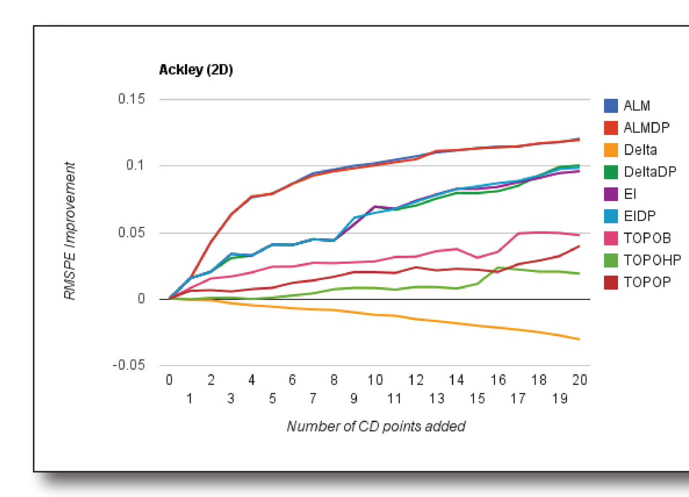
Proposed Topological Functions

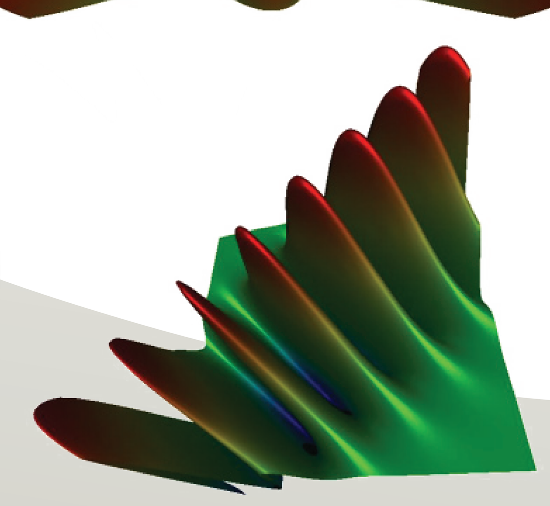
- TOPOB - Measure of the bottleneck persistence between MSC before and after adding CD point.
- TOPOHP - Select CD point which has the highest persistence in MSC including TD and all CD points.
- TOPOP - Average change in persistence of all points in MSC before and after adding the CD point.

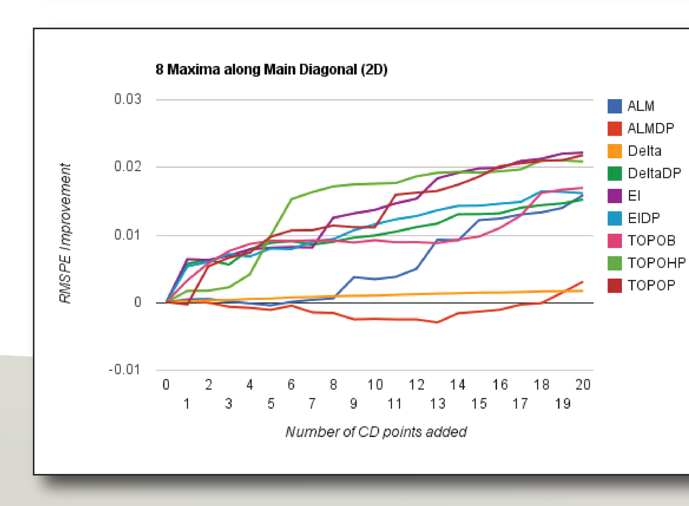










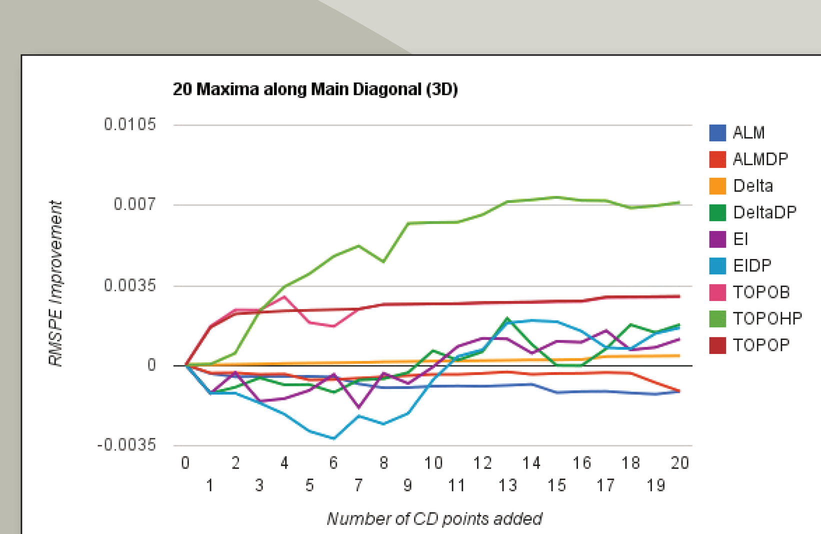


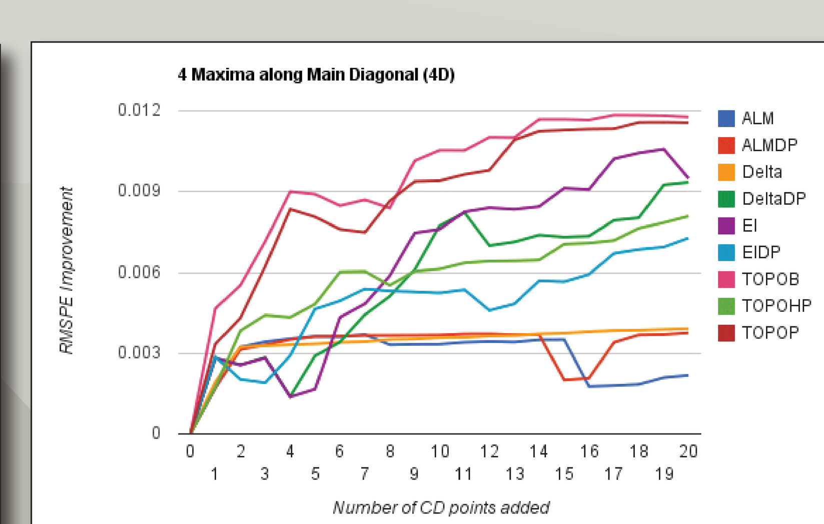
Topology Background

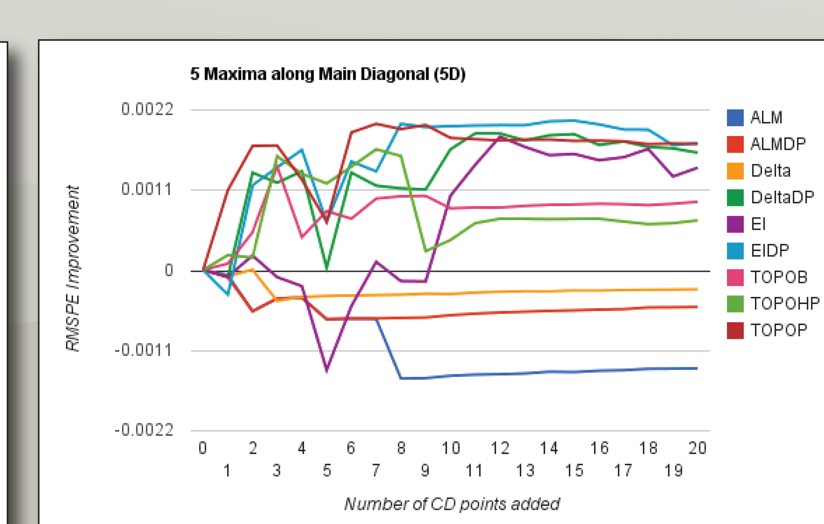
In order to quantize the topology into a metric we construct the Morse-Smale Complex (MSC) of the dataset.

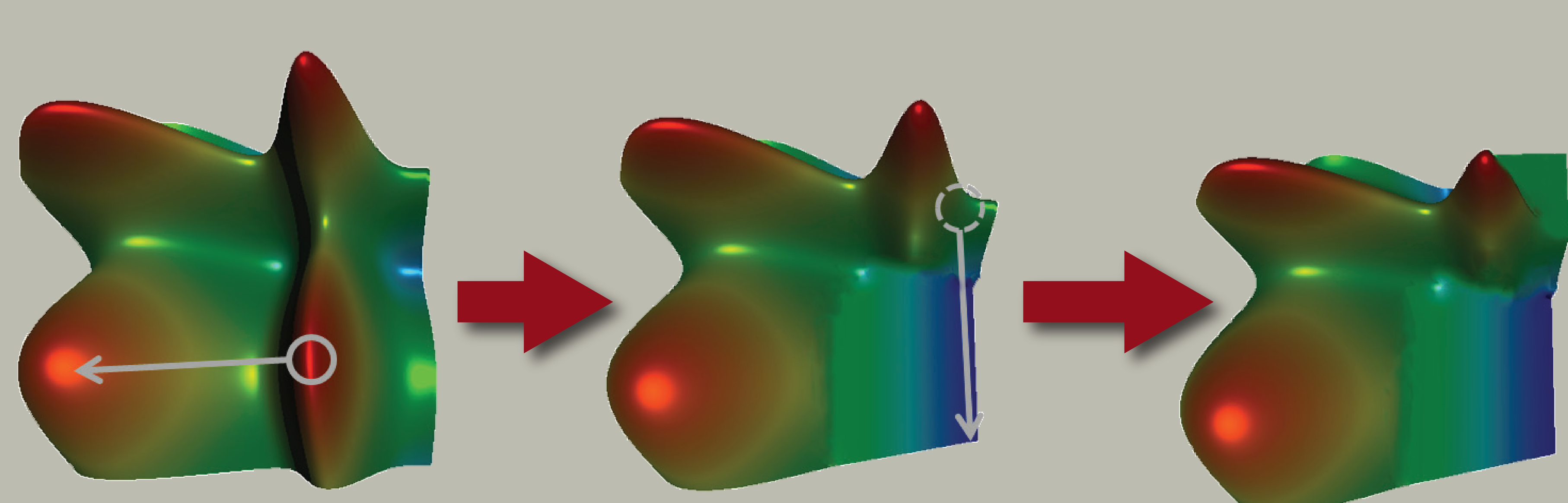
The MSC breaks the domain into monotonic regions and can be done in arbitrary dimensions allowing these methods to be used for scoring high-dimension datasets.

Furthermore, we can control the resolution of the MSC by simplifying one extremum to an adjacent one. The order in which we simplify is based on an extremum’s persistence, the difference in value between an extremum and its closest (response value) saddle.









Results

The above images show the surface we are trying to model for the 2D cases. The graphs show the improvement of the root-mean-square prediction error (RMSPE) as we select 20 CD points to add.

Future Work

- Investigate using different PMs
- Further investigate how to measure topological importance of a CD point
- Optimize for even higher dimension simulations

SCI INSTITUTE • EXHIBIT • EXPLORE • EXCITE • EXPERIENCE • EXCHANGE • SCI X

THE UNIVERSITY OF UTAH   www.sci.utah.edu   LLNL