

Possible Research Projects

Computational Aspects of Data Mining

Data mining is a new area of research, applying methods from statistics, machine learning and computational mathematics to the analysis of very large data sets. Some of the main challenges in data mining include

- High dimensional data
- Large number of data points (trillions)
- Ill-defined goals of searches

We have access to very large data sets and collaborate with several institutions like the NSW Health Department, the HIC, NRMA and the Taxation Office. The honours projects will deal with the computational challenges which are connected with finding information in such large data sets.

The aim of the projects is to develop, analyze, and implement data mining algorithms which can extract useful information from very large data sets. The data size may be in the order of gigabytes or even terabytes. Particular requirements for the algorithms are scalability with respect to the number of data points even in the event of high dimensionality.

Depending on the interests of the students the work can have components of

- experimental exploration of computational performance using both real and simulated data
- development and analysis of algorithms and software
- mathematical error and performance analysis

Of particular interest to our group is the use of Sparse grids for Data Mining.

Sparse Grids

Consider an example of describing the price of a house as a function of the number of rooms, the build of house, location, accessibility and parameters describing the neighbourhood. Such a function is useful to estimate the value of a house new on the market. A practical available data set, the Boston Housing Data contains 14 parameters from the 1977 US Census.

In choosing a function class one needs to consider the complexity of function evaluation, the approximation power, and the complexity of the class. In this project we will study an approximation class, the sparse grids, which balance these requirements. Consider first the one-dimensional case of real functions. The linear space of piece-wise linear functions \mathbb{V}_m (with grid spacing $\frac{1}{m}$ defined on a uniform grid) lead to functions which are very fast to evaluate, are continuous and have good approximation power for twice continuously differentiable (C^2) functions. The so-called hat-functions (which are Lagrangian, i.e., they equal 1 in one grid point and zero in all the others) form a convenient basis of \mathbb{V}_m . From this one-dimensional case one derives a class for multidimensional functions using the tensor product basis where the basis functions in d dimensions are products of basis functions in one dimension. This is a choice which is very popular in surface fitting, and the finite element method in general and has been used numerous times for 2 and 3 dimensional problems. However, this turns out to be an impractical choice in high dimensions. Suppose we have a grid with m basis functions in (every) one dimension, then there would be m^d tensor basis functions for a d -dimensional space. The approximation of C^2 functions is proportional to m^{-2} , independent of the dimension. In this case one can improve any given approximation by a factor of 4 by doubling the number of basis functions in all dimensions. This would increase the number of tensor product basis functions by a factor of $2^{14} = 16,384$. Note that the smallest (nontrivial) space of tensor product piecewise linear functions has dimension 2^{14} , the case of 16 grid points in each dimension would lead to a $16^{14} = 2^{56} \approx 72 * 10^{15}$ dimensional function space and the storage of a general function in that space would require $72 * 10^{15}$ parameters which is far beyond what any current storage space can handle. This is why tensor product basis functions, and in general, piecewise multi-linear functions have not been used for high-dimensional functions.

Sparse grids are “low”-dimensional subspaces of this tensor-product space. They filter out multi-dimensional high-frequency or strongly oscillating terms which do not contribute much to the approximation power of the functions. They are based on hierarchical basis functions, which, like wavelets, resolve the function both in space and in the scale domain, and consequently, the coefficients of these basis functions are much smaller for the basis functions associated with small scales than for the ones associated with large scales. This effect is strongly amplified in higher dimensions and the case of products of functions of small scales. The sparse grid function spaces are obtained from the tensor product space

by setting all coefficients which are bound to be small for any C^2 function to zero. This leads to a space, which, instead of having m^d dimensions, has asymptotically in m only $\log_2(m)^{d-1}m$ dimensions. The approximation error for C^2 functions in this space drops from $O(m^{-2})$ to $O(\log_2(m)^{d-1}m^{-2})$ (asymptotically in m). This *sparse grid* function space has been introduced several times since the 1960s but has made its debut as a computational tool to solve partial differential equations in 1991 by Zenger. It has been shown that these sparse grid spaces are flexible enough to deal with problems of up to around $d = 10$ dimensions.

Examples of Projects

Here is a range of possible projects. Please come and talk to us about them and other possibilities.

Investigating the use of Sparse Grids to approximate high dimensional functions

Possible Supervisor: Dr Markus Hegland

Investigating the use of Sparse Grids for Probability Density Estimation

Possible Supervisor: Dr Steve Roberts

Producing histogram estimates of high dimensional data sets provides an important application of sparse grid techniques. There are opportunities for mathematical analysis as well as practical implementation issues.

Solving high Dimension Schrödinger's equation using Sparse Grids

Possible Supervisor: Dr Steve Roberts

Standard PDE methods are restricted to dimensions of 3 to 4. A typical n quantum particle simulation involves the solution of a $3n$ dimensional pde, Schrödinger's equation,

$$i\frac{\partial\phi}{\partial t} = \sum_{i=1}^n \Delta_{\mathbf{x}_i}\phi + V(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$$

where $\mathbf{x}_i \in \mathbb{R}^3$. We can investigate the application of sparse grid methods to the solution of this equation.

Clustering techniques

Possible Supervisor: Dr Markus Hegland, Dr Steve Roberts

Clustering, i.e., finding groups of similar objects is a central theme in data mining. While the k-means algorithm is one of the most popular at the moment, strong contenders are based on the estimation of density. Following a review of the literature the various methods should be compared and a new method based on finite element techniques and sparse grids investigated.

Parallel high-dimensional density estimation

Possible Supervisor: Dr Markus Hegland, Dr Linda Stals, Dr Jochen Garcke

The modelling of high-dimensional data sets poses a particular challenge both for approximation and the parallel implementation. This project we investigate the parallelization of sparse grid methods.

Comparative studies of predictive models

Possible Supervisor: Dr Steve Roberts, Dr Markus Hegland, Dr Jochen Garcke

Comparing the relative merits of different predictive models can tedious and is often not done in a systematic way. This is largely due to different implementations, different control parameters and different assumptions about data and data formats. This project will set up a unified framework for predictive models, data and meta data. A wide range of predictive models such as neural nets, support vector machines, MARS, CART, sparse grids etc will be included in the framework and studies on test data sets as well as real-world datasets will be carried out. The outcome will be a suite of predictive models as well as an assessment of their strengths, weaknesses and scopes of applicability.

Adaptive sparse grids

Possible Supervisor: Dr Markus Hegland, Dr Jochen Garcke

Sparse grids in their basic form requires a maximal resolution to be specified. In reality, however, problems are not likely to be isotropic i.e. require the same resolution in all dimensions nor is it likely that the 'right' resolution is known a-priori. Adaptive sparse grids automatically search for the 'best' resolution in each dimension and builds a model of the data. This project will study techniques, implementations and merits of such adaptive schemes for use in predictive modelling using sparse grids.

Parallel sparse grids

Possible Supervisor: Dr Jochen Garcke

See parallel high-dimensional density estimation.

Data Mining Medical Data

Possible Supervisor: Dr Linda Stals

Bon Secours Hampton Roads (BSHR) is a non-profit health care organisation covering three institutions in the Hampton Roads area of Virginia, USA. One of the services offer by BSHR is physical rehabilitation, which is provided in various settings including in-patient rehabilitation, transitional care, skilled nursing centres, home health and outpatient therapy.

The US Medicare administration is currently in the process of phasing in a new system to determine the amount of payment based on expected resource consumption. This new payment system means that the financial risk has shifted from Medicare to the provider.

In accordance with the phasing in of the new Medicare payment the rehabilitation centre of BSHR wants to determine a treatment path for all new patients that fall under its administrative umbrella. The amount of money reimbursed by Medicare depends upon which setting the patient goes to and how long they spend there.

One parameter of interest is the FPC (Functional Patient Categories) score. The FPC is designed to measure a patient's functionality and is recorded for each patient in each setting. It is a numeric value and, ideally, should increase during a patient's treatment. As well as the FPC a patient is assigned a SA (Severity Adjustment) score when they first enter the system. The SA measures a patient's overall health.

The FPC and SA are two numerical values that are designed to measure the *quality of care*.

To determine an appropriate method of treatment for a patient that gives the best clinical care and maximum reimbursement the hospital needs to be able to predict a patient's FPC given their SA.

In this project we shall explore various data fitting technique using the BSHR data as our model problem.

Environmental Flows

Shallow Water Wave Equations

Depth averaged models of shallow water flow over discontinuous terrain would provide for the efficient simulation of many practical physical situations. Steps are used in spillways to provide an efficient method for dissipating energy. Rivers which breach their banks can be modelled as a flow over discontinuous terrain. In estuaries, tidal flows play an important environmental role providing nutrients to plants and organisms in the tidal channels and on the mud flats. Flows from the well defined channels that incise the flat and wide estuary can be considered as flow over a discontinuity.

The standard shallow water wave equation

$$\begin{bmatrix} h \\ uh \end{bmatrix}_t + \begin{bmatrix} uh \\ u^2h + \frac{1}{2}gh^2 \end{bmatrix}_x = \begin{bmatrix} 0 \\ -gh \left(\frac{dz}{dx}\right) \end{bmatrix} \quad (1)$$

is a conservation law for fluid mass and horizontal momentum, together with a momentum dissipation term dependent on the slope of the terrain. This equation is obtained as a depth averaged approximation of the full two or three dimensional Euler equation. Flow over discontinuous terrain can be obtained as a limit of infinite slope. On the other hand, it has been reported by Alcrudo and Benkhaldoun that an alternative form of the shallow water wave equation

$$\begin{bmatrix} h \\ u/g \end{bmatrix}_t + \begin{bmatrix} uh \\ h + \frac{1}{2}u^2/g + z \end{bmatrix}_x = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (2)$$

in which energy is the conserved quantity, provides a good approximation to the flow over discontinuous terrain. The alternative forms of the equations lead to different prospects for the numerical solution of the shallow water equation over discontinuous terrain.

Free Surface Euler Equation Solver

The equations of motion for incompressible two-phase flow can be written in the form

$$\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} + \frac{\nabla p}{\rho(c)} = \frac{1}{\rho(c)} \nabla \cdot (\mu(c)[\nabla \mathbf{u} + \nabla \mathbf{u}^T]) - \frac{1}{\rho(c)} \gamma \kappa(c) \nabla c - g \mathbf{e}_g \quad (3)$$

$$\nabla \cdot \mathbf{u} = 0 \quad (4)$$

together with a volume-of-fluid equation

$$c_t + \nabla(\mathbf{u}c) = 0 \quad (5)$$

where $\mathbf{u} = (u, v)$ is fluid velocity, c is the volume-of-fluid or the void fraction function (1 if liquid, 0 if gas) and $\kappa(c)$ is the mean curvature of the interface between the liquid and gas, as defined by c .

Examples of Projects

Here is a range of possible projects. Please come and talk to us about them and other possibilities.

Reconcile the use of Momentum and Energy Conserving Equations

Possible Supervisor: Dr Steve Roberts

Investigate the use of depth averaged equations used to model shallow water wave equations. Investigate the derivation of the equations. Consider numerical methods. Investigate the use of Momentum and Energy Conserving Equations when approximating flow over a step.

Solution of the Navier Stokes Equation with Free Boundaries

Possible Supervisor: Dr Steve Roberts

Investigate the methods for solving the Navier Stokes equation. Consider implementation issues. Look at general computational Fluid Dynamical methods and apply this to the solution of the Navier Stokes equation with free boundaries

Simulating Radiation Transport

Radiation transport problems dominate the budgets of many U.S. government lab production supercomputers.

Under certain assumptions, such as isotropic radiation, optically thick material and temperature equilibrium, radiation transport may be modeled by the equation:

$$\frac{\partial E}{\partial t} - \nabla \cdot (D(E) \nabla E) = 0 \quad \text{on } \Omega \times I, \quad (6)$$

where E is the radiation energy.

More general models that remove some of the physical assumptions are given by systems of non-linear equations.

One definition of the diffusivity term, $D(E)$, is:

$$D_1(E) = Z^\alpha E^\beta, \quad (7)$$

with $\alpha < 0$, $0 \leq \beta \leq 1$. Typically α is taken to be between -1 and -3 while β is taken to be 1/4 or 3/4. Z is the atomic mass number and may vary within the domain to reflect

inhomogeneities in the material. The constant β controls the strength of the nonlinearity while α affects the size of the jumps in the coefficients.

The definition of $D(E)$ given in Equation (??) may produce results that shows the energy moving through the system at a rate faster than the speed of light. Consequently a flux limiter is included to slow down the movement and the diffusivity term is rewritten as:

$$D_2(E) = \left(\frac{1}{Z^\alpha E^\beta} + \frac{|\nabla E|}{E} \right)^{-1}. \quad (8)$$

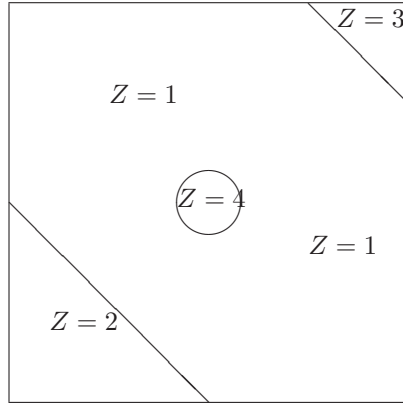


Figure 1: The values for the atomic mass number Z depend on the topology of the material. In our model problem, we define Z as shown above.

The domain, Ω , is a square domain ($[0, 1] \times [0, 1]$) with the following mixture of Newton and Neumann boundary conditions;

$$\begin{aligned} \partial E / \partial n &= 0 && \text{on } \Gamma_N \times I, \\ n^T D(E) \nabla E + E/2 &= 2 && \text{on } \Gamma_{F_0} \times I, \\ n^T D(E) \nabla E + E/2 &= 0 && \text{on } \Gamma_{F_1} \times I, \end{aligned}$$

where Γ_N represents the lines $y = 0$ and $y = 1$, Γ_{F_0} is the line $x = 0$ and Γ_{F_1} is the line $x = 1$, n is the outward unit normal and I is the time interval.

In our model problem we take Z to be 1.0 except in the following regions:
 $\sqrt{(x - 0.5)^2 + (y - 0.5)^2} \leq 0.125$, $y \leq 0.5 - x$ and $y \geq 1.75 - x$ where $Z = 4$, $Z = 2$ and $Z = 3$ respectively. See Figure ??.

Initially, the energy, E is set to be the constant value $E = 10^{-5}$.

We have developed and studied the performance of highly parallelisable solvers and grid adaptation procedures for unclassified nonlinear problems in radiation transport.

Example Project

Possible Supervisor: Dr Linda Stals

In addition to the standard grid resolution and parallel scalability studies, we propose a temporal accuracy study. For cases where the non-linear problem at each time-step is solved to a predetermined tolerance, higher temporal accuracy can be obtained using a second-order backwards differencing scheme. Even higher accuracy (third or fourth order) can be obtained by resorting to a Runge-Kutta time stepping schemes. However, the benefits of higher order time accurate methods must be compared with the simple strategy of time step refinement for lower

order time discretisations. We will develop a time error estimation procedure to use as a driver for an adaptive time-step refinement strategy.