Sparse grids and optimisation

Lecture 1: Challenges and structure of multidimensional problems

Markus Hegland[1], ANU

10.-16. July, MATRIX workshop on approximation and optimisation

[1]support by ARC DP and LP and by Technical University of Munich and DFG

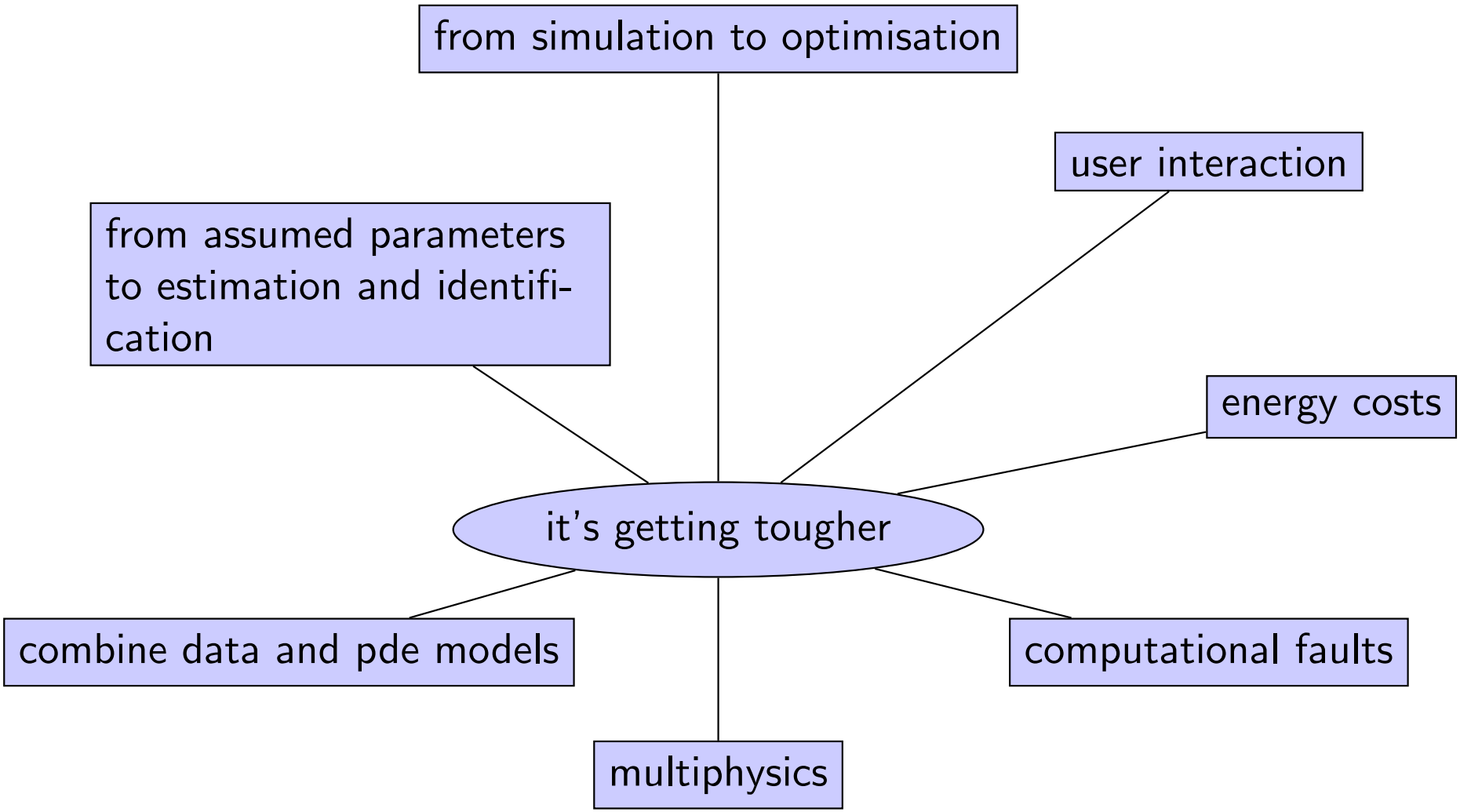
Outline

Outline

Challenges of numerical analysis

- numerical techniques are major driver of innovation in industrial societies and indispensable for design of aeroplanes, weather forecast, environmental monitoring, medical diagnostics, robotics and image processing
- fundamental techniques and their foundations well established
- techniques include finite elements, finite differences and finite volumes, they are widely available but do not work for
  - ill-posed problems
  - high-dimensional problems
- in both cases one requires specially adapted techniques and theory and practice of these techniques are active area of research
- recent developments in High Performance Computing (HPC) and new algorithms allow solution of new multidimensional problems – but introduce new challenges

Computational challenges in HPC



Examples: PDEs, data, parameters

- PDEs $u = \operatorname{argmin}_{v \in V} J(v)$
  - elliptic PDEs $J(v) = \frac{1}{2} a(v, v) - f(v)$
  - least squares solution $J(v) = \int (Lv(x) - f(x))^2 \, dx$
  - eigenvalues $J(v) = \frac{a(v,v)}{b(v,v)}$ (Rayleigh quotient)
- fitting data $u = \operatorname{argmin}_{v \in V} L(v)$
  - penalised least squares $L(v) = \frac{1}{N} \sum_{i=1}^{N} (v(x_i) - y_i)^2 + a(v, v)$
  - MAP for density $p(x) = \exp u(x)$

$$L(v) = \frac{1}{N} \sum_{i=1}^{N} v(x_i) + \log \int \exp(v(x)) \, dx + a(v, v)$$

- parametric problems combine PDEs and data fitting

$$u = \operatorname{argmin}\{L(u(\mu); \mu) \mid v = u(\mu), \mu \in M\}$$

  with PDE constraint $u(\mu) = \operatorname{argmin}_v J(v; \mu)$
- quantities of interest $q = s(u)$ target of approximation, e.g. energy, moments, likelihood, cost, risk

Integrating multiplicities

HPC tackles a "multi-challenge"

- multi-disciplinary domains and education
- multi-physics models
- multi-scale models
- multi-dimensional numerics
- multi-level numerics
- multi-core systems

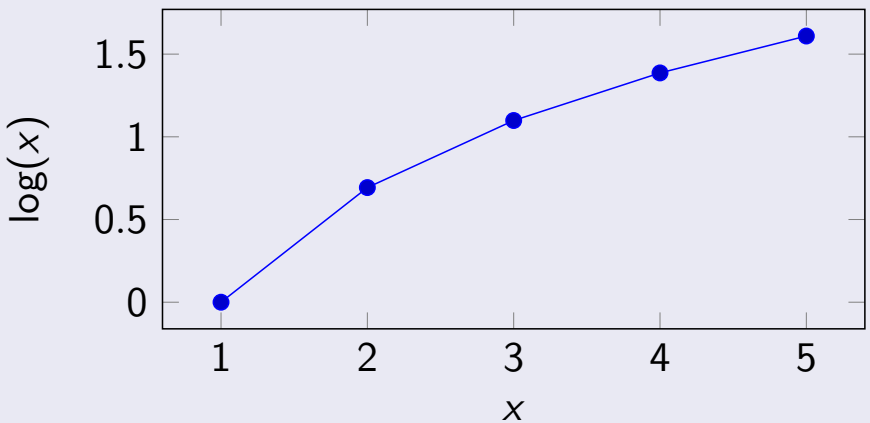
The prevailing paradigm in modern computational science and HPC combines multiple resources and approaches with a wide range of different properties to gain new insights into immensely complex systems in the natural, engineering and social sciences.
This reflects the multi-skilled and multi-cultural societies in which modern science is developed.

Outline

Interpolation

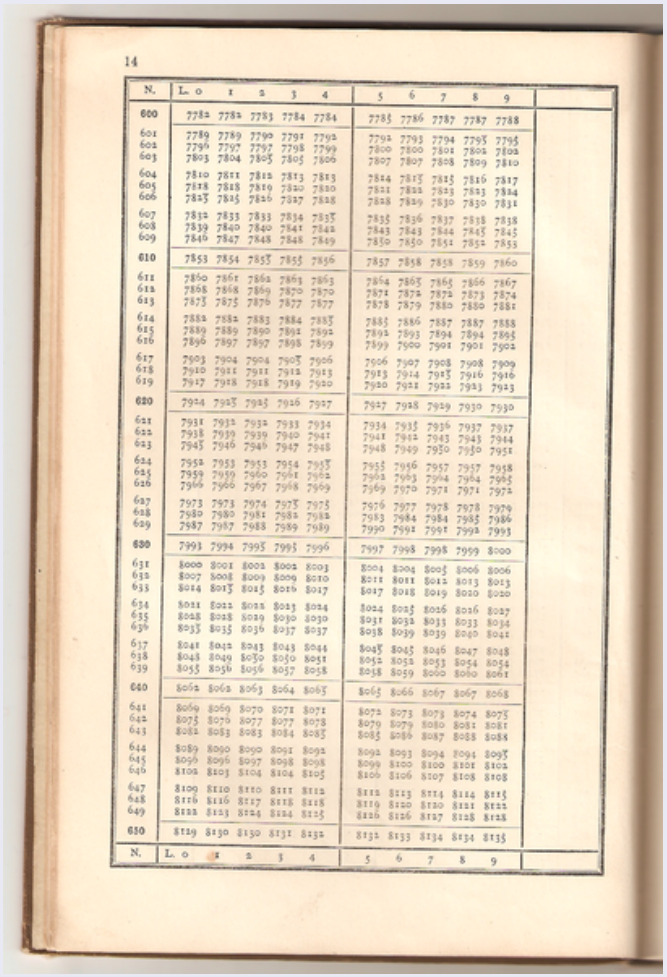
- evaluation of function expensive
- compute some, interpolate
- logarithm tables by H. Briggs 1617

piecewise linear interpolant



- fast evaluation
- reasonable accuracy
- stable, positive

logarithm table

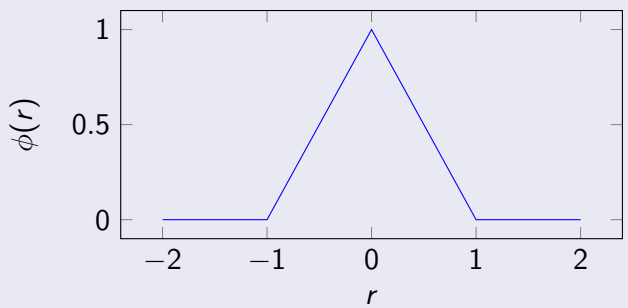


from: Wikipedia

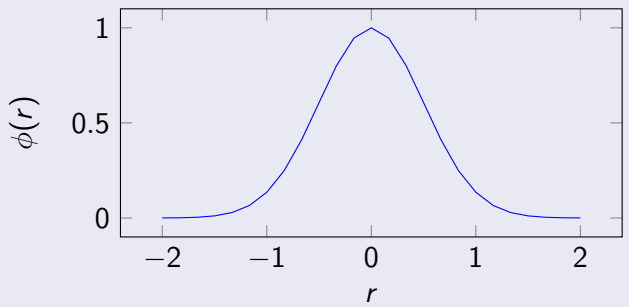
A more accurate and flexible approach

$$f_I(x) = c_1\,\phi(|x - x_1|) + \cdots + c_m\,\phi(|x - x_m|) \qquad \text{interpolation function}$$

piecewise linear



Gaussian $\phi(r) = e^{-r^2/\gamma}$



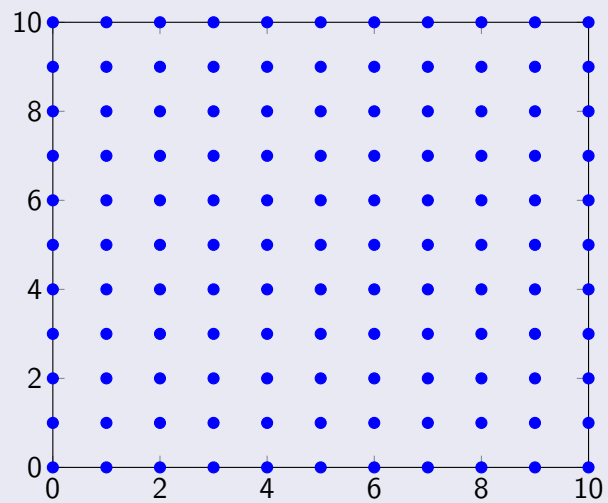
interpolation equations

$$\begin{bmatrix} \phi(0) & \phi(|x_1 - x_2|) & \cdots & \phi(|x_1 - x_m|) \\ \phi(|x_2 - x_1|) & \phi(0) & \cdots & \phi(|x_2 - x_m|) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(|x_m - x_1|) & \phi(|x_m - x_2|) & \cdots & \phi(0) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{bmatrix}$$

Multidimensional interpolation

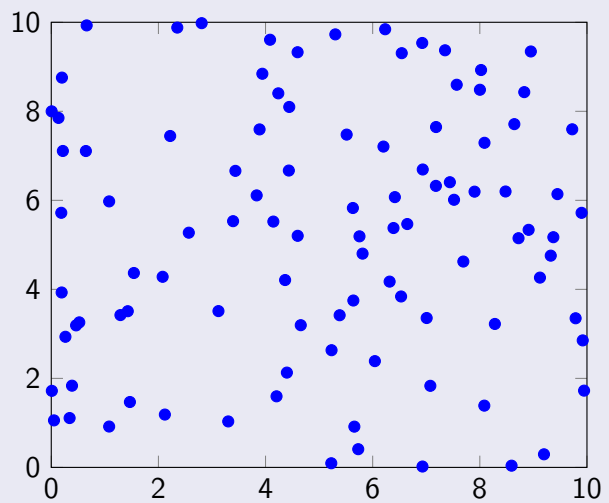
use $\phi(\|x - x_i\|)$ where $\|x - x_i\|$ is Euclidean distance of x and x_i

x_i on regular grid



error $O(m^{-2/d})$

random points x_i



error $O(m^{-1/2})$

- up to $d = 4$ dimensions and smooth functions regular grid competitive
- for higher dimensions random interpolation points better
- theory for random points uses *law of large numbers*

The concentration of measure

- in high dimensions any pair of random points have same distance
- consequently interpolant is close to constant with high probability

interpolation matrix for $d = 100$

$$[\phi(\|x_i - x_j\|)]_{i,j=1,\ldots,n} = \begin{bmatrix} 1 & 0.79 & 0.77 & 0.74 & 0.78 & 0.79 \\ 0.79 & 1 & 0.80 & 0.77 & 0.77 & 0.80 \\ 0.77 & 0.80 & 1 & 0.77 & 0.76 & 0.77 \\ 0.74 & 0.77 & 0.77 & 1 & 0.78 & 0.78 \\ 0.78 & 0.77 & 0.76 & 0.78 & 1 & 0.77 \\ 0.79 & 0.80 & 0.77 & 0.78 & 0.77 & 1 \end{bmatrix}$$

other instances of concentration of measure

- most of volume of sphere (Earth) close to surface
- law of large numbers, statistical convergence theory

[Lévy, 20s, Milman 70s, Gromov, Talagrand 90s+]

When concentration is not a problem for interpolation

- when the points of interest are on low-dimensional sub-manifold
- when function which is to be interpolated has known simple structure, e.g., is linear or additive:

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{d} f_i(x_i)$$

  or is close to such a function
- when function only depends on few dimensions

$$f(x_1, \ldots, x_d) = g(x_1, x_2, x_3)$$

dimension is not only a curse

in high dimensions any non-empty neighbourhood contains large numbers of points which can be used for error reduction by averaging [Anderssen, H. 1999]

Outline

Density estimation

- large data sets, queries expensive
- data set = probability measure over feature space
- histogram = piecewise constant approximation of measure
- extract relevant information fast from histogram

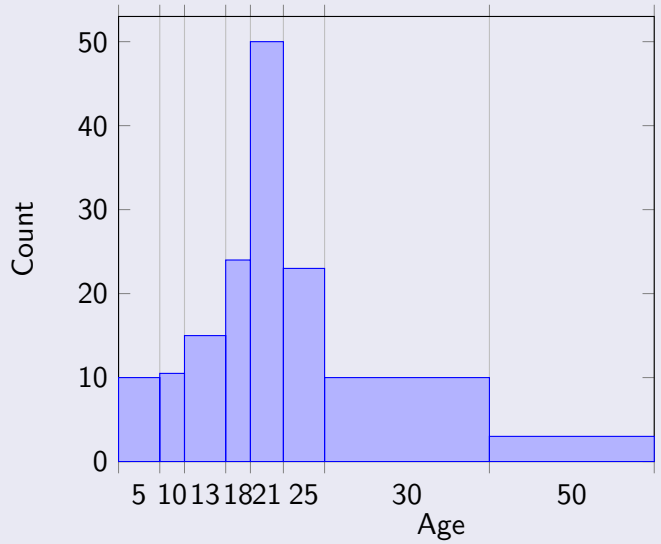
application

- mean, variance, moments
- number and location of modes
- skewness and tail behaviour

All modern theories of statistical inference take as their starting point the idea of the probability density function of the observations.

E. Parzen (1961) in An Approach to Time Series Analysis

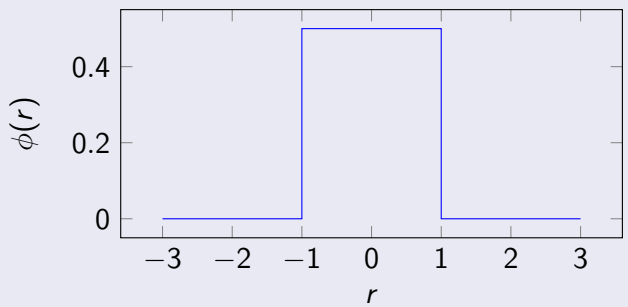
histogram



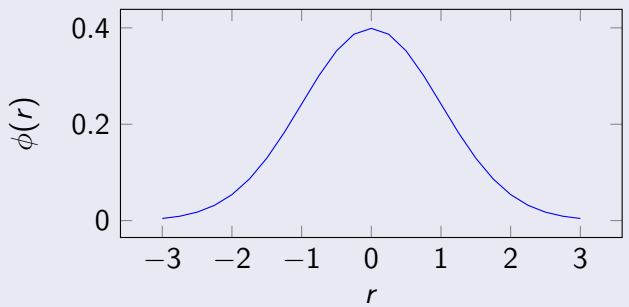
A more accurate and flexible approach

$$p_K(x) = \frac{\phi(|x - x_1|/\sigma)}{m\sigma} + \cdots + \frac{\phi(|x - x_m|/\sigma)}{m\sigma}$$ kernel density estimator

piecewise constant



Gaussian $\phi(r) = \frac{e^{-r^2/2}}{\sqrt{2\pi}}$



- more accurate representation of smooth densities
- control smoothness with width parameter σ, can depend on x
- no need to solve linear system of equations
- more flexible: also for multidimensional distributions

Challenges of multidimensional density estimation

low dimensional case

for any x only the $\phi(|x - x_i|/\sigma)$ for neighbouring x_i are nonzero, gives efficient estimator as every point has only few neighbours

$$p_K(x) = \sum_{x_i \in \mathcal{N}(x)} \phi(|x - x_i|/\sigma)/m\sigma$$

high dimensional case

all x_i are neighbours, need to consider all data points to evaluate density

$$p_K(x) = \sum_{i=1}^{m} \phi(|x - x_i|/\sigma)/m\sigma$$

very high dimensional case

if $x, x_1, \ldots, x_m$ are i.i.d. then all components $\phi(|x - x_i|/\sigma)/m\sigma$ of the same size, density asymptotically uniform $p_K(x) \approx E(\phi(|X_i - X_j|/\sigma))/\sigma$

When concentration is not a problem for density estimation

- when the points of interest are on low-dimensional submanifold
- when unknown p has known simple structure, e.g.

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p_i(x_i)$$

- more generally, the density is described by *graphical model* which leads to a factorisation as in

$$p(x_1, x_2, x_3) = \frac{p(x_1, x_2)\, p(x_2, x_3)}{p(x_2)}$$

- mixture model

$$p(x) = \sum_{i=1}^{K} p_i(x)\pi_i$$

where $p_i(x) = p(x|x \in \Omega_i)$ has some known form and $\pi_i = p(\Omega_i)$

Outline

Partial differential equations

Partial differential equations are a very widely used tool in computational science

examples of equations

$$ih\frac{\partial \psi}{\partial t} = -\frac{h^2}{2m}\Delta\psi + V\psi$$ Schrödinger equation, quantum chemistry

$$\frac{dp}{dt} = \sum_z (S_z - I)\Lambda_z p$$ Chemical master equation, molecular biology

$$\frac{\partial f}{\partial t} + v^T \nabla_x f + q(E + v \times B)\nabla_p f = 0$$ Vlasov equation, plasma physics

dimensionality

ψ and p can depend on hundreds of variables, f depends on five variables

Controlling the function values

Sobolev norms

important tool for PDE theory

$$\|u\|_k^2 = (-1)^k \int_\Omega u(x)\Delta^k u(x)\,dx$$

for $u \in C_0^\infty(\Omega)$ and completion

bounded solutions for $d \leq 3$

- PDE regularity theory

$$\|u\|_2 < \infty$$

- Sobolev embedding

$$|u(x)| \leq C\|u\|_2$$

case $d > 3$

- embedding for $k \geq \lfloor \frac{d}{2} \rfloor + 1$

$$|u(x)| \leq C\|u\|_k$$

- $k = 2$ from regularity theory

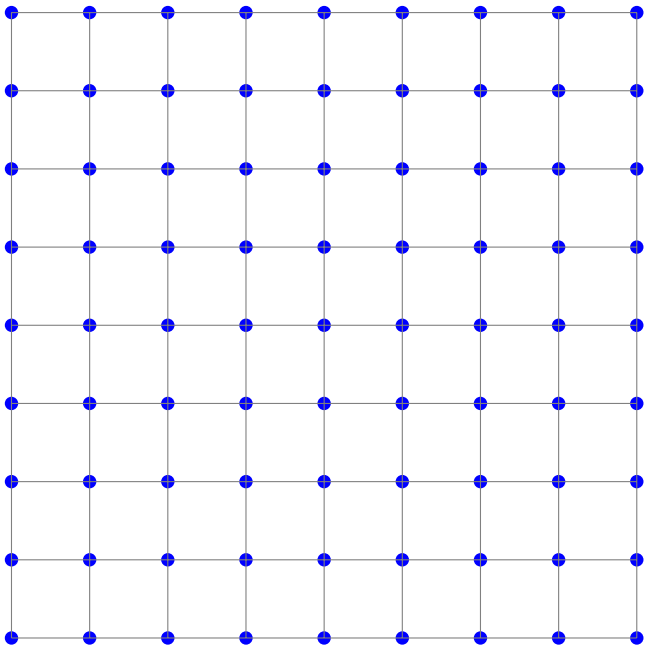
mixed norms

$$\|u\|_{\text{mix}}^2 = \int \left| \frac{\partial^d u(x)}{\partial x_1 \cdots \partial x_d} \right|^2 dx$$

and so $|u(x)| \leq C^d \|u\|_{\text{mix}}^2$

Outline

The grid



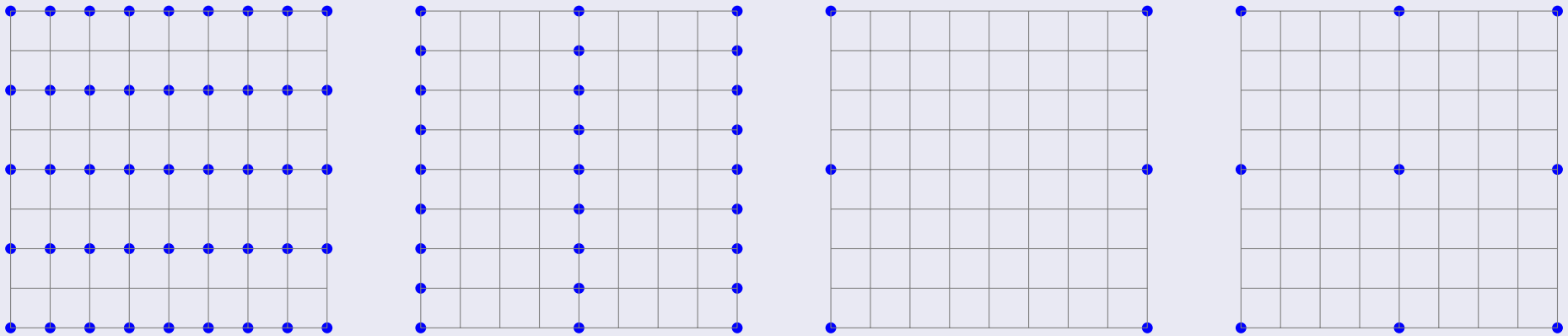
- approximate unknown function $u(x, y)$
- compute only values $u(x_i, y_j)$ on discrete grid points
- interpolate values $u(x, y)$ for other points (x, y)
- regular isotropic grid: $x_i = ih$ and $y_j = jh$

the challenge: curse of dimension

In two dimensions $1/h^2$ grid points, in d dimensions $1/h^d$ grid points but accuracy proportional to h^2

Anisotropic grids

more general regular grids

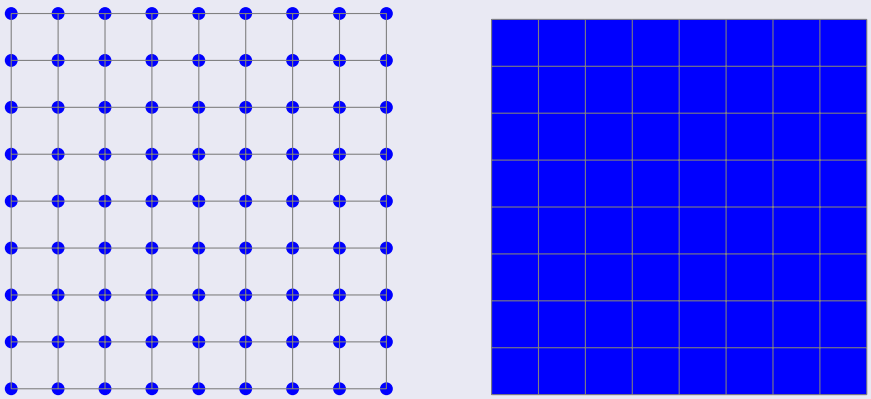


- choose fine grid when $u(x, y)$ has large gradients
- choose coarse grid when $u(x, y)$ is smooth
- gradients may be different in different directions
- choose anisotropic grid when $u(x, y)$ varies differently in different directions

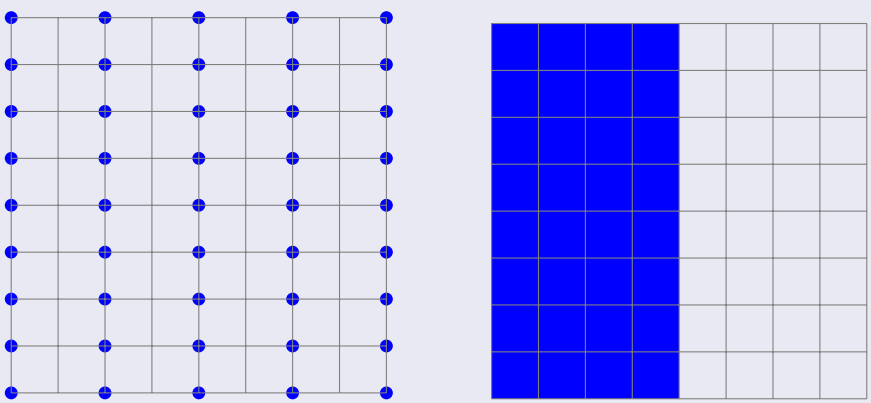
with anisotropic grids one can approximate multi-dimensional $u(x_1, \ldots, x_d)$ if u very smooth in most x_k

Grids and sampling

full grid captures all scales



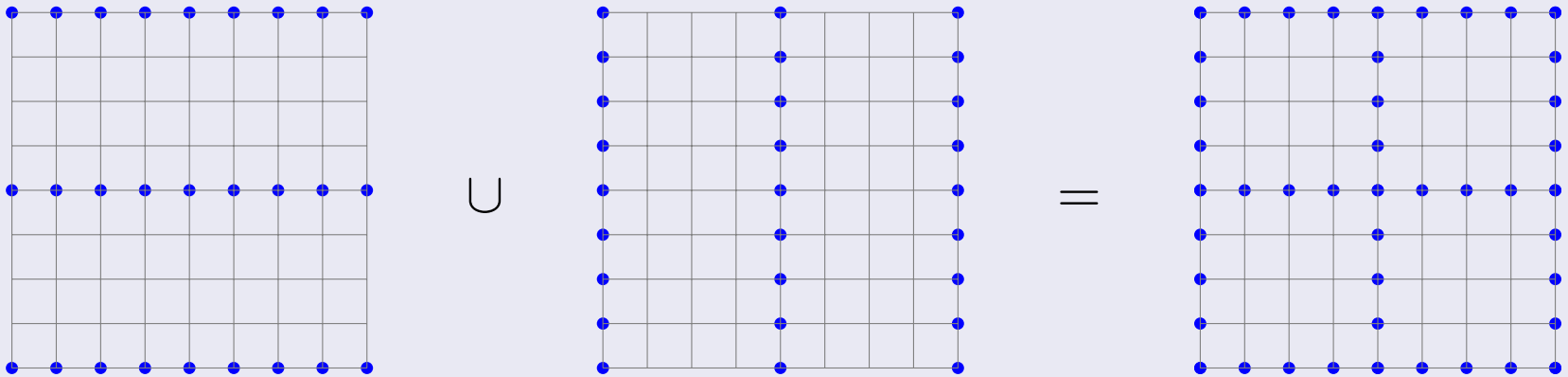
subgrid captures less scales



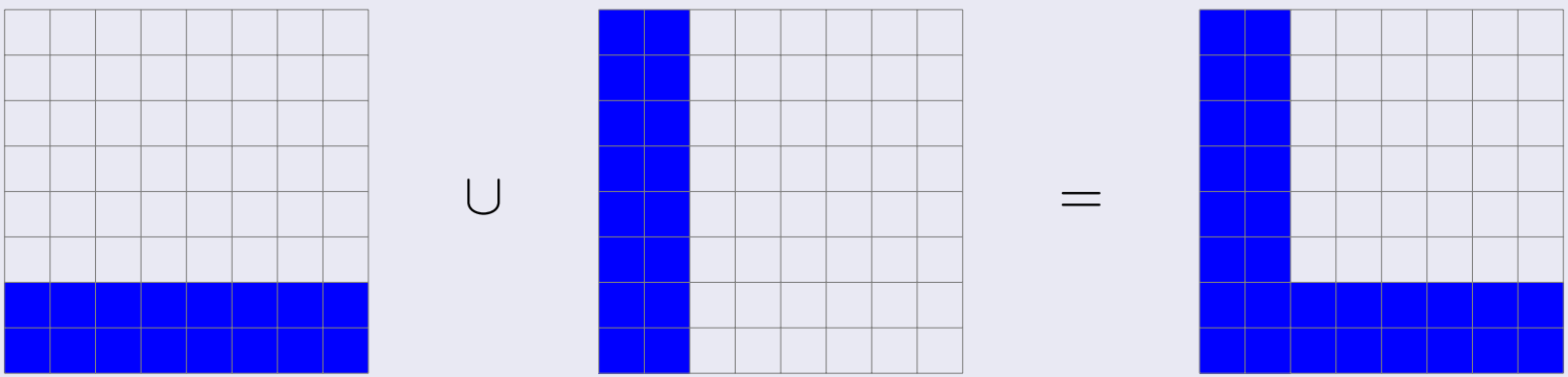
- evaluation of $u(x, y)$ on the grid corresponds to sampling u on the grid points
- sampling on a fine grid captures high frequencies – small scale fluctuations (Nyqvist/Shannon)
- with anisotropic grids one can capture small scales in one dimension and different scales in another

Sparse grid = union of regular anisotropic grids

a simple sparse grid



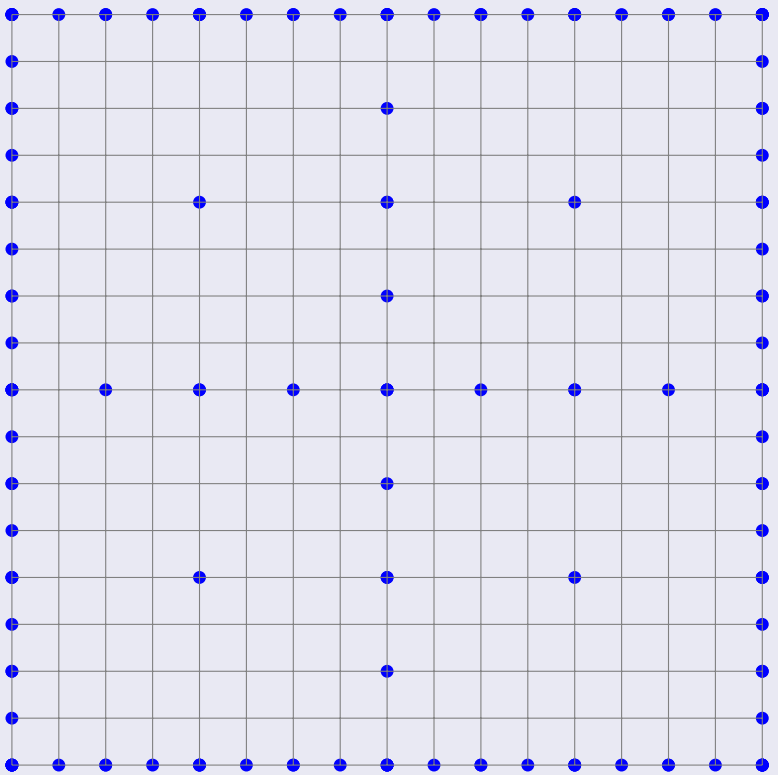
sparse grid in frequency / scale space



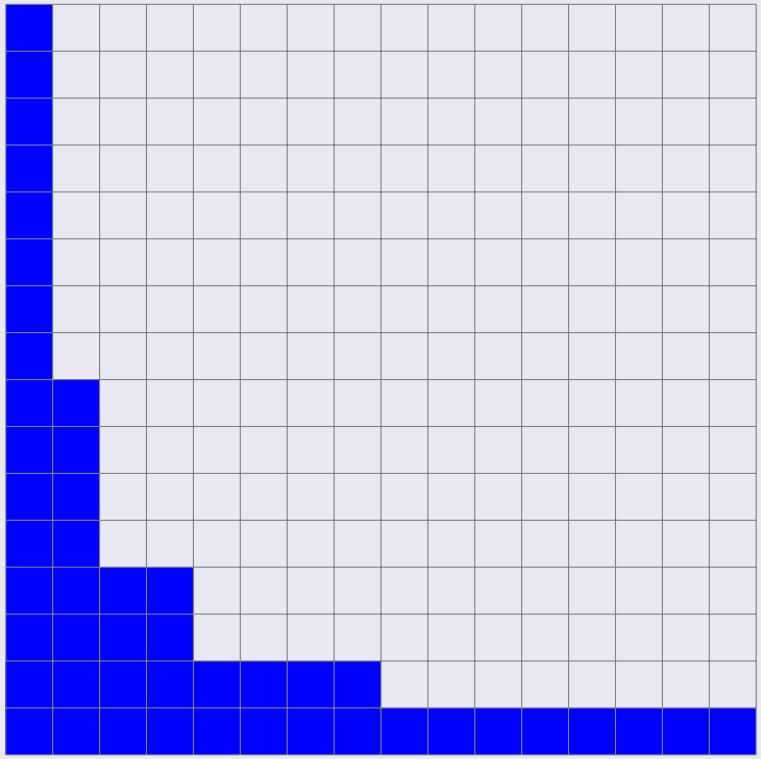
captures fine scales in both dimensions but not joint fine scales

Another sparse grid

sparse grid points



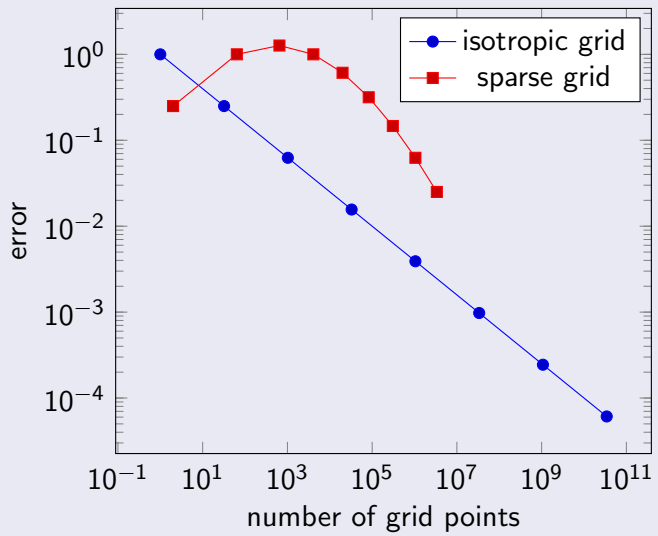
sparse grid frequency diagram



the frequency diagram displays 1/4 of a hyperbolic cross

Sparse grids and the curse of dimension

four dimensional case



- only asymptotic error rates given here
- constants and preasymptotics also depend on dimension
- practical experience: with sparse grids up to 10 dimensions
- Zenger 1991

	number of points	error
regular isotropic grids	h^{-d}	h^2
sparse grids	$h^{-1} \lvert \log_2 h \rvert^{d-1}$	$h^2 \lvert \log_2 h \rvert^{d-1}$

The big plan – dimension independence

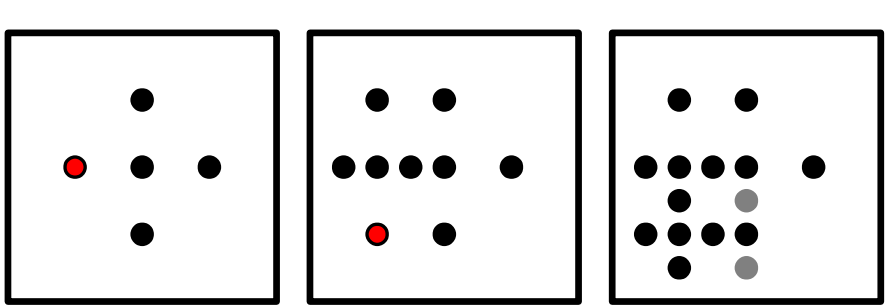
- problem of sparse grids: exponential d-dependence of time and error through
  - factors $|\log_2(h)|^{d-1}$
  - factors of the form C^d
- aim: remove all exponential d-dependence so that
  - error $\sim h^2$
  - time $\sim 1/h$

  as in the case $d = 1$
- ideas:
  - parallel solution on subgrids (see next section) gives $1/h$ time
  - stronger (energy) sparse grids give h^2 error
  - weighted mixed norms and special basis functions to deal with C^d dependence

Spatially adaptive (sparse) grids

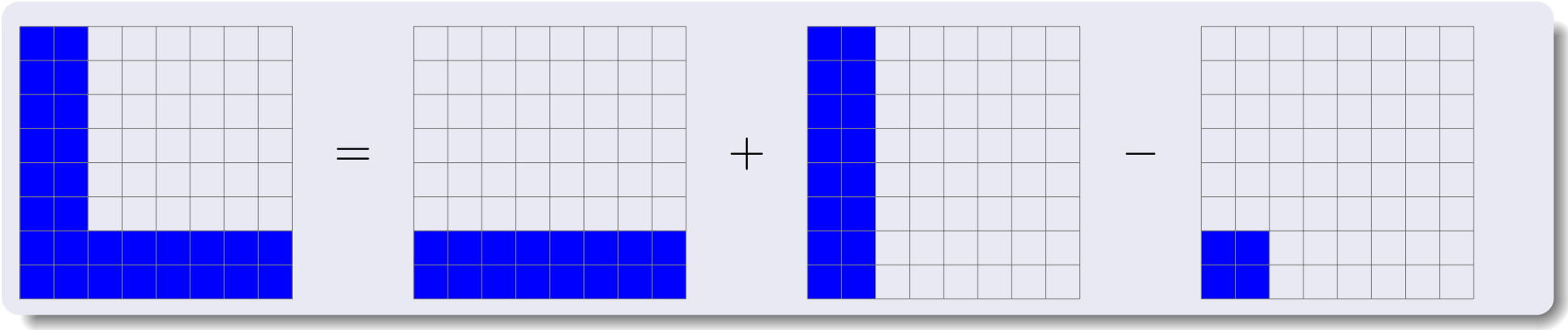
choosing a sparse sub-grid of the sparse grid

- adaptively choose necessary sparse grid points and corresponding (hierarchical) basis functions
- requires an error indicator function
- grid points inserted only where necessary
- acts as extra regularisation (like smoothing) for machine learning applications
- modified basis functions for boundary to remove the C^d
- implemented in SG++ software package by Dirk Pflüger (Universität Stuttgart), 2010



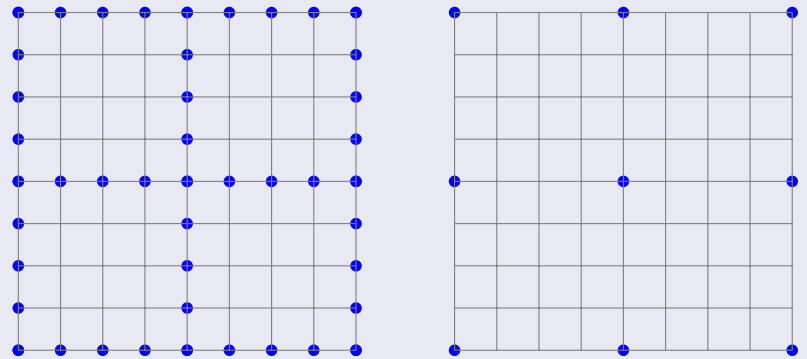
Outline

Combining two regular grids



- solution on combined grid is approximated as a linear combination of the solution on the regular component grids
- the components include the maximal generators and all intersections

union and intersection grids



Weak solutions of boundary value problems

boundary value problem

$$-\Delta u(x) = f(x), \quad x \in \Omega$$
$$u(x) = 0, \quad x \in \partial\Omega$$

weak solution

$$a(u, v) = \langle f, v \rangle, \quad v \in H_0^1(\Omega)$$

where

$$a(u, v) = \int_\Omega \nabla u(x)^T \nabla v(x)\, dx$$
$$\langle f, v \rangle = \int_\Omega f(x) v(x)\, dx$$

approximate solution $u_h \in V_h$

$$a(u_h, v_h) = \langle f, v_h \rangle, \quad v_h \in V_h$$

Approximate solution can be viewed as a projection

$$u_h = P_h u$$

which is orthogonal with respect to the energy norm

Combination approximations

regular grid approximation

- regular grid $G_{\mathbf{h}}$
- function space $V_{\mathbf{h}}$
- Galerkin equations for $u_{\mathbf{h}}$

$$a(u_{\mathbf{h}}, v_{\mathbf{h}}) = \langle f, v_{\mathbf{h}} \rangle$$

for all $v_{\mathbf{h}} \in V_{\mathbf{h}}$

sparse grid approximation

- sparse grid $G_{\mathrm{SG}} = \bigcup_{\mathbf{h}} G_{\mathbf{h}}$
- function space $V_{\mathrm{SG}} = \sum_{\mathbf{h}} V_{\mathbf{h}}$
- Galerkin equations for u_{SG}

$$a(u_{\mathrm{SG}}, v_{\mathrm{SG}}) = \langle f, v_{\mathrm{SG}} \rangle$$

for all $v_{\mathrm{SG}} \in V_{\mathrm{SG}}$

combination technique – where HPC comes in

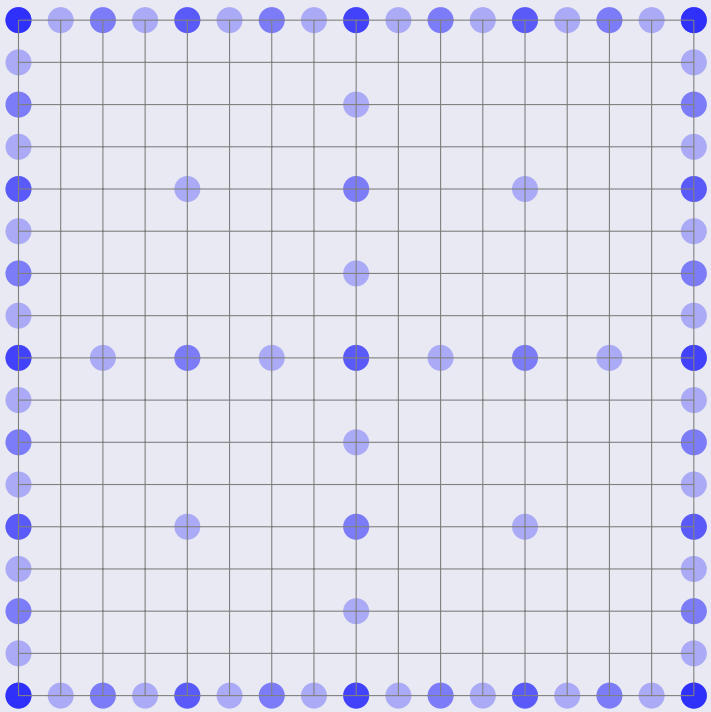
compute all $u_{\mathbf{h}}$ in parallel and combine solutions using parallel reduction:

$$u_C = \sum_{\mathbf{h}} c_{\mathbf{h}} u_{\mathbf{h}}$$

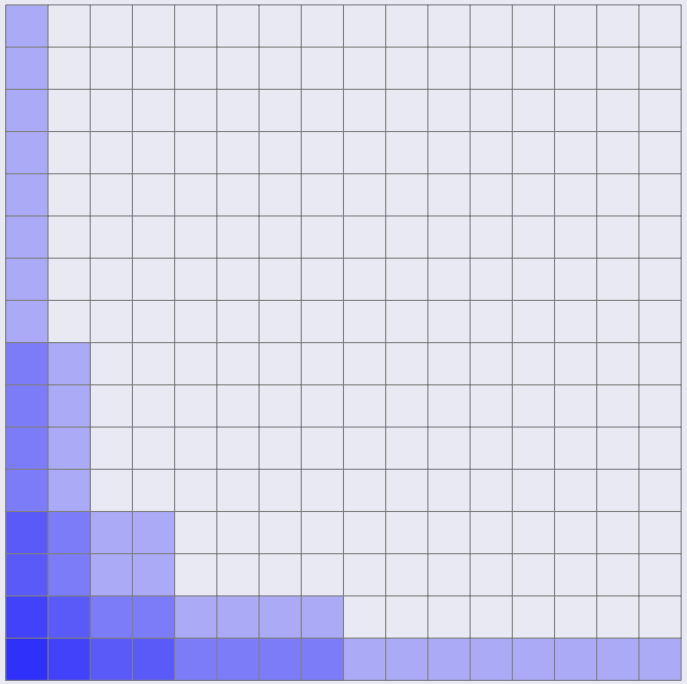
Big question: when is $u_C \approx u_{\mathrm{SG}}$?

Sparse grid combination technique

sparse grid points



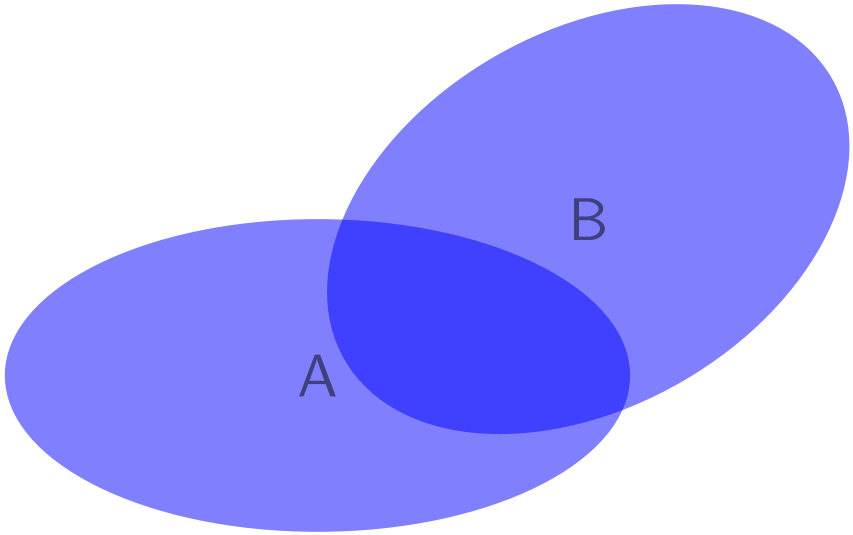
sparse grid frequency diagram



combination formula

$$u_C = u_{1,16} + u_{2,8} + u_{4,4} + u_{8,2} + u_{16,1} - u_{1,8} - u_{2,4} - u_{4,2} - u_{8,1}$$

Inclusion / exclusion principle in combinatorics



for the cardinality of sets

$$|A \cup B| = |A| + |B| - |A \cap B|$$

more general for additive α:

$$\alpha(A \cup B) = \alpha(A) + \alpha(B) - \alpha(A \cap B)$$

Theorem (de Moivre)

If $A_1, \ldots, A_m$ form intersection structure then

$$\alpha\left(\bigcup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} c_i\, \alpha(A_i), \quad \text{for some } c_i \in \mathbb{Z}$$

When the combination approximation is the sparse grid solution

Lemma

- *if the grids* $G_{\mathbf{h}}$ *and the spaces* $V_{\mathbf{h}}$ *form an intersection structure*
- *if the Galerkin projections* $P_{\mathbf{h}}$ *commute, i.e.,*

$$P_{\mathbf{h}} P_{\mathbf{h}'} = P_{\mathbf{h}'} P_{\mathbf{h}}, \quad \text{for all } \mathbf{h}, \mathbf{h}'$$

then

$$u_C = u_{SG}$$

i.e., the combination technique provides the sparse grid solution

Proof.

This is a consequence of the inclusion-exclusion principle as it follows from the commutativity that $P_{\mathbf{h}}$ is additive □

Tensor products – the classical sparse grid

$V_1 \subset V_2 \subset \cdots \subset V_m \subset V$
hierarchy of functions of one variable

classical sparse grid space

$$V_{SG} = \sum_{i+j=n} V_i \otimes V_j$$

tensor product function space

$V_i \otimes V_j$ space of functions generated by products
$u_1 \otimes u_2(x_1, x_2) = u_i(x_1)u_j(x_2)$
where $u_i \in V_i$ and $u_j \in V_j$

combination coefficients

$$c_{ij} = \begin{cases} 1 & i+j = n \\ -1 & i+j = n-1 \\ 0 & \text{else} \end{cases}$$

$V_i \otimes V_j$ form an intersection structure as

$$(V_{i_1} \otimes V_{j_1}) \cap (V_{i_2} \otimes V_{j_2}) = V_{\min(i_1,i_2)} \otimes V_{\min(j_1,j_2)}$$

and combination formula exact if $a(u_1 \otimes u_2, v_1 \otimes v_2) = a(u_1, v_1)a(u_2, v_2)$

[Griebel, Schneider, Zenger 1992]

Extrapolation

assumption: error model

error of approximation in $V_{ij} = V_i \otimes V_j$ is of form

$$e_{ij} = e_i^{(1)} + e_j^{(2)} + r_{ij}$$

is type of ANOVA decomposition for the error

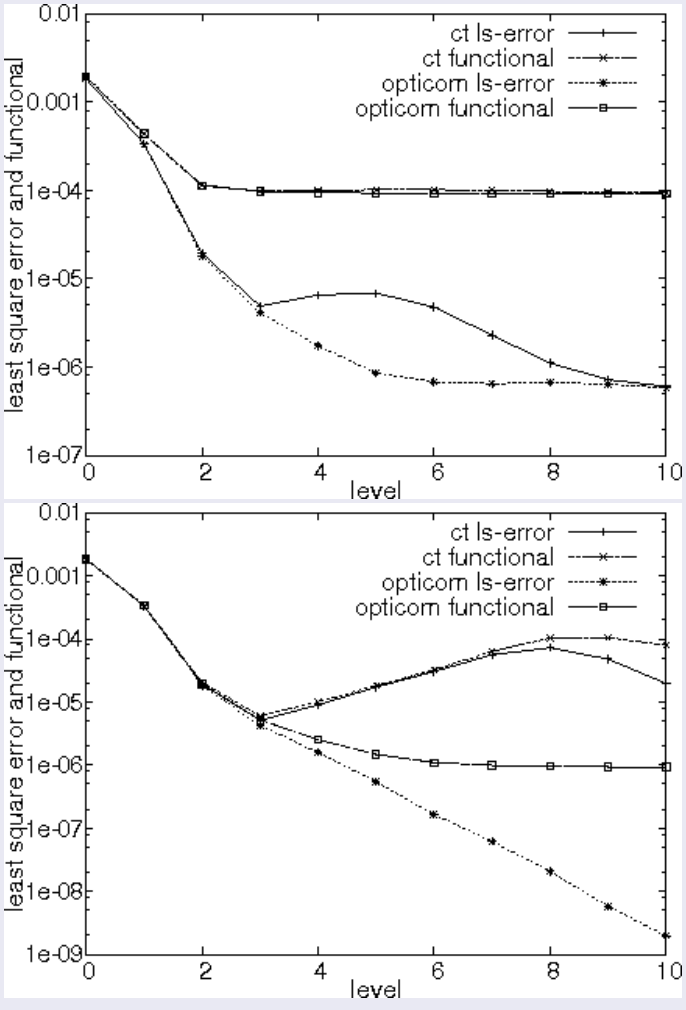
consequence: error of combination technique

$$e_h = \sum_{i+j \leq n} c_{ij} e_{ij} = e_n^{(1)} + e_n^{(2)} + \sum_{i+j \leq n} c_{ij} r_{ij}$$

if last term very small then $e_h \approx e_{n,n}$ i.e., the combination technique approximation using only components in $V_i \otimes V_j$ with $i + j \leq n$ get a similar approximation order as the one in V_{nn}

[Bungartz et al 1994, Pflaum and Zhou 1999, Liem, Lu Shih 1995 (splitting extrapolation)]

Breakdown of the combination technique



regression problem

minimise

$$\frac{1}{M}\sum_{i=1}^{M}(u(x_i)-y_i)^2+\lambda\|\nabla u\|^2$$

with $\lambda = 10^{-4}$ (left) and $\lambda = 10^{-6}$ (right)

combination approximation is not necessarily better for finer grids

[Garcke 2004, H. 2003, H., Garcke, Challis 2007]

Opticom

"Optimal combination technique": choose the coefficients c_i such that J is optimised with

$$J(c_1, \ldots, c_m) = \|u - \sum_{i=1}^{m} c_i u_i\|_E^2$$
$$= \sum_{i,j=1}^{m} c_i c_j \, a(u_i, u_j) - 2 \sum_{i=1}^{m} c_i \|u_i\|_E^2 + \|u\|_E^2$$

normal equations

$$\begin{bmatrix} \|u_1\|_E^2 & a(u_1, u_2) & \cdots & a(u_1, u_m) \\ a(u_2, u_1) & \|u_2\|_E^2 & \cdots & a(u_2, u_m) \\ \vdots & \vdots & \ddots & \vdots \\ a(u_m, u_1) & a(u_m, u_2) & \cdots & \|u_m\|_E^2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} \|u_1\|_E^2 \\ \|u_2\|_E^2 \\ \vdots \\ \|u_m\|_E^2 \end{bmatrix}$$

Solution of the system of order $O(m^3)$ at most but one needs to determine the $O(m^2)$ matrix elements each costing $O(n)$ in machine learning

Opticom is better than the sub-grid solutions

Let $V_{SG} = \sum_{i=1}^{n} V_i \subset V$, $a(\cdot,\cdot)$ be V-elliptic and bounded symmetric bilinear form, $u_i \in V_i$ defined by

$$a(u_i, v_i) = a(u, v_i), \quad \text{for all } v_i \in V_i$$

and c_i be the Opticom coefficients. Then the error in the energy norm satisfies

$$\|u - u_{SG}\|_E \leq \|u - \sum_{i=1}^{n} c_i u_i\|_E \leq \min_{i=1,\ldots,n} \|u - u_i\|_E$$

The standard combination technique does not have this property

The optimality of Opticom

Proposition

Let $V_{SG} = \sum_{i=1}^{n} V_i \subset V$, $a(\cdot,\cdot)$ *be V-elliptic and bounded bilinear form,* $u_i \in V_i$ *defined by*

$$a(u_i, v_i) = a(u, v_i), \quad \text{for all } u_i \in V_i$$

and c_i *be the Opticom coefficients. Then for some* $\kappa > 0$ *one has*

$$\|u - \sum_{i=1}^{n} c_i u_i\|_V \leq \kappa \|u - \sum_{i=1}^{n} \tilde{c}_i u_i\|_V \quad \text{for any } \tilde{c}_i \in \mathbb{R}$$

Proof.

This is a direct application of Céa's Lemma □

Norm reduction with Opticom

Proposition

Let $V_{SG} = \sum_{i=1}^{n} V_i \subset V$ $a(\cdot,\cdot)$ *be V-elliptic and bounded symmetric bilinear form,* $u_i \in V_i$ *defined by*

$$a(u_i, v_i) = a(u, v_i), \quad \text{for all } v_i \in V_i$$

and c_i *be the Opticom coefficients. Then one has for the energy norm defined by* $a(\cdot,\cdot)$ *the bound*

$$\|u - \sum_{i=1}^{n} c_i u_i\|_E \leq \|u\|_E$$

and either $\|f - \sum_{i=1}^{n} c_i u_i\|_E < \|u\|_E$ *or* $f \perp V_h$ *thus* $u_i = 0$, $i, \ldots, n$.

Proof.

$$\|u\|_E^2 = \|u - \sum_{i=1}^{n} c_i u_i\|_E^2 + \|\sum_{i=1}^{n} c_i u_i\|_E^2$$

If the best approximation is zero then u has to be orthogonal to all u_i. As $u - u_i$ is orthogonal to V_h it follows that all the v which are orthogonal to u_i are also orthogonal to u and it follows that u is orthogonal to V_i □

An iterative method

Opticom iterative refinement

$$u^{(0)} = 0$$

$$a(u_i^{(k+1)}, v_i) = a(u - u^{(k)}, v_i), \quad v_i \in V_i$$

$$c_i^{(k+1)} \quad \text{such that} \quad \left\| \sum_{i=1}^{n} c_i^{(k+1)} u_i^{(k+1)} - (u - u^{(k)}) \right\|_E \quad \text{minimal}$$

$$u^{(k+1)} = u^{(k)} + \sum_{i=1}^{n} c_i u_i^{(k+1)}$$

- algorithm converges to the sparse grid solution
- variant of parallel subspace correction [Xu 1992]
- also combine with Newton [Griebel, H. 2010]

the machine learning problem

given data $x_1, \dots x_N$ in $\mathbb{R}^d$ find density $f(x)$ such that

$$f \approx \frac{1}{N} \sum_{k=1}^{N} \delta_{x_k}$$

[Tapia & Thompson 1978, Silverman 1986, Scott 1992]

from data smoothing to sums of simpler approximations

- function approximation: RBF $\Longrightarrow$ sums of products

$$f(x) = \sum_{i=1}^{N} c_i \kappa(x - x_i) \quad \Longrightarrow \quad f(x) = \sum_{k=1}^{K} c_k \prod_{j=1}^{K} f_{j,k}(\xi_j)$$

  where $x = (\xi_1, \ldots, \xi_d)$ and x_i are data points

- density estimation: kernel density estimators $\Longrightarrow$ mixture models

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \kappa(x - x_i) \quad \Longrightarrow \quad f(x) = \sum_{k=1}^{K} \pi_k \, N(x \mid \mu_k, C_k)$$

[McLachlan and Peel, 2000]

Outline

Bayesian inference

- given data and models
  - data, e.g. $x = (x_1, \ldots, x_N)$ typically $x \in \mathbb{R}^{Nd}$
  - likelihood of data: $p(x \mid z)$
  - prior of parameters z: $p(z)$ – models "reasonable assumptions" about z
- Bayes' rule – how to adapt $p(z)$ in light of the evidence

$$p(z \mid x) = \frac{p(x \mid z)p(z)}{p(x)} \quad \text{where} \quad p(x) = \int p(x \mid z)p(z)\, dz$$

  $p(x)$ is $p_X(X = x)$, $p(z \mid x)$ is $p_{X|Z}(X = x \mid Z = z)$ etc
- what to do with the posterior
  - expectations $E(Y) = \int y p(y \mid z)p(z \mid x)\, dx\, dy$
  - probability distributions $p(y) = \int p(y \mid z)p(z \mid x)\, dy$
  - maximum $z_{\max} = \operatorname{argmax}_z p(z \mid x) = \operatorname{argmax}_z p(x \mid z)p(z)$
- tractability

  the computation of $p(x), p(y)$ and $E(y)$ require in general highdimensional integrals $\Rightarrow$ approximation

density estimation

- data: $x = (x_1, \ldots, x_N)$ drawn randomly from some unknown probability distribution
- probability density model: $f(x_k) = p(x_k \mid u) = \exp(u(x_k) - \gamma(u))$ where $\gamma(u)$ is such that $\int p(x_k \mid u)\, dx_k = 1$
- estimation problem: for given data $x_1, \ldots, x_N$ find $\hat{u}(\hat{x})$ such that $p(x_k \mid \hat{u})$ approximates underlying density
- likelihood:

$$p(x \mid u) = \exp\left(\sum_{i=1}^{n} u(x_i) - n\,\gamma(u)\right)$$

  choose $\hat{u}$ such that likelihood large
- parametric case: maximum likelihood method
- problem underdetermined in nonparametric case

MAP with Gaussian process priors

- prior for u: Gaussian probability measure ν over space of functions = Gaussian process prior – we consider covariance $C = C_1 \otimes \cdots \otimes C_d$
- posterior based on likelihood $\rho(u) = p(x \mid u)$:

$$d\mu = \rho \, d\nu$$

  is a well defined measure if $\rho \in L_1(\nu)$
- maximum a-posteriori (MAP) method: estimate u as mode of posterior
- Laplace approximation of posterior: Gaussian process with expectation u

a variational problem

- characterisation of mode u of μ:

$$\rho(u) \geq \frac{d\lambda_v}{d\lambda}(u)\,\rho(u+v), \quad \text{for all } v \in H$$

  where $\lambda_v(A) = \lambda(v + A)$

- this leads to minimisation of functional

$$j(u) = \frac{1}{n}\|u\|^2_{CM} - \frac{1}{n}\sum_{i=1}^{n} u(x_i) + \log \int_X \exp(u(x))\,dx$$

  where $\|\cdot\|_{CM}$ is Cameron-Martin norm defined by prior

[H. 2007, Griebel, H. 2010]

Newton-Galerkin Opticom method

- Newton Galerkin $u_{n+1} = u_n + \Delta u_n, \quad \Delta u_n$ minimises

$$J(\Delta u) = \frac{1}{2} H_{u_n}(\Delta u, \Delta u) + (F(u_n), \Delta u)_H$$

- Sparse grid space $V = \sum_j V^{(j)}$
- Sparse grid combination technique $\Delta u_n = \sum_{j=1}^{k} c_j \Delta u_n^{(i)}$ where components $\Delta u_n^{(j)}$ minimise $J(\Delta u)$ over $V^{(j)}$
- Opticom method: choose combination coefficients c_j to minimise $J(\sum_{j=1}^{k} c_j \Delta u_n^{(j)}) \Rightarrow$ descent method, converges to sparse grid solution, not some combination approximation
- inexact Newton method [Deuflhard, Weiser 1996, Deuflhard 2004] alternative: nonlinear additive Schwarz [Dryja, Hackbusch 1997]

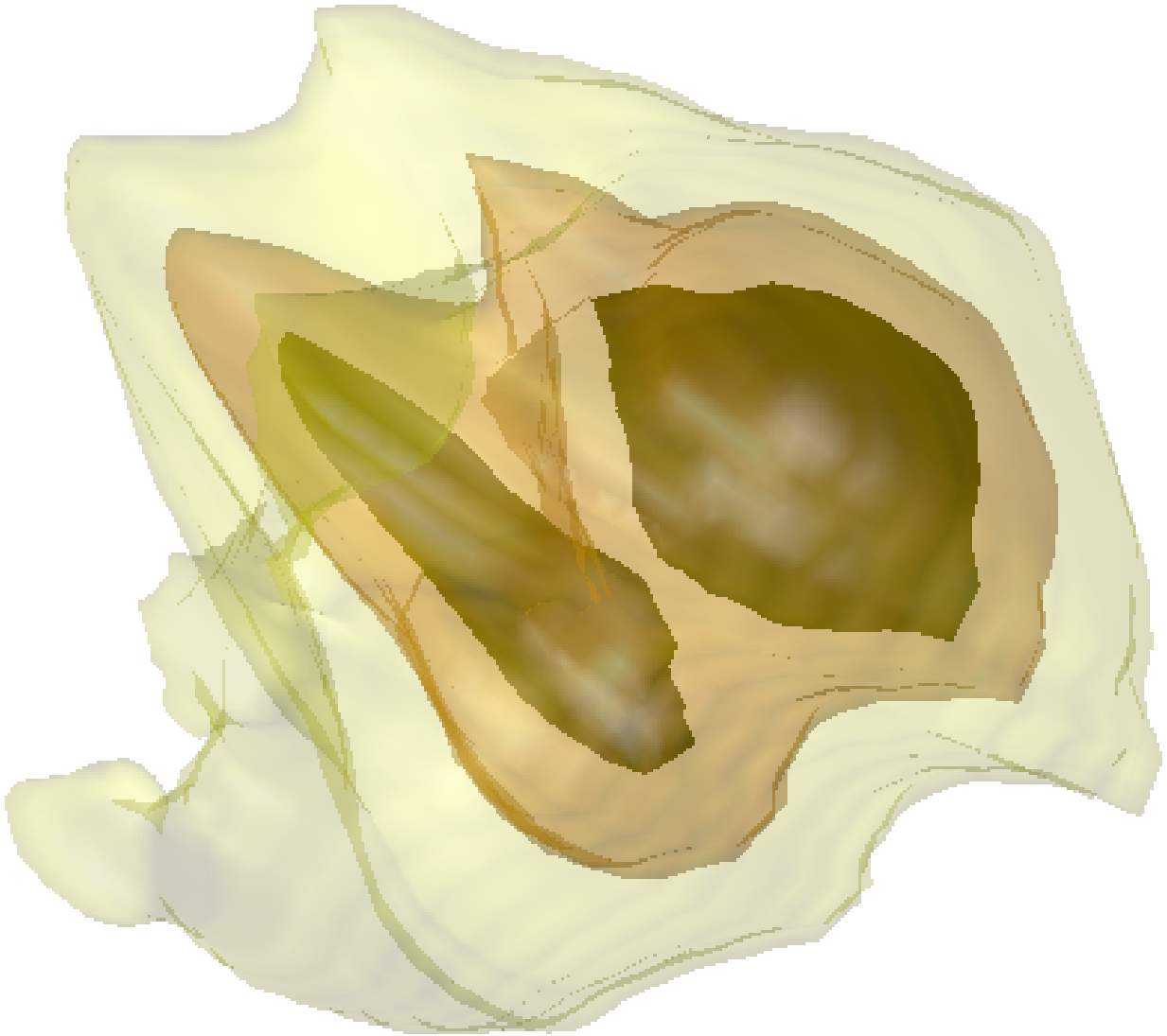
[H., Griebel 2007]

errors of sparse grid approximation for 2D case

approximation of the normal distribution: maximum likelihood projection and our estimator

l	$e^{(1)}_{1,l}$	$\frac{e^{(1)}_{1,l}}{e^{(1)}_{1,l+1}}$	$e^{(1)}_{2,l}$	$\frac{e^{(1)}_{2,l}}{e^{(1)}_{2,l+1}}$	$e^{(3)}_{1,l}$	$\frac{e^{(3)}_{1,l}}{e^{(3)}_{1,l+1}}$	$e^{(3)}_{2,l}$	$\frac{e^{(3)}_{2,l}}{e^{(3)}_{2,l+1}}$
1	1.42e+00	–	–	–	8.05e-02	–	1.33e-01	–
2	3.12e-01	4.55e+00	1.16e+00	–	7.97e-02	1.01e+00	1.27e-01	1.04e+00
3	7.37e-02	4.23e+00	2.44e-01	4.75e+00	3.11e-02	2.56e+00	6.39e-02	1.99e+00
4	1.94e-02	3.81e+00	6.34e-02	3.85e+00	9.63e-03	3.23e+00	1.89e-02	3.38e+00
5	4.92e-03	3.93e+00	1.60e-02	3.96e+00	3.13e-03	3.08e+00	6.14e-03	3.08e+00
6	1.23e-03	4.00e+00	4.17e-03	3.83e+00	8.04e-04	3.89e+00	1.72e-03	3.56e+00

3D density



[Griebel, H. 2010]

Outline

two variational problems

method	Ritz-Galerkin for u_h	variational Bayes for q
error	energy norm $\|u - u_h\|_E = \sqrt{a(u-u_h, u-u_h)}$	KL-divergence $KL(q \,\|\| \, p(\cdot \mid x)) = \int q(z) \log\left(\frac{p(z\|x)}{q(z)}\right) dz$
optimisation problem	minimise $J(u_h) = \frac{1}{2}a(u_h, u_h) - \langle f, u_h \rangle$	maximise $\mathcal{L}(q) = \int q(z) \log\left(\frac{p(x,z)}{q(z)}\right) dz$
property	V-ellipticity	convexity

right column: use

$$KL(q \,\|\| \, p(\cdot \mid x)) - \mathcal{L}(q) = \log p(x)$$

data: f on left and $p(x, z)$ on right
[Beal 2003, MacKay 2003, Bishop 2006]

fix point characterisation of best product

Proposition (characterisation)

If $q = \prod_{i=1}^{m} q_j$ *is best approximant then*

$$u_j(z_j) = \int \log(p(x,z)) \prod_{i \neq j} q_i(z_i)\, dz_i$$

$$q_j(z_j) = \frac{\exp(u_j(z_j))}{\int \exp(u_j(w_j))\, dw_j}$$

Proof.

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_{i=1}^{m} q_i(z_i) \left(\log(p(x,z) - \sum_{i=1}^{m} \log q_i(z_i) \right) dz_i \\ &= \int (u_j(z_j) - \log q_j(z_j)\,)\, q_j(z_j)\, dz_j + F(\{q_i\}_{i \neq j})\end{aligned}$$

iterative solver

start with $q_1^{(0)}, \ldots, q_m^{(0)}$

$n = 1, 2, \ldots$

$j = 1, \ldots, m$

$$u_j^{(n)}(z_j) = \int \log p(x, z) \prod_{i=1}^{j-1} q_i^{(n)}(z_i)\, dz_i \prod_{i=j+1}^{m} q_i^{(n-1)}(z_i)\, dz_i$$

$$q_j^{(n)}(z_j) = \frac{\exp u_j^{(n)}(z_j)}{\int \exp u_j^{(n)}(w_j)\, dw_i}$$

convergence as KL-divergence convex in u_j

mixture models

- probability distribution for n-th observation

$$p(x_n \mid u) = \sum_{k=1}^{K} \pi_k \, p_k(x_n \mid u_k)$$

components p_k are separable $\Rightarrow$ sums of separable functions

- problem: estimation of p given data $x = (x_1, \ldots, x_N)$
- difficulty: while each p_k has product structure which is adapted to KL minimisation the sum is a problem
- idea: introduce latent (or hidden) variables $z_1, \ldots, z_N$ which are binary vectors indicating the class k of observation n thus

$$p(x_n \mid u) = \sum_{k=1}^{K} p(x_n \mid u_k, z_n = e_k) \, p(z_n = e_k) = \sum_{z_n} p(x_n, z_n \mid u)$$

is interpreted as a marginal distribution and $u = (u_1, \ldots, u_K)$

likelihood, priors and posterior

- likelihood of $x = (x_1, \ldots, x_n)$

$$p(x \mid z, u) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(x_n \mid u_k)^{z_{nk}} \quad \text{– sum disappeared}$$

- prior for latent variables z_n

$$p(z \mid \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

- priors for π and u: $p(\pi)$ and $p(u)$
- posterior distribution $p(z, \pi, u \mid x) = p(x, z, \pi, u)/p(x)$ where

$$p(x, z, \pi, u) = p(x \mid z, u)\, p(z \mid \pi)\, p(\pi)\, p(u)$$

variational Bayes for mixture models

- aim: $q(z, \pi, u)$ approximation of posterior $p(z, \pi, u \mid x)$
- product Ansatz: $q(z, \pi, u) = q(z)q(\pi, u)$
- fix point formulation from minimal KL divergence

$$q(z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

$$q(\pi, u) = C\, p(\pi) \prod_{k=1}^{K} p(u_k) \prod_{n=1}^{N} \prod_{k=1}^{K} (\pi_k\, p(x_n \mid u_k))^{r_{nk}}$$

where $r_{nk} = \rho_{nk} / (\sum_k \rho_{nk})$ and $\rho_{nk} = E_\pi[\log \pi_k] + E_u[\log p(x_n \mid u)]$ and

- approximate $p(x_n \mid u_k)$ as product to get tractability

Conclusions

- with the wide availability of new computational resources high performance computing ideas now enter mainstream computational science
- HPC is getting increasingly complex with a shift towards new problems and approaches characterised by the "multi-challenge"
- an increasingly important challenge originates from the multi and high dimensionality of many models in physics, chemistry, biology, statistics and engineering
- new theory and algorithms are needed
- sparse grid combination technique deals with dimensionality and is ideally suited for HPC
- the Opticom method is able to overcome a stability issue of the original combination technique
- next: high-dimensional inverse problems?