Data Mining with Spare Grids

Seminar: Computational Aspects of Machine Learning

Sebastian Kreisel
October 17, 2015

Technische Universität München

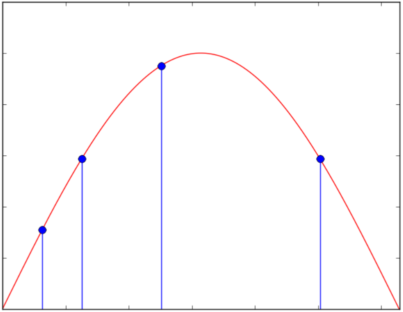
Overview

- Motivation for Sparse Grids
- Sparse Grids: Basics
- Sparse Grids: Machine Learning
- Examples with Data Sets
- Parallelization and Implementation

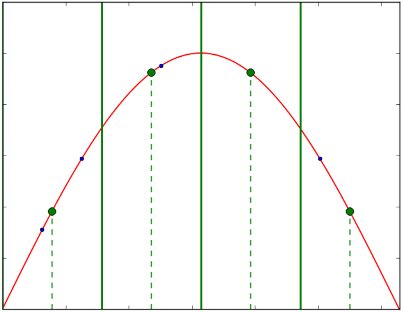
Motivation for Sparse Grids

Grid based approaches in ML

- Discretizes the space into a grid
- Basis–functions around grid points, not data points



Point based



Grid based

Motivation for Sparse Grids

Suitable for

- Big datasets
- Easily/automatically classifiable data
- Medical, seismic, commercial data

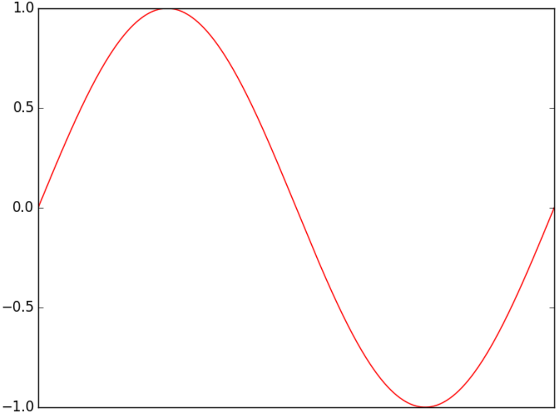
Curse of dimensionality

- The volume of a space is exponential in it's dimensions
- The amount of training data required becomes unmanageable
  - because of lacking computational/storage capacities
  - because data acquisition is expensive
- Becomes relevant for $d > 3$
- **Applies to full-grid discretization**

Full Grid Discretization

1

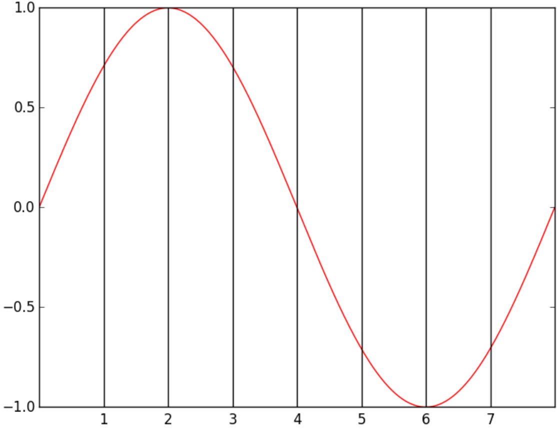
1. Function to discretize



$f(x) = \sin(2\pi \cdot x)$

Full Grid Discretization

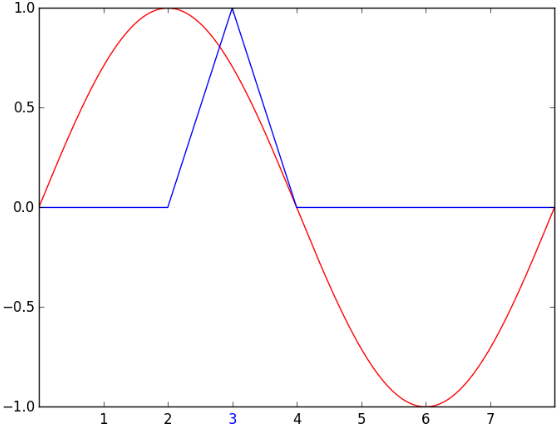
2. Full, regular grid



7 grid points $i \in \{1..7\}$

Full Grid Discretization

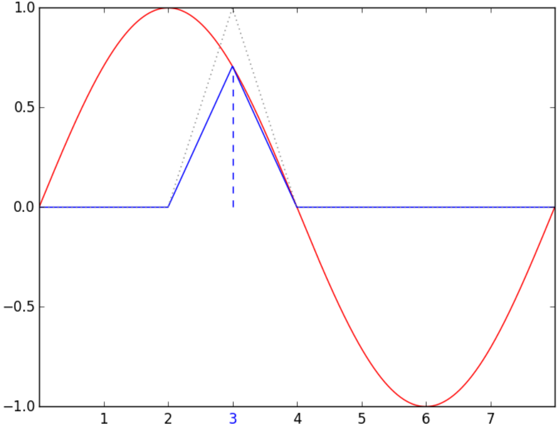
3. Basis function (standard hat function)



$\phi(x-3)$ with $\phi(x) = \max(|1-x|, 0)$

Full Grid Discretization

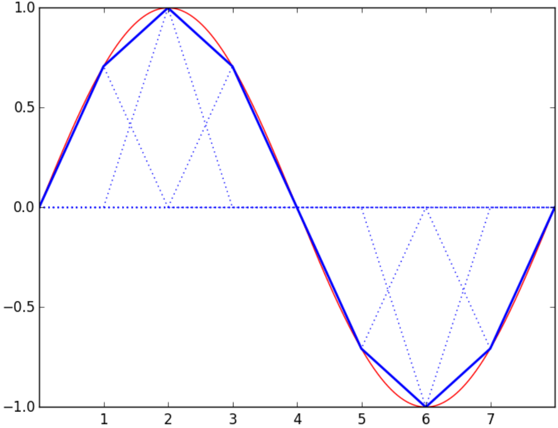
4. Coefficient α (Surplus)



$\alpha_3 \cdot \phi_3(x)$

Full Grid Discretization

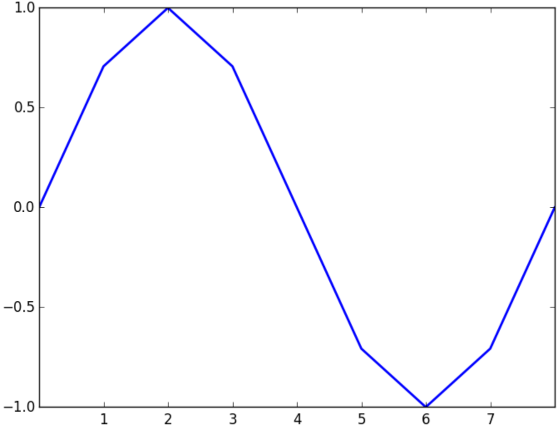
Sum over all basis functions



$$f(x) \approx \hat{f}(x) = \sum_i \alpha_i \phi_i$$

Full Grid Discretization

Sum over all basis functions



$f(x) \approx \hat{f}(x) = \sum_i \alpha_i \phi_i$

Full Grid Discretization

Full grid discretization in one dimension

1. A function $f(x)$ to discretize
2. Grid points indexed by $i \in \{1, 2, \dots\}$
3. Basis functions; i.e. hat function $\phi(x) = \max(1 - |x|, 0)$
4. Coefficients α_i (Surpluses)

$$f(x) \approx \hat{f}(x) = \sum_i \alpha_i \phi_i(x)$$

Full Grid Discretization

$d > 1$ dimensions

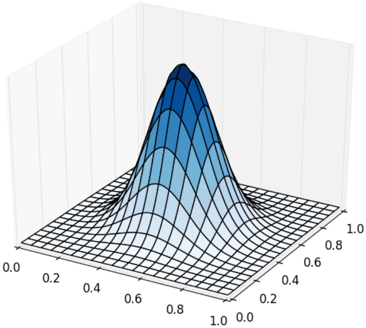
- Grid points as d–tuple, i.e. $(1, 3, 1)$
- Tensor product over one dimensional basis functions

$$\phi_i(\vec{x}) = \prod_{j=1}^{d} \phi_{i,j}(x_j)$$

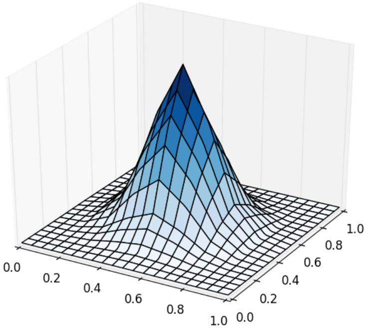
$$f(\vec{x}) \approx \hat{f}(\vec{x}) = \sum_i \alpha_i \phi_i(\vec{x})$$

Full Grid Discretization

Full grid with $d = 2$



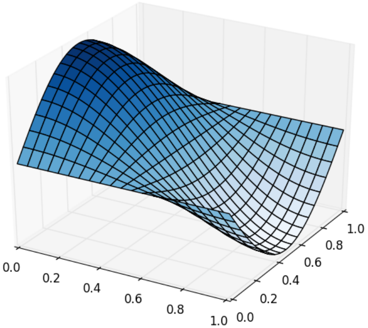
$f(x) = \mathcal{N}(\vec{x})$



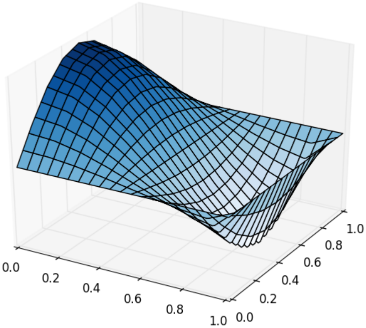
Full grid discretization $\hat{f}(\vec{x})$

Full Grid Discretization

Full grid with $d = 2$



$f(x) = \sin(x_1) \cdot \cos(x_2)$



Full grid discretization $\hat{f}(\vec{x})$

Full Grid Discretization

Summary

- Grid points $i \in \{1, 2, \dots, N\}^d$ defining $\phi_i(x)$
- For $d > 1$: **product** of 1D basis functions:

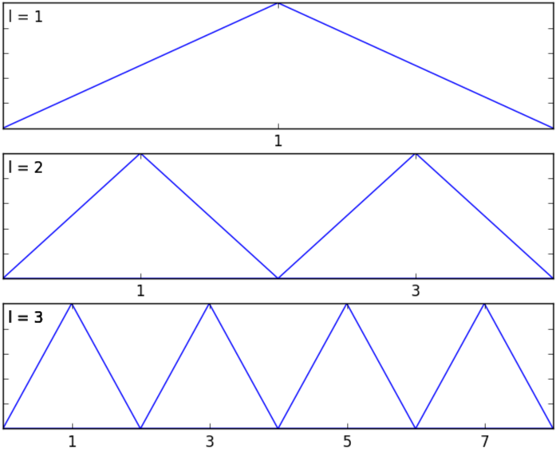
$$\phi_i(\vec{x}) = \prod_{j=1}^{d} \phi_{i,j}(x_j)$$

- **Sum** over all weighted basis functions:

$$\hat{f}(\vec{x}) = \sum_{i=1}^{N} \alpha_i \phi_i(\vec{x})$$

Sparse Grids – Basics

Hierarchical Basis



Sparse Grids – Basics

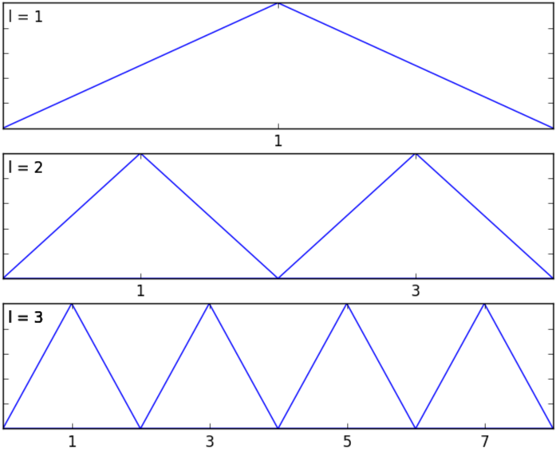
Hierarchical basis (vs nodal basis)

- Grouping grid points into levels $l \in \{1, 2, 3 \ldots, n\}$
- Basis function by index **and** level: $\phi_{l,i}(x)$

$$\hat{f}(x) = \sum_{l \le n, i \in G_l} \alpha_{l,i} \cdot \phi_{l,i}(x)$$

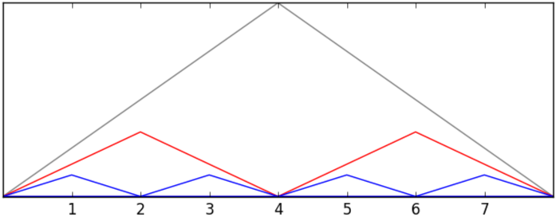
Sparse Grids – Basics

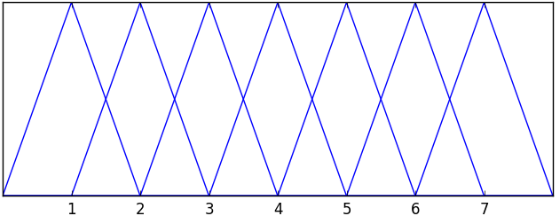
Hierarchical Basis



Sparse Grids – Basics

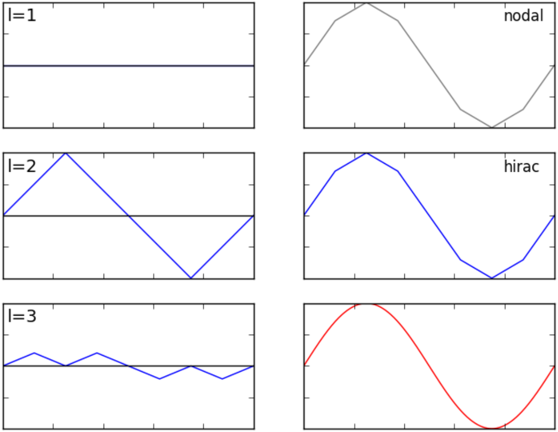
Hierarchical vs. nodal basis





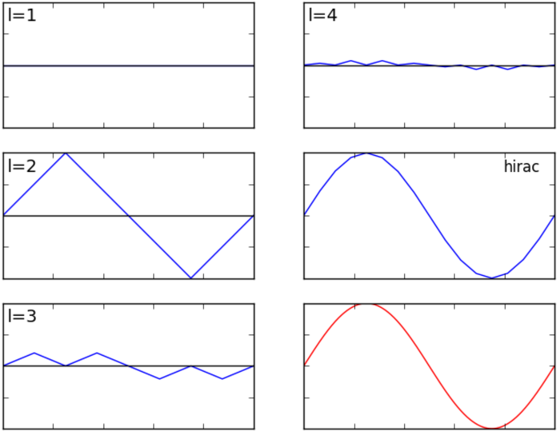
Sparse Grids – Basics

Full grid discretization: Hierarchical basis



Sparse Grids – Basics

Full grid discretization: Hierarchical basis



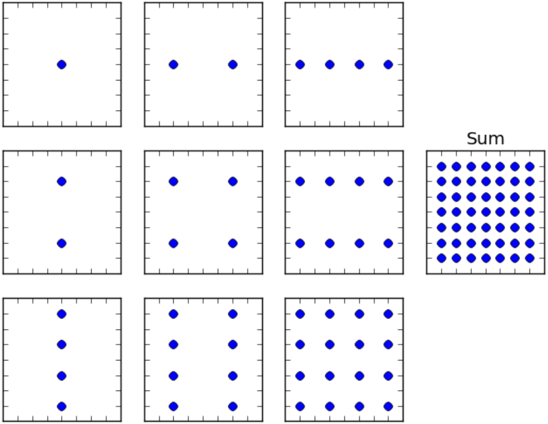
Sparse Grids – Basics

Basis function subspaces

- Combination of levels and dimensions
- Notion of hierarchical subspaces
- Defined by the levels of detail in all dimensions $(l_x, l_y, ...)$

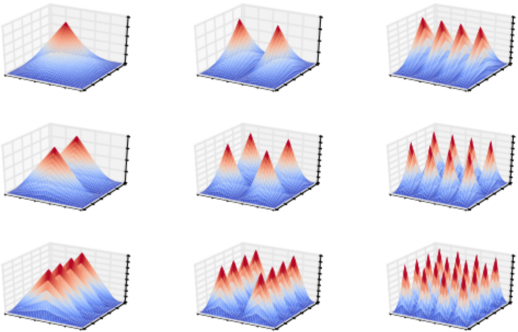
Sparse Grids – Basics

Hierarchical grid points



Sparse Grids – Basics

Hierarchical subspaces



Sparse Grids – Basics

Sparse grid – Changes

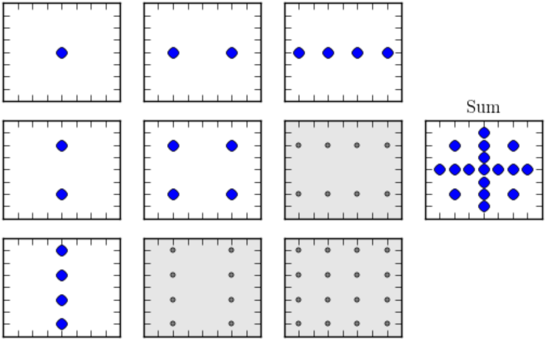
- Throwing away certain subspaces
- Finding those is a *a-priori* solvable optimization problem
- The resulting grid is now **sparse**

Profit

- Reducing the computational effort “a lot”
- Maintaining “high” accuracy

Sparse Grids – Basics

A *sparse* grid



Sparse Grids – Basics

Boundary and smoothness

- Boundaries need special treatment
- The function needs to be sufficiently smooth
  $D^2 f$ needs to be bounded

Adaptivity

- *A-posteriori* modifications to better fit the function
- Picking a single grid point and adding level of detail around it
- Prone to overfitting and huge computational effort

Sparse Grids – Basics

Summary

- Hierarchical basis through grouping grid points into levels
- Creating "subspaces" through combination of levels in dimensions
- Selecting and combining subspaces

To keep in mind

- Smoothness requirement for $f(x)$
- Boundary treatment
- Accuracy–effort trade-off
- Adaptivity options (*a-posteriori*)

Sparse Grids – Data Mining

Machine learning tasks

- Classification
- Regression

Training-data

$$X = \{x^{(i)} \mid x^{(i)} \in [0,1]^d\}_{i=1}^{M}$$

$$Y = \{y^{(i)} \mid y^{(i)} \in \mathbb{R}\}_{i=1}^{M}$$

Sparse Grids – Data Mining

Least squares

$$\hat{c} = \arg\min_f \left(\frac{1}{M} \sum_{i=1}^{M} (y_i - f(x_i))^2 \; + \; \lambda ||\nabla f|| \right)$$

Sparse Grids – Data Mining

Sparse grid setting

- Do least squares in a sparse grid setting ("space")
- Discretize $\hat{c}$ using a sparse grid

Least squares: Sparse gird discretized

$$\hat{c} = \arg\min_{\alpha} \left(\frac{1}{M} \sum_{i=1}^{M} \left(y_i - \sum_j \alpha_j \phi_j(x_i) \right)^2 + \lambda \sum_j \alpha_j^2 \right)$$

Sparse Grids – Data Mining

Matrix formulation

$$\left(\frac{1}{M}BB^T + \lambda I\right)\alpha = \frac{1}{M}By$$

$$\mathbb{R}^{N\times M} \ni B = \begin{bmatrix} \phi_1(x^{(1)}) & \dots & \phi_1(x^{(M)}) \\ \vdots & \ddots & \vdots \\ \phi_N(x^{(1)}) & \dots & \phi_N(x^{(M)}) \end{bmatrix}$$

Sparse Grids – Data Mining

Observations

- $BB^T \in \mathbb{R}^{N \times N}$ where $N =$ number of grid points
- Number of freedom not dependent on M
- Linear scaling in M