Project C4
Regression approaches for large-scale high-dimensional data
Prof. Dr. Katja Ickstadt, Prof. Dr. Christian Sohler

Goals: develop highly efficient regression approaches for massive data
obtain a common and unified understanding of sketching and sampling methods

Three cornerstones to develop unified solutions with guarantees for regression models

Problem

Geometric Relaxations
- Overlap of data
- Small perturbations
- Avoid worst-case
- Complexity parameter
- Captures geometric structure

Algorithmic Techniques
- Streaming and distributed computing
- Sketching and sampling
- Random projections
- Sensitivity framework quantifies importance of each point by its contribution in the worst case
- Sensitivity framework quantifies importance of each point by its contribution in the worst case

Statistical Modelling
- Exponential family
  \[ f(x \mid \theta) = h(x)g(\theta) \exp(\eta(\theta) \cdot t(x)) \]
- Common form of distributions
- Prior as relaxation technique
- Prior as penalisation

WP1: Generalised Linear Models
- Common loss function over the exponential family
  \[ \ell(\beta \mid X, Y) = \sum_{i=1}^{n} g(x_i \beta) - Y_i X \beta + C \]
- Sensitivity sampling
  \[ \bar{g} = \sup_{\beta \in \mathbb{R}^d} \ell(\beta \mid X, Y) \]
- Bound sensitivities using convex layers and Tukey depth
- Smoothed analysis

WP2: Bayesian Generalised Linear Models
Bayesian Sensitivities
- Prior as relaxation admits sublinear sampling scheme
- Bound error in expectation, \((1 + \varepsilon)\)-guarantee in expectation
- Prior models \(\ell_1\) and \(\ell_2\) penalization
- Graph models and generalised linear models
- Evaluate GLM coresets in a Bayesian setting
- Develop coresets for sum-product-networks

WP3: Importance Measures
Aims:
- Impact on sampling
- Variable selection
- Study properties and relationships
Techniques:
- Leverage scores, cross-leverage scores (LS)
- Jacobian leverage (JL)
- Tukey depth, Tukey median (TD)
- Convex-layer depth (CL)
- Sensitivity (S)
- Variable importance measures (VIM)

WP4: Mixtures of Normal Distributions
Motivation:
- Regression on normal mixture models
- Approximate arbitrary continuous distributions
- Handle outliers
Challenges:
- Allocation to mixture components and regression problem
- Trade-off: number of components ↔ model accuracy
- Relaxed the allocation problem through oracle techniques