technische universität dortmund

SFB 876 Providing Information by Resource-Constrained Data Analysis

Project A2

Algorithmic aspects of learning methods in embedded systems

Prof. Dr. Christian Sohler, Prof. Dr. Jens Teubner

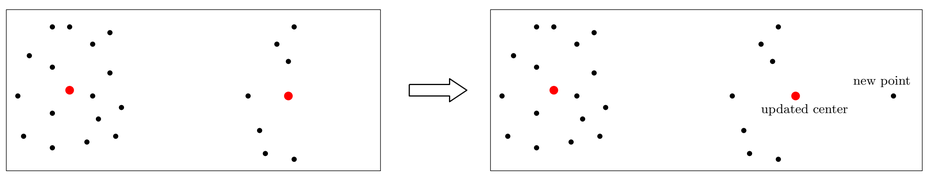
Problem

Big Data and Limited Resources

How to **learn** on **embedded systems**?

- **Large** and demanding computations
- Bounded and/or **distributed** resources
- Not possible to store whole data
- Data is **dynamic** – insertions and deletions allowed (e.g. in a database)

How much information should we keep?



Research question:

- Given dynamic data stream of points from space $\{1, \ldots, \Delta\}^d$, how much time and space is needed to keep track of k-median clustering of these points?

Energy-Efficient Computations

Energy: Most critical factor for computer systems

- Chip design (e.g. dark silicon), data centers

CPU power consumption

$$P_{dyn} = C_L V_{dd}^2 f$$

- Frequency is critical because of strong effect on power consumption

State of the Art

- Energy management is done in hardware (Thermal Design Power, dark silicon)
- Algorithm design usually aims at performance → Low static power consumption

Can software help to lower dynamic power consumption?

Methodology

Data Reduction

Find **coresets**, **sketches**, or other **summary** processes, that enable (e.g.)

- $(1+\varepsilon)$-approximation algorithms for k-center/k-median/k-means clustering in dynamic data stream environment.
- a distributed algorithm for logistic regression problem, under some input niceness assumption.

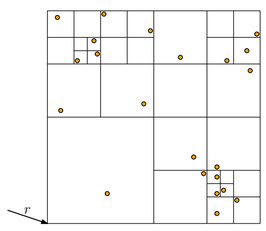
C4

Previously known:

- Coresets for k-median clustering in dynamic data streams required space exponential in dimension d of the input space.

Techniques:

- 2^d**-ary tree** structure – store the input points within tree cells.
- **Randomisation** – random shift yields results *with high probability*;
- **Sampling** – uniformly from highly populated cells ⇒ the cost does not become too large.



Highly populated cells in a shifted 2^d-ary tree

Coreset for k-median clustering in dynamic data streams
[Braverman et al., ICML'17]

- Time and space polynomial in input dimension

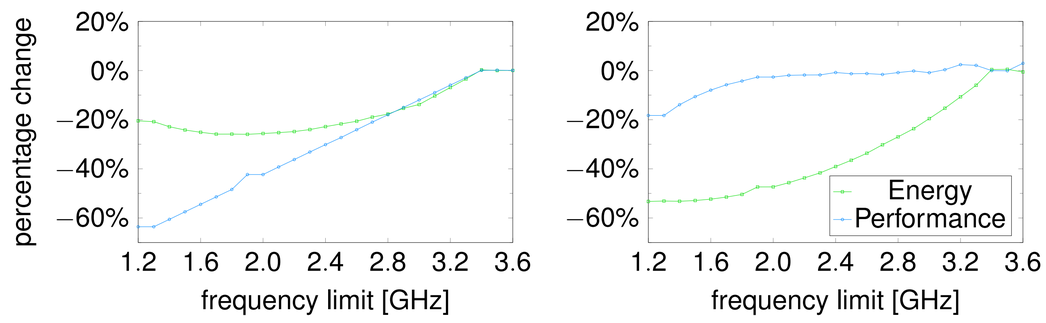
Balancing Resource Utilisation

- Provisioning of **resources always comes at a cost**, e.g. energy, money, occupation
- **Method: Balance resource demands** to benefit from **each investment**
- E.g. CPU and dth

CPU and bandwidth for database operators
[Noll et al., BTW'17], [Noll et al., DBSpektrum'17]

- High bandwidth demand, low compute demand (e.g. scan)
- Control frequency to lower provisioning of compute resource

Characterisation



CPU heavy **Memory heavy**

Results

Sketching-based Summary Algorithms

[Sohler and Woodruff, FOCS'18]

- **Strong coresets for the k-median** and subspace approximation problem, on n points in d dimensions (stored as $n \times d$ matrix A) of **size independent of n and d**.
- Running time

$$\mathcal{O}(\mathrm{nnz}(A) + (n+d)\mathrm{poly}(k/\varepsilon) + \exp(\mathrm{poly}(k/\varepsilon)))$$

[Driemel, Krivošija and Sohler, SODA'16]

- $(1+\varepsilon)$-approximation algorithm for k**-center and** k**-median** clustering of **time series** in $\mathbb{R}$, under **Fréchet distance**.
- Running time **nearly linear in the input size**, for constant k, ℓ, and ε, where ℓ is the length of the center curves.

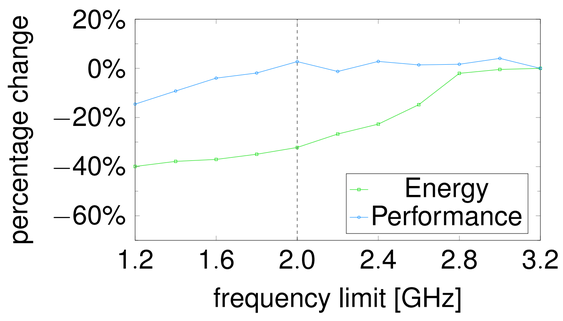
Energy-Efficient Sketching

- Convert **large datasets** into **approximate sketches** → Linear regression with **precision guarantees**
- **Bandwidth-hungry** consumption of original datasets

C4

Approach: Better **energy-efficiency** with **frequency scaling**

Save >30% processor energy with same performance



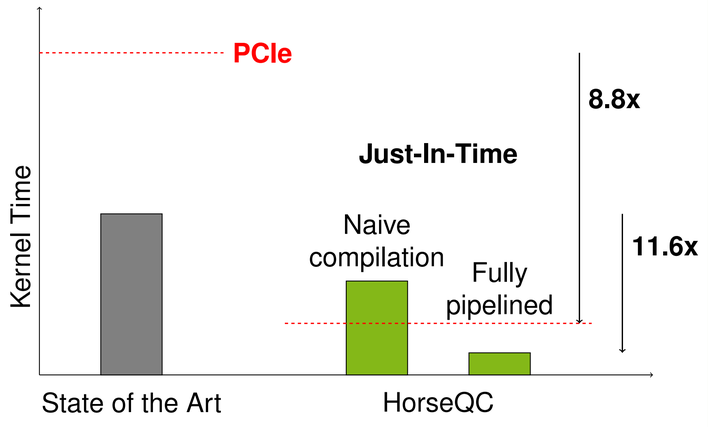
Computing a 536.9 MB sketch for 17.2 GB input data

Resource-Aware Query Coprocessing

- **Multi-level bandwidth constraints:** PCIe interconnect, GPU memory, main-memory
- Use **JIT-compilation** to **balance compute and bandwidth** demand
- Tailored **prefix sum** implementation **extends pipeline length**

C5

→ **Bandwidth-efficiency** [Funke et al., SIGMOD'18]



Lehrstuhl Informatik 2
Effiziente Algorithmen und Komplexitätstheorie

Lehrstuhl Informatik 6
Datenbanken und Informationssysteme