Generic Coreset for Scalable Learning of Monotonic Kernels: Logistic Regression, Sigmoid and more

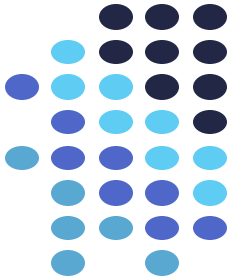
Elad Tolochinsky

Ibrahim Jubran

Dan Feldman

The Robotics & Big Data Lab,
Department of Computer Science,
University of Haifa,
Israel

ICML
International Conference
On Machine Learning



Logistic Regression

- Input dataset:

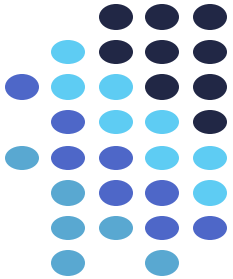
$$D = \{(p_1, y_1), \cdots, (p_n, y_n)\}$$

$$p_i \in \mathbb{R}^d, y_i \in \{1,0\}$$

- Loss function: the log likelihood of the data

$$\mathcal{L}(x) = \sum_{i=1}^{n} \log p(y_i | p_i, x)$$

$$\log p(y|p,x) = y \cdot \log(1 + e^{-p \cdot x}) + (1 - y) \cdot \log(1 + e^{p \cdot x})$$

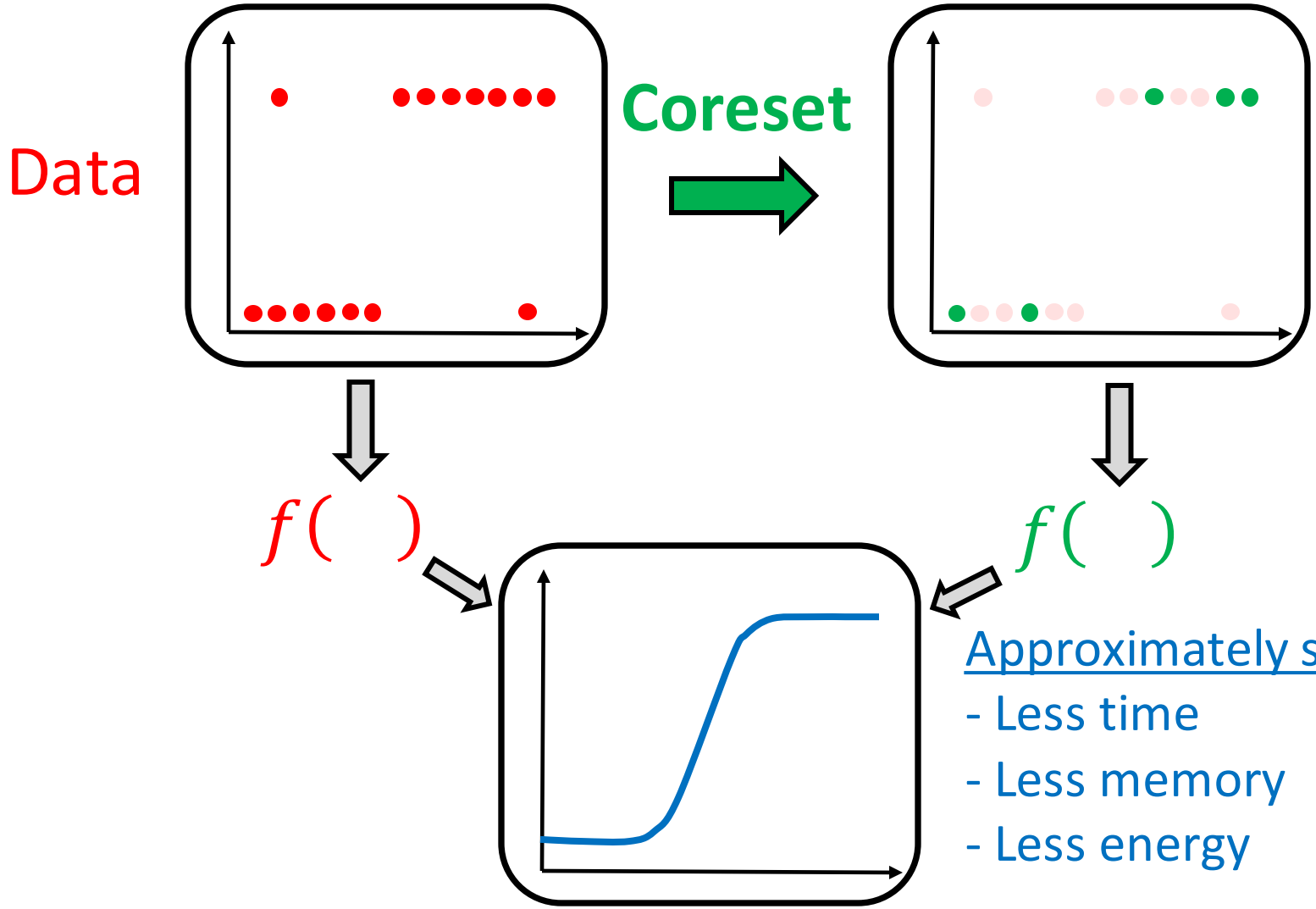


Motivation

While efficient solvers for the optimization problem exist, **huge amounts of data lead to high costs due to**:

- Huge memory consumption.
- Infeasible training times (e.g., when using hyperparameter tuning).
- High communication times.

Main Technique: Coresets

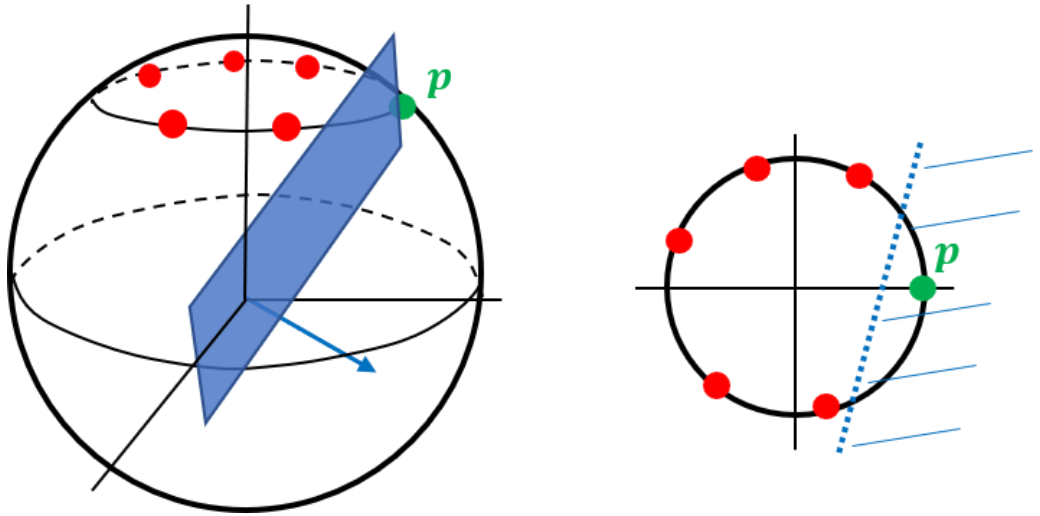


Approximately same output, but with:

- Less time
- Less memory
- Less energy

Contribution I:
No Coreset for the General Case

- We provide example synthetic datasets $\boldsymbol{D}$ for which there is **no coreset** of size smaller than $|\boldsymbol{D}|$.



Following this lower bound, the only hope remaining for constructing coresets resides in adding additional assumptions.

Contribution II: Coreset for regularized logistic regression

- We add a standard ℓ_2 regularization.

- We then prove that, with probability at least $1-\delta$, a small coreset Q exists for any normalized input set $P \subseteq \mathbb{R}^d$. Coreset size

$$|Q| \in \Omega\left(\frac{t}{\varepsilon^2}\left(d^2 \ln t + \ln\frac{1}{\delta}\right)\right)$$

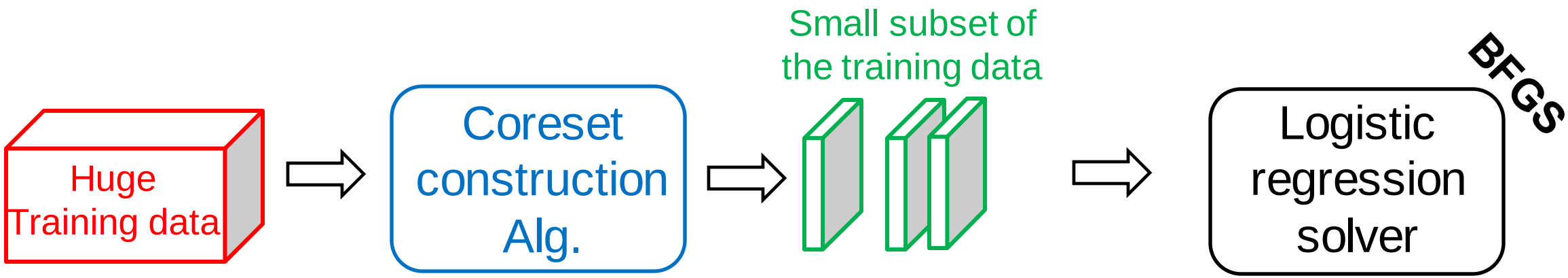
Approximation error

depends linearly on the regularization parameter and logarithmically on $|P|$

Failure probability

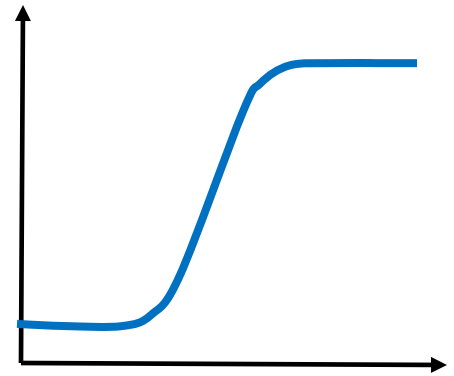
- A coreset construction algorithm that runs in near linear expected time.

Contribution II: Coreset for regularized logistic regression



The coreset is guaranteed to approximate:

- Loss function value.
- Marginal likelihood.
- Log likelihood ratio.



Contribution III: Generalization to additional loss functions

- We show that our coreset construction scheme holds for any other loss function $\mathbf{loss}(\mathbf{p}, \mathbf{x}) = \boldsymbol{f}(\boldsymbol{p} \cdot \boldsymbol{x})$ which satisfies:

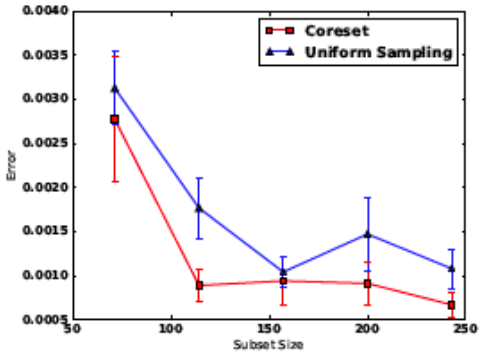
$$f(\|p\|\|x\|) + \frac{g(\|x\|)}{k} \leq b_p \left(f(-\|p\|\|x\|) + \frac{g(\|x\|)}{k} \right)$$

- For example, the non-convex sigmoid loss function:

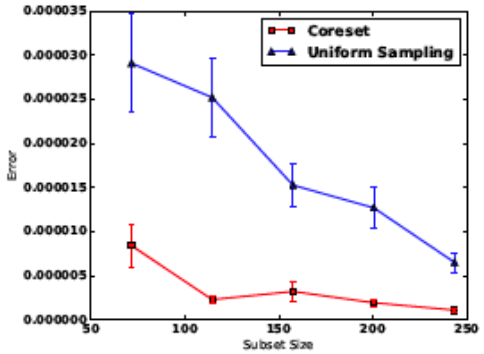
$$f(p \cdot x) = \frac{1}{1 + e^{-p \cdot x}}$$

Experimental Results

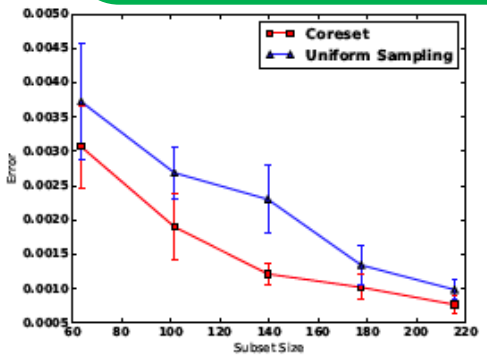
Empirically, a coreset of size $< 1\%$ can produce a small error of $\varepsilon = 0.001$.



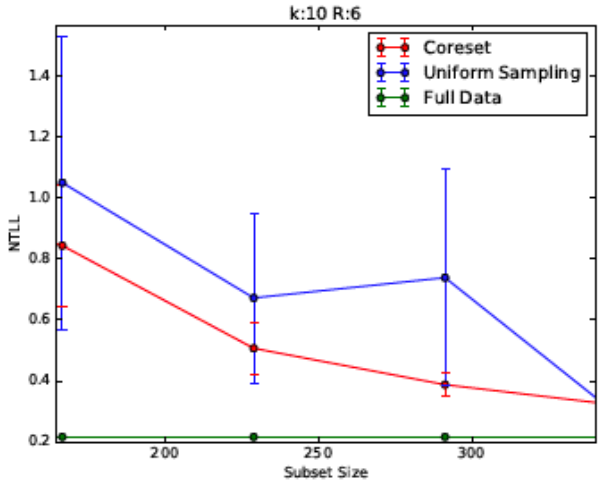
(a) Bank Marketing dataset, $k = 100$



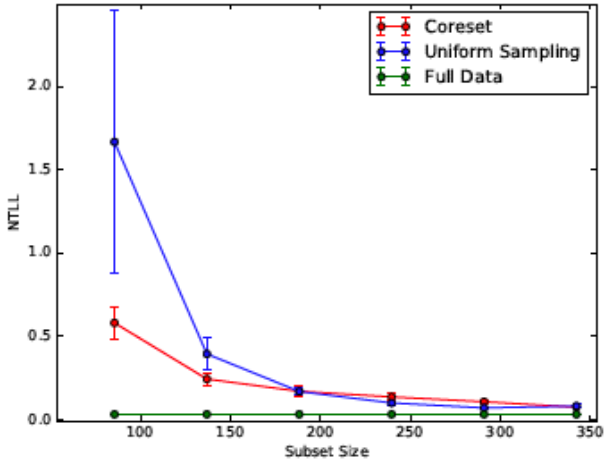
(b) Synthetic dataset, $k = 500$



(c) Wine dataset, $k = 1000$

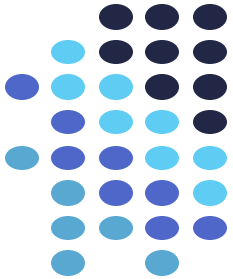


(d) Bank Marketing dataset, $k = 10, R = 6$



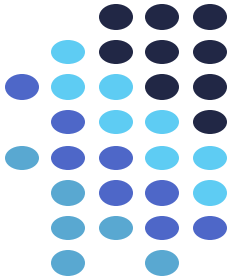
(e) Wine dataset, $k = 500, R = 4$

Figure 3. Experimental results. **Fig. 3(a)-3(c):** The error of maximizing sum of sigmoids using coreset and uniform sampling. **Fig. 3(d)-3(e):** Negative test log-likelihood. Lower is better in all figures.



Future Work

- In future work we hope to extend the work to handle an even wider range of loss functions.
- We also hope to reduce the dependency of the coreset size on the various parameters.



Thank you

Elad Tolochinsky
eladt26@gmail.com

Ibrahim Jubran
ibrahim.jub@gmail.com

Dan Feldman
dannyf.post@gmail.com