technische universität
dortmund

SFB 876 Providing Information
by Resource-Constrained Data Analysis

Almost Linear Constant-Factor Sketching for ℓ_1 and Logistic Regression

joint work with A. Munteanu (TU Dortmund) and D. Woodruff (CMU)

Simon Omlor | ICLR 2023

Setting: Logistic regression

- Input: Dataset $\{x_1, \dots, x_n\}$, $x_i \in \mathbb{R}^d$ with labels $y_i \in \{-1, 1\}$.
- We search for the empirical risk minimizer $\beta \in \mathbb{R}^d$ minimizing

$$f(\beta) = \sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i \beta)).$$

Logistic regression for massive data

Problems:

- Too much data to store in the working memory
- Limited access to data: data streams
- Data is given in pieces or updated dynamically (turnstile streams, vertically distributed)

Massive data analysis

Sketch and solve paradigm

$$\begin{array}{ccc} X & \xrightarrow{\Pi} & \Pi X \\ \downarrow & & \downarrow \\ f(\beta \mid X) & \approx & f(\beta \mid \Pi X) \end{array}$$

Massive data analysis

Sketch and solve paradigm

$$\begin{array}{ccc} X & \xrightarrow{\Pi} & \Pi X \\ \downarrow & & \downarrow \\ f(\beta \mid X) & \approx & f(\beta \mid \Pi X) \end{array}$$

Canonical approach

1. Data reduction $X \to \Pi X$ (fast linear sketch), where $|\Pi X| \ll |X|$
2. Time- and space efficient calculations on ΠX
3. Approximation guarantee: solution is close to optimal

Sketching matrix

$$\Pi = \begin{pmatrix} S_0 \\ S_1 \\ \vdots \\ S_{h_{\max}} \end{pmatrix}$$

- Idea: Subsample the points at different rates: for each entry randomly choose its level $h \in \mathbb{N}$ at rate 2^{-h}
- $S_{h_{\max}}$ is a uniform sample and S_0 is a CountMin-Sketch of the full data;

Sketching matrix

$$\Pi = \begin{pmatrix} S_0 \\ S_1 \\ \vdots \\ S_{h_{\max}} \end{pmatrix}$$

- Idea: Subsample the points at different rates: for each entry randomly choose its level $h \in \mathbb{N}$ at rate 2^{-h}
- $S_{h_{\max}}$ is a uniform sample and S_0 is a CountMin-Sketch of the full data;
- To reflect the rarity of entries, buckets of high levels (with few elements) get a high weight;
- Sketch the points at each level: inside the levels we use the CountMin-Sketch, i.e., we hash points into buckets uniformly at random and add up entries in the same bucket.

Analysis

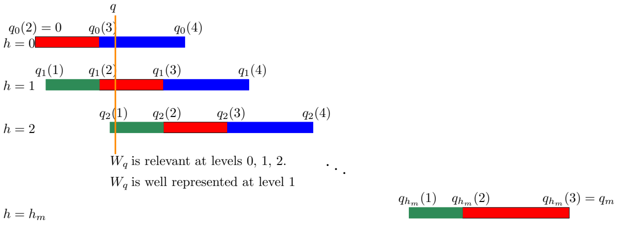
Fix β and set $z = X\beta$, prove that with high probability following holds:

- Contraction bounds: $f(\Pi z) \geq (1-\varepsilon)f(z)$ (apply a net argument to show that this holds for any $z' = X\beta'$);
- Dilation bounds: $f(\Pi z) \leq kf(z)$.

Analysis

Idea: Divide the entries of z into weight classes $W_q^+ = \{i \mid 2^{-q} \geq \frac{z_i}{\|z\|_1} > 2^{-(q+1)}\}$.

- For any 'relevant' weight class W_q^+ there is a level where the contribution of W_q^+ is preserved
- The expected contribution of any weight class W_q^+ to any level is at most $\|W_q^+\|_1$ and W_q^+ contributes to at most k levels



Main result

Theorem 1

There is a distribution over sketching matrices $\Pi \in \mathbb{R}^{r \times n}$ such that with constant probability

1. ΠX *has* $r = O(\mu d^{1+c} \ln(n)^{2+4c})$ *rows, can be computed in time* $O(d \ln(n) \mu \cdot \mathtt{nnz}(X))$ *and yields an* $O(1)$ *approximation.*
2. ΠX *has* $r = \mu^2 (\varepsilon^{-1} \ln(n) d)^{O(\varepsilon^{-1})}$ *rows, can be computed in time* $O(\mathtt{nnz}(X))$ *and yields an* $1 + \varepsilon$ *approximation.*

$\mu = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\sum_{x_i \beta > 0} |x_i \beta|}{\sum_{x_i \beta < 0} |x_i \beta|}$

$\mathtt{nnz}(X)$ = number of non-zero entries of X.

Other target functions

We get similar results for ℓ_1-regression

$$f(X\beta) = \|X\beta - y\|_1$$

and logistic regression with variance-based regularization

$$\begin{aligned} f(X\beta) &= \frac{1}{n}\sum_{i=1}^{n}\ell(x_i\beta) + \frac{\lambda}{2n}\sum_{i=1}^{n}\ell(x_i\beta)^2 - \frac{\lambda}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\ell(x_i\beta)\right)^2 \\ &= \mathbb{E}(\ell(x\beta)) \qquad + \frac{\lambda}{2}\cdot \mathrm{Var}(\ell(x\beta)). \end{aligned}$$