Efficient Symmetric Norm Regression via Linear Sketching

Zhao Song — University of Washington
Ruosong Wang — Carnegie Mellon University
Lin Yang — University of California, Los Angeles
Hongyang Zhang — Toyota Technological Institute at Chicago
Peilin Zhong — Columbia University

Linear Regression

Given: $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$, and a loss function $L : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$.
Output: $x \in \mathbb{R}^d$ which minimizes

$$L(Ax - b).$$

Approximate solution: $\hat{x}$ satisfies

$$L(A\hat{x} - b) \leq \alpha \cdot \min_{x\in\mathbb{R}^d} L(Ax - b)$$

for some approximation ratio $\alpha \geq 1$.

Motivations

Some classic choices of loss functions:

- $L(x) \equiv \sum x_i^2$
  - ℓ_2 regression (Least Squares Regression).
- $L(x) \equiv \sum |x_i|$
  - ℓ_1 regression (Least Absolute Deviation Regression).
- $L(x) \equiv \sum |x_i|^p$
  - ℓ_p regression.

Is it possible to design fast algorithms for linear regression, that work for a wide range of loss functions?

- Prior work studied this problem for the loss function $L(x) \equiv \sum_{i=1}^n M(x_i)$ for some function M:
  - $M(\cdot)$ is an M-estimator:

Table 1: Some of M-estimators.

Huber	$\begin{cases} x^2/2 & \lvert x\rvert \leq c \\ c(\lvert x\rvert - c/2) & \lvert x\rvert > c \end{cases}$
$\ell_1 - \ell_2$	$2(\sqrt{1 + x^2/2} - 1)$
"Fair"	$c^2\left(\lvert x\rvert/c - \log(1 + \lvert x\rvert/c)\right)$

- However, much less is known for the case where the loss function $L(\cdot)$ is a norm, except for ℓ_p norms.
- A recent work gives an $\widetilde{O}(\mathrm{nnz}(A) + \mathrm{poly}(d))$ time approximation algorithm when $L(\cdot)$ is an Orlicz norm.
- Two problems left open:
  - Prior work for Orlicz norm has approximation ratio $d \cdot \mathrm{poly}(\log n)$. Can we obtain $(1 + \varepsilon)$-approximation for arbitrary small ε?
  - Is it possible to have an $\widetilde{O}(\mathrm{nnz}(A) + \mathrm{poly}(d))$ time approximation algorithm for a wider class of norms?

Symmetric Norm

A norm $\|\cdot\|_\ell$ is called a *symmetric norm*, if $\|(y_1, y_2, \ldots, y_n)\|_\ell = \|(s_1 y_{\sigma_1}, s_2 y_{\sigma_2}, \ldots, s_n y_{\sigma_n})\|_\ell$ for any permutation σ and any assignment of $s_i \in \{-1, 1\}$.

- Symmetric norm includes ℓ_p norms and Orlicz norms as special cases. Other examples include top-k norms, max-mix of ℓ_p norms, sum-mix of ℓ_p norms, the k-support norm and the box-norm, etc.

Orlicz norm:

- In our work, we consider a function $G : \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfies the following properties:
  - G is a strictly increasing convex function on $[0, \infty)$;
  - $G(0) = 0$, and for all $x \in \mathbb{R}$, $G(x) = G(-x)$;
  - There exists some $C_G > 0$, such that for all $0 < x < y$, $G(y)/G(x) \leq C_G (y/x)^2$.
- For a function G and a vector $y \in \mathbb{R}^n$ with $y \neq 0$, the corresponding Orlicz norm $\|y\|_G$ is defined as the unique value α such that

$$\sum_{i=1}^n G(|y_i|/\alpha) = 1.$$

When $y = 0$, we define $\|y\|_G$ to be 0.

Our Results

Theorem 1. *There exists an algorithm that, on any input $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, finds a vector x^* in time $\widetilde{O}(\mathrm{nnz}(A) + \mathrm{poly}(d/\varepsilon))$, such that with probability at least 0.9, $\|Ax^* - b\|_G \leq (1 + \varepsilon)\min_{x\in\mathbb{R}^d} \|Ax - b\|_G$.*

Theorem 2. *Given a symmetric norm $\|\cdot\|_\ell$, there exists an algorithm that, on any input $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, finds a vector x^* in time $\widetilde{O}(\mathrm{nnz}(A) + \mathrm{poly}(d))$, such that with probability at least 0.9, $\|Ax^* - b\|_\ell \leq \sqrt{d} \cdot \mathrm{polylog}n \cdot \mathrm{mmc}(\ell) \cdot \min_{x\in\mathbb{R}^d} \|Ax - b\|_\ell$.*

- $\mathrm{mmc}(\ell)$ is a characteristic of $\|\cdot\|_\ell$, which has been proven to be essential in streaming algorithms for symmetric norms.
- Examples with $\mathrm{mmc}(\ell) \leq \mathrm{polylog}n$:
  - ℓ_p norms with $p \leq 2$, top-k norms with $k \geq n/\mathrm{polylog}n$, max-mix of ℓ_2 norm and ℓ_1 norm ($\max\{\|x\|_2, c\|x\|_1\}$ for some $c > 0$), sum-mix of ℓ_2 norm and ℓ_1 norm ($\|x\|_2 + c\|x\|_1$ for some $c > 0$), the k-support norm, and the box-norm.
  - Our algorithm has approximation ratio $\sqrt{d} \cdot \mathrm{polylog}n$ for all these norms.

Our Techniques

- Orlicz norm regression:
  - Our algorithm is based on row sampling.
  - For a given matrix $A \in \mathbb{R}^{n\times d}$, our goal is to output a *sparse* weight vector $w \in \mathbb{R}^n$ with at most $\mathrm{poly}(d \log n/\varepsilon)$ non-zero entries, such that with high probability, for all $x \in \mathbb{R}^d$,

$$(1-\varepsilon)\|Ax-b\|_G \leq \|Ax-b\|_{G,w} \leq (1+\varepsilon)\|Ax-b\|_G.$$

  - For $w \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, the *weighted Orlicz norm* $\|y\|_{G,w}$ is defined as the unique value α such that $\sum_{i=1}^n w_i G(|y_i|/\alpha) = 1$.
  - It suffices to solve

$$\min_{x\in\mathbb{R}^d} \|Ax - b\|_{G,w}.$$

  - We want the number of non-zero entries of w to be at most $\mathrm{poly}(d \log n/\varepsilon)$.
  - Let $\bar{A}$ be $[A\ b]$. We want $\forall x \in \mathbb{R}^{d+1}$,

$$(1 - \varepsilon)\|\bar{A}x\|_G \leq \|\bar{A}x\|_{G,w} \leq (1 + \varepsilon)\|\bar{A}x\|_G.$$

  - Well-conditioned basis: for all $x \in \mathbb{R}^d$,

$$\|x\|_2 \leq \|Ux\|_G \leq \kappa_G\|x\|_2.$$

  - Orlicz norm leverage score of row i: $G(\|U_i\|_2)$. The summation of leverage scores will be $O(d\kappa_G^2)$.
  - Sample each row with probability

$$p_i \geq \min\{1, d\varepsilon^{-2}\log(1/\varepsilon)G(\|U_i\|_2)\}.$$

  Set the weight $w_i = 1/p_i$.
- General symmetric norm:
  - Want to construct Π such that $\forall x$, $\|\Pi\bar{A}x\|_2$ is a good approximation to $\|\bar{A}x\|_G$.
  - $\Pi = S \cdot \widetilde{D} = S \cdot \begin{bmatrix} w_0 D_0 \\ w_1 D_1 \\ \vdots \\ w_t D_t \end{bmatrix}.$
  - S is an ℓ_2 subspace embedding. Each diagonal entry of D_i is 1 w.p. $1/2^i$. $w_i = \|(1, 1, \cdots, 1, 0, 0, \cdots, 0)\|_\ell$ and there are 2^i 1s.