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:
- \(L(x) \equiv \sum_i x_i^2\), \(\ell_2\) regression (Least Squares Regression).
- \(L(x) \equiv \sum_i |x_i|\), \(\ell_1\) regression (Least Absolute Deviation Regression).
- \(L(x) \equiv \sum_i |x_i|^p\), \(\ell_p\) regression.
- \(L(x) \equiv \sum_i \ell_i(x_i)\), \(\ell_i\) 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:

| Huber | \(f(x^2) \leq c\) | \(|x| - c/2\)| | \(|x| > c\) |
|-------|-----------------|----------------|--------------|
| \(\ell_1 - \ell_2\) | \(2\sqrt{1 + x^2/2 - 1}\) | \(x/\sqrt{1 + x^2/2 - 1}\) |

- However, much less is known for the case where the loss function \(L(\cdot)\) is a norm, except for \(\ell_p\) norms.
- A recent work gives an \(O(\text{nnz}(A) + \text{poly}(d))\) time approximation algorithm when \(L(\cdot)\) is an Orlicz norm.

**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 \(O(\text{nnz}(A) + \text{poly}(d/\varepsilon))\), such that with probability at least 0.9, \(\|Ax^* - b\|_2 \leq (1 + \varepsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|_2\).

**Theorem 2.** Given a symmetric norm \(\|\cdot\|\), 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 \(O(\text{nnz}(A) + \text{poly}(d))\), such that with probability at least 0.9, \(\|Ax^* - b\|_2 \leq \sqrt{d} \cdot \text{polylog}(n) \cdot \text{mmse}(\ell) \cdot \min_{x \in \mathbb{R}^d} \|Ax - b\|_2\).

**Our Results**

**Symmetric Norm**

A norm \(\|\cdot\|_\ell\) is called a symmetric norm, if for any permutation \(\sigma\) and any assignment of \(s_i \in \{-1, 1\}\)
- Symmetric norm includes \(\ell_p\) norms and Orlicz norms as special cases.
- Other examples include top-\(k\) norms, max-mix of \(\ell_p\) norms, sum-mix of \(\ell_p\) 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 \(y \in \mathbb{R}^d\) with \(y \neq 0\), the corresponding Orlicz norm \(\|y\|_G\) is defined as the unique value \(\alpha\) such that
  \[
  \sum_{i=1}^n G(|y_i|/\alpha) = 1.
  \]
- When \(y = 0\), we define \(\|y\|_G = 0\).

**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 vector weight \(w \in \mathbb{R}^n\) with at most \(\text{poly}(d \log n/\varepsilon)\) non-zero entries, such that with high probability, for all \(x \in \mathbb{R}^d\),
    \[
    (1 - \varepsilon)\|Ax - b\|_2 \leq \|Ax - b\|_{G,w} \leq (1 + \varepsilon)\|Ax - b\|_2.
    \]
  - 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 \(\alpha\) 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 \(\text{poly}(d \log n/\varepsilon)\).
- Let \(A = [A \ b]\). We want \(\forall x \in \mathbb{R}^{d+1}\),
  \[
  (1 - \varepsilon)\|Ax\|_2 \leq \|Ax\|_{G,w} \leq (1 + \varepsilon)\|Ax\|_2.
  \]
- Well-conditioned basis: for all \(x \in \mathbb{R}^d\),
  \[
  \|x\|_2 \leq \|Ux\|_2 \leq \kappa_G \|x\|_2.
  \]
- Orlicz norm leverage score of row \(i\): \(G(U_{i,\cdot})\). The summation of leverage scores will be \(O(d\kappa_G^2)\).
- Sample each row with probability
  \[
  p_i \geq \min\{1, d/\varepsilon^2 \log(1/\varepsilon) G(U_{i,\cdot})\}.
  \]
- Set the weight \(w_i = 1/p_i\).
- General symmetric norm:
  - Want to construct \(I\) such that \(\forall x, \|I Ax\|_2\) is a good approximation to \(\|Ax\|_2\).
  - \(I = S \cdot \widetilde{D} = S \cdot \begin{bmatrix} w_0 D_0 \\ w_1 D_1 \\ \vdots \\ w_t D_t \end{bmatrix}
    \)
  - \(S\) is an \(\ell_2\) subspace embedding. Each diagonal entry of \(D_i\) is 1 w.p. \(1/2^i\).
  - \(w_i = \max\{1, 1, \ldots, 1, 0, \ldots, 0\}\) for some \(c > 0\), sum-mix of \(\ell_2\) norms and \(\ell_1\) norm \(\|x\|_{1+\varepsilon/\sqrt{2\cdot n}}\) for some \(c > 0\), the \(k\)-support norm, and the box-norm.
  - Our algorithm has approximation ratio \(\sqrt{d} \cdot \text{polylog}(n)\) for all these norms.