Subspace Embedding and Linear Regression with Orlicz Norm

Alexandr Andoni[1] Chenyu Lin[1] Ying Sheng[1] Peilin Zhong[1]
Ruiqi Zhong[1](Speaker)

[1]Department of Computer Science
Columbia University in the City of New York
NY 10027, U.S.A.

International Conference on Machine Learning, 2018

General Linear Regression Formulation

- General Linear Regression as an optimization problem:
  $x^* = \text{argmin}_{x \in \mathbb{R}^d} L(Ax - b)$, where:
  $A \in \mathbb{R}^{n \times d}, x \in \mathbb{R}^d, b \in \mathbb{R}^n, L : \mathbb{R}^n \to \mathbb{R}, r = Ax - b$

General Linear Regression Formulation

- General Linear Regression as an optimization problem:
  $x^* = \text{argmin}_{x \in \mathbb{R}^d} L(Ax - b)$, where:
  $A \in \mathbb{R}^{n \times d}, x \in \mathbb{R}^d, b \in \mathbb{R}^n, L : \mathbb{R}^n \to \mathbb{R}, r = Ax - b$
- ℓ_2 Linear Regression (a.k.a Least Square Regression)
  - $L(r) = \|r\|_2^2 = \sum_i r_i^2 = \sum_i (A_i x - b_i)^2$
  - unbiased mean estimator; easy to optimize.
  - sensitive to large outliers.

General Linear Regression Formulation

- General Linear Regression as an optimization problem:
  $x^* = \text{argmin}_{x \in \mathbb{R}^d} L(Ax - b)$, where:
  $A \in \mathbb{R}^{n \times d}, x \in \mathbb{R}^d, b \in \mathbb{R}^n, L : \mathbb{R}^n \to \mathbb{R}, r = Ax - b$
- ℓ_2 Linear Regression (a.k.a Least Square Regression)
  - $L(r) = \|r\|_2^2 = \sum_i r_i^2 = \sum_i (A_i x - b_i)^2$
  - unbiased mean estimator; easy to optimize.
  - sensitive to large outliers.
- ℓ_1 Linear Regression (a.k.a Least Absolute Value Regression)
  - $L(r) = ||r||_1 = \sum_i |r_i|$
  - median estimator; robust against outliers.
  - large MSE; relatively unstable

Robust Statistics: Huber Loss

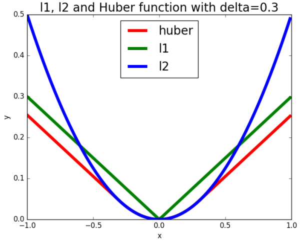
- Huber Loss: $L(r) = \sum_{i=1}^{n} f(r_i)$, where

$$f(r_i) = \begin{cases} \frac{r_i^2}{2} & |r_i| \leq \delta \\ \delta(|r_i| - \delta/2) & \text{otherwise} \end{cases}$$

Robust Statistics: Huber Loss

- Huber Loss: $L(r) = \sum_{i=1}^{n} f(r_i)$, where

$$f(r_i) = \begin{cases} \frac{r_i^2}{2} & |r_i| \leq \delta \\ \delta(|r_i| - \delta/2) & \text{otherwise} \end{cases}$$

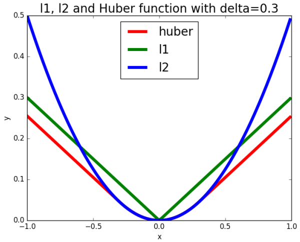


- Exactly ℓ_2 loss near 0 and similar to ℓ_1 asymptotically.

Robust Statistics: Huber Loss

- Huber Loss: $L(r) = \sum_{i=1}^{n} f(r_i)$, where

$$f(r_i) = \begin{cases} \frac{r_i^2}{2} & |r_i| \le \delta \\ \delta(|r_i| - \delta/2) & \text{otherwise} \end{cases}$$



- Exactly ℓ_2 loss near 0 and similar to ℓ_1 asymptotically.
- Stable solution, robust against outliers.

Problem: Choosing an Appropriate δ

- $$f(r_i) = \begin{cases} \frac{r_i^2}{2} & |r_i| \leq \delta \\ \delta(|r_i| - \delta/2) & \text{otherwise} \end{cases}$$
- $\delta = 1.345\sigma$ is a common choice, where σ is the variance of the prediction error.

Problem: Choosing an Appropriate δ

- $$f(r_i) = \begin{cases} \frac{r_i^2}{2} & |r_i| \leq \delta \\ \delta(|r_i| - \delta/2) & \text{otherwise} \end{cases}$$
- $\delta = 1.345\sigma$ is a common choice, where σ is the variance of the prediction error.
- σ is unknown.

Problem: Choosing an Appropriate δ

- $$f(r_i) = \begin{cases} \frac{r_i^2}{2} & |r_i| \le \delta \\ \delta(|r_i| - \delta/2) & \text{otherwise} \end{cases}$$
- $\delta = 1.345\sigma$ is a common choice, where σ is the variance of the prediction error.
- σ is unknown.
- f is not scale-invariant.

Our Approach: via Orlicz Norm

- Orlicz Norm

$$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$

where G is a non-negative convex function $R_+ \to R_+, G(0) = 0$

Our Approach: via Orlicz Norm

- Orlicz Norm
$$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$
where G is a non-negative convex function $R_+ \to R_+, G(0) = 0$
- $G(x) = x^2$, $\|x\|_G = \|x\|_2 = \alpha = \sqrt{\sum_{i=1}^n x_i^2}$,
$\sum_{i=1}^n G(\frac{x_i}{\alpha}) = \sum_{i=1}^n (\frac{x_i}{\sqrt{\sum_i^n x_i^2}})^2 = 1$

Our Approach: via Orlicz Norm

- Orlicz Norm
$$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$
where G is a non-negative convex function $R_+ \to R_+$, $G(0) = 0$
- $G(x) = x^2$, $\|x\|_G = \|x\|_2 = \alpha = \sqrt{\sum_{i=1}^n x_i^2}$, $\sum_{i=1}^n G(\frac{x_i}{\alpha}) = \sum_{i=1}^n (\frac{x_i}{\sqrt{\sum_i^n x_i^2}})^2 = 1$
- If $G(x) = x^p$, then $\|x\|_G = \|x\|_p$

Our Approach: via Orlicz Norm

- Orlicz Norm

$$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$

where G is a non-negative convex function $R_+ \to R_+$, $G(0) = 0$

- $G(x) = x^2$, $\|x\|_G = \|x\|_2 = \alpha = \sqrt{\sum_{i=1}^n x_i^2}$, $\sum_{i=1}^n G(\frac{x_i}{\alpha}) = \sum_{i=1}^n (\frac{x_i}{\sqrt{\sum_i^n x_i^2}})^2 = 1$
- If $G(x) = x^p$, then $\|x\|_G = \|x\|_p$
- We will focus on G as huber function.

$$G(x) = \begin{cases} \frac{x^2}{2} & |x| \leq \delta \\ \delta(|x| - \delta/2) & \text{otherwise} \end{cases}$$

Our Approach: via Orlicz Norm

- Orlicz Norm

$$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$

where G is a non-negative convex function $R_+ \to R_+$, $G(0) = 0$

- $G(x) = x^2$, $\|x\|_G = \|x\|_2 = \alpha = \sqrt{\sum_{i=1}^n x_i^2}$,
$\sum_{i=1}^n G(\frac{x_i}{\alpha}) = \sum_{i=1}^n (\frac{x_i}{\sqrt{\sum_i^n x_i^2}})^2 = 1$
- If $G(x) = x^p$, then $\|x\|_G = \|x\|_p$
- We will focus on G as huber function.

$$G(x) = \begin{cases} \frac{x^2}{2} & |x| \leq \delta \\ \delta(|x| - \delta/2) & \text{otherwise} \end{cases}$$

- $\|\cdot\|_G$ is scale-invariant

Outline

Outline

Outline

1. Orlicz Norm and the Motivation
   - General Linear Regression
   - Huber Loss
   - Orlicz Norm
2. Solving Tasks Defined on Orlicz Norms
   - Subspace embedding method for “nice” Orlicz Norm
   - Orlicz Norm Linear Regression
3. Experiments on Orlicz Norm Linear Regression
   - Comparison with ℓ_1 and ℓ_2 regression
   - Best δ under Different Outlier Size
   - Beyond Huber function - A General Framework

Outline

Outline

Conditions for “nice” Orlicz Norm

- $$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$

Conditions for "nice" Orlicz Norm

- $$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$
- $G : \mathbb{R}_+ \to \mathbb{R}_+$ is convex, non-decreasing, $G(0) = 0$

Conditions for "nice" Orlicz Norm

- $$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$
- $G : \mathbb{R}_+ \to \mathbb{R}_+$ is convex, non-decreasing, $G(0) = 0$
- $\exists C_G > 0, \forall 0 < x < y, \frac{G(y)}{G(x)} \leq C_G(\frac{y}{x})^2$

Conditions for "nice" Orlicz Norm

- $$\|x\|_G = \alpha \text{ s.t. } \sum_{i=1}^{n} G(|x_i|/\alpha) = 1$$
- $G : \mathbb{R}_+ \to \mathbb{R}_+$ is convex, non-decreasing, $G(0) = 0$
- $\exists C_G > 0, \forall 0 < x < y, \frac{G(y)}{G(x)} \leq C_G(\frac{y}{x})^2$
- $\exists \delta_G$ s.t. G is twice-differentiable on interval $(0, \delta_G)$

Our Main Tool: Subspace Embedding

- $A \in \mathbb{R}^{n \times d}$, for embedding matrix $S \in \mathbb{R}^{m \times n}$ satisfies

Our Main Tool: Subspace Embedding

- $A \in \mathbb{R}^{n\times d}$, for embedding matrix $S \in \mathbb{R}^{m\times n}$ satisfies $\forall x \in \mathbb{R}^d$, $\|Ax\|_G/\alpha \leq \|SAx\|_2 \leq \beta\|Ax\|_G$

Our Main Tool: Subspace Embedding

- $A \in \mathbb{R}^{n\times d}$, for embedding matrix $S \in \mathbb{R}^{m\times n}$ satisfies $\forall x \in \mathbb{R}^d$, $\|Ax\|_G/\alpha \le \|SAx\|_2 \le \beta\|Ax\|_G$, then we say S embeds the column space of A with Orlicz norm into the column space of SA with 2-norm.

Our Main Tool: Subspace Embedding

- $A \in \mathbb{R}^{n \times d}$, for embedding matrix $S \in \mathbb{R}^{m \times n}$ satisfies $\forall x \in \mathbb{R}^d$, $\|Ax\|_G/\alpha \leq \|SAx\|_2 \leq \beta\|Ax\|_G$, then we say S embeds the column space of A with Orlicz norm into the column space of SA with 2-norm. The distortion is $\alpha\beta$.

Our Main Tool: Subspace Embedding

- $A \in \mathbb{R}^{n\times d}$, for embedding matrix $S \in \mathbb{R}^{m\times n}$ satisfies $\forall x \in \mathbb{R}^d$, $\|Ax\|_G/\alpha \leq \|SAx\|_2 \leq \beta\|Ax\|_G$, then we say S embeds the column space of A with Orlicz norm into the column space of SA with 2-norm. The distortion is $\alpha\beta$.
- For a matrix $A \in \mathbb{R}^{n\times m}$ with rank d, $\Omega(1/(d \log n))\|Ax\|_G \leq \|D^{-1}Ax\|_2 \leq O(d^2 \log n)\|Ax\|_G$

Our Main Tool: Subspace Embedding

- $A \in \mathbb{R}^{n \times d}$, for embedding matrix $S \in \mathbb{R}^{m \times n}$ satisfies $\forall x \in \mathbb{R}^d$, $\|Ax\|_G/\alpha \leq \|SAx\|_2 \leq \beta\|Ax\|_G$, then we say S embeds the column space of A with Orlicz norm into the column space of SA with 2-norm. The distortion is $\alpha\beta$.
- For a matrix $A \in \mathbb{R}^{n \times m}$ with rank d, $\Omega(1/(d\log n))\|Ax\|_G \leq \|D^{-1}Ax\|_2 \leq O(d^2\log n)\|Ax\|_G$
- Our main technique is to embed this norm to ℓ_2

Our Main Tool: Subspace Embedding

- $A \in \mathbb{R}^{n \times d}$, for embedding matrix $S \in \mathbb{R}^{m \times n}$ satisfies $\forall x \in \mathbb{R}^d$, $\|Ax\|_G/\alpha \leq \|SAx\|_2 \leq \beta\|Ax\|_G$, then we say S embeds the column space of A with Orlicz norm into the column space of SA with 2-norm. The distortion is $\alpha\beta$.
- For a matrix $A \in \mathbb{R}^{n \times m}$ with rank d, $\Omega(1/(d \log n))\|Ax\|_G \leq \|D^{-1}Ax\|_2 \leq O(d^2 \log n)\|Ax\|_G$
- Our main technique is to embed this norm to ℓ_2
- $D \in \mathbb{R}^{n \times n}$ be a diagonal matrix, each entry an i.i.d. random variable drawn from CDF $1 - e^{-G(t)}$.

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n \times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n\times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \le O(d^2 \log(n)\|Ax\|_G)$

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n\times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \leq O(d^2 \log(n)\|Ax\|_G)$
  Proof sketch: $f : \mathbb{R}^n \to \mathbb{R}^n$:
  $(x_1, x_2, ..., x_n) \to (x_1/u_1, x_2/u_2, ..., x_n/u_n)$. We want to bound the probability of $\|f(x)/\gamma\alpha\|_G \geq 1$, and thus $\mathbb{E}[\sum_{i=1}^n G(x_i\gamma)/(\alpha u_i)]$.
  $1/G(u_i)$ is roughly $O(\log n)$.

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n\times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \leq O(d^2 \log(n)\|Ax\|_G)$
  Proof sketch: $f : \mathbb{R}^n \to \mathbb{R}^n$:
  $(x_1, x_2, ..., x_n) \to (x_1/u_1, x_2/u_2, ..., x_n/u_n)$. We want to bound the probability of $\|f(x)/\gamma\alpha\|_G \geq 1$, and thus $\mathbb{E}[\sum_{i=1}^n G(x_i\gamma)/(\alpha u_i)]$.
  $1/G(u_i)$ is roughly $O(\log n)$.
  used ℓ_1 well-conditioned basis argument to prove for all x.

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n\times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \leq O(d^2 \log(n)\|Ax\|_G)$
  Proof sketch: $f : \mathbb{R}^n \to \mathbb{R}^n$:
  $(x_1, x_2, ..., x_n) \to (x_1/u_1, x_2/u_2, ..., x_n/u_n)$. We want to bound the probability of $\|f(x)/\gamma\alpha\|_G \geq 1$, and thus $\mathbb{E}[\sum_{i=1}^n G(x_i\gamma)/(\alpha u_i)]$.
  $1/G(u_i)$ is roughly $O(\log n)$.
  used ℓ_1 well-conditioned basis argument to prove for all x.
- Contraction bound: $w.h.p.$ $\Omega(1/(d \log n)\|Ax\|_G) \leq \|D^{-1}Ax\|_\infty$

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n \times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \leq O(d^2 \log(n)\|Ax\|_G)$
  Proof sketch: $f : \mathbb{R}^n \to \mathbb{R}^n$:
  $(x_1, x_2, ..., x_n) \to (x_1/u_1, x_2/u_2, ..., x_n/u_n)$. We want to bound the probability of $\|f(x)/\gamma\alpha\|_G \geq 1$, and thus $\mathbb{E}[\sum_{i=1}^n G(x_i\gamma)/(\alpha u_i)]$.
  $1/G(u_i)$ is roughly $O(\log n)$.
  used ℓ_1 well-conditioned basis argument to prove for all x.
- Contraction bound: $w.h.p.$ $\Omega(1/(d \log n)\|Ax\|_G) \leq \|D^{-1}Ax\|_\infty$
  Proof sketch: with probability at least $1 - e^{-\alpha}$, $\|f(x)\|_\infty \geq \|x\|_G/\alpha$.

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n\times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \leq O(d^2 \log(n)\|Ax\|_G)$
  Proof sketch: $f : \mathbb{R}^n \to \mathbb{R}^n$:
  $(x_1, x_2, ..., x_n) \to (x_1/u_1, x_2/u_2, ..., x_n/u_n)$. We want to bound the probability of $\|f(x)/\gamma\alpha\|_G \geq 1$, and thus $\mathbb{E}[\sum_{i=1}^n G(x_i\gamma)/(\alpha u_i)]$.
  $1/G(u_i)$ is roughly $O(\log n)$.
  used ℓ_1 well-conditioned basis argument to prove for all x.
- Contraction bound: $w.h.p.$ $\Omega(1/(d \log n)\|Ax\|_G) \leq \|D^{-1}Ax\|_\infty$
  Proof sketch: with probability at least $1 - e^{-\alpha}$, $\|f(x)\|_\infty \geq \|x\|_G/\alpha$.
  ϵ net argument to prove for all x.

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n\times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \leq O(d^2 \log(n)\|Ax\|_G)$
  Proof sketch: $f : \mathbb{R}^n \to \mathbb{R}^n$:
  $(x_1, x_2, ..., x_n) \to (x_1/u_1, x_2/u_2, ..., x_n/u_n)$. We want to bound the probability of $\|f(x)/\gamma\alpha\|_G \geq 1$, and thus $\mathbb{E}[\sum_{i=1}^n G(x_i\gamma)/(\alpha u_i)]$.
  $1/G(u_i)$ is roughly $O(\log n)$.
  used ℓ_1 well-conditioned basis argument to prove for all x.
- Contraction bound: $w.h.p.$ $\Omega(1/(d \log n)\|Ax\|_G) \leq \|D^{-1}Ax\|_\infty$
  Proof sketch: with probability at least $1 - e^{-\alpha}$, $\|f(x)\|_\infty \geq \|x\|_G/\alpha$.
  ϵ net argument to prove for all x.
- $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \|\cdot\|_G$, we can achieve:

$$\Omega((1/d \log n)\|Ax\|_G) \leq \|D^{-1}Ax\|_2 \leq O(d^2 \log(n)\|Ax\|_G)$$

Our Subspace Embedding and Proof Sketch

- Given $A \in \mathbb{R}^{n\times m}$ with rank d. Generate diagonal matrix D with each entry from CDF $1 - e^{-G(t)}$, then:
- Dilation bound: $w.h.p., \forall x \in \mathbb{R}^m$, $\|D^{-1}Ax\|_G \leq O(d^2 \log(n)\|Ax\|_G)$
  Proof sketch: $f : \mathbb{R}^n \to \mathbb{R}^n$:
  $(x_1, x_2, ..., x_n) \to (x_1/u_1, x_2/u_2, ..., x_n/u_n)$. We want to bound the probability of $\|f(x)/\gamma\alpha\|_G \geq 1$, and thus $\mathbb{E}[\sum_{i=1}^n G(x_i\gamma)/(\alpha u_i)]$.
  $1/G(u_i)$ is roughly $O(\log n)$.
  used ℓ_1 well-conditioned basis argument to prove for all x.
- Contraction bound: $w.h.p.$ $\Omega(1/(d \log n)\|Ax\|_G) \leq \|D^{-1}Ax\|_\infty$
  Proof sketch: with probability at least $1 - e^{-\alpha}$, $\|f(x)\|_\infty \geq \|x\|_G/\alpha$.
  ϵ net argument to prove for all x.
- $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \|\cdot\|_G$, we can achieve:

$$\Omega((1/d \log n)\|Ax\|_G) \leq \|D^{-1}Ax\|_2 \leq O(d^2 \log(n)\|Ax\|_G)$$

Linear Regression with Orlicz norm $\|\cdot\|_G$

- Input: $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$

Linear Regression with Orlicz norm $\|\cdot\|_G$

- Input: $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$
- $\operatorname{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_G$

Linear Regression with Orlicz norm $\|\cdot\|_G$

- Input: $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$
- $\mathrm{argmin}_{x\in\mathbb{R}^d} \|Ax - b\|_G$
- $S = D^{-1}$, $A' = SA, b' = Sb$

Linear Regression with Orlicz norm $\|\cdot\|_G$

- Input: $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$
- $\text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_G$
- $S = D^{-1}$, $A' = SA, b' = Sb$
- Solve ℓ_2 linear regression with A', b', a.k.a. $\hat{x} = (A'^T A')^{-1} A'^T b'$

Linear Regression with Orlicz norm $\|\cdot\|_G$

- Input: $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$
- $\operatorname{argmin}_{x\in\mathbb{R}^d} \|Ax - b\|_G$
- $S = D^{-1}$, $A' = SA, b' = Sb$
- Solve ℓ_2 linear regression with A', b', a.k.a. $\hat{x} = (A'^T A')^{-1} A'^T b'$
- Let $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} \|Ax - b\|_G$, we have:
  $\|A\hat{x} - b\|_G \le O(d\log^2 n)\|Ax^* - b\|_G$

Outline

1. Orlicz Norm and the Motivation
   - General Linear Regression
   - Huber Loss
   - Orlicz Norm
2. Solving Tasks Defined on Orlicz Norms
   - Subspace embedding method for “nice” Orlicz Norm
   - Orlicz Norm Linear Regression
3. Experiments on Orlicz Norm Linear Regression
   - Comparison with ℓ_1 and ℓ_2 regression
   - Best δ under Different Outlier Size
   - Beyond Huber function - A General Framework
4. Conclusion and Future Work

Comparison with ℓ_1 and ℓ_2 regression

- Let $\|\cdot\|_G$ be induced by huber function with $\delta = 0.75$

$$G(x) = \begin{cases} \frac{x^2}{2} & |x| \le \delta \\ \delta(|x| - \delta/2) & \text{otherwise} \end{cases}$$

Comparison with ℓ_1 and ℓ_2 regression

- Let $\|\cdot\|_G$ be induced by huber function with $\delta = 0.75$

$$G(x) = \begin{cases} \frac{x^2}{2} & |x| \leq \delta \\ \delta(|x| - \delta/2) & \text{otherwise} \end{cases}$$

- 3 types of noises:
  - Gaussian noise
  - Sparse outliers on 3% of the entries
  - Gaussion noise + sparse outliers

Comparison with ℓ_1 and ℓ_2 regression

- Let $\|\cdot\|_G$ be induced by huber function with $\delta = 0.75$

$$G(x) = \begin{cases} \frac{x^2}{2} & |x| \le \delta \\ \delta(|x| - \delta/2) & \text{otherwise} \end{cases}$$

- 3 types of noises:
  - Gaussian noise
  - Sparse outliers on 3% of the entries
  - Gaussion noise + sparse outliers

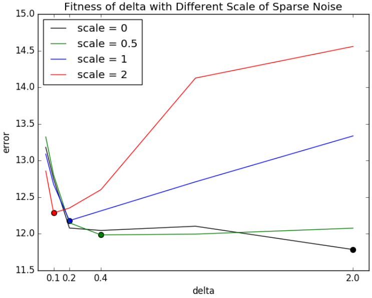
	Gaussian	Sparse	Mixed
best performing	ℓ_2	ℓ_1	$G_{\delta=0.75}$ (Orlicz)
worst performing	ℓ_1	ℓ_2	ℓ_1

Best δ under Different Outlier Size

- Mixed noise with sparse noise of size $[-s\|Ax^*\|_2, s\|Ax^*\|_2]$, scale s from [0, 0.5, 1, 2]. Comparisons between Orlicz norm regression δ varying from [0.05, 0.1, 0.2, 0.4, 1, 2].

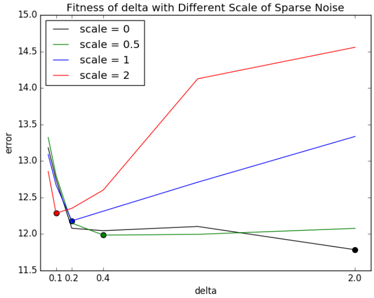
Best δ under Different Outlier Size

- Mixed noise with sparse noise of size $[-s\|Ax^*\|_2, s\|Ax^*\|_2]$, scale s from [0, 0.5, 1, 2]. Comparisons between Orlicz norm regression δ varying from [0.05, 0.1, 0.2, 0.4, 1, 2].



-

Best δ under Different Outlier Size

- Mixed noise with sparse noise of size $[-s\|Ax^*\|_2, s\|Ax^*\|_2]$, scale s from [0, 0.5, 1, 2]. Comparisons between Orlicz norm regression δ varying from [0.05, 0.1, 0.2, 0.4, 1, 2].
- 

scale s	best δ
0	2
0.5	1
1	0.5
2	0

A General Framework: Picking "whichever G" you want

- Consider the Orlicz norm G induced by the following function:

$$G_{\ell_{1.5}}(x) = \begin{cases} x^{1.5}/1.5 & |x| \le \delta \\ \delta^{0.5}(|x| - \delta/3) & \text{otherwise} \end{cases}$$

A General Framework: Picking "whichever G" you want

- Consider the Orlicz norm G induced by the following function:

$$G_{\ell_{1.5}}(x) = \begin{cases} x^{1.5}/1.5 & |x| \leq \delta \\ \delta^{0.5}(|x| - \delta/3) & \text{otherwise} \end{cases}$$

- 1 sparse outlier with size scale $s = 100$.

A General Framework: Picking "whichever G" you want

- Consider the Orlicz norm G induced by the following function:

$$G_{\ell_{1.5}}(x) = \begin{cases} x^{1.5}/1.5 & |x| \leq \delta \\ \delta^{0.5}(|x| - \delta/3) & \text{otherwise} \end{cases}$$

- 1 sparse outlier with size scale $s = 100$.
- $\mathbf{G}_{\ell_{1.5}} \leq \ell_1 \leq \ell_{1.5} \leq G_{\delta=0.25} \leq G_{\delta=0.75} \leq \ell_2$

A General Framework: Picking “whichever G” you want

- Consider the Orlicz norm G induced by the following function:

$$G_{\ell_{1.5}}(x) = \begin{cases} x^{1.5}/1.5 & |x| \leq \delta \\ \delta^{0.5}(|x| - \delta/3) & \text{otherwise} \end{cases}$$

- 1 sparse outlier with size scale $s = 100$.
- $\mathbf{G}_{\ell_{1.5}} \leq \ell_1 \leq \ell_{1.5} \leq G_{\delta=0.25} \leq G_{\delta=0.75} \leq \ell_2$
- Orlicz Norm Regression is a general and flexible framework.

Conclusion and Future Work

- Distortion factor $O(d \log^2 n)$ is large; can we improve this if we assume the input distribution are under some statistical distribution?

Conclusion and Future Work

- Distortion factor $O(d \log^2 n)$ is large; can we improve this if we assume the input distribution are under some statistical distribution?
- Calculate the actual approximation ratio with some slower but more accurate optimization algorithm.

Conclusion and Future Work

- Distortion factor $O(d \log^2 n)$ is large; can we improve this if we assume the input distribution are under some statistical distribution?
- Calculate the actual approximation ratio with some slower but more accurate optimization algorithm.
- Examine whether the sketching preserves the statistical properties of the regression error.

Conclusion and Future Work

- Distortion factor $O(d \log^2 n)$ is large; can we improve this if we assume the input distribution are under some statistical distribution?
- Calculate the actual approximation ratio with some slower but more accurate optimization algorithm.
- Examine whether the sketching preserves the statistical properties of the regression error.
- $1 + \epsilon$ approximation with non-oblivious embedding.

Thanks for watching! Questions?
Poster: # 18