TIME/ACCURACY TRADEOFFS FOR LEARNING A RELU WITH RESPECT TO GAUSSIAN MARGINALS

Surbhi Goel **Sushrut Karmalkar** **Adam Klivans**

The University of Texas at Austin

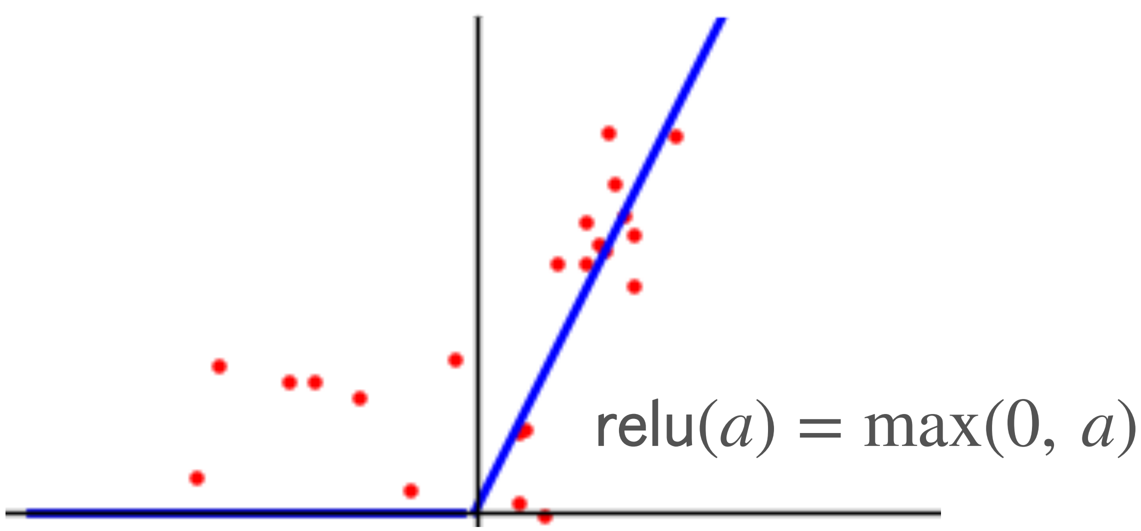
WHAT IS RELU REGRESSION?

Given: Samples drawn from distribution $\mathscr{D}$ with arbitrary labels

Output: $\widehat{w} \in \mathbb{R}^d$ such that

$$\mathop{\mathbb{E}}_{\mathscr{D}}\left[\left(\mathrm{relu}(\widehat{w}\cdot x) - y\right)^2\right] \le \mathrm{opt} + \epsilon$$

test error



$$\mathrm{opt} := \min_{w}\left(\mathop{\mathbb{E}}_{\mathscr{D}}\left[\left(\mathrm{relu}(w\cdot x) - y\right)^2\right]\right)$$

loss of the best-fitting ReLU

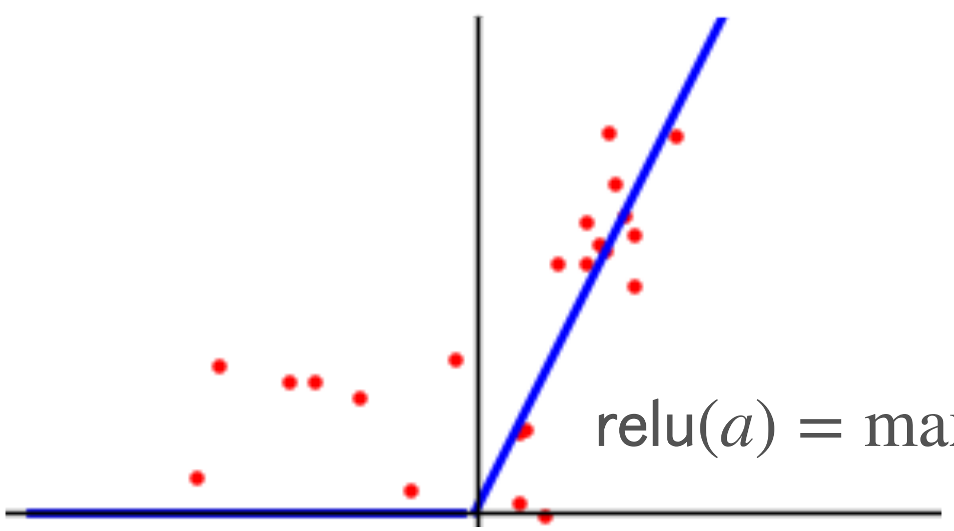
WHAT IS RELU REGRESSION?

Given: Samples drawn from distribution $\mathscr{D}$ with arbitrary labels

Output: $\widehat{w} \in \mathbb{R}^d$ such that

$$\mathop{\mathbb{E}}_{\mathscr{D}}\left[\left(\mathrm{relu}(\widehat{w}\cdot x) - y\right)^2\right] \leq \mathrm{opt} + \epsilon$$

test error



$\mathrm{relu}(a) = \max(0, a)$

$$\mathrm{opt} := \min_{w}\left(\mathop{\mathbb{E}}_{\mathscr{D}}\left[\left(\mathrm{relu}(w\cdot x) - y\right)^2\right]\right)$$

loss of the best-fitting ReLU

The underlying optimization problem is non-convex!

PRIOR WORK - POSITIVE

Mean-zero noise: Isotonic regression over Sphere [Kalai-Sastry'08, Kakade-Kalai-Kanade-Shamir'11]

Noiseless: Gradient descent over Gaussian input [Soltanolkotabi'17]

PRIOR WORK - POSITIVE

Mean-zero noise: Isotonic regression over Sphere [Kalai-Sastry'08, Kakade-Kalai-Kanade-Shamir'11]

Noiseless: Gradient descent over Gaussian input [Soltanolkotabi'17]

Results require strong restrictions on the input or the label

PRIOR WORK - NEGATIVE

Minimizing training loss is **NP-hard** [Manurangsi-Reichman'18]

Hardness over uniform on the **boolean cube** [**G**-Kanade-**K**-Thaler'17]

PRIOR WORK - NEGATIVE

Minimizing training loss is **NP-hard** [Manurangsi-Reichman'18]

Hardness over uniform on the **boolean cube** [**G**-Kanade-**K**-Thaler'17]

Results use special discrete distributions to prove hardness

DISTRIBUTION ASSUMPTION

Assumption: For all $(x, y) \sim \mathscr{D}$, $x \sim \mathcal{N}(0, I_d)$ and $y \in [0,1]$

DISTRIBUTION ASSUMPTION

Assumption: For all $(x, y) \sim \mathscr{D}$, $x \sim \mathscr{N}(0, I_d)$ and $y \in [0,1]$

Gaussian input allows for positive results in noiseless setting

[Tian'17, Soltanolkotabi'17, Li-Yuan'17, Zhong-Song-Jain-Bartlett-Dhillon'17, Brutzkus-Globerson'17, Zhong-Song-Dhillon17, Du-Lee-Tian-Poczos-Singh'18, Zhang-Yu-Wang-Gu'19, Fu-Chi-Liang'19……]

DISTRIBUTION ASSUMPTION

Assumption: For all $(x, y) \sim \mathcal{D}$, $x \sim \mathcal{N}(0, I_d)$ and $y \in [0,1]$

Gaussian input allows for positive results in noiseless setting

[Tian'17, Soltanolkotabi'17, Li-Yuan'17, Zhong-Song-Jain-Bartlett-Dhillon'17, Brutzkus-Globerson'17, Zhong-Song-Dhillon17, Du-Lee-Tian-Poczos-Singh'18, Zhang-Yu-Wang-Gu'19, Fu-Chi-Liang'19…….]

Explicitly compute closed-form expressions for loss/gradient

HARDNESS RESULT

There exists NO algorithm for ReLU regression up to error ϵ in time $d^{o\left(\log(1/\epsilon)\right)}$ under standard computational hardness assumptions.

HARDNESS RESULT

There exists NO algorithm for ReLU regression up to error ϵ in time $d^{o\left(\log(1/\epsilon)\right)}$ under standard computational hardness assumptions.

The problem is as hard as learning sparse parities with noise!

HARDNESS RESULT

There exists NO algorithm for ReLU regression up to error ϵ in time $d^{o\left(\log(1/\epsilon)\right)}$ under standard computational hardness assumptions.

The problem is as hard as learning sparse parities with noise!

First hardness result under the Gaussian assumption!

HARDNESS FOR GRADIENT DESCENT

Unconditionally, NO statistical query (SQ) algorithm with bounded norm queries can perform ReLU regression up to error ϵ with less than $d^{o\left(\log(1/\epsilon)\right)}$ queries.

HARDNESS FOR GRADIENT DESCENT

Unconditionally, NO statistical query (SQ) algorithm with bounded norm queries can perform ReLU regression up to error ϵ with less than $d^{o\left(\log(1/\epsilon)\right)}$ queries.

Gradient Descent (GD) is well-known to be an SQ algorithm

HARDNESS FOR GRADIENT DESCENT

Unconditionally, NO statistical query (SQ) algorithm with bounded norm queries can perform ReLU regression up to error ϵ with less than $d^{o\left(\log(1/\epsilon)\right)}$ queries.

Gradient Descent (GD) is well-known to be an SQ algorithm

GD can NOT solve ReLU regression in polynomial time

HARDNESS FOR GRADIENT DESCENT

Unconditionally, NO statistical query (SQ) algorithm with bounded norm queries can perform ReLU regression up to error ϵ with less than $d^{o\left(\log(1/\epsilon)\right)}$ queries.

Gradient Descent (GD) is well-known to be an SQ algorithm

GD can NOT solve ReLU regression in polynomial time

Recall GD works in noiseless setting [Soltanokotabi'17]

APPROXIMATION RESULT

There exists an an algorithm for ReLU regression with error $O\left(\mathrm{opt}^{2/3}\right)+\epsilon$ in time poly $\left(d, 1/\epsilon\right)$.

APPROXIMATION RESULT

There exists an an algorithm for ReLU regression with error $O\left(\mathrm{opt}^{2/3}\right) + \epsilon$ in time $\mathrm{poly}\left(d, 1/\epsilon\right)$.

Can get $O\left(\mathrm{opt}\right) + \epsilon$ in time $\mathrm{poly}\left(d, 1/\epsilon\right)$

[Diakonikolas-**G**-**K**-**K**-Soltanolkotabi'TBD]

APPROXIMATION RESULT

There exists an an algorithm for ReLU regression with error $O\left(\mathrm{opt}^{2/3}\right) + \epsilon$ in time poly $\left(d, 1/\epsilon\right)$.

Can get $O\left(\mathrm{opt}\right) + \epsilon$ in time poly $\left(d, 1/\epsilon\right)$

[Diakonikolas-**G**-**K**-**K**-Soltanolkotabi'TBD]

Finding approximate solutions is tractable!

THANK YOU!

Poster @ East Exhibition Hall B + C #235