We give the first dimension-efficient algorithms for learning Rectified Linear Units (ReLUs), which are functions of the form $\max(0, w \cdot x)$ with w a unit vector (2-norm equal to 1). Our algorithm works in the challenging Reliable Agnostic learning model of Kalai, Kanade, and Mansour where the learner is given access to a distribution D on labeled examples, but the labeling may be arbitrary. We construct a hypothesis that simultaneously minimizes the false-positive rate and the $l_p$ loss of inputs given positive labels by $D$.

It runs in polynomial-time (in $n$) with respect to *any* distribution on $\mathbb{S}^{n-1}$ (the unit sphere in n dimensions) and for any error parameter $\epsilon = \Omega(1/\log n)$. These results are in contrast to known efficient algorithms for reliably learning linear threshold functions, where epsilon must be $\Omega(1)$ and strong assumptions are required on the marginal distribution. We can compose our results to obtain the first set of efficient algorithms for learning constant-depth networks of ReLUs.

Surprisingly, the following question remains open: does there exists a fully polynomial-time algorithm (polynomial in the dimension and accuracy parameter) for learning even a single ReLU.

Joint work with Surbhi Goel, Varun Kanade, and Justin Thaler

# 1 Intro

The RELU is a function $\max(0, w \cdot x)$. We think about it as a function $\mathbb{S}^{n-1} \to [0, 1]$. This is a popular activation function in deep nets. We talk about the simplest net: just one activation function. There are simple questions involving simple architectures that we don't know what to solve.

Consider linear regression.

$$\min_w \mathop{\mathbb{E}}_{(x,y)\sim D}[(\langle w, x\rangle - y)^2]$$

where $D$ is on $\mathbb{S}^{n-1} \times [0, 1]$. This is easy.

On the other hand, there is the question of learning a linear separator.

$$\min_w \mathbb{P}_{(x,y)\sim D}[\text{sign}(w \cdot x) \neq y]$$

This problem is hard—there are computationally intractability results.

Suppose the goal is to output $h$ such that

$$\mathbb{P}_{(x,y)\sim D}[h(x) \neq y] \leq OPT + \varepsilon.$$

(This is improper learning.) This is computationally intractable in a variety of cases. Amit Daniely showed that assuming certain CSP's are hard there is no polytime algorithm that works for every distribution. Klivans and Sherstove showed that if you can solve this you can break cryptographic schemes based on learning with errors.

One natural thing is to make an assumption about the marginal distribution, for example

1. Suppose $D$ is uniform distribution on $\{0, 1\}^n$. Assuming LWE is hard, you need time $n^{\Omega(\frac{1}{\varepsilon^{1.999}})}$.

2. For $D$ Gaussian (Klivans, Kothari), $n^{\Omega(\ln(\frac{1}{\varepsilon}))}$.

3. For $D$ uniform on $\mathbb{S}^{n-1}$, PTAS obtaining $(1+\gamma)OPT+\varepsilon$ but exponential in $\gamma$, $\text{poly}(n, \frac{1}{\varepsilon}2^{\frac{1}{\gamma}})$.

Consider $C$ the class of disjunctions. The goal is to output $h$ such that

$$\mathbb{P}_{(x,y)\sim D}(h(x)\neq y)\leq OPT+\varepsilon.$$

The best is $n^{\sqrt{n\ln(\frac{1}{\varepsilon})}}$. For polytime algorithms, you can get $n^{\frac{1}{3}}OPT+\varepsilon$. No distributional assumptions.

You're tempted to say, "Your model is wrong, has no relation to anything in ML."

(These results allow arbitrary labels. Moving away from agnostic is my least favorite thing to do. Ben-David 2000: if you have margins, exponent is with respect to margin.)

I will show how to learn $\max(0, w\cdot x)$ in time $2^{O(\frac{1}{\varepsilon})}\,\text{poly}(n)$ over any distribution, with $OPT+\varepsilon$.

This is considerably better than what we can do for halfspace. Think of this as 1-sided halfspace. If it's negative, threshold it; if it's positive, don't change the value.

Minimize both

1. False positive rate $\mathbb{P}_{(x,y)\sim D}[h(x)\neq 0\wedge y=0]$

2. loss (square-loss) on points given non-values.

In reliable agnostic learning (Kalai, Kanade, Mansour), minimize

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(h(x)-y)^2\mathbb{1}(y>0)].$$

First consider the vanilla agnostic model:

$$OPT=\min_{w}\mathop{\mathbb{E}}_{(x,y)\sim D}[(\max(0,w\cdot x)-y)^2] \tag{1}$$

$$\mathbb{E}[(h(x)-y)^2]\leq OPT+\varepsilon. \tag{2}$$

This can be done in $2^{O(\frac{1}{\varepsilon})}\,\text{poly}(n)$.

Let $C^+$ be ReLUs with zero false positive rate wrt $D$. In reliable agnostic learning, we want

$$OPT=\min_{c\in C^+}\mathbb{E}[(\cdot-y)^2].$$

The goal is to output $h$ such that $\mathbb{P}_{(x,y)\sim D}[h(0)\neq 0\wedge y=0]\leq\varepsilon$. and $\mathbb{E}_{(x,y)\sim D}[(h(x)-y)^2\mathbb{1}(y>0)]\leq OPT+\varepsilon$.

Ex. comment moderation: some comments are trolls. For the remaining, you want to score/rank according to relevance.

Mote this is not a convex minimization problem because you have a max inside a square or absolute value; this makes the problem nonconvex.

We want to solve

$$\min_w \frac{1}{m} \sum_{1 \le i \le m, y^i > 0} l(\max(0, w \cdot x'), y^i)$$

under constraints $\max(0, w \cdot x^j) = 0$ where $j$ is such that $y^j = 0$. Constraints minimize false-positive while simultaneously minimizing loss.

We can write $\max(0, w \cdot x) = \frac{|w \cdot x|}{2} + \frac{w \cdot x}{2}$. We first find a polynomial $p$ that approximates $|z|$. The great thing about ReLU is that it is continuous. I can apply Jackson's theorem from approximation theory:

**Theorem 1.1.** *For any continuous function on $[-1, 1]$ with modulus of continuity 1, there exists a polynomial $p$ such that for all $x \in [-1, 1]$, $|p(x) - f(x)| \le \varepsilon$ and $p$ has degree $O\left(\frac{1}{\varepsilon}\right)$.*

Write $p = \sum_{m=1}^{\frac{1}{\varepsilon}} p_i x^i$, $\|p\|_2^2 \le 2^{O\left(\frac{1}{\varepsilon}\right)}$.

Cf. SSS used kernels to learn halfspaces.

We now want

$$\min_{p \text{ in } n \text{ variables of degree } O\left(\frac{1}{\varepsilon}\right)} \frac{1}{m} \sum_{1 \le i \le m, y^i > 0} l(p(\cdot, x^i, y^i))$$

under constraints $p(\cdot, x^j) \le \varepsilon$ for $j$ such that $y^j = 0$.

Outline: embed $x$ into a feature space via $\varphi$. view coefficients of polynomial $p$ as vector $v_w$,

$$p_w(x) = \langle \varphi(x), v_w \rangle \tag{3}$$

$$MK(x, x') = \sum_{i=1}^{d} \langle x, x' \rangle^i \tag{4}$$

Different orderings of monomials will be different terms.

Noisy polynomial reconstruction: $\min_{\deg p \le d, \|p\|_2^2 \le B} \mathbb{E}_{(x,y) \sim D, x \sim \mathbb{S}^{n-1}}[(p(x) - y)^2]$ for any distribution, in time $\text{poly}(n, d, B, \frac{1}{\varepsilon}, \ldots)$.

Hardest is showing the loss function generalizes. Use a Rademacher bound.

Some other results.

- Sum of $k$ ReLUs with $\|w\|_2 = 1$ in time $2^{O\left(\frac{\sqrt{k}}{\varepsilon}\right)} \text{poly}(n)$.
- For the sum of $k$ sigmoids, $\text{poly}(h, k, \frac{1}{\varepsilon})$ where $\sum_i w_i \sigma_i$, $\|w\| = 1$.