Learning Distributions Generated by One-Layer ReLU Networks

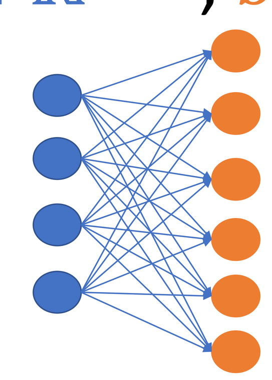
Shanshan Wu, Alex Dimakis, Sujay Sanghavi

University of Texas at Austin

[Background]

- A popular generative model these days is passing a standard Gaussian distribution thru a Neural Net.
- Consider a one-layer ReLU generative model with parameters $W \in \mathbb{R}^{d\times k}$, $b \in \mathbb{R}^d$:

$z \in \mathbb{R}^k$
$z \sim N(0, I_{k\times k})$



$x = \mathrm{ReLU}(Wz+b) \in \mathbb{R}^d$

A distribution over $\mathbb{R}^d$

- Given i.i.d. samples of x, parameters can be learned by training a GAN or VAE, but no guarantees are known.

[Problem Formulation]

$z \sim N(0, I_{k\times k})$

Given n i.i.d. samples $x_1, x_2, \dots x_n \sim \mathrm{ReLU}(Wz+b)$

🤔

Can we estimate W, b?

Note: This is an unsupervised learning problem.
We only observe x (the variable z is hidden from us).

[Identifiability]

- Is $W \in \mathbb{R}^{k\times d}$ identifiable from the distribution $\mathrm{ReLU}(Wz+b)$?

Only WW^T can be possibly identified.

Fact: $W_1W_1^T = W_2W_2^T$
$\Downarrow$
distr of $\mathrm{ReLU}(W_1z+b)$ = distr of $\mathrm{ReLU}(W_2z+b)$

- Is $b \in \mathbb{R}^d$ identifiable from the distribution $\mathrm{ReLU}(Wz+b)$?

Yes, but if b is negative, estimate b needs $\Omega(\exp(\|b\|_\infty^2))$ samples.

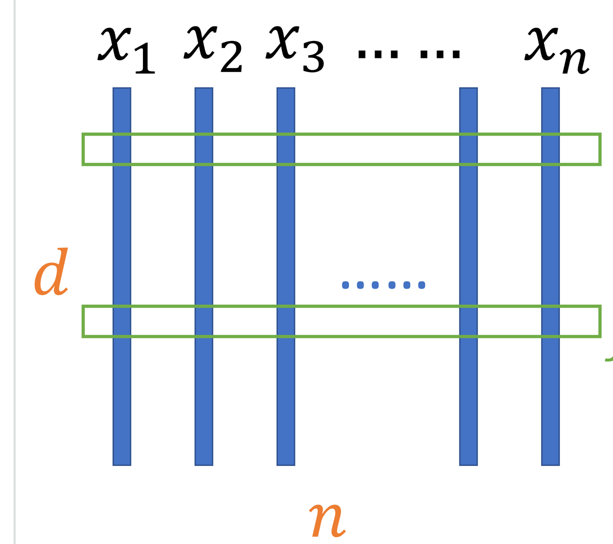
Assumption: b is non-negative.

[Our Algorithm]

Overview The (i,j)-th entry of WW^T is:

$$\langle W(i,:), W(j,:)\rangle = \|W(i,:)\|_2\|W(j,:)\|_2 \cos\theta_{ij}$$

Our algorithm:



i-th coordinate → **Step 1**: for $i \in [d]$
Estimate $\|W(i,:)\|_2$ & $b(i)$

j-th coordinate → **Step 2**: for $i \neq j \in [d]$
Estimate θ_{ij}

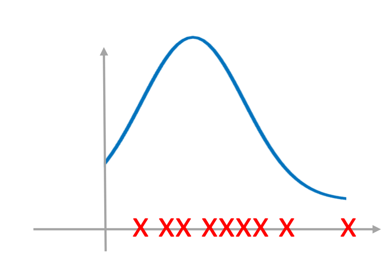
Step 1

Goal: estimate $\|W(i,:)\|_2$ & $b(i)$

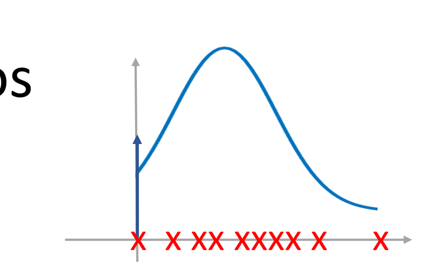
[Daskalakis et al., 2018]

The i-th coordinate
$x(i) \sim \mathrm{ReLU}(W(i,:)z + b(i))$

PDF of $x(i)$



Remove zeros



Samples ~ 1D truncated Gaussian

$\{x \in \mathbb{R}: x > 0\}$ $\quad N(b(i), \|W(i,:)\|_2^2)$

Samples ~ $x(i)$

Step 2

The i-th and j-th coordinate
$x(i) \sim \mathrm{ReLU}(W(i,:)z + b(i))$
$x(j) \sim \mathrm{ReLU}(W(j,:)z + b(j))$

Goal: estimate θ_{ij} := angle b/t $W(i,:)$ & $W(j,:)$

Fact: $\mathbb{P}_x[x(i) > b(i) \,\&\, x(j) \geq b(j)] = \frac{\pi - \theta_{ij}}{2\pi}$

$b(i), b(j) \geq 0$ $\quad z \sim N(0, I_{k\times k})$

$\mathbb{P}_z[W(i,:)z > 0 \,\&\, W(j,:)z > 0]$

Given $\hat{b}(i), \hat{b}(j)$ from Step 1:
$\hat{b}(i) \approx b(i)$, $\hat{b}(j) \approx b(j)$ ➡ $\mathbb{P}_x[x(i) > \hat{b}(i) \,\&\, x(j) \geq \hat{b}(j)] \approx \frac{\pi - \theta_{ij}}{2\pi}$

[Sample Complexity]

- **Parameter Estimation:**

[Main Theorem] Assuming $b^* \in \mathbb{R}^d$ is non-negative, then our algo takes $\tilde{O}(\frac{1}{\epsilon^2}\ln(\frac{d}{\delta}))$ samples ~ $\mathrm{ReLU}(W^*z+b^*)$ and its output satisfies w.p. at least $1-\delta$,

$$\|\widehat{WW^T} - W^*W^{*T}\|_F \leq \epsilon\|W^*\|_F^2, \quad \|\hat{b} - b^*\|_2 \leq \epsilon\|W^*\|_F$$

Our algorithm	Lower bound
$\tilde{O}(\frac{1}{\epsilon^2}\ln(\frac{d}{\delta}))$	$\Omega(\frac{1}{\epsilon^2})$

Our algorithm is optimal

- **Total Variation Distance**: $\mathrm{TV}(D, \mathrm{ReLU}(W^*z+b^*)) \leq \epsilon$

Our algorithm	Lower bound
$\tilde{O}(\frac{\kappa^2 d^2}{\epsilon^2}\ln(\frac{d}{\delta}))$	$\Omega(\frac{d}{\epsilon^2})$

This gap comes from:
1) Param est → TV distance
2) Lower bound is not tight

[Experiments]

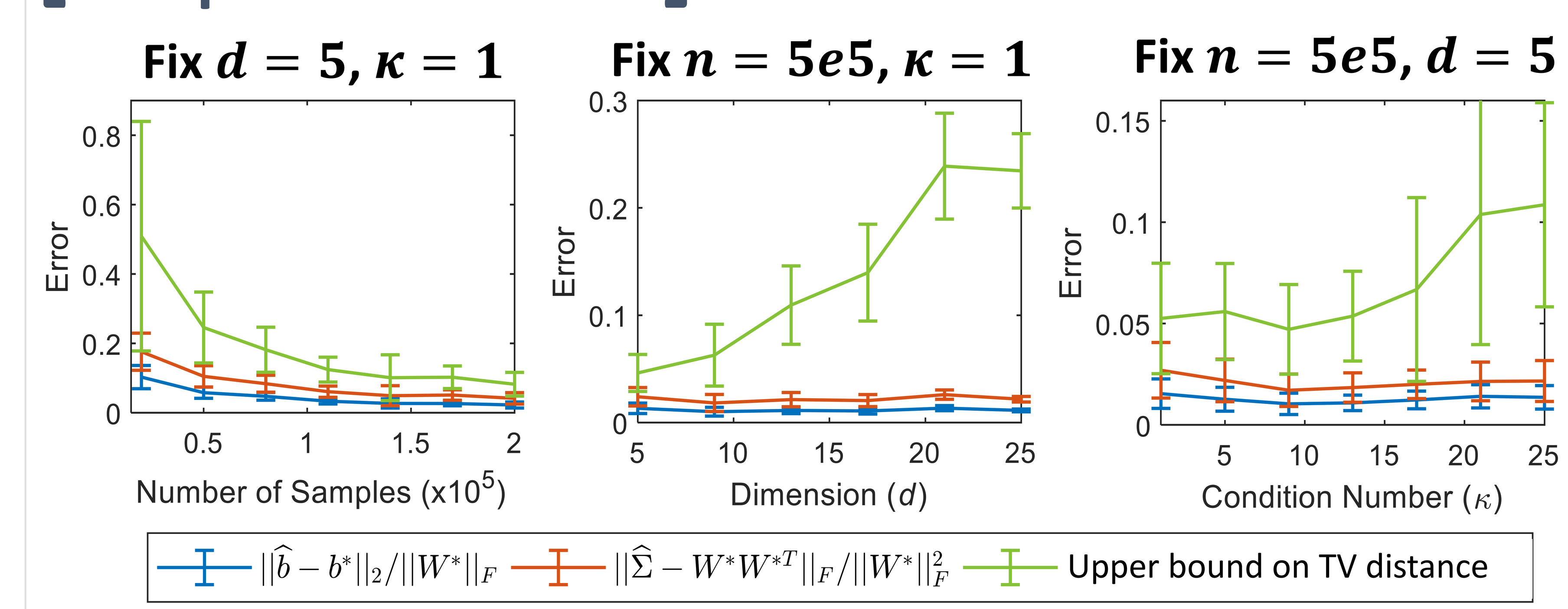


Figure 1. Empirical performance of our algorithm w.r.t. three parameters: number of samples n, dimension d, and condition number κ. Every point is the mean and standard deviation over 10 runs.

[Code] *https://github.com/wushanshan/densityEstimation*

[Open Problems]

- What if b^* has negative values?
- Two-layer generative model?
- Noisy samples?……

See our paper for more discussions.