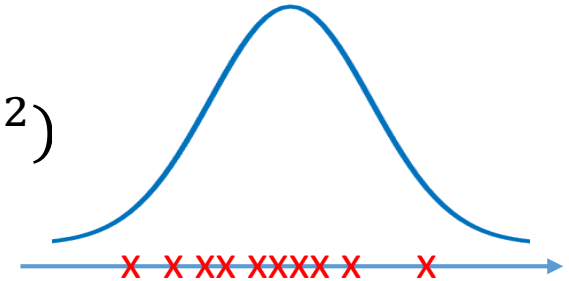
Learning Distributions Generated by One-Layer ReLU Networks

Shanshan Wu, Alex Dimakis, Sujay Sanghavi

[NeurIPS 2019]

Background: Density Estimation

Input: i.i.d. samples ~ unknown distribution



Goal: estimate the distribution

E.g., samples ~ $N(\mu, \sigma^2)$

Estimate μ, σ^2

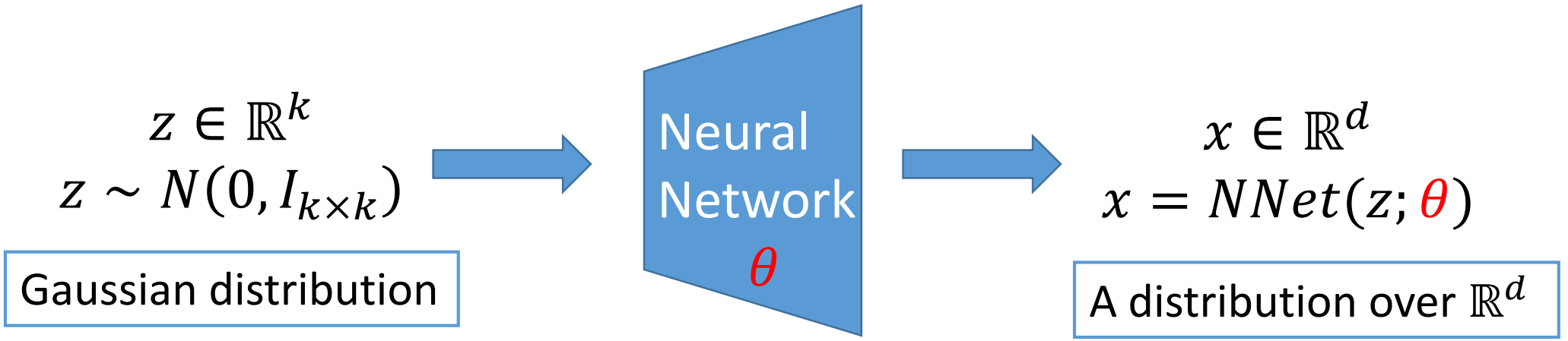
E.g., samples ~ natural images

ImageNet

Estimate the parameters of a generative model

Background: Generative Model

- A popular way to model high-dimensional complex distributions:



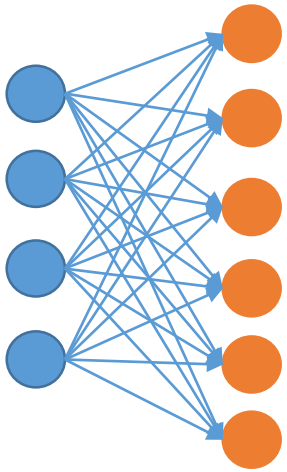
- Given i.i.d. samples, θ can be learned by
  - GAN [Goodfellow et al.'2014]
  - VAE [Kingma and Welling'2013]
  - …….

No guarantees & sample complexity

Problem Formulation

- One-layer ReLU generative model with params $W \in \mathbb{R}^{d\times k}$, $b \in \mathbb{R}^d$:

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

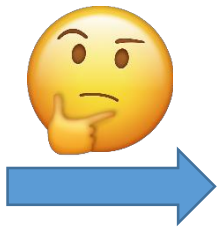


$$\mathrm{ReLU}(Wz + b) \in \mathbb{R}^d$$

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

- Our problem:

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

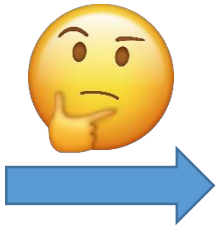


Can we estimate W, b?

Problem Formulation

- Our problem is unsupervised learning:

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



Can we estimate W, b?

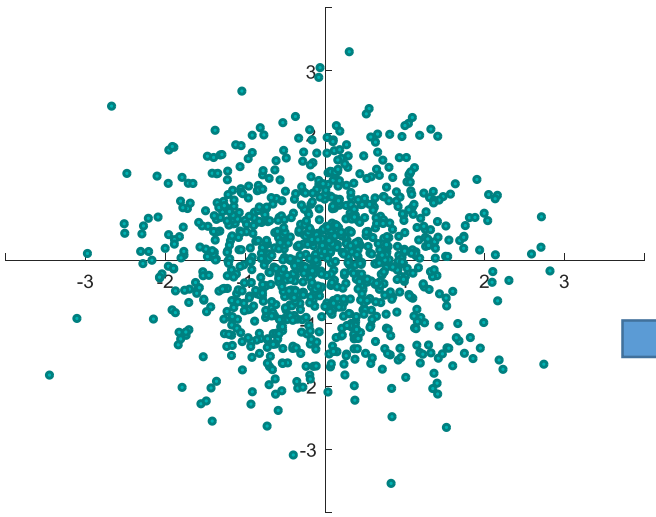
- Different from supervised learning problems [Ge et al., Goel et al., …]

Given n (input, output) pairs
$(z_1, x_1), \dots, (z_n, x_n)$ where
$x_i = \mathrm{ReLU}(Wz_i + b)$

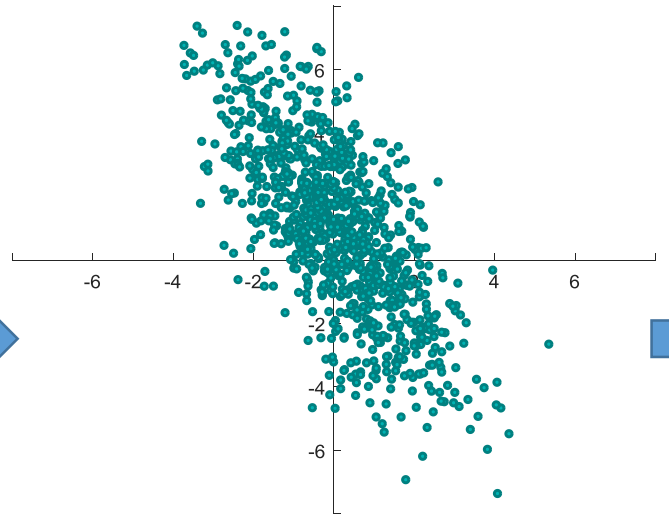
Can we estimate W, b?

Visualization of Our Problem

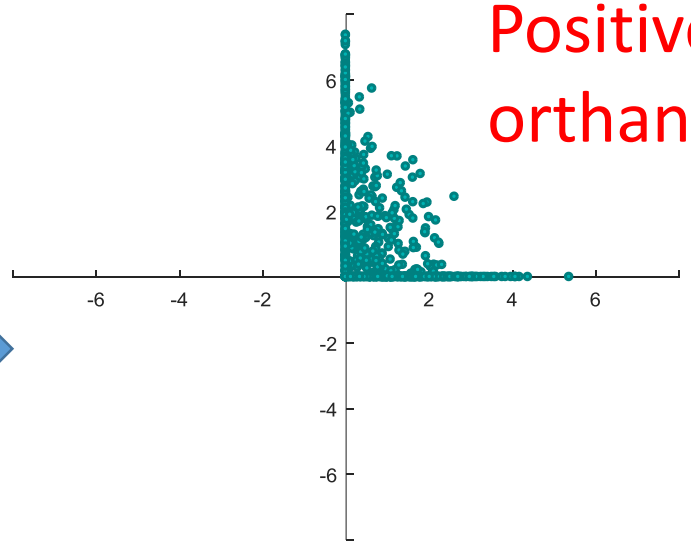
- $W^* = \begin{bmatrix} -0.9 & -1.2 \\ 0.4 & 2.7 \end{bmatrix}$, $b^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, distribution of $\mathrm{ReLU}(W^* z + b^*)$:



$z \sim N(0, I_{2\times 2})$



$W^* z + b^*$



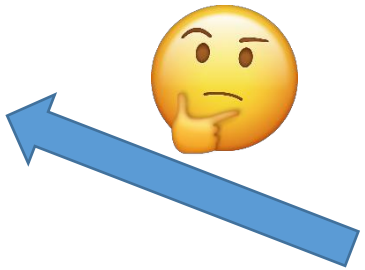
$\mathrm{ReLU}(W^* z + b^*)$

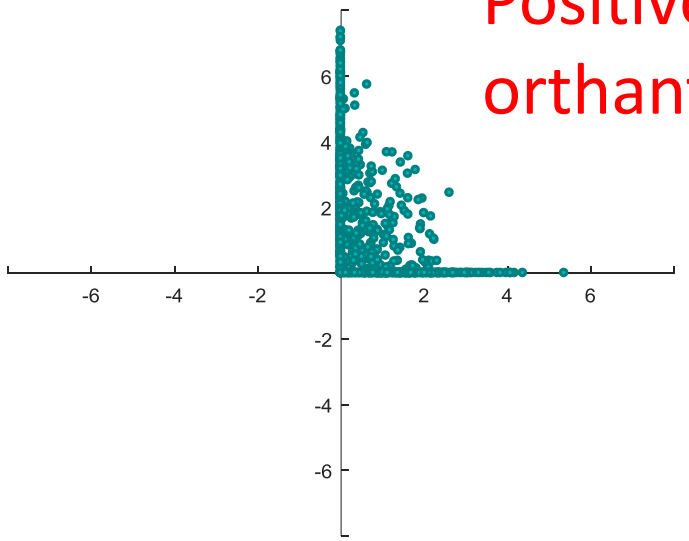
Positive orthant

Visualization of Our Problem

- $W^* = \begin{bmatrix} -0.9 & -1.2 \\ 0.4 & 2.7 \end{bmatrix}, b^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Can we recover W^*, b^*?





Positive orthant

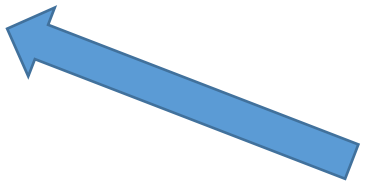
Samples $\sim \mathrm{ReLU}(W^* z + b^*)$

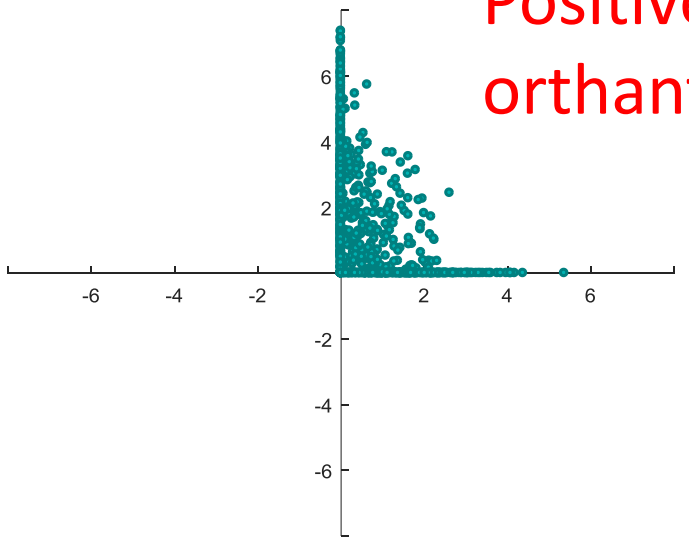
Visualization of Our Problem

- $W^* = \begin{bmatrix} -0.9 & -1.2 \\ 0.4 & 2.7 \end{bmatrix}, b^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Can we recover W^*, b^*?

Are they identifiable?





Positive orthant

Samples $\sim \mathrm{ReLU}(W^* z + b^*)$

Identifiability

- Is $W \in \mathbb{R}^{k \times d}$ identifiable from the distribution $\mathrm{ReLU}(Wz + b)$? No!

Only WW^T can be possibly identified.

- Fact:

$$W_1 W_1^T = W_2 W_2^T$$

$$\text{distr of } \mathrm{ReLU}(W_1 z + b) = \text{ distr of } \mathrm{ReLU}(W_2 z + b)$$

Identifiability

- Is $W \in \mathbb{R}^{k \times d}$ identifiable from the distribution $\mathrm{ReLU}(Wz + b)$? No!

Only WW^T can be possibly identified.

- Fact:

$$W_1 W_1^T = W_2 W_2^T$$

$$\text{distr of } \mathrm{ReLU}(\underbrace{W_1 z + b}_{N(b, W_1 W_1^T)}) = \text{distr of } \mathrm{ReLU}(\underbrace{W_2 z + b}_{N(b, W_2 W_2^T)})$$

Identifiability

- 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

Identifiability

- 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

Claim:

Distinguish $\mathrm{ReLU}(z - |b|)$ & $\mathrm{ReLU}(z - |b| - 0.1)$ needs $\Omega(\exp(b^2))$ samples.

Identifiability

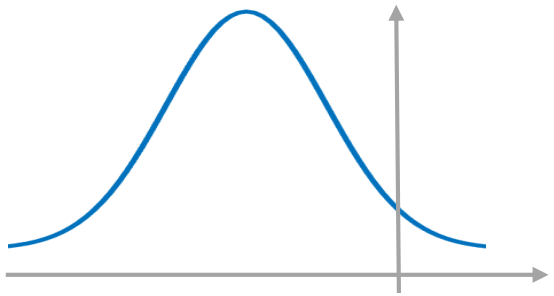
- Is $b \in \mathbb{R}^d$ identifiable from the distribution $\mathrm{ReLU}(Wz + b)$? Yes, but…

If b is negative

Estimate b needs $\Omega(\exp(\|b\|_\infty^2))$ samples

Claim:

Distinguish $\mathrm{ReLU}(z - |b|)$ & $\mathrm{ReLU}(z - |b| - 0.1)$ needs $\Omega(\exp(b^2))$ samples.



pdf of $z - |b|$

Identifiability

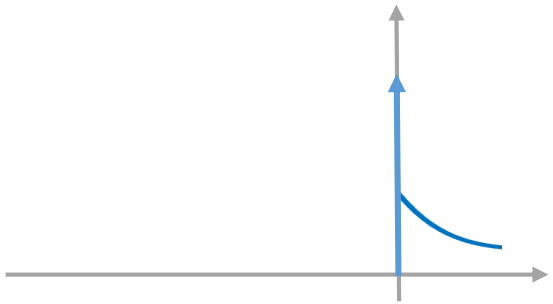
- Is $b \in \mathbb{R}^d$ identifiable from the distribution $\mathrm{ReLU}(Wz + b)$? Yes, but…

If b is negative

Estimate b needs $\Omega(\exp(\|b\|_\infty^2))$ samples

Claim:

Distinguish $\mathrm{ReLU}(z - |b|)$ & $\mathrm{ReLU}(z - |b| - 0.1)$ needs $\Omega(\exp(b^2))$ samples.



pdf of $\mathrm{ReLU}(z - |b|)$

$\mathbb{P}[\mathrm{ReLU}(z - |b|) > 0] \leq \exp(-b^2)$

Identifiability

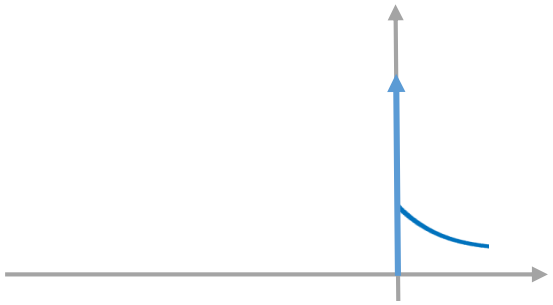
- Is $b \in \mathbb{R}^d$ identifiable from the distribution $\mathrm{ReLU}(Wz + b)$? Yes, but…

If b is negative

Estimate b needs $\Omega(\exp(\|b\|_\infty^2))$ samples

Claim:

Distinguish $\mathrm{ReLU}(z - |b|)$ & $\mathrm{ReLU}(z - |b| - 0.1)$ needs $\Omega(\exp(b^2))$ samples.



$\mathbb{P}[\mathrm{ReLU}(z - |b| - 0.1) > 0] \leq \exp(-b^2)$

pdf of $\mathrm{ReLU}(z - |b| - 0.1)$

Identifiability

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

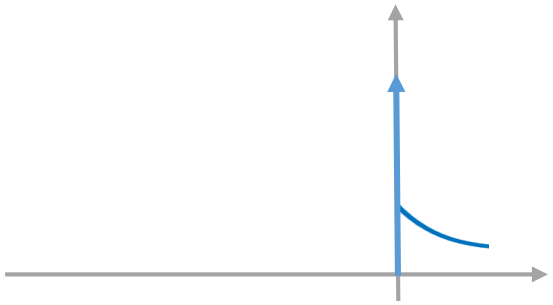
If b is negative

Estimate b needs $\Omega(\exp(\|b\|_\infty^2))$ samples

Assumption: b is non-negative

Claim:

Distinguish $\mathrm{ReLU}(z - |b|)$ & $\mathrm{ReLU}(z - |b| - 0.1)$ needs $\Omega(\exp(b^2))$ samples.



$\mathbb{P}[\mathrm{ReLU}(z - |b| - 0.1) > 0] \leq \exp(-b^2)$

pdf of $\mathrm{ReLU}(z - |b| - 0.1)$

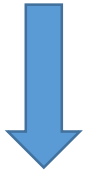
Our Problem

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

Observe $x_1, x_2, \ldots, x_n \sim \mathrm{ReLU}(W^* z + b^*)$

Unknown $\mathbb{R}^{d\times k}$

Unknown $\mathbb{R}^d$ & non-negative

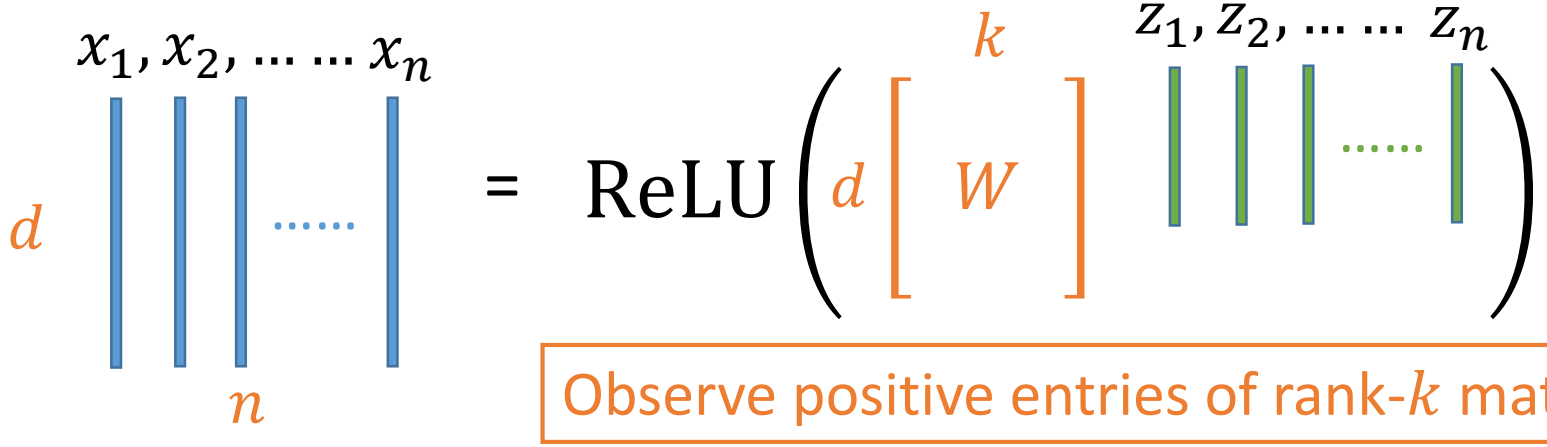


Goal: Estimate $W^* W^{*T}$ and b^*

Initial Attempts

- **Idea 1:** Maximum likelihood estimation
- Problem: 1) no closed-form
  2) non-convex

- **Idea 2:** Matrix completion



Observe positive entries of rank-k matrix

Experiments do not work!

Idea 3: Truncated Gaussian

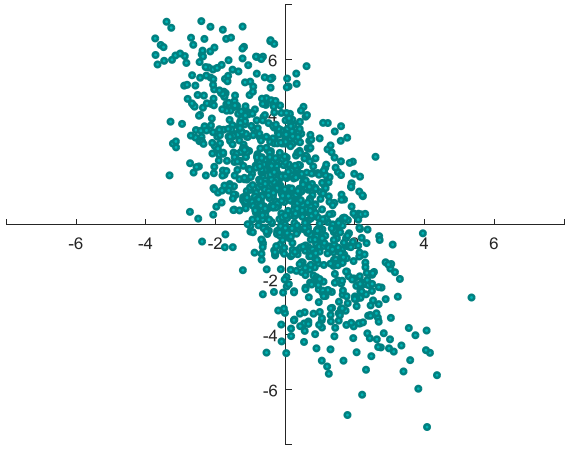
- Given a set $S \subset \mathbb{R}^d$, the pdf of an S-truncated Gaussian distribution

pdf of $N(\mu, \Sigma)$

$$q(x; \mu, \Sigma, S) = \begin{cases} \dfrac{p(x; \mu, \Sigma)}{\int_S p(y; \mu, \Sigma)\, \mathrm{d}y}, & \text{if } x \in S \\ 0, & \text{if } x \notin S \end{cases}$$

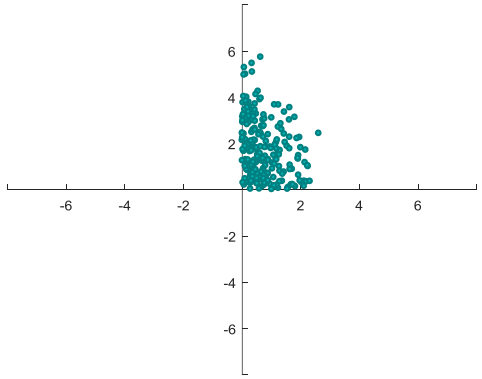
- Given truncated samples, estimate μ, Σ [Daskalakis et al., 2018]

Truncation versus ReLU

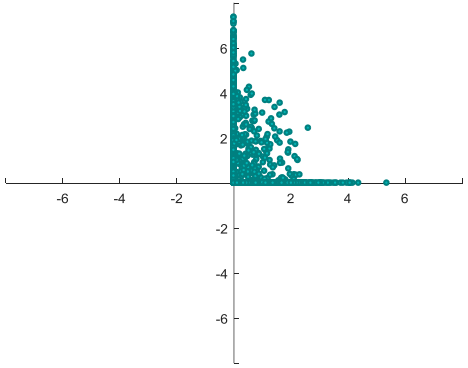


samples $\sim N(\mu, \Sigma)$

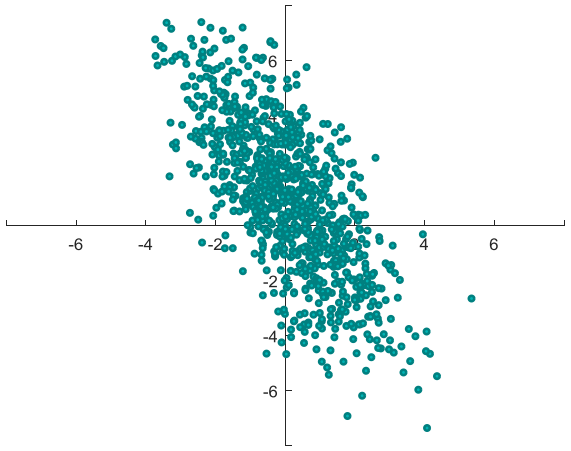
Truncation with $S = \{x \in \mathbb{R}^d : x > 0\}$



ReLU(samples)



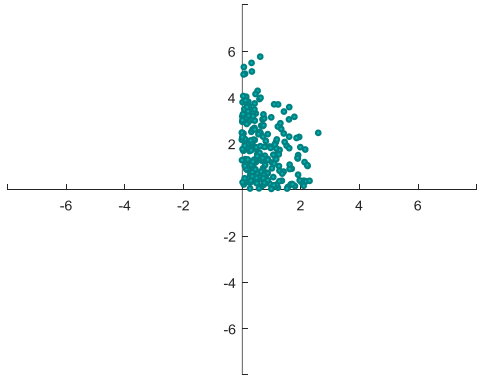
Convert ReLU to Truncated Samples

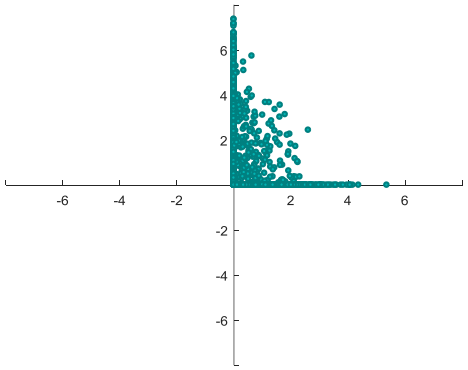


samples $\sim N(\mu, \Sigma)$

Truncation with
$S = \{x \in \mathbb{R}^d : x > 0\}$

ReLU(samples)





Remove samples outside S

Convert ReLU to Truncated Samples

Samples $\sim \mathrm{ReLU}(Wz + b)$

Remove the samples outside $S = \{x \in \mathbb{R}^d : x > 0\}$

Samples $\sim S$-truncated $N(b, WW^T)$

Run [Daskalakis et al., 2018]

Estimate b, WW^T

Convert ReLU to Truncated Samples

Samples $\sim \mathrm{ReLU}(Wz + b)$

Remove the samples outside $S = \{x \in \mathbb{R}^d : x > 0\}$

$$x = \begin{pmatrix} \mathrm{ReLU}(v^T z) \\ \mathrm{ReLU}(-v^T z) \end{pmatrix}$$

$\mathbb{P}[x \in S] = 0$

Samples $\sim S$-truncated $N(b, WW^T)$

Run [Daskalakis et al., 2018]

Estimate b, WW^T

Our Algorithm: Overview

- Our problem:

  Given $x_1, \dots, x_n \sim \mathrm{ReLU}(Wz + b)$, recover $WW^T \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$.

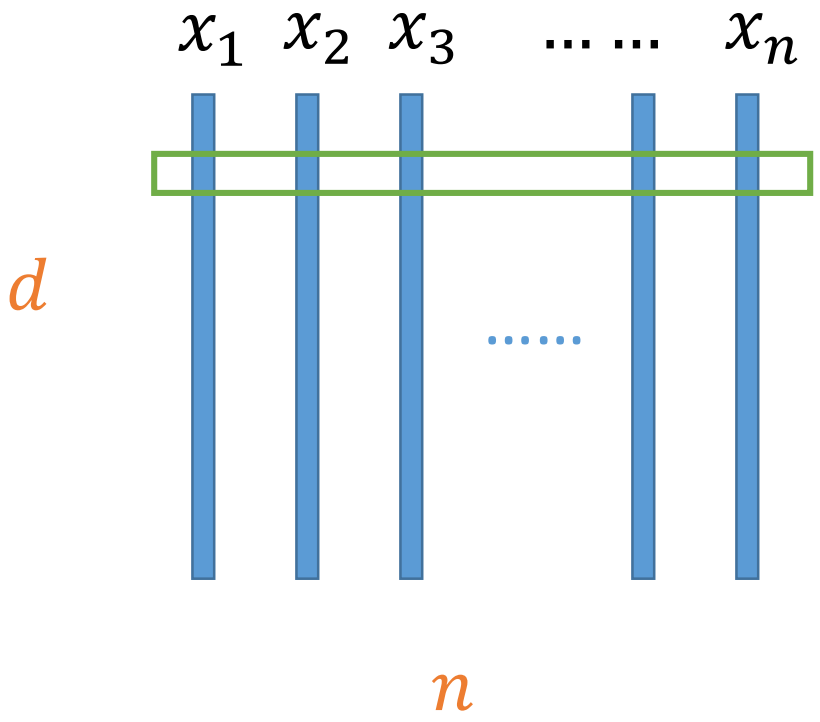
- (i, j)-th entry of WW^T is

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

- Our algorithm:
  - Step 1: estimate $\|W(i,:)\|_2$ & $b(i)$ for $i \in [d]$
  - Step 2: estimate θ_{ij} for $i \neq j \in [d]$

Our Algorithm: Overview

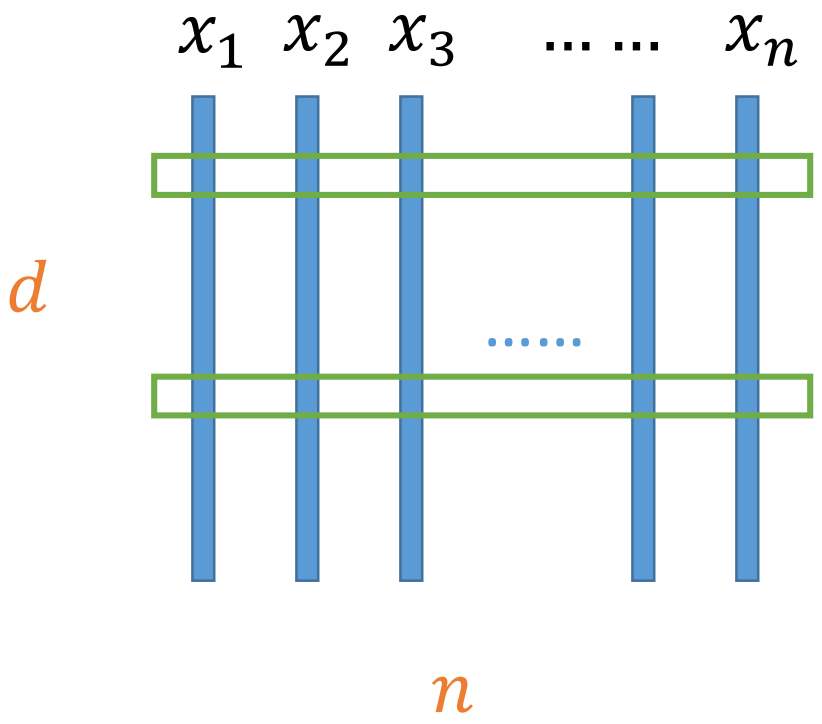
- Our problem: Given $x_1, \ldots, x_n \sim \mathrm{ReLU}(Wz + b)$, recover $WW^T,\ b$.



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

Our Algorithm: Overview

- Our problem: Given $x_1, \ldots, x_n \sim \mathrm{ReLU}(Wz + b)$, recover $WW^T,\ b$.

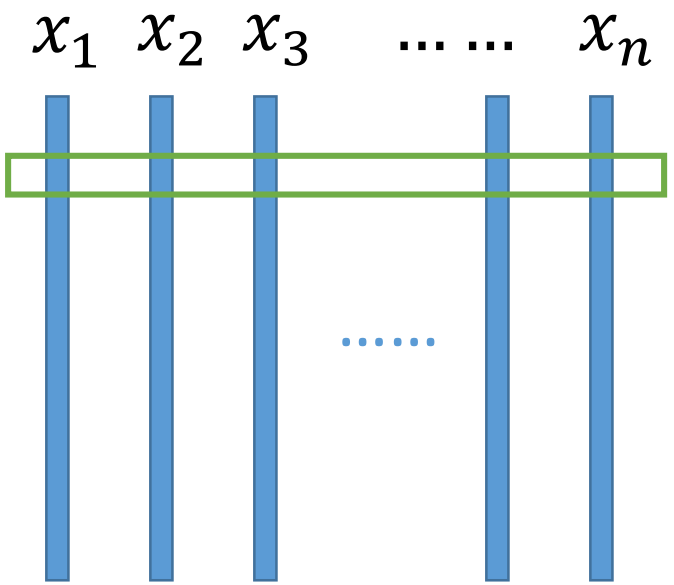


i-th coordinate

j-th coordinate

Step 2: for $i \neq j \in [d]$
Estimate θ_{ij}

Step 1: estimate $\|W(i,:)\|_2$ & $b(i)$



i-th coordinate

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

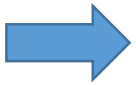
Step 1: estimate $\|W(i,:)\|_2$ & $b(i)$

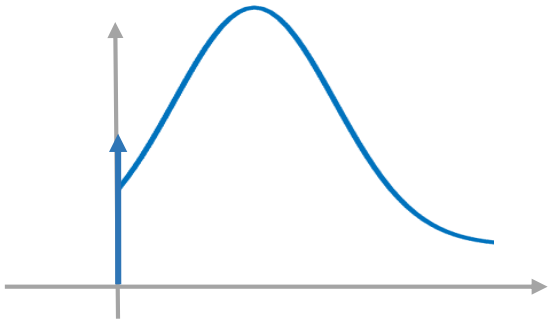
x_1 x_2 x_3 …… x_n

i-th coordinate

$$x(i) \sim \text{ReLU}(\underbrace{W(i,:)z + b(i)}_{N(b(i),\|W(i,:)\|_2^2)})$$

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





Pdf of $x(i)$

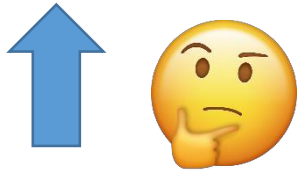
Step 1: estimate $\|W(i,:)\|_2$ & $b(i)$

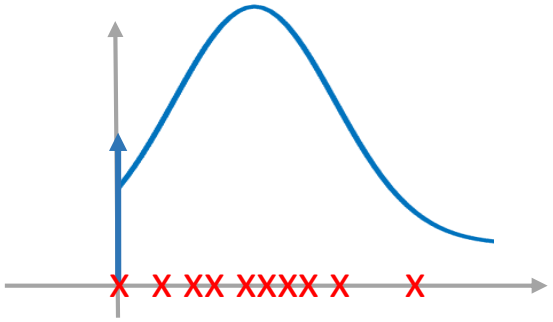
x_1 x_2 x_3 …… x_n

i-th coordinate

$$x(i) \sim \text{ReLU}(\underbrace{W(i,:)z + b(i)}_{N(b(i), \|W(i,:)\|_2^2)})$$

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





Samples $\sim x(i)$

Step 1: estimate $\|W(i,:)\|_2$ & $b(i)$

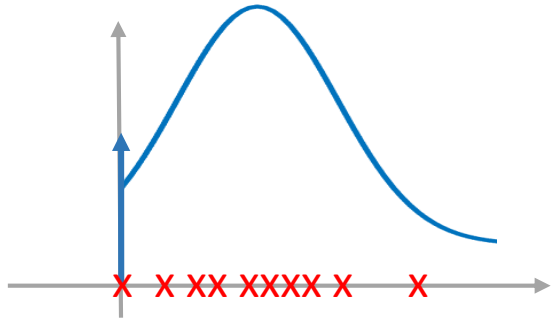
x_1 x_2 x_3 …… x_n

i-th coordinate

$$x(i) \sim \mathrm{ReLU}(\underbrace{W(i,:)z + b(i)}_{N(b(i), \|W(i,:)\|_2^2)})$$

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

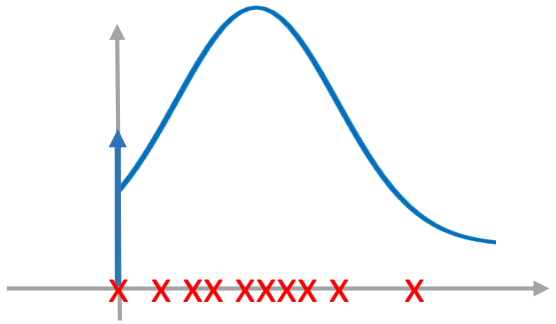
Truncated Gaussian



Samples $\sim x(i)$

Step 1: estimate $\|W(i,:)\|_2$ & $b(i)$

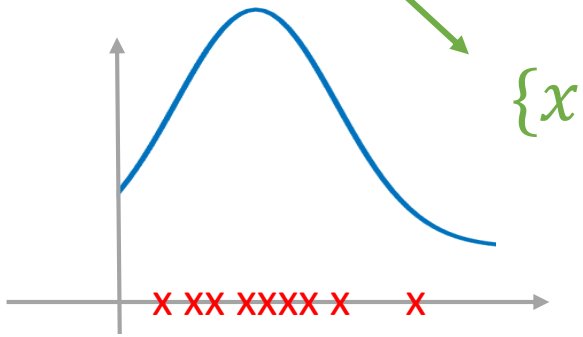
Samples $\sim x(i)$



Remove zeros

Samples $\sim$ truncated $N(b(i), \|W(i,:)\|_2^2)$

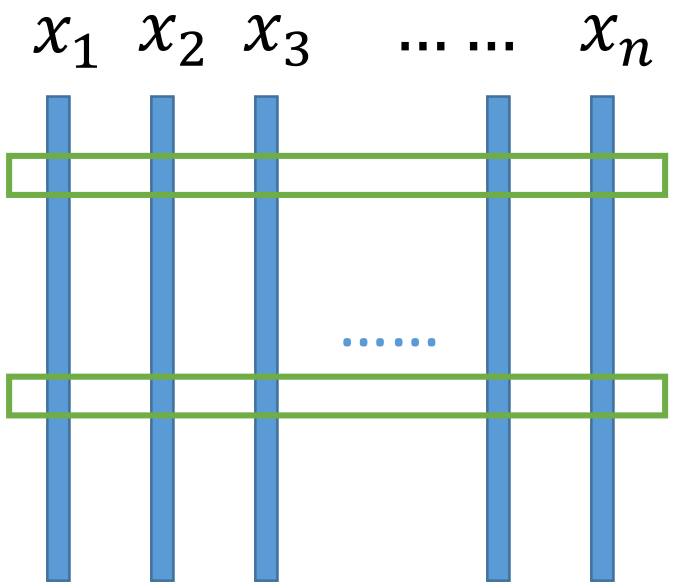
$\{x \in \mathbb{R}: x > 0\}$



Run algo [Daskalakis et al., 2018]

Estimate $b(i), \|W(i,:)\|_2$

Step 2: estimate θ_{ij}



i-th coordinate

j-th coordinate

Goal: Estimate θ_{ij}

Step 2: estimate θ_{ij}

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

Goal: Estimate
$\theta_{ij} :=$ angle b/t $W(i,:)$ & $W(j,:)$

Step 2: estimate θ_{ij}

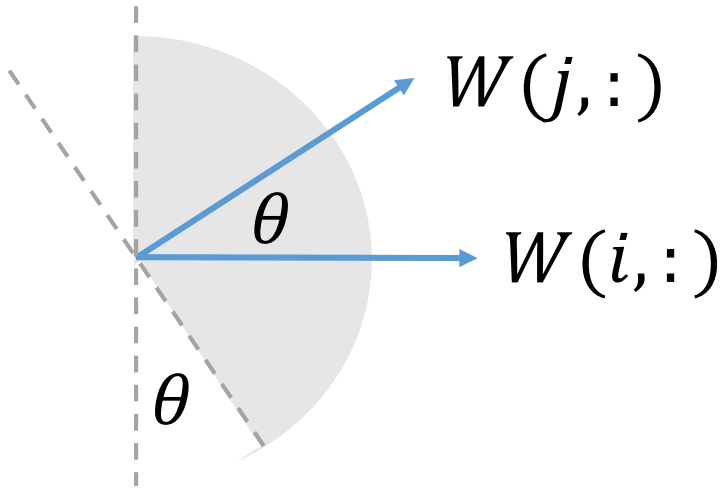
- $x(i) \sim \mathrm{ReLU}(\, W(i,:)z + b(i)\,)$
- $x(j) \sim \mathrm{ReLU}(\, W(j,:)z + b(j)\,)$

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

- Fact [Williamson and Shmoys'11]:

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

$$= \frac{\pi - \theta_{ij}}{2\pi}$$



Step 2: estimate θ_{ij}

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

Goal: Estimate
$\theta_{ij} :=$ angle b/t $W(i,:)$ & $W(j,:)$

- Suppose we know $b(i), b(j)$:

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

Assumption:
$b(i), b(j) \geq 0$

$$= \frac{\pi - \theta_{ij}}{2\pi}$$

Step 2: estimate θ_{ij}

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

Goal: Estimate
$\theta_{ij} :=$ angle b/t $W(i,:)$ & $W(j,:)$

- Suppose we know $b(i), b(j)$:

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

Assumption:
$b(i), b(j) \geq 0$

$$= \frac{\pi - \theta_{ij}}{2\pi}$$

If $b(i) < 0$, since $x(i) \geq 0$, $\mathbb{P}_x[\, x(i) > b(i)\,] = 1$

Step 2: estimate θ_{ij}

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

Goal: Estimate
$\theta_{ij} :=$ angle b/t $W(i,:)$ & $W(j,:)$

- Given $\hat{b}(i)$ and $\hat{b}(j)$ from Step 1:

$$\hat{b}(i) \approx b(i),\ \hat{b}(j) \approx b(j)$$

$$\mathbb{P}_x[x(i) > \hat{b}(i)\ \&\ x(j) > \hat{b}(j)] \approx \frac{\pi - \theta_{ij}}{2\pi}$$

Step 2: estimate θ_{ij}

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

Goal: Estimate
$\theta_{ij} :=$ angle b/t $W(i,:)$ & $W(j,:)$

- Given $\hat{b}(i)$ and $\hat{b}(j)$ from Step 1:

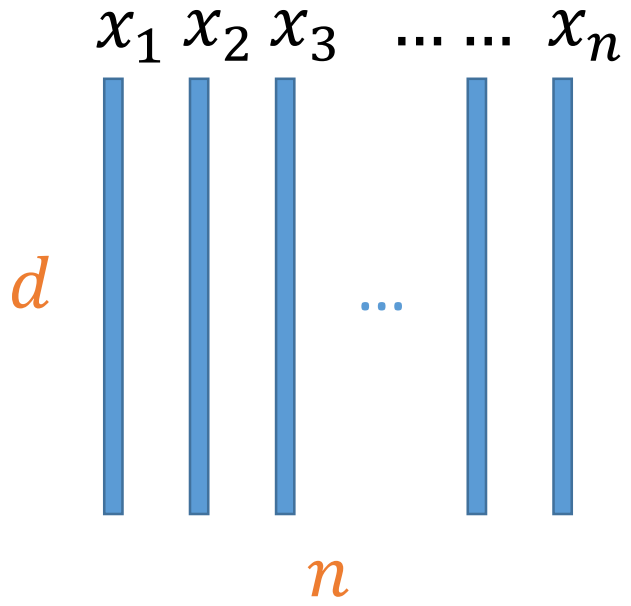
$$\hat{b}(i) \approx b(i),\ \hat{b}(j) \approx b(j)$$

$$\mathbb{P}_x[x(i) > \hat{b}(i)\ \&\ x(j) > \hat{b}(j)] \approx \frac{\pi - \theta_{ij}}{2\pi}$$

$$\hat{\theta}_{ij} = \pi - 2\pi \frac{\sum_{k=1}^{n} \mathbf{1}[x_k(i) > \hat{b}(i)\ \&\ x_k(j) > \hat{b}(j)]}{n}$$

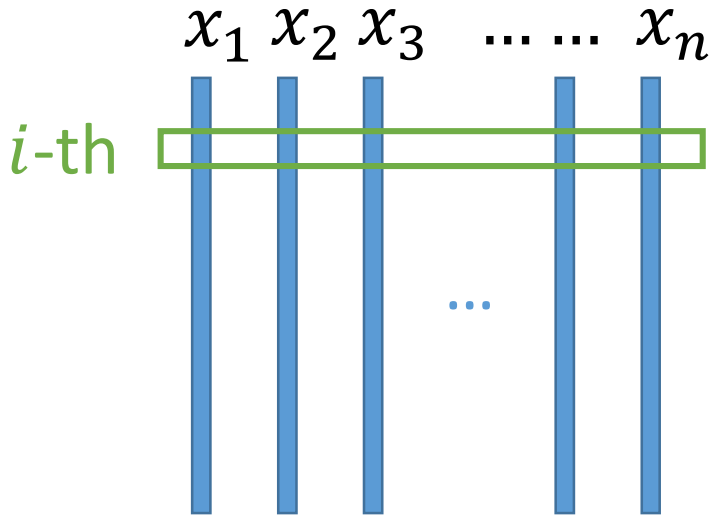
Our Algorithm: Overall

- **Input:** $x_1, \dots, x_n \sim \mathrm{ReLU}(Wz + b)$, b non-negative
- **Output:** $\widehat{\mathrm{WW}^{\mathrm{T}}} \in \mathbb{R}^{d \times d}$, $\hat{b} \in \mathbb{R}^d$
- **For** $i = 1, 2, \dots, d$
  - Remove zero samples from i-th coordinates
  - $\hat{b}(i), \|\widehat{W(i,:)}\|_2^2 \leftarrow$ Run [Daskalakis et al.'2018]
- **For** $i \neq j \in [d]$
  - $\hat{\theta}_{ij} \leftarrow \pi - 2\pi \frac{\sum_{k=1}^{n} \mathbf{1}[x_k(i) > \hat{b}(i) \,\&\, x_k(j) > \hat{b}(j)]}{n}$
  - $\widehat{WW^T}(i,j) = \|\widehat{W(i,:)}\|_2 \|\widehat{W(j,:)}\|_2 \cos\hat{\theta}_{ij}$



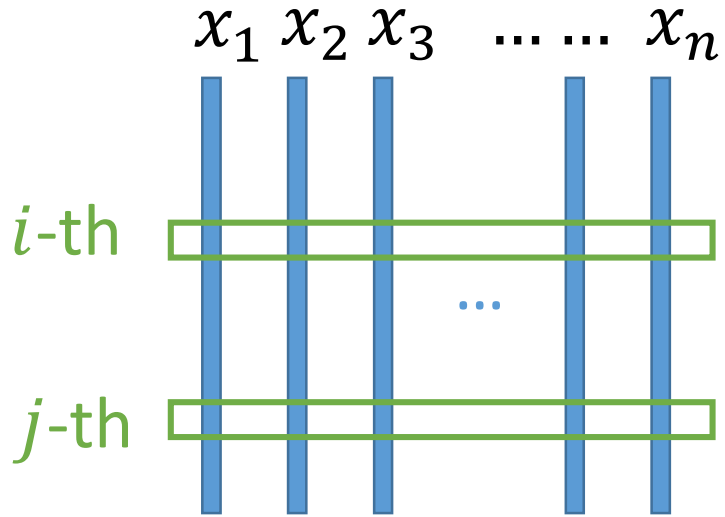
Our Algorithm: Overall

- **Input:** $x_1, \dots, x_n \sim \mathrm{ReLU}(Wz + b)$, b non-negative
- **Output:** $\widehat{WW^{\mathrm{T}}} \in \mathbb{R}^{d \times d}$, $\hat{b} \in \mathbb{R}^d$
- **For** $i = 1, 2, \dots, d$
  - Remove zero samples from i-th coordinates
  - $\hat{b}(i), \|\widehat{W(i,:)}\|_2^2 \leftarrow$ Run [Daskalakis et al.'2018]
- **For** $i \neq j \in [d]$
  - $\hat{\theta}_{ij} \leftarrow \pi - 2\pi \frac{\sum_{k=1}^{n} \mathbf{1}[x_k(i) > \hat{b}(i) \,\&\, x_k(j) > \hat{b}(j)]}{n}$
  - $\widehat{WW^T}(i,j) = \|\widehat{W(i,:)}\|_2 \|\widehat{W(j,:)}\|_2 \cos\hat{\theta}_{ij}$



Our Algorithm: Overall

- **Input:** $x_1, \dots, x_n \sim \mathrm{ReLU}(Wz + b)$, b non-negative
- **Output:** $\widehat{\mathrm{WW}^{\mathrm{T}}} \in \mathbb{R}^{d \times d}$, $\hat{b} \in \mathbb{R}^d$
- **For** $i = 1, 2, \dots, d$
  - Remove zero samples from i-th coordinates
  - $\hat{b}(i), \|\widehat{W(i,:)}\|_2^2 \leftarrow$ Run [Daskalakis et al.'2018]
- **For** $i \neq j \in [d]$
  - $\hat{\theta}_{ij} \leftarrow \pi - 2\pi \frac{\sum_{k=1}^{n} \mathbf{1}[x_k(i) > \hat{b}(i) \,\&\, x_k(j) > \hat{b}(j)]}{n}$
  - $\widehat{WW^T}(i,j) = \|\widehat{W(i,:)}\|_2 \|\widehat{W(j,:)}\|_2 \cos\hat{\theta}_{ij}$



Analysis

- **Theorem**: Assuming that $b^* \in \mathbb{R}^d$ is non-negative, then our algorithm takes $\tilde{O}(\frac{1}{\epsilon^2}\ln(\frac{d}{\delta}))$ samples $\sim \mathrm{ReLU}(W^* z + b^*)$ and outputs

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

  with probability at least $1 - \delta$.

Analysis: non-degenerate case

- **Corollary:** Let $\kappa :=$ condition number of $W^*{W^*}^T$. Our algorithm takes $\tilde{O}\left(\frac{\kappa^2 d^2}{\epsilon^2}\ln\left(\frac{d}{\delta}\right)\right)$ samples $\sim \mathrm{ReLU}(W^* z + b^*)$ and outputs

$$\mathrm{TV}\left(\mathrm{ReLU}\left(\widehat{W} z + \hat{b}\right), \mathrm{ReLU}(W^* z + b^*)\right) \leq \epsilon$$

with probability at least $1 - \delta$.

Sample Complexity Lower Bounds

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

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

Our algorithm is optimal

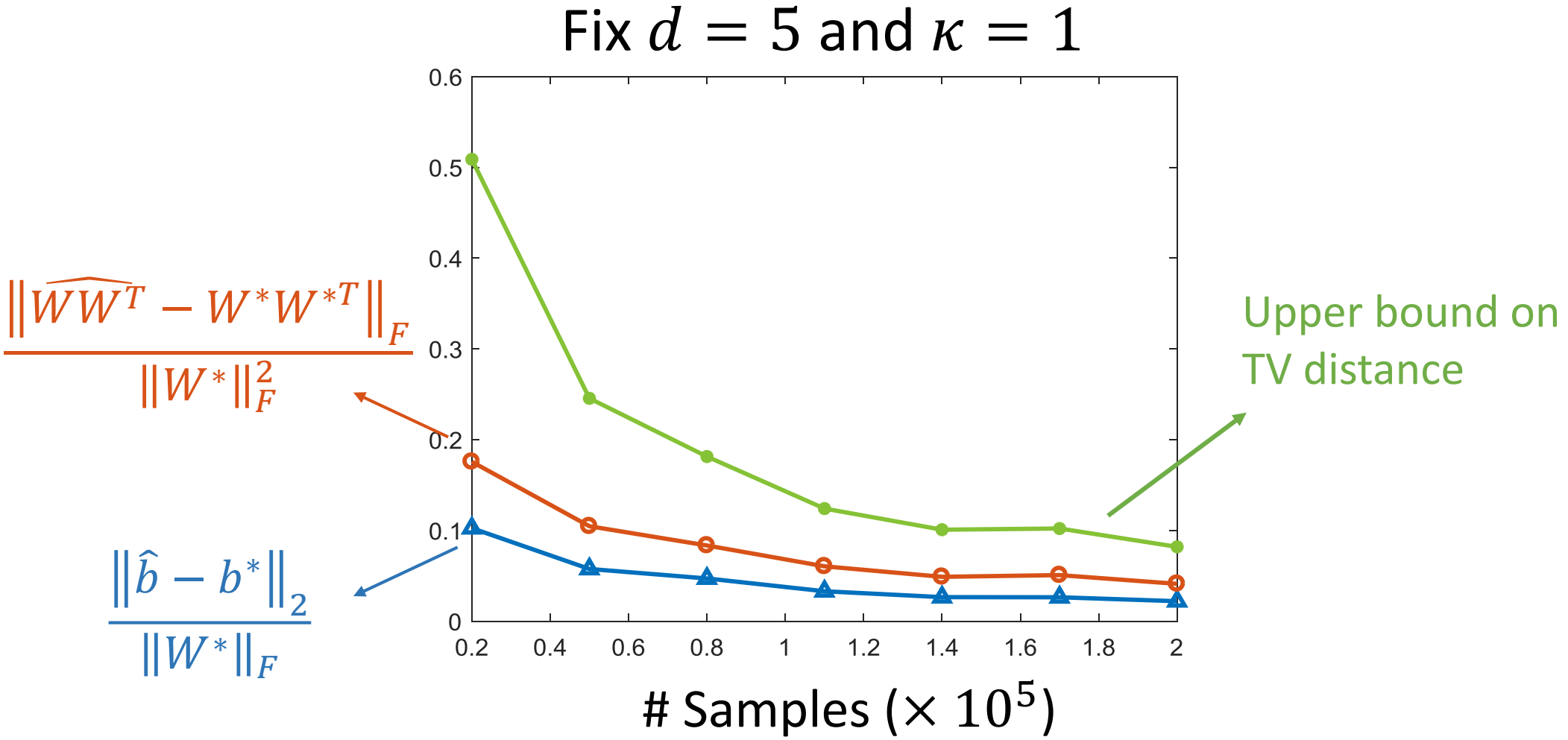
- Total Variation Distance: $\mathrm{TV}(D, \mathrm{ReLU}(W^*z + b^*)) \leq \epsilon$

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:

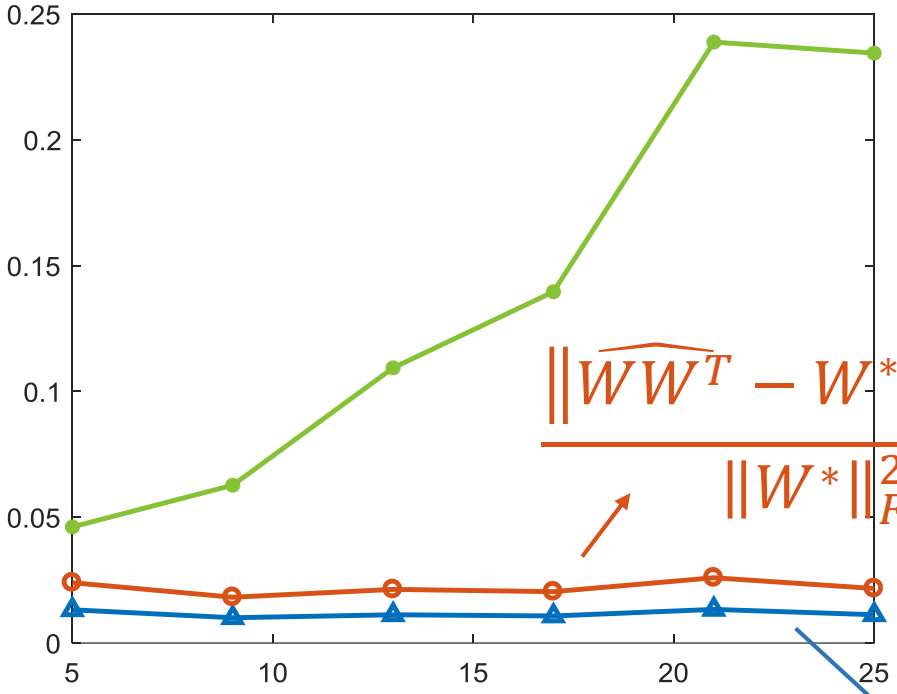
1) Param est → TV distance
2) Lower bound is not tight

Experimental Results



Experimental Results

Fix # samples = 2e5

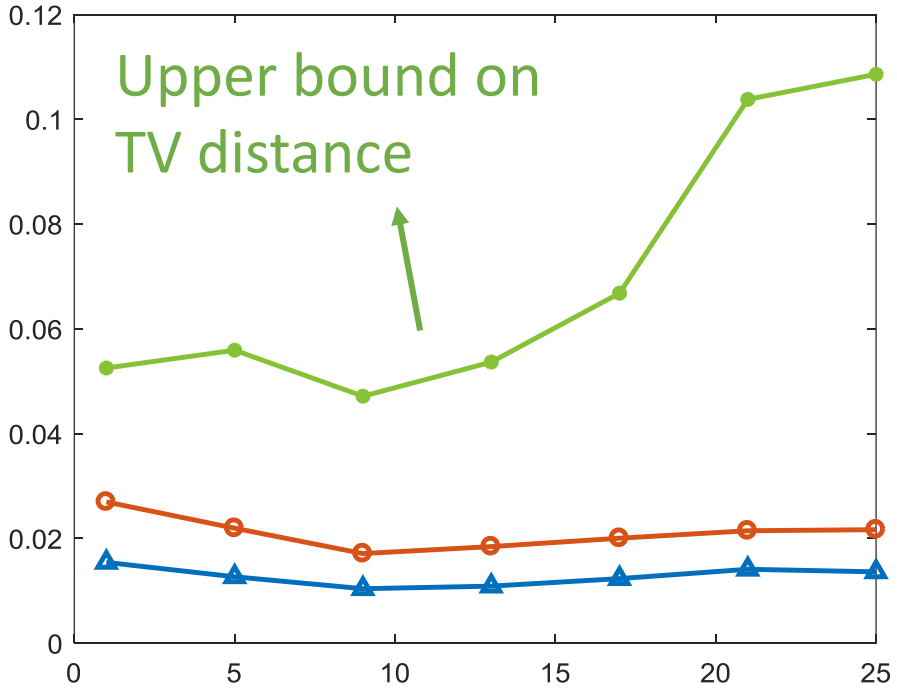


$$\frac{\left\|\widehat{WW}^T - W^*W^{*T}\right\|_F}{\|W^*\|_F^2}$$

$$\frac{\left\|\hat{b} - b^*\right\|_2}{\|W^*\|_F}$$

Dimension (d)

Fix # samples = 2e5



Condition number (κ)

What if b has negative values?

$$\exists \eta \geq 0,\ b(i) \geq -\eta \|W(i,:)\|_2, \forall i \in [d]$$

Given $\tilde{O}(\frac{1}{\epsilon^2}\ln(\frac{d}{\delta}))$ samples, our algorithm outputs:

$$\left\|\widehat{WW}^T - W^*{W^*}^T\right\|_F \leq \max(\epsilon, \eta)\|W^*\|_F^2$$

$$\left\|\hat{b} - b^*\right\|_2 \leq \max(\epsilon, \eta)\|W^*\|_F$$

Related Work

- Learning ReLU neural networks [GKLW'19, GKM'18, LY'17,…….]

Supervised v.s. Unsupervised

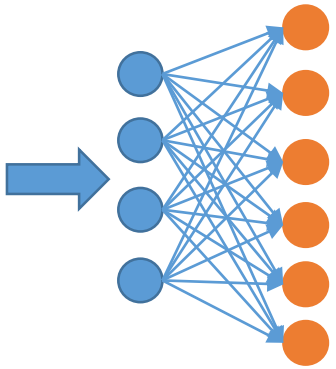
- Provably learning a generative model
  - [Nguyen, Wong, and Hedge'18]

Linear generative model

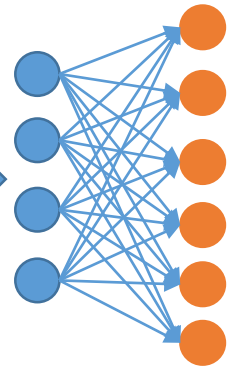
  - [Mazumdar and Rawat'19]

Random b, recover W's column space

Summary

$z \sim N(0, I_{k\times k})$  $x \sim \mathrm{ReLU}(Wz + b)$  Recover WW^T, b

Summary

$z \sim N(0, I_{k\times k})$  $x \sim \mathrm{ReLU}(Wz + b)$  Recover $WW^T,\ b$

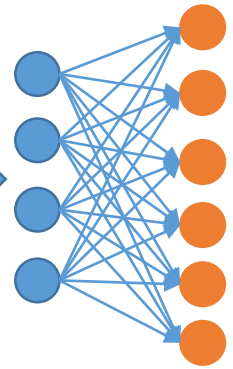
Our algorithm:
Step 1: est $\|W(i,:)\|_2,\ b(i)$
Step 2: est θ_{ij}

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

Assumption:
b non-negative
Otherwise:
$\Omega(\exp(\|b\|_\infty^2))$

Experimental results are consistent with analysis.

Summary

$z \sim N(0, I_{k\times k})$  $x \sim \mathrm{ReLU}(Wz + b)$  Recover WW^T, b

- Open problems:
  - Two-layer Generative model
  - Negative bias vector
  - Agnostic setting
  - GAN/VAE……

When we use neural network to model distributions, what is the structure of those distributions?