How to Hoodwink a Halfspace

A survey done in partial fulfillment of the requirements of the Comprehensive Examination for doctoral candidates

Purushottam Kar
Y8111062

January 12, 2010

Halfspaces

Definition (Halfspaces)

A halfspace in a d dimensional Euclidean space is a dichotomy characterized by a weight vector $w \in \mathbb{R}^d$ and a threshold $\theta \in \mathbb{R}$. More specifically $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)\ \forall x \in \mathbb{R}^d$

Halfspaces

Definition (Halfspaces)

A halfspace in a d dimensional Euclidean space is a dichotomy characterized by a weight vector $w \in \mathbb{R}^d$ and a threshold $\theta \in \mathbb{R}$. More specifically $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)\ \forall x \in \mathbb{R}^d$

Studied extensively in learning theory, geometry, game theory, complexity theory ...

Halfspaces

Definition (Halfspaces)

A halfspace in a d dimensional Euclidean space is a dichotomy characterized by a weight vector $w \in \mathbb{R}^d$ and a threshold $\theta \in \mathbb{R}$. More specifically $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)\ \forall x \in \mathbb{R}^d$

Studied extensively in learning theory, geometry, game theory, complexity theory ...

Halfspaces

Definition (Halfspaces)

A halfspace in a d dimensional Euclidean space is a dichotomy characterized by a weight vector $w \in \mathbb{R}^d$ and a threshold $\theta \in \mathbb{R}$. More specifically $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)\ \forall x \in \mathbb{R}^d$

Studied extensively in learning theory, geometry, game theory, complexity theory ...

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- Halfspaces are weak

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- ... very weak

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- ... very weak
- Cannot separate strings based on parity [MP69]

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- ... very weak
- ... let alone regular languages

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- ... very weak
- ... let alone regular languages
- Cannot separate most interesting language classes due to high (read infinite) VC dimension

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- ... very weak
- ... let alone regular languages
- Cannot separate most interesting language classes due to high (read infinite) VC dimension
- Have been shown to be able to adapt to *piecewise testable languages* using large margin methods methods (no PAC guarantees though) [CKM07]

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- ... very weak
- ... let alone regular languages
- Cannot separate most interesting language classes due to high (read infinite) VC dimension
- Have been shown to be able to adapt to *piecewise testable languages* using large margin methods methods (no PAC guarantees though) [CKM07]
- Whatever they can represent, can be learnt pretty fast

Pre [DGJ+09] ...

The Learning Theory part ...

- ... very weak
- ... let alone regular languages
- Cannot separate most interesting language classes due to high (read infinite) VC dimension
- Have been shown to be able to adapt to *piecewise testable languages* using large margin methods methods (no PAC guarantees though) [CKM07]
- Whatever they can represent, can be learnt in “soft” quadratic time

Pre [DGJ+09] ...

The Learning Theory part ...

- ... very weak
- ... let alone regular languages
- Cannot separate most interesting language classes due to high (read infinite) VC dimension
- Have been shown to be able to adapt to *piecewise testable languages* using large margin methods methods (no PAC guarantees though) [CKM07]
- Whatever they can represent, can be learnt in “soft” quadratic time given the presence of a teacher

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- ... very weak
- ... let alone regular languages
- Cannot separate most interesting language classes due to high (read infinite) VC dimension
- Have been shown to be able to adapt to *piecewise testable languages* using large margin methods methods (no PAC guarantees though) [CKM07]
- Whatever they can represent, can be learnt in "soft" quadratic time given the presence of a teacher
- There is a quadratic lower bound on the learning time [MT94]

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- Thresholded polynomials are stronger

Pre [DGJ+09] ...

The Learning Theory part ...

- Thresholded polynomials are stronger
- can be used to represent DNFs of exponentially larger sizes

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- Thresholded polynomials are stronger
- can be used to represent DNFs of exponentially larger sizes
- [KS04] show that s-term DNFs can be computed by polynomial threshold functions of degree $\mathcal{O}\left(n^{1/3} \log s\right)$

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- Thresholded polynomials are stronger
- can be used to represent DNFs of exponentially larger sizes
- [KS04] show that s-term DNFs can be computed by polynomial threshold functions of degree $\mathcal{O}\left(n^{1/3}\log s\right)$
- Matches a lower bound of $\Omega\left(n^{1/3}\right)$ by [MP69]

Pre [DGJ$^+$09] ...

The Learning Theory part ...

- Thresholded polynomials are stronger
- can be used to represent DNFs of exponentially larger sizes
- [KS04] show that s-term DNFs can be computed by polynomial threshold functions of degree $\mathcal{O}\left(n^{1/3}\log s\right)$
- Matches a lower bound of $\Omega\left(n^{1/3}\right)$ by [MP69]
- The construction gives a $2^{\mathcal{O}\left(n^{1/3}\log s\log n\right)}$-time algorithm to learn DNFs by extending halfspace learning algorithms to ones that learn polynomial threshold functions over boolean valued attributes

Pre [DGJ$^+$09] ...

The Complexity Theory part ...

- Halfspaces are resilient

Pre [DGJ$^+$09] ...

The Complexity Theory part ...

- ... very resilient

Pre [DGJ$^+$09] ...

The Complexity Theory part ...

- ... very resilient
- Cannot be simulated by low-degree polynomials or AC^0

Pre [DGJ$^+$09] ...

The Complexity Theory part ...

- ... very resilient
- Cannot be simulated by low-degree polynomials or AC^0
- A separation like $\mathrm{NP} \not\subset \mathrm{HALFSPACE}^2$ still eludes us

Pre [DGJ$^+$09] ...

The Complexity Theory part ...

- ... very resilient
- Cannot be simulated by low-degree polynomials or AC^0
- A separation like $\mathrm{NP} \not\subset \mathrm{HALFSPACE}^2$ still eludes us
- Circuits composed of halfspaces can be simulated by circuits of majority gates of almost same depth

Pre [DGJ$^+$09] ...

The Complexity Theory part ...

- ... very resilient
- Cannot be simulated by low-degree polynomials or AC^0
- A separation like $\mathrm{NP} \not\subset \mathrm{HALFSPACE}^2$ still eludes us
- Circuits composed of halfspaces can be simulated by circuits of majority gates of almost same depth
- Representational Complexity : Integer weights of size $\frac{(n+1)\log(n+1)}{2} - n$ bits suffice and $\frac{n \log n}{2} - n$ are necessary [Hås94]

Pre [DGJ$^+$09] ...

The Complexity Theory part ...

- ... very resilient
- Cannot be simulated by low-degree polynomials or AC^0
- A separation like $\mathrm{NP} \not\subset \mathrm{HALFSPACE}^2$ still eludes us
- Circuits composed of halfspaces can be simulated by circuits of majority gates of almost same depth
- Representational Complexity : Integer weights of size $\frac{(n+1)\log(n+1)}{2} - n$ bits suffice and $\frac{n \log n}{2} - n$ are necessary [Hås94]
- If approximate representations are all we want then $\sqrt{n}2^{\tilde{O}(1/\epsilon^2)}$ bits suffice to get a halfplane that begs to differ only on an ϵ fraction of the inputs [Ser07]

The Art and Mathematics of Deception

Definition (Fooling a Function)

A distribution $\mathcal{D}$ on strings over $\{-1, 1\}$ of length n is said to ϵ-fool a boolean function $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ if $|E_{x \leftarrow \mathcal{D}}[f(x)] - E_{x \leftarrow \mathcal{U}}[f(x)]| \leq \epsilon$

The Art and Mathematics of Deception

Definition (Fooling a Function)

A distribution $\mathcal{D}$ on strings over $\{-1, 1\}$ of length n is said to ϵ-fool a boolean function $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ if $|E_{x \leftarrow \mathcal{D}}[f(x)] - E_{x \leftarrow \mathcal{U}}[f(x)]| \leq \epsilon$

The uniform distribution $\mathcal{U}$ fools every function - but it requires too many random bits to implement

The Art and Mathematics of Deception

Definition (Fooling a Function)

A distribution $\mathcal{D}$ on strings over $\{-1, 1\}$ of length n is said to ϵ-fool a boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ if $|E_{x \leftarrow \mathcal{D}}[f(x)] - E_{x \leftarrow \mathcal{U}}[f(x)]| \leq \epsilon$

The uniform distribution $\mathcal{U}$ fools every function - but it requires too many random bits to implement

Can we fool certain functions using distributions that we can "create" ourselves given smaller amount of randomness ?

The Art and Mathematics of Deception

Definition (Fooling a Function)

A distribution $\mathcal{D}$ on strings over $\{-1,1\}$ of length n is said to ϵ-fool a boolean function $f : \{-1,1\}^n \to \{-1,1\}$ if $|E_{x \leftarrow \mathcal{D}}[f(x)] - E_{x \leftarrow \mathcal{U}}[f(x)]| \leq \epsilon$

But why would one want to indulge in such a trivial pursuit ?

The Art and Mathematics of Deception

Definition (Fooling a Function)

A distribution $\mathcal{D}$ on strings over $\{-1,1\}$ of length n is said to ϵ-fool a boolean function $f : \{-1,1\}^n \to \{-1,1\}$ if $|E_{x \leftarrow \mathcal{D}}[f(x)] - E_{x \leftarrow \mathcal{U}}[f(x)]| \leq \epsilon$

Can we fool certain functions using distributions that we can "create" ourselves given smaller amount of randomness ?

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Example taken from `http://www.nada.kth.se/~johanh/verktyg/lecture3.pdf`

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Example taken from `http://www.nada.kth.se/~johanh/verktyg/lecture3.pdf`

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Example taken from `http://www.nada.kth.se/~johanh/verktyg/lecture3.pdf`

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Example taken from `http://www.nada.kth.se/~johanh/verktyg/lecture3.pdf`

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Example taken from `http://www.nada.kth.se/~johanh/verktyg/lecture3.pdf`

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Construction

Optimal constructions of such distributions exist

Example taken from `http://www.nada.kth.se/~johanh/verktyg/lecture3.pdf`

Less than random distributions

Definition (k-wise Independence)

A distribution $\mathcal{D}$ on $\{-1,+1\}^n$ is said to be k-wise independent if the projection of $\mathcal{D}$ on any k indices is uniformly distributed over $\{-1,+1\}^k$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Construction

Optimal constructions of such distributions exist

Randomness Requirement

Given a sequence of $k \log n$ random bits, one can generate a sequence of n random bits that is k-wise independent.

Example taken from `http://www.nada.kth.se/~johanh/verktyg/lecture3.pdf`

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- We know how to fool low-degree polynomials, constant depth boolean circuits , ...

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- We know how to fool low-degree polynomials, constant depth boolean circuits , ...
- Some of these constructions imply that halfspaces with small weights can be fooled

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- We know how to fool low-degree polynomials, constant depth boolean circuits , ...
- Some of these constructions imply that halfspaces with small weights can be fooled
- The question of fooling general halfspaces ...

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- We know how to fool low-degree polynomials, constant depth boolean circuits , ...
- Some of these constructions imply that halfspaces with small weights can be fooled
- The question of fooling general halfspaces ... [DGJ$^+$09]

Pre [DGJ+09] ...

The Complexity Theory part ... contd

- We know how to fool low-degree polynomials, constant depth boolean circuits , ...
- Some of these constructions imply that halfspaces with small weights can be fooled
- The question of fooling general halfspaces ... [DGJ+09]
- The question investigated by [DGJ+09] is not directly related to construction of pseudo-random generators for halfspaces

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- We know how to fool low-degree polynomials, constant depth boolean circuits , ...
- Some of these constructions imply that halfspaces with small weights can be fooled
- The question of fooling general halfspaces ... [DGJ$^+$09]
- The question being asked is that of a property fooling a class of functions rather than a distribution doing so

A Key Result

Theorem ([Baz07])

A boolean function $f : \{-1,1\}^n \to \{-1,1\}$ *can be* ϵ*-fooled by the class of k-wise independent distributions iff there exist multivariate polynomials* $u : \{-1,1\}^n \to \{-1,1\}$, $l : \{-1,1\}^n \to \{-1,1\}$, *such that*

- $\deg(u), \deg(l) \leq k$
- $u(x) \geq f(x) \geq l(x)\ \forall x \in \{-1,1\}^n$
- $\mathbf{E}_{x \leftarrow \mathcal{U}}[u(x) - f(x)], \mathbf{E}_{x \leftarrow \mathcal{U}}[f(x) - l(x)] \leq \epsilon$

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- Has been used very productively to fool

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- Has been used very productively to fool
    - DNFs [Baz07] [Raz08]

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- Has been used very productively to fool
  - DNFs [Baz07] [Raz08]
  - AC^0 functions [Bra09]

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- Has been used very productively to fool
  - DNFs [Baz07] [Raz08]
  - AC^0 functions [Bra09]
  - halfspaces [DGJ$^+$09][GOWZ10][KNW10]

Pre [DGJ$^+$09] ...

The Complexity Theory part ... contd

- Has been used very productively to fool
  - DNFs [Baz07] [Raz08]
  - AC^0 functions [Bra09]
  - halfspaces [DGJ$^+$09][GOWZ10][KNW10]
- Note : Servedio's construction in [Ser07] gives us PRGs for halfspaces if $\epsilon = \Omega(1/\sqrt{\log n})$. The [DGJ$^+$09] construction itself stops working if $\epsilon = \mathcal{O}(1/\sqrt{n})$

Now [DGJ+09]

Plan of attack

Goal : Find two low-degree polynomials that sandwich our halfspace function while closely approximating it

Plan of attack

Plan of attack :

Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function

Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it

Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it

Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it
- Wait ... what happened to the halfspace ??

Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it
- Probably need to restate some of the goals

Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it under the Gaussian distribution
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it

Plan of attack

Plan of attack :

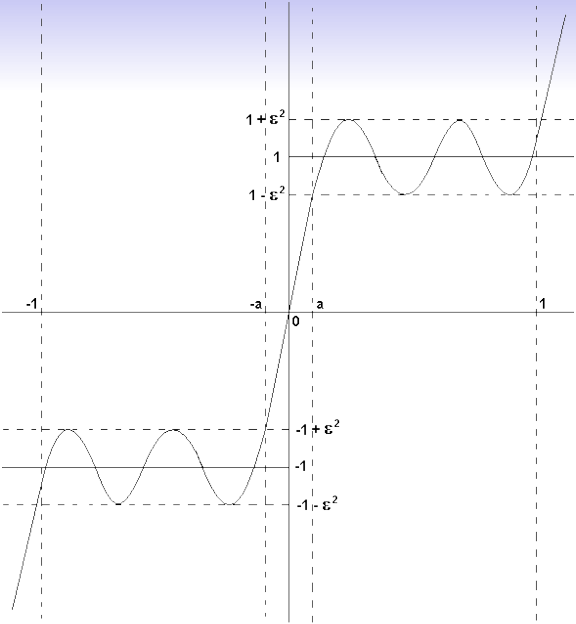
- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it under the Gaussian distribution
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it under the Gaussian distribution

Plan of attack

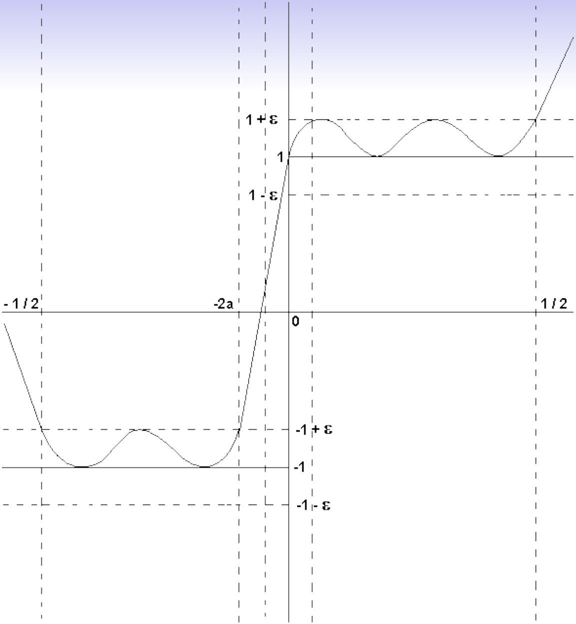
Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it under the Gaussian distribution
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it under the Gaussian distribution
- Use the fact that values taken by homogeneous 'regular' linear polynomials are distributed normally

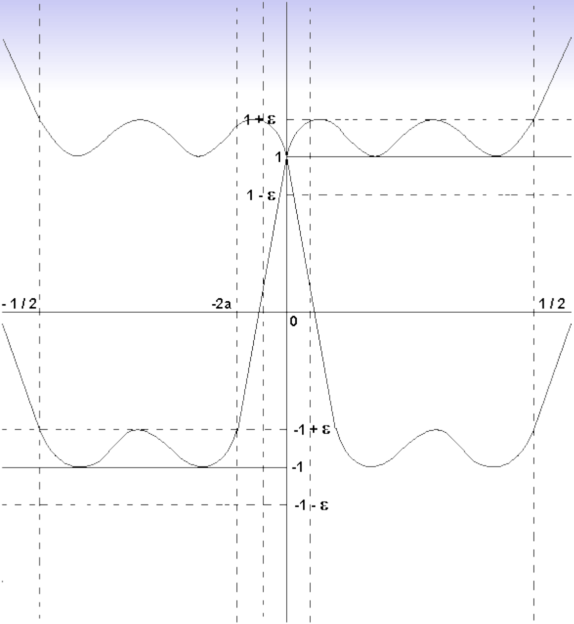
Step 1



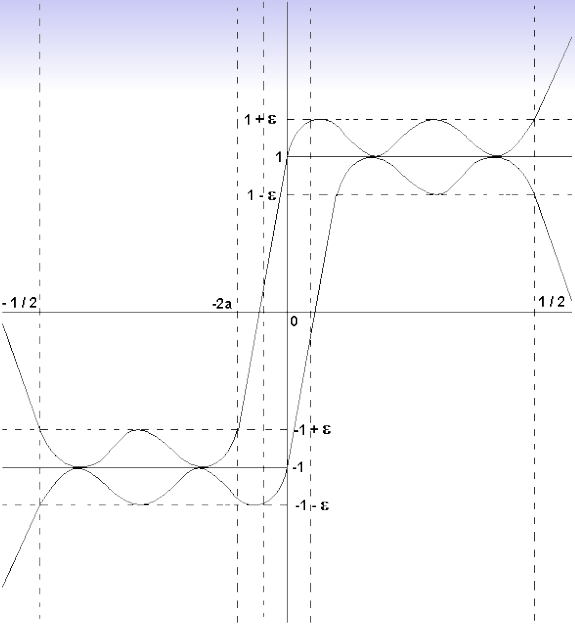
Step 2



Step 3



Step 3



Approximating Real-valued Functions - I

Theorem (Jackson)

Any bounded continuous function $f : [-1,1] \to \mathbb{R}$ admits a $6\omega_f\left(\frac{1}{\ell}\right)$-pointwise approximation by a degree-ℓ polynomial in the domain $[-1,1]$.

Approximating Real-valued Functions - I

Theorem (Jackson)

Any bounded continuous function $f : [-1, 1] \rightarrow \mathbb{R}$ *admits a* $6\omega_f\left(\frac{1}{\ell}\right)$*-pointwise approximation by a degree-*ℓ *polynomial in the domain* $[-1, 1]$.

- Use Jackson's theorem to $\mathcal{O}(1)$-approximate sgn by a degree $\mathcal{O}(1/a)$ polynomial ($a = \epsilon^2/log(1/\epsilon)$)

Approximating Real-valued Functions - I

Theorem (Jackson)

Any bounded continuous function $f : [-1, 1] \to \mathbb{R}$ *admits a* $6\omega_f\left(\frac{1}{\ell}\right)$*-pointwise approximation by a degree-*ℓ *polynomial in the domain* $[-1, 1]$.

- Use Jackson's theorem to $\mathcal{O}(1)$-approximate sgn by a degree $\mathcal{O}(1/a)$ polynomial ($a = \epsilon^2/log(1/\epsilon)$)
- Use an amplifying polynomial of degree $\mathcal{O}(log(1/\epsilon))$ to reduce the error to ϵ^2

Approximating Real-valued Functions - I

Theorem (Jackson)

Any bounded continuous function $f : [-1,1] \to \mathbb{R}$ *admits a* $6\omega_f\left(\frac{1}{\ell}\right)$*-pointwise approximation by a degree-*ℓ *polynomial in the domain* $[-1,1]$.

- Use Jackson's theorem to $\mathcal{O}(1)$-approximate sgn by a degree $\mathcal{O}(1/a)$ polynomial ($a = \epsilon^2/log(1/\epsilon)$)
- Use an amplifying polynomial of degree $\mathcal{O}(log(1/\epsilon))$ to reduce the error to ϵ^2

Lemma

There is a polynomial $p_1(x)$ *of degree* $2m = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ *which gives a pointwise* ϵ^2*-approximation to the* sgn *function in the range* $[-1,-a] \cup [a,1]$.

Approximating Real-valued Functions - II

Theorem (Chebyshev)

For any bounded continuous function $f : [k, l] \to \mathbb{R}$ *and any non-zero continuous function* $f : [k, l] \to \mathbb{R}$*, for every* m*, there is a unique degree-*m *polynomial* $r(z)$ *that minimizes the maximum pointwise error* $\max_{x \in [k,l]} |f(x) - s(x)r(x)|$ *and is characterized by the fact that the function* $s(x)r(x)$ *achieves this maximum error* $m + 2$ *times in the interval* $[k, l]$ *with alternating signs.*

Approximating Real-valued Functions - II

Theorem (Chebyshev)

For any bounded continuous function $f : [k, l] \to \mathbb{R}$ and any non-zero continuous function $f : [k, l] \to \mathbb{R}$, for every m, there is a unique degree-m polynomial $r(z)$ that minimizes the maximum pointwise error $\max_{x \in [k,l]} |f(x) - s(x)r(x)|$ and is characterized by the fact that the function $s(x)r(x)$ achieves this maximum error $m + 2$ times in the interval $[k, l]$ with alternating signs.

- Use Chebyshev's theorem to get the best degree m approximation $r(x)$ which minimizes $\max_{x \in [a^2,1]} |1 - \sqrt{x}r(x)|$

Approximating Real-valued Functions - II

Theorem (Chebyshev)

For any bounded continuous function $f : [k, l] \to \mathbb{R}$ *and any non-zero continuous function* $f : [k, l] \to \mathbb{R}$*, for every* m*, there is a unique degree-*m *polynomial* $r(z)$ *that minimizes the maximum pointwise error* $\max_{x \in [k,l]} |f(x) - s(x)r(x)|$ *and is characterized by the fact that the function* $s(x)r(x)$ *achieves this maximum error* $m + 2$ *times in the interval* $[k, l]$ *with alternating signs.*

- Use Chebyshev's theorem to get the best degree m approximation $r(x)$ which minimizes $\max_{x \in [a^2,1]} |1 - \sqrt{x} r(x)|$
- Let $p(x) = x \cdot r(x^2)$.

Completing Step 1

- Write $p_1(x)$ in the form $x \cdot r_1(x^2)$

Completing Step 1

- Write $p_1(x)$ in the form $x \cdot r_1(x^2)$
- Use it to bound the error of $p(x)$ in the interval $[-1, a] \cup [a, 1]$ by ϵ^2

Completing Step 1

- Write $p_1(x)$ in the form $x \cdot r_1(x^2)$
- Use it to bound the error of $p(x)$ in the interval $[-1, a] \cup [a, 1]$ by ϵ^2
- and get some more properties ...

Completing Step 1

- Write $p_1(x)$ in the form $x \cdot r_1(x^2)$
- Use it to bound the error of $p(x)$ in the interval $[-1, a] \cup [a, 1]$ by ϵ^2
- and get some more properties ...

Lemma

There is a polynomial $p(x)$ of degree $2m + 1 = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

Completing Step 1

- Write $p_1(x)$ in the form $x \cdot r_1(x^2)$
- Use it to bound the error of $p(x)$ in the interval $[-1, a] \cup [a, 1]$ by ϵ^2
- and get some more properties ...

Lemma

There is a polynomial $p(x)$ of degree $2m + 1 = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

- $p(x) \in \mathrm{sgn}(x) \pm \epsilon^2$ *for all* $|x| \in [a, 1]$

Completing Step 1

- Write $p_1(x)$ in the form $x \cdot r_1(x^2)$
- Use it to bound the error of $p(x)$ in the interval $[-1, a] \cup [a, 1]$ by ϵ^2
- and get some more properties ...

Lemma

There is a polynomial $p(x)$ of degree $2m + 1 = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

- $p(x) \in \mathrm{sgn}(x) \pm \epsilon^2$ *for all* $|x| \in [a, 1]$
- $p(x) \in \pm(1 + \epsilon^2)$ *for all* $|x| \in [0, a]$

Completing Step 1

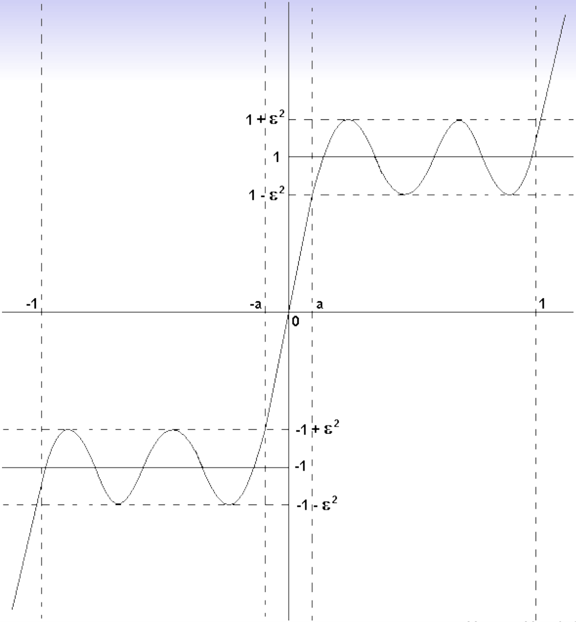
- Write $p_1(x)$ in the form $x \cdot r_1(x^2)$
- Use it to bound the error of $p(x)$ in the interval $[-1, a] \cup [a, 1]$ by ϵ^2
- and get some more properties ...

Lemma

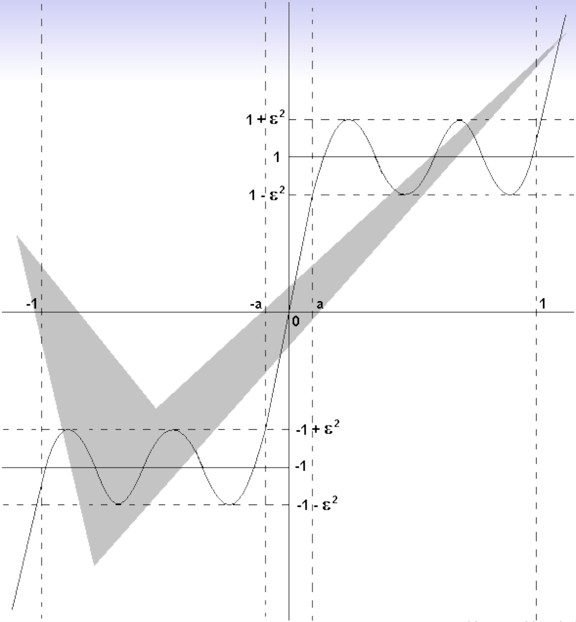
There is a polynomial $p(x)$ of degree $2m + 1 = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

- $p(x) \in \mathrm{sgn}(x) \pm \epsilon^2$ *for all* $|x| \in [a, 1]$
- $p(x) \in \pm(1 + \epsilon^2)$ *for all* $|x| \in [0, a]$
- $p(x)$ *is increasing in* $(\infty, -1] \cup [1, \infty)$.

Step 1



Step 1



Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x + a)\right)^2 - 1$

Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x + a)\right)^2 - 1$
- Use simple case analyses

Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x + a)\right)^2 - 1$
- ... and the fact that a polynomial of degree d taking values in $[-b, b]$ on $[-1, 1]$ is bounded by $b|2x|^d$ for all $|x| > 1$

Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x+a)\right)^2 - 1$
- ... and the fact that a polynomial of degree d taking values in $[-b, b]$ on $[-1, 1]$ is bounded by $b|2x|^d$ for all $|x| > 1$
- ... to get the following result

Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x+a)\right)^2 - 1$
- ... and the fact that a polynomial of degree d taking values in $[-b, b]$ on $[-1, 1]$ is bounded by $b|2x|^d$ for all $|x| > 1$
- ... to get the following result

Lemma

There is a polynomial $P(x)$ of degree $K = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x+a)\right)^2 - 1$
- ... and the fact that a polynomial of degree d taking values in $[-b, b]$ on $[-1, 1]$ is bounded by $b|2x|^d$ for all $|x| > 1$
- ... to get the following result

Lemma

There is a polynomial $P(x)$ of degree $K = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

- $P(x) \geq \mathrm{sgn}(x)$ *for all* $x \in \mathbb{R}$

Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x+a)\right)^2 - 1$
- ... and the fact that a polynomial of degree d taking values in $[-b, b]$ on $[-1, 1]$ is bounded by $b|2x|^d$ for all $|x| > 1$
- ... to get the following result

Lemma

There is a polynomial $P(x)$ of degree $K = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

- $P(x) \geq \mathrm{sgn}(x)$ *for all* $x \in \mathbb{R}$
- $P(x) \in [\mathrm{sgn}(x), \mathrm{sgn}(x) + \epsilon]$ *for all* $x \in [-1/2, -2a] \cup [0, 1/2]$

Completing Step 2

- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x+a)\right)^2 - 1$
- ... and the fact that a polynomial of degree d taking values in $[-b, b]$ on $[-1, 1]$ is bounded by $b|2x|^d$ for all $|x| > 1$
- ... to get the following result

Lemma

There is a polynomial $P(x)$ of degree $K = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

- $P(x) \geq \mathrm{sgn}(x)$ *for all* $x \in \mathbb{R}$
- $P(x) \in [\mathrm{sgn}(x), \mathrm{sgn}(x) + \epsilon]$ *for all* $x \in [-1/2, -2a] \cup [0, 1/2]$
- $P(x) \in [-1, 1 + \epsilon]$ *for all* $x \in (-2a, 0)$

Completing Step 2

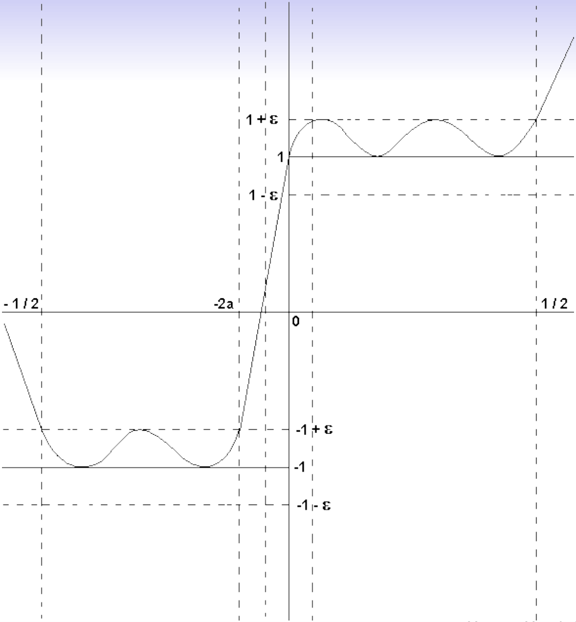
- Let $P(x) = \frac{1}{2}\left(1 + \epsilon^2 + p(x+a)\right)^2 - 1$
- ... and the fact that a polynomial of degree d taking values in $[-b, b]$ on $[-1, 1]$ is bounded by $b|2x|^d$ for all $|x| > 1$
- ... to get the following result

Lemma

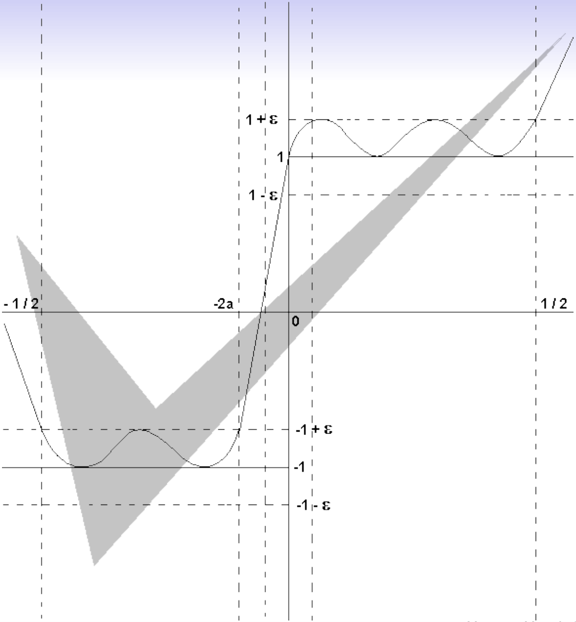
There is a polynomial $P(x)$ of degree $K = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ such that

- $P(x) \geq \mathrm{sgn}(x)$ *for all* $x \in \mathbb{R}$
- $P(x) \in [\mathrm{sgn}(x), \mathrm{sgn}(x) + \epsilon]$ *for all* $x \in [-1/2, -2a] \cup [0, 1/2]$
- $P(x) \in [-1, 1+\epsilon]$ *for all* $x \in (-2a, 0)$
- $|P(x)| \leq 2 \cdot (4x)^K$ *for all* $|x| \geq 1/2$.

Step 2



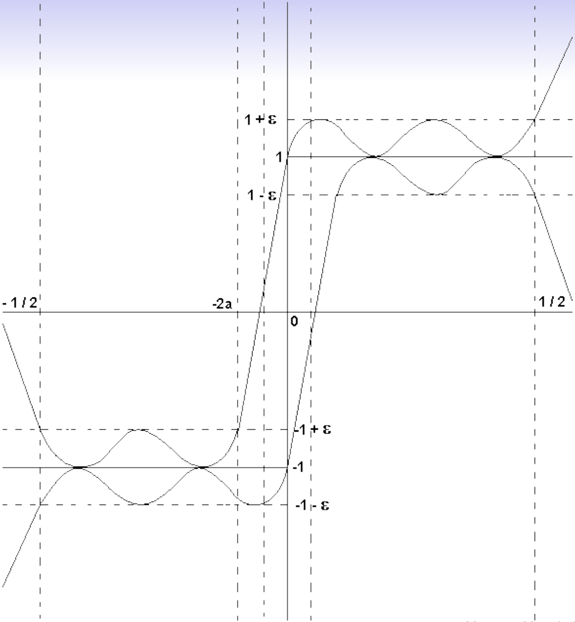
Step 2



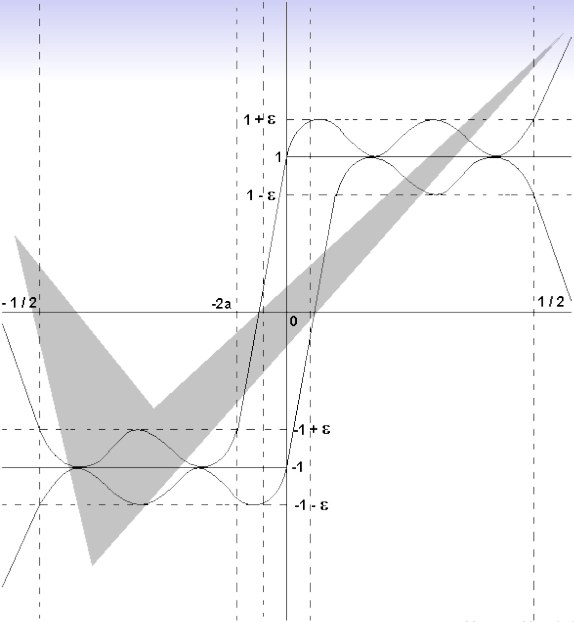
Completing Step 3(i)/3(ii)

Left as an exercise ☻

Step 3



Step 3



Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it under the Gaussian distribution
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it under the Gaussian distribution
- Now use the fact that values taken by homogeneous 'regular' linear polynomials are distributed normally

Plan of attack

Plan of attack :

- Construct a polynomial that gives a nice point wise approximation to the sgn function
- Use it to construct a polynomial that upper bounds the sgn function while closely approximating it under the Gaussian distribution
- Use it to construct a polynomial that lower bounds the sgn function while closely approximating it under the Gaussian distribution
- Now use the fact that values taken by homogeneous 'regular' linear polynomials are distributed normally
- A regular halfspace is one in which no weight is "large", i.e. if $w_i \leq \epsilon \|w\|_2$ for all i, then we call the halfspace ϵ-regular

An Effective Central Limit Theorem

Theorem (Berry-Esséen)

Let $X_1, \ldots, X_n$ be a sequence of independent random variables satisfying $\mathbf{E}[X_i] = 0$ for all i, $\sqrt{\sum_i \mathbf{E}\left[X_i^2\right]} = \sigma$ and $\sum_i \mathbf{E}\left[\left|X_i^3\right|\right] = \rho$. Let $S = (X_1 +, \ldots, + X_n)/\sigma$ and let F be the cumulative distribution function of S and Φ be the same for $N(0,1)$. Then

$$\sup_x |F(x) - \Phi(x)| \leq \rho/\sigma^3.$$

Regular Halfspaces generate Normally distributed outputs

Theorem

Let $x_1, \ldots, x_n \in_R -1, 1, w_1, \ldots, w_n \in \mathbb{R}$*. Let* $\sigma = \|w\|_2$ *and assume* $w_i \leq \tau \cdot \sigma$*. Then for any* $[a, b] \subset \mathbb{R}$*,*

$$\left| \Pr[a \leq w_1 x_1 + \ldots + w_n x_n \leq b] - \Phi\left(\frac{a}{\sigma}, \frac{b}{\sigma}\right) \right| \leq 2\tau.$$

Regular Halfspaces generate Normally distributed outputs

Theorem

Let $x_1, \ldots, x_n \in_R -1, 1, w_1, \ldots, w_n \in \mathbb{R}$. *Let* $\sigma = \|w\|_2$ *and assume* $w_i \leq \tau \cdot \sigma$. *Then for any* $[a, b] \subset \mathbb{R}$,

$$\left| \Pr[a \leq w_1 x_1 + \ldots + w_n x_n \leq b] - \Phi\left(\frac{a}{\sigma}, \frac{b}{\sigma}\right) \right| \leq 2\tau.$$

- Let $X_i = w_i x_i$, then $\mathbf{E}[X_i] = 0, \mathbf{E}[X_i^2] = w_i^2, \mathbf{E}[|X_i|^3] = |w_i|^3$

Regular Halfspaces generate Normally distributed outputs

Theorem

Let $x_1, \ldots, x_n \in_R -1, 1, w_1, \ldots, w_n \in \mathbb{R}$. *Let* $\sigma = \|w\|_2$ *and assume* $w_i \leq \tau \cdot \sigma$. *Then for any* $[a, b] \subset \mathbb{R}$,

$$\left|\Pr[a \leq w_1 x_1 + \ldots + w_n x_n \leq b] - \Phi\left(\frac{a}{\sigma}, \frac{b}{\sigma}\right)\right| \leq 2\tau.$$

- Let $X_i = w_i x_i$, then $\mathbf{E}[X_i] = 0, \mathbf{E}[X_i^2] = w_i^2, \mathbf{E}[|X_i|^3] = |w_i|^3$

Theorem (Hoeffding)

For any $w \in \mathbb{R}^n$. *For any* $\gamma > 0$, *we have*

$$\Pr_{x \leftarrow \mathcal{U}}[|w \cdot x| > \gamma \|w\|] \leq e^{-\gamma^2/2}$$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold $(|\theta| \leq Z/4)$, $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 1**: $x \in [-\epsilon/Z, 0]$, Error: $2 + \epsilon$, Probability : $\leq 3\epsilon$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold $(|\theta| \leq Z/4)$, $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 2**: $|x| \leq 1/2$, Error: ϵ, Probability : ≤ 1

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold $(|\theta| \leq Z/4)$, $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 3(i)**: $x \in [1/2, 1]$, Error: $2 \cdot 4^K - 1$, Probability : $\leq e^{-Z^2/32}$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 3(ii)**: $x \in [1, 3/2]$, Error: $2 \cdot 6^K - 1$, Probability : $e^{-4Z^2/32}$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 3(iii)**: $x \in [3/2, 2]$, Error: $2 \cdot 8^K - 1$, Probability : $e^{-9Z^2/32}$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 3(iv)**: ...

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold $(|\theta| \leq Z/4)$, $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 3(iv)**: ...
- In all we get $\mathbf{E}_x[u(x) - h(x)] \leq \mathcal{O}(\epsilon)$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 3(iv)**: ...
- In all we get $\mathbf{E}_x[u(x) - h(x)] \leq \mathcal{O}(\epsilon)$
- Note : The normalization by Z only required to bound the contribution of events $3(\dots)$

Let the con begin !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with small threshold ($|\theta| \leq Z/4$), $Z = \epsilon/2a = \mathcal{O}(1/\epsilon \log(1/\epsilon))$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - \theta}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] \leq \mathbf{E}_{x \in [-\epsilon/Z, 0]} + \mathbf{E}_{|x| \leq 1/2} + \mathbf{E}_{|x| \geq 1/2}[u(x) - h(x)]$
- **Event 3(iv)**: ...
- In all we get $\mathbf{E}_x[u(x) - h(x)] \leq \mathcal{O}(\epsilon)$
- Note : The normalization by Z only required to bound the contribution of events $3(\ldots)$
- One can lower bound the halfspace using $l(x) = -u(-x)$

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x\rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.
- Let $g(x) = \mathrm{sgn}(\langle w \cdot x\rangle - Z/4)$, note $g(x) \geq h(x)$

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.
- Let $g(x) = \mathrm{sgn}(\langle w \cdot x \rangle - Z/4)$, note $g(x) \geq h(x)$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - Z/4}{Z}\right)$

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.
- Let $g(x) = \mathrm{sgn}(\langle w \cdot x \rangle - Z/4)$, note $g(x) \geq h(x)$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - Z/4}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] = \mathbf{E}_x[u(x) - g(x)] + \mathbf{E}_x[g(x) - h(x)]$

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.
- Let $g(x) = \mathrm{sgn}(\langle w \cdot x \rangle - Z/4)$, note $g(x) \geq h(x)$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - Z/4}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] = \mathbf{E}_x[u(x) - g(x)] + \mathcal{O}(\epsilon)$

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.
- Let $g(x) = \mathrm{sgn}(\langle w \cdot x \rangle - Z/4)$, note $g(x) \geq h(x)$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - Z/4}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] = \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon)$

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.
- Let $g(x) = \mathrm{sgn}(\langle w \cdot x \rangle - Z/4)$, note $g(x) \geq h(x)$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - Z/4}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] = \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon)$
- In all we get $\mathbf{E}_x[u(x) - h(x)] \leq \mathcal{O}(\epsilon)$

Let the con continue !

- ... for a regular halfspace $h(x) = \mathrm{sgn}(\langle w \cdot x \rangle - \theta)$ with large threshold ($|\theta| > Z/4$) - assume that $\theta > Z/4$ w.l.o.t.m.g.
- Let $g(x) = \mathrm{sgn}(\langle w \cdot x \rangle - Z/4)$, note $g(x) \geq h(x)$
- Upper bound the halfspace with $u(x) = P\left(\frac{\langle w \cdot x \rangle - Z/4}{Z}\right)$
- $\mathbf{E}_x[u(x) - h(x)] = \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon)$
- In all we get $\mathbf{E}_x[u(x) - h(x)] \leq \mathcal{O}(\epsilon)$
- Lower bound the halfspace using $l(x) = -1$: it works since the halfspace almost always outputs -1

Goal Accomplished !

Theorem

Any $K(\epsilon)$*-wise distribution* $\mathcal{O}(\epsilon)$*-fools any* ϵ*-regular halfspace where* $K(\epsilon) = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$.

Goal Accomplished !

Theorem

Any $K(\epsilon)$-wise distribution $\mathcal{O}(\epsilon)$-fools any ϵ-regular halfspace where $K(\epsilon) = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$.

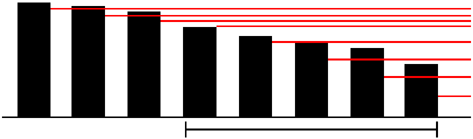
Wait till the end for some fun facts about this statement ...

Non-regular Halfspaces and Critical Indices

- Assume $|w_1| \geq |w_2| \geq \ldots |w_n|$ i.e. in decreasing order

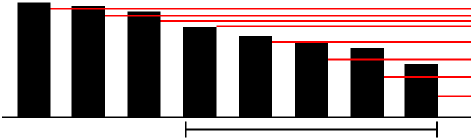
Non-regular Halfspaces and Critical Indices

- Assume $|w_1| \geq |w_2| \geq \ldots |w_n|$ i.e. in decreasing order
- The first point from where the (sub)-halfspace $(w_i, \ldots . w_n)$ becomes ϵ-regular is the critical index at ϵ



Non-regular Halfspaces and Critical Indices

- Assume $|w_1| \geq |w_2| \geq \ldots |w_n|$ i.e. in decreasing order
- The first point from where the (sub)-halfspace $(w_i, \ldots . w_n)$ becomes ϵ-regular is the critical index at ϵ



- We shall condition on how far do we need to go in order to get a regular halfspace

Small Critical Index

- “Most” of the halfspace is ϵ-regular

Small Critical Index

- "Most" of the halfspace is ϵ-regular
- Feed in full independence for the non-regular part to fool it - hopefully not much would be needed

Small Critical Index

- "Most" of the halfspace is ϵ-regular
- Feed in full independence for the non-regular part to fool it - hopefully not much would be needed
- If the critical index at ϵ is less than $L(\epsilon) = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ then we are done

Small Critical Index

- "Most" of the halfspace is ϵ-regular
- Feed in full independence for the non-regular part to fool it - hopefully not much would be needed
- If the critical index at ϵ is less than $L(\epsilon) = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$ then we are done

Theorem

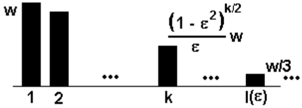
Any $K(\epsilon) + L(\epsilon)$-wise distribution $\mathcal{O}(\epsilon)$-fools any halfspace with critical index less than $L(\epsilon)$.

Large Critical Index

- Exploit "structural properties" of non-regular halfspaces

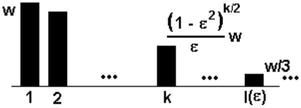
Large Critical Index

- Exploit "structural properties" of non-regular halfspaces
- Weights decrease rather rapidly in non-regular regions of the halfspace

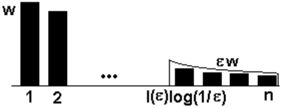


Large Critical Index

- Exploit "structural properties" of non-regular halfspaces
- Weights decrease rather rapidly in non-regular regions of the halfspace

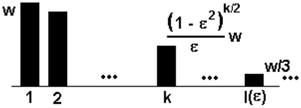


- ... and so do the norms of the weight vectors (i.e. $\sqrt{\sum w_i^2}$)

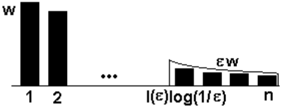


Large Critical Index

- Exploit "structural properties" of non-regular halfspaces
- Weights decrease rather rapidly in non-regular regions of the halfspace



- ... and so do the norms of the weight vectors (i.e. $\sqrt{\sum w_i^2}$)



- $l(\epsilon) = \mathcal{O}(1/\epsilon^2 \log(1/\epsilon))$

Some Technical Results

Intuition later ...

Some Technical Results

Intuition later ...

Theorem

Let $v_1 > v_2 > \ldots > v_t > 0$ *such that* $v_i \geq 3v_{i+1}$, *then for any* $x, y \in \{-1, 1\}^t, x \neq y$, *we have* $|\langle v \cdot x \rangle - \langle v \cdot y \rangle| \geq v_t$.

Some Technical Results

Intuition later ...

Theorem

Let $v_1 > v_2 > \ldots > v_t > 0$ *such that* $v_i \geq 3v_{i+1}$, *then for any* $x, y \in \{-1, 1\}^t, x \neq y$, *we have* $|\langle v \cdot x \rangle - \langle v \cdot y \rangle| \geq v_t$.

Theorem

Let $k = 4/\epsilon^2 \log^2(10/\epsilon)$, *then with probability at least* $1 - \epsilon/10$,

$$\left| \theta - \sum_{i=1}^{L(\epsilon)} w_i x_i \right| \geq |w_k|/4.$$

Some Technical Results

Intuition later ...

Theorem

Let $v_1 > v_2 > \ldots > v_t > 0$ *such that* $v_i \geq 3v_{i+1}$, *then for any* $x, y \in \{-1, 1\}^t, x \neq y$, *we have* $|\langle v \cdot x \rangle - \langle v \cdot y \rangle| \geq v_t$.

Theorem

Let $k = 4/\epsilon^2 \log^2(10/\epsilon)$, *then with probability at least* $1 - \epsilon/10$,

$$\left| \theta - \sum_{i=1}^{L(\epsilon)} w_i x_i \right| \geq |w_k|/4.$$

Theorem (Chebyshev)

For any random variable X *with* $\mathbf{E}[X] = \mu, \mathbf{Var}[X] = \sigma^2$, *for any* $k > 0$, $\Pr[|X - \mu| > k\sigma] \leq 1/k^2$.

Just a few more steps ...

- If $\sigma_T = \sqrt{\sum_{L(\epsilon)}^{n} w_i^2}$, then w.h.p. $\left|\theta - \sum_{i=1}^{L(\epsilon)} w_i x_i\right| \geq \sigma_T / 4\epsilon$

Just a few more steps ...

- If $\sigma_T = \sqrt{\sum_{L(\epsilon)}^{n} w_i^2}$, then w.h.p. $\left|\theta - \sum_{i=1}^{L(\epsilon)} w_i x_i\right| \geq \sigma_T / 4\epsilon$
- In such a situation unless $\left|\sum_{L(\epsilon)}^{n} w_i x_i\right| > \sigma_T / 4\epsilon$, the output of the halfspace is completely decided by the first $L(\epsilon)$ variables

Just a few more steps ...

- If $\sigma_T = \sqrt{\sum_{L(\epsilon)}^{n} w_i^2}$, then w.h.p. $\left|\theta - \sum_{i=1}^{L(\epsilon)} w_i x_i\right| \geq \sigma_T / 4\epsilon$
- In such a situation unless $\left|\sum_{L(\epsilon)}^{n} w_i x_i\right| > \sigma_T / 4\epsilon$, the output of the halfspace is completely decided by the first $L(\epsilon)$ variables
- But Chebyshev tells us that $\left|\sum_{L(\epsilon)}^{n} w_i x_i\right| \leq \sigma_T / 4\epsilon$ w.h.p.

Just a few more steps ...

- If $\sigma_T = \sqrt{\sum_{L(\epsilon)}^{n} w_i^2}$, then w.h.p. $\left|\theta - \sum_{i=1}^{L(\epsilon)} w_i x_i\right| \geq \sigma_T / 4\epsilon$
- In such a situation unless $\left|\sum_{L(\epsilon)}^{n} w_i x_i\right| > \sigma_T / 4\epsilon$, the output of the halfspace is completely decided by the first $L(\epsilon)$ variables
- But Chebyshev tells us that $\left|\sum_{L(\epsilon)}^{n} w_i x_i\right| \leq \sigma_T / 4\epsilon$ w.h.p.

Theorem

Any $L(\epsilon) + 2$-wise distribution $\mathcal{O}(\epsilon)$-fools any halfspace with critical index more than $L(\epsilon)$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution $\mathcal{O}(\epsilon)$-fools any halfspace where $K(\epsilon) = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution ϵ -fools any halfspace where $K(\epsilon) = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = \mathcal{O}(1/\epsilon^2 \log^2(1/\epsilon))$.

Done !

Theorem

Any $K(\epsilon)$*-wise distribution* 12ϵ *-fools any halfspace where*
$K(\epsilon) = 1/\epsilon^2 \log^2(1/\epsilon)$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

- i.e. the result is non-trivial only if $n > 2^{32}$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

- i.e. the result is non-trivial only if $n > 4294967296$.

Done !

Theorem

Any $K(\epsilon)$*-wise distribution* 12ϵ *-fools any halfspace where* $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

- i.e. the result is non-trivial only if $n >$ wait ... forgot something.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

- i.e. the result is non-trivial only if $n > 2^{42}$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

- i.e. the result is non-trivial only if $n > 4398046511104$.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

- i.e. the result is non-trivial only if $n > 4398046511104$.
- The results are tight :

Theorem ([BGGP])

There exists a $C > 0$ such that for every $k \geq 2$,

$$\max_{\mathcal{D} \in \mathcal{A}(n,k)} \left| \Pr_{x \in \mathcal{D}} [\mathrm{Maj}(x) = 1] - \frac{1}{2} \right| \geq \frac{C}{\sqrt{k \log k}}$$

.

Done !

Theorem

Any $K(\epsilon)$-wise distribution 12ϵ -fools any halfspace where $K(\epsilon) = 181923848 \cdot 1/\epsilon^2 \log^2(1/\epsilon)$.

- i.e. the result is non-trivial only if $n > 4398046511104$.
- The results are tight :

Theorem ([BGGP])

There exists a $C > 0$ *such that for every* $k \geq 2$,

$$\max_{\mathcal{D} \in \mathcal{A}(n,k)} \left| \Pr_{x \in \mathcal{D}} [\mathrm{Maj}(x) = 1] - \frac{1}{2} \right| \geq \frac{C}{\sqrt{k \log k}}$$

.

- Easier to verify for $k = n - 1$

Post [DGJ$^+$09] ...

- [KNW10] give an alternate proof of the [DGJ$^+$09] based on new techniques - there is some worsening of parameters $K(\epsilon) = \epsilon^{-2} \log^{2+o(1)}(1/\epsilon)$

Post [DGJ$^+$09] ...

- [KNW10] give an alternate proof of the [DGJ$^+$09] based on new techniques - there is some worsening of parameters $K(\epsilon) = \epsilon^{-2} \log^{2+o(1)}(1/\epsilon)$
- [DKN] extend ideas used in [KNW10] to show that thresholded quadratic polynomials can be ϵ-fooled by $\tilde{\Omega}(\epsilon^{-9})$ independence

Post [DGJ$^+$09] ...

- [KNW10] give an alternate proof of the [DGJ$^+$09] based on new techniques - there is some worsening of parameters $K(\epsilon) = \epsilon^{-2} \log^{2+o(1)}(1/\epsilon)$
- [DKN] extend ideas used in [KNW10] to show that thresholded quadratic polynomials can be ϵ-fooled by $\tilde{\Omega}(\epsilon^{-9})$ independence
- the result extends to intersection of constant number of halfspaces - dependence on number of halfspaces is polynomial

Post [DGJ⁺09] ...

- [MZ] give explicit pseudorandom generators with seed length $2^{\mathcal{O}(d)} \log n/\epsilon^{8d+3}$ for thresholded polynomials of degree d

Post [DGJ$^+$09] ...

- [MZ] give explicit pseudorandom generators with seed length $2^{\mathcal{O}(d)} \log n/\epsilon^{8d+3}$ for thresholded polynomials of degree d
- The construction gives improved PRG constructions for halfspaces with seed length $\mathcal{O}(\log n \log(1/\epsilon))$ for $\epsilon = \Omega(1/\mathrm{poly}(n))$

Post [DGJ$^+$09] ...

- [MZ] give explicit pseudorandom generators with seed length $2^{\mathcal{O}(d)} \log n/\epsilon^{8d+3}$for thresholded polynomials of degree d
- The construction gives improved PRG constructions for halfspaces with seed length $\mathcal{O}(\log n \log(1/\epsilon))$ for $\epsilon = \Omega(1/\text{poly}(n))$
- and seed length $\mathcal{O}(\log n)$ for $\epsilon = \Omega(1/\text{poly}(\log n))$

Post [DGJ$^+$09] ...

- [MZ] give explicit pseudorandom generators with seed length $2^{\mathcal{O}(d)} \log n/\epsilon^{8d+3}$ for thresholded polynomials of degree d
- The construction gives improved PRG constructions for halfspaces with seed length $\mathcal{O}(\log n \log(1/\epsilon))$ for $\epsilon = \Omega(1/\mathrm{poly}(n))$
- and seed length $\mathcal{O}(\log n)$ for $\epsilon = \Omega(1/\mathrm{poly}(\log n))$
- However non-explicit arguments show the existence of $\mathcal{O}(d \log n + \log(1/\epsilon))$ seed length PRGs to fool degree d Polynomial threshold functions [MZ]

Post [DGJ$^+$09] ...

- [GOWZ10] consider fooling functions of halfspaces

Post [DGJ$^+$09] ...

- [GOWZ10] consider fooling functions of halfspaces
- Give a modification of the [MZ] construction to yield a $\mathcal{O}((d\log(ds/\epsilon) + \log n) \cdot \log(ds/\epsilon))$ seed length PGR for arbitrary decision trees of halfspaces of size s and depth d

Post [DGJ$^+$09] ...

- [GOWZ10] consider fooling functions of halfspaces

- Give a modification of the [MZ] construction to yield a $\mathcal{O}((d \log(ds/\epsilon) + \log n) \cdot \log(ds/\epsilon))$ seed length PGR for arbitrary decision trees of halfspaces of size s and depth d
- i.e. TC^0 can be fooled by a seed length of $\mathcal{O}(\log^2(n/\epsilon))$

Post [DGJ$^+$09] ...

- [GOWZ10] consider fooling functions of halfspaces

- Give a modification of the [MZ] construction to yield a $\mathcal{O}((d \log(ds/\epsilon) + \log n) \cdot \log(ds/\epsilon))$ seed length PGR for arbitrary decision trees of halfspaces of size s and depth d
- i.e. TC^0 can be fooled by a seed length of $\mathcal{O}(\log^2(n/\epsilon))$
- Also extend the construction given in [DGJ$^+$09] to show that $\tilde{\mathcal{O}}(d^4 s^2/\epsilon^2)$-wise independence fools arbitrary decision trees of halfspaces of size s and depth d

Post [DGJ$^+$09] ...

- [GOWZ10] consider fooling functions of halfspaces under various Product Distributions
- Give a modification of the [MZ] construction to yield a $\mathcal{O}((d \log(ds/\epsilon) + \log n) \cdot \log(ds/\epsilon))$ seed length PGR for arbitrary decision trees of halfspaces of size s and depth d
- i.e. TC^0 can be fooled by a seed length of $\mathcal{O}(\log^2(n/\epsilon))$
- Also extend the construction given in [DGJ$^+$09] to show that $\tilde{\mathcal{O}}(d^4 s^2/\epsilon^2)$-wise independence fools arbitrary decision trees of halfspaces of size s and depth d

Post [DGJ$^+$09] ...

- [GOWZ10] consider fooling functions of halfspaces under various Product Distributions
- Give a modification of the [MZ] construction to yield a $\mathcal{O}((d \log(ds/\epsilon) + \log n) \cdot \log(ds/\epsilon))$ seed length PGR for arbitrary decision trees of halfspaces of size s and depth d
- i.e. TC^0 can be fooled by a seed length of $\mathcal{O}(\log^2(n/\epsilon))$
- Also extend the construction given in [DGJ$^+$09] to show that $\tilde{\mathcal{O}}(d^4 s^2/\epsilon^2)$-wise independence fools arbitrary decision trees of halfspaces of size s and depth d
- [HKM09] do slightly better at fooling intersection of k regular halfspaces using seed length $\mathcal{O}\left(\epsilon^{-5} \log n \log^{9.1} k \log(1/\epsilon)\right)$

References

Louay Bazzi.
Polylogarithmic independence can fool DNF formulas.
In *IEEE Symposium on Foundations of Computer Science*, 2007.

Itai Benjamini, Ori Gurel-Gurevich, and Ron Peled.
On k-wise independent events and percolation.
Available at `http://www.cims.nyu.edu/~peled/homepage_files/K-wise_extended_abstract_2.pdf`.

Mark Braverman.
Poly-logarithmic independence fools AC0 circuits.
In *IEEE Conference on Computational Complexity*, 2009.

Corinna Cortes, Leonid Kontorovich, and Mehryar Mohri.
Learning Languages with Rational Kernels.
In *Computational Learning Theory*, 2007.

References

Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco Servedio, and Emanuele Viola.
Bounded Independence fools Halfspaces.
In *IEEE Symposium on Foundations of Computer Science*, 2009.

Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson.
Bounded Independence Fools Degree-2 Threshold Functions.
Available at
`http://math.harvard.edu/~dankane/deg2ptf.pdf`.

Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman.
Fooling functions of halfspaces under product distributions.
Technical Report TR10-006, Electronic Colloquium on Computational Complexity, 2010.

References

Johan Håstad.
On the size of weights for threshold gates.
SIAM Journal on Discrete Mathematics, 7(3):484–492, 1994.

Prahladh Harsha, Adam Klivans, and Raghu Meka.
An Invariance Principle for Polytopes.
Technical Report TR09-144, Electronic Colloquium on Computational Complexity, 2009.

Daniel Kane, Jelani Nelson, and David Woodruff.
On the Exact Space Complexity of Sketching and Streaming Small Norms.
In *21st Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010.

Adam R. Klivans and Rocco A. Servedio.
Learning DNF in time $2^{\tilde{O}(n^{1/3})}$.
Journal of Computer and System Sciences, 68(2):303–318, 2004.

References

Wolfgang Maass and György Turan.
How fast can a threshold gate learn ?
In *Computational Learning Theory and Natural Learning Systems*, volume I: Constraints and Prospects, pages 381–414. The MIT Press, 1994.

Raghu Meka and David Zuckerman.
Pseudorandom Generators for Polynomial Threshold Functions.
Available at http://arxiv.org/abs/0910.4122.

Marvin Minsky and Seymour Papert.
Perceptrons.
The MIT Press, 1969.

Rocco A. Servedio.
Every linear threshold function has a low-weight approximator.
Computational Complexity, 16(2):180–209, 2007.

References

Alexander A. Razborov.
A Simple Proof of Bazzi's theorem.
Technical Report TR08-081, Electronic Colloquium on Computational Complexity, 2008.