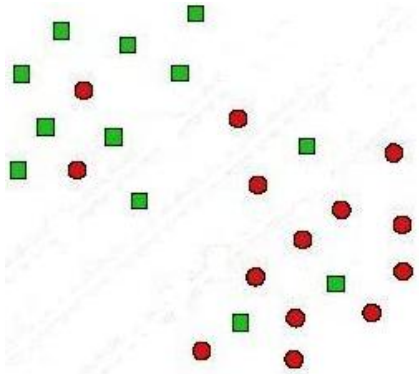
The complexity of learning halfspaces using generalized linear methods

Amit Daniely, *Nati Linial, Shai Shalev-Shwartz*

COLT 2014

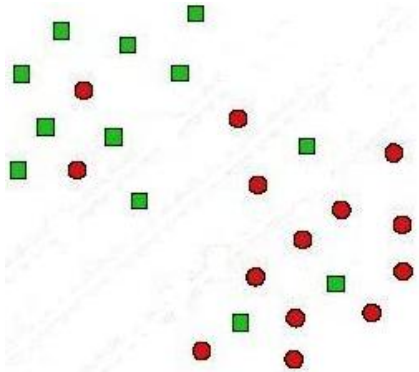
The generalized linear method for classification



The classification problem

- $\mathcal{D}$ is a distribution on $\mathcal{X} \times \{0,1\}$
- Based on examples $(x,y) \sim \mathcal{D}$, find good $h : \mathcal{X} \to \{0,1\}$

The generalized linear method for classification



The classification problem

- $\mathcal{D}$ is a distribution on $\mathcal{X} \times \{0, 1\}$
- Based on examples $(x, y) \sim \mathcal{D}$, find good $h : \mathcal{X} \to \{0, 1\}$

A popular approach:

The generalized linear method

- Embed $\mathcal{X}$ in a high dimensional space.
- Find a halfspace with a small (convex surrogate) loss.

The generalized linear method for classification

- Very popular:
  - Kernel SVM.
  - Regression based algorithms.
  - Learning using polynomials.
  - Fourier transform based methods.
  - Lasso
  - ...

The generalized linear method for classification

- Very popular:
  - Kernel SVM.
  - Regression based algorithms.
  - Learning using polynomials.
  - Fourier transform based methods.
  - Lasso
  - ...
- Best guarantees for many problems:
  - Large margin halfspaces.
  - Decision trees.
  - DNFs.
  - Intersection of halfspaces.
  - ...

The generalized linear method for classification

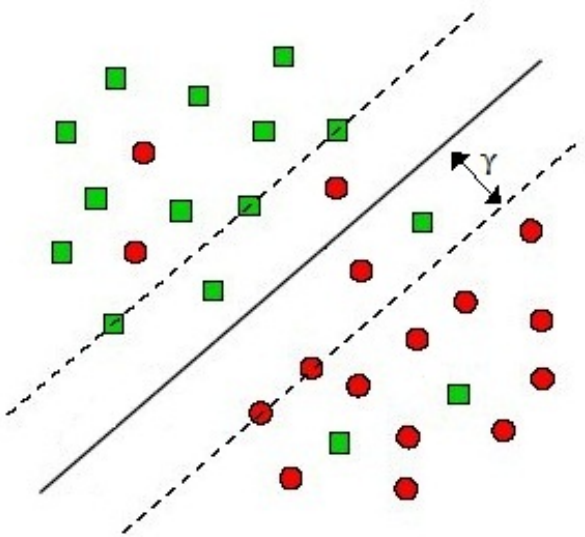
- Very popular:
  - Kernel SVM.
  - Regression based algorithms.
  - Learning using polynomials.
  - Fourier transform based methods.
  - Lasso
  - ...
- Best guarantees for many problems:
  - Large margin halfspaces.
  - Decision trees.
  - DNFs.
  - Intersection of halfspaces.
  - ...

In this work

Lower bounds for learning large-margin halfspaces.

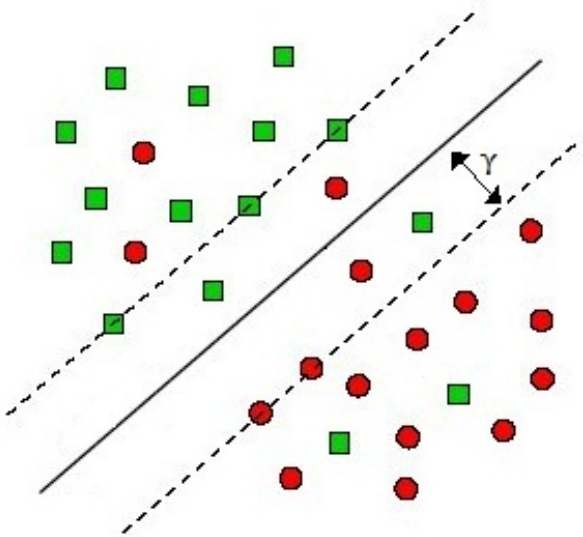
Learning large margin halfspaces

- $\mathcal{X}$ – vectors with norm ≤ 1.



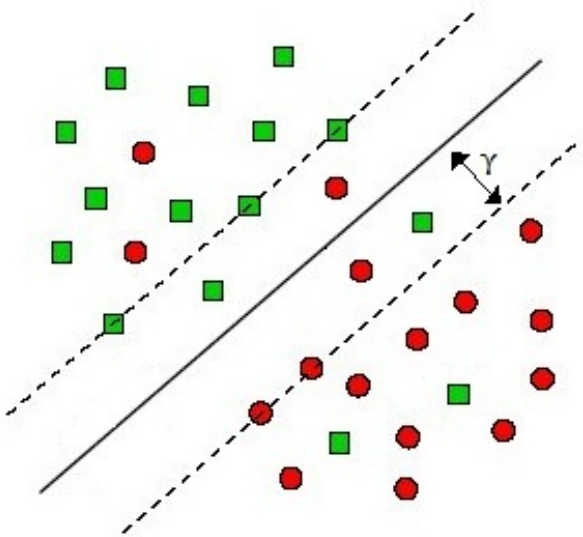
Learning large margin halfspaces

- $\mathcal{X}$ – vectors with norm ≤ 1.
- The γ-error of a hyperplane h is the probability that an example falls
  - At the wrong side of h. *or*
  - At distance $\leq \gamma$ from h.



Learning large margin halfspaces

- $\mathcal{X}$ – vectors with norm ≤ 1.
- The γ-error of a hyperplane h is the probability that an example falls
  - At the wrong side of h. *or*
  - At distance $\leq \gamma$ from h.
- $\mathrm{Err}_\gamma(\mathcal{D})$ is the γ-error of the best hyperplane.



Approximation and Efficient algorithms

- The learner's **input**:
  - $\gamma > 0$, $\epsilon > 0$ and oracle access to $\mathcal{D}$.

Approximation and Efficient algorithms

- The learner's **input**:
  - $\gamma > 0$, $\epsilon > 0$ and oracle access to $\mathcal{D}$.
- A learner enjoys $\alpha \geq 1$ approximation ratio if it outputs (w.p. $\geq \frac{9}{10}$)
  - $f : \mathcal{X} \to \{0, 1\}$ with $\mathrm{Err}_{\mathcal{D}}(f) < \alpha \cdot \mathrm{Err}_{\gamma}(\mathcal{D}) + \epsilon$.
  - Here, $\mathrm{Err}_{\mathcal{D}}(f) := \Pr_{(x,y)\sim\mathcal{D}}(f(x) \neq y)$

Approximation and Efficient algorithms

- The learner's **input**:
  - $\gamma > 0$, $\epsilon > 0$ and oracle access to $\mathcal{D}$.
- A learner enjoys $\alpha \geq 1$ approximation ratio if it outputs (w.p. $\geq \frac{9}{10}$)
  - $f : \mathcal{X} \to \{0, 1\}$ with $\mathrm{Err}_{\mathcal{D}}(f) < \alpha \cdot \mathrm{Err}_{\gamma}(\mathcal{D}) + \epsilon$.
  - Here, $\mathrm{Err}_{\mathcal{D}}(f) := \Pr_{(x,y)\sim\mathcal{D}}(f(x) \neq y)$
- A learner is efficient if it
  - Uses $\mathrm{poly}\left(\frac{1}{\gamma}, \frac{1}{\epsilon}\right)$ examples.
  - Runs in polynomial time.

Previous results

Upper bounds (appr. ratio)

Previous results

Upper bounds (appr. ratio)

- Plain SVM (no kernel): $1/\gamma$

Previous results

Upper bounds (appr. ratio)

- Plain SVM (no kernel): $1/\gamma$
- (SB 12, using kernel SVM): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$

Previous results

Upper bounds (appr. ratio)

- Plain SVM (no kernel): $1/\gamma$
- (SB 12, using kernel SVM): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$
- (LS 11): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$

Previous results

Upper bounds (appr. ratio)

- Plain SVM (no kernel): $1/\gamma$
- (SB 12, using kernel SVM): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$
- (LS 11): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$

Lower bounds (appr. ratio)

- General algorithms: 1 (KS,FGKP 06, SSS 10 under crypto. assumptions)

Previous results

Upper bounds (appr. ratio)

- Plain SVM (no kernel): $1/\gamma$
- (SB 12, using kernel SVM): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$
- (LS 11): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$

Lower bounds (appr. ratio)

- General algorithms: 1 (KS,FGKP 06, SSS 10 under crypto. assumptions)
- General algorithms: Any constant (DLS 14 under average case complexity assumptions)

Previous results

Upper bounds (appr. ratio)

- Plain SVM (no kernel): $1/\gamma$
- (SB 12, using kernel SVM): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$
- (LS 11): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$

Lower bounds (appr. ratio)

- General algorithms: 1 (KS,FGKP 06, SSS 10 under crypto. assumptions)
- General algorithms: Any constant (DLS 14 under average case complexity assumptions)
- Plain SVM (no kernel): $1/\gamma$ (BLSS 13, LS 11)

Previous results

Upper bounds (appr. ratio)

- Plain SVM (no kernel): $1/\gamma$
- (SB 12, using kernel SVM): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$
- (LS 11): $\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$

Lower bounds (appr. ratio)

- General algorithms: 1 (KS,FGKP 06, SSS 10 under crypto. assumptions)
- General algorithms: Any constant (DLS 14 under average case complexity assumptions)
- Plain SVM (no kernel): $1/\gamma$ (BLSS 13, LS 11)
- Kernel SVM: ?
- Generalized linear algorithm: ?

Main result

Main result (informal)

Every efficient generalized linear learner has appr. ratio $\geq \frac{1/\gamma}{\log^8(1/\gamma)}$

Main result

Main result (informal)

Every efficient generalized linear learner has appr. ratio $\geq \frac{1/\gamma}{\log^8(1/\gamma)}$

- Essentially matches the best known upper bound ($\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$).

Main result

Main result (informal)

Every efficient generalized linear learner has appr. ratio $\geq \frac{1/\gamma}{\log^8(1/\gamma)}$

- Essentially matches the best known upper bound ($\frac{1/\gamma}{\log^{0.5}(1/\gamma)}$).
- Applies to many techniques (Kernel SVM, Regression, Polynomials learning, Fourier transform, Lasso,...)

Outline

1. Our Technique
2. Summery and open questions

Our technique

- **Main challenge:** rule out **every** feature mapping.

Our technique

- **Main challenge:** rule out **every** feature mapping.
- Established by a new technique to lower bound generalized linear learners.

Our technique

- **Main challenge:** rule out **every** feature mapping.
- Established by a new technique to lower bound generalized linear learners.
- We will use various mathematical machinery:
  - Functional analysis
  - Harmonic analysis and representation theory over the d-dimensional sphere
  - Geometry
  - Reproducing Kernel Hilbert Spaces
  - Orthogonal polynomials
  - New symmetrization technique

Kernel based learners

- Let H be a Hilbert space with a unit ball B.

Kernel based learners

- Let H be a Hilbert space with a unit ball B.
- Fix $\psi : \mathcal{X} \to B$ and convex surrogate $l : \mathbb{R} \to \mathbb{R}_+$.

Kernel based learners

- Let H be a Hilbert space with a unit ball B.
- Fix $\psi : \mathcal{X} \to B$ and convex surrogate $l : \mathbb{R} \to \mathbb{R}_+$.
- For $w \in H, b \in \mathbb{R}$ consider the classifier

$$h_{w,b}(x) = \operatorname{sign}\left(\langle w, \psi(x)\rangle + b\right)$$

and denote $\operatorname{Err}_l(h_{w,b}) = \mathrm{E}_{(x,y)\sim\mathcal{D}} l(y(\langle w, x\rangle + b))$

Kernel based learners

- Let H be a Hilbert space with a unit ball B.
- Fix $\psi : \mathcal{X} \to B$ and convex surrogate $l : \mathbb{R} \to \mathbb{R}_+$.
- For $w \in H, b \in \mathbb{R}$ consider the classifier

$$h_{w,b}(x) = \text{sign}\left(\langle w, \psi(x)\rangle + b\right)$$

  and denote $\text{Err}_l(h_{w,b}) = \text{E}_{(x,y)\sim\mathcal{D}} l(y(\langle w, x\rangle + b))$
- A kernel based learner corresponding to ψ, l:

$$\begin{aligned} \min \quad & \text{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

Kernel based learners

- Let H be a Hilbert space with a unit ball B.
- Fix $\psi : \mathcal{X} \to B$ and convex surrogate $l : \mathbb{R} \to \mathbb{R}_+$.
- For $w \in H, b \in \mathbb{R}$ consider the classifier

$$h_{w,b}(x) = \text{sign}\left(\langle w, \psi(x)\rangle + b\right)$$

  and denote $\text{Err}_l(h_{w,b}) = \text{E}_{(x,y)\sim\mathcal{D}} l(y(\langle w, x\rangle + b))$
- A kernel based learner corresponding to ψ, l:

$$\begin{aligned} \min \quad & \text{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

Main result

Every efficient kernel based learner has appr. ratio $\geq \frac{1/\gamma}{\log^8(1/\gamma)}$

Kernel based learners

- Let H be a Hilbert space with a unit ball B.
- Fix $\psi : \mathcal{X} \to B$ and convex surrogate $l : \mathbb{R} \to \mathbb{R}_+$.
- For $w \in H, b \in \mathbb{R}$ consider the classifier

$$h_{w,b}(x) = \text{sign}\left(\langle w, \psi(x)\rangle + b\right)$$

  and denote $\text{Err}_l(h_{w,b}) = \text{E}_{(x,y)\sim\mathcal{D}} l(y(\langle w, x\rangle + b))$
- A kernel based learner corresponding to ψ, l:

$$\begin{aligned} \min \quad & \text{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

Main result

Every efficient kernel based learner has appr. ratio $\geq \frac{1/\gamma}{\log^8(1/\gamma)}$

Must show: for every ψ, l, C there is $\mathcal{D}$ on which the algorithm performs poorly.

An equivalent formulation

- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

An equivalent formulation

- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

- Consider the **Hilbert space**

$$\mathcal{F} = \{h_{w,0} \mid w \in H\}, \qquad h_{w,0}(x) = \langle w, \psi(x) \rangle$$

with the norm $\|h_{w,0}\|_{\mathcal{F}} = \|w\|_H$

An equivalent formulation

- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

- Consider the **Hilbert space**

$$\mathcal{F} = \{h_{w,0} \mid w \in H\}, \qquad h_{w,0}(x) = \langle w, \psi(x) \rangle$$

  with the norm $\|h_{w,0}\|_{\mathcal{F}} = \|w\|_H$
- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f + b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

An equivalent formulation

- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

- Consider the **Hilbert space**

$$\mathcal{F} = \{h_{w,0} \mid w \in H\}, \qquad h_{w,0}(x) = \langle w, \psi(x) \rangle$$

with the norm $\|h_{w,0}\|_{\mathcal{F}} = \|w\|_H$

- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f + b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

- We can quantify over $\mathcal{F}$ instead of ψ.

An equivalent formulation

- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(h_{w,b}) \\ s.t. \quad & \|w\| \leq C(\gamma) \end{aligned}$$

- Consider the **Hilbert space**

$$\mathcal{F} = \{h_{w,0} \mid w \in H\}, \qquad h_{w,0}(x) = \langle w, \psi(x) \rangle$$

with the norm $\|h_{w,0}\|_{\mathcal{F}} = \|w\|_H$

- A **kernel based learner** corresponding to $\psi : \mathcal{X} \to B$:

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f + b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

- We can quantify over $\mathcal{F}$ instead of ψ.
- We'll use both formulations.

The proof – Bird's eye view

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f+b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

The proof – Bird's eye view

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f+b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

- **Step 1:** Polynomial sample $\Rightarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$.

The proof – Bird's eye view

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f+b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

- **Step 1:** Polynomial sample $\Rightarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$.
  - Converse of the well known: Polynomial sample $\Leftarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$

The proof – Bird's eye view

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f+b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

- **Step 1:** Polynomial sample $\Rightarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$.
  - Converse of the well known: Polynomial sample $\Leftarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$
  - Proved using a geometric argument.

The proof – Bird's eye view

$$\begin{array}{ll} \min & \mathrm{Err}_l(f+b) \\ s.t. & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{array}$$

- **Step 1:** Polynomial sample $\Rightarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$.
  - Converse of the well known: Polynomial sample $\Leftarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$
  - Proved using a geometric argument.
- **Step 2:** $\mathcal{X} = [-1,1]$ and $\mathcal{F} = \{p : [-1,1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

The proof – Bird's eye view

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f+b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

- **Step 1:** Polynomial sample $\Rightarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$.
  - Converse of the well known: Polynomial sample $\Leftarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$
  - Proved using a geometric argument.
- **Step 2:** $\mathcal{X} = [-1,1]$ and $\mathcal{F} = \{p : [-1,1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.
- **Step 3:** $\mathcal{X} = B^d$ and $\mathcal{F} = \{p : B^{d-1} \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

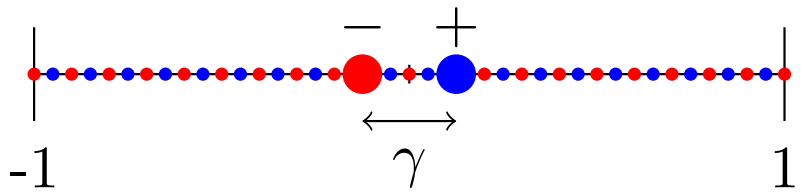
The proof – Bird's eye view

$$\begin{aligned} \min \quad & \mathrm{Err}_l(f+b) \\ s.t. \quad & \|f\|_{\mathcal{F}} \leq C(\gamma),\ b \in \mathbb{R} \end{aligned}$$

- **Step 1:** Polynomial sample $\Rightarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$.
  - Converse of the well known: Polynomial sample $\Leftarrow C(\gamma) \leq \mathrm{poly}\,(1/\gamma)$
  - Proved using a geometric argument.
- **Step 2:** $\mathcal{X} = [-1,1]$ and $\mathcal{F} = \{p : [-1,1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.
- **Step 3:** $\mathcal{X} = B^d$ and $\mathcal{F} = \{p : B^{d-1} \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.
- **Step 4:** General case.

Step 2 – The one dimensional case

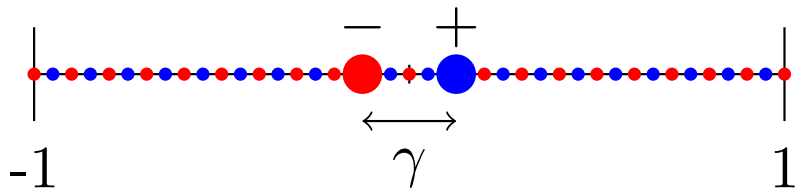
$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.



$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

Step 2 – The one dimensional case

$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

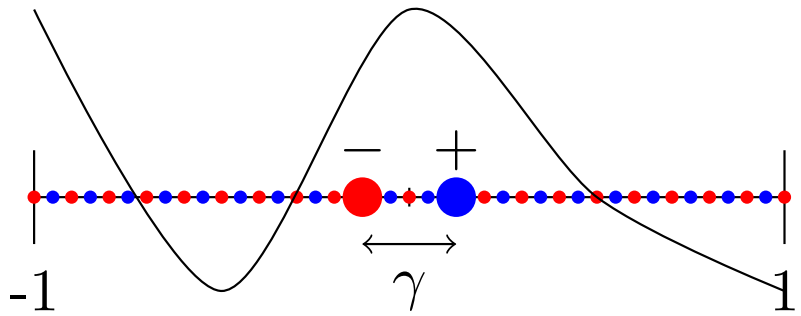


$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

- $\mathrm{Err}_\gamma(\mathcal{D}) \leq 2\gamma$.

Step 2 – The one dimensional case

$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

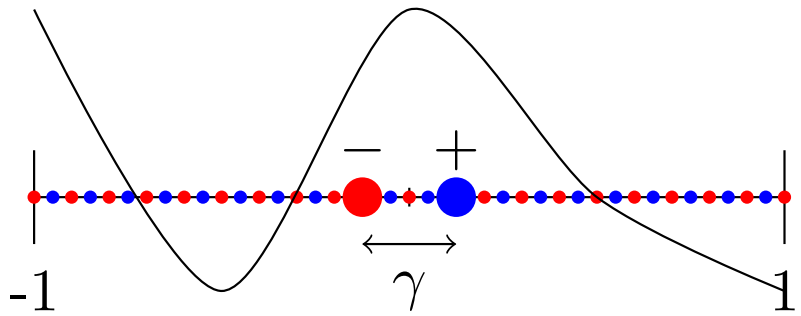


$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

- $\mathrm{Err}_\gamma(\mathcal{D}) \leq 2\gamma$.
- Let p be the polynomial returned by the algorithm.

Step 2 – The one dimensional case

$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

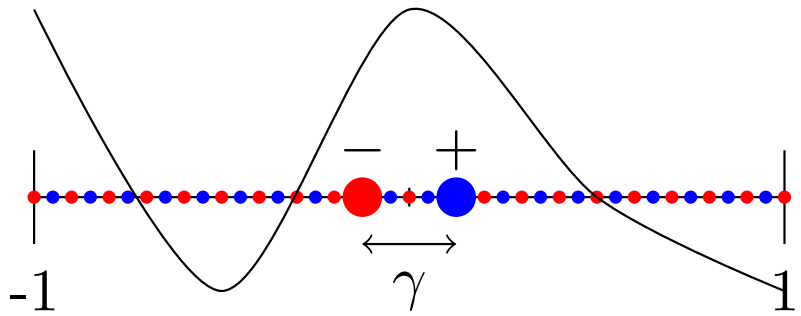


$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

- $\mathrm{Err}_\gamma(\mathcal{D}) \leq 2\gamma$.
- Let p be the polynomial returned by the algorithm.
- $\|p\|_2$ must be small.

Step 2 – The one dimensional case

$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

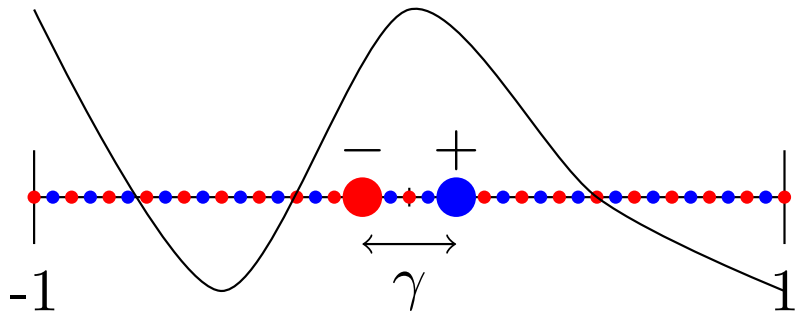


$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

- $\mathrm{Err}_\gamma(\mathcal{D}) \leq 2\gamma$.
- Let p be the polynomial returned by the algorithm.
- $\|p\|_2$ must be small.
- Since $\deg(p) \leq \log(1/\gamma)$, $\left\|\frac{\partial p}{\partial x}\right\|_\infty$ is small as well!

Step 2 – The one dimensional case

$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

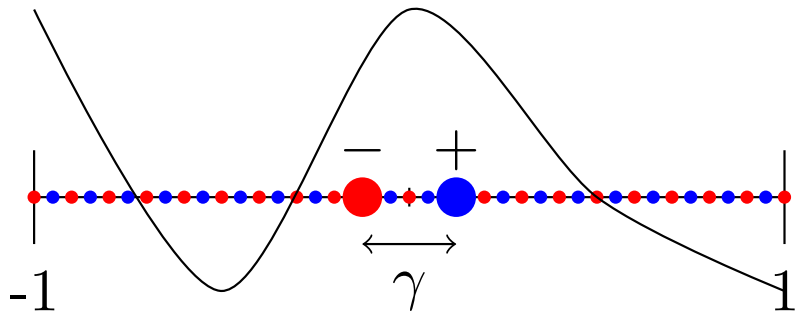


$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

- $\mathrm{Err}_\gamma(\mathcal{D}) \leq 2\gamma$.
- Let p be the polynomial returned by the algorithm.
- $\|p\|_2$ must be small.
- Since $\deg(p) \leq \log(1/\gamma)$, $\left\|\frac{\partial p}{\partial x}\right\|_\infty$ is small as well!
- $\Rightarrow$ The label of the heavy points is the same

Step 2 – The one dimensional case

$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.

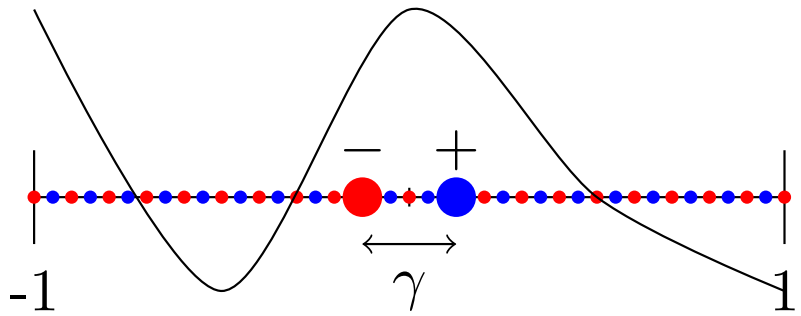


$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

- $\mathrm{Err}_\gamma(\mathcal{D}) \leq 2\gamma$.
- Let p be the polynomial returned by the algorithm.
- $\|p\|_2$ must be small.
- Since $\deg(p) \leq \log(1/\gamma)$, $\left\|\frac{\partial p}{\partial x}\right\|_\infty$ is small as well!
- $\Rightarrow$ The label of the heavy points is the same
- $\Rightarrow \mathrm{Err}_{\mathcal{D}}(p) \approx \frac{1}{2}$

Step 2 – The one dimensional case

$\mathcal{X} = [-1, 1]$ and $\mathcal{F} = \{p : [-1, 1] \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$.



$\Pr(\bullet) = \Pr(\bullet) = 0.5 - \gamma$

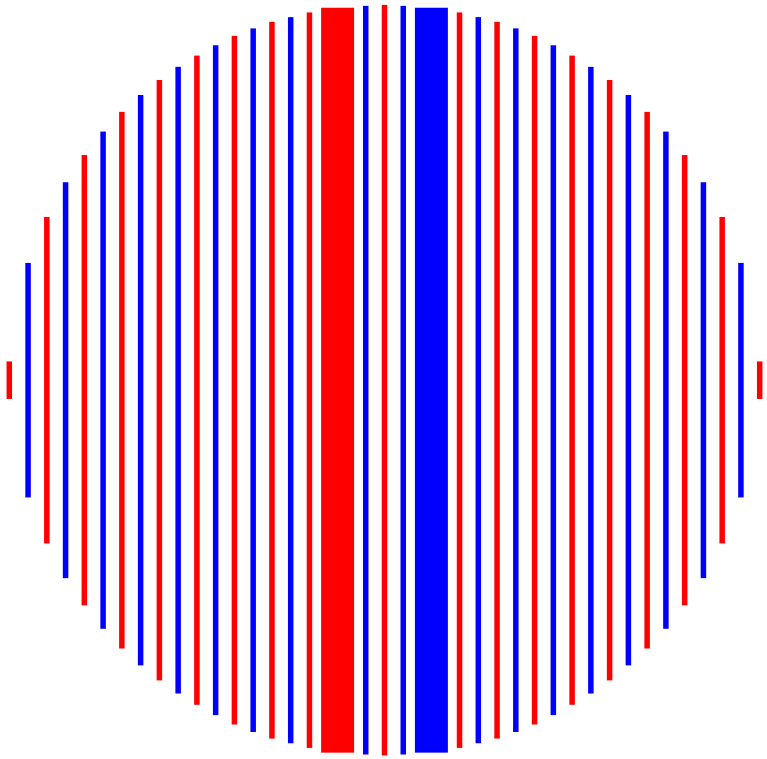
- $\mathrm{Err}_\gamma(\mathcal{D}) \leq 2\gamma$.
- Let p be the polynomial returned by the algorithm.
- $\|p\|_2$ must be small.
- Since $\deg(p) \leq \log(1/\gamma)$, $\left\|\frac{\partial p}{\partial x}\right\|_\infty$ is small as well!
- $\Rightarrow$ The label of the heavy points is the same
- $\Rightarrow \mathrm{Err}_{\mathcal{D}}(p) \approx \frac{1}{2}$
- Actual analysis is done using Chebyshev polynomials.

Step 3 – Lifting to the d-dimensional ball

$\mathcal{X} = B^d$ and $\mathcal{F} = \{p : B^{d-1} \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$

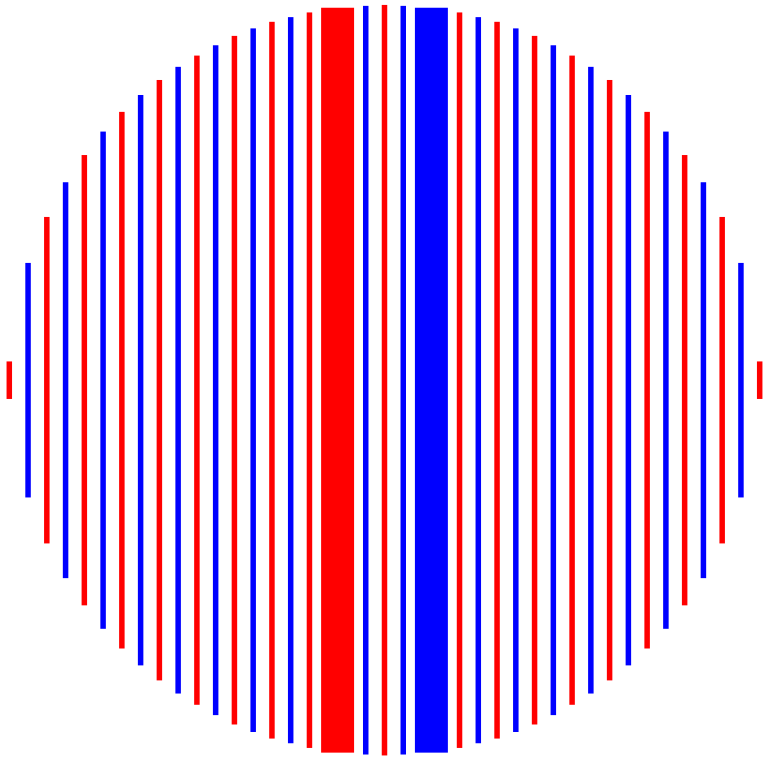
Step 3 – Lifting to the d-dimensional ball

$\mathcal{X} = B^d$ and $\mathcal{F} = \{p : B^{d-1} \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$



Step 3 – Lifting to the d-dimensional ball

$\mathcal{X} = B^d$ and $\mathcal{F} = \{p : B^{d-1} \to \mathbb{R} \mid \deg(p) \leq \log(1/\gamma)\}$



Analysed using Representation theory and Harmonic analysis over the sphere.

Step 4 – Symmetrization

- $\mathbb{O}(d)$ – the group of linear isometries $U : \mathbb{R}^d \to \mathbb{R}^d$.

Step 4 – Symmetrization

- $\mathbb{O}(d)$ – the group of linear isometries $U : \mathbb{R}^d \to \mathbb{R}^d$.
- The Hilbert space $\mathcal{F}$ is symmetric if

$$\forall f \in \mathcal{F}, U \in \mathbb{O}(d), \quad \|f\|_{\mathcal{F}} = \|f \circ U\|_{\mathcal{F}}$$

Step 4 – Symmetrization

- $\mathbb{O}(d)$ – the group of linear isometries $U : \mathbb{R}^d \to \mathbb{R}^d$.
- The Hilbert space $\mathcal{F}$ is symmetric if

$$\forall f \in \mathcal{F}, U \in \mathbb{O}(d), \quad \|f\|_{\mathcal{F}} = \|f \circ U\|_{\mathcal{F}}$$

- By "averaging over $\mathbb{O}(d)$", we introduce a symmetrization operator

$$\mathcal{F} \mapsto \mathrm{Sym}(\mathcal{F})$$

Step 4 – Symmetrization

- $\mathbb{O}(d)$ – the group of linear isometries $U : \mathbb{R}^d \to \mathbb{R}^d$.
- The Hilbert space $\mathcal{F}$ is symmetric if

$$\forall f \in \mathcal{F}, U \in \mathbb{O}(d), \quad \|f\|_{\mathcal{F}} = \|f \circ U\|_{\mathcal{F}}$$

- By "averaging over $\mathbb{O}(d)$", we introduce a symmetrization operator

$$\mathcal{F} \mapsto \mathrm{Sym}(\mathcal{F})$$

- We show that:
  - $\mathrm{Sym}(\mathcal{F})$ is as good as $\mathcal{F}$.

Step 4 – Symmetrization

- $\mathbb{O}(d)$ – the group of linear isometries $U : \mathbb{R}^d \to \mathbb{R}^d$.
- The Hilbert space $\mathcal{F}$ is symmetric if

$$\forall f \in \mathcal{F}, U \in \mathbb{O}(d), \quad \|f\|_{\mathcal{F}} = \|f \circ U\|_{\mathcal{F}}$$

- By "averaging over $\mathbb{O}(d)$", we introduce a symmetrization operator

$$\mathcal{F} \mapsto \mathrm{Sym}(\mathcal{F})$$

- We show that:
  - $\mathrm{Sym}(\mathcal{F})$ is as good as $\mathcal{F}$.
  - $\mathrm{Sym}(\mathcal{F})$ "almost coincides" with the space of polynomials of degree $\log(1/\gamma)$.

Step 4 – Symmetrization

- $\mathbb{O}(d)$ – the group of linear isometries $U : \mathbb{R}^d \to \mathbb{R}^d$.
- The Hilbert space $\mathcal{F}$ is symmetric if

$$\forall f \in \mathcal{F}, U \in \mathbb{O}(d), \quad \|f\|_{\mathcal{F}} = \|f \circ U\|_{\mathcal{F}}$$

- By "averaging over $\mathbb{O}(d)$", we introduce a symmetrization operator

$$\mathcal{F} \mapsto \mathrm{Sym}(\mathcal{F})$$

- We show that:
  - $\mathrm{Sym}(\mathcal{F})$ is as good as $\mathcal{F}$.
  - $\mathrm{Sym}(\mathcal{F})$ "almost coincides" with the space of polynomials of degree $\log(1/\gamma)$.
- Analysed using Functional analysis, Representation theory and Harmonic analysis over the sphere.

Outline

1 Our Technique

2 Summery and open questions

Summery and open questions

- New technique to lower bound generalized linear methods.

Summery and open questions

- New technique to lower bound generalized linear methods.
  - Upper bounds for other algorithms?

Summery and open questions

- New technique to lower bound generalized linear methods.
  - Upper bounds for other algorithms?
  - Lower bounds for general algorithms?

Summery and open questions

- New technique to lower bound generalized linear methods.
  - Upper bounds for other algorithms?
  - Lower bounds for general algorithms?
  - **Advertisement:** A new hope for such results (D, Linial, Shalev-Shwartz 2013, 2014).

Summery and open questions

- New technique to lower bound generalized linear methods.
  - Upper bounds for other algorithms?
  - Lower bounds for general algorithms?
  - **Advertisement:** A new hope for such results (D, Linial, Shalev-Shwartz 2013, 2014).
- Lower bounds for other problems?

Summery and open questions

- New technique to lower bound generalized linear methods.
  - Upper bounds for other algorithms?
  - Lower bounds for general algorithms?
  - **Advertisement:** A new hope for such results (D, Linial, Shalev-Shwartz 2013, 2014).
- Lower bounds for other problems?
- Lower bounds for more algorithms? Neural nets? Boosting?

Summery and open questions

- New technique to lower bound generalized linear methods.
  - Upper bounds for other algorithms?
  - Lower bounds for general algorithms?
  - **Advertisement:** A new hope for such results (D, Linial, Shalev-Shwartz 2013, 2014).
- Lower bounds for other problems?
- Lower bounds for more algorithms? Neural nets? Boosting?