Nearly Tight Bounds for Robust Proper Learning of Halfspaces with a Margin

Ilias Diakonikolas
UW Madison

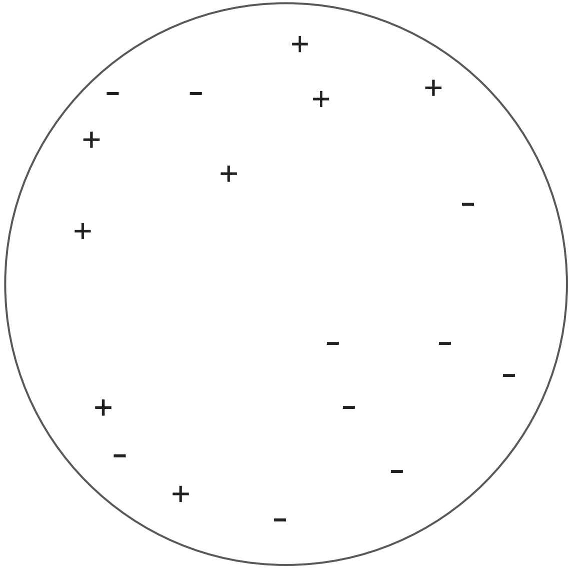
Daniel M. Kane
UC San Diego

Pasin Manurangsi
Google

Agnostic Proper Learning of Halfspaces

Input

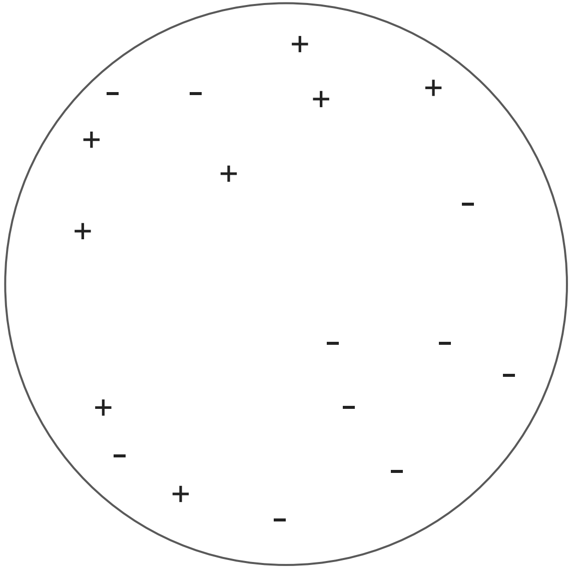
- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$



Agnostic Proper Learning of Halfspaces

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε



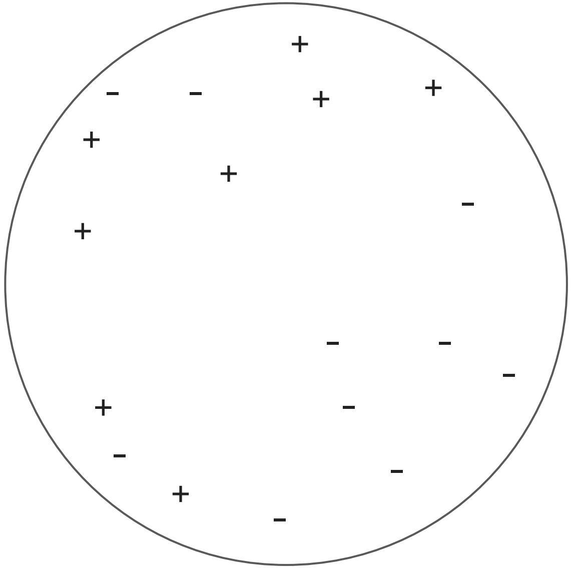
Agnostic Proper Learning of Halfspaces

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with “small” classification error



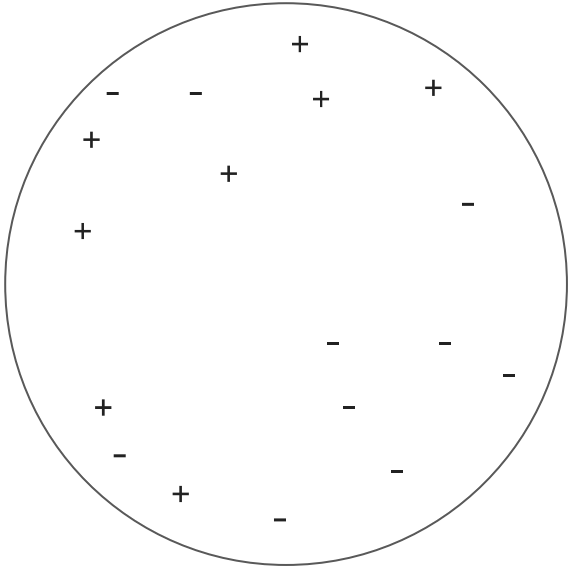
Agnostic Proper Learning of Halfspaces

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error



OPT = Min classification error among all halfspaces

$= \min_w \Pr_{(x, y) \sim \mathcal{D}} [\langle w, x \rangle \cdot y < 0]$

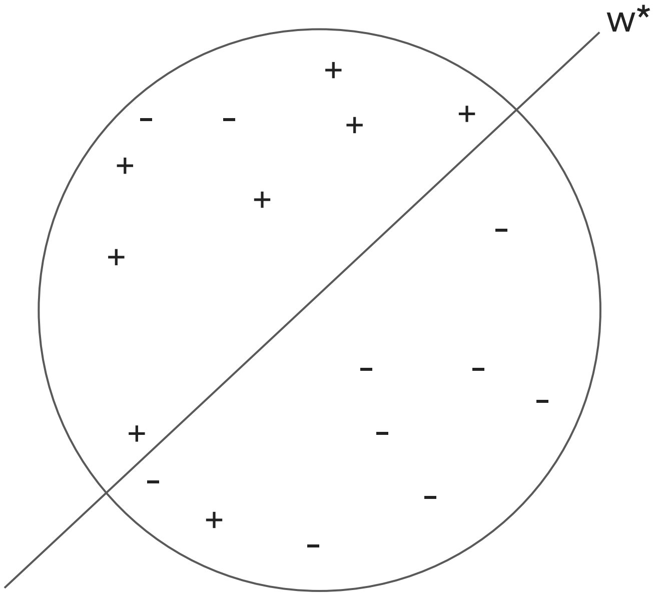
Agnostic Proper Learning of Halfspaces

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error



OPT = Min classification error among all halfspaces

$= \min_w \Pr_{(x, y) \sim \mathcal{D}} [\langle w, x \rangle \cdot y < 0]$

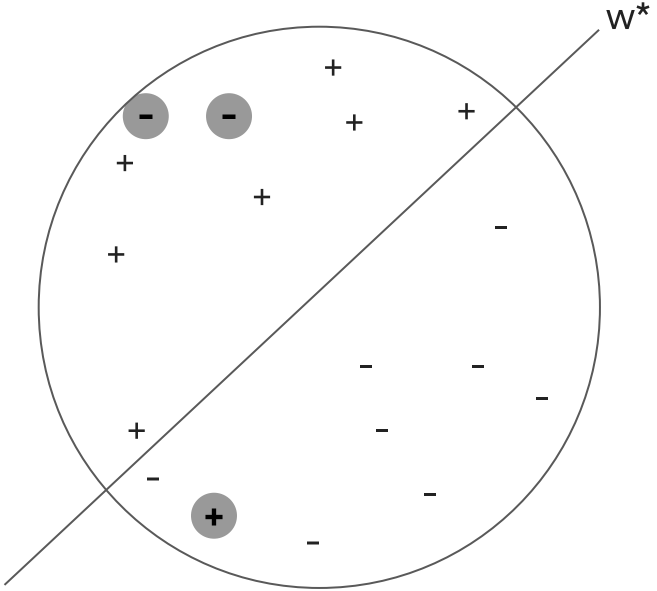
Agnostic Proper Learning of Halfspaces

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error



OPT = Min classification error among all halfspaces

$= \min_w \Pr_{(x, y) \sim \mathcal{D}} [\langle w, x \rangle \cdot y < 0]$

Agnostic Proper Learning of Halfspaces

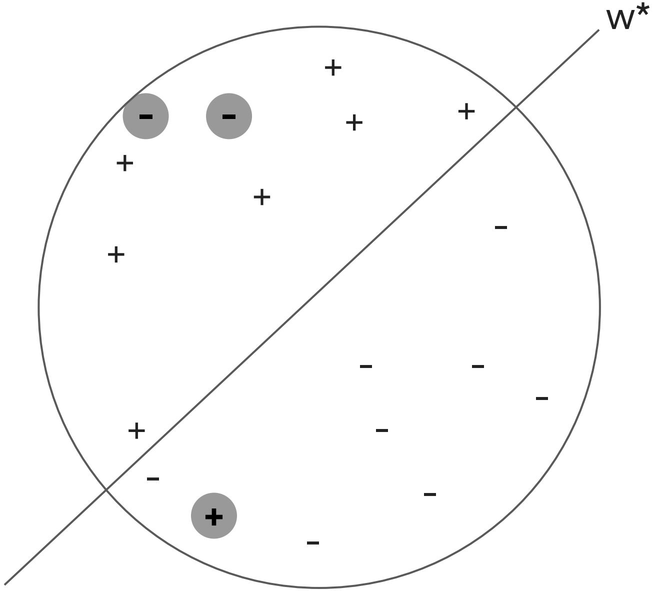
Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with “small” classification error

An algorithm is a α-learner if it outputs w with classification error at most $\alpha \cdot$ OPT + ε



OPT = Min classification error among all halfspaces

$= \min_w \Pr_{(x, y)\sim\mathcal{D}} [\langle w, x\rangle \cdot y < 0]$

Agnostic Proper Learning of Halfspaces

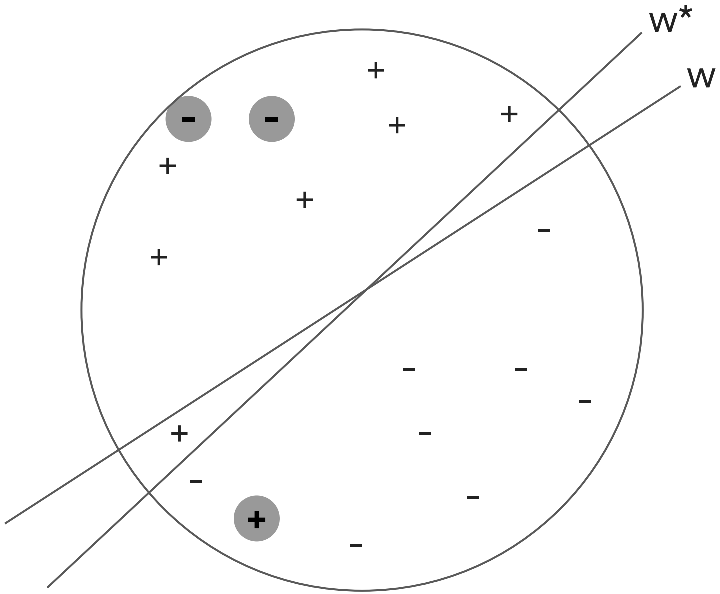
Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error

An algorithm is a α-learner if it outputs w with classification error at most $\alpha \cdot \mathrm{OPT} + \varepsilon$



OPT = Min classification error among all halfspaces

$= \min_w \Pr_{(x, y) \sim \mathcal{D}} [\langle w, x \rangle \cdot y < 0]$

Agnostic Proper Learning of Halfspaces

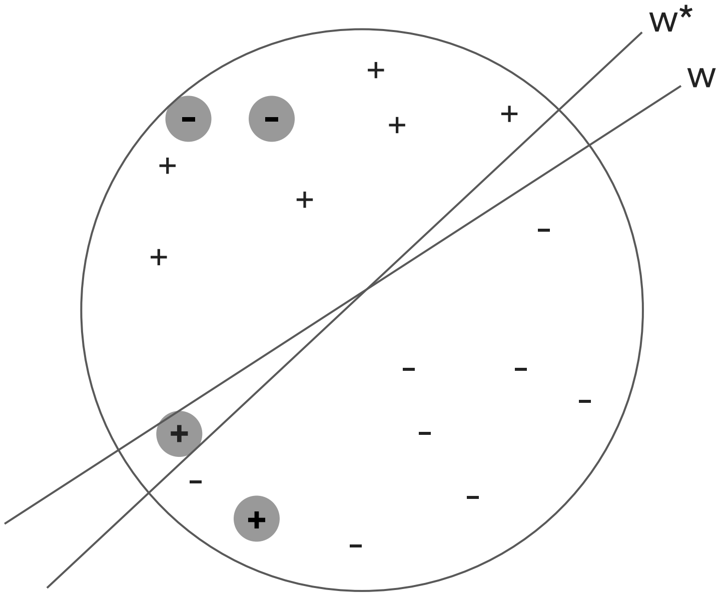
Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with “small” classification error

An algorithm is a α-learner if it outputs w with classification error at most $\alpha \cdot \mathrm{OPT} + \varepsilon$



OPT = Min classification error among all halfspaces

$= \min_w \Pr_{(x, y) \sim \mathcal{D}} [\langle w, x \rangle \cdot y < 0]$

Agnostic Proper Learning of Halfspaces

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

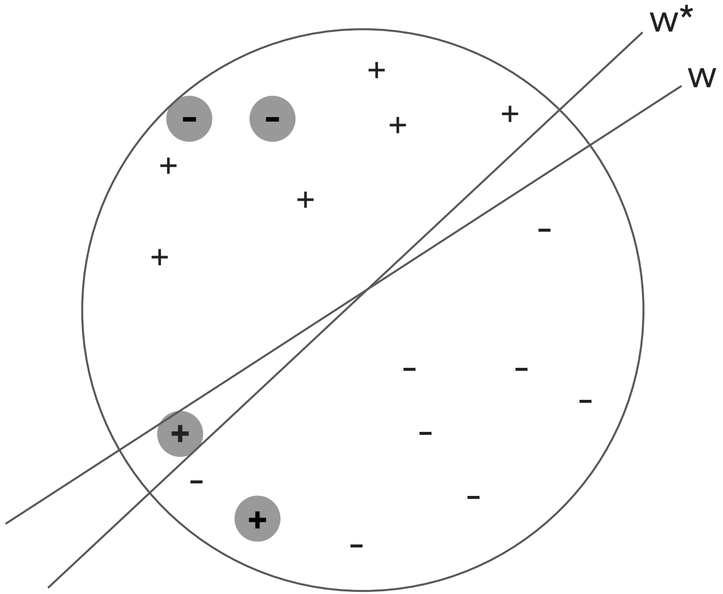
A halfspace w with “small” classification error

An algorithm is a α-learner if it outputs w with classification error at most $\alpha \cdot \mathrm{OPT} + \varepsilon$

Bad news:

[Arora et al.’97] Unless NP = RP, no poly-time α-learner for all constants α.

[Guruswami-Raghavendra’ 06, Feldman et al.’06] Even weak learning is NP-hard.



$$\mathrm{OPT} = \text{Min classification error among all halfspaces}$$

$$= \min_w \Pr_{(x,y)\sim\mathcal{D}} [\langle w, x\rangle \cdot y < 0]$$

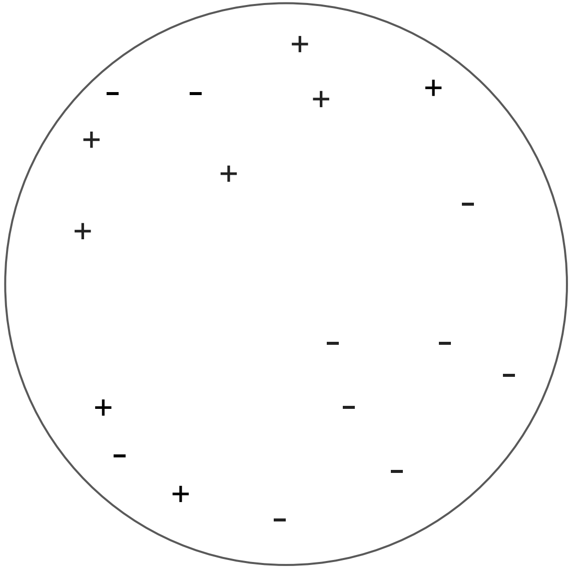
Agnostic Proper Learning of Halfspaces *with a Margin*

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error



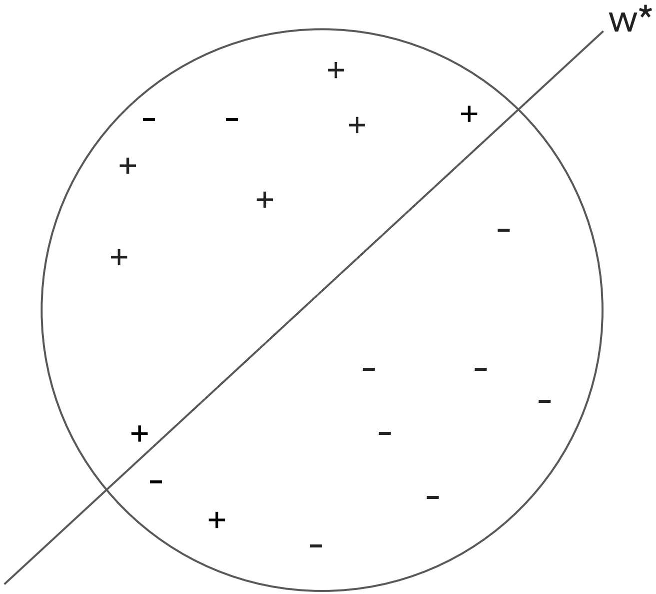
Agnostic Proper Learning of Halfspaces *with a Margin*

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with “small” classification error



OPT_γ = Min *γ-margin error* among all halfspaces

$= \min_w \Pr_{(x, y)\sim\mathcal{D}} [\langle w, x\rangle \cdot y < \gamma]$

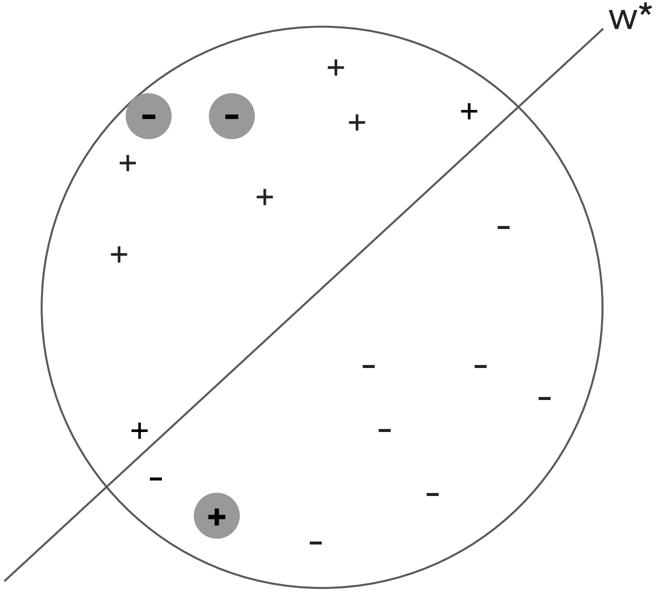
Agnostic Proper Learning of Halfspaces *with a Margin*

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error



OPT_γ = Min *γ-margin error* among all halfspaces

$= \min_w \Pr_{(x, y) \sim \mathcal{D}} [\langle w, x \rangle \cdot y < \gamma]$

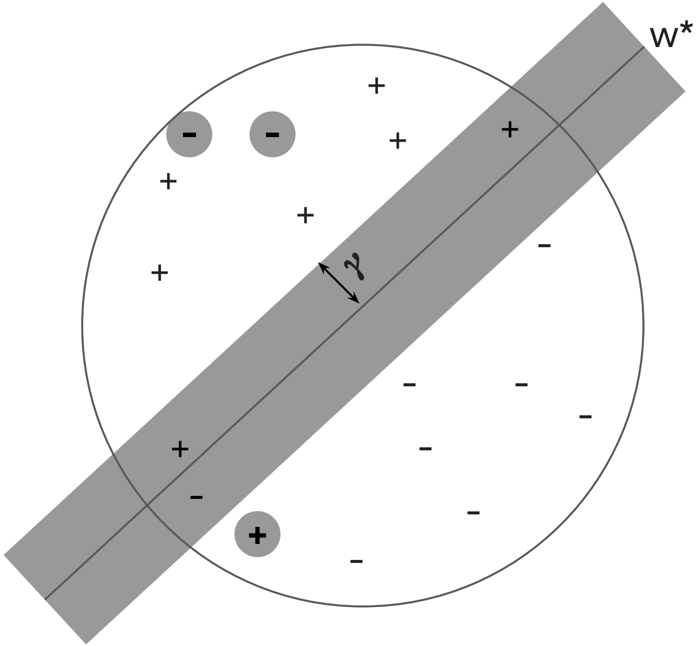
Agnostic Proper Learning of Halfspaces *with a Margin*

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error



OPT_γ = Min *γ-margin error* among all halfspaces

$= \min_w \Pr_{(x, y)\sim\mathcal{D}} [\langle w, x\rangle \cdot y < \gamma]$

Agnostic Proper Learning of Halfspaces *with a Margin*

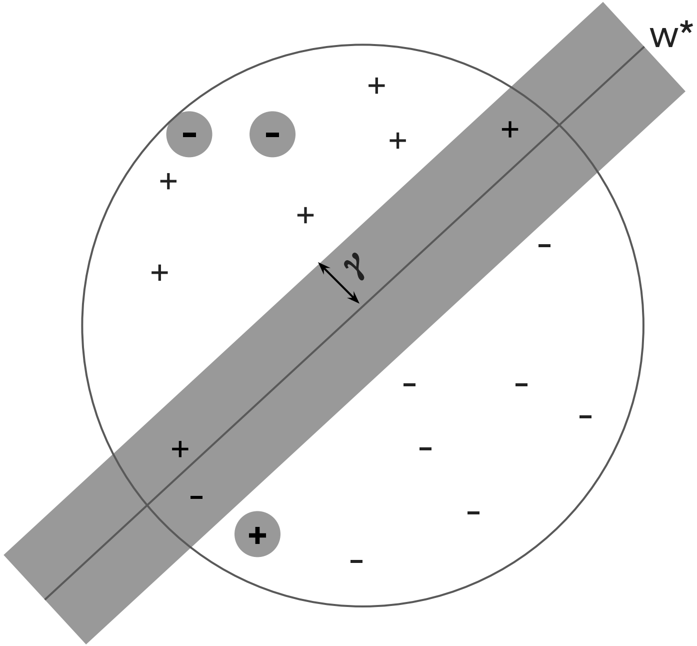
Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

Output

A halfspace w with "small" classification error

An algorithm is a α-learner if it outputs w with classification error at most $\alpha \cdot \mathrm{OPT}_\gamma + \varepsilon$



OPT_γ = Min *γ-margin error* among all halfspaces

$= \min_w \Pr_{(x, y)\sim\mathcal{D}} [\langle w, x\rangle \cdot y < \gamma]$

Agnostic Proper Learning of Halfspaces *with a Margin*

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

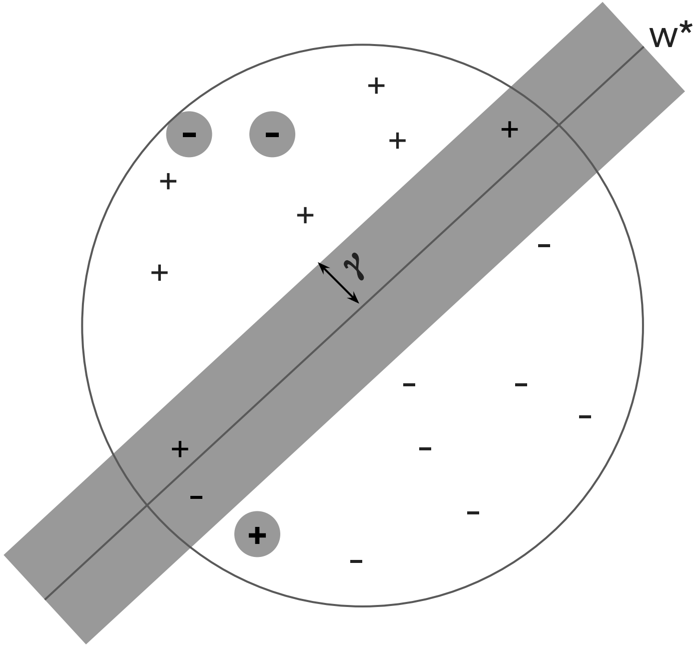
Output

A halfspace w with "small" classification error

An algorithm is a α-learner if it outputs w with classification error at most $\alpha \cdot \mathrm{OPT}_\gamma + \varepsilon$

Margin Assumption

- "Robustness" of the optimal halfspace to ℓ_2 noise



OPT_γ = Min *γ-margin error* among all halfspaces

$= \min_w \Pr_{(x, y)\sim\mathcal{D}} [\langle w, x\rangle \cdot y < \gamma]$

Agnostic Proper Learning of Halfspaces *with a Margin*

Input

- Labeled samples $(x_1, y_1), (x_2, y_2), \ldots \in \mathcal{B}(d) \times \{\pm 1\}$ from distribution $\mathcal{D}$
- Positive real number ε

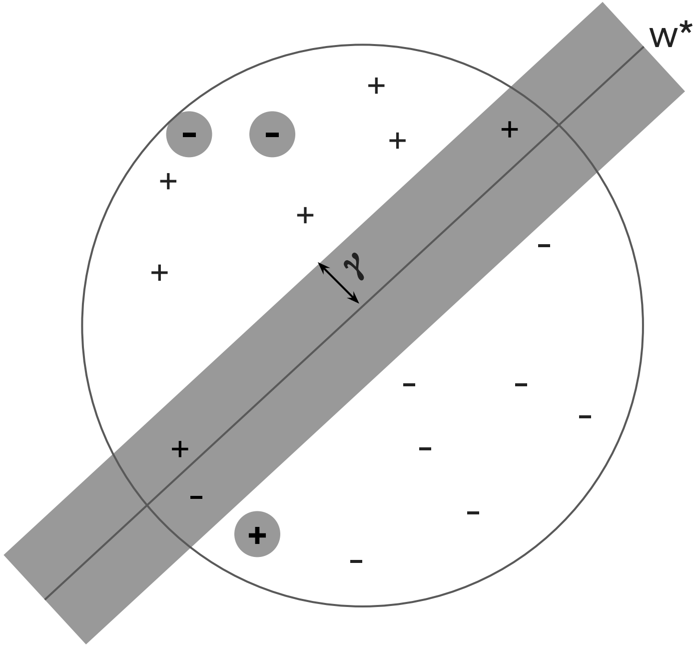
Output

A halfspace w with "small" classification error

An algorithm is a α-learner if it outputs w with classification error at most $\alpha \cdot \mathrm{OPT}_\gamma + \varepsilon$

Margin Assumption

- "Robustness" of the optimal halfspace to ℓ_2 noise
- Variants used in Perceptron, SVMs



OPT_γ = Min *γ-margin error* among all halfspaces

$= \min_w \Pr_{(x, y)\sim\mathcal{D}} [\langle w, x\rangle \cdot y < \gamma]$

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d)·exp($\tilde{O}(\log(1/\varepsilon)/\gamma^2)$) time,
takes $O(1/\varepsilon^2\gamma^2)$ samples

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that
runs in poly(d/ε)·exp($\tilde{O}(1/\gamma)$) time,
takes poly(d/ε)·exp($\tilde{O}(1/\gamma)$) samples

Our Results

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d) · exp($\tilde{O}$(log(1/ε)/γ^2)) time, takes O(1/$\varepsilon^2\gamma^2$) samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that runs in poly(d/ε) · exp($\tilde{O}$(1/γ)) time, takes poly(d/ε) · exp($\tilde{O}$(1/γ)) samples

Our Results

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d) $\cdot$ $\exp(\tilde{O}(\log(1/\varepsilon)/\gamma^2))$ time, takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that runs in poly(d/ε) $\cdot$ $\exp(\tilde{O}(1/\gamma))$ time, takes poly(d/ε) $\cdot$ $\exp(\tilde{O}(1/\gamma))$ samples

Our Results

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d)·exp($\tilde{O}$(log(1/ε)/γ^2)) time, takes O(1/$\varepsilon^2\gamma^2$) samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that runs in poly(d/ε)·exp($\tilde{O}$(1/γ)) time takes poly(d/ε)·exp($\tilde{O}$(1/γ)) samples

Our Results

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: α = 1

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d) · exp($\tilde{O}$(log(1/ε)/γ^2)) time, takes O(1/$\varepsilon^2\gamma^2$) samples

Approximation ratio: α = 1

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that runs in poly(d/ε) · exp($\tilde{O}$(1/γ)) time takes poly(d/ε) · exp($\tilde{O}$(1/γ)) samples

Output hypothesis is not a halfspace

Our Results

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: α = 1

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d) · exp($\tilde{O}$(log(1/ε)/γ^2)) time, takes O(1/$\varepsilon^2\gamma^2$) samples

Approximation ratio: α = 1

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that runs in poly(d/ε) · exp($\tilde{O}$(1/γ)) time takes poly(d/ε) · exp($\tilde{O}$(1/γ)) samples

Output hypothesis is not a halfspace

Our Results

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d)·exp($\tilde{O}(\log(1/\varepsilon)/\gamma^2)$) time,
takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that
runs in poly(d/ε)·exp($\tilde{O}(1/\gamma)$) time
takes poly(d/ε)·exp($\tilde{O}(1/\gamma)$) samples

Output hypothesis is not a halfspace

Our Results

Theorem 1 *proper* 1.01-learner that
runs in poly(d/ε)·exp($\tilde{O}(1/\gamma^2)$)-time
takes $O(1/\varepsilon^2\gamma^2)$ samples

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d)$\cdot$exp($\tilde{O}(\log(1/\varepsilon)/\gamma^2)$) time, takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that
runs in poly(d/ε)$\cdot$exp($\tilde{O}(1/\gamma)$) time
takes poly(d/ε)$\cdot$exp($\tilde{O}(1/\gamma)$) samples

Output hypothesis is not a halfspace

Our Results

Approximation ratio: any $\alpha > 1$

Theorem 1 *proper* 1.01-learner that runs in poly(d/ε)$\cdot$exp($\tilde{O}(1/\gamma^2)$)-time takes $O(1/\varepsilon^2\gamma^2)$ samples

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in $\mathrm{poly}(d) \cdot \exp(\tilde{O}(\log(1/\varepsilon)/\gamma^2))$ time, takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ time takes $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ samples

Output hypothesis is not a halfspace

Our Results

Approximation ratio: any $\alpha > 1$

Theorem 1 *proper* 1.01-learner that runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma^2))$-time takes $O(1/\varepsilon^2\gamma^2)$ samples

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d) $\cdot$ $\exp(\tilde{O}(\log(1/\varepsilon)/\gamma^2))$ time, takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that runs in poly(d/ε) $\cdot$ $\exp(\tilde{O}(1/\gamma))$ time takes poly(d/ε) $\cdot$ $\exp(\tilde{O}(1/\gamma))$ samples

Output hypothesis is not a halfspace

Our Results

Approximation ratio: any $\alpha > 1$

Theorem 1 *proper* 1.01-learner that runs in poly(d/ε) $\cdot$ $\exp(\tilde{O}(1/\gamma^2))$-time takes $O(1/\varepsilon^2\gamma^2)$ samples

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in poly(d)·$\exp(\tilde{O}(\log(1/\varepsilon)/\gamma^2))$ time,
takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that
runs in poly(d/ε)·$\exp(\tilde{O}(1/\gamma))$ time
takes poly(d/ε)·$\exp(\tilde{O}(1/\gamma))$ samples

Output hypothesis is not a halfspace

Our Results

Approximation ratio: any $\alpha > 1$

Theorem 1 *proper* 1.01-learner that
runs in poly(d/ε)·$\exp(\tilde{O}(1/\gamma^2))$-time
takes $O(1/\varepsilon^2\gamma^2)$ samples

Theorem 2 Assuming *Exponential Time Hypothesis*,
for any constant $\alpha > 1$, no *proper* α-learner
runs in poly(d/ε)·$\exp(O(1/\gamma^{2-o(1)}))$ time

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in $\mathrm{poly}(d) \cdot \exp(\tilde{O}(\log(1/\varepsilon)/\gamma^2))$ time,
takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ time
takes $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ samples

Output hypothesis is not a halfspace

Our Results

Approximation ratio: any $\alpha > 1$

Theorem 1 *proper* 1.01-learner that
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma^2))$-time
takes $O(1/\varepsilon^2\gamma^2)$ samples

Theorem 2 Assuming *Exponential Time Hypothesis*,
for any constant $\alpha > 1$, no *proper* α-learner
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(O(1/\gamma^{2-o(1)}))$ time

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in $\mathrm{poly}(d) \cdot \exp(\tilde{O}(\log(1/\varepsilon)/\gamma^2))$ time,
takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ time
takes $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ samples

Output hypothesis is not a halfspace

Our Results

Approximation ratio: any $\alpha > 1$

Theorem 1 *proper* 1.01-learner that
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma^2))$-time
takes $O(1/\varepsilon^2\gamma^2)$ samples

Theorem 2 Assuming *Exponential Time Hypothesis*,
for any constant $\alpha > 1$, no *proper* α-learner
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(O(1/\gamma^{2-o(1)}))$ time

Approximation ratio: $\alpha = 1$

Theorem 3 Assuming W[1] ≠ FPT,
for any function f, no *proper* 1-learner
runs in $\mathrm{poly}(d/\varepsilon) \cdot f(1/\gamma)$ time

Agnostic Proper Learning of Halfspaces *with a Margin*

Previous Works

Approximation ratio: $\alpha = 1$

[Ben-David & Simon'00]
proper 1-learner that runs in $\mathrm{poly}(d) \cdot \exp(\tilde{O}(\log(1/\varepsilon)/\gamma^2))$ time,
takes $O(1/\varepsilon^2\gamma^2)$ samples

Approximation ratio: $\alpha = 1$

[Shalev-Shwartz, Shamir & Sridharan'09]
improper 1-learner that
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ time
takes $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma))$ samples

Output hypothesis is not a halfspace

Our Results

Approximation ratio: any $\alpha > 1$

Theorem 1 *proper* 1.01-learner that
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(\tilde{O}(1/\gamma^2))$-time
takes $O(1/\varepsilon^2\gamma^2)$ samples

Theorem 2 Assuming *Exponential Time Hypothesis*,
for any constant $\alpha > 1$, no *proper* α-learner
runs in $\mathrm{poly}(d/\varepsilon) \cdot \exp(O(1/\gamma^{2-o(1)}))$ time

Approximation ratio: $\alpha = 1$

Theorem 3 Assuming W[1] ≠ FPT,
for any function f, no *proper* 1-learner
runs in $\mathrm{poly}(d/\varepsilon) \cdot f(1/\gamma)$ time

Also results for large approximation ratio α