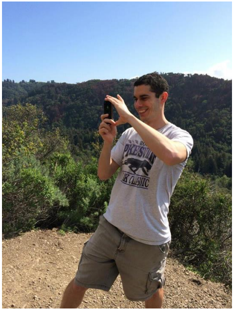
Distribution-Independent Reliable Learning

Varun Kanade   Justin Thaler

UC Berkeley   Yahoo Labs



June 15, 2014

Outline

1. Introduction
2. Framework
   - Agnostic Learning Framework
   - Positive Reliable Learning
   - Fully Reliable Learning
3. Main Results
   - Polynomial Approximations
   - Learning Results
   - One-sided Approximations
4. Conclusion

Some Learning Scenarios

SPAM Classification

- Lots of SPAM messages—annoying to deal with unimportant emails in Inbox
- Very costly if an important mail gets marked as spam
- False positives much worse than false negatives

Detecting Network Failures

- Failure to detect very costly; incorrect detection relatively small cost
- False negative errors very harmful

Medical Diagnosis

- All kinds of errors are bad!
- Want to have (almost) no errors, at the cost of sometimes predicting "*don't know*"

We call these reliable learning problems!

Some Learning Scenarios

SPAM Classification

- Lots of SPAM messages—annoying to deal with unimportant emails in Inbox
- Very costly if an important mail gets marked as spam
- False positives much worse than false negatives

Detecting Network Failures

- Failure to detect very costly; incorrect detection relatively small cost
- False negative errors very harmful

Medical Diagnosis

- All kinds of errors are bad!
- Want to have (almost) no errors, at the cost of sometimes predicting "*don't know*"

We call these reliable learning problems!

Some Learning Scenarios

SPAM Classification

- Lots of SPAM messages—annoying to deal with unimportant emails in Inbox
- Very costly if an important mail gets marked as spam
- False positives much worse than false negatives

Detecting Network Failures

- Failure to detect very costly; incorrect detection relatively small cost
- False negative errors very harmful

Medical Diagnosis

- All kinds of errors are bad!
- Want to have (almost) no errors, at the cost of sometimes predicting "*don't know*"

We call these reliable learning problems!

Prior Work

- Minimize asymmetric loss function:

$$\min_{f \in F} \mathrm{false}_-(f) + 1000\mathrm{false}_+(f)$$

- Classical Statistics: Neyman-Pearson Lemma
  - Framed in language of hypothesis testing
- Lots of other work: cautious classifiers, abstaining classifiers
  [Domingos '99], [Elkan '01], [Bartlett, Wegkamp '08], [El-Yaniv, Wiener '10]
- Question: What is the computational complexity for these problems?

Outline

1. Introduction
2. Framework
   - Agnostic Learning Framework
   - Positive Reliable Learning
   - Fully Reliable Learning
3. Main Results
   - Polynomial Approximations
   - Learning Results
   - One-sided Approximations
4. Conclusion

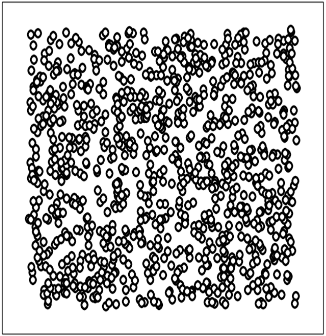
PAC Learning Framework [Valiant 1984]

- $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from D over $\{-1, 1\}^n$
- Labels $y_i = f(\mathbf{x}_i)$ for some f in class F, *e.g.* linear separators, DNF
- Goal: Find hypothesis: $h : \{-1, 1\}^n \to \{-1, 1\}$, s.t.

$$\mathrm{err}(h) = \Pr_{\mathbf{x} \sim D}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$$

- Want learning algorithm to succeed for all distributions D

PAC Learning Framework [Valiant 1984]

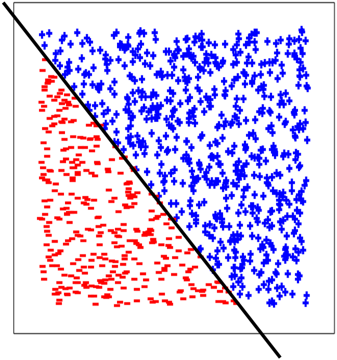


- $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from D over $\{-1, 1\}^n$
- Labels $y_i = f(\mathbf{x}_i)$ for some f in class F, *e.g.* linear separators, DNF
- Goal: Find hypothesis: $h : \{-1, 1\}^n \to \{-1, 1\}$, s.t.

$$\mathrm{err}(h) = \Pr_{\mathbf{x} \sim D}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$$

- Want learning algorithm to succeed for all distributions D

PAC Learning Framework [Valiant 1984]

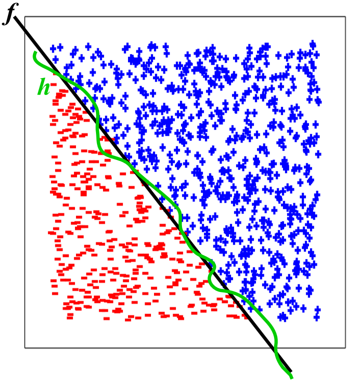


- $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from D over $\{-1,1\}^n$
- Labels $y_i = f(\mathbf{x}_i)$ for some f in class F, *e.g.* linear separators, DNF
- Goal: Find hypothesis: $h : \{-1,1\}^n \to \{-1,1\}$, s.t.

$$\mathrm{err}(h) = \Pr_{\mathbf{x}\sim D}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$$

- Want learning algorithm to succeed for all distributions D

PAC Learning Framework [Valiant 1984]

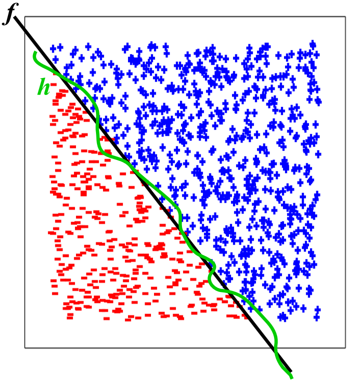


- $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from D over $\{-1, 1\}^n$
- Labels $y_i = f(\mathbf{x}_i)$ for some f in class F, *e.g.* linear separators, DNF
- Goal: Find hypothesis: $h : \{-1, 1\}^n \to \{-1, 1\}$, s.t.

$$\mathrm{err}(h) = \Pr_{\mathbf{x} \sim D}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$$

- Want learning algorithm to succeed for all distributions D

PAC Learning Framework [Valiant 1984]



- $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from D over $\{-1, 1\}^n$
- Labels $y_i = f(\mathbf{x}_i)$ for some f in class F, *e.g.* linear separators, DNF
- Goal: Find hypothesis: $h : \{-1, 1\}^n \rightarrow \{-1, 1\}$, s.t.

$$\mathrm{err}(h) = \Pr_{\mathbf{x} \sim D}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$$

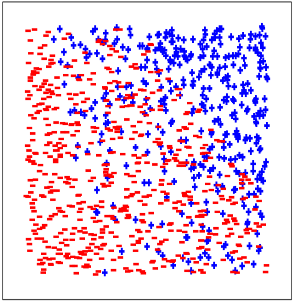
- Want learning algorithm to succeed for all distributions D

Agnostic Learning [Haussler '92, Kearns, Schapire, Sellie '94]

- Generalization of Valiant's PAC framework
- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, (say linear separators), find $h : \{-1,1\}^n \to \{-1,1\}$ such that

$$\mathrm{err}(h) \leq \min_{f \in F} \mathrm{err}(f) + \epsilon$$

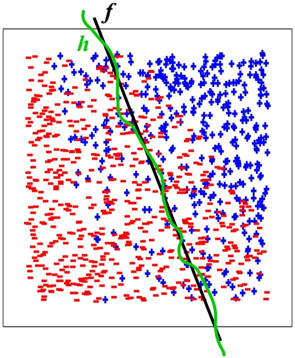
Agnostic Learning [Haussler '92, Kearns, Schapire, Sellie '94]



- Generalization of Valiant's PAC framework
- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, (say linear separators), find $h : \{-1,1\}^n \to \{-1,1\}$ such that

$$\mathrm{err}(h) \leq \min_{f \in F} \mathrm{err}(f) + \epsilon$$

Agnostic Learning [Haussler '92, Kearns, Schapire, Sellie '94]



- Generalization of Valiant's PAC framework
- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1, 1\}^n \times \{-1, 1\}$
- Goal: For some class F, (say linear separators), find $h : \{-1, 1\}^n \to \{-1, 1\}$ such that

$$\mathrm{err}(h) \leq \min_{f \in F} \mathrm{err}(f) + \epsilon$$

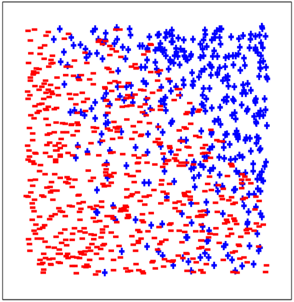
Positive Reliable Learning [Kalai, K., Mansour '09]

- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, find $h : \{-1,1\}^n \to \{-1,1\}$ such that:

$$\mathrm{false}_+(h) \le \epsilon$$
$$\mathrm{false}_-(h) \le \min_{f \in F^+} \mathrm{false}_-(f) + \epsilon,$$

where F^+ denotes the classifiers in F for which $\mathrm{false}_+(f) = 0$

Positive Reliable Learning [Kalai, K., Mansour '09]

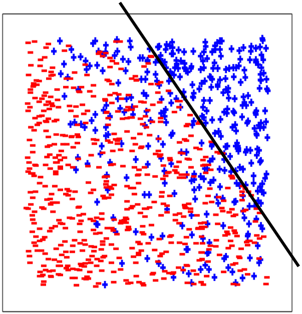


- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, find $h : \{-1,1\}^n \to \{-1,1\}$ such that:

$$\mathrm{false}_+(h) \le \epsilon$$
$$\mathrm{false}_-(h) \le \min_{f \in F^+} \mathrm{false}_-(f) + \epsilon,$$

where F^+ denotes the classifiers in F for which $\mathrm{false}_+(f) = 0$

Positive Reliable Learning [Kalai, K., Mansour '09]

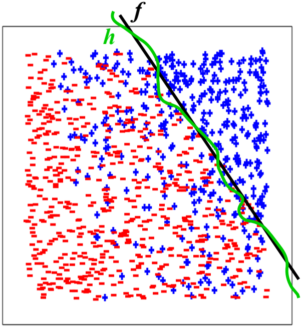


- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, find $h : \{-1,1\}^n \to \{-1,1\}$ such that:

$$\mathrm{false}_+(h) \leq \epsilon$$
$$\mathrm{false}_-(h) \leq \min_{f \in F^+} \mathrm{false}_-(f) + \epsilon,$$

where F^+ denotes the classifiers in F for which $\mathrm{false}_+(f) = 0$

Positive Reliable Learning [Kalai, K., Mansour '09]

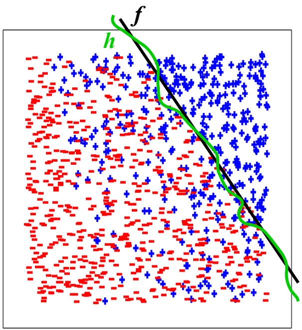


- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, find $h : \{-1,1\}^n \to \{-1,1\}$ such that:

$$\mathrm{false}_+(h) \leq \epsilon$$
$$\mathrm{false}_-(h) \leq \min_{f \in F^+} \mathrm{false}_-(f) + \epsilon,$$

where F^+ denotes the classifiers in F for which $\mathrm{false}_+(f) = 0$

Positive Reliable Learning [Kalai, K., Mansour '09]



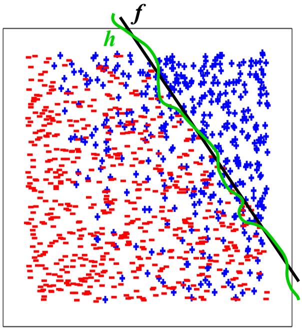
- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, find $h : \{-1,1\}^n \to \{-1,1\}$ such that:

$$\mathrm{false}_+(h) \leq \epsilon$$
$$\mathrm{false}_-(h) \leq \min_{f \in F^+} \mathrm{false}_-(f) + \epsilon,$$

where F^+ denotes the classifiers in F for which $\mathrm{false}_+(f) = 0$

Models situations, such as SPAM classification, where false positives are very harmful

Positive Reliable Learning [Kalai, K., Mansour '09]



- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, find $h : \{-1,1\}^n \to \{-1,1\}$ such that:

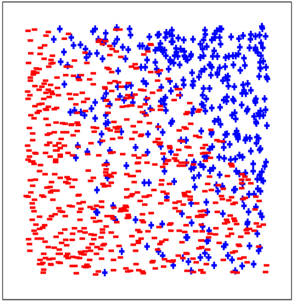
$$\mathrm{false}_+(h) \leq \epsilon$$
$$\mathrm{false}_-(h) \leq \min_{f \in F^+} \mathrm{false}_-(f) + \epsilon,$$

where F^+ denotes the classifiers in F for which $\mathrm{false}_+(f) = 0$

Models situations, such as SPAM classification, where false positives are very harmful

Negative Reliable Learning is defined analogously

Fully Reliable Learning [Kalai, K., Mansour 2009]

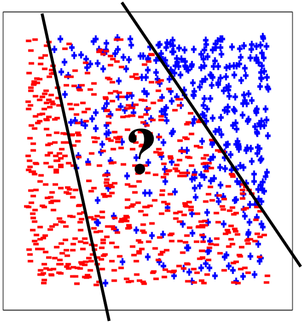


- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1, 1\}^n \times \{-1, 1\}$
- Goal: For some class F, find $h : \{-1, 1\}^n \to \{-1, 1, ?\}$ such that:

$$\mathrm{err}(h) \leq \epsilon$$

$$\Pr[h(\mathbf{x}) = ?] \leq \mathrm{opt} + \epsilon$$

Fully Reliable Learning [Kalai, K., Mansour 2009]



- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1,1\}^n \times \{-1,1\}$
- Goal: For some class F, find $h : \{-1,1\}^n \to \{-1,1,?\}$ such that:

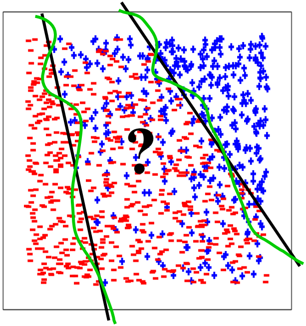
$$\mathrm{err}(h) \leq \epsilon$$

$$\Pr[h(\mathbf{x}) = ?] \leq \mathrm{opt} + \epsilon$$

For each (f^+, f^-) in class F, define $g : \{-1,1\}^n \to \{-1,+1,?\}$, as $g(\mathbf{x}) = f^+(\mathbf{x})$, if $f^+(\mathbf{x}) = f^-(\mathbf{x})$, and $g(\mathbf{x}) = ?$ otherwise

$$\mathrm{opt} = \min_{\substack{g, s.t.\\ \mathrm{err}(g)=0}} \Pr_{\mathbf{x}\sim D}[g(\mathbf{x}) = ?]$$

Fully Reliable Learning [Kalai, K., Mansour 2009]



- Like in the agnostic setting: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ from D over $\{-1, 1\}^n \times \{-1, 1\}$
- Goal: For some class F, find $h : \{-1, 1\}^n \to \{-1, 1, ?\}$ such that:

$$\mathrm{err}(h) \leq \epsilon$$

$$\Pr[h(\mathbf{x}) = ?] \leq \mathrm{opt} + \epsilon$$

For each (f^+, f^-) in class F, define $g : \{-1, 1\}^n \to \{-1, +1, ?\}$, as $g(\mathbf{x}) = f^+(\mathbf{x})$, if $f^+(\mathbf{x}) = f^-(\mathbf{x})$, and $g(\mathbf{x}) = ?$ otherwise

$$\mathrm{opt} = \min_{\substack{g, s.t. \\ \mathrm{err}(g)=0}} \Pr_{\mathbf{x} \sim D}[g(\mathbf{x}) = ?]$$

Models situations such as medical diagnosis, where abstaining is preferred to making errors

Prior Results

Theorem [Kalai, K., Mansour 2009]

If F is agnostically learnable, then F is positive and negative reliably learnable. In fact, disjunctions of functions in F are positive reliably learnable.

Theorem [Kalai, K., Mansour 2009]

If F is positive and negative reliably learnable, then F is fully reliably learnable.

- Reliable learning no harder than agnostic learning
- Some evidence that positive/negative realiable learning easier than agnostic learning
- Is fully reliable learning strictly easier than agnostic learning?

Prior Results

Theorem [Kalai, K., Mansour 2009]

If F is agnostically learnable, then F is positive and negative reliably learnable. In fact, disjunctions of functions in F are positive reliably learnable.

Theorem [Kalai, K., Mansour 2009]

If F is positive and negative reliably learnable, then F is fully reliably learnable.

- Reliable learning no harder than agnostic learning
- Some evidence that positive/negative realiable learning easier than agnostic learning
- Is fully reliable learning strictly easier than agnostic learning?

Prior Results

Theorem [Kalai, K., Mansour 2009]

If F is agnostically learnable, then F is positive and negative reliably learnable. In fact, disjunctions of functions in F are positive reliably learnable.

Theorem [Kalai, K., Mansour 2009]

If F is positive and negative reliably learnable, then F is fully reliably learnable.

- Reliable learning no harder than agnostic learning
- Some evidence that positive/negative realiable learning easier than agnostic learning
- Is fully reliable learning strictly easier than agnostic learning?

Outline

1 Introduction

2 Framework
- Agnostic Learning Framework
- Positive Reliable Learning
- Fully Reliable Learning

3 Main Results
- Polynomial Approximations
- Learning Results
- One-sided Approximations

4 Conclusion

General Approach

- ERM: Find a function $f \in F$ that minimizes appropriate zero-one loss
- PAC Learning: $\forall i, f(\mathbf{x}_i) = y_i$
- Agnostic Learning:

$$f^* = \text{argmin}_{f \in F} \sum_{i=1}^{m} \mathbb{I}(f(\mathbf{x}_i) \neq y_i)$$

- Positive Reliable Learning: Find f such that

$$\forall i, y_i = -1, f(\mathbf{x}_i) = -1,$$

and subject to above f minimizes

$$\sum_{i:y_i=+1} \mathbb{I}(f(\mathbf{x}_i) = -1)$$

General Approach

- ERM: Find a function $f \in F$ that minimizes appropriate zero-one loss
- PAC Learning: $\forall i, f(\mathbf{x}_i) = y_i$
- Agnostic Learning:

$$f^* = \mathrm{argmin}_{f \in F} \sum_{i=1}^{m} \mathbb{I}(f(\mathbf{x}_i) \neq y_i)$$

- Positive Reliable Learning: Find f such that

$$\forall i, y_i = -1, f(\mathbf{x}_i) = -1,$$

and subject to above f minimizes

$$\sum_{i: y_i = +1} \mathbb{I}(f(\mathbf{x}_i) = -1)$$

General Approach

- ERM: Find a function $f \in F$ that minimizes appropriate zero-one loss
- PAC Learning: $\forall i, f(\mathbf{x}_i) = y_i$
- Agnostic Learning:

$$f^* = \text{argmin}_{f \in F} \sum_{i=1}^{m} \mathbb{I}(f(\mathbf{x}_i) \neq y_i)$$

- Positive Reliable Learning: Find f such that

$$\forall i, y_i = -1, f(\mathbf{x}_i) = -1,$$

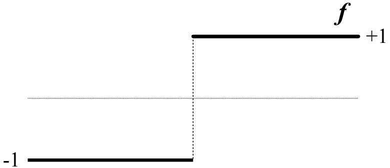
and subject to above f minimizes

$$\sum_{i : y_i = +1} \mathbb{I}(f(\mathbf{x}_i) = -1)$$

General Approach

- Problems are typically not convex, computationally hard
- Consider larger class H such that
  - For each $f \in F$, some $h \in H$ "approximates" f
  - Find h in H that empirically minimizes a suitable loss function
- (Various types of) polynomial approximations give suitable algorithms
- Focus on distribution-independent learning

Polynomial Threshold Approximations



- Want polynomial p such that

$$\text{sign}(p(\mathbf{x})) = f(\mathbf{x})$$

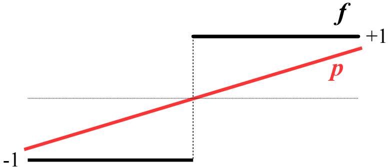
- Suffices for PAC learning
- Linear Programming: Find p s.t.

$$\forall i, p(x_i)y_i \geq 0$$

- Yields some of best known results

DNF learning in $2^{\tilde{O}(n^{1/3})}$ time
[Klivans, Servedio 2001]

Polynomial Threshold Approximations



- Want polynomial p such that

$$\text{sign}(p(\mathbf{x})) = f(\mathbf{x})$$

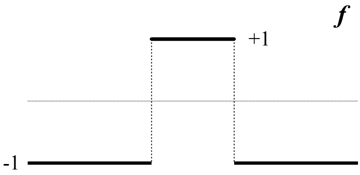
- Suffices for PAC learning
- Linear Programming: Find p s.t.

$$\forall i, p(x_i)y_i \geq 0$$

- Yields some of best known results

  DNF learning in $2^{\tilde{O}(n^{1/3})}$ time
  [Klivans, Servedio 2001]

Polynomial Threshold Approximations



- Want polynomial p such that

$$\mathrm{sign}(p(\mathbf{x})) = f(\mathbf{x})$$

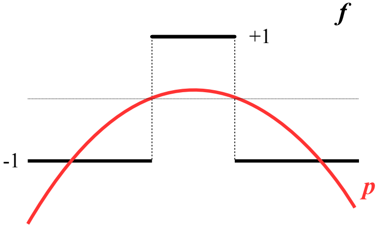
- Suffices for PAC learning
- Linear Programming: Find p s.t.

$$\forall i, p(x_i)y_i \geq 0$$

- Yields some of best known results

DNF learning in $2^{\tilde{O}(n^{1/3})}$ time
[Klivans, Servedio 2001]

Polynomial Threshold Approximations



- Want polynomial p such that

$$\mathrm{sign}(p(\mathbf{x})) = f(\mathbf{x})$$

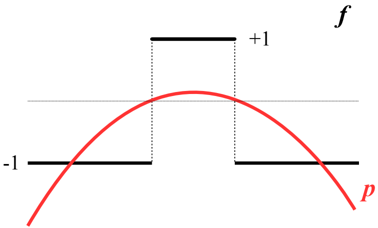
- Suffices for PAC learning
- Linear Programming: Find p s.t.

$$\forall i, p(x_i)y_i \geq 0$$

- Yields some of best known results

DNF learning in $2^{\tilde{O}(n^{1/3})}$ time
[Klivans, Servedio 2001]

Polynomial Threshold Approximations



- Want polynomial p such that

$$\operatorname{sign}(p(\mathbf{x})) = f(\mathbf{x})$$

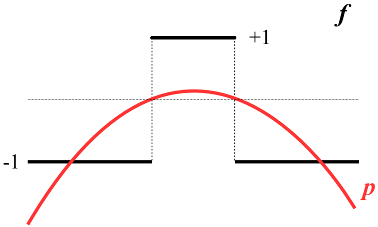
- Suffices for PAC learning
- Linear Programming: Find p s.t.

$$\forall i, p(x_i)y_i \geq 0$$

- Yields some of best known results

  DNF learning in $2^{\tilde{O}(n^{1/3})}$ time
  [Klivans, Servedio 2001]

Polynomial Threshold Approximations



- Want polynomial p such that

$$\operatorname{sign}(p(\mathbf{x})) = f(\mathbf{x})$$

- Suffices for PAC learning
- Linear Programming: Find p s.t.

$$\forall i, p(x_i)y_i \geq 0$$

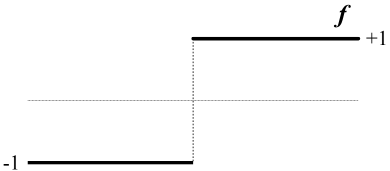
- Yields some of best known results

DNF learning in $2^{\tilde{O}(n^{1/3})}$ time
[Klivans, Servedio 2001]

Degree d approximations gives algorithms with running time $O(n^d)$

Sample complexity related to weight of approximating polynomial

Pointwise Approximations



- Want polynomial p such that

$$\forall \mathbf{x} \in \{-1, 1\}^n, |f(\mathbf{x}) - p(\mathbf{x})| \leq \epsilon$$

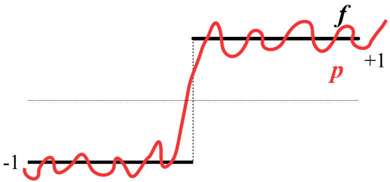
- Suffices (required?) for agnostic learning
- L1 Regression: Find p that minimizes

$$\sum_i |p(\mathbf{x}_i) - y_i|$$

[Kalai, Klivans, Mansour, Servedio 2005]

- Pointwise approximations typically requires much larger degree compared to threshold approximations

Pointwise Approximations



- Want polynomial p such that

$$\forall \mathbf{x} \in \{-1, 1\}^n, |f(\mathbf{x}) - p(\mathbf{x})| \leq \epsilon$$

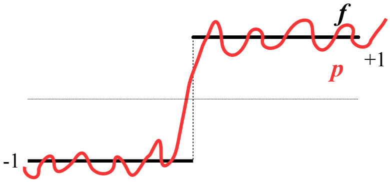
- Suffices (required?) for agnostic learning
- L1 Regression: Find p that minimizes

$$\sum_i |p(\mathbf{x}_i) - y_i|$$

[Kalai, Klivans, Mansour, Servedio 2005]

- Pointwise approximations typically requires much larger degree compared to threshold approximations

Pointwise Approximations

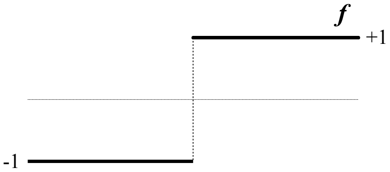


- Want polynomial p such that
$$\forall \mathbf{x} \in \{-1, 1\}^n, |f(\mathbf{x}) - p(\mathbf{x})| \leq \epsilon$$
- Suffices (required?) for agnostic learning
- L1 Regression: Find p that minimizes
$$\sum_i |p(\mathbf{x}_i) - y_i|$$
[Kalai, Klivans, Mansour, Servedio 2005]
- Pointwise approximations typically requires much larger degree compared to threshold approximations

Degree d approximations gives algorithms with running time $O(n^d)$.

Sample complexity related to weight of polynomial approximation

One-sided Approximations



- Want polynomial p such that

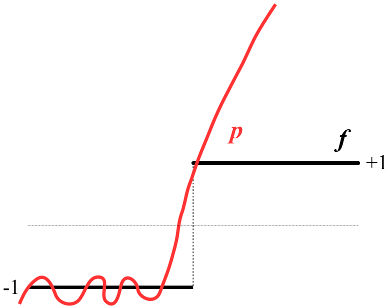
$$\forall \mathbf{x} \text{ s.t. } f(\mathbf{x}) = -1, |f(\mathbf{x}) - p(\mathbf{x})| \leq \epsilon$$

and

$$\forall \mathbf{x} \text{ s.t. } f(\mathbf{x}) = +1, p(\mathbf{x}) \geq 1 - \epsilon$$

- Call this positive one-sided polynomial approximation
- Theorem: Suffices for positive-reliable learning
- One-sided approximate degree can be much lower than approximate degree

One-sided Approximations



- Want polynomial p such that

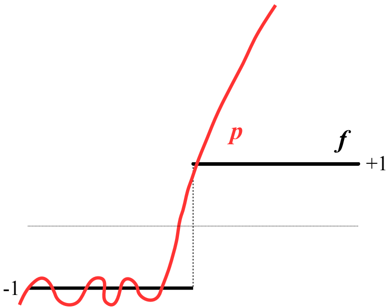
$$\forall \mathbf{x} \text{ s.t. } f(\mathbf{x}) = -1, |f(\mathbf{x}) - p(\mathbf{x})| \leq \epsilon$$

and

$$\forall \mathbf{x} \text{ s.t. } f(\mathbf{x}) = +1, p(\mathbf{x}) \geq 1 - \epsilon$$

- Call this positive one-sided polynomial approximation
- Theorem: Suffices for positive-reliable learning
- One-sided approximate degree can be much lower than approximate degree

One-sided Approximations



- Want polynomial p such that

$$\forall \mathbf{x} \text{ s.t. } f(\mathbf{x}) = -1, |f(\mathbf{x}) - p(\mathbf{x})| \leq \epsilon$$

and

$$\forall \mathbf{x} \text{ s.t. } f(\mathbf{x}) = +1, p(\mathbf{x}) \geq 1 - \epsilon$$

- Call this positive one-sided polynomial approximation
- Theorem: Suffices for positive-reliable learning
- One-sided approximate degree can be much lower than approximate degree

Introduced recently in [Bun, Thaler 2013], [Sherstov 2014] to prove lower bounds in complexity theory

Main Result

Theorem

Any class F that has positive one-sided polynomial approximations of degree d and weight W, can be learned by an algorithm with:

- Running time $n^{O(d)}$
- Sample complexity polynomial in n, W, $1/\epsilon$

An analogous result is true for negative reliable learning.

Convex Program:

Find a polynomial p that minimizes, $\sum_{i:y_i=+1} (1 - p(\mathbf{x}_i))_+$ (hinge loss)

subject to: $\forall i$ such that $y_i = -1$, $p(x_i) \leq -1 + \epsilon$

Main Result

Theorem

Any class F that has positive one-sided polynomial approximations of degree d and weight W, can be learned by an algorithm with:

- Running time $n^{O(d)}$
- Sample complexity polynomial in n, W, $1/\epsilon$

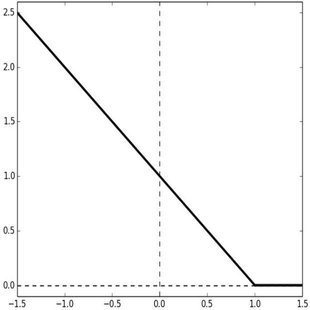
An analogous result is true for negative reliable learning.

Convex Program:

Find a polynomial p that minimizes, $\sum_{i:y_i=+1} (1 - p(\mathbf{x}_i))_+$ (hinge loss)

subject to: $\forall i$ such that $y_i = -1$, $p(x_i) \leq -1 + \epsilon$

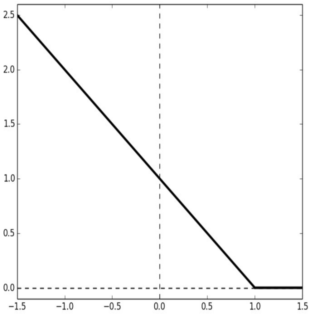
Proof Sketch



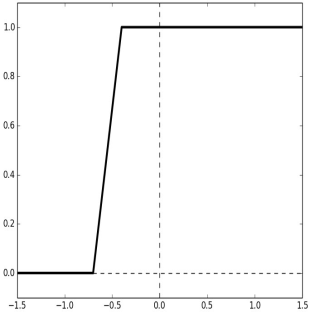
- For positive examples: hinge loss
- Convex loss function (objective)

- For negative examples: (almost) zero-one loss
- Posed as constraints

Proof Sketch

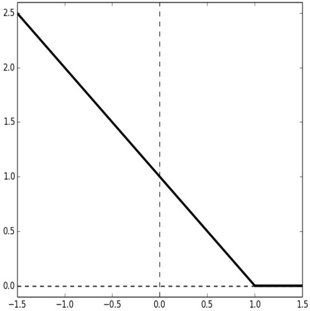


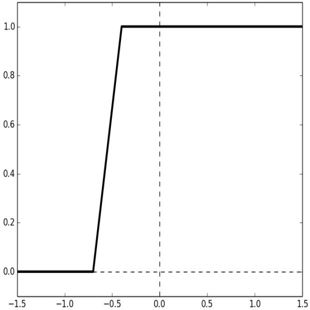
- For positive examples: hinge loss
- Convex loss function (objective)



- For negative examples: (almost) zero-one loss
- Posed as constraints

Proof Sketch





- For positive examples: hinge loss
- Convex loss function (objective)

- For negative examples: (almost) zero-one loss
- Posed as constraints

Existence of one-sided approximating polynomial implies that good solution to the convex program gives a good positive reliable classifier

One-sided approximations for low-weight thresholds

Consider the class of functions of the form:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} w_i x_i\right),$$

where w_i are integers. Let $W = \sum_i |w_i|$ denote the total weight.

Theorem

The class of threshold functions of weight W has (positive and negative) one-sided approximation degree $\tilde{O}(\sqrt{W})$

- Proof using Chebychev polynomials
- Majority has (pointwise) approximate-degree $\Omega(n)$.
- Majorities can be positive, negative and fully reliably learned in time $2^{\tilde{O}(\sqrt{n})}$
- Current best known algorithm for agnostic learning majority has running time $2^{O(n)}$.

One-sided approximations for low-weight thresholds

Consider the class of functions of the form:

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{n} w_i x_i\right),$$

where w_i are integers. Let $W = \sum_i |w_i|$ denote the total weight.

Theorem

The class of threshold functions of weight W has (positive and negative) one-sided approximation degree $\tilde{O}(\sqrt{W})$

- Proof using Chebychev polynomials
- Majority has (pointwise) approximate-degree $\Omega(n)$.
- Majorities can be positive, negative and fully reliably learned in time $2^{\tilde{O}(\sqrt{n})}$
- Current best known algorithm for agnostic learning majority has running time $2^{O(n)}$.

One-sided approximations for low-weight thresholds

Consider the class of functions of the form:

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{n} w_i x_i\right),$$

where w_i are integers. Let $W = \sum_i |w_i|$ denote the total weight.

Theorem

The class of threshold functions of weight W has (positive and negative) one-sided approximation degree $\tilde{O}(\sqrt{W})$

- Proof using Chebychev polynomials
- Majority has (pointwise) approximate-degree $\Omega(n)$.
- Majorities can be positive, negative and fully reliably learned in time $2^{\tilde{O}(\sqrt{n})}$
- Current best known algorithm for agnostic learning majority has running time $2^{O(n)}$.

One-sided approximations for low-weight thresholds

Consider the class of functions of the form:

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i x_i\right),$$

where w_i are integers. Let $W = \sum_i |w_i|$ denote the total weight.

Theorem

The class of threshold functions of weight W has (positive and negative) one-sided approximation degree $\tilde{O}(\sqrt{W})$

- Proof using Chebychev polynomials
- Majority has (pointwise) approximate-degree $\Omega(n)$.
- Majorities can be positive, negative and fully reliably learned in time $2^{\tilde{O}(\sqrt{n})}$
- Current best known algorithm for agnostic learning majority has running time $2^{O(n)}$.

One-sided approximations for low-weight thresholds

Consider the class of functions of the form:

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{n} w_i x_i\right),$$

where w_i are integers. Let $W = \sum_i |w_i|$ denote the total weight.

Theorem

The class of threshold functions of weight W has (positive and negative) one-sided approximation degree $\tilde{O}(\sqrt{W})$

- Proof using Chebychev polynomials
- Majority has (pointwise) approximate-degree $\Omega(n)$.
- Majorities can be positive, negative and fully reliably learned in time $2^{\tilde{O}(\sqrt{n})}$
- Current best known algorithm for agnostic learning majority has running time $2^{O(n)}$.

One-sided approximations: Composition Results

Theorem

Let F be a class of functions that has positive one-sided polynomial approximations of degree d and weight W, then if

$$g = f_1 \vee f_2 \vee \cdots \vee f_m$$

g has positive one-sided polynomial approximation of degree d and weight mW

- Thus, disjunctions of majority are positive reliably learnable
- Analogously, conjunctions of majority are negative reliably learnable

- Weight vs degree tradeoff in (one-sided) polynomial approximation results in sample complexity vs running time tradeoff

One-sided approximations: Composition Results

Theorem

Let F be a class of functions that has positive one-sided polynomial approximations of degree d and weight W, then if

$$g = f_1 \vee f_2 \vee \cdots \vee f_m$$

g has positive one-sided polynomial approximation of degree d and weight mW

- Thus, disjunctions of majority are positive reliably learnable
- Analogously, conjunctions of majority are negative reliably learnable

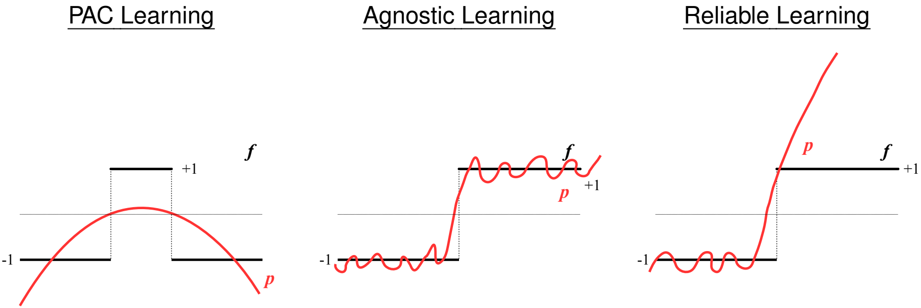
- Weight vs degree tradeoff in (one-sided) polynomial approximation results in sample complexity vs running time tradeoff

Outline

1. Introduction
2. Framework
   - Agnostic Learning Framework
   - Positive Reliable Learning
   - Fully Reliable Learning
3. Main Results
   - Polynomial Approximations
   - Learning Results
   - One-sided Approximations
4. Conclusion

Conclusion

Polynomial approximations play a fundamental role in learning!



- Algorithmic application of one-sided polynomial approximations
- Previously only used for lower-bounds in complexity theory
- Evidence that (fully) reliable learning easier than agnostic learning

Open Questions

- What can be said about one-sided degree of thresholds with larger weight?
  - For halfspaces with weight $2^{\Omega(n)}$, one-sided approximate degree is $\Omega(n)$.
- Other applications of one-sided polynomial approximations?

Thank you!