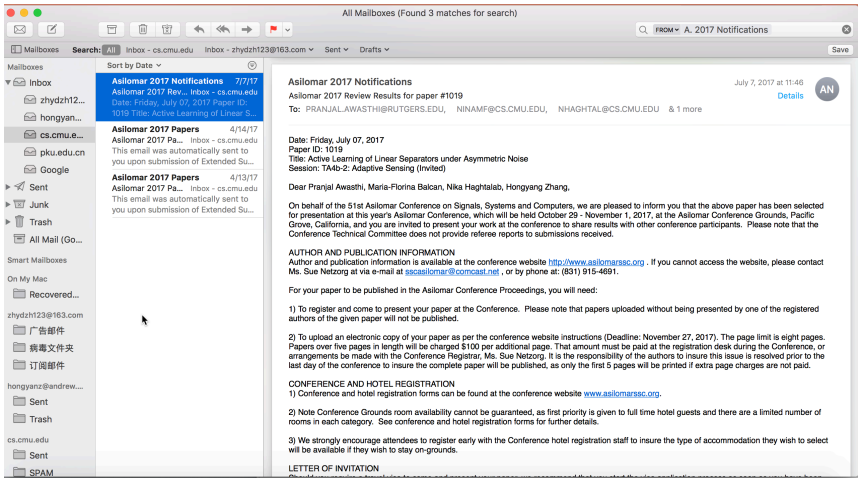
Active Learning of Linear Separators under Asymmetric Noise

Hongyang Zhang, Carnegie Mellon University
With Pranjal Awasthi, Maria-Florina Balcan, and Nika Haghtalab

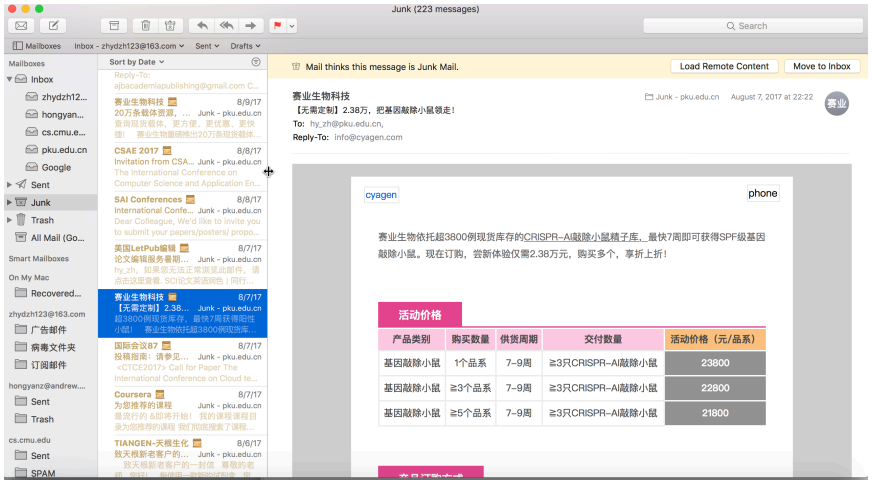
Asilomar Conference on Signals, Systems, and Computers

Learning of Halfspaces and 1-bit CS

Not Spam

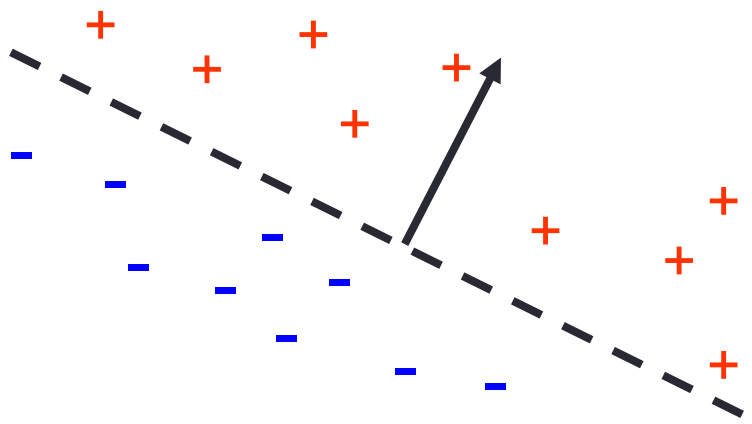


Spam



Goal: use emails seen so far to produce good prediction rule for future data.

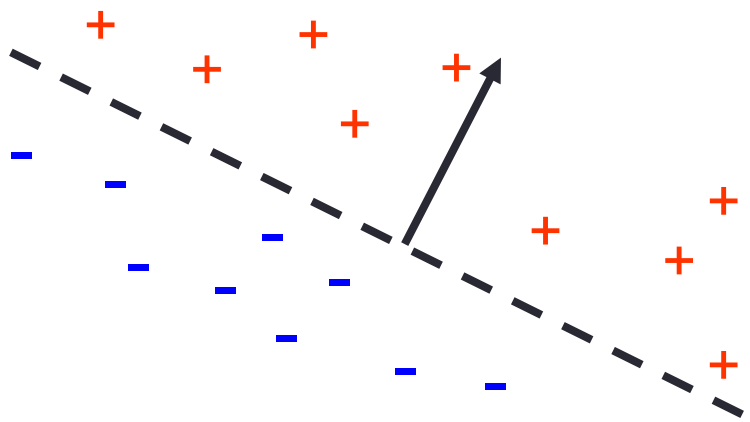
Learning of Halfspaces



What if we know the classifier is sparse? Is it possible that we require fewer samples?

[ABL] The Power of Localization for Efficiently Learning Linear Separators with Noise, JACM'17
[KLS] Learning halfspaces with malicious noise, JMLR'09
[KKMS] Agnostically learning halfspaces, FOCS'05

1-Bit Compressed Sensing



What if we know the classifier is sparse? Is it possible that we require fewer samples?

Difference with learning: Impose additional sparsity constraint

[PV] Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach, IEEE TIT'13

Optimization Formulation

- No Noise: Easy – solve ERM via a linear program

Find w such that $\forall i,\ y_i(w \cdot x_i) \geq 0$

- With Noise: Solve a non-convex problem

$\min_w \Pr_{(x,y)\sim\tilde{D}}[\mathrm{sgn}(w \cdot x) \neq y],\ (\text{s.t. } \|w\|_0 \leq t \text{ for 1-bit CS})$ (log-concave dist.)

- Sparsity (1-bit CS): Use a number of samples poly*(t,* log*(d/δ), 1/ε)*

Can we minimize the objective function to the accuracy of the information-theoretic limit under asymmetric noise model, although its formulation is non-convex?

The answer is affirmative!

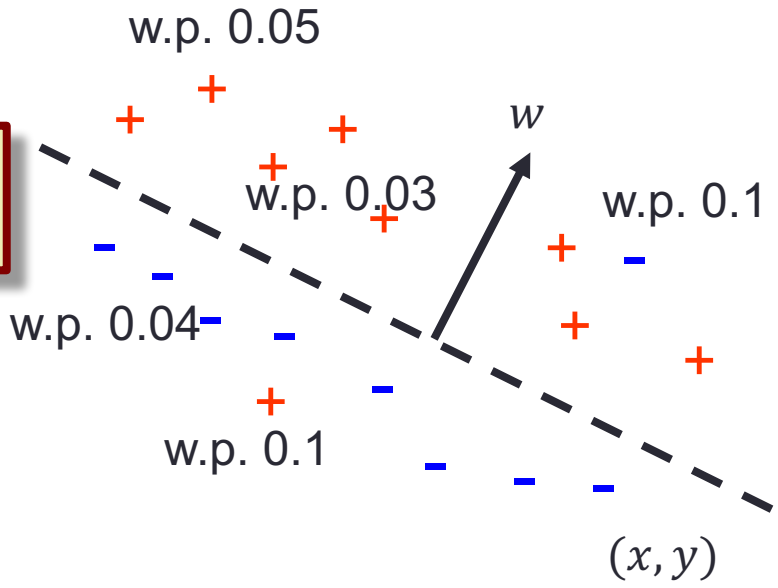
Outline

- Motivation and examples
- **Our settings**
- Our algorithms
- Our hardness results
- Conclusions

Asymmetric Noise model – Bounded Noise

Bounded Noise (a.k.a. Massart Noise):

For a fixed $\eta \leq ½$, for each x, the adversary flips the label of x with probability $\eta(x) \leq \eta$.

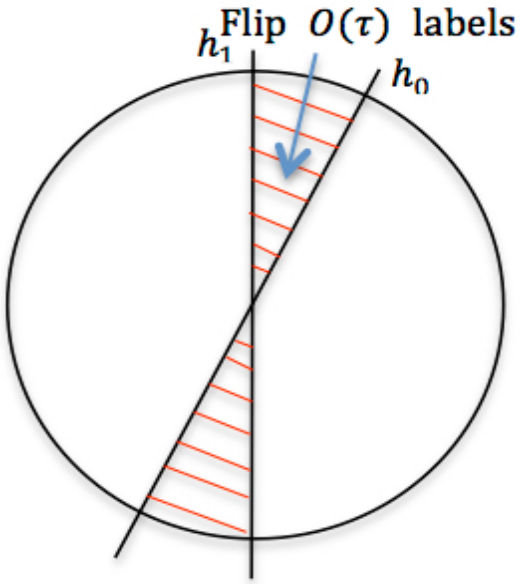


- Generalization of the RCN model
- Prior Result: $\eta \approx 10^{-6}$
- No result is known when $w \in \Re^d$ is *t*-sparse
- Information-theoretic limit: $OPT + \varepsilon$

Asymmetric Noise model – Adversarial Noise

Adversarial Noise:

The adversary can flip any τ fraction of labels of x.



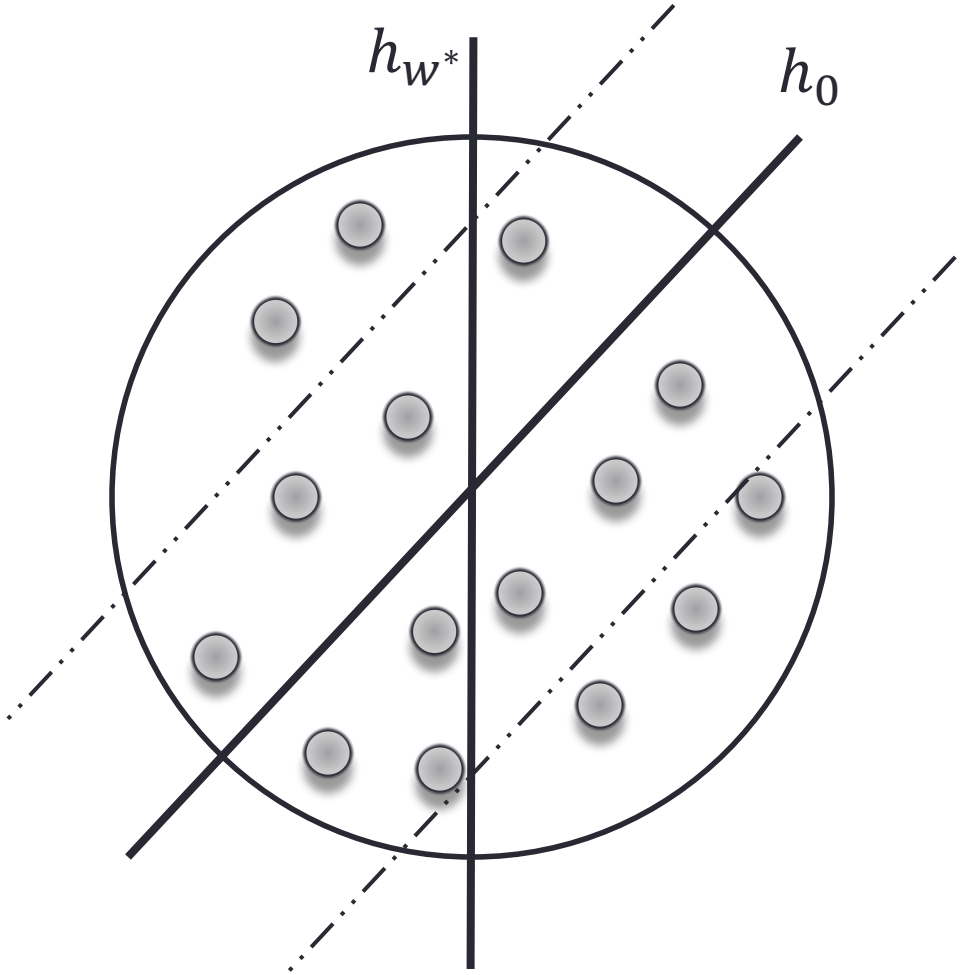
- No result is known when $w \in \Re^d$ is *t*-sparse
- Information-theoretic limit: $OPT + \tau + \varepsilon$

Outline

- Motivation and examples
- Our settings
- **Our algorithms**
- Our hardness results
- Conclusions

Algorithm

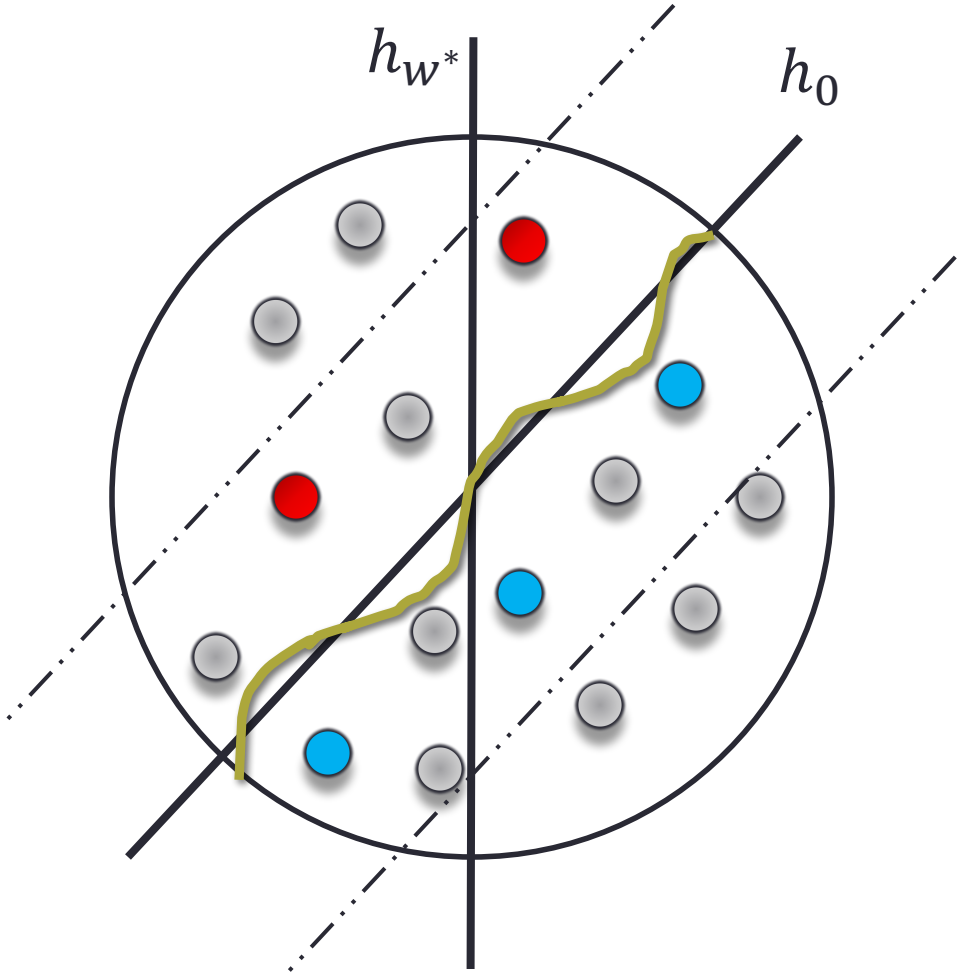
- Idea: Adaptively solve a sequence of convex programs



Sample unlabeled data
and have an initial guess

Algorithm

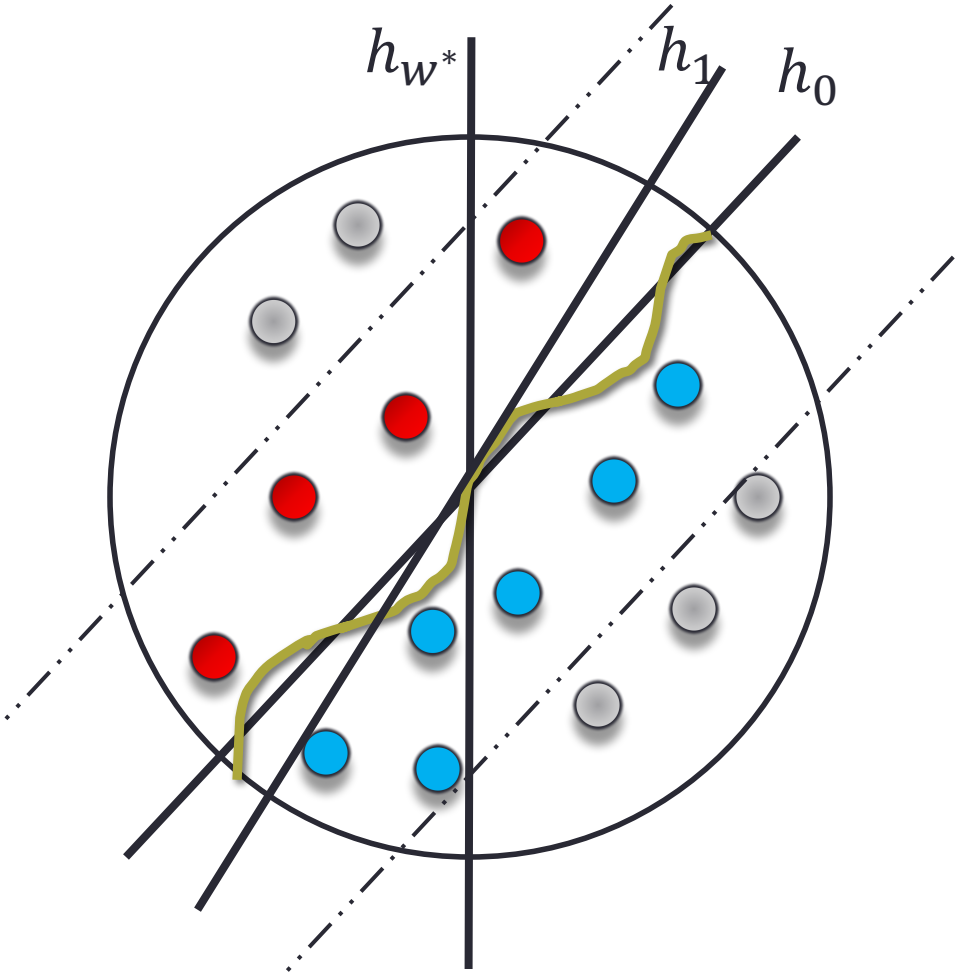
- Idea: Adaptively solve a sequence of convex programs



Ask some of labels in the band,
fit a polynomial to constant error
(require exp. time on 1/error)

Algorithm

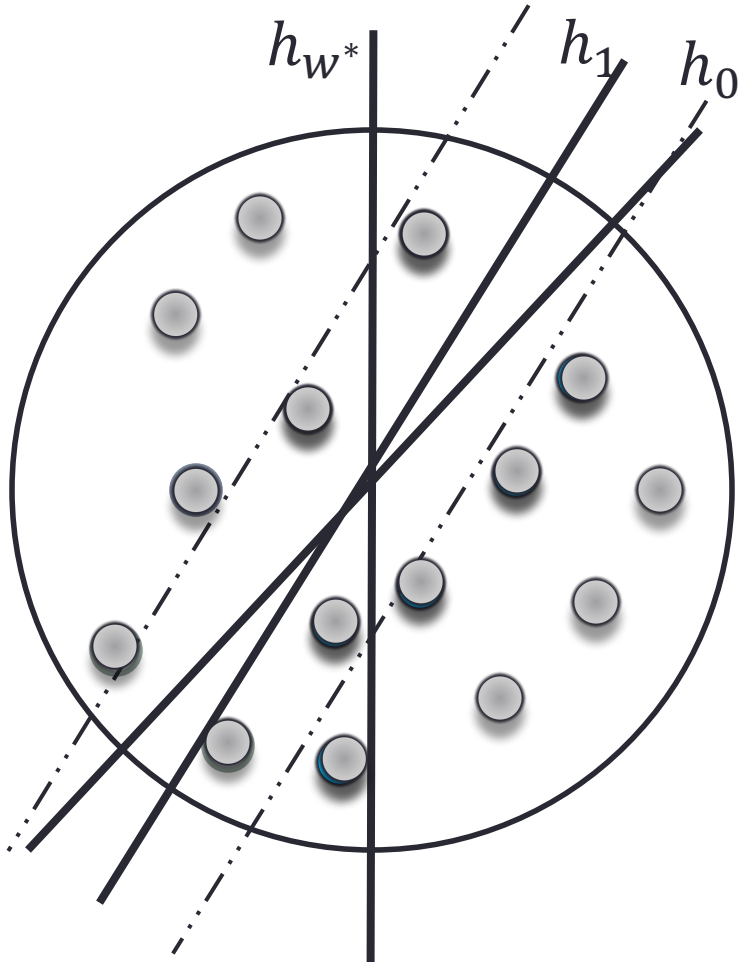
- Idea: Adaptively solve a sequence of convex programs



Label points in band by the polynomial, do hinge loss minimization to constant error, and obtain h_1

Algorithm

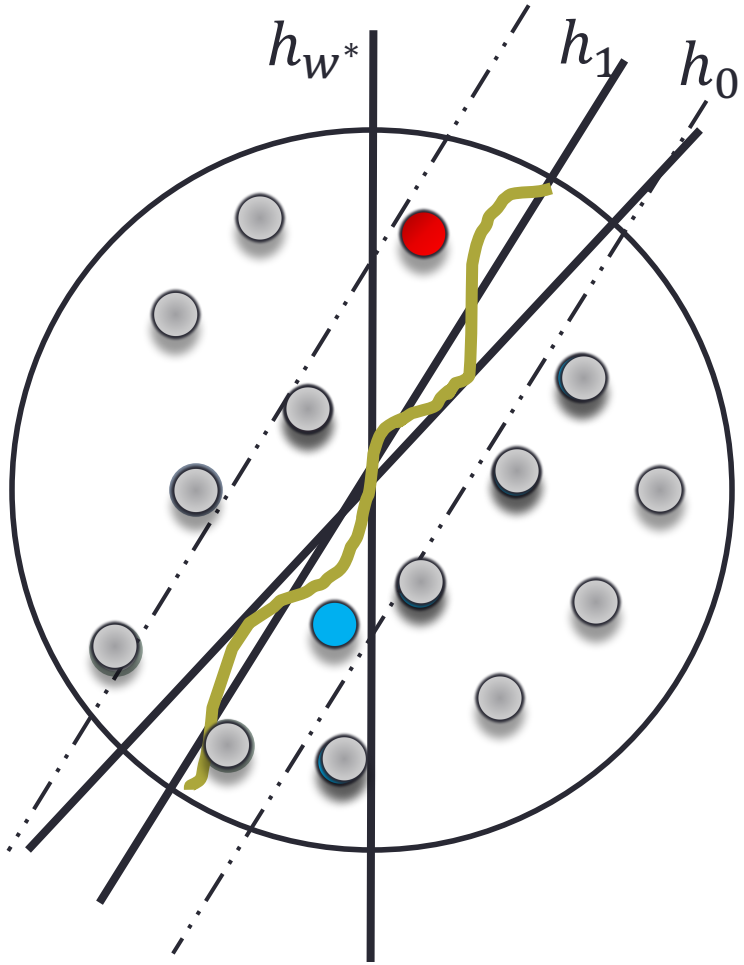
- Idea: Adaptively solve a sequence of convex programs



Halve the bandwidth around h_1, ask labels in the band, fit polynomial

Algorithm

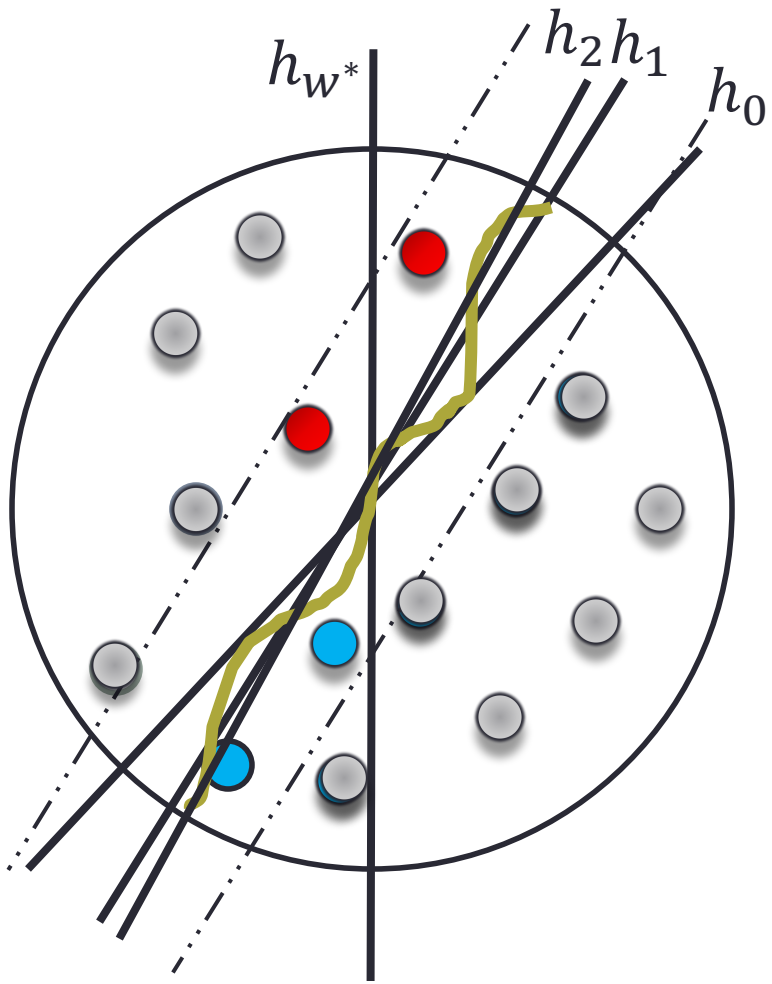
- Idea: Adaptively solve a sequence of convex programs



Halve the bandwidth around h_1, ask labels in the band, fit polynomial

Algorithm

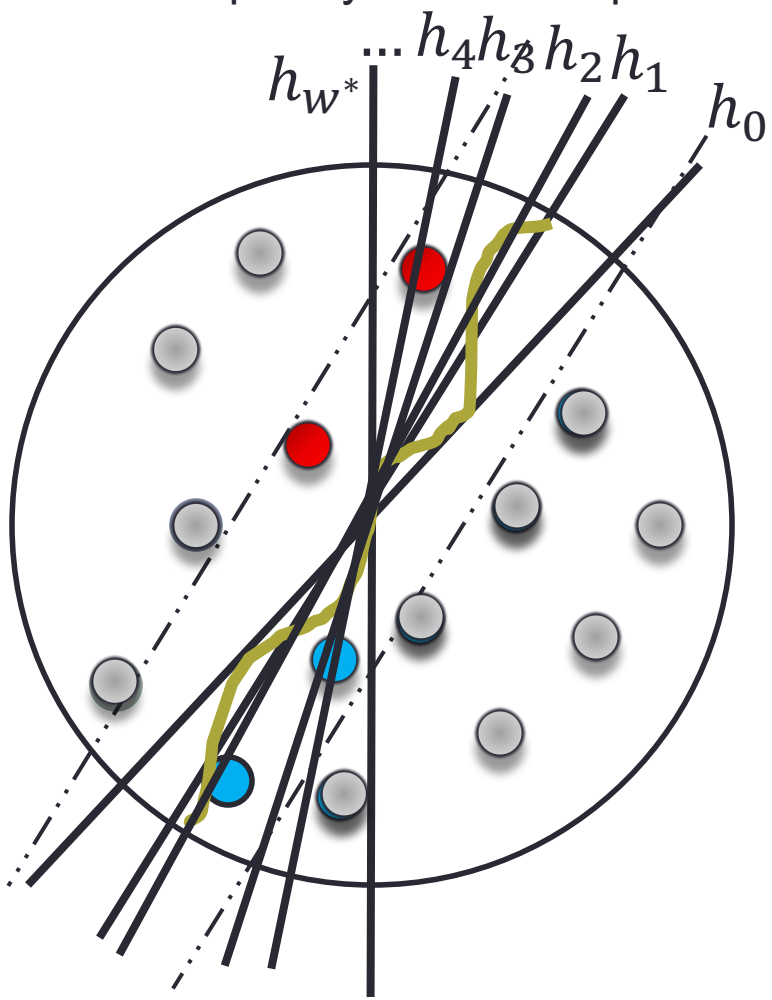
- Idea: Adaptively solve a sequence of convex programs



Label points in band by polynomial, do hinge loss minimization to constant error, and obtain h_2

Algorithm

- Idea: Adaptively solve a sequence of convex programs



Repeat $\log(1/\varepsilon)$ rounds

Main Results

In $\Re^d$, for the log-concave dist. with polynomial time and probability at least 1- δ:

Theorem 1 (Bounded Noise, Learning):

Label Complexity: poly*(d,* log*(1/δ),* log*(1/ε))*
Guarantee: $OPT + \varepsilon$

Theorem 2 (Bounded Noise, 1-bit CS):

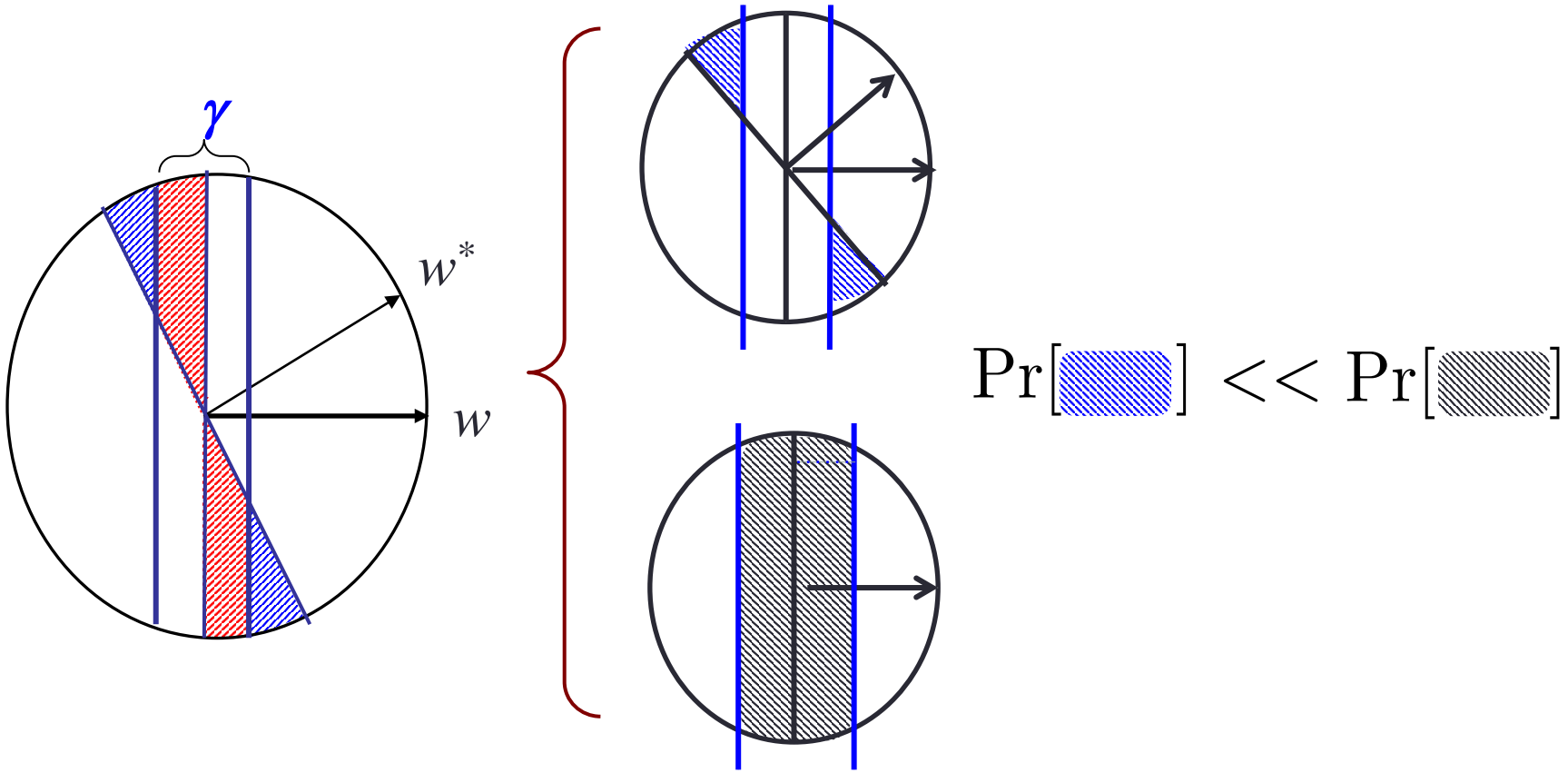
Label Complexity: poly*(t,* log*(d/δ), 1/ε)*
Guarantee: $OPT + \varepsilon$

Theorem 3 (Adversarial Noise, 1-bit CS):

Label Complexity: *O (t,* polylog*(d/δ), 1/ε)*
Guarantee: $OPT + O(\tau) + \varepsilon$

Intuition and Analysis

Most of the errors are near the decision boundary:



Intuition and Analysis

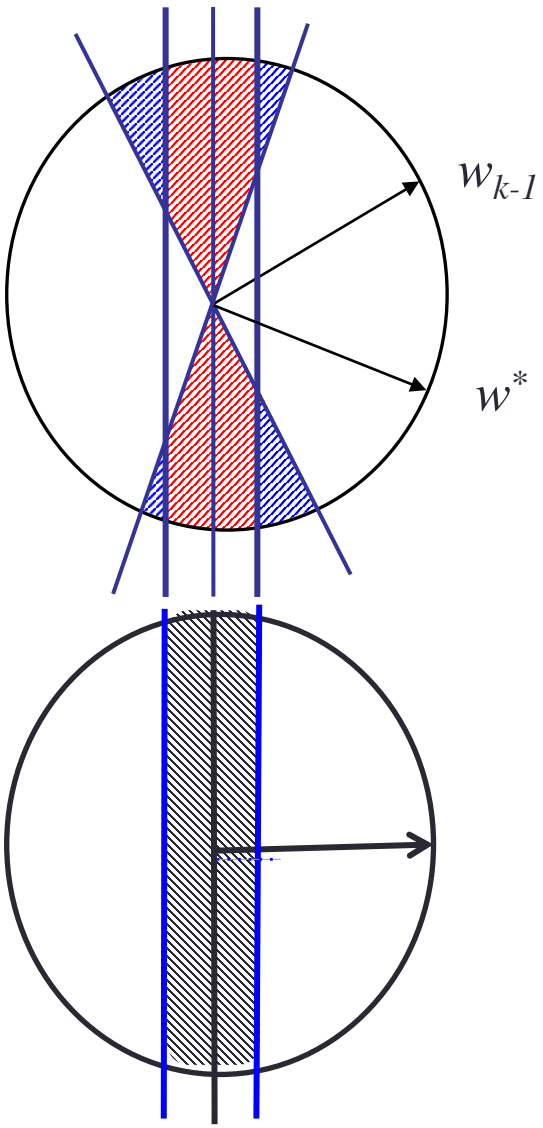
$err(w) = \Pr[\,] + \Pr[\,]$

$\Pr[\,] = \Pr[\,] \times \mathrm{err}_{band}(w)$

$\Pr[\,]$ small

How to find w?

- Hinge loss minimization
- Works only when $\eta \approx 10^{-6}$
- Poly Regression [Kalai et al.] with constant error
- Return a poly, rather than a halfspace
- Combine two together



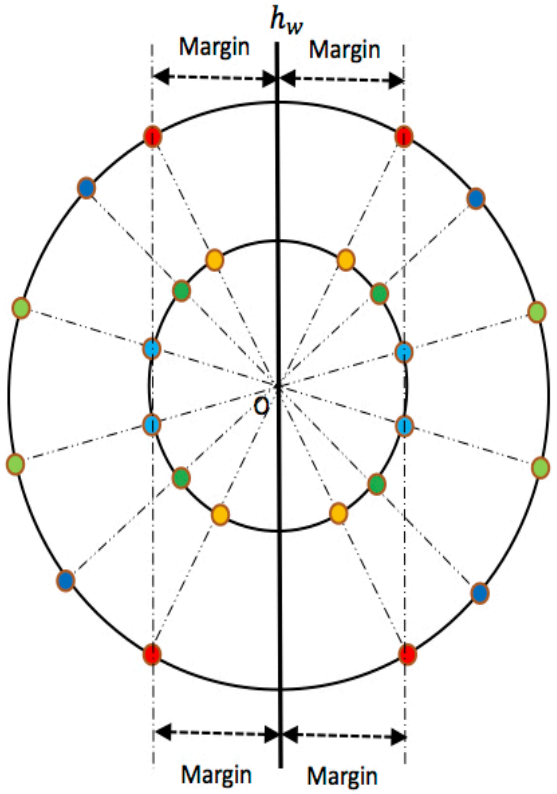
Outline

- Motivation and examples
- Our settings
- Our algorithms
- **Our hardness results**
- Conclusions

Hardness – One shot minimization

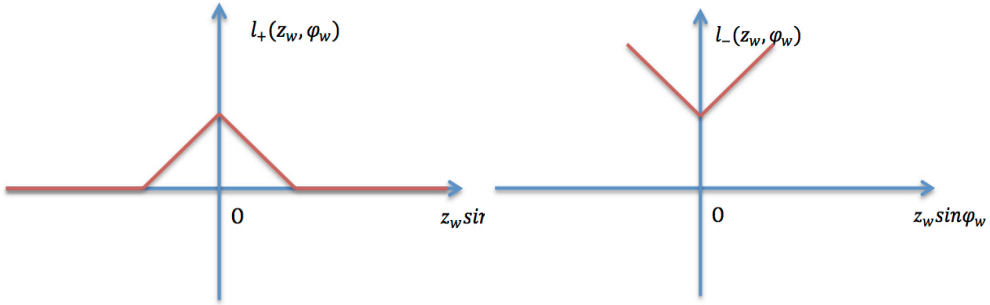
Continuous loss function on h_w satisfies:

- Symmetric w.r.t. h_w
- The loss is larger if h_w is inconsistent with the true label



A couple of examples:

- Hinge, Logistic, Square, Exponential Loss, etc.
  E.g. Hinge:



Hardness – One shot minimization

Theorem 4 (for bounded noise):

Any one-shot minimization of function satisfying above properties cannot achieve $OPT + \varepsilon$ error under log-concave distribution with bounded noise.

Outline

- Motivation and examples
- Our settings
- Our algorithms
- Our hardness results
- **Conclusions**

Conclusions

- Learning of halfspaces and 1-bit CS
  - Polynomial-time algorithm
  - Noise-tolerant for bounded and adversarial noise models
  - Achieve information-theoretic limits
  - Solve a non-convex problem via a sequence of convex programs
- Hardness results
  - One-shot minimization does not work
- Future Work
  - Explore the localization technique to the other applications

Thank You