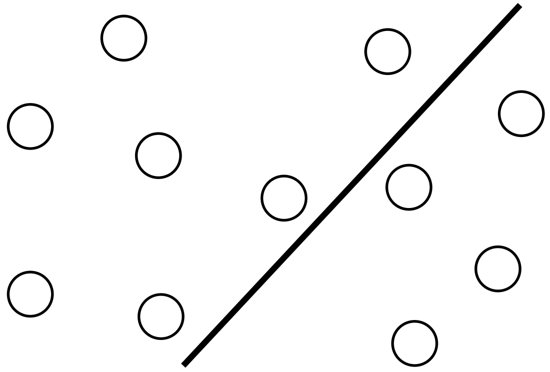
Improved algorithms for efficient active learning halfspaces with Massart and Tsybakov noise

Chicheng Zhang

University of Arizona

Joint work with Yinan Li (University of Arizona)

Active learning for classification



- Given: $(x_1, y_1), \ldots (x_n, y_n)$

  features   **interactive** label queries

- Find: Classifier h in a class H to predict y from x
  - With few interactive label queries

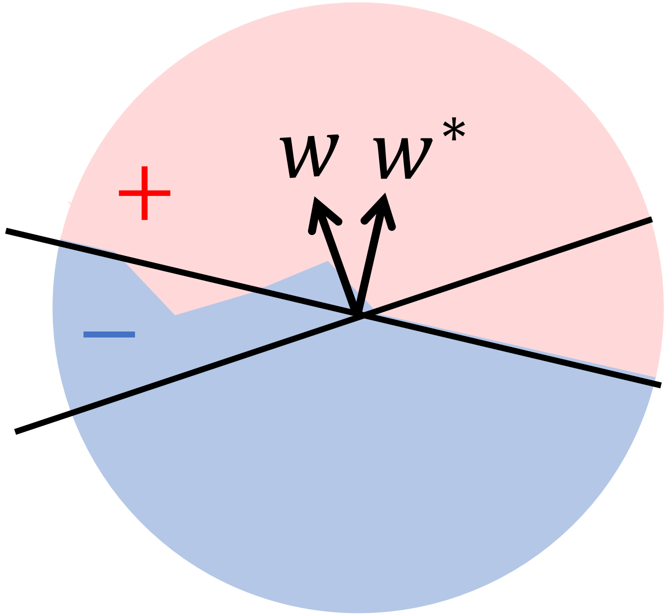
- Useful in practical settings where labels are expensive to obtain

Outline

- Active learning halfspaces with noise
- The algorithm
- Conclusion and open problems

Active learning in the PAC model [V84,BBL06]

- Setting:
  - (x, y) drawn from a distribution D
  - x drawn from a ``structured'' distribution [DKKTZ20] (e.g. Gaussian, Laplace, ..)
  - Linear classifiers: $H = \{\text{sign}(w \cdot x) : w \in \mathrm{R}^d\}$
  - Error $\text{err}(w) = P\big(y \neq \text{sign}(w \cdot x)\big)$
  - Optimal linear classifier $w^* = \text{argmin}_w \, \text{err}(w)$

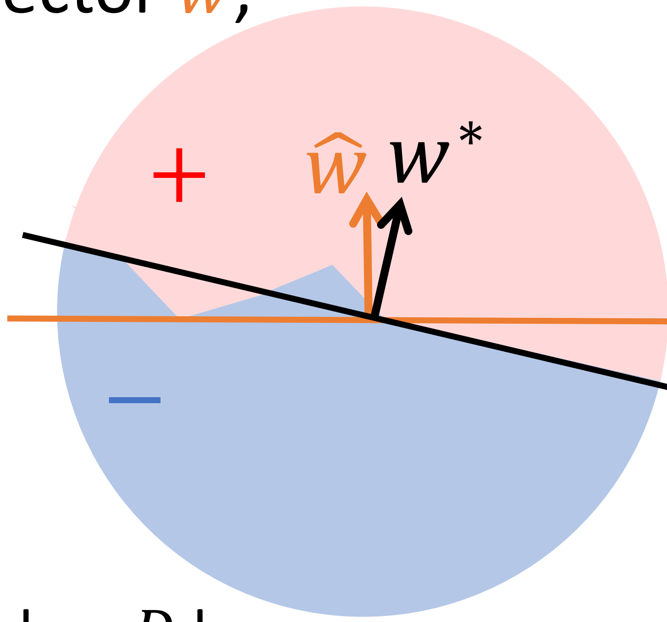


Active learning in the PAC model [V84,BBL06]

- Goal: computationally efficient algorithm that returns a vector $\hat{w}$, such that

$$\mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon,$$

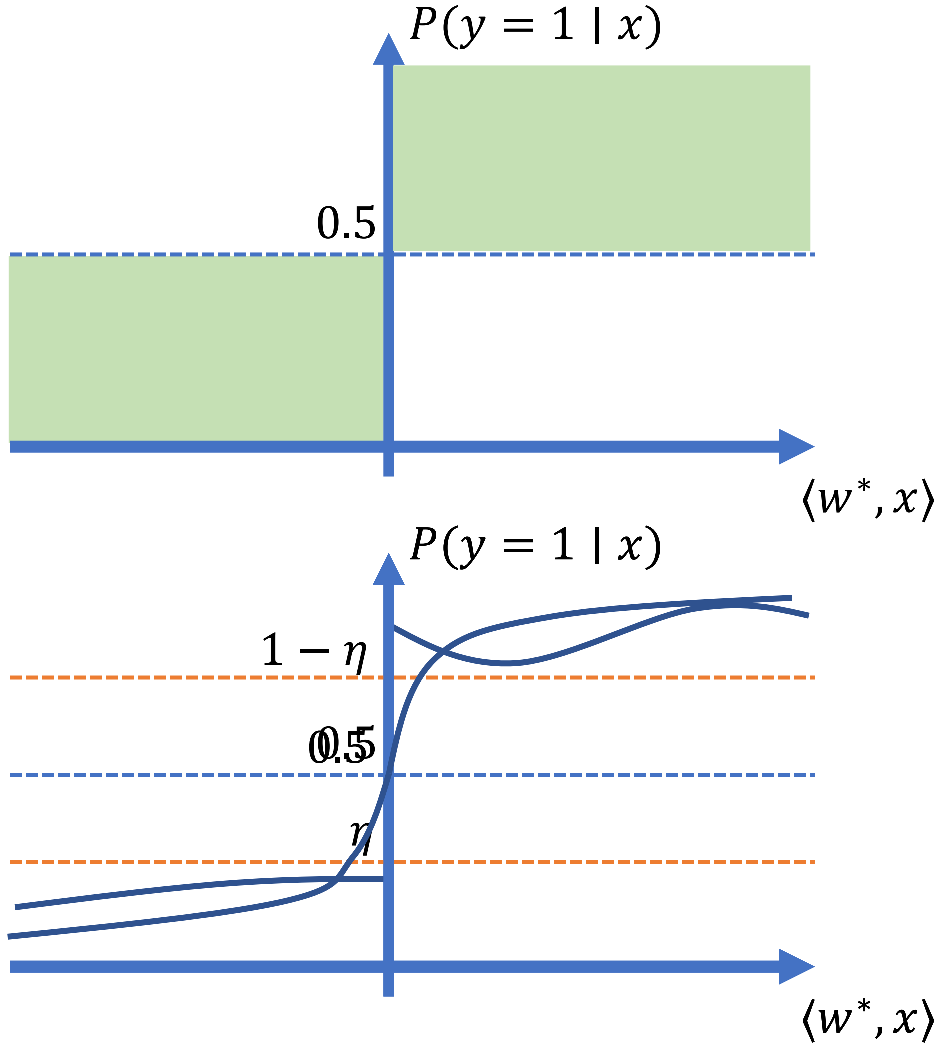
  using a few label queries



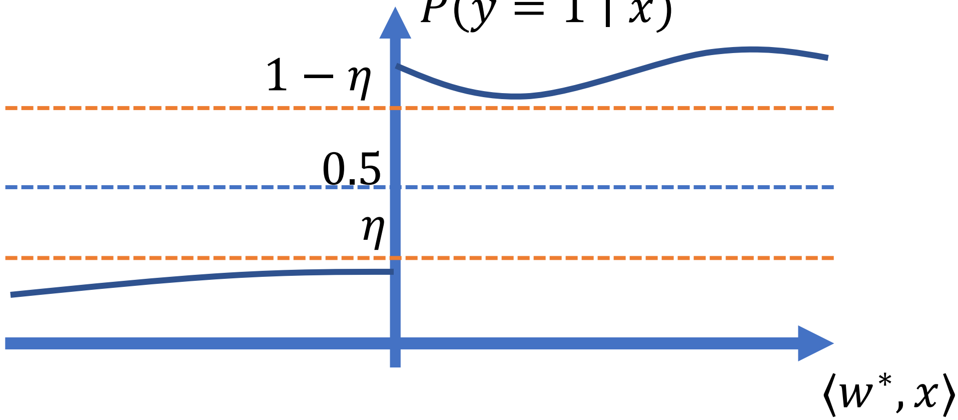
- Challenge: noise tolerance
  - Agnostically learning halfspaces is computationally hard even when D has ``nice'' unlabeled data distribution [KK14, DKZ20]
  - Benign noise conditions

Learning halfspaces under benign noise

- Main assumption: there exists some halfspace w^* that is Bayes optimal, i.e. for all x,
  $$\eta(x) := P_D(y \neq \text{sign}(w^* \cdot x)|x) \leq 1/2$$
- η-Massart [MN06]: for all x, $\eta(x) \leq \eta < \frac{1}{2}$
- α-Tsybakov [T04] for $\alpha \in (0,1)$: for all t,
  $$P_{\mathrm{D}}(1/2 - \eta(x) \leq t) \leq O\left(t^{\alpha/(1-\alpha)}\right)$$
- α-Geometric Tsybakov [e.g., CN08]: for all x,
  $$\frac{1}{2} - \eta(x) \geq |w^* \cdot x|^{\frac{1-\alpha}{\alpha}}$$



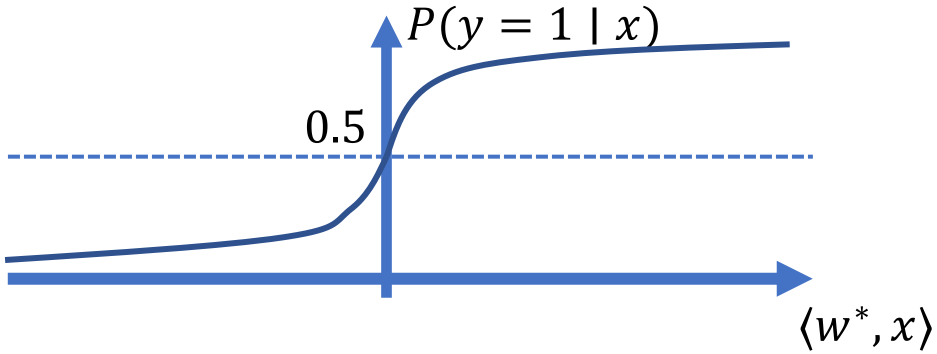
Main results - Massart noise



Algorithm	Efficient?	Label complexity in $\widetilde{O}$
[BL13]	No	$\frac{d}{(1-2\eta)^2}\text{polylog}(1/\epsilon)$
[ZSA20]	Yes	$\frac{d}{(1-2\eta)^4}\text{polylog}(1/\epsilon)$
This work	Yes	$\frac{d}{(1-2\eta)^2}\text{polylog}(1/\epsilon)$

- Such efficient and label-optimal results for learning Massart halfspaces were previously only known for uniform distribution [YZ17]
  - Our work significantly relaxed the distributional requirements
- Some assumptions on unlabeled distribution seem necessary [CKMY20, DK20]

Main results – Tsybakov noise



Algorithm	Efficient?	Label complexity in $\tilde{O}$
[BL13]	No	$d\left(\frac{1}{\epsilon}\right)^{2-2\alpha}$
[DKKTZ20]	Yes	$\mathrm{poly}(d)\left(\frac{1}{\epsilon}\right)^{O(1/\alpha)}$
This work ($\alpha \in \left(\frac{1}{2}, 1\right]$)	Yes	$d\left(\frac{1}{\epsilon}\right)^{\frac{2-2\alpha}{2\alpha-1}}$
This work (Geometric Tsybakov)	Yes	$d\left(\frac{1}{\epsilon}\right)^{\frac{2-2\alpha}{\alpha}}$

- Our label complexity results improve over passive learning for a range of α values

Outline

- Active learning sparse halfspaces with noise
- The algorithm
- Conclusion and open problems

The algorithm: overview

- Main idea: maintain iterate $\{w_k\}$ such that $\theta(w_k, w^*)$ shrinks geometrically

// Initialization
$w_1 \leftarrow \text{Initialize}()$.

// Refinement
In phases $k = 1,2,..,k_0 = \log(1/\epsilon)$:

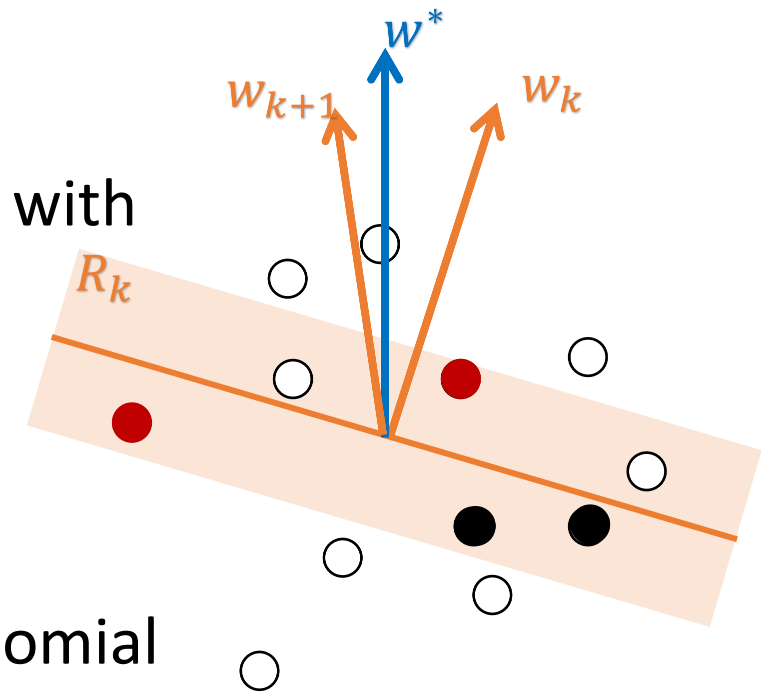
$$w_{k+1} \leftarrow \text{Refine}(w_k, 2^{-(k+1)}).$$

Return w_{k_0+1}.

Acute initialization

Ensuring w_k has angle $\leq 2^{-k}$ with w^*

Refine: design challenges



- A series of prior works combine margin-based sampling with loss minimization techniques to design Refine
- [BL13]: 0-1 loss minimization
  - Computationally inefficient
- [ABHU15, ABHZ16]: surrogate loss minimization + polynomial regression
  - Analysis only tolerates $\eta \leq$ small constant, or requires high label complexity
- [ZSA20]: SGD-like update rule + iteration-dependent sampling
  - Specialized to Massart noise (needs to know η)

The algorithm: Refine

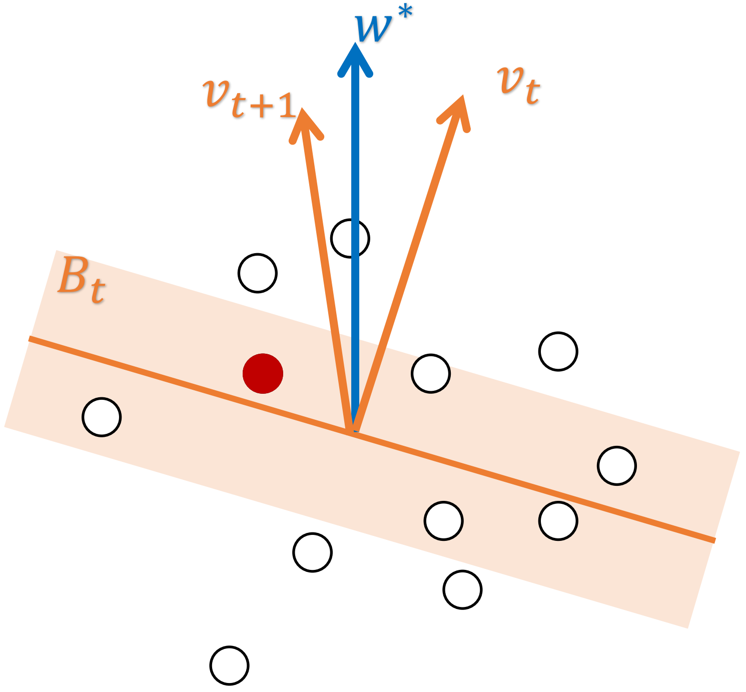
Input: halfspace v_1, target angle θ
Output: halfspace v (that has angle $\leq \theta$ to w^*)

For $t = 1,2,\ldots,T$:

1. **Sample:** $(x_t, y_t) \leftarrow$ example drawn from $D|_{B_t}$,
   where $B_t = \{x: |v_t \cdot x| \leq b\}$.

2. **Update:** $v_{t+1} \leftarrow v_t - \alpha g_t$, where $g_t = -y_t x_t$

Return average: $v \leftarrow \frac{1}{T}\sum_{t=1}^{T} v_t$



Key difference from [ZSA20]: simpler definition of g_t leads to broader noise tolerance

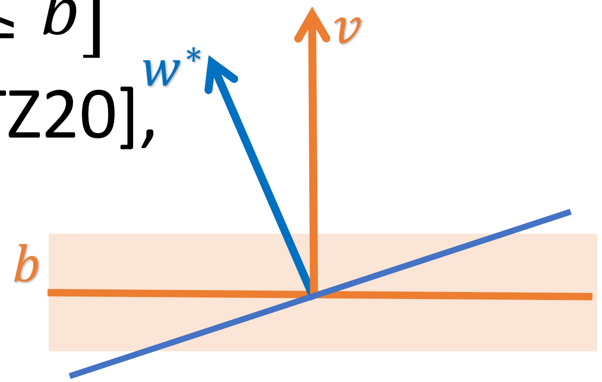
Refine: theoretical properties

- **Theorem:** If $\theta(v_1, w^*) \leq 2\theta$, then with high probability, $\mathrm{Refine}(v_1, \theta)$ returns a vector v with $\theta(v, w^*) \leq \theta$,if T is of order:
  - $\frac{d}{(1-2\eta)^2}$, under η-Massart noise;
  - $d\left(\frac{1}{\theta}\right)^{\frac{2-2\alpha}{2\alpha-1}}$, under α-Tsybakov noise with $\alpha \in \left(\frac{1}{2}, 1\right]$;
  - $d\left(\frac{1}{\theta}\right)^{\frac{2-2\alpha}{\alpha}}$, under α-Geometric Tsybakov noise.

Refine: analysis

- **Key observation:** Refine can be viewed as optimizing the following ``proximity function'' in a nonstandard way:

$$\psi_b(v) = \mathrm{E}\left[\left(1 - 2\eta(x)\right) |w^* \cdot x| \mid |v \cdot x| \le b\right]$$



- Different from ``nonconvex optimization'' views [GCB09, DKTZ20], although algorithmically similar

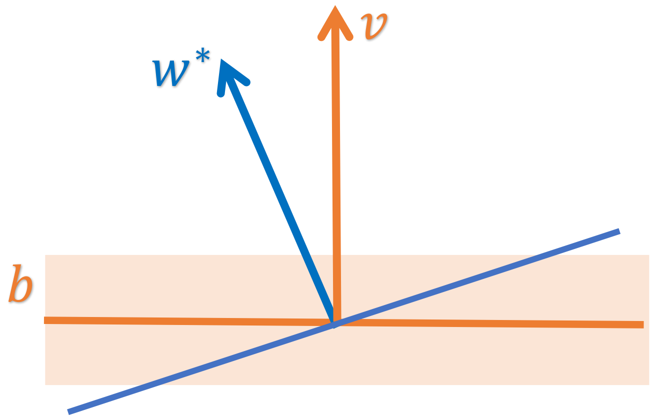
- Idea: rewriting OGD's regret guarantees over g_t's:

$$\frac{1}{T}\sum_{t=1}^{T}\langle -w^*, g_t\rangle \le \frac{1}{T}\sum_{t=1}^{T}\langle -v_t, g_t\rangle + O\left(\frac{1}{\sqrt{T}}\right)$$

Concentrates to $\frac{1}{T}\sum_{t=1}^{T}\psi_b(v_t)$ Can be made small by tuning b, T

The ``proximity function'' ψ_b

- $\psi_b(v) = \mathrm{E}\left[\left(1 - 2\eta(x)\right) |w^* \cdot x| \mid |v \cdot x| \le b\right]$



- **Lemma (simplified):** For ``structured'' D, $\psi_b(v)$ is at least (of order):
  - $(1 - 2\eta)\theta(v, w^*)$, under η-Massart noise;
  - $b^{(1-\alpha)/\alpha}\theta(v, w^*)$, under α-Tsybakov noise;
  - $\theta(v, w^*)^{1/\alpha}$, under α-Geometric Tsybakov noise.

- Optimizing $\psi_b(v) \Rightarrow$ optimizing $\theta(v, w^*)$

Initialize: design challenges and resolution

- [ZSA20]: average-based initialization – label inefficient ☹
  - e.g. results in $O\left(\frac{d}{(1-2\eta)^4}\right)$ label complexity under η-Massart noise

- This work: a new initialization procedure
  - Key observation: Refine *with arbitrary initialization* label-efficiently returns a halfspace with acute angle with w^*, with constant probability
  - ``Boosting the confidence'' using a repeat-and-select procedure
  - Results in optimal label complexity under η-Massart noise ☺

Outline

- Active learning sparse halfspaces with noise
- The algorithm
- Discussions

Discussions

- Under Massart noise, our work significantly relaxes the distributional requirements for efficient and label-optimal learning halfspaces
  - Can they be further relaxed, e.g., to s-concave distributions [BZ17]?
- Under (Geometric) Tsybakov noise, our analysis pays a large price when doing angle-excess error conversion
  - Can we get around this?
- Under Tsybakov noise, our algorithm has a higher label complexity than computationally inefficient algorithms, and cannot handle $\alpha \leq 1/2$
  - Can we achieve efficiency and label-optimality simultaneously?

References

- [ABHU15] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning halfspaces with bounded noise. COLT 2015.
- [ABHZ16] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. COLT 2016.
- [BBL06] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. ICML 2006.
- [BL13] Maria-Florina Balcan and Philip M. Long. Active and passive learning of linear separators under logconcave distributions.
- [CKMY20] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. NeurIPS 2020.
- [DK20] Ilias Diakonikolas and Daniel M Kane. Hardness of learning halfspaces with massart noise.arXiv preprintarXiv:2012.09720, 2020.
- [DKKTZ20] Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, Nikos Zarifis. A Polynomial Time Algorithm for Learning Halfspaces with Tsybakov Noise. ArXiv 2020.
- [DKTZ20] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. COLT 2020.
- [DKZ20] Ilias Diakonikolas, Daniel M Kane, and Nikos Zarifis. Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals. NeurIPS 2020.
- [GCB09] A Guillory, E Chastain, J Bilmes, Active learning as non-convex optimization. AISTATS 2009.
- [KK14] Adam Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space.
- [V84] Valiant. A Theory of the Learnable. JACM, 1984
- [YZ17] Songbai Yan and Chicheng Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces, NeurIPS 2017.
- [ZSA20] Chicheng Zhang, Jie Shen, and Pranjal Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. NeurIPS 2020.

Thank you!

arXiv: 2102.05312