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

Chicheng Zhang and Yinan Li
The University of Arizona

COLT 2021

THE UNIVERSITY OF ARIZONA

Abstract

We give an efficient PAC active halfspace learning algorithm that has improved noise-tolerance and label efficiency under benign noise conditions, given that the unlabeled data distribution satisfies certain structural properties [DKKTZ20]. Specifically:

1. Under Massart noise, it achieves optimal label complexity; such efficient and label-optimal results were previously only known when the unlabeled data distribution is uniform [YZ17].
2. Under two subfamilies of Tsybakov noise, it achieves improved label complexities compared to passive learning algorithms.

Problem: efficient active learning halfspaces with benign noise

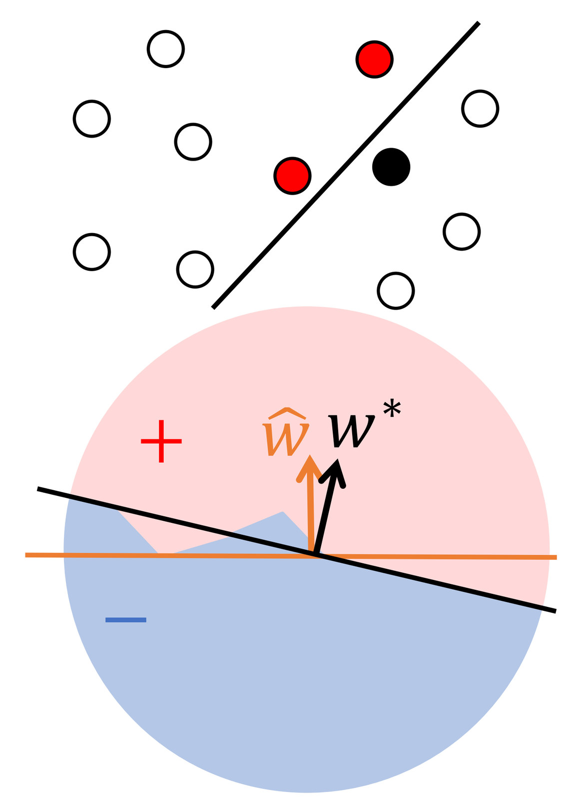
- ($\boxed{x}$, $\boxed{y}$) drawn from a distribution D

  features   **interactive** label queries

- x drawn from a structured distribution [DKKTZ20] (e.g. isotropic log-concave distributions)

- Linear classifiers: $H = \{\mathrm{sign}(w \cdot x): w \in \mathrm{R}^d\}$
- Error $\mathrm{err}(w) = P(y \neq \mathrm{sign}(w \cdot x))$
- Optimal linear classifier $w^* = \mathrm{argmin}_w \mathrm{err}(w)$

- 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

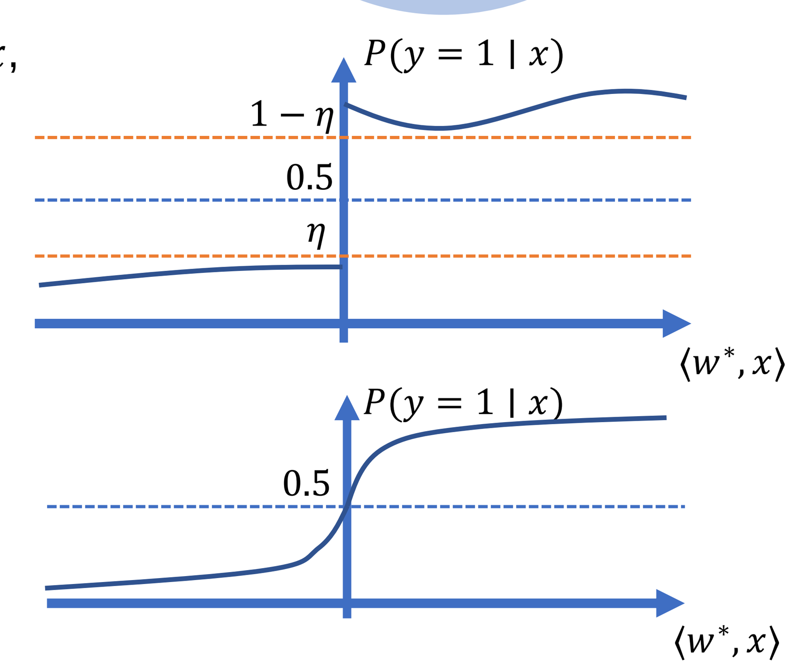


- Main assumption: there exists some halfspace w^* that is Bayes optimal, i.e., for all x,
  $\eta(x) := P_D(y \neq \mathrm{sign}(w^* \cdot x) | x) \leq 1/2$

- η-Massart [MN06]: for all x, $\eta(x) \leq \eta < \frac{1}{2}$

- α-Tsybakov [T04] for $\alpha \in (0,1)$: for all t, $P_D(1/2 - \eta(x) \leq t) \leq O\left(t^{\alpha/(1-\alpha)}\right)$

- α-Geometric Tsybakov [e.g., CN08]: for all x, $\frac{1}{2} - \eta(x) \geq |w^* \cdot x|^{\frac{1-\alpha}{\alpha}}$



Main result: Massart noise

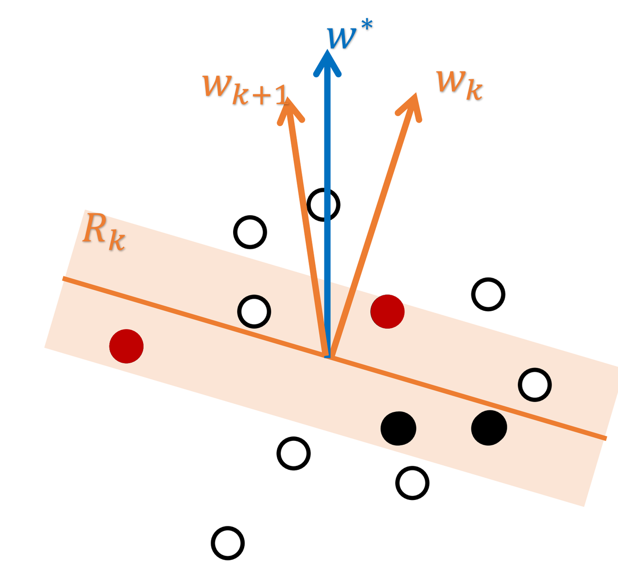
Algorithm	Efficient?	Label complexity in $\tilde{O}$
[BL13]	No	$\frac{d}{(1-2\eta)^2}\mathrm{polylog}(1/\epsilon)$
[ZSA20]	Yes	$\frac{d}{(1-2\eta)^4}\mathrm{polylog}(1/\epsilon)$
This work	Yes	$\frac{d}{(1-2\eta)^2}\mathrm{polylog}(1/\epsilon)$

Main result: 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	$d\left(\frac{1}{\epsilon}\right)^{\frac{2-2\alpha}{\alpha}}$

Algorithm skeleton

$w_1 \leftarrow \mathrm{Initialize}()$.   //Acute Initialization
In phases $k = 1,2,..,k_0 = \log(1/\epsilon)$:
    $w_{k+1} \leftarrow \mathrm{Refine}(w_k, 2^{-(k+1)})$.   // Refinement
Return w_{k_0+1}.



Refine: design challenges and related work

A series of prior works combine margin-based sampling with loss minimization techniques to design Refine:

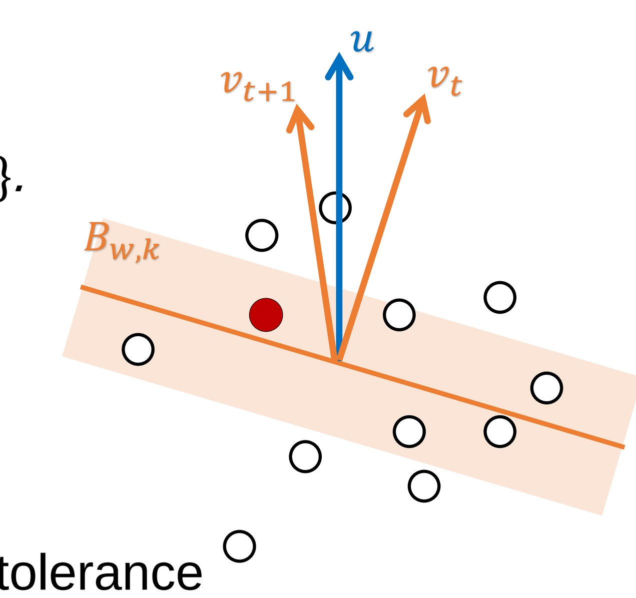
- [BL13]:computationally inefficient (0-1 loss minimization)
- [ABHU15, ABHZ16]: analysis only tolerates $\eta \leq$ small constant, or requires high label complexity
- [ZSA20]: specialized to Massart noise (needs to know η)

Refine: our design

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

Algorithmically similar to ``nonconvex optimization'' view [GCB09, DKTZ20], but analysis very different (see next)

Analysis: key ideas

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.

Key observation: Refine optimizes the following ``proximity function'' in a nonstandard way:

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

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

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

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

Lemma (simplified): For ``structured'' D, under one of the three noise conditions, $\psi_b(v)$ is lower bounded by an increasing function of $\theta(v, w^*)$.

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