Revisiting Perceptron: Efficient and Label-Optimal Learning of Halfspaces

Songbai Yan yansongbai@eng.ucsd.edu UC San Diego

Chicheng Zhang chicheng.zhang@microsoft.com UC San Diego -> Microsoft Research New York City

ABSTRACT

- We propose an efficient Perceptron-based algorithm for actively learning homogeneous halfspaces. Specifically:
  - Under the bounded noise condition, our algorithm achieves computational efficiency and label-optimality, improving over the state-of-the-art algorithms [1,3].
  - Under the adversarial noise condition, our algorithm achieves a near-optimal label complexity and requires less time than the state-of-the-art method [2].
  - In addition, our algorithm can be converted to an efficient passive learning algorithm with near-optimal sample requirement.

SETTING

- **Active Learning**

  Given:

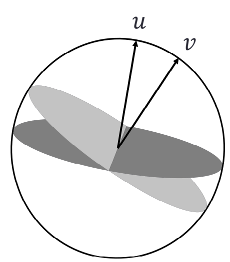
  (1) A distribution D over $X \times Y$, a set of classifiers H;

  (2) Ability to draw unlabeled examples $x \sim D_X$;

  (3) Ability to make interactive queries to get label $y \sim D_{Y|X=x}$ for example x;

  Goal:

  Find an $h \in H$ such that $P_D(h(X) \neq Y)$ is small while making only a few label queries.



- Learning homogeneous halfspaces: $H = \{\text{sign}(v \cdot x) : v \in R^d, \|v\| = 1\}$.
- Unlabeled distribution D_X: uniform over the unit sphere $\{x \in R^d : \|x\| = 1\}$.
- **Noise Models:**
  - η-bounded noise ($0 \le \eta < \frac{1}{2}$): there is a halfspace u such that for all x, $P_D(Y \neq \text{sign}(u \cdot x) \mid X = x) \le \eta$.
  - ν-adversarial noise ($0 \le \nu < 1$): there is a halfspace u such that $P_D\left(Y \neq \text{sign}(u \cdot X)\right) \le \nu$.
- **Label Complexity**: the number of labels required to output a halfspace v such that $P_D\left(\text{sign}(v \cdot X) \neq \text{sign}(u \cdot X)\right) \le \epsilon$.

RELATED WORK

- **Noise-free ($\eta = 0$ or $\nu = 0$)**
  - Efficient and label-optimal solutions have been proposed (e.g. [3,5])
- **Bounded noise**
  - [3]: a margin-based algorithm which is label-optimal but computationally inefficient.
  - [1]: combining the idea of [3] and polynomial regression. Efficient but requires $\tilde{O}\left(d^{(1-2\eta)^{-4}} \ln \frac{1}{\epsilon}\right)$ labels.
- **Adversarial noise**
  - [4]: learning halfspaces with agnostic noise is computationally hard with unbounded ν, even if the unlabeled distribution is uniform.
  - [2]: computationally efficient and label-optimal algorithms that tolerates a noise level of $\nu = \Theta(\epsilon)$.

REFERENCES

[1] P. Awasthi, M.-F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. COLT 2016.

[2] S. Hanneke, V. Kanade, and L. Yang. Learning with a drifting target concept. ALT 2015.

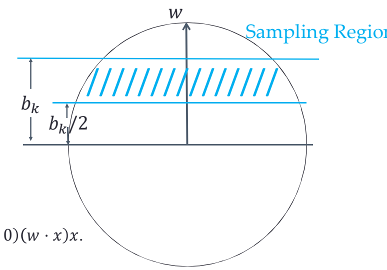
[3] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. COLT 2013.

[4] A. Klivans and P. Kothari. Embedding Hard Learning Problems Into Gaussian Space. APPROX/RANDOM 2014.

[5] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. COLT 2005.

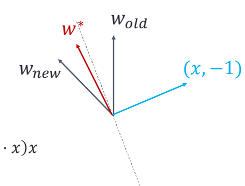
ALGORITHM

- **Input**: target error ϵ;
- **Output**: learned halfspace w.



- 1. Initialize w uniformly at random from the unit sphere.
- 2. Set sample schedule $m_k, b_k, k \ge 1$.
- 3. In phases $k = 1,2, \ldots, \lceil \log \frac{1}{\epsilon} \rceil$:
  - Repeat m_k times:
    - Sample x from $D_X|_{\{x: b_k/2 \le w \cdot x \le b_k\}}$ and query its label y;
    - Perform modified Perceptron update [5]: $w \leftarrow w - 2I(y\, w \cdot x \le 0)(w \cdot x)x$.
- 4. Return w.

- **Sample Schedule:**
  - (η-Bounded Noise): $m_k = \tilde{O}\left(\frac{d}{(1-2\eta)^2}\right)$, $b_k = \tilde{\Theta}\left(\frac{(1-2\eta)2^{-k}}{\sqrt{d}}\right)$;
  - (Adversarial Noise): $m_k = \tilde{O}(d)$, $b_k = \tilde{\Theta}\left(\frac{2^{-k}}{\sqrt{d}}\right)$.



- **The Modified Perceptron Update:** $w_{new} \leftarrow w_{old} - 2I(y\, w_{old} \cdot x \le 0)(w_{old} \cdot x)x$
  - Perceptron update with a careful tuning of step size
  - In the noiseless setting, the angle between w and w^* never increases; in the noisy setting, the angle never increases in expectation.

PERFORMANCE GUARANTEES

- **η-Bounded Noise**

Algorithm	Label Complexity	Time Complexity
[3]	$\tilde{O}\left(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon}\right)$	$\tilde{O}(\text{superpoly}(d, 1/\epsilon))$
[1]	$\tilde{O}\left(d^{(1-2\eta)^{-4}} \ln \frac{1}{\epsilon}\right)$	$\tilde{O}\left(d^{(1-2\eta)^{-4}} \ln \frac{1}{\epsilon}\right)$
This Work	$\tilde{O}\left(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon}\right)$	$\tilde{O}\left(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon}\right)$
Lower Bound	$\Omega\left(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon}\right)$	-

- Our algorithm achieves optimal label complexity and computational efficiency simultaneously

- **ν-Adversarial Noise**

Algorithm	Noise Tolerance	Label Complexity	Time Complexity
[2]	$\nu = \Theta(\epsilon)$	$\tilde{O}\left(d \ln \frac{1}{\epsilon}\right)$	$\tilde{O}\left(\frac{\text{poly}(d)}{\epsilon}\right)$
This Work	$\nu = \tilde{\Theta}(\epsilon)$	$\tilde{O}\left(d \ln \frac{1}{\epsilon}\right)$	$\tilde{O}\left(\frac{d^2}{\epsilon}\right)$
Lower Bound	$\nu = \Theta(\epsilon)$	$\Omega\left(d \ln \frac{1}{\epsilon}\right)$	-

- Our algorithm has a lower running time than the state-of-the-art algorithms

OPEN PROBLEMS

- Design efficient and label-optimal halfspace learning algorithms that:
  - adapt to unknown bounded noise parameter η
  - Work under broader unlabeled distributions, e.g. log-concave distributions
  - Work under weaker noise assumptions, e.g. Tsybakov noise condition