Proper Agnostic Learning of Functions of Halfspaces

Sergei Tikhonov (UT Austin) Arsen Vasilyan (Aarhus University)

NEURAL INFORMATION PROCESSING SYSTEMS

Institute for Foundations of MACHINE LEARNING

Introduction

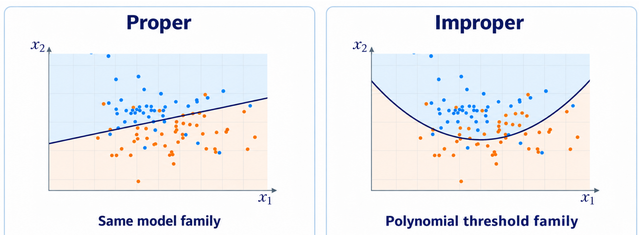
We study **distribution-specific agnostic learning**: covariates $x \sim \mathcal{N}(0, I_d)$, while the labels $y \in \{\pm 1\}$ are arbitrary.

Let $\mathcal{C}_K$ be the class of all Boolean functions of K halfspaces. **Our goal** is to learn $h \in \mathcal{C}_K$ satisfying, with high probability,

$$\Pr_{(x,y)\sim D}[h(x) \neq y] \leq \mathrm{OPT}_K + \varepsilon$$

where OPT_K is the best error in $\mathcal{C}_K$.

Proper learning returns a member of $\mathcal{C}_K$, while **improper learning** returns a function that is not necessarily in $\mathcal{C}_K$.



Why proper algorithm is better than improper:

Simpler models	Lower costs	Robustness
Same error guarantee	Storage and prediction	Less sensitive to perturbations

Results

Proper learning matches the **dimension dependence:**

Concept class	Our runtime	Best improper algorithm	SQ lower bound
Halfspaces	$d^{O(1/\varepsilon^2)} + S_H$	$d^{O(1/\varepsilon^2)}$	$d^{\Omega(1/\varepsilon^2)}$
Intersections of K halfspaces	$d^{O\left(\frac{\log K \log(1/\varepsilon)}{\varepsilon^2}\right)} + S_I$	$d^{O\left(\frac{\log K \log(1/\varepsilon)}{\varepsilon^2}\right)}$	$d^{\Omega\left(\frac{\log K}{\log\log K} + \frac{1}{\varepsilon^2}\right)}$
Boolean functions of K halfspaces	$d^{O\left(\frac{K^2 \log(1/\varepsilon)}{\varepsilon^2}\right)} + S_B$	$d^{O\left(\frac{K^2 \log(1/\varepsilon)}{\varepsilon^2}\right)}$	$d^{\Omega(K^2/\varepsilon^2)}$

Additional search cost (independent of d):

$$S_H = (1/\varepsilon)^{O(1/\varepsilon^{5/2})}$$

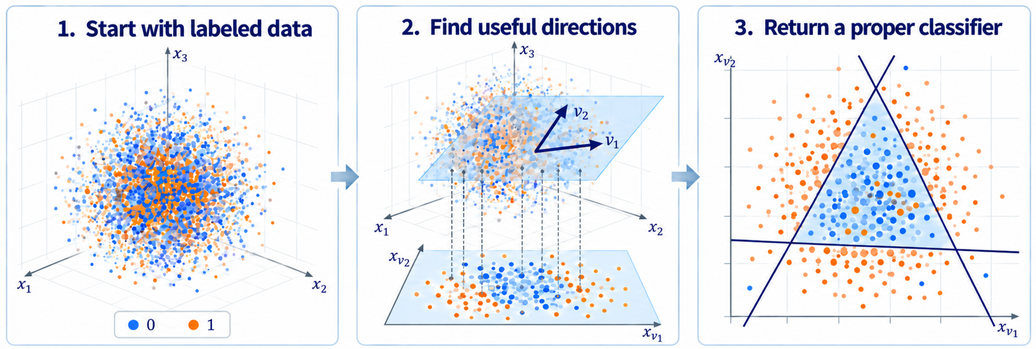
$$S_I = \left(\sqrt{\log K}/\varepsilon\right)^{O(K^2\sqrt{\log K}/\varepsilon^{5/2})}$$

$$S_B = (K/\varepsilon)^{O(K^3/\varepsilon^{5/2})}$$

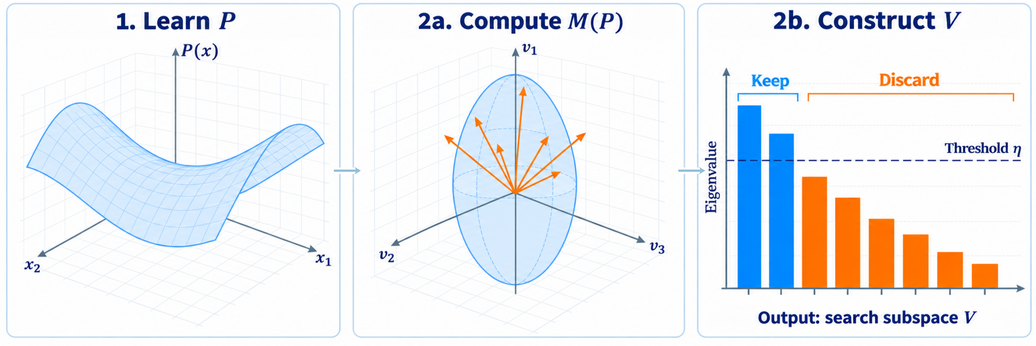
Improper: Pesenti et al. (2026).
SQ: Diakonikolas et al. (2021), Hsu et al. (2022).

Algorithmic Overview

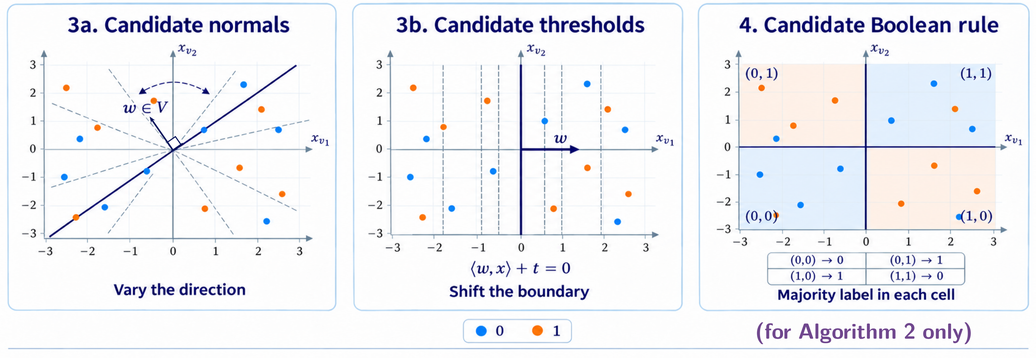
Idea: use improper classifier to find a low-dimensional subspace for near-optimal proper classifier.



Steps 1–2 of Algorithms: Learn the polynomial and find useful directions



Steps 3–4 of Algorithms: Search within V



(for Algorithm 2 only)

Algorithms

Objective

$$\mathcal{L}_{\mu,\nu}(P) = \underbrace{\mathbb{E}[\log(1 + e^{-yP(x)})]}_{\text{fit the labels}} + \underbrace{\mu\,\mathbb{E}[P(x)^2]}_{\text{ensure strong convexity}} + \underbrace{\nu\,\mathrm{tr}(M(P)^{1/2})}_{\text{control the search dimension}}.$$

Algorithm 1: Halfspaces

Goal: Return a halfspace with error $\mathrm{OPT} + \varepsilon$.

1. **Learn P.** Fit a low-degree polynomial P to the labels using regularized logistic loss.
2. **Construct V.** Compute the influence matrix
$$M(P) = \mathbb{E}[\nabla P(x) \nabla P(x)^\top].$$
Keep its eigenvectors with eigenvalues at least η; let V be their span.
3. **Search within V.** Try normals from a finite cover of the unit sphere in V and thresholds from a finite grid.
4. **Select the best candidate.** Draw one fresh labeled sample and evaluate every normal–threshold pair on it. Return the halfspace with the fewest mistakes.

Algorithm 2: Functions of K halfspaces

Goal: Return a Boolean function of K halfspaces with error $\mathrm{OPT}_K + \varepsilon$.

1. **Learn P.** Fit a low-degree polynomial P to the labels using regularized logistic loss, with parameters chosen for K and ε.
2. **Construct V.** Compute the influence matrix
$$M(P) = \mathbb{E}[\nabla P(x) \nabla P(x)^\top].$$
Keep its eigenvectors with eigenvalues at least η; let V be their span.
3. **Search within V.** Try normals from a finite cover of the unit sphere in V and thresholds from a finite grid.
4. **Select the best candidate.** Draw one fresh labeled sample S. For each joint choice of K normals and thresholds, group S by the outputs $(h_1(x), \ldots, h_K(x))$. Fit the Boolean rule by assigning each group its majority label. Compare the resulting classifiers on the same S and return the one with the fewest mistakes.

A key mathematical contribution is proving that the constructed low-dimensional subspace V supports a near-optimal proper classifier.

Future Work

- Extension to other covariate distributions D, e.g. uniform over the hypercube or log-concave marginals.
- Proper learning of other concept classes.
- Match the dimension dependence with search term.