Tester-Learners for Learning Halfspaces under Adversarial Label Noise (Literature Review)*

Deep Patel
Department of Computer Science
University of Wisconsin-Madison
dbpatel5@wisc.edu

Abstract

We will summarize two recent, NeurIPS'23 papers that propose different tester-learners for halfspaces. Paper 1 proposes tester-learners for halfspaces under the assumption of Gaussian marginal distributions and adversarial label noise. On the other hand, Paper 2 proposes tester-learners for halfspaces under adversarial label noise as well but, here, the marginal distribution can be *any* of a broad-class of distributions that satisfy a Poincaré inequality. This includes log-concave distributions.

It is known in the learning theory community that learning halfspaces in the agnostic (adversarial label noise) and distribution-free setting is computationally intractable. However, for learning halfspaces in the agnostic learning and distribution-specific settings (where the marginal distribution belongs to a particular family of distributions, say Gaussian or log-concave), the halfspace learner has an error of the form opt $+ \epsilon$, where opt denotes the optimal 0-1 error. The problem here is that the sample and time complexity is d^{1/ϵ^2} and the exponential dependence is tight!

1 Summary of Paper 1 – Efficient Testable Learning of Halfspaces with Adversarial Label Noise

Thus, authors in Paper 1 look at designing algorithms for efficient learning of halfspaces for the distribution-specific setting where marginal is Gaussian under the recently proposed framework of 'Tester-learner'. In particular, authors here propose a halfspace learning algorithm that achieves error $f(\text{opt}) + \epsilon$ (where $f(t) \to 0$ as $t \to 0$) and the time-complexity being poly(d/ϵ). Authors ultimately achieve a constant-factor approximation error guarantee.

The ideas in the paper can be summarised by first asking the questions shown below. Following that, we look at a summary of the additional criterion required and to be proved to achieve the aforementioned goal of efficient learning of halfspaces under adversarial label noise.

1. Come up with a *proper*, testable, weak agnostic learner w.r.t. the Gaussian distribution.
   - The tester-learner runs in polynomial time and either
     i.) reports that the $\mathbf{x}$-marginal is not $\mathcal{N}(0, I)$
     ii.) outputs a unit vector $\mathbf{w}$ with small constant distance to the target $\mathbf{v}^*$.
2. Question: How to output the vector $\mathbf{w}$?
   - The weak proper tester-learner first verifies that the given $\mathbf{x}$-marginal approximately matches constantly many low-degree moments with the standard Gaussian. If it does, return the vector defined by the degree-1 Chow parameters, i.e., $\mathbf{c} = \mathbb{E}_{(\mathbf{x},y)\sim D}[y\mathbf{x}]$

*This report is for the CS 880 (Fall'23) course

3. Question: But why is that enough?
    - Note that if $D_{\mathbf{x}}$ approximately matches its low-degree moments with the standard Gaussian, then the Chow parameters of any homogeneous LTF w.r.t $D_{\mathbf{x}}$ are close to its Chow parameters w.r.t. $\mathcal{N}(0, I)$. That is, for any homogeneous LTF we have $\mathbb{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[f(\mathbf{x})\mathbf{x}] \approx \mathbb{E}_{\mathbf{x}^*\sim\mathcal{N}(0,I)}[f(\mathbf{x}^*)\mathbf{x}^*]$.
    - Since the Chow vector of a homogeneous LTF w.r.t. Gaussian distribution is PARALLEL to its normal vector $\mathbf{v}^*$, we can show that the Chow vector of the LTF w.r.t. $D_{\mathbf{x}}$ will not be very far from $\mathbf{v}^*$ and will satisfy the (weak) learning guarantee of $\|\mathbf{c} - \mathbf{v}^*\|_2 \leq$ very small
4. Question: But what if there's label noise?
    - If $\mathbf{x}'$ has bounded second moments (a condition that we can efficiently test), we can robustly estimate the Chow parameters $\mathbb{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathbf{x}f(\mathbf{x})]$ with samples from D up to error $\sqrt{\text{opt}}$
    - Lemma 2.7 in the paper helps estimate the Chow parameters robustly under adversarial label noise. In particular, we estimate them coordinate-wise robustly using a 1-d median estimator.
5. Question: But we want a vector $\mathbf{w}$ that has a small 0-1 disagreement with the target halfspace $\mathbf{v}^*$. All we have obtained so far is a vector $\mathbf{w}$ that is close to $\mathbf{v}^*$ in ℓ_2-norm. If and why is that enough for minimizing the 0-1 disagreement?
    - If the underlying $\mathbf{x}$-marginal is a standard Gaussian and if we have $\|\mathbf{w} - \mathbf{v}^*\|_2 = \delta$, then we have
    $$\mathbb{P}_{\mathbf{x}\sim D_{\mathbf{x}}}[\text{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \text{sign}(\mathbf{v}^*\cdot\mathbf{x})] = \mathcal{O}(\delta)$$
    - Thus, achieving ℓ_2-distance $\mathcal{O}(\text{opt}) + \epsilon$ suffices.

Outline of the proof for efficient testable learning of halfspaces:

1.) Construct a TESTER which certifies that the probability of the diagreement region between two halfspaces whose defining vectors are close to each other is small under $D_{\mathbf{x}}$. Lemma 4.2 of [DKS18] tells us that
$$\mathbb{P}_{\mathbf{x}\sim D_{\mathbf{x}}}[h_{\mathbf{u}}(\mathbf{x}) \neq h_{\mathbf{v}}(\mathbf{x})] \leq \mathcal{O}(\|\mathbf{u} - \mathbf{v}\|_2) \tag{1}$$
   - Hence, learning homogeneous halfspaces under Gaussian marginals can often be reduced to approximately learnign the defining vector of some optimal halfspace h^*.
   - But this is no longer the case if $D_{\mathbf{x}}$ is an arbitrary distribution.
   - Thus, we will show that it's possible to "certify" whether some relationship similar to the one in Equation 1 above holds.
   - Specifically, we will prove that we will either be able to say with a high probability whether
     * "$\|\mathbf{w}-\mathbf{v}\|_2 \leq \eta \implies \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\text{sign}(\mathbf{v}\cdot\mathbf{x}) \neq \text{sign}(\mathbf{w}\cdot\mathbf{x})] \leq C\eta$" for some absolute constant $C > 1$
       · We first show "$\|\mathbf{w} - \mathbf{v}\|_2 \leq \eta \implies \mathbb{P}_{\mathbf{x}\sim\tilde{\mathcal{D}}}[\text{sign}(\mathbf{v}\cdot\mathbf{x}) \neq \text{sign}(\mathbf{w}\cdot\mathbf{x})] \leq C\eta$" for some absolute constant $C > 1$.
       · Proof of this crucially relies on the following: the empirical distribution of sampled data in the direction of $\mathbf{v}$ is close enough to the distribution of $\mathcal{N}(0, 1)$ and that in the orthogonal complement of $\mathbf{v}$ the empirical distribution has bounded covariance. (Both of these are steps that need to be tested by the tester.)
       · Finally, using the standard VC-theory bounds for linear threshold functions (dimension d) we can claim with high probability that
       $$\mathbb{P}_{\mathbf{x}\sim D_{\mathbf{x}}}[\text{sign}(\mathbf{v}\cdot\mathbf{x}) \neq \text{sign}(\mathbf{w}\cdot\mathbf{x})] \leq (C+1)\eta$$
     OR
     * $D_{\mathbf{x}}$ is NOT the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$.
     * Note: By its inherent distributional properties, if $D_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$, then, with high probability, this tester will yield $\mathbf{w}$ s.t. "$\|\mathbf{w} - \mathbf{v}\|_2 \leq \eta \implies \mathbb{P}_{\mathbf{x}\sim D_{\mathbf{x}}}[\text{sign}(\mathbf{v}\cdot\mathbf{x}) \neq \text{sign}(\mathbf{w}\cdot\mathbf{x})] \leq (C+1)\eta$"

2.) Construct a LEARNER as follows:

- **Assume for now:** That we have obtained a vector $\mathbf{v}$ s.t. $\|\mathbf{v} - \mathbf{v}^*\|_2 \leq \delta$ (where δ is "sufficiently small").
  - Given this, we transform the marginal $D_{\mathbf{x}}$ to a Gaussian whose covariance is $\mathcal{O}(\delta^2)$ in the direction of $\mathbf{v}$ and identity in the orthogonal (to $\mathbf{w}$) directions via the following rejection sampling procedure that was proposed in [DKS18]: "*Draw samples* $\mathbf{x} \sim D_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ *and accept them with probability* $e^{-(\mathbf{v}\cdot\mathbf{x})^2\cdot(\sigma^{-2}-1)}/2$ where $\mathbf{v} \in \mathbf{R}^d$ is a unit vector and $\sigma \in (0, 1)$". This rejection sampling will ensure that the distribution of $\mathbf{x}$ conditioned on acceptance in the rejection sampling procedure outlined above will be $D_{\mathbf{v},\sigma} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \mathbf{I} - (1 - \sigma^2)\mathbf{v}\mathbf{v}^T$.
- We will now learn a halfspace with normal $\mathbf{w}$ which achieves small constant error w.r.t. the new distribution $D_{\mathbf{v},\sigma}$.
- Use $\mathbf{w}$ and $\mathbf{v}$ to obtain a new halfspace that has small error w.r.t. the original distribution D.

3.) Combining Steps 1 and 2 listed above, authors obtain the final algorithm:

- Use moment-matching as described in Question 3 above to obtain a vector $\mathbf{v}$ that is sufficiently close to $\mathbf{v}^*$ in ℓ_2-distance. This is not enough as we get error $\mathcal{O}(\sqrt{\text{opt} + \eta})$ in $\mathcal{O}(d^{\tilde{\mathcal{O}}(1/\eta^2)})$ time.
- If we are unable to return such a $\mathbf{v}$, then we report that $D_{\mathbf{x}}$ is not $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and exit.
- If not, we 'boost' the weak learner obtained above by iteratively do the following:
  - Transform the marginal $D_{\mathbf{x}}$ to a Gaussian whose covariance is $\mathcal{O}(\delta^2)$ in the direction of $\mathbf{v}$ and identity in the orthogonal directions via the rejection sampling as outlined in Step-2 above.
  - This step is called "localizing to the learned halfspace".
  - Now, we apply the approximate moment-matching subroutine again to obtain a new, updated halfspace $\mathbf{v}^{(t)}$
  - Now, if the new halfspace $\mathbf{v}^{(t)}$ is sub-optimal, then, we can say that one of the following two happens thanks to our construction of tester-learner and the fact that $D_{\mathbf{x}}$ is $\mathcal{N}(\mathbf{0}, \mathbf{I})$:
    - $D_{\mathbf{x}}$ is NOT $\mathcal{N}(\mathbf{0}, \mathbf{I})$
      OR
    - Check if the acceptance probability for the rejection sampling procdure above is close to δ. If it is NOT, then report $D_{\mathbf{x}}$ is NOT $\mathcal{N}(\mathbf{0}, \mathbf{I})$.
      Or else, obtain a new halfspace $\mathbf{v}^{(t+1)}$ by localising (i.e. transforming the $D_{\mathbf{x}}$ marginal) to the neighbourhood of the decision boundary of current halfspace sign($\mathbf{v}^{(t)} \cdot \mathbf{x}$) as mentioned above and run the approximate moment-matching subroutine again.
- The above iterative procedure will take LOGARITHMIC number of rounds until the probability mass of the disagreement region is small between the halfspace $\mathbf{v}^{(t)}$ obtained at the end of above iterative routine and the optimal halfspace $\mathbf{v}^*$. Here, by 'close', we mean the ℓ_2-error between $\mathbf{v}^{(t)}$ and $\mathbf{v}^*$ is small.
- Since we have obtained logarithmically many such halfspaces in the iterative subroutine above, we choose the best halfspace $\hat{\mathbf{v}}^*$ among these by checking for smallest empirical test error using samples from the original distribution D.

2 Summary of Paper 2 – Tester-Learners for Halfspaces: Universal Algorithms

Outline of the porposed approach – that is intended to work for distributions that are "nice"[2][3] – is as follows:

[2] λ-nice [DKTZ20b]: that is, the density function's projection on any 2D subspace has concentration, anti-concentration and anti-anti-concentration properties where the upper and lower bounds are λ, $1/\lambda$

[3] γ-Poincaré: A distribution over $\mathbb{R}^d$ is γ-Poincaré, if var($f(x)$) $\leq \gamma \cdot \mathbb{E}[\|\nabla f(x)\|_2^2]$ for any differentiable $f : \mathbb{R}^d \to \mathbb{R}$

- Use non-convex SGD as done in prior works [DKTZ20a, DKTZ20b] for robust learning of half-spaces under adversarial label noise along with a set of testers that check if distribution is "nice" enough to proceed or not.
- Authors use a smoothened-version of ramp loss as a non-convex surrogate, L_σ, for the 0-1 loss, namely a smooth version of the ramp loss.
- First, compute a stationary point $\mathbf{w}$ of the loss L_σ.
- The tester-learner (specifically, Tester #2 as defined later in the Section) will now check distributional properties of the unknown marginal D that will ensure that $\measuredangle(\mathbf{w}, \mathbf{w}^*)$ is small. (This is done because it can be shown that the probability of disagreement is upper-bounded in terms of $\measuredangle(\mathbf{w}, \mathbf{w}^*)$). In particular, this is shown by ensuring that any $\mathbf{w}$ having large gradient norm ($\|\nabla L_\sigma(\mathbf{w})\|$) must be having large $\measuredangle(\mathbf{w}, \mathbf{w}^*)$.
  - This aforementioned condition can be further reduced to a test of a anti-concentration property: Let $\mathbf{v}$ denote any unit vector orthogonal to $\mathbf{w}$, and let D_T denote D restricted to the σ-band $T = \{\mathbf{x} | |\langle \mathbf{w}, \mathbf{x}\rangle| \leq \sigma\}$ (where the width σ is carefully selected according to certain constraints). Then the property we need is that $\mathbb{P}_{\mathbf{x}\sim D_T}[|\langle \mathbf{v}, \mathbf{x}\rangle| \geq \Theta(1)] \geq \Theta(1)$.
- If Tester #2 accepts the data S[4], we are assured that the stationary point $\mathbf{w}$ has small $\measuredangle(\mathbf{w}, \mathbf{w}^*)$.
- Run Tester #1 and if it returns "ACCEPT"[5], we are assured that $\mathbb{P}_{\mathbf{x}\sim D}[\text{sign}(\langle \mathbf{w}, \mathbf{x}\rangle \neq \langle \mathbf{w}^*, \mathbf{x}\rangle)] \leq \mathcal{O}(\measuredangle(\mathbf{w}, \mathbf{w}^*))$.
- Running this procedure for a list of $\mathbf{w}$'s obtained by running (Projected) SGD for minimizing the non-convex surrogate L_σ ultimately yields a set of $\mathbf{w}$ out of which we pick the one with the smallest empirical test error as our final learnt halfspace normal.

With that outline in mind, we can summarize the author's proposed testers as follows:

1. `(Universal) Tester #1`: It helps ensures that if the candidate halfspace $\mathbf{w}$ that is close to the optimal one $\mathbf{w}^*$ in terms of $\measuredangle(\mathbf{w}, \mathbf{w}^*)$; then, for a set of samples S drawn from any of the "nice" distributions [6], $\mathbf{w}$ has low test error. i.e., it's also an approximate empirical risk minimizer. In other words, we have a tester that satisfies the following:
   i.) If the tester accepts S, then $\forall \mathbf{w}' \in \mathbb{R}^d$ ($\|\mathbf{w}'\|_2 = 1$) satisfying $\measuredangle(\mathbf{w}', \mathbf{w}) \leq \theta$, we have
   $$\mathbb{P}_{x\sim S}[\text{sign}(\langle \mathbf{w}', \mathbf{x}\rangle \neq \langle \mathbf{w}, \mathbf{x}\rangle)] \leq C \cdot \theta \cdot \lambda^C$$
   ii.) If the marginal is λ-nice, the tester accepts S with probability $1 - \delta$.
   - Proof sketch for why this tester works: Unlike in Paper-1 – where authors obtain the bound on disagreement probability by considering probability of falling in rectangular slabs that are orthogonal to the target vector ($\mathbf{w}$ here) and utilising the anti-concentration properties of the Gaussian distribution and that it has light tails – authors here propose constructing a $(d-1)\times(d-1)$ covariance matrix M_S of the projection of the datapoints in S onto the $(d-1)$-dimensional subspace orthogonal to $\mathbf{w}$. Tester rejects S if $\|M_S\|_{\text{op}}$ is not small[7]. Otherwise, by construction of M_S, we can then show that if $D_{\mathbf{x}}$ is λ-nice, the tester accepts S with high probability and if a tester accepts S, then we have
   $$\mathbb{P}_{\mathbf{x}\sim S}[\text{sign}(\langle \mathbf{w}', \mathbf{x}\rangle \neq \langle \mathbf{w}, \mathbf{x}\rangle)] \leq C_1\theta\lambda^{C_1} + 4\|M_S\|_{\text{op}} \leq 5C_1\theta\lambda^{C_1}$$
2. `(Universal) Tester #2`: For a given vector $\mathbf{w}$, a set of samples S and an unknown vector $\mathbf{v} \perp \mathbf{w}$, we want to check if the conditional distribution[8] is weakly anti-concentrated in every direction. Tester #2 helps us test for this. Basically, we want the tester to satisfy the following:

[4] acceptance w.h.p. if underlying distribution is "nice"

[5] acceptance is w.h.p if distribution is "nice"

[6] λ-nice and γ-Poincaré as described above in the paper outline

[7] This can be efficiently checked with pre-existing testers as long as the elements of the matrix have bounded moments; this is true for our M_S

[8] conditioned on the samples of S falling withing a slab orthogonal to the target vector $\mathbf{w}$

i.) If the tester accepts S, then for any $\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1$ with $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ we have (for appropriate constants $\lambda, C, \gamma, \sigma$)

$$\mathbb{P}_{\mathbf{x} \sim S} \left[|\mathbf{v}, \mathbf{x}| \geq \frac{1}{C\lambda^C} | |\mathbf{w}, \mathbf{x}| \leq \sigma \right] \geq \frac{1}{C\lambda^C \gamma^4}$$

ii.) If D is γ-Poincaré and λ-nice, then the tester accepts S with high probability.

- Proof sketch for why Tester #2 works: This check relies on utilizing Paley-Zygmund inquality for any non-negative random variable Z:

$$\mathbb{P}[Z > \mathbb{E}[Z]/2] \geq \frac{1}{4} \cdot \frac{(\mathbb{E}[Z])^2}{\mathbb{E}[Z^2]}$$

where we will consider the Z following the distribution of $\langle \mathbf{v}, \mathbf{x} \rangle^2$ conditioned on $\langle \mathbf{w}, \mathbf{x} \rangle \leq \sigma$ for some unit, orthogonal vectors $\mathbf{w}, \mathbf{v}$ and constant $\sigma > 0$ and $\mathbf{x}$ whose distribution is λ-nice. To check if the RHS above is bounded away from zero, we essentially need to test whether $\mathbb{E}[Z^2]$ is bounded uniformly over all $\mathbf{v}$ s.t. $\mathbf{v} \perp \mathbf{w}$. This is equivalent to testing boundedness of fourth-order empirical moment as $Z^2 = \langle \mathbf{v}, \mathbf{x} \rangle^4$ where $\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1$. This bound can be obtained by existing results from [KS17] which uses a Sum-of-Squares relaxation to obtain upper bound of $C\gamma^4$ for $\mathbb{E}[\langle \mathbf{v}, \mathbf{x} \rangle^4]$ ($\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1$) when underlying distribution D is γ-Poincaré. Thus, the tester will do the following:

  - Similar to the case of Tester #1, we construct a $(d-1) \times (d-1)$ covariance matrix M_S of the projection of only those datapoints in S that are within a σ-band around $\mathbf{w}$ onto the $(d-1)$-dimensional subspace orthogonal to $\mathbf{w}$. REJECT if the minimum singular value of M_S is "too small". Otherwise, continue.
  - Now, for the same set, S', of projection of only those datapoints in S that are within a σ-band around $\mathbf{w}$ onto the $(d-1)$-dimensional subspace orthogonal to $\mathbf{w}$, find $\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{E}[\langle \mathbf{v}, \mathbf{x} \rangle^4]$. REJECT the set S if this value is $> C'\gamma^4$. Or else ACCEPT S and RETURN ONLY S' – the set of samples that falls inside the σ-band around $\mathbf{w}$.
  - If D is γ-Poincaré, then with high probability the max value obtained above is bounded above by $C\gamma^4$ with high probability. In order for the tester to accept S with high probability we can set $C = C'$.

3 Notes on the differences between approaches of Paper 1 and Paper 2

Differences between the two papers' approaches that I have noticed so far:

1. Bound on disagreement probability for vectors $\mathbf{w}$ & $\mathbf{v}$: The disagreement probability is upper-bounded by the ℓ_2-distance, $\|\mathbf{w} - \mathbf{v}\|_2$ in Paper-1 and the angle between the two vectors, $\measuredangle(\mathbf{w}, \mathbf{v})$ in Paper-2. The bound in terms of angle between the two vectors is a more-generally-applicable bound since it doesn't assume marginal to be a Gaussian.
2. Concentration, anti-concentration, anti-anti-concentration: Paper-1 is able to use anti-concentration and light-tails property of the marginal distribution which is assumed to be Gaussian. Paper-2, on the other hand, wants to deal with λ-nice and γ-Poincaré distributions where these constants (that help determine the concentration, anti-concentration, anti-anti-concentration properties) are unknown. Thus, authors are required to construct (Universal) Tester #2 check for certifiable hypercontractivity in order to test for anti-concentration explicitly. The authors rely on an existing result from [KS17] for certifiable hypercontractivity via Sum-of-Squares method.
3. Obtaining halfspace weight vector updates: Paper 1 uses multiple 'soft'-localization steps to modify the Gaussian marginal and use these modified distributions to obtain the list of candidate halfspace vectors. On the other hand, in case of Paper-2, authors use samples from the underlying distribution alone and obtain the list of candidate halfspace vectors by looking at the various iterative updates of (Projected) SGD on the non-convex surrogate loss L_σ[9].

[9] following which, Testers #2 and Testers #1 are used to prune the list of candidates; and after that, with the remaining candidates we search the one with smallest empirical test set errors

References

[DKS18] Diakonikolas, I., Kane, D. M., & Stewart, A. (2018). Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (pp. 1061-1073).

[DKTZ20a] Diakonikolas, Ilias, Vasilis Kontonis, Christos Tzamos, & Nikos Zarifis. (2020). Learning halfspaces with massart noise under structured distributions. *In Conference on Learning Theory*, pp. 1486-1513. PMLR.

[DKTZ20b] Diakonikolas, I., Kontonis, V., Tzamos, C., & Zarifis, N. (2020). Non-convex SGD learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33, 18540-18549.

[KS17] Kothari, P. K., & Steinhardt, J. (2017). Better agnostic clustering via relaxed tensor norms. *arXiv preprint arXiv:1711.07465*.