Semi-Verified PAC Learning from the Crowd

Shiwei Zeng
Stevens Institute of Technology

Jie Shen
Stevens Institute of Technology

Abstract
We study the problem of crowdsourced PAC learning of threshold functions. This is a challenging problem and only recently have query-efficient algorithms been established under the assumption that a noticeable fraction of the workers are perfect. In this work, we investigate a more challenging case where the majority may behave adversarially and the rest behave as the Massart noise – a significant generalization of the perfectness assumption. We show that under the semi-verified model of Charikar et al. (2017), where we have (limited) access to a trusted oracle who always returns correct annotations, it is possible to PAC learn the underlying hypothesis class with a manageable amount of label queries. Moreover, we show that the labeling cost can be drastically mitigated via the more easily obtained comparison queries. Orthogonal to recent developments in semi-verified or list-decodable learning that crucially rely on data distributional assumptions, our PAC guarantee holds by exploring the wisdom of the crowd.

1 INTRODUCTION
Efficient and robust learning of threshold functions is arguably one of the most important problems in machine learning, and has been broadly investigated under different label noise models, such as the random classification noise (Angluin and Laird, 1987), Massart noise (Sloan 1988, Massart and Nédélec 2006), Tsybakov noise (Tsybakov 2004), and agnostic noise (Haussler 1992, Kearns et al. 1992). In recent years, under certain distributional assumptions on the unlabeled data, a variety of computationally efficient algorithms have been established (Awasthi et al. 2017a, Zhang et al. 2020, Diakonikolas et al. 2020c,b) under the standard probably approximately correct (PAC) learning model of Valiant (1984), where there is a single adversary who generates all the noisy labels.

However, the standard PAC learning model of Valiant (1984) does not fully characterize the process of data collection and learning in modern machine learning applications. In fact, when constructing a large-scale data set, researchers often appeal to a crowdsourcing platform to hire multiple crowd workers for annotation, with the hope of obtaining a set of high-quality labels. To be more concrete, a crucial feature of crowdsourced learning is that for each instance \( x \), the learner has a discretion to collect a set of \( k \) labels, denoted by \( y_i(x), i = 1, \ldots, k \), from a large pool of heterogeneous crowd workers. Here, each single worker in the pool may behave as a certain type of label noise models in the standard PAC model, and some workers can even collude with others to decide which instances to corrupt among those assigned to them. A common practice to address such noisy annotations is to aggregate the labels via majority vote, which turns out to be a successful remedy provided that the majority are correct (Dekel and Shamir 2009, Vaughan 2017, Awasthi et al. 2017b, Zeng and Shen 2022a).

In this paper, nevertheless, we consider learning from an extremely noisy pool of workers, where the majority might be adversarial, under which using majority voting will make the result even noisier, and more seriously, rendering PAC learning impossible. This problem, in addition to being theoretically interesting and has been broadly studied under the non-crowdsourcing setting (Charikar et al. 2017), also finds a variety of real-world applications (Steinhardt et al., 2016, Prelec et al., 2017, Meister and Valiant 2018). For example, it is observed that when confronted with the question “is Philadelphia the capital of Pennsylvania?” more than 60% of the respondents endorse the incorrect answer with high confidence (Prelec et al. 2017). This raises a pressing question:

If the majority of the crowd might be malicious, can we still achieve the PAC guarantees? If the answer is positive, what is the price for it?

\[ \text{In the standard PAC learning model, the label of an instance is always persistent.} \]
1.1 Warmup: Reduction to Standard PAC Learning

We begin our exploration by considering a quite fundamental problem setup: the underlying hypothesis class is the one of homogeneous halfspaces among which there exists a hypothesis that correctly classifies all instances \( \text{Rosenblatt} [1958] \) \( \text{Valiant} [1984] \). For each instance we assign one worker for labeling.

Formally, let \( \mathcal{X} \subset \mathbb{R}^d \) be the instance space, \( \mathcal{Y} = \{-1, 1\} \) be the label space, and \( D \) be the marginal distribution on \( \mathcal{X} \). The class of homogeneous halfspaces is given by \( \mathcal{H}_\text{hs} = \{ h : x \mapsto \text{sign}(w \cdot x), \ w \in \mathbb{R}^d \} \), and the true label of \( x \) is given by \( h^*(x) \) for some \( h^* \in \mathcal{H}_\text{hs} \). The error rate of any hypothesis \( h \) is defined as \( \text{err}_D(h) = \Pr_{x \sim D}(h(x) \neq h^*(x)) \). The learner has access to a large pool of crowd workers who provide labels, and we denote by \( P^t \) the uniform distribution on it. Given \( x \in \mathcal{X} \), the quantity \( \eta_t(x) := \Pr_x(h_t(x) \neq h^*(x)) \) characterizes how likely the learner gets an incorrect label when having worker \( t \) annotate the instance \( x \), where the probability is taken over the internal random bits of the worker.

To further simplify the problem, we assume that a \((1 - \alpha)\)-fraction of workers may be adversarial but the remaining \( \alpha \)-fraction are perfect in the sense that they always provide correct labels across the domain.

**Assumption 1.** With probability \( 1 - \alpha \) over the draw of a worker \( t \), \( t \) is adversarial, i.e. \( \eta_t(x) \) can be an arbitrary function; otherwise, \( t \) is perfect, i.e. \( \eta_t(x) = 0, \forall x \in \mathcal{X} \).

We aim to approach the problem by reducing it to some well-studied standard PAC learning model, such that there exists a learner with non-trivial PAC guarantees. One promising solution is the agnostic learner, which works under the scenario where an adversary can corrupt an arbitrary fraction of the labels in an adversarial manner (Hausdorff 1992; Kearns et al. 1992; Valiant 1992; Kearns et al. 1993; Kalai et al. 2008; Diakonikolas et al. 2021). We give the following theorem using results from Kalai et al. (2008); Diakonikolas et al. (2021).

**Theorem 1.** Suppose Assumption 1 is satisfied, and further assume the marginal distribution \( D \) is standard Gaussian. Given a set of \( \Theta(d^{\text{poly}}(1/\epsilon)) \) independent instances from \( D \), each of which is labeled by one worker, there exists an algorithm that runs in time \( O((d/\epsilon)^{\text{poly}}(1/\epsilon)) \) and learns \( \mathcal{H}_\text{hs} \) by returning a hypothesis with error rate \( \leq (1 - \alpha) + \epsilon \).

**Remark 1.** There are a few limitations of such method. First, the error rate is non-vanishing (as \( \alpha < 1/2 \)), though this is the best one can hope for without extra information. Second, such reduction scheme only works for the very special case of learning homogeneous halfspaces with respect to Gaussian marginals, and with quite high computational complexity and sample complexity. Last, the success of the reduction hinges on the strong assumption that an \( \alpha \)-fraction of the workers are perfect.

In this paper, we show that it is possible to break all of the above barriers by leveraging the wisdom of the crowd and by querying a trusted oracle with a small number of times.

1.2 Main Results

Observe that using a single worker to annotate an instance is not a common practice. Thus, from now on, we consider the more realistic scenario where the learner would assign an instance to multiple workers. An immediate caveat in our setting is that, the majority vote becomes even noisier. As a remedy, we follow a recent research line of Steinhardt et al. (2016); Awasthi et al. (2017b); Meister and Valiant (2018) and assume that there exists a trusted oracle \( O_T \) who always provides the correct annotation for any instance that we query on, and we call a query to \( O_T \) as a verified query. To make it practical, we must restrict the number of verified queries the learner can make. We now consider a more general hypothesis class \( \mathcal{H} \) of threshold functions, i.e. \( \mathcal{H} = \{ h : x \mapsto \text{sign}(f(x)), \ f \in \mathcal{F} \} \) where \( \mathcal{F} = \{ f : \mathcal{X} \rightarrow \mathbb{R} \} \); e.g. \( f(x) = w \cdot x \) recovers the class of linear functions and hence \( \mathcal{H} \) is the class of halfspaces. Note that the target classifier \( h^* \) can be written as \( h^*(x) = \text{sign}(f^*(x)) \) for some \( f^* \in \mathcal{F} \). We also drop Assumption 1 on the existence of perfect workers, and instead assume the following:

**Assumption 2.** With probability \( 1 - \alpha \) over the draw of a worker \( t \), \( t \) is adversarial, i.e. \( \eta_t(x) \) can be an arbitrary function; otherwise, \( t \) is Massart, i.e. \( \eta_t(x) \leq \eta, \forall x \in \mathcal{X} \) for some noise rate \( \eta \in [0, 1/2] \).

The non-adversarial workers now act as the Massart noise (Sloan 1988; Massart and Nédélec 2006) which is a challenging semi-random model that has attracted a surge of recent research interests under the standard PAC model; see e.g. Diakonikolas et al. (2019); Zhang et al. (2020); Diakonikolas et al. (2020c). Our model on the crowd workers is thus a mixture of adversarial and semi-random, with the adversarial consisting of the majority. The model itself is already new and challenging; to the best of our knowledge, no known results have been established under neither standard nor crowdsourced setting.

Now the principle of our algorithmic design is three-fold: 1) offering a PAC guarantee, i.e. finding a hypothesis \( h \) such that \( \text{err}_D(h) \leq \epsilon \); 2) the number of verified queries, \( m_V \), is few; and 3) the number of queries to the crowd, \( m_L \), is moderate. In particular, we hope that the overhead of the algorithm, \( \Lambda_L \), defined as the averaged number of crowd queries on a given instance, behaves as a constant.

We note that the last (implicit) condition is that the hypothesis class \( \mathcal{H} \) has finite VC-dimension \( d \), which is a minimum requirement for PAC learnability even without noise (Kearns and Vazirani 1994; Anthony and Bartlett 1999).

**Assumption 3.** There exists a computationally efficient algorithm \( A_H \) satisfying the following: for any \( \epsilon, \delta \in (0, 1) \), \( A_H \) draws \( m_{\epsilon, \delta} \) correctly labeled instances and learns \( \mathcal{H} \) by
Formally, we consider the class $\mathcal{H}$ of threshold functions. For any $(x, x') \in \mathcal{X} \times \mathcal{X}$, the underlying comparison function is given by $Z^*(x, x') = \text{sign}(f^*(x) - f^*(x'))$. Denote by $\mathcal{P}^C$ the uniform distribution over the pool of workers who will provide comparison tags, and a worker $i$ is defined by a comparison function $Z_i : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$.

**Assumption 4.** When providing comparison tags, a $(1 - \beta)$ fraction of the workers may behave adversarially, and the rest $\beta \leq 1/2$ fraction are such that $\Pr(Z_i(x, x') \neq Z^*(x, x') \mid (x, x')) \leq \eta$ for some $\eta \in [0, 1/2)$.

Let $m_C$ be the total number of comparison queries to the crowd, $\Lambda_C$ the comparison overhead, and $m_V$ be the total number of verified label and comparison queries. Our second set of main results is a *label-efficient* crowdsourced PAC algorithm.

**Theorem 3** (Theorem 1, informal). Suppose that Assumptions 2 and 3 are satisfied, and $\eta < \frac{\alpha}{16}$. Given any $\epsilon, \delta \in (0, 1)$, there exists a polynomial-time algorithm (Algorithm 3) that PAC learns $\mathcal{H}$. In addition, $\Lambda_C = O_{\alpha}(1)$ and $m_V = O_{\alpha}(1)$.

**Remark 5.** This is the first label-efficient algorithm that works in the semi-verified crowdsourced learning setting where adversarial workers can form a strong majority. Moreover, both the overheads and the number of verified data are independent of the sample size, meaning that the algorithm is query-efficient even for large-scale learning problems. The label efficiency follows from that $\Lambda_C$ may tend to zero; alternatively, it also follows from the fact that $m_L = \tilde{O}(\log \frac{1}{\epsilon})$, where in label-only learning, this quantity scales as $\tilde{O}\left(\frac{1}{\epsilon}\right)$ (Theorem 4). We note that a label-efficient crowdsourced PAC algorithm was obtained in [Zeng and Shen (2022a)] but their analysis only works for $\alpha, \beta > \frac{1}{2}$.

**Remark 6.** The merit of comparison-based PAC learning algorithms is to tradeoff the labeling and comparison complexity ([Xu et al. (2017)]), rather than obtaining a lower overall query complexity. Indeed, the latter is information-theoretically impossible in general: a lower bound of $\Omega(d + 1/\epsilon)$ was shown in [Kane et al. (2017)].

### 1.3 Overview of Techniques

**Win-win pruning of adversarial workers.** When the adversarial workers form a majority in the crowd, a natural idea is to query the trusted oracle so as to identify and prune them. The main question here lies in how to select appropriate instances as testing cases such that a significant fraction of the adversarial workers can be pruned every time the trusted oracle is called. Our main contribution here is a *win-win argument* even when no perfect worker is present: the adversarial workers can either collude and provide incorrect annotations, for which the algorithm can identify and prune a noticeable fraction of them; or they are forced to provide correct annotations, for which we can be sure that the majority are correct. The key insight lies in a nat-
ural threshold of the majority size (the fraction of workers that agree with majority) being \((1 - (1 - \eta)\alpha)\), as on all accounts the adversarial workers cannot form a majority larger than that. This is also our condition for selecting the instances as testing cases, which generalizes the analysis in Awasthi et al. (2017b). See Section 3.1.

**Query and label-efficient semi-verified filtering.** To obtain an algorithm that is both label and query-efficient (Theorem 3), we need to resolve a more serious issue of majority voting over a constant number of workers; if the sample size increases, this majority voting can be incorrect with high probability. If we were to increase the constant number, the query complexity grows beyond what we can afford. To this end, we introduce a new testing stage to the comparison-equipped algorithm which draws a moderate set of informative samples and then queries comparison tags. This stage would give us enough information about how likely the crowd workers would form a strong majority, i.e. the majority size \(\geq \eta\). If we were to increase the constant size, this majority voting can be incorrect with high probability. We say a crowdsourced PAC learner is label-efficient if \(\Lambda L = o(1)\). Our goal is to design a polynomial-time PAC learning algorithm that is both query-efficient and label-efficient; in addition, we aim to control the number of calls to the trusted expert as otherwise the problem is trivial.

Given an input \(z\) and a set of workers \(W\), we define \(\text{Maj}_{W}(z)\) to be the majority vote over the annotations from the workers in \(W\), where \(z\) can either be an instance or a pair of instances. We denote by \(\text{Maj-size}_{W}(z)\) the fraction of the workers in \(W\) that agree with \(\text{Maj}_{W}(z)\), which is referred to as the empirical majority size. Similarly, we can consider the population version \(\text{Maj-size}_{\rho}(z) = \max \{\Pr_{z \sim \rho}(h_{1}(z) = 1), \Pr_{z \sim \rho}(h_{2}(z) = -1)\}\). Specifically, let \(h_{1}, h_{2}, h_{3}\) be three classifiers. The function \(\text{Maj}(h_{1}, h_{2}, h_{3})\) maps any instance \(z\) to the label \(y = \text{Maj}(h_{1}, h_{2}, h_{3})(z)\). Given a distribution \(\mathcal{P}\) over workers and an annotated set \(S\) (either a set with labels or one with comparison tags), denote by \(\mathcal{P}_{S}\) the distribution \(\mathcal{P}\) conditioned on workers that agreed with the annotations in \(S\); this will be useful to prune the adversarial workers. Given any region \(R \subseteq \mathcal{X}\), denote by \(D[R]\) the density of instances \(x \in R\) under distribution \(D\).

### 2 RELATED WORKS

Crowdsourcing has been broadly utilized as a tool for data annotation with provable algorithms (Vaughan, 2017). More in line with this work is crowdsourcing in the presence of highly noisy workers, i.e. the majority can be adversarial. Steinhardt et al. (2016); Meister and Valiant (2018) considered recovery of true values of a finite data set and Awasthi et al. (2017b) examined PAC learnability which aligns more closely with our work. Our algorithm shares the merits with Awasthi et al. (2017b) but we develop a novel analysis to address the label corruption from Massart workers, and we carefully leverage comparison queries to achieve label efficiency. We emphasize that crowdsourced PAC learning was also investigated in Zeng and Shen (2022a), but their algorithm fails immediately when the majority are incorrect.

The semi-verified learning model was recently proposed in Charikar et al. (2017). This model is also known as list-decodable learning where a learner is allowed to output a finite list of hypotheses among which at least one is guaranteed to be close to the true hypothesis. Such model has been broadly studied under the non-crowdsourcing setting for various problems such as mean estimation (Kothari and Steinhardt, 2017; Diakonikolas et al., 2020a; Zeng and Shen, 2022b), learning mixtures of Gaussian (Diakonikolas et al., 2018), linear regression (Karmalkar et al.,...
3 MAIN ALGORITHMS AND GUARANTEES

In this section, we present our main algorithms, one that uses label-only queries (Algorithm 1) and the other that combines label and comparison queries (Algorithm 3). In the label-only learning model, we will build upon the algorithm of Awasthi et al. (2017b) and show how a new pruning method (specifically, a new pruning threshold) can be leveraged into their algorithm to obtain Theorem 2. In the comparison-based learning model, we will extend the algorithm of Zeng and Shen (2022a) by presenting an ensemble of new algorithmic subroutines that delicately combine verified queries, label queries, and comparison queries.

First of all, we introduce the core idea of these algorithms, assuming for now that the adversarial workers only consisting a minority. We emphasize that obtaining distribution-free PAC guarantees under label noise in the standard PAC model is highly nontrivial, and among prevalent noise models, the only known computationally efficient PAC learner works for the very basic random classification noise (Blum et al. 1996), while for Massart noise or adversarial noise, hardness results have been established (Diaconis and Kane 2022). Guruswami and Raghavendra (2006). Yet, it turns out that as far as the majority is correct, by using majority vote, it is possible to PAC learn the underlying hypothesis class without distributional assumptions, where a key step is to gather sufficient labels for each instance so that with overwhelming probability, all the aggregated labels are correct, for which one can apply the realizable learner stated in Assumption 3. However, a naive label-and-train approach, as observed in Awasthi et al. (2017b), leads to a labeling overhead that grows with the sample size. Therefore, they proposed an elegant algorithm that interleave data annotation and learning. In particular, there are three phases in the main algorithm: in phase 1, it aims to learn a hypothesis \( h_1 \) with \( \text{err}_D(h_1) \leq \frac{1}{2} \sqrt{\epsilon} \); in phase 2, it identifies a distribution \( D_2 \) where the performance of \( h_1 \) is almost comparable to random guess, and learns another \( h_2 \) with \( \text{err}_{D_2}(h_2) \leq \frac{1}{2} \sqrt{\epsilon} \); in phase 3, it concentrates on a distribution \( D_3 \) where \( h_1 \) and \( h_2 \) disagree, and learns the last hypothesis \( h_3 \) with \( \text{err}_{D_3}(h_3) \leq \frac{1}{2} \sqrt{\epsilon} \). The final output of the algorithm, \( \hat{h} \), is a majority voting function over these three hypotheses. This is essentially the boosting algorithm of Schapire (1990) which enjoys a guarantee that \( \text{err}_D(\hat{h}) \leq \epsilon \).

Our algorithm in the semi-verified setting will be built upon the above, and the primary idea is to design query-efficient pruning methods to remove most adversarial workers, such that the Massart ones form the majority. To be more detailed, we will incorporate new pruning methods in all three phases to learn the desired hypotheses, while in some cases, additional technical efforts are needed to ensure the correctness of our results. We remark that the difference to obtain Theorem 3 compared to Theorem 2 is on how to use label queries in an economic way through the introduction of comparison queries.

3.1 Label-Only Algorithm: Algorithm 1

Now let us consider that the majority are adversarial. A similar semi-verified crowdsourcing model is also studied in Awasthi et al. (2017b) but their analysis, in particular, the success of their paradigm to prune adversarial workers, crucially relies on Assumption 4 to ensure that the perfect workers will always be retained during pruning, and thus eventually form the majority. This is because perfect workers make no mistake, so that none of them will be removed during the pruning step and hence the fraction of the perfect ones will increase and eventually form the majority. We here leverage a key observation that, even when some perfect workers are removed, it is still possible to increase the fraction \( \alpha \), so long as the fraction of the adversarial workers being removed is always strictly larger than that of the perfect workers. Hence, we can relax the assumption of perfect workers to Massart workers (Assumption 2).

The main algorithm is given in Algorithm 1 which contains three phases corresponding to our previous discussion. The most important component of the main algorithm is PRUNE-AND-LABEL (Algorithm 2). For each given instance \( x_t \), in Step 3 of PRUNE-AND-LABEL, we check whether the current fraction of Massart workers, \( \alpha \), has formed the majority (note that we will update \( \alpha \), thus it has potential to increase), or the size of workers that agree with the majority vote is greater than a threshold. If either condition is satisfied, we take the majority vote as the label of \( x_t \). What is subtle here is our condition on the majority size, which is essentially our pruning rule. To see why it
makes sense, observe that at the beginning, the adversarial workers form a $(1 - \alpha)$ fraction, while the Massart workers form the $\alpha$ fraction among which, in expectation, at most $\alpha\eta$ will return incorrect labels. Therefore, the maximum number of incorrect labels will be $1 - \alpha + \alpha\eta = 1 - (1 - \eta)\alpha$. So in expectation, if we see that the majority size is greater, it must be the case that the Massart workers who are committed to providing correct labels contribute to the voting, hence the majority vote is correct. It is easy to see that there is a constant factor difference in our algorithm, which is due to the conversion from expected value to high probability argument. When Step 3 of PRUNE-AND-LABEL is not satisfied, it is likely the case that majority vote gives incorrect labels, they will likely be pruned away. Hence, the fraction among which, in expectation, at most \((1 - \eta)\alpha_i\) fraction of Massart workers in the pool \(\forall i \geq 0\). Assume the algorithm works properly and \(\alpha_i \geq \alpha_{i-1}\), \(\forall i\). Then, we require that \(\eta < \frac{\alpha_0}{16}\) by telescoping, where \(\alpha\) is the fraction of Massart workers in the crowd before running the algorithm.

**Restarting times.** Note that each time the main algorithm restarts, it removes at least \((1 - \eta)\alpha_i\) fraction of the workers, with at most an \(\eta\alpha\) fraction being Massart. Therefore, denote by \(N_i\) the total number of workers in the pool after \(i\) restartings of the main algorithm.

\[
\alpha_i \geq \frac{\alpha_0 N_0 \cdot (1 - \eta)^i}{N_0 - \frac{(1 - \eta)\alpha_0}{8} \cdot N_0 - \cdots - \frac{(1 - \eta)\alpha_{i-1}}{8} \cdot N_1 - 1}.
\]

By simple deductions on the above inequality, we can show that after \(O\left(\log \frac{1}{\delta\alpha} \left(\frac{\rho_0}{\alpha - \alpha_0}\right)\right)\) updates, \(\alpha_i\) would surpass any given constant \(C \in [0.7, 1]\). In other words, the main algorithm is restarted at most \(O(T_{\alpha, \eta})\) times, using one verified label in each restart. Hence, the number of verified labels required from the trusted oracle \(O_T\) is also bounded by \(O(T_{\alpha, \eta})\).

Our main result for Algorithm 1 is summarized in the following; the proof is deferred to Appendix B.

**Theorem 4.** Suppose that Assumptions 2 and 3 are satisfied, and \(\eta < \frac{\alpha}{16}\). Denote \(T_{\alpha, \eta} = \log \frac{1}{\delta\alpha} \left(\frac{\rho_0}{\alpha - \alpha_0}\right)\). The following holds with probability \(1 - \delta\). There exists an algorithm that runs in \(O(\text{poly}(d, \frac{1}{\delta}, T_{\alpha, \eta}, \frac{1}{\alpha}))\) time and learns \(\mathcal{H}\) by returning a hypothesis \(h : \mathcal{X} \rightarrow \mathcal{Y}\) with \(err_D(h) \leq \epsilon\).

In addition, \(m_L = O\left(d T_{\alpha, \eta} \cdot \log \frac{1}{\delta}\right)\), and \(m_V = O(T_{\alpha, \eta})\), where \(\hat{O}\) hides the logarithmic factors. Therefore, \(\Lambda_L = O\left(T_{\alpha, \eta} \cdot \log^2(T_{\alpha, \eta})\right)\) where \(c \in (0, \log^{-2} d)\). In particular, when \(\alpha\) is constant, \(\Lambda_L = O(1)\) and \(m_V = O(1)\).

**Remark 7.** Recall that [Awasthi et al., 2017b] studied the problem with the existence of perfect workers. Their work requires a number of \(O\left(\frac{1}{\alpha}\right)\) verified label from \(O_T\), while in our analysis, the amount is \(O(T_{\alpha, \eta}) = O\left(\log \frac{1}{\delta\alpha} \left(\frac{\rho_0}{\alpha - \alpha_0}\right)\right)\). It is not hard to show that \(\lim_{\eta \rightarrow 0} T_{\alpha, \eta} = \frac{1}{\alpha}\). Thus, our re-
**Algorithm 2 Prune-and-Label**

**Require:** An instance set $S$, parameters $\alpha$, $\delta_1$.

**Ensure:** A labeled set $\hat{S}$.

1. $k_0 \leftarrow \frac{1}{2(1-\eta)\alpha} \cdot \log \frac{1000m}{\delta_1}$
2. Draw a set of $k_0$ workers $W'_{L}$ from $PL$.
3. if $\alpha \geq 0.7$ or Maj-size$_{W'_{L}}(x_i) \geq 1 - \frac{(1-\eta)\alpha}{4}$ then
4. $y_i \leftarrow \text{Maj}_{W'_{L}}(x_i)$.
5. else
6. Get a verified label $y_i^\ast \leftarrow h^\ast(x_i)$. Restart Algorithm 1 with $PL \leftarrow PL_{[x_i,y_i^\ast]}$ and $\alpha \leftarrow \frac{(1-\eta)\alpha}{1-(1-\eta)\alpha}$.
7. end if

return $\hat{S}$.

result is a new analysis on the structure of the Massart workers which guides an algorithmic refinement and derives a generalization of [Awashio et al. (2017b)].

### 3.2 Comparison-Equipped Algorithm: Algorithm 3

In this section, the goal is to largely reduce the labeling cost using noisy pairwise comparisons from the crowd. We note that there exists a label-efficient algorithm in [Zeng and Shen (2022a)] that achieves the above tradeoff, but only works with $\alpha, \beta > \frac{1}{2}$. We will use their algorithm as the starting point and we elaborate on the technical challenges and our design details.

First, since there exists $1 - \beta \geq \frac{1}{2}$ fraction of adversarial workers in the crowd, a natural idea is to add a pruning step when collecting comparison tags from the crowd. Similar to the analysis in Section 5.1, it is easy to show that if a condition $\epsilon < \frac{\beta}{10}$ is satisfied, the win-win pruning scheme can also apply to $PC$ and increase the fraction of Massart workers who provide pairwise comparisons, i.e. $\beta$. Furthermore, it requires at most $O(T_{\beta, \eta}) = O\left(\log \frac{1}{\epsilon} \left(\frac{\beta - \eta}{\beta - \eta}\right)\right)$ verified comparison tags from $\Theta_T$ to increase $\beta$ to any specified constant $C \in [0.7, 1]$. This leads to our design of algorithm prune-and-label (see Section 3.2.1).

A more serious issue with $\beta < \frac{1}{2}$ roots in the filtering process of [Zeng and Shen (2022a)], which roughly speaking, will perform majority voting over a constant number of workers. In this regard, an empirical majority size may deviate far from the population one, thus the pruning fails. We also note that if we were to increase this constant number, we would suffer high query complexity. We thus develop a new testing stage that samples a large enough set of test cases and query the crowd for their majority size. If less than an $\epsilon$ fraction of the test cases have a small majority vote size, we can show that the total fraction of such test cases in the underlying distribution $D$ is smaller than $\epsilon$, hence the filtering process correctly filters all the desired instances except for an $\epsilon$ fraction; this is sufficient to establish a robust version of the performance guarantee. This algorithmic design is presented in **SEMI-VERIFIED-FILTER** (see Section 3.2.2).

As a result, we obtain a label-efficient semi-verified crowdsourced learning algorithm as presented in Algorithm 3.

#### 3.2.1 prune-and-label

The comparison-based PAC learning algorithms of [Xu et al. (2017); Zeng and Shen (2022a)] aid label efficiency with a “compare-and-label” approach, which first sorts all the instances in a set $S$ and then query the label of the instances during binary search. We fortify such method by adding certain check points for both the “compare” and “label” steps to ensure that the majority vote is always correct (with high probability). We summarize the performance guarantees of Algorithms 1 and 6 below.

**Lemma 5.** Consider Algorithm 1. Given any set of instances $S$, with probability $1 - \frac{\epsilon^4}{m}$, the algorithm either increases the fraction of Massart workers from $\beta$ to $\frac{(1-\eta)\beta}{1-(1-\eta)\beta}$, or correctly sorts $S$.

**Lemma 6.** Assume $\hat{S}$ is correctly
Algorithm 4 Prune-Compare-and-Label
Require: An instance set $S$, parameters $\alpha$, $\beta$, $\delta_1$.
Ensure: A sorted and labeled set $\tilde{S}$.
1: $\tilde{S} \leftarrow$ Semi-Verified-QuickSort($S, \beta, \delta_1$).
2: $\overline{S} \leftarrow$ Semi-Verified-BinarySearch($\tilde{S}, \alpha, \delta_1$).
   return $\overline{S}$.

Algorithm 5 Semi-Verified-QuickSort
Require: An instance set $S = \{x_i\}_{i=1}^n$, $\beta$, $\delta_1$.
Ensure: A sorted list $\tilde{S}$.
1: $k_1 \leftarrow \frac{\log \frac{n \cdot \log n}{\delta_1}}{\frac{3006 \cdot \log n}{\delta_1}}$.
2: Apply randomized Quicksort to $S$: for each pair $(x, x')$ being compared, draw a set of $k_1$ workers $\mathcal{W}_C$ from $\mathcal{P}_C$ and do the following:
3: if $\beta \geq 0.7$ or Maj-size$\mathcal{W}_C (x, x') \geq 1 - \frac{(1-\eta)\beta}{4}$ then
4: Use Maj$\mathcal{W}_C (x, x')$ as the comparison tag.
5: else
6: Get a verified comparison tag $z^* \leftarrow Z^*(x, x')$.
   Restart Algorithm 1 with $\mathcal{P}_C \leftarrow \mathcal{P}_C((x, x', z^*))$ and $\beta \leftarrow \frac{(1-\eta)\beta}{1 - \frac{(1-\eta)\beta}{4}}$.
7: end if
end for

Proposition 7. Consider Algorithm 2. If it does not restart and $|S| \geq (\frac{1}{\alpha})^{1/1000}$, then with probability $1 - \delta_1$, it correctly sorts and labels all the instances in $S$ with $O\left(\frac{\log |S|}{\alpha} \cdot \log \log |S|\right)$ label queries and $O(\frac{1}{\beta^2} \cdot |S| \cdot \log^2 |S|)$ comparison tags.

In light of Theorem 23 (i.e., boosting [Schapire, 1990]), the two weak hypotheses $h_1$ and $h_3$ are trained on distributions $D$ and $D_3$ which are not hard to construct if $h_2$ is given: $D$ is the original marginal distribution, and $D_3$ can be obtained by rejection sampling. By Assumption 5, drawing a set of $\Theta(n_r \tau_n x)$ instances on each distribution respectively and feeding them to $A_H$ gives the following corollary.

Corollary 8. With probability $1 - 4\epsilon^2$, $err_D(h_1) \leq \frac{4\epsilon^2}{\beta}$. With probability $1 - 4\epsilon^2$, $err_{D_3}(h_3) \leq \frac{4\epsilon^2}{\beta}$.

3.2.2 Semi-Verified-Filter

A core component of Algorithm 5 is a correct filtering process. [Awasthi et al., 2017b] achieved this by executing a Phase 0 that examines the crowd labels for a randomly drawn instance set. However, this approach does not apply in the comparison-equipped setting because it relies on two support instances $x^-, x^+$ as reference to identify the informative instances. To be more concrete, the filtering scheme incorporates a similar idea to that of “compare-and-label” in order to achieve label efficiency. Given a large enough sample set $S$, the goal is to identify two support instances $x^-, x^+$ that are close enough to the “threshold” where the labels in $S$ shift. By spending a few label queries to infer the labels of $x^-, x^+$, it helps to filter a large number of instances in $S$ by querying pairwise comparisons. This is because $\forall x \cdot [h^*(x^-) = -1, x < x^-] \implies h^*(x) = -1$ (similar holds for $x^+$). The pairwise comparisons of any $x$ with respect to $x^-, x^+$ become exceptionally informative.

We hence build on the non-semi-verified filter of [Zeng and Shen, 2022a] and integrate a new component, Test, that ensures to recover a distribution $D'$ that is close enough to $D_2$ even in presence of the adversarial workers. In particular, we are concerned with two regions, $R_1$ and $R_2$

$$R_1 := \{x : \text{Maj-size}_{PC}(x, x^-) \leq 1 - \frac{\beta}{2}\},$$

$$R_2 := \{x : \text{Maj-size}_{PC}(x, x^+) \leq 1 - \frac{\beta}{2}\},$$

(3.1)

which include any instance $x$ that would form a low-confidence pair with either support instance $x^-$ or $x^+$. Observe that both $x^-$, $x^+$ are fixed. If the probability mass of $R_1 \cup R_2$ under distribution $D$, denoted by $D[R_1 \cup R_2]$, is smaller than $\epsilon$, it means the filter uses $x^-, x^+$ to correctly identify the instances except for the $\epsilon$ fraction in $R_1 \cup R_2$. Meanwhile, Test works on a smaller set of instances than
Now we have three hypotheses $h_1, h_2, h_3$ that satisfy the requirement for boosting; this suffices to produce a hypothesis $h$ with the desired PAC guarantee. We further note that each pruning step helps to remove a significant fraction of the adversarial workers but only a small fraction of the Massart workers. Hence, our main theorem, Theorem 11 follows. See Appendix C for the full proof.

**Theorem 11 (Comparison-equipped learning).** Suppose that Assumptions 2, 3, and 4 are satisfied. Given any $\alpha, \beta, \epsilon, \delta \in (0, 1)$, assume $\eta = \frac{\alpha}{\min(\alpha, \beta)}$. Denote $T_{\alpha, \eta} = \log \frac{\log \frac{1}{\alpha \eta}}{\beta (\frac{\alpha}{\min(\alpha, \beta)}}$ and $T_{\beta, \eta} = \log \frac{\beta \eta}{\beta (\frac{\alpha}{\min(\alpha, \beta)}}. The following holds with probability $1 - \delta$. There exists an algorithm (Algorithm 7) that runs in $O(\text{poly}(d, \frac{1}{\epsilon}, T_{\alpha, \eta}, T_{\beta, \eta}, \frac{1}{\delta}, \frac{1}{\epsilon}))$ time and learns $H$ by returning a hypothesis $h : X \rightarrow Y$ with error $D(h) \leq \epsilon$. In addition, $m_L = T_{2\alpha, \eta} (T_{\alpha, \eta} + T_{\beta, \eta}) \cdot \widetilde{O} \left( \log \frac{d T_{2\alpha, \eta}}{\epsilon (T_{\alpha, \eta} + T_{\beta, \eta})} \right)$, $m_C = T_{2\beta, \eta} \cdot \widetilde{O} \left( \frac{1}{d (T_{\alpha, \eta} + T_{\beta, \eta})} \right)$, and $m_V = T_{\alpha, \eta} + T_{\beta, \eta}$. Therefore, $\Lambda_L = \widetilde{O}(T_{2\alpha, \eta} (T_{\alpha, \eta} + T_{\beta, \eta}))$ and $\Lambda_C = \widetilde{O}(T_{2\beta, \eta} (T_{\alpha, \eta} + T_{\beta, \eta}))$ when $\epsilon \in (0, (\log d)^{-1})$. In particular, when $\alpha$ and $\beta$ are constants, we have $\Lambda_L = o(1)$, $\Lambda_C = O(1)$, and $m_V = O(1)$.

## 4 CONCLUSION

In this paper, we studied the problem of semi-verified learning threshold functions from the crowd. We showed that when the majority of the crowd workers are adversarial and the rest behave as Massart noise, it is still possible to achieve PAC guarantees without distributional assumptions. In addition, our algorithms enjoy both query and label-efficiency, and run in polynomial time.

**Acknowledgements**

We thank the anonymous reviewers and meta-reviewer for valuable feedback. This work is supported by NSF-IIS-1948133 and by the startup funding from Stevens Institute of Technology.

**References**

Arpit Agarwal, Shivani Agarwal, Sanjeev Khanna, and Prathamesh Patil. Rank aggregation from pairwise comparisons in the presence of adversarial corruptions.
Semi-Verified PAC Learning from the Crowd


A OMITTED PROOF OF THEOREM 1

Proof. Recall that in the agnostic model, the adversary can flip an arbitrary \((1 - \alpha)\)-fraction of the labels in an adversarial manner while retaining the marginal distribution on \(X\). It is shown in [Kalai et al., 2008; Diakonikolas et al., 2021] that if \(\mathcal{H}_{hs}\) is the class of homogeneous halfspaces and \(D\) is the standard Gaussian, then there exists a learning algorithm which takes as input \(O(d/poly(1/\epsilon))\) samples generated in such a way, runs in time \(O((d/\epsilon)poly(1/\epsilon))\), and returns a hypothesis with error rate less than \((1 - \alpha) + \epsilon\) with overwhelming probability.

Now we note that in our crowdsourcing model, each instance is assigned one worker which is randomly chosen from the pool and may be adversarial with probability \(1 - \alpha\), while otherwise he is perfect. Therefore, the labels gathered in such a way satisfy the condition of agnostic noise model, and we apply the above results directly which gives Theorem 1. \(\square\)

B OMITTED PROOF FOR THEOREM 4

The key idea of the proof is the following: we show that with the carefully designed pruning approach, the adversarial workers either corrupt the labels, under which a noticeable fraction of them will be pruned away and Algorithm 1 will be restarted with the cleaner pool of workers, or they provide labels in such a way that the majority vote is correct and Algorithm 1 is exactly mirroring the easier case where the majority is correct. Our goal here is to show that in the former case, the algorithm must make significant progress such that it obtains a pool of workers with most of them, say 70%, that are not Massart. Taking the union bound over the labeling of \(L\) instances, we have that with probability at least \(1 - \delta_1\), for all \((x, y) \in \mathcal{S}\), \(y = h^*(x)\).

Proof. Recall that Algorithm \(\mathcal{A}\) queries a set \(\mathcal{W}_L\) of \(k_0 = \frac{1}{2(1-\eta)2\alpha^2} \log \frac{1000m}{\delta_1}\) workers for labeling each instance. Since \(1 - \eta \in (\frac{1}{2}, 1]\), by applying the Chernoff bound, we have

\[
\Pr \left( |\text{Maj-size}_{\mathcal{W}_L}(x) - \text{Maj-size}_{\mathcal{W}_L}(x)| \leq \frac{(1-\eta)\alpha}{8} \right) \leq 2 \cdot e^{-2k_0 \left( \frac{\delta}{8} \right)^2} \leq \frac{\delta_1}{1000m}.
\]

Taking the union bound over the labeling of \(\log n\) instances, we have that with probability at least \(1 - \delta_1\), for all \(x \in \mathcal{S}\), \(|\text{Maj-size}_{\mathcal{W}_L}(x) - \text{Maj-size}_{\mathcal{W}_L}(x)| \leq \frac{(1-\eta)\alpha}{8}\). Thus, if \(\text{Maj-size}_{\mathcal{W}_L}(x) \geq 1 - \frac{(1-\eta)\alpha}{4}\), we have the population majority size \(\text{Maj-size}_{\mathcal{W}_L}(x) \geq 1 - 3\left(\frac{1-\eta)\alpha}{8}\right) > 1 - (1 - \eta)\alpha\) with high probability, indicating that the majority voting is correct, i.e. \(\text{Maj}_{\mathcal{W}_L}(x) = h^*(x)\). As a result, if the algorithm finishes without restarting, we are sure that \(\forall (x, y) \in \mathcal{S}, y = h^*(x)\).

On the flip side, if the algorithm prunes due to \(\text{Maj-size}_{\mathcal{W}_L}(x) < 1 - \frac{(1-\eta)\alpha}{4}\), we have \(\text{Maj-size}_{\mathcal{W}_L}(x) < 1 - \frac{1-\eta)\alpha}{8}\) with high probability, indicating that if we query a verified label for \(x\), it is guaranteed that more than a \(\frac{(1-\eta)\alpha}{8}\)-fraction of the workers can be pruned (among which only \(\eta\alpha\) could be Massart workers), and \(\alpha \leftarrow \frac{(1-\eta)\alpha}{1-\frac{1-\eta)\alpha}{8}}\). \(\square\)

Proposition 13. Consider Algorithm \(\mathcal{A}\). If it does not restart, then with probability \(1 - \delta_1\), it correctly labels all the instances in \(S\) with \(O(m \log m \alpha)\) label queries.

Proof. Due to Lemma 12, for any input set \(S\) of size \(m\), the total number of crowd labels is \(k_0 \cdot O(|S|) = O(\frac{m \log m}{\alpha^2})\). \(\square\)

Lemma 14. Consider Algorithm \(\mathcal{A}\) \(\forall \alpha \in (0, 1]\), if \(\eta < \frac{\alpha}{12}\), then with probability \(1 - \delta\), Algorithm \(\mathcal{A}\) will restart \(O(T_\alpha, \eta)\) times. In addition, the required number of verified labels from \(O_T\) is \(m_V = O(T_\alpha, \eta)\).

Proof. By Lemma 12, each time we get a verified label \(y^*\) from \(O_T\), we prune a \(\frac{(1-\eta)\alpha}{8}\)-fraction of the workers with high probability. Since it is not guaranteed that all Massart workers would give the correct labels, some of them would be pruned. Given that the pool is large, with probability \(1 - \delta / \tau\), at most an \(\eta\) fraction of the Massart workers will be pruned.
Denote by $\alpha_i$ the fraction of Massart workers and $N_i$ the total number of workers in the pool after $i$ prunings. Clearly, we have $\alpha_0 = \alpha$. Note that

$$\alpha_i N_i \geq \alpha_{i-1} N_{i-1} \cdot (1 - \eta) \geq \cdots \geq \alpha_0 N_0 \cdot (1 - \eta)^i.$$

Then, we have

$$\alpha K \geq \frac{\alpha_0 N_0 \cdot (1 - \eta)^K}{N_0 - \frac{(1-\eta)\alpha_0}{8} \cdot N_0 - \frac{(1-\eta)\alpha_1}{8} \cdot N_1 - \cdots - \frac{(1-\eta)\alpha_{K-1}}{8} \cdot N_{K-1}}.$$

From Eq. (B.1), to pick a sufficient number of verified labels $K$ that increases $\alpha_K$ to any given constant $C \in [0.7, 1]$, it suffices to choose

$$K \geq \log_{1-\eta} \frac{C\alpha_0(1-\eta) - 8C\eta}{C\alpha(1-\eta)^2 - 8\eta\alpha_0}.$$ 

Due to $C \in [0.7, 1]$, $1 - \eta \in (\frac{1}{2}, 1]$, $\alpha_0 < 1$, it requires at most

$$O\left(\frac{\alpha_0 - \eta\alpha_0}{\alpha_0 - \eta}\right) = O\left(T(\alpha_0, \eta)\right)$$

verified label from the trusted oracle $O_T$ for $\alpha_K$ to surpass 0.7. By union bound, with probability at least $1 - \delta$, it requires at most $O(T_{\alpha, \eta})$ verified labels to increase the fraction of Massart workers such that they form a strong majority, where the algorithm no longer prunes and must return the desired hypothesis. Therefore, the main algorithm only restarts for $O(T_{\alpha, \eta})$ times.

**Proof of Theorem 4.** Given that in all three phases, Algorithm I gathers a sample of size $O(m_{\alpha, \lambda})$ and is correctly labeled with high probability. In addition, the algorithm is restarted at most $O(T_{\alpha, \eta})$ times (Lemma 14). Then, by union bound, Assumption 2, Lemma 24 and Theorem 23 we conclude that with probability at least $1 - \delta$, Algorithm I returns $h$ such that $\text{err}_D(h) \leq \epsilon$.

The number of required verified labels from $O_T$ is $m_{\omega} = O(T_{\alpha, \eta})$ by Lemma 14. It remains to show the label complexity and label overhead. From Lemma 4.9 of [Awasthi et al., 2017] and Proposition 13, it requires

$$m_{L} = O\left(T_{\alpha, \eta} \cdot m_{\alpha, \delta} + T_{\alpha, \eta} \cdot \frac{1}{\epsilon} \cdot \log \frac{1}{\epsilon} \cdot \log \frac{1}{\epsilon\delta^2} + T_{\alpha, \eta} \cdot \frac{m_{\sqrt{\alpha\delta^2}} \log m_{\sqrt{\alpha\delta^2}}}{\alpha^2}\right)$$

label queries from the crowd, where we recall that $\delta' = \frac{2}{8\sqrt{\alpha_{\eta}}} \alpha_{\eta}$ as defined in Algorithm I. This combined with the definition of $m_{\alpha, \delta}$ (see Eq. (1.1)) gives the announced label complexity of $O\left(\frac{dT_{\alpha, \eta}}{\alpha_{\eta}^2} \cdot \log \frac{1}{\delta}\right)$.

Finally, the calculation of labeling overhead follows from the definition, i.e.

$$\Lambda_{L} = \frac{m_{L}}{m_{\alpha, \delta}} = O\left(T_{\alpha, \eta} \log T_{\alpha, \eta} + T_{\alpha, \eta} \log^2 T_{\alpha, \eta} + T_{\alpha, \eta} \cdot \frac{m_{\sqrt{\alpha\delta^2}} \log m_{\sqrt{\alpha\delta^2}}}{\alpha^2 \cdot m_{\alpha, \delta}}\right) = O\left(T_{\alpha, \eta} \log^2 T_{\alpha, \eta} + \frac{T_{\alpha, \eta}}{\alpha^2} \sqrt{\epsilon} \log d \cdot \log^2 T_{\alpha, \eta}\right).$$
Thus, when \( \epsilon < \log^{-2} d \), the labeling overhead reads as \( O\left(\frac{T_{\alpha, \eta}}{\epsilon^2} \log^2 T_{\alpha, \eta}\right) \), which is upper bounded by a constant as far as \( \alpha \) is a small constant. The proof is complete.

C OMITTED PROOF FOR THEOREM[11]

Lemma 15 (Restatement of Lemma[5]). Consider Algorithm[5]. Given any set of instances \( S \), with probability \( 1 - \frac{2\delta}{3} \), the algorithm either increases the fraction of Massart workers from \( \beta \) to \( \frac{1 - (1 - \eta)\beta}{1 - (1 - \eta)\beta} \), or correctly sorts \( S \).

Proof. Given that \( k_1 = \frac{1}{2\beta} \cdot \log \left(\frac{1000\alpha n \cdot \log n}{\delta_1}\right) \) in Algorithm[5]. By the guarantee of celebrated algorithm RANDOMIZED QUICKSORT, with probability \( 1 - \frac{1}{n^2} \), the total number of pairs that needs to be compared is \( (c + 2)n \log n \). By the sample size of input set \( S \) and setting \( c = 1000 \), with probability \( 1 - \frac{1}{n^2} \cdot (c + 2)n \log n = 1002n \log n \). In addition, similar to the proof of[6], we have \( \text{Maj-size}_{pc}(x, x') = \mathbb{E}[\text{Maj-size}_{\mathcal{W}C}(x, x')] \) for any \( (x, x') \), and

\[
\Pr \left[ \bigcup_{l=1}^{1002n \log n} \left| \text{Maj-size}_{pc}(x, x')_l - \text{Maj-size}_{\mathcal{W}C}(x, x')_l \right| \leq \frac{(1 - \eta)\beta}{8} \right] \leq \frac{\delta_1}{3}.
\]

In other words, with probability \( 1 - \frac{\delta}{3} \), the following is guaranteed. If \( \text{Maj-size}_{\mathcal{W}C}(x, x') \geq 1 - \frac{(1 - \eta)\beta}{4} \), we have the population majority size \( \text{Maj-size}_{pc}(x, x') \geq 1 - \frac{3(1 - \eta)\beta}{4} > 1 - (1 - \eta)\beta \), indicating that the majority voting is correct, i.e. \( \text{Maj}_{\mathcal{W}C}(x, x') = Z^*_{\beta}(x, x') \). As a result, if the algorithm never restarts, \( \hat{S} \) is correctly sorted according to \( Z^* \). On the other hand, if the algorithm prunes because \( \text{Maj-size}_{\mathcal{W}C}(x, x') < 1 - \frac{(1 - \eta)\beta}{4} \), we have \( \text{Maj-size}_{pc}(x, x') \geq 1 - \frac{(1 - \eta)\beta}{8} \) with high probability, meaning that by querying a verified comparison for \( (x, x') \) we can remove at least a \( \frac{(1 - \eta)\beta}{8} \)-fraction of the workers (with at most \( \eta \beta \) being Massart) and update \( \beta \leftarrow \frac{(1 - \eta)\beta}{1 - (1 - \eta)\beta} \).

Lemma 16 (Restatement of Lemma[6]). Consider Algorithm[6]. Assume \( \hat{S} \) is correctly sorted. With probability \( 1 - \frac{\delta}{3} \), the algorithm either increases the fraction of Massart workers to \( \frac{1 - (1 - \eta)\beta}{1 - (1 - \eta)\beta} \), or correctly labels \( \hat{S} \), namely, for all \( (x, y) \in \hat{S} \), \( y = h^*(x) \).

Proof. The proof follows the same pipeline as that of Lemma[12] with parameter \( k_2 \) for SEMI-VERIFIED-BINARYSEARCH algorithm and the fact that binary search only queries labels on at most \( \log n \) instances.

Proposition 17 (Restatement of Proposition[7]). Consider Algorithm[6]. If it does not restart and \( |S| \geq \left(\frac{1}{\delta^2}\right)^{1/1000} \), then with probability \( 1 - \delta_1 \), it correctly sorts and labels all the instances in \( S \) with \( O\left(\frac{\log n}{\alpha^2} \cdot \log \log n\right) \) label queries and \( O\left(\frac{1}{\alpha^2} \cdot n \log^2 n\right) \) comparison tags.

Proof. Due to Lemma[15] and[16], for any input set \( S \) of size \( n \), the total number of crowd labels is

\[
k_2 \cdot O(\log |S|) = O\left(\frac{1}{\alpha^2} \log \left(\frac{\log n}{\delta_1}\right) \cdot \log n\right) = O\left(\frac{\log n}{\alpha^2} \cdot \log \log n\right),
\]

and the total number of crowd comparison tags is

\[
k_1 \cdot O(|S| \log |S|) = O\left(\frac{1}{\beta^2} \cdot \log \frac{n \cdot \log n}{\delta_1} \cdot n \log n\right) = O\left(\frac{1}{\beta^2} \cdot n \log^2 n\right).
\]

Corollary 18 (Restatement of Corollary[8]). With probability \( 1 - \frac{\delta}{3} \), \( err_{D}(h_1) \leq \frac{\sqrt{2}}{2} \). With probability \( 1 - \frac{\delta}{3} \), \( err_{D_3}(h_3) \leq \frac{\sqrt{2}}{2} \).

Proof. By drawing a sample \( S_1 \) of \( \sqrt{n/2, \delta'/6} \) from \( D \) and labeling it by PRUNE-COMPARE-AND-LABEL, with probability at least \( 1 - \frac{\delta}{3} \), \( S_1 \) is labeled correctly according to \( h^* \) (Proposition[17]). Furthermore, we note that \( n_{\sqrt{7/2, \delta'}/6} \geq m_{\sqrt{7/2, \delta'}/6} \). By Assumption[3] \( err_{D}(h_1) \leq \frac{\sqrt{2}}{2} \), with probability \( 1 - \frac{\delta}{3} \).
Similarly, in Phase 3 of Algorithm 3, by rejection sampling we can successfully sample a set $S_3$ of $n^{\sqrt{2}\beta/6}$ from $D_3$. Again by Proposition 7 and Assumption 3, $\text{err}_{D_3}(h_3) \leq \frac{\sqrt{2}}{\beta}$ with probability $1 - \frac{\delta'}{\beta}$.

Lemma 19. \( \forall \alpha, \beta \in (0, 1] \), with probability $1 - \delta$, Algorithm 3 will restart $O(T_{\alpha, \eta} + T_{\beta, \eta})$ times. In addition, the required number of verified labels from $O_{\beta}$ is $O(T_{\alpha, \eta})$ and that of the verified comparison tags is $O(T_{\beta, \eta})$.

Proof. The deduction for removing adversarial workers from the pool of workers who provide comparison tags is similar to the one who provide labels. By union bound, with probability at least $1 - \delta$, it requires at most $O(T_{\alpha, \eta})$ verified labels and $O(T_{\beta, \eta})$ verified comparison tags to increase the fraction of Massart workers such that they form a strong majority, where the algorithm no longer prunes and must return the desired hypothesis. Therefore, the main algorithm only restarts for $O(T_{\alpha, \eta} + T_{\beta, \eta})$ times.

C.1 Performance Guarantee of Phase 2

Lemma 20 (Restatement of Lemma 9). If Algorithm 8 terminates without restarting Algorithm 3, we have the probability mass $D[R_1 \cup R_2] \leq \frac{\epsilon}{4}$ with probability $1 - \frac{\delta}{4}$. In addition, the comparison complexity is $O\left(\frac{1}{\epsilon^2} \log \left(\frac{1}{\delta} \log \left(\frac{1}{\epsilon\beta}\right)\right)\right)$.

Proof. Recall that by definition, $R_1 := \{x : \text{Maj-size}_{PC}(x, x^-) \leq 1 - \frac{\delta}{4}\}$ and $R_2 := \{x : \text{Maj-size}_{PC}(x, x^+) \leq 1 - \frac{\delta}{4}\}$. Without loss of generality, we prove the lemma for $x^-$. We remark that the following guarantee holds for $x^+$ as well.

For any instance $x \in S_r$, let

$$\ell_i = \begin{cases} 1, & \text{if worker } i \text{ agrees with } \text{Maj}_{PC}(x, x^-), \\ 0, & \text{otherwise}. \end{cases}$$

Since we query a set $W$ of $N = \frac{32}{\beta^2} \cdot \log \left(\frac{32S_r}{\delta^2}\right)$ workers from the crowd to compare $(x, x^-)$. Note that $\text{Maj-size}_{PC}(x, x^-) = E[\text{Maj-size}_W(x, x^-)] = \mu$, which is also the probability that $\ell_i = 1$. Therefore, by Hoeffding’s inequality we have

$$\Pr \left[ \text{Maj-size}_W(x, x^-) \geq 1 - \frac{(1 - \eta)\beta}{4} \right] \leq \Pr \left[ \left| \text{Maj-size}_W(x, x^-) - \text{Maj-size}_{PC}(x, x^-) \right| \geq \frac{(1 - \eta)\beta}{8} \right]$$

$$= \Pr \left[ \left| \frac{1}{N} \sum_{i=1}^{N} \ell_i - \mu \right| \geq \frac{(1 - \eta)\beta}{8} \right]$$

$$\leq 2 \cdot e^{-\frac{2N \left(\frac{\mu}{\eta}\right)^2}{(1 - \eta)^2}}$$

$$\leq 2 \cdot e^{-2} \cdot \frac{32}{\beta^2} \cdot \log \left(\frac{32S_r}{\delta^2} \right) \cdot \left(\frac{\beta}{\epsilon\eta}\right)^2$$

$$\leq \frac{\delta}{16[S_r]}.$$

By union bound, with probability $1 - \frac{\delta}{16}$, if there exists some $x \in R_1$ in set $S_r$, the algorithm detects and removes it. In other words, if TEST terminates without restarting Algorithm 3, the probability mass $D[R_1] \leq \frac{\epsilon}{8}$ with probability $1 - \frac{\delta}{8}$. Therefore, if Algorithm 7 reaches its Step 7, we have $D[R_1 \cup R_2] \leq \frac{\epsilon}{4}$ with probability at least $1 - \frac{\delta'}{4}$.

In addition, the total number of comparison tags in TEST is

$$\frac{8}{\epsilon} \log \frac{16}{\delta} \cdot \frac{32}{\beta^2} \cdot \log \left(\frac{8S_r}{\delta^2}\right) = O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta} \log \left(\frac{1}{\epsilon\beta}\right)\right).$$

(C.2)

Lemma 21 (Restatement of Lemma 10). Consider Phase 2 of Algorithm 3 with probability $1 - \frac{\delta'}{4}$, $\text{err}_{D_2}(h_2) \leq \frac{\sqrt{2}}{\beta}$.

Proof. We acknowledge that some of the deductions in this proof follow directly from that of Awasthi et al. (2017b) and are included for completeness.
Let $R := R_1 \cup R_2$. By Lemma $\text{[7]}$ and $\delta_2 = \frac{\delta}{12}$, with probability at least $1 - \frac{\delta}{12}$, $D[R] \leq \frac{\varepsilon}{4}$. We first argue that for any $x \notin R$, $\text{SEMI-VERIFIED-FILTER}$ does a good job to simulate $D_I$. Consider distribution $D'$ that has equal probability on the distributions induced by $\overline{W}_I$ and $\overline{W}_C$ and let $d'(x)$ denote the density of point $x$ in this distribution. Likewise let $d_2(x)$ be the density of points in $D_2$. We want to show that for any $x \notin R$, $d'(x) = \Theta(d_2(x))$.

Recall that $\text{err}_D(h_1) = \Theta(\frac{\varepsilon}{\sqrt{d}})$. Let $d(x), d_C(x), \text{and } d_I(x)$ be the density of instance $x$ in distributions $D, D_C, \text{and } D_I$, respectively. Note that, for any $x$ such that $h_1(x) = h^*(x)$, we have $d(x) = d_C(x)(1 - \frac{1}{2}\sqrt{\varepsilon})$. Similarly, for any $x$ such that $h_1(x) \neq h^*(x)$, we have $d(x) = d_I(x)\frac{1}{2}\sqrt{\varepsilon}$. $N_C(x), N_I(x), M_C(x)$ and $M_I(x)$ be the number of occurrences of $x$ in the sets $S_C, S_I, \overline{W}_C$ and $\overline{W}_I$, respectively. For any $x$, there are two cases:

If $h_1(x) = h^*(x)$: Then, there exist absolute constants $c_1$ and $c_2$ according to Lemma $\text{[26]}$ such that

$$d'(x) = \frac{1}{2} E \left[ \frac{M_C(x)}{|W_C|} \right] \geq \frac{E[M_C(x)]}{c_1 \cdot m_{\sqrt{\varepsilon}}} \geq \frac{E[N_C(x)]}{c_1 \cdot m_{\sqrt{\varepsilon}}} = \frac{|S_C| \cdot d(x)}{c_1 \cdot m_{\sqrt{\varepsilon}}}$$

where the second and sixth transitions are by the sizes of $\overline{W}_C$ and $|S_C|$ and the third transition is by the fact that if $h(x) = h^*(x), M_C(x) > N_C(x)$.

If $h_1(x) \neq h^*(x)$: Then, there exist absolute constants $c'_1$ and $c'_2$ according to Lemma $\text{[26]}$ such that

$$d'(x) = \frac{1}{2} E \left[ \frac{M_I(x)}{|W_I|} \right] \geq \frac{E[M_I(x)]}{c'_1 \cdot m_{\sqrt{\varepsilon}}} \geq \frac{E[N_I(x)]}{c'_1 \cdot m_{\sqrt{\varepsilon}}} \geq \frac{1}{2} \frac{d(x)}{|S|} \geq \frac{c'_2 d_I(x)}{2}$$

where the second and sixth transitions are by the sizes of $\overline{W}_I$ and $|S_I|$, the third transition is by the fact that if $h(x) \neq h^*(x), M_I(x) > N_I(x)$, and the fourth transition holds by Lemma $\text{[25]}$.

For $x \in R$, we have $D[R] \leq \frac{\varepsilon}{4}$. Therefore, $D_2[R] \leq \frac{\varepsilon}{4}$ because $D = \frac{\varepsilon}{4} D_I + (1 - \frac{\varepsilon}{4}) D_C$ and $D_2 = \frac{1}{2} D_I + \frac{1}{2} D_C$. As a result, except for a $\frac{\varepsilon}{4}$ fraction under $D_2$, $\forall x \ d'(x) \geq \Theta(d_2(x))$, meaning that $D'$ is a good simulation of $D_2$. By Lemma $\text{[26]}$, since $|\overline{W}| = \Theta(n_{\sqrt{\varepsilon}, \varepsilon})$, applying the super-sampling lemma (Lemma $\text{[24]}$), we know that $\text{err}_{D_2}(h_2) \leq O(\frac{\varepsilon}{\sqrt{d}})$ with probability $1 - \frac{\delta}{4}$. 

\begin{theorem}[Restatement of Theorem $\text{[11]}$]
Given any $\alpha, \beta, \epsilon, \delta \in (0, 1)$, assume $\eta < \frac{\min(\alpha, \beta)}{100}$. Denote $T_{a, \eta} = \log \left( \frac{1}{\epsilon} \left( \frac{\alpha - \eta}{\alpha - \gamma} \right) \right)$ and $T_{b, \eta} = \log \left( \frac{1}{\epsilon} \left( \frac{\beta - \eta}{\beta - \gamma} \right) \right)$. The following holds with probability $1 - \delta$. There exists an algorithm (Algorithm $\text{[3]}$) that runs in $O(\log(d, T_{a, \eta}, T_{b, \eta}))$ time and returns a hypothesis $h: \mathcal{X} \rightarrow \mathcal{Y}$ with $\text{err}_D(h) \leq \epsilon$. In addition, $m_L = \Theta(\frac{T_{a, \eta} + T_{b, \eta}}{\alpha^2} \cdot \log \left( \frac{\log d + T_{a, \eta} + T_{b, \eta}}{\epsilon} \right)), m_C = \frac{1}{\beta} \cdot \log \left( \frac{T_{a, \eta} + T_{b, \eta}}{\alpha^2} \cdot n_{\epsilon, \delta}, \right)$, and $m_V = O(T_{\alpha, \eta} + T_{\beta, \eta})$. Therefore, $\Lambda_L = o\left(\frac{T_{a, \eta} + T_{b, \eta}}{\alpha^2}\right)$ and $\Lambda_C = o\left(\frac{1}{\beta} \left( T_{a, \eta} + T_{b, \eta} \right) \right)$ when $\epsilon \in (0, (\log d)^{-4})$. In particular, when $\alpha$ and $\beta$ are constants, $\Lambda_L = o(1), \Lambda_C = O(1)$, and $m_V = O(1)$.

\end{theorem}

\begin{remark}
The above theorem presents tighter bounds for $m_L, m_C, \Lambda_L, \Lambda_C$ than that in Theorem $\text{[1]}$. Notice that $\frac{1}{\alpha} = O(T_{a, \eta})$ and $\frac{1}{\beta} = O(T_{b, \eta})$. Hence, Theorem $\text{[1]}$ in our main paper is more general. In addition, the quantities $T_{a, \eta}, T_{b, \eta}, \frac{1}{\alpha}, \frac{1}{\beta}, \eta$ are independent of the dimension $d$, the desired error rate $\epsilon$, and the confidence parameter $\delta$. Hence, when consider $m_V$ and the overheads, our algorithm is query-efficient. 

\end{remark}

\begin{proof}
Following Corollary $\text{[18]}$, Lemma $\text{[19]}$ and $\text{[21]}$, if the main algorithm outputs a hypothesis $h$ without restarting, we have the following guarantees with probability $1 - \delta$: (i) $\text{err}_D(h_1) \leq \frac{\varepsilon}{\sqrt{d}}$; $\text{err}_{D_2}(h_2) \leq \frac{\varepsilon}{\sqrt{d}}$; $\text{err}_{D_3}(h_3) \leq \frac{\varepsilon}{\sqrt{d}}$ hold simultaneously; (ii) the algorithm queries at most $O(T_{a, \eta})$ verified labels and $O(T_{b, \eta})$ verified comparison tags from $O_T$.

Applying Theorem $\text{[23]}$, the hypothesis $h$ returned by Algorithm $\text{[1]}$ is such that $\text{err}_D(h) \leq \epsilon$ with probability $1 - \delta$

In Lemma $\text{[19]}$, we show that the success of our pruning scheme crucially relies on a condition that $\eta < \frac{\alpha}{2}$ and $\eta < \frac{\beta}{2}$ for some large enough constant $c > 0$. Set $c = 16$. We require $\eta < \frac{\min(\alpha, \beta)}{16}$ for the guarantees in Theorem $\text{[1]}$ to hold. Moreover, $m_V = O(T_{a, \eta} + T_{b, \eta})$.


It remains to show the label and comparison complexity, \(m_L, m_C\), and corresponding overheads \(\Lambda_L, \Lambda_C\). The label complexity follows from Proposition \([17]\) by setting \(n = n_{\sqrt{\alpha \beta'}}\) and the fact that main algorithm is only restarted by a total of \(O(T_{\alpha \eta} + T_{\beta \eta})\) times. Note that \(O\left(\frac{1}{\epsilon^2}\right) = O\left(\frac{1}{\epsilon^2}\right) = O((T_{\alpha \eta} + T_{\beta \eta}) \cdot \frac{1}{\delta})\). We have
\[
m_L = (T_{\alpha \eta} + T_{\beta \eta}) \cdot \mathcal{O}\left(\frac{\log n_{\sqrt{\alpha \beta'}}}{\alpha^2} \cdot \log \log n_{\sqrt{\alpha \beta'}}\right)
\]
\[
= \frac{T_{\alpha \eta} + T_{\beta \eta}}{\alpha^2} \cdot \mathcal{O}\left(\frac{d + (T_{\alpha \eta} + T_{\beta \eta}) \frac{1}{2}}{\epsilon}\right).
\]

On the other hand, the comparison complexity is a summation of the results from Lemma \([20]\) and that of \(\text{Zeng and Shen} \, (2022a)\).
\[
m_C = (T_{\alpha \eta} + T_{\beta \eta}) \cdot \mathcal{O}\left(\frac{n_{\sqrt{\alpha \beta'}}}{\beta^2} \cdot \log^2 n_{\sqrt{\alpha \beta'}} + n_{e \delta} + \frac{1}{\epsilon \beta^2} \log \left(\frac{1}{\beta'}\right) \log \left(\frac{1}{\epsilon \beta'}\right)\right)
\]
\[
= \frac{T_{\alpha \eta} + T_{\beta \eta}}{\beta^2} \cdot \mathcal{O}\left(\frac{d \log \left(\frac{1}{\epsilon}\right) + \left((T_{\alpha \eta} + T_{\beta \eta}) \frac{1}{2}\right) \frac{1}{\epsilon} \log \left(\frac{1}{\epsilon \beta'}\right) \log \left(\frac{1}{\epsilon \beta'}\right)}{\epsilon}\right)
\]
\[
= \frac{1}{\beta^2} \cdot \mathcal{O}\left(\frac{(T_{\alpha \eta} + T_{\beta \eta}) \frac{1001}{\epsilon}}{\epsilon} \cdot \frac{d + \left(\frac{1}{\epsilon}\right) \frac{1001}{\epsilon}}{\epsilon}\right) = \frac{1}{\beta^2} \cdot \mathcal{O}\left(\frac{(T_{\alpha \eta} + T_{\beta \eta}) \frac{1001}{\epsilon}}{\epsilon} \cdot \frac{d + \left(\frac{1}{\epsilon}\right) \frac{1001}{\epsilon}}{\epsilon}\right)
\]

These in allusion to \(m_{e \delta} = K \cdot \left(\frac{1}{\epsilon} (d \log (1/\epsilon) + \log (1/\delta))\right)\) immediately give the overheads as follows:
\[
\Lambda_L = \mathcal{O}\left(\frac{T_{\alpha \eta} + T_{\beta \eta}}{\alpha^2} \cdot \log \log n_{\sqrt{\alpha \beta'}}\right)
\]
\[
\leq \frac{T_{\alpha \eta} + T_{\beta \eta}}{\alpha^2} \cdot \frac{\epsilon}{d + \log (1/\delta)} \cdot \mathcal{O}\left(\log \left(\frac{d + (T_{\alpha \eta} + T_{\beta \eta}) \frac{1}{2}}{\epsilon}\right)\right)
\]
\[
= \frac{T_{\alpha \eta} + T_{\beta \eta}}{\alpha^2} \log \left(\frac{T_{\alpha \eta} + T_{\beta \eta}}{\epsilon} \frac{d}{d + \log (1/\delta)} \right)
\]

and
\[
\Lambda_C = \frac{T_{\alpha \eta} + T_{\beta \eta}}{\beta^2} \cdot \mathcal{O}\left(\frac{n_{\sqrt{\alpha \beta'}}}{m_{e \delta}} \cdot \log n_{\sqrt{\alpha \beta'}} + \frac{1}{\epsilon \beta^2} \log \left(\frac{1}{\beta'}\right) \log \left(\frac{1}{\epsilon \beta'}\right)\right)
\]
\[
\leq \frac{T_{\alpha \eta} + T_{\beta \eta}}{\beta^2} \cdot \mathcal{O}\left(\sqrt{\epsilon} \cdot \frac{d \log \frac{1}{\epsilon} + \left((T_{\alpha \eta} + T_{\beta \eta}) \frac{1}{2}\right) \frac{1}{\epsilon} \log \left(\frac{1}{\epsilon \beta'}\right) \log \left(\frac{1}{\epsilon \beta'}\right)}{d \log \frac{1}{\epsilon} + \log \left(\frac{1}{\delta}\right)} \times \mathcal{O}\left(\frac{d + (T_{\alpha \eta} + T_{\beta \eta}) \frac{1}{2}}{\epsilon}\right)\right)
\]
\[
\leq \frac{1}{\beta^2} \left(T_{\alpha \eta} + T_{\beta \eta} \right) \frac{1001}{\epsilon} \log^2 \left(\frac{T_{\alpha \eta} + T_{\beta \eta}}{d \log \frac{1}{\epsilon} + \log \left(\frac{1}{\delta}\right)} \right) \cdot \mathcal{O}\left(\sqrt{\epsilon} \cdot \frac{d \log \frac{1}{\epsilon} + (\frac{1}{\delta}) \frac{1001}{\epsilon}}{d \log \frac{1}{\epsilon} + \log \left(\frac{1}{\delta}\right)} \cdot \frac{d + (T_{\alpha \eta} + T_{\beta \eta}) \frac{1}{2}}{\epsilon}\right)
\]

Recall that by the definition of \(n_{e \delta}\) and \(m_{e \delta}\) in Section \([4]\) we have \(m_{e \delta} = \Theta(n_{e \delta}) = \Theta\left(d \log \frac{1}{\delta}\right)\) when \(\delta\) is a constant. In this case, we can see that
\[
\Lambda_C \leq \frac{1}{\beta^2} \left(T_{\alpha \eta} + T_{\beta \eta} \right) \frac{1001}{\epsilon} \log^2 \left(\frac{T_{\alpha \eta} + T_{\beta \eta}}{d \log \frac{1}{\epsilon} + \log \left(\frac{1}{\delta}\right)} \right) \cdot \mathcal{O}\left(\sqrt{\epsilon} \cdot \log^2 \frac{d}{\epsilon} + 2\right).
\]

When \(\epsilon \in (0, (\log d)^{-4}), \Lambda_C = \mathcal{O}\left(\frac{1}{\beta^2} \left(T_{\alpha \eta} + T_{\beta \eta} \right) \frac{1001}{\epsilon} \log^2 \left(\frac{T_{\alpha \eta} + T_{\beta \eta}}{d \log \frac{1}{\epsilon} + \log \left(\frac{1}{\delta}\right)} \right)\right). \) In addition, \(\Lambda_L = o\left(\frac{T_{\alpha \eta} + T_{\beta \eta}}{\alpha^2} \cdot \log (T_{\alpha \eta} + T_{\beta \eta})\right)\), because \(\Lambda_L\) goes to 0 when \(\epsilon\) goes to 0.
D USEFUL LEMMAS

**Theorem 23** (Boosting, Schapire [1990]). For any $p < \frac{1}{2}$ and distribution $D$, consider three classifiers $h_1(x), h_2(x), h_3(x)$ satisfying the following. 1) $\text{err}_D(h_1) \leq p$; 2) $\text{err}_D(h_2) \leq p$ where $D_2 := \frac{1}{2}D_C + \frac{1}{2}D_1$, $D_C$ denotes the distribution $D$ conditioned on $\{x : h_1(x) = h^*(x)\}$, and $D_1$ denotes $D$ conditioned on $\{x : h_1(x) \neq h^*(x)\}$; 3) $\text{err}_D(h_3) \leq p$ where $D_3$ is $D$ conditioned on $\{x : h_1(x) \neq h_2(x)\}$. Then $\text{err}_D(\text{Maj}(h_1, h_2, h_3)) \leq 3p^2 - 2p^3$.

**Lemma 24** (Robust Super-Sampling Lemma, Lemma 4.12 in Awasthi et al. [2017b]). Given a hypothesis class $\mathcal{H}$ consider any two discrete distributions $D$ and $D'$ over $X$ such that except for an $\epsilon$ fraction of the mass under $D$, we have that for all $x, d'(x) \geq c \cdot d(x)$ for an absolute constant $c > 0$ and both distributions are labeled according to $h^* \in \mathcal{H}$. There exists a constant $c' > 1$ such that for any $\epsilon, \delta$, with probability $1 - \delta$ over a labeled sample set $S$ of size $c'm_{\epsilon, \delta}$ drawn from $D'$, $\mathcal{A}_\mathcal{H}(S)$ has error of at most $2\epsilon$ with respect to $D$.

**Lemma 25** (Lemma 14 in Zeng and Shen [2022a]). Consider Algorithm 7. Assume that the subset $U$ is correctly labeled. Consider any given instance $x \in S_2$ except for an $\frac{\epsilon}{2}$-fraction, we have the following guarantee. If $h(x) = h^*(x)$, it will be added to $S_1$ with probability at most $\frac{1}{4}\sqrt{\epsilon}$; if $h(x) \neq h^*(x)$, it goes to $S_1$ with probability at least $\frac{1}{4}$.

**Proof.** First, for any instance $x \in S_2 \cup U$, the guarantee follows from that of Prune-Compare-and-Label.

Now consider any instance $x \in S_2 \setminus (S_2 \cup U)$. Recall that Lemma 9 guarantees the probability mass $D[R_1 \cup R_2] \leq \frac{\epsilon}{8}$ with probability $1 - \frac{\epsilon}{12}$. Given that the input size of $S$ is $\Theta(n_{\epsilon, \delta})$, with probability at least $1 - \frac{\epsilon}{2}$, the fraction of the instances $x \in S_2$ that falls in region $R_1 \cup R_2$ is less than $\frac{\epsilon}{2}$. Therefore, at least $1 - \frac{\epsilon}{2}$ fraction of the instances in $S$ falls outside $R_1 \cup R_2$ such that $\text{Maj-size}_C(x, x^-) \geq 1 - \frac{\epsilon}{2}$ and $\text{Maj-size}_C(x, x^+) \geq 1 - \frac{\epsilon}{2}$ (Eq. (3.1)). For these instances, we can consider that the majority is correct. Now if $x$ is outside the interval $[x^-, x^+]$, when $\beta < 0.7$, $\text{Maj-size}_C(x, x') \geq \frac{1}{2} + \Theta(1)$ which suffices for the filtering scheme from Awasthi et al. [2017b] to work; when $\beta \geq 0.7$, lemma follows from that of Zeng and Shen [2022a]. If $x$ falls into the interval $[x^-, x^+]$, the probability that it will be added to $S_1$ but $h(x) = h^*(x)$, or it will not be added to $S_1$ but $h(x) \neq h^*(x)$ are both less than $\frac{1}{4}\sqrt{\epsilon}$. The proof is complete.

**Lemma 26** (Lemma 4.7 in Awasthi et al. [2017b]). Consider Algorithm 7. With probability at least $1 - \exp(-\Omega(m_{\epsilon, \delta}^{\tau}, \delta'))$, $W_I$, $W_C$ and $S_1$ all have size $\Theta(m_{\epsilon, \delta}^{\tau}, \delta')$. Consider Algorithm 5. With probability at least $1 - \exp(-\Omega(n_{\epsilon, \delta}^{\tau}, \delta'))$, $W_I$, $W_C$ and $S_1$ all have size $\Theta(n_{\epsilon, \delta}^{\tau}, \delta')$.

**Proof.** Notice that in both Algorithm 1 and 5, we have the input sample size $m_{\epsilon, \delta}, n_{\epsilon, \delta}$ lower bounded by $\Omega(m_{\epsilon, \delta})$, the lemma follows from Lemma 4.7 of Awasthi et al. [2017b].

**Proposition 27** (Proposition 4 in Zeng and Shen [2022a]). Suppose $\alpha \geq 0.7, \beta \geq 0.7$. Consider the Prune-Compare-and-Label algorithm, i.e. Algorithm 4. If $|S| \geq (\frac{\epsilon}{3})^\epsilon / 1000$, then with probability at least $1 - \delta$, it correctly sorts and labels all the instances in $S$. The label complexity is $O\left(\log|S| \cdot \log\log|S|\right)$, and the comparison complexity is given by $O\left(|S| \cdot \log^2|S|\right)$.

**Theorem 28** (Theorem 8 in Zeng and Shen [2022a]). Suppose $\alpha \geq 0.7, \beta \geq 0.7$. With probability $1 - \delta$, Algorithm 3 runs in time $\text{poly}(d, \frac{1}{\epsilon})$ and returns a classifier $h \in \mathcal{H}$ with error rate $\text{err}_D(h) \leq \epsilon$. In addition, the label complexity is $O\left(\log n_{\epsilon, \delta}^{\tau} \cdot \log \log n_{\epsilon, \delta}^{\tau}\right)$, and the comparison complexity is $O\left(n_{\epsilon, \delta}^{\tau} \log^2 n_{\epsilon, \delta}^{\tau} + n_{\epsilon, \delta}\right)$.