Efficient Semi-supervised and Active Learning of Disjunctions: Supplementary Material

Maria-Florina Balcan NINAMF@CC.GATECH.EDU
Christopher Berlind CBERLIND@GATECH.EDU
Steven Ehrlich SEHRLICH@CC.GATECH.EDU
Yingyu Liang YLIANG39@GATECH.EDU
School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

1. Bounding T by $\log |C_{D,\chi}(\epsilon)|$

The following lemma bounds the number of connected components in the commonality graph by the number of compatible hypotheses. For notational definitions, refer to Section 2 in the main body of the paper.

Lemma 1. *Let G be the graph that results from removing all non-indicators from $G_{\mathrm{com}}(U)$, and suppose G is divided into T connected components. If $m_u \geq \frac{2n^2}{\epsilon} \log \frac{n}{\delta}$, then $T \leq \log_2 |C_{D,\chi}(\epsilon)|$ with probability at least $1 - \delta$.*

Proof. Since G has no non-indicators, a hypothesis is compatible with U if and only if every component is made entirely of indicators of the same type. There are two possible choices for each component, so the number of fully compatible hypotheses is $|C_{U,\chi}(0)| = 2^T$.

To complete the proof, it is sufficient to show that $C_{U,\chi}(0) \subseteq C_{D,\chi}(\epsilon)$. Since any hypothesis in $C_{U,\chi}(0)$ is compatible with any example containing variables from only one component, we only need to show that there is at most ϵ probability mass of examples that contain variables from multiple components. All such examples correspond to edges that are absent from $G_{\mathrm{com}}(U)$, so we only need to show that $G_{\mathrm{com}}(U)$ was constructed with enough examples so that nearly all significant edges appear in the graph.

To see this, fix any pair of variables x_i, x_j. If $\Pr_{x\sim D}[x_i = 1 \wedge x_j = 1] < \epsilon/n^2$, we can ignore this pair since all such pairs together constitute a probability mass strictly less than ϵ. Now suppose $\Pr_{x\sim D}[x_i = 1 \wedge x_j = 1] \geq \epsilon/n^2$. The probability that x_i and x_j do not appear together in any of the examples in U is at most $(1 - \frac{\epsilon}{n^2})^{m_u}$, so if $m_u \geq \frac{n^2}{\epsilon} \log \frac{n^2}{\delta}$ then this failure probability is at most δ/n^2. By a union bound over all such pairs, with probability at least $1 - \delta$ all corresponding edges appear in $G_{\mathrm{com}}(U)$, and the probability mass of examples containing variables from multiple components is at most ϵ. This means that every fully compatible hypothesis has unlabeled error at most ϵ, so we have $T = \log_2 |C_{U,\chi}(0)| \leq \log_2 |C_{D,\chi}(\epsilon)|$. □

2. Finding a Consistent Compatible Hypothesis is NP-hard

The following theorem formalizes the computational difficulty of finding a fully consistent and compatible two-sided disjunction in the semi-supervised setting.

Theorem 4. *Given data sets L and U, finding a hypothesis $h \in C$ that is both consistent with L and compatible with U is* **NP***-hard.*

Proof sketch. The proof is by reduction from 3-SAT. Given a 3-SAT instance φ on variables $x_1, \ldots, x_n$ we produce the following data sets L and U containing examples on the $4n$ variables $x_1^+, x_1^-, \bar{x}_1^+, \bar{x}_1^-, \ldots, x_n^+, x_n^-, \bar{x}_n^+, \bar{x}_n^-$. The labeled set L contains examples of the form $(\{x_i^+, \bar{x}_i^+\}, +1)$ and $(\{x_i^-, \bar{x}_i^-\}, -1)$ for $1 \leq i \leq n$. In addition, for each clause in φ of the form $(\ell_i \vee \ell_j \vee \ell_k)$ where ℓ_i, ℓ_j, ℓ_k can each be positive or negative literals, L contains the example $(\{\ell_i^+, \ell_j^+, \ell_k^+\}, +1)$. The unlabeled set U contains examples of the form $\{x_i^+, x_i^-\}$ and $\{\bar{x}_i^+, \bar{x}_i^-\}$ for $1 \leq i \leq n$. The labelings that are consistent and compatible with all the non-clause examples correspond precisely to assignments of $x_1, \ldots, x_n$, and the clauses are compatible with a given hypothesis only if they are satisfied in the underlying assignment. The set of positive indicators of any hypothesis $h = (h_+, h_-) \in C$ that is both consistent with L and compatible with U corresponds to a truth assignment to $x_1, \ldots, x_n$ that satisfies φ, therefore finding such a hypothesis is **NP**-hard. □

3. Random Classification Noise

Here we consider the problem of learning two-sided disjunctions under random classification noise, where the label of each example is flipped with probability $0 \leq \alpha < 1/2$ independently. Our goal is to extend our algorithms to this setting so that they still successfully output a low error hypothesis without significant increase in sample complexity.

More specifically, we have a distribution $D_{X,Y}$ over labeled examples $(X, \ell(X))$, and the Bayes decision rule is a two-sided disjunction $h^* \in C$, which we also refer to as the target concept. We have $\chi(h^*, D) = 1$ where D is the margin of $D_{X,Y}$ over X, and

$$\Pr[\ell(X) = -h^*(x) | X = x] = \alpha.$$

For a hypothesis h, let $\mathsf{err}_D(h)$ denote its error over the distribution $D_{X,Y}$, i.e.

$$\mathsf{err}_D(h) = \Pr_{(x,\ell(x))\sim D_{X,Y}} [h(x) \neq \ell(x)].$$

Let $\mathsf{err}_L(h)$ denote its empirical error on the noisy labeled examples L, i.e.

$$\mathsf{err}_L(h) = \frac{1}{|L|} \sum_{(x,\ell(x))\in L} I[h(x) \neq \ell(x)].$$

For convenience, we define the distance between h and h^* to be

$$d(h, h^*) = \Pr[h(x) \neq h^*(x)].$$

We aim to find a hypothesis h with error

$$\mathsf{err}_D(h) \leq \mathsf{err}_D(h^*) + \epsilon = \alpha + \epsilon.$$

Note that it is sufficient to have $d(h, h^*) \leq \epsilon$, since by triangle inequality

$$\begin{aligned} \mathsf{err}_D(h) &= \Pr[h(x) \neq \ell(x)] \\ &\leq \Pr[h(x) \neq h^*(x)] + \Pr[h^*(x) \neq \ell(x)] \\ &\leq d(h, h^*) + \mathsf{err}_D(h^*) = d(h, h^*) + \alpha. \end{aligned}$$

In the following two subsections, we show how to extend the algorithms (Algorithm 2 and 3) for semi-supervised and active learning respectively. In the last subsection, we include the extension of Algorithm 1. Note that due to the noise in the labeled examples, we cannot hope to find a consistent and compatible hypothesis. The extension of Algorithm 1 only outputs a hypothesis that has low error. However, it achieves better sample complexity bound than the extension of Algorithm 2.

3.1. Semi-supervised Learning

In the noise-free setting, we build a hypothesis based on the commonality graph, and then check and update the hypothesis until it has low error rate. The key idea in the noisy setting is that we label variables (i.e. identify variables as positive/negative potential indicators) by majority labels of sufficiently many examples containing the variables, and we use a sufficiently large set of labeled examples each time we check the hypothesis. A brief description is provided below, while the details are provided in Algorithm 4 and Theorem 5.

First, we build the commonality graph and the potential indicator sets. We only label a variable if it is present in at least $\tilde{O}(\frac{1}{(1-2\alpha)^2})$ examples so that we can use majority label of the examples to correctly decide its type. We call such variables *significant*. If we draw $\tilde{O}(\frac{1}{\epsilon_0(1-2\alpha)^2})$ examples, then with high probability each non-indicator is significant and thus labeled.

Then we build a hypothesis and draw a set of labeled examples to check it. If the empirical error is small, we output the hypothesis since it is guaranteed to have small error. Otherwise, either a component without any labeled variables in it or a non-indicator causes large error. In the first case, we need sufficiently many examples fall in the component so that we can use majority voting to decide the type of the variables in it. This requires $\tilde{O}(\frac{T}{\epsilon(1-2\alpha)^2})$ examples where T is the number of connected components in the graph that results from removing all non-indicators from the commonality graph. In the second case, to reveal the non-indicator, we must be able to distinguish between an error rate of α (the error rate caused by noise) and $\alpha + \Theta(\epsilon(1-2\alpha)/k)$ (the error rate caused by noise and the non-indicator leading to large error). This requires $\tilde{O}(\frac{k^2}{\epsilon^2(1-2\alpha)^2})$ labeled examples for each significant variable. Therefore, we draw $\tilde{O}\left(\frac{1}{(1-2\alpha)^2}\left[\frac{k^3}{\epsilon^3} + \frac{T}{\epsilon}\right]\right)$ examples at each check, so that either a component that previously contained no labeled variables is labeled, or a non-indicator is revealed. After at most $(k + T)$ updates, we are guaranteed to have a hypothesis with small error.

Theorem 5. *For any distribution $D_{X,Y}$ over $\{0,1\}^n \times \{-1,1\}$ and target concept $h^* \in C$ in the random classification noise model such that h^* has at most k non-indicators, and the minimum non-indicator probability is ϵ_0, if $m_u = O(\frac{n^2}{\epsilon}\log\frac{n}{\delta})$ and*

$$\begin{aligned} m_l &= O\Bigg(\frac{k + \log|C_{D,\chi}(\epsilon)|}{(1-2\alpha)^2}\log^2\frac{n}{\delta} \\ &\qquad \left[\frac{1}{\epsilon_0} + \frac{k^3}{\epsilon^3} + \frac{\log|C_{D,\chi}(\epsilon)|}{\epsilon}\right]\Bigg) \end{aligned}$$

Algorithm 4 Learning a Low-error Hypothesis for Two-Sided Disjunctions under Random Classification Noise

```
Input: data sets $U$ and $L$, parameters $\epsilon$, $\delta$, $k$, $\epsilon_0$, $T$
Set $G = G_{\rm com}(U), V_+^0 = V_-^0 = \emptyset$
Set $L' = $ sample $\left(\frac{100}{\epsilon_0(1-2\alpha)^2}\log^2\frac{n}{\delta}, L\right)$ and $L = L \backslash L'$

Set $size(v) = |\{x \in L' : x \ni v\}|, \forall v \in V$
Set $SIG = \{v \in V : size(v) > \frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}\}$
for each $v \in SIG$ do
    Set $l = \text{sign}(\sum_{x \ni v} \ell(x))$ and $V_l^0 = V_l^0 \cup \{v\}$
Set $V_+ = V_+^0, V_- = V_-^0$ and $h = h_{G,V_+\cup V_-}$
$S = $ sample $\left(\frac{100}{(1-2\alpha)^2}\left[\frac{k^3}{\epsilon^3} + \frac{T}{\epsilon}\right]\log^2\frac{n}{\delta}, L\right)$, $L = L \setminus S$

while $L \neq \emptyset$ and $\mathsf{err}_S(h) > \alpha + \frac{1-2\alpha}{2}\epsilon$ do
    Let $\mathcal{R}$ denote the set of the components in $G$ that have no labeled variables
    Let $S(R) = \{x \in S : x$ falls into $R\}, \forall R \in \mathcal{R}$
    Let $S(v) = \{x \in S : \text{nn}_{G,V_+\cup V_-}(x) = v\}, \forall v \in V_+ \cup V_-$
    if $\exists R \in \mathcal{R}$ such that $|S(R)| \geq \frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}$ then
        Set $l = \text{sign}(\sum_{x \in S(R)} \ell(x))$ and $V_l = V_l \cup R$
    if $\exists v \in V_+^0 \cup V_-^0$ such that $|S(v)| \geq \frac{10k^2}{\epsilon^2(1-2\alpha)^2}\log\frac{n}{\delta}$ and $\mathsf{err}_{S(v)}(h) \geq \alpha + (1-2\alpha)\frac{\epsilon}{16k}$ then
        Set $G = G \setminus \{v\}$
    Set $S = $ sample $\left(\frac{100}{(1-2\alpha)^2}\left[\frac{k^3}{\epsilon^3} + \frac{T}{\epsilon}\right]\log^2\frac{n}{\delta}, L\right)$,
    Set $L = L \setminus S$ and $h = h_{G,V_+\cup V_-}$
Output: the hypothesis $h$
```

then with probability at least $1-\delta$, Algorithm 4 outputs a hypothesis h in polynomial time such that $\mathsf{err}_D(h) \leq \alpha + \epsilon$.

Proof. **Generalization Error:** Suppose we check the hypothesis at most $k+T$ times (proved later), where T is the number of connected components in the graph that results from removing all non-indicators from $G_{\rm com}(U)$. Then when

$$|S| \geq \frac{100}{(1-2\alpha)^2\epsilon^2}\log\frac{k+T}{\delta}$$

w.h.p. all hypotheses h with $d(h,h^*) > \epsilon$ will have empirical error on S larger than $\alpha + \frac{1-2\alpha}{2}\epsilon$. So when the algorithm stops the hypothesis satisfies $d(h,h^*) \leq \epsilon$ and thus $\mathsf{err}_D(h) \leq \alpha + \epsilon$.

Bounding the Number of Updates: We now show that the hypothesis is indeed updated at most $(k+T)$ times. We begin by proving that when

$$|L'| \geq \frac{100}{\epsilon_0(1-2\alpha)^2}\log^2\frac{n}{\delta}$$

w.h.p. every non-indicator is labeled and every labeled indicator gets the correct label, so that the hypothesis is updated correctly when its error is large. First, by Chernoff and union bounds, the probability that there exists a non-indicator that appears in less than $\frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}$ examples is bounded by $k\exp\{O(\epsilon_0|L'|)\} \leq O(\delta)$. So w.h.p. every non-indicator appears in enough examples and thus is labeled. Second, the type of an indicator is decided by the majority label of $O(\frac{1}{(1-2\alpha)^2}\log\frac{n}{\delta})$ examples. By Hoeffding and union bounds, w.h.p. the types of all indicators appearing in enough examples are decided correctly.

We now prove that when the error of the hypothesis is large, we can make progress by either labeling a previously unlabeled component or identifying a non-indicator, and thus it is updated at most $(k+T)$ times. When

$$|S| > \frac{100}{(1-2\alpha)^2\epsilon^2}\log\frac{k+T}{\delta},$$

if $\mathsf{err}_S(h) > \alpha + \frac{1-2\alpha}{2}\epsilon$, then w.h.p. $d(h,h^*) \geq \epsilon/4$. For each $v \in V_+ \cup V_-$, let $X(v)$ denote those examples whose nearest labeled variable is v, i.e.

$$X(v) = \{x \in X : \text{nn}_{G,V_+\cup V_-}(x) = v\}.$$

For each $R \in \mathcal{R}$, let $X(R)$ denote those examples fall into the component R, i.e.

$$X(R) = \{x \in X : x \text{ falls into } R\}.$$

Note that for any indicator $v \in V_+ \cup V_-$, its type is correctly decided, so the hypothesis makes no mistake on $X(v)$. Since $|\mathcal{R}| \leq T$ and the number of non-indicators is bounded by k, either there is a component R' such that

$$\Pr[h(x) \neq h^*(x) \wedge x \in X(R')] > \epsilon/(8T)$$

or there is a non-indicator v' such that

$$\Pr[h(x) \neq h^*(x) \wedge x \in X(v')] > \epsilon/(8k).$$

In the first case, w.h.p. there are more than $\frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}$ examples in $S(R')$ when

$$|S| \geq \frac{100T}{\epsilon(1-2\alpha)^2}\log^2\frac{n}{\delta}.$$

These examples are sufficient to decide the type of the indicators in the component correctly. Also, w.h.p. for any $R \in \mathcal{R}$ with more than $\frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}$ examples in $S(R)$, the type of the indicators in the component are correctly decided. This means that we correctly label at least one component not labeled previously. This type of updates happen at most T times.

In the second case, we have

$$\Pr[h(x) \neq h^*(x) | x \in X(v')] > \epsilon/(8k)$$

and thus

$$\Pr[h(x) \neq \ell(x) | x \in X(v')] > \alpha + (1-2\alpha)\epsilon/(8k).$$

When

$$|S| \geq \frac{100k^3}{\epsilon^3(1-2\alpha)^2} \log^2 \frac{n}{\delta}$$

w.h.p. in $S(v')$ there are more than $\frac{10k^2}{\epsilon^2(1-2\alpha)^2} \log \frac{n}{\delta}$ examples and we have

$$\mathsf{err}_{S(v')}(h) > \alpha + (1-2\alpha)\epsilon/(16k).$$

Also, for any indicator $v \in V_+ \cup V_-$,

$$\Pr[h(x) \neq \ell(x) | x \in X(v)] = \alpha.$$

Then w.h.p. we have

$$\mathsf{err}_{S(v)}(h) < \alpha + (1-2\alpha)\epsilon/(16k)$$

for any indicator v such that $|S(v)| \geq \frac{10k^2}{\epsilon^2(1-2\alpha)^2} \log \frac{n}{\delta}$. This means that we can correctly identify a non-indicator. This type of updates happen at most k times. Hence, we update the hypothesis at most $(k+T)$ times.

Sample Complexity and Running Time: When building the commonality graph, we need

$$|L'| = O\left(\frac{1}{\epsilon_0(1-2\alpha)^2} \log^2 \frac{n}{\delta}\right).$$

Each time we check the hypothesis, we need

$$|S| = O\left(\frac{1}{(1-2\alpha)^2}\left[\frac{k^3}{\epsilon^3} + \frac{T}{\epsilon}\right] \log^2 \frac{n}{\delta}\right).$$

The number of labeled examples then follows from bounding the number of checks by $(k+T)$ and bounding T by $\log |C_{D,X}(\epsilon)|$. The algorithm runs in polynomial time since building the commonality graph and checking the hypothesis take polynomial time. □

3.2. Active Learning

There are two main changes in extending our algorithm to the noisy setting. First, instead of simply determining the type of an indicator with one example, we need to check many examples that contain it. A majority vote will correctly identify the type of the indicator. Second, instead of doing binary search over nodes in the path that connect positive and negative examples, we need to search over edges. This is because with noise, an indicator may appear both in negative and positive examples similar to a non-indicator, so it is not straightforward to identify a non-indicator merely by the types of examples it appears in. On the other hand, an edge that contains an indicator will have its true label match that of the indicator, so we can reliably determine the type of an edge if there are enough examples that contain both variables, and then identify a vertex in edges of two different types to be a non-indicator. More specifically, Subroutine 5 can be used to decide the type of a pair of variables (when $u \neq v$) or that of one variable (when $u = v$), which is a building block for the extension of the active learning algorithm to the noisy setting. A brief description of the extension is provided below, while the details are provided in Algorithm 6 and Theorem 6.

First, we build the commonality graph and an initial hypothesis. Let F be the set of all examples that contain some pair of variables appearing together in fewer than $O(\frac{1}{(1-2\alpha)^2} \log \frac{n}{\delta})$ examples. We only use the unlabeled data $U \setminus F$ to construct the commonality graph, so that every edge corresponds to a pair of variables whose indicator type can be decided. Then we pick a variable in each component in the graph and decide its type. This requires $\tilde{O}(\frac{T}{(1-2\alpha)^2})$ queries, where T is the number of connected components in the graph that results from removing all non-indicators from the commonality graph $G_{\text{com}}(U \setminus F)$. A hypothesis is then constructed which labels an input example by the type of the nearest labeled variable.

Second, we check and update the hypothesis on a set of examples. We randomly sample a set S of $\tilde{O}(\frac{1}{(1-2\alpha)^2\epsilon^2})$ examples from $U \setminus F$, and compute $\mathsf{err}_S(h)$. If $\mathsf{err}_S(h)$ is at most $\alpha + \frac{1-2\alpha}{2}\epsilon$, we output the hypothesis since it has small error. Otherwise, it can be shown that on $\Omega(\epsilon)$ fraction of examples in S the hypothesis has different labels from the target concept h^*. This fact can be used to identify a non-indicator. We randomly sample $\tilde{O}(\frac{1}{\epsilon})$ examples from S, and for each example x, pick $\min(k+1, |x|_1)$ variables and decide their types, where $|x|_1$ is the number of variables appearing in x. This ensures that we will eventually pick an indicator in an example x such that $h(x) \neq h^*(x)$. Then we find a path connecting a positive indicator and a negative indicator, and thus can identify a non-indicator by binary search on the edges along the path. Therefore, after at most k updates, we are guaranteed to have a hypothesis with small error.

Theorem 6. *For any distribution $D_{X,Y}$ over $\{0,1\}^n \times \{-1,1\}$ and target concept $h^* \in C$ in the random classification noise model such that h^* has at most k non-*

Subroutine 5 DecideType(U, u, v)

```
Input: unlabeled data $U$, variables $u$ and $v$
Set $S = \text{sample}(\frac{100}{(1-2\alpha)^2}\log\frac{n}{\delta}, \{x \in U : \{u,v\} \subseteq x\})$
Set $t = \text{sign}(\sum_{x \in S} \ell(x))$
Output: $\{(u,t),(v,t)\}$
```

Algorithm 6 Actively Learning Two-Sided Disjunctions under Random Classification Noise

```
Input: unlabeled data $U$, parameters $\alpha, \epsilon, \delta, k$
Set $U(u,v) = |\{x \in U : \{u,v\} \subseteq x\}|, \forall u,v \in V$.
Set $F = \{x \in U : \exists u,v \in x, U(u,v) < \frac{100}{(1-2\alpha)^2}\log\frac{n}{\delta}\}$
Set $U' = U \setminus F$, $G = G_{\text{com}}(U')$, and $L = \emptyset$
for each connected component $R$ of $G$ do
  Set $L = L \cup \text{DecideType}(U,v,v)$ for any $v \in R$
Set $h = h_{G,L}$ and $S = \text{sample}(\frac{10}{(1-2\alpha)^2\epsilon^2}\log\frac{k}{\delta}, U')$
while $\mathsf{err}_S(h) > \alpha + \frac{1-2\alpha}{2}\epsilon$ do
  for $i = 1$ to $\frac{100}{\epsilon}\log\frac{k}{\delta}$ do
    Set $x = \text{sample}(1, S)$
    for each of $\min(k+1, |x|_1)$ variables $v \in x$ do
      Set $L = L \cup \text{DecideType}(U,v,v)$
    if $\exists (u,1),(v,-1) \in L$ such that $u \leftrightarrow_G v$ then
      Set $v = \text{BinarySearch}_{G,L}(x)$
      Set $G = G \setminus \{v\}$
      for each new component $R$ of $G$ do
        Set $L = L \cup \text{DecideType}(U,v,v)$ for $v \in R$
      Set $h = h_{G,L}$
      break
  Set $S = \text{sample}(\frac{10}{(1-2\alpha)^2\epsilon^2}\log\frac{k}{\delta},\ U')$
Output: the hypothesis $h$
```

indicators, if $|U| = O\left(\frac{n^2}{\epsilon(1-2\alpha)^2}\log^2\frac{n}{\delta}\right)$*, after at most*

$$m_q = O\left(\frac{1}{(1-2\alpha)^2}\left[\log|C_{D,\mathcal{X}}(\epsilon)| + \frac{k^2}{\epsilon} + \frac{k}{\epsilon^2}\right]\log^2\frac{n}{\delta}\right)$$

label queries, with probability at least $1-\delta$*, Algorithm 6 outputs a hypothesis* h *in polynomial time such that* $\mathsf{err}_D(h) \le \alpha + \epsilon$*.*

Proof. **Generalization Error:** Assuming the hypothesis is updated at most k times (proved later), we bound the probability that the output hypothesis h has $d(h,h^*) \le \epsilon$. We begin by showing that the ignored examples F have small probability mass. When U is sufficiently large, w.h.p. all pairs of variables that appear together with probability at least $\frac{\epsilon}{8n^2}$ will appear in sufficiently many examples in U. Assuming this is true, we have

$$\Pr[x \in F] \le \epsilon/8.$$

This means when $d(h,h^*) > \epsilon$,

$$\Pr[h(x) \neq h^*(x) | x \in X \setminus F] > 3\epsilon/4$$

and

$$\Pr[h(x) \neq \ell(x) | x \in X \setminus F] > \alpha + (1-2\alpha)\frac{3\epsilon}{4}.$$

Then we have

$$\Pr\left[\mathsf{err}_S(h) \le \alpha + \frac{1-2\alpha}{2}\epsilon\right] \le \frac{\delta}{8k}.$$

Union bounding over the k updates, we have that w.h.p. the hypothesis h output has $d(h,h^*) \le \epsilon$, which leads to $\mathsf{err}_D(h) \le \alpha + \epsilon$.

Correctness of Subroutine 5: Here we show that w.h.p. the majority voting method always decides correctly the type of the indicators, so that we build and update the hypothesis correctly. Fix a pair of variables (u,v) containing at least one indicator. Let $B_{u,v}$ denote the event that (u,v) appear in at least $\frac{100}{(1-2\alpha)^2}\log\frac{n}{\delta}$ examples in U but the algorithm fails to decide the type. This happens when the labels of more than half of the examples queried are flipped. We have by Hoeffding bound

$$\Pr[B_{u,v}] \le \exp\left\{-2(1-2\alpha)^2\frac{100}{(1-2\alpha)^2}\log\frac{n}{\delta}\right\} \le \frac{\delta}{4n^2}$$

and thus $\Pr[\cup_{u,v}B_{u,v}] \le \frac{\delta}{4}$.

Queries per Stage: To build the hypothesis, we decide the type of one variable for each connected component of G. The number of components is bounded by T, so here we need $O(\frac{T}{(1-2\alpha)^2}\log\frac{n}{\delta})$ queries.

We now show that by using sufficient many queries at each check, we make sure that when the hypothesis has large error a non-indicator is identified, so that the hypothesis is updated at most k times. If $d(h,h^*) \le \epsilon/4$, then

$$\Pr[h(x) \neq h^*(x) | x \in X \setminus F] \le \epsilon/3$$

and thus

$$\Pr[h(x) \neq \ell^*(x) | x \in X \setminus F] \le \alpha + (1-2\alpha)\frac{\epsilon}{3}.$$

Then w.h.p. when $|S| = O(\frac{1}{(1-2\alpha)^2\epsilon^2}\log\frac{k}{\delta})$,

$$\mathsf{err}_S(h) \le \alpha + \frac{1-2\alpha}{2}\epsilon.$$

Therefore, if $\mathsf{err}_S(h) > \alpha + \frac{1-2\alpha}{2}\epsilon$, we have $d(h,h^*) \ge \epsilon/4$. This means on at least $\epsilon/16$ fraction of the examples in S we have $h(x) \neq h^*(x)$. By sampling $\frac{100}{\epsilon}\log\frac{k}{\delta}$ times from S and then picking $\min(k+1, |x|_1)$ variables in the sampled x, w.h.p. we will eventually pick such an example, and pick at least one indicator in it, whose type is different from the nearest indicator. Then we find a path connecting positive and negative indicators, and discover a non-indicator by binary search. So we need

$$O\left(\frac{1}{(1-2\alpha)^2\epsilon^2}\log\frac{k}{\delta}+\frac{k}{(1-2\alpha)^2\epsilon}\log\frac{k}{\delta}\right)$$

queries each time we check and update the hypothesis.

Query Complexity and Running Time: Since when the hypothesis has large error a non-indicator is identified, it is checked and updated at most k times. Then the number of queries follows by bounding T by $\log|C_{D,\mathcal{X}}(\epsilon)|$. Notice that T is the number of connected components in $G_{\mathrm{com}}(U \setminus F)$ (instead of $G_{\mathrm{com}}(U)$) after removing the non-indicators. However, when U is sufficiently large, w.h.p. the probability mass of F is at most $\epsilon/8$, i.e. all significant edges appear in the graph, so we still have $T \le \log|C_{D,\mathcal{X}}(\epsilon)|$. Building, checking and updating the hypothesis all take polynomial time, so the algorithm runs in polynomial time. □

3.3. Extension of Algorithm 1

The key idea is that we can still build the indicator graph by majority voting, and then enumerate over compatible hypotheses. A description is provided below, while the details are provided in Algorithm 7 and Theorem 7.

First, we draw enough examples to make sure that all non-indicators are significant, and use majority voting to correctly decide the types of significant variables. This requires $\tilde{O}(\frac{1}{\epsilon_0(1-2\alpha)^2})$ labeled examples. We could construct the indicator graph as we did in the noise-free setting. However, since we only label significant variables (i.e. identify them as potential indicators), possibly not enough variables are labeled. Then many of the components in the graph that results from removing all non-indicators from the commonality graph are not connected to any labeled variables, and thus all the hypotheses built according to minimal vertex covers in the indicator graph do not have small errors.

To address this, we take an additional step to add more variables to the potential indicator sets before building the the indicator graph. More precisely, after removing the significant variables from the commonality graph, we know that every component must contain only indicators of one type. If there are $\tilde{O}(\frac{1}{(1-2\alpha)^2})$ examples in a component, we can safely decide the type of the variables in it to be the majority label of these examples, and add these variables to the corresponding potential indicator sets. So we draw $\tilde{O}(\frac{T}{\epsilon(1-2\alpha)^2})$ examples to make sure only components with probability mass at most $\epsilon/(8T)$ are not labeled. Here T is the number of connected components in the graph that results from removing all non-indicators from $G_{\mathrm{com}}(U)$, and can be bounded by $\log|C_{D,\mathcal{X}}(\epsilon)|$. Now hypotheses built on the indicator graph that are compatible with the unlabeled data can only make mistakes on these components.

Algorithm 7 Semi-supervised Learning with Random Classification Noise via Enumeration

```
Input: data sets $U$ and $L$, parameters $\epsilon$, $\delta$, $\epsilon_0$, $\alpha$
Set $G = G_{\mathrm{com}}(U), V_+^0 = V_-^0 = \emptyset$
Set $L^0 = \mathrm{sample}\left(\frac{100}{\epsilon_0(1-2\alpha)^2}\log^2\frac{n}{\delta}, L\right)$, $L = L \setminus L^0$
for each variable $v$ appearing in more than $\frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}$ examples in $L^0$ do
  Let $l$ be the majority label of these examples
  Set $V_l^0 = V_l^0 \cup \{v\}$
$V_+ = V_+^0, V_- = V_-^0$
for each component $R$ in $G \setminus (V_+^0 \cup V_-^0)$ do
  if $\exists \ge \frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}$ examples from $L$ in $R$ then
    Let $l$ be the majority label of these examples
    Set $V_l = V_l \cup R$
Set $G_I = G_{\mathrm{ind}}(G, V_+, V_-)$
for each minimal vertex cover $S$ of $G_I$ do
  Set $G' = G \setminus S$, $V'_+ = V_+ \setminus S$, $V'_- = V_- \setminus S$
  Set $h_+ = \{v \in G' : \exists u \in V'_+,\ u \leftrightarrow_{G'} v\}$
  if $h = (h_+, G' \setminus h_+)$ is compatible and $\mathsf{err}_L(h) \le \alpha + (1-2\alpha)\frac{\epsilon}{2}$ then
    break
Output: hypothesis $h = (h_+, G' \setminus h_+)$
```

After the additional step, we build the indicator graph and enumerate over compatible hypotheses built according to minimal vertex covers. To ensure that the output hypothesis has small error, we must be able to distinguish under the noise between a hypothesis h with $d(h, h^*) > \epsilon$ and one with $d(h, h^*) \le \epsilon/4$, i.e. to distinguish between a hypothesis h with

$$\Pr[h(x) \ne \ell(x)] \ge \alpha + (1-2\alpha)\epsilon$$

and one with

$$\Pr[h(x) \ne \ell(x)] \le \alpha + (1-2\alpha)\epsilon/4.$$

So we need to bound the deviation of the empirical error by $O(\epsilon(1-2\alpha))$, which requires $\tilde{O}(\frac{1}{(1-2\alpha)^2\epsilon^2})$ labeled examples. Union bounding over all compatible hypothesis introduces an extra term $O(\log|C_{D,\mathcal{X}}(\epsilon)|)$.

Theorem 7. *For any distribution $D_{X,Y}$ over $\{0,1\}^n \times \{-1,1\}$ and target concept $h^* \in C$ in the random classification noise model such that h^* has at most k non-indicators, and the minimum non-indicator probability is ϵ_0, if $m_u = O(\frac{n^2}{\epsilon}\log\frac{n}{\delta})$ and*

$$m_l = O\left(\frac{1}{(1-2\alpha)^2}\left[\frac{1}{\epsilon_0} + \frac{\log|C_{D,\mathcal{X}}(\epsilon)|}{\epsilon^2}\right]\log^2\frac{n}{\delta}\right)$$

then with probability at least $1-\delta$, Algorithm 7 outputs a two-sided disjunction $h \in C$ such that $\chi(h,U)=1$, and $\mathsf{err}_D(h) \le \alpha+\epsilon$. Furthermore, the algorithm runs in polynomial time when $k = O(\log n)$.

Proof. **Reducing to the Noise-free Setting:** We show that all non-indicators are in $V_+ \cup V_-$, and the types of all indicators in $V_+ \cup V_-$ are decided correctly, so that the indicator graph and the hypotheses are built correctly as in the noise-free setting. When

$$|L^0| \ge \frac{100}{\epsilon_0(1-2\alpha)^2}\log^2\frac{n}{\delta}$$

all non-indicators are in $V_+^0 \cup V_-^0$ and thus in $V_+ \cup V_-$. Also, for any indicator in $V_+^0 \cup V_-^0$, its type is not decided correctly only when the labels of more than half of the examples containing it are flipped, which happens with probability at most $\delta/(8n)$. By union bound, we have that with probability at least $1-\delta/8$, the types of all indicators in $V_+^0 \cup V_-^0$ are decided correctly.

Furthermore, the types of all indicators added to $V_+ \cup V_-$ in later steps are also correct. Since all non-indicators are in $V_+^0 \cup V_-^0$, any component R in $G \setminus (V_+^0 \cup V_-^0)$ must contain only one type of indicators. Then with probability at least $1-\delta/8$, the types of all R with $O(\frac{1}{(1-2\alpha)^2}\log\frac{n}{\delta})$ labeled examples are decided correctly.

Generalization Error: First note that when

$$m_u \ge \frac{10n^2}{\epsilon}\log\frac{n}{\delta}$$

all hypotheses compatible with U fall in $C_{D,\chi}(\epsilon)$ with probability at least $1-\delta/8$. Fix a hypothesis $h' \in C_{D,\chi}(\epsilon)$ with $d(h',h^*) > \epsilon$. By Hoeffding bound, when

$$|L| \ge \frac{10}{(1-2\alpha)^2\epsilon^2}\log\frac{|C_{D,\chi}(\epsilon)|}{\delta}$$

we have

$$\Pr\left[\mathsf{err}_L(h') \le \alpha + \frac{1-2\alpha}{2}\epsilon\right] \le \frac{\delta}{8|C_{D,\chi}(\epsilon)|}.$$

Then by union bound, with probability at least $1-\delta/8$, all $h \in C_{D,\chi}(\epsilon)$ with $d(h,h^*) > \epsilon$ have $\mathsf{err}_L(h) > \alpha + \frac{1-2\alpha}{2}\epsilon$. So a hypothesis h satisfying the exit condition satisfies $d(h,h^*) \le \epsilon$ and thus $\mathsf{err}_D(h) \le \alpha+\epsilon$.

Now it suffices to show that we can always find a suitable minimal vertex cover that leads to such a hypothesis. Note that at least one endpoint of every edge in G_I must be a non-indicator, so there must be a subset $\tilde{S}$ of non-indicators that is a minimal vertex cover of G_I. Let $\tilde{h} = (\tilde{h}_+, \tilde{h}_-)$ be the hypothesis formed from the minimal vertex cover $\tilde{S}$. We show that $\tilde{h}$ is compatible and $\mathsf{err}_L(\tilde{h}) \le \alpha + \frac{1-2\alpha}{2}\epsilon$.

If an example contained both positive and negative indicators, this would imply an edge still present in G_I, which is impossible. So $\tilde{h}$ is compatible with U. Next we show that $\mathsf{err}_L(\tilde{h}) \le \alpha + \frac{1-2\alpha}{2}\epsilon$. Consider the components in $G \setminus (V_+ \cup V_-)$. Suppose when building the hypothesis, the variables in some components $R_1, R_2, ..., R_t$ are not correctly decided. First, there are just a few such components. Each such component is either connected to non-indicators in $G \setminus \tilde{S}$ that have label of a different type, or is not connected to any labeled variable but the variables in it are positive indicators. Then such components are components in the graph $\hat{G}$ that results from removing all non-indicators from the commonality graph G. So t is no larger than the number T of components in $\hat{G}$, and we have

$$t \le T \le \log|C_{D,\chi}(\epsilon)|$$

when $m_u = O(\frac{n^2}{\epsilon}\log\frac{n}{\delta})$. Second, each such component has small probability mass. These components are also components in $G \setminus (V_+^0 \cup V_-^0)$. When

$$|L| \ge \frac{10T}{\epsilon(1-\alpha)^2}\log^2\frac{n}{\delta}$$

all components in $G \setminus (V_+^0 \cup V_-^0)$ with probability more than $\epsilon/(8T)$ have at least $\frac{10}{(1-2\alpha)^2}\log\frac{n}{\delta}$ labeled examples and thus are added to $V_+ \cup V_-$. Then each of $R_1, ..., R_t$ has probability at most $\epsilon/(8T)$. This means

$$d(\tilde{h}, h^*) \le \Pr[x \in \cup_i R_i] \le \epsilon/8$$

and

$$\Pr[\tilde{h}(x) \ne \ell(x)] \le \alpha + (1-2\alpha)\epsilon/8.$$

Then w.h.p. we have

$$\mathsf{err}_L(\tilde{h}) \le \alpha + \frac{1-2\alpha}{2}\epsilon.$$

Therefore, we are guaranteed to find such a hypothesis when the algorithm stops.

Sample Complexity and Running Time: To ensure the indicator graph is built correctly, we need

$$|L^0| = O\left(\frac{1}{\epsilon_0(1-2\alpha)^2}\log^2\frac{n}{\delta}\right).$$

To ensure the hypothesis output has small error, we need

$$|L| = O\left(\frac{T}{(1-2\alpha)^2\epsilon^2}\log^2\frac{n}{\delta}\right).$$

Then the sample complexity of the algorithm follows from $T \le |C_{D,\chi}(\epsilon)|$. The time for building the indicator graph is clearly polynomial. The running time for checking the hypotheses is the same as in the noise-free setting, so it is polynomial when $k = O(\log n)$. Therefore, the algorithm runs in polynomial time when $k = O(\log n)$. □