Appendix

A LOWER BOUNDS

A.1 Noise Sensitivity Lower Bound for Some Symmetric Functions

In this section we show that some families of symmetric functions on subsets of half the bits are hard to improperly learn in an information-theoretic sense, and prove Theorem 1.2.

It can be easily checked that the distribution $\mathcal{D}$ has the following properties:

- For every $z \in \{0,1\}^{n/2}$ and a random string $\boldsymbol{x} \sim \mathcal{D}$, $\boldsymbol{x}^z$ is distributed as a uniformly random string over $\{0,1\}^{n/2}$.
- For every pair of strings $z, z' \in \{0,1\}^{n/2}$ and a random string $\boldsymbol{x} \sim \mathcal{D}$, the random strings $\boldsymbol{x}^z$ and $\boldsymbol{x}^{z'}$, restricted to the coordinates where z and z' disagree, are ρ-noisy copies of each other.
- To construct the distribution of $\boldsymbol{x}^{z'}$ from $\boldsymbol{x}^z$, one can apply ρ-noise to the coordinates of $\boldsymbol{x}^z$ in those coordinates where z and z' differ (and just read off the coordinates of $\boldsymbol{x}^z$ where they are the same).
- In fact, $\mathcal{D}^z$ is identical to $\mathcal{D}$ for every $z \in \{0,1\}^{n/2}$. However, the distribution of labeled examples $\langle \boldsymbol{x}, f^z(\boldsymbol{x})\rangle$ where $\boldsymbol{x} \sim \mathcal{D}^z$ depends on z. The distribution of labeled examples after attribute noise $\langle N_\rho^z(\boldsymbol{x}), f^z(\boldsymbol{x})\rangle$ is independent of z; the marginal distribution on $N_\rho^z(\boldsymbol{x})$ is $\mathcal{D} = \mathcal{D}^z$.

A.1.1 Noise Sensitivity

For concreteness, recall that the **noise operator at** ρ **on** S is denoted by $N_{S,\rho}(x)$ is a random string such that $N_{S,\rho}(x)_i$ is a uniform random bit ρ-correlated with x_i if $i \in S$, and $N_{m,\rho}(x)_i = x_i$ with probability 1 for $i \notin S$ (cf. O'Donnell (2014)). We now prove Claim 3.1:

Proof of Claim 3.1. Note that, for every x, $N_{\rho/15}(x)$ is distributed as $N_{\boldsymbol{T},\rho}(x)$, where $\boldsymbol{T}$ is a set where each coordinate is included independently with probability 1/15. It follows that

$$\begin{aligned}
\mathbb{NS}_{\rho/15}(f) &= \Pr_{\boldsymbol{y}\sim\mathcal{U}_{n/2}}[f(\boldsymbol{y}) \neq f(N_{\rho/15}(\boldsymbol{y}))] \\
&= \Pr_{\boldsymbol{y}\sim\mathcal{U}_{n/2}}[f(\boldsymbol{y}) \neq f(N_{\boldsymbol{T},\rho}(\boldsymbol{y}))] \\
&= \mathrm{E}_{\boldsymbol{T}}[\mathbb{NS}_{\boldsymbol{T},\rho}(f)].
\end{aligned}$$

By a Chernoff bound, $\Pr[|\boldsymbol{T}| \leq n/14] \geq 1 - 2^{-\Omega(n)}$. Thus, for a set S such that $|S| = n/14$, we have

$$\begin{aligned}
\mathbb{NS}_{\rho/15}(f) &= \mathrm{E}_{\boldsymbol{T}}[\mathbb{NS}_{\boldsymbol{T},\rho}(f)] \\
&= \mathrm{E}_{\boldsymbol{T}}[\mathbb{NS}_{\boldsymbol{T},\rho}(f) \mid |\boldsymbol{T}| \leq n/14] \Pr_{\boldsymbol{T}}[|\boldsymbol{T}| \leq n/14] \\
&+ \mathrm{E}_{\boldsymbol{T}}[\mathbb{NS}_{\boldsymbol{T},\rho}(f) \mid |\boldsymbol{T}| > n/14] \Pr_{\boldsymbol{T}}[|\boldsymbol{T}| > n/14] \\
&\leq \mathbb{NS}_{S,\rho}(f) \Pr[|\boldsymbol{T}| \leq n/14] + 2^{-\Omega(n)} \\
&\leq \mathbb{NS}_{S,\rho}(f)(1 + o(1)),
\end{aligned}$$

where we used the fact that $\mathbb{NS}_{S,\rho}$ is nondecreasing as $|S|$ increases. (Since we assumed that f is symmetric, only $|S|$ matters.) Dividing both sides by the $(1 + o(1))$ factor yields the claim. □

Remark A.1. *The symmetric assumption can be relaxed by noting that the bound works for any function that is roughly balanced over the uniform distribution, since the noise sensitivity of such functions is* $\Omega(\min\{\Pr[f(\boldsymbol{x}) = 0], \Pr[f(\boldsymbol{x}) = 1]\})$. *Roughly speaking, this result asserts that we cannot learn with error smaller than the noise sensitivity.*

A.2 Maximum Sensitivity Lower Bound for Conjunctions

In this section we show a lower bound for improper list learning of conjunctions and by proving a more specific version of Theorem 1.5. We will use the same notation as in Section A.1.

Theorem A.2. *Let $k > 0$ be an integer, $\epsilon > 0$, and let $\mathcal{C}_k$ be the set of all conjunctions over k bits out of n bits $f : \{0,1\}^n \to \{0,1\}$. If the attribute noise is $\rho = \frac{1}{k} > 8\epsilon$, then any net $\mathcal{H}$ of functions satisfying*

$$\max_{z\in\{0,1\}^{n/2}} \min_{h\in\mathcal{H}} \Pr_{\boldsymbol{x}\sim\mathcal{D}^z}[f^z(\boldsymbol{x}) \neq h(\boldsymbol{x})] < \epsilon$$

must have $|\mathcal{H}| > 2^{\Omega(k)}$.

Proof. Suppose that the distribution $\mathcal{D}^z$ over $\{0,1\}^{2k}$ is such that

- The coordinates in $\boldsymbol{x}^z$ are drawn independently at random with bias $1/k$. That is, $\boldsymbol{x}^z \sim \mu_{k,1/k}$, where $\mu_{n,p}$ denotes the p-biased distribution over $\{0,1\}^n$.
- The coordinates in $\boldsymbol{x}^{\overline{z}}$ are ρ-noisy copies of $\boldsymbol{x}^z$; specifically, each bit $\boldsymbol{x}_i^{\overline{z_i}}$ is a ρ-noisy copy of $\boldsymbol{x}_i^{z_i}$.

We will show that if z is unknown, and we see labeled examples according to f^z under $\mathcal{D}^z$ with ρ-bounded attribute noise, then list-learning to small accuracy requires an exponential size list. That is, for every set of functions $\mathcal{H}$ (our proposed net), the quantity

$$\max_{z\in\{0,1\}^{n/2}} \min_{h\in\mathcal{H}} \Pr_{\boldsymbol{x}\sim\mathcal{D}^z}[f^z(\boldsymbol{x}) \neq h(\boldsymbol{x})]$$

is "large" if $|\mathcal{H}|$ is sub-exponential in k.

For f^z with respect to $\mathcal{D}^z$, given x, the attribute noise $N_\rho^z(x)$ is as follows: we apply ρ-noise to each $x_i^{z_i}$, and no noise to $x_i^{\overline{z_i}}$. It follows that for *every* $\mathcal{D}^z$, the resulting distribution over the *labeled* examples is the same. We define $\mathcal{D}$ to be distribution[5] on $\{0,1\}^n$ such that, for each i, x_i^0 and x_i^1 are ρ-correlated random bits with bias $(1-\rho)(1/k) + \rho(1-1/k)$, and the k pairs (x_i^0, x_i^1) are chosen independently. It can be easily checked that the distribution $\mathcal{D}$ has the following properties:

- For every $z \in \{0,1\}^{n/2}$ and a random string $\boldsymbol{x} \sim \mathcal{D}$, $\boldsymbol{x}^z$ is distributed as a uniformly random string over $\{0,1\}^{n/2}$.
- For every pair of strings $z, z' \in \{0,1\}^{n/2}$ and a random string $\boldsymbol{x} \sim \mathcal{D}$, the random strings $\boldsymbol{x}^z$ and $\boldsymbol{x}^{z'}$, restricted to the coordinates where z and z' disagree, are ρ-noisy copies of each other.
- To construct $\boldsymbol{x}^{z'}$ from $\boldsymbol{x}^z$, one can apply ρ-noise to the coordinates of $\boldsymbol{x}^z$ in those coordinates where z and z' differ (and just read off the coordinates of $\boldsymbol{x}^z$ where they are the same).
- In fact, $\mathcal{D}^z$ is identical to $\mathcal{D}$ for every $z \in \{0,1\}^{n/2}$. However, the distribution of labeled examples $\langle \boldsymbol{x}, f^z(\boldsymbol{x})\rangle$ where $\boldsymbol{x} \sim \mathcal{D}^z$ depends on z. The distribution of labeled examples after attribute noise $\langle N_\rho^z(\boldsymbol{x}), f^z(\boldsymbol{x})\rangle$ is independent of z; the marginal distribution on $N_\rho^z(\boldsymbol{x})$ is $\mathcal{D} = \mathcal{D}^z$.

Unlike the uniform distribution case, when we consider the accuracy of a function in the net on a conjunction, the distribution under which we calculate the error depends on the conjunction. We compute the following quantities first:

- The probability of the all-0's string in the true distribution is $(1-1/k)^k(1-\rho)^k$; the all 0's string in drawn in the conjunction bits, and no flips occur in the noisy version.

[5] Actually, this is the same as $\mathcal{D}^z$.

- The probability of a string of all-0's, except for $x_i^b = 1$ depends on the conjunction. If $z_i = b$ (x_i^b is in the conjunction), then the probability mass assigned is $(1-1/k)^{k-1}(1/k)(1-\rho)^{k-1}\rho$. If $z_i = 1-b$ (x_i^b is not in the conjunction), then the probability mass assigned is $(1-1/k)^k(1-\rho)^{k-1}\rho$.

Consider the values of a function f on these standard basis strings.

- If $f(e_{i,b}) = 1$ ($x_i^b = 1$) and $z_i = b$ (x_i^b is in the conjunction), f incorrectly computes the conjunction. The contribution to the error is $(1-1/k)^{k-1}(1/k)(1-\rho)^{k-1}\rho$.
- If $f(e_{i,1-b}) = 0$ ($x_i^b = 0$) and $z_i = b$ (x_i^b is in the conjunction), f incorrectly computes the conjunction. The contribution to the error is $(1-1/k)^k(1-\rho)^{k-1}\rho$.

So for every conjunction, a false 0 is roughly k times as costly as a false 1. To make the error less than $(1-1/k)^{k-1}(1/k)(1-\rho)^{k-1}\rho\cdot(99k/100)$, there must be a function in the net that has no false 0's and at most $99k/100$ false 1's on these strings. A function in the net covers the most conjunctions by taking f to be 1 on $k+99k/100 = 199k/100$ of these strings and 0 on the other $k/100$. A function is covered if its bits are correspond to those with ones. There are $2^{99k/100}$ conjunctions covered, but 2^k conjunctions in total, so any net must have $2^{k/100}$ functions in it to achieve error below $(1-1/k)^{k-1}(1/k)(1-\rho)^{k-1}\rho\cdot(99k/100)$. Taking $\rho = 1/k$, this is at least

$$\begin{aligned}&(1-1/k)^{k-1}(1/k)(1-1/k)^{k-1}(1/k)\cdot(99k/100)\\&=99(1-1/k)^{2k-2}/(100k)\\&\geq 99/(100e^2k)\\&\geq 1/(8k),\end{aligned}$$

so the error is at least $\rho/8$. We need $\rho < 8\epsilon$ for a sub-exponential size net. □

B PROOF OF THEOREM 1.6

Our main theorem (the formal version of Theorem 1.6) is the following

Theorem B.1. *For any positive integer k and any real numbers $0 < \epsilon, \delta < 1$, $0 < \gamma \leq 1/2$, there exists a randomized algorithm which, with probability at least $1-\delta$, list-learns k-conjunctions with accuracy $1-\epsilon$, with sample complexity $\tilde{O}(k^4\log(1/\delta)/(\epsilon^9\gamma^4))$ and time complexity $\max\{\tilde{O}(n^2k^4\log(1/\delta)/(\epsilon^9\gamma^4)), O((32k^2/\epsilon^5\gamma^2)^k)\}$, in the attribute-noise model with bit noise rate $0 \leq \nu_i < \frac{1}{2}-\gamma$ for every $1 \leq i \leq n$, under the assumption that the ground-truth distribution is k'-wise independent for some $k' \geq 2$.*

In the rest of this section, we set $m := 32k^2/(\epsilon^5\gamma^2)$. Also, by a simple application of Chernoff bound, if we draw $M := O(k^4\log n\log(1/\delta)/(\epsilon^9\gamma^4))$ random examples from the noisy example oracle $\tilde{\mathrm{EX}}(c, D)$, then with probability at least $1-\delta$, we can estimate quantities such as $\mathrm{E}_{\tilde{D}_1}[x_i]$, $\mathrm{E}_{\tilde{D}_1}[x_i\cdot x_j]$ with additive accuracy $O(1/(\epsilon m))$ for every $1 \leq i, j \leq n$. To ease exposition, from now on, we condition our arguments on this event happening.

Since every k'-wise independent distribution for $k' \geq 2$ is also pairwise independent, it is enough to prove Theorem B.1 for $k' = 2$. The algorithm for the pairwise independence test is given by Algorithm 2.

In the rest of this section, we use the notation $\widehat{H}$ to denote the estimate of a quantity H using random examples sampled from the noisy example oracle $\tilde{\mathrm{EX}}(c, D)$.

First of all, since we include the trivial functions **0** and **1** in the output list, our learning algorithm succeeds trivially whenever the target concept is ϵ-close to either **0** or **1**. Therefore, from now on, we assume that $\epsilon \leq \Pr_D[c(x) = 1] \leq 1-\epsilon$.

B.1 Conjunction Bits with Low Label-Sensitivity

The next lemma shows that using bits in S we can get a conjunction which approximates the target concept well.

Lemma B.2. *Let $c = \wedge_{i \in c} x_i$ be the target concept, and let c' be the set of bits obtained by removing from c the set of bits eliminated in Line 8 of Algorithm 1. Then conjunction c' is $\epsilon/2$-close to c, i.e.* $\Pr_D[c(x) \neq c'(x)] \leq \epsilon$.

Proof. First note that eliminating non-conjunction bits can not worsen the performance of our learning algorithm, so we can focus on the effect of eliminating a conjunction bit from S in Line 8.

Since c' is a subset of c,

$$\begin{aligned}\Pr_D[c(x) \neq c'(x)] &= \Pr_D[c'(x) = 1 \text{ and } \exists i \in c \setminus c' \text{ such that } x_i = 0] \\ &\leq \Pr_D[\exists i \in c \setminus c' \text{ such that } x_i = 0] \\ &\leq \sum_{i \in c \setminus c'} \Pr_D[x_i = 0]. \qquad \text{(by union bound)} \end{aligned} \tag{1}$$

We can upper bound $\Pr_D[x_i = 0]$ for any $i \in c \setminus c'$ as

$$\begin{aligned}\Pr_D[x_i = 0] &= \Pr_D[c(x) = 0] \cdot \Pr_{D_0}[x_i = 0] + \Pr_D[c(x) = 1] \cdot \Pr_{D_1}[x_i = 0] \\ &= \Pr_D[c(x) = 0] \cdot \Pr_{D_0}[x_i = 0] \leq \Pr_{D_0}[x_i = 0].\end{aligned}$$

On the other hand, in terms of quantities over the observed distribution $\tilde{D}$, we have

$$\begin{aligned}\Pr_{\tilde{D}_0}[x_i = 0] &= (1 - \nu_i) \Pr_{D_0}[x_i = 0] + \nu_i \Pr_{D_0}[x_i = 1] \\ &= (1 - \nu_i) \Pr_{D_0}[x_i = 0] + \nu_i (1 - \Pr_{D_0}[x_i = 0]) \\ &= (1 - 2\nu_i) \Pr_{D_0}[x_i = 0] + \nu_i,\end{aligned}$$

and

$$\Pr_{\tilde{D}_1}[x_i = 0] = (1 - \nu_i) \Pr_{D_1}[x_i = 0] + \nu_i \Pr_{D_1}[x_i = 1] = \nu_i \Pr_{D_1}[x_i = 1] \leq \nu_i.$$

Using $O(\log n \log(1/\delta) k^2 / \epsilon^3 \gamma^2) = o(M)$ random examples, we can, with probability at least $1 - \delta$, obtain $\Omega(\log n \log(1/\delta) k^2 / \epsilon^2 \gamma^2)$ random negative examples and $\Omega(\log n \log(1/\delta) k^2 / \epsilon^2 \gamma^2)$ random positive examples, and get an estimate of $\widehat{\mathrm{LS}}_i$ with $|\widehat{\mathrm{LS}}_i - \mathrm{LS}_i| \leq \epsilon\gamma/(2k)$ for every $1 \leq i \leq n$. Since bit-i was eliminated from S, we

$$\mathrm{LS}_i \leq \widehat{\mathrm{LS}}_i + \epsilon\gamma/(2k) < 2\epsilon\gamma/k.$$

Combining this with bounds on $\Pr_{\tilde{D}_0}[x_i = 0]$ and $\Pr_{\tilde{D}_1}[x_i = 0]$, we have

$$2\epsilon\gamma/k > \mathrm{LS}_i = \Pr_{\tilde{D}_0}[x_i = 0] - \Pr_{\tilde{D}_1}[x_i = 0] \geq (1 - 2\nu_i) \Pr_{D_0}[x_i = 0] > 2\gamma \Pr_{D_0}[x_i = 0],$$

where the last step follows from the fact that $\nu_i < \frac{1}{2} - \gamma$. Therefore we have $\Pr_{D_0}[x_i = 0] < \epsilon/k$.

Finally, plugging the above upper bound on $\Pr_{D_0}[x_i = 0]$ into inequality (1) completes the proof. □

B.2 Pairwise Independent Bits

A simple but important observation is that, if the target concept conjunction is $c = \wedge_{i \in c} x_i$, then in the observed distribution $\tilde{D}_1$ of positive examples, the bits in c are totally independent. This is because, when restricting to bits in c, D_1 is supported on a single vector 1^k. After applying the (bit-wise independent) attribute noise, $\tilde{D}_1$ is a product distribution when restricting to bits in c.

As it is computationally expensive to check total independence among the conjunction bits on $\tilde{D}_1$, and pairwise independence suffices for our concentration argument, we check pairwise independence in Algorithm 2 by estimating the covariances between each pair of bits.

Lemma B.3. *With probability at least $1 - \delta$, the followings hold: the output S of Algorithm 2 includes every bit in c; and conversely, every pair of bits X_i and X_j in S are close to being pairwise independent in the sense that* $|\mathbf{Cov}_{\tilde{D}_1}(X_i, X_j)| \leq 1/(4\epsilon m)$.

Claim B.4. *Let $D' : \{0,1\}^n \to \mathbb{R}^{\geq 0}$ be a distribution and let $X \in \{0,1\}^n$ be the random variable obtained from sampling according to D'. Then, for any $0 \leq \epsilon \leq 1/2$, if $\Pr[X_i = 1] \leq \epsilon$ for some $1 \leq i \leq n$, then $|\mathbf{Cov}(X_i, X_j)| \leq \epsilon$ for every $i \neq j$. The same bound holds when $\Pr[X_i = 0] \leq \epsilon$.*

Proof. Let $p_0 = \Pr[X_i = 0 \wedge X_j = 0]$, $p_1 = \Pr[X_i = 0 \wedge X_j = 1]$, $p_2 = \Pr[X_i = 1 \wedge X_j = 0]$, and $p_3 = \Pr[X_i = 1 \wedge X_j = 1]$. Then $p_2 + p_3 = \Pr[X_i = 1] \leq \epsilon$ and $\mathbf{Cov}(X_i, X_j) = p_3 - (p_2 + p_3)(p_1 + p_3)$. Therefore, $\mathbf{Cov}(X_i, X_j) \geq -(p_2 + p_3)(p_1 + p_3) \geq -(p_2 + p_3) = -\epsilon$. On the other hand, $\mathbf{Cov}(X_i, X_j) \leq p_3 - p_3^2 \leq \epsilon - \epsilon^2 \leq \epsilon$, as $x - x^2$ is increasing for $0 \leq x \leq 1/2$.

The case of $\Pr[X_i = 0] \leq \epsilon$ follows directly from the identity $\mathbf{Cov}(1 - X_i, 1 - X_j) = \mathbf{Cov}(X_i, X_j)$. □

Claim B.5. *Let distribution D' and random variable X be the same as in Claim B.4. For any pair of distinct bits i and j, let $\widehat{\mathbf{Cov}(X_i, X_j)} := \widehat{\mathrm{E}[X_i \cdot X_j]} - \widehat{\mathrm{E}[X_i]} \cdot \widehat{\mathrm{E}[X_j]}$ be the estimated covariance of X_i and X_j. Then the estimate error can be upper bounded as*

$$|\widehat{\mathbf{Cov}(X_i, X_j)} - \mathbf{Cov}(X_i, X_j)| \leq |\widehat{\mathrm{E}[X_i \cdot X_j]} - \mathrm{E}[X_i \cdot X_j]| + 2|\widehat{\mathrm{E}[X_i]} - \mathrm{E}[X_i]| + 2|\widehat{\mathrm{E}[X_j]} - \mathrm{E}[X_j]|.$$

Proof. Let $\Delta X_i = \widehat{\mathrm{E}[X_i]} - \mathrm{E}[X_i]$ and $\Delta X_j = \widehat{\mathrm{E}[X_j]} - \mathrm{E}[X_j]$. Then we have

$$\begin{aligned}\left|\widehat{\mathrm{E}[X_i]} \cdot \widehat{\mathrm{E}[X_j]} - \mathrm{E}[X_i] \cdot \mathrm{E}[X_j]\right| &= |\Delta X_i \mathrm{E}[X_j] + \Delta X_j \mathrm{E}[X_i] + \Delta X_i \Delta X_j| \\ &\leq |\Delta X_i|(\mathrm{E}[X_j] + |\Delta X_j|) + |\Delta X_j|(\mathrm{E}[X_i] + |\Delta X_i|) \\ &\leq 2|\Delta X_i| + 2|\Delta X_j|,\end{aligned}$$

because both $\widehat{\mathrm{E}[X_i]}$ and $\mathrm{E}[X_i]$ are real numbers between 0 and 1. Now the bound in the claim follows directly from

$$\begin{aligned}\left|\widehat{\mathbf{Cov}(X_i, X_j)} - \mathbf{Cov}(X_i, X_j)\right| &= \left|\widehat{\mathrm{E}[X_i \cdot X_j]} - \widehat{\mathrm{E}[X_i]} \cdot \widehat{\mathrm{E}[X_j]} - \mathrm{E}[X_i \cdot X_j] + \mathrm{E}[X_i] \cdot \mathrm{E}[X_j]\right| \\ &\leq \left|\widehat{\mathrm{E}[X_i]} \cdot \widehat{\mathrm{E}[X_j]} - \mathrm{E}[X_i] \cdot \mathrm{E}[X_j]\right| + \left|\widehat{\mathrm{E}[X_i \cdot X_j]} - \mathrm{E}[X_i \cdot X_j]\right|.\end{aligned}$$ □

Proof of Lemma B.3. As mentioned earlier, if we draw enough examples from the noisy example oracle, we can esitmate quantities such as $E_{\tilde{D}_1}[X_i]$ and $E_{\tilde{D}_1}[X_i \cdot X_j]$ accurately enough. More specifically, using $O(\log(1/\delta) \log n (\epsilon m)^2/\epsilon) = \tilde{O}(k^4 \log(1/\delta)/(\epsilon^9 \gamma^4))$ random samples, with probability at least $1 - \delta$, we have $|\widehat{\mathrm{E}_{\tilde{D}_1}[X_i]} - \mathrm{E}_{\tilde{D}_1}[X_i]| \leq 1/(48\epsilon m)$ for every $1 \leq i \leq n$ and $|\widehat{\mathrm{E}_{\tilde{D}_1}[X_i \cdot X_j]} - \mathrm{E}_{\tilde{D}_1}[X_i \cdot X_j]| \leq 1/(24\epsilon m)$ for every pair of distinct $1 \leq i, j \leq n$. Then for every pair of conjunction bits $i, j \in c$ or a pair of conjunction bit $i \in c$ and and a non-conjunction bit $j \in [n] \setminus c$, we always have $\mathbf{Cov}_{\tilde{D}_1}(X_i, X_j) = 0$. By Claim B.5, $|\widehat{\mathbf{Cov}_{\tilde{D}_1}(X_i, X_j)}| \leq 1/(8\epsilon m)$, so any conjunction bit can never be removed from S in line 12 of Algorithm 2. On the other hand, by Claim B.4 and Claim B.5 and analogous calculations, for any pair of bits X_i and X_j that are in the output S of Algorithm 2, it must be the case that $|\mathbf{Cov}_{\tilde{D}_1}(X_i, X_j)| \leq 1/(4\epsilon m)$. □

B.3 Bounding the Size of S

Claim B.6. *For every surviving bit X_i in S, we have $\mathrm{E}_{\tilde{D}_1}[X_i] - \mathrm{E}_{\tilde{D}}[X_i] > \epsilon^2\gamma/(2k)$.*

Proof. If x_i is in S, then by a similar argument as in the proof of Lemma B.2, $\mathrm{LS}_i \geq \widehat{\mathrm{LS}}_i - \epsilon\gamma/(2k) \geq \epsilon\gamma/(2k)$.

Now, by the definitions of $\mathrm{E}_{\tilde{D}}[X_i]$ and $\mathrm{E}_{\tilde{D}_1}[X_i]$,

$$\begin{aligned}
\mathrm{E}_{\tilde{D}_1}[X_i] - \mathrm{E}_{\tilde{D}}[X_i] &= \mathrm{E}_{\tilde{D}_1}[X_i] - (\Pr_{\tilde{D}}[c=0]\cdot \mathrm{E}_{\tilde{D}_0}[X_i] + \Pr_{\tilde{D}}[c=1]\cdot \mathrm{E}_{\tilde{D}_1}[X_i]) \\
&= (1-\Pr_{\tilde{D}}[c=1])(\mathrm{E}_{\tilde{D}_1}[X_i] - \mathrm{E}_{\tilde{D}_0}[X_i]) \\
&\geq (1-\Pr_{\tilde{D}}[c=1])\frac{\epsilon\gamma}{4k} \\
&> \epsilon\cdot\frac{\epsilon\gamma}{4k} \qquad (\text{since } \Pr_{\tilde{D}}[c=1] = \Pr_{D}[c=1] \leq 1-\epsilon) \\
&= \frac{\epsilon^2\gamma}{2k}.
\end{aligned}$$

□

Lemma B.7. *Suppose the size of S at line 9 in Algorithm 1 is at least m. Then the target concept c is ϵ-close to the all-zero function* $\mathbf{0}$.

Proof. Suppose $|S| \geq m$. Let $S' \subseteq S$ be any subset of S of size exactly m. Without loss of generality, assume that $S' = \{1,\ldots,m\}$.

Let X and X^+ be the random variables obtained by sampling from $\{0,1\}^n$ according to distributions $\tilde{D}$ and $\tilde{D}_1$ respectively. Let random variable $Z(X) := X_1 + \cdots + X_m$ and $Z^+(X^+) := X_1^+ + \cdots + X_m^+$.

Since D is pairwise independent, then by Claim 3.5, distribution $\tilde{D}$ is pairwise independent as well. Therefore,

$$\mathbf{Var}(Z) = \mathbf{Var}(X_1) + \cdots + \mathbf{Var}(X_m) = \sum_{i=1}^{m} \mathrm{E}_{\tilde{D}}[X_i](1-\mathrm{E}_{\tilde{D}}[X_i]) \leq \frac{m}{4}.$$

On the other hand, using the bound on covariances in Lemma B.3, we have

$$\mathbf{Var}(Z^+) = \sum_{i=1}^{m}\mathbf{Var}(X_i^+) + \sum_{i\neq j}\mathbf{Cov}(X_i^+, X_j^+) < \frac{m}{4} + m^2\frac{1}{4\epsilon m} \leq \frac{m}{2\epsilon}.$$

Let $\bar{Z} = \mathrm{E}_{\tilde{D}}[Z]$ and $\bar{Z}^+ = \mathrm{E}_{\tilde{D}_1}[Z^+]$. Then by Claim B.6,

$$\Delta Z := \bar{Z}^+ - \bar{Z} > \frac{\epsilon^2\gamma m}{2k}.$$

Now, by setting $\Delta_1 = \sqrt{\frac{m}{2\epsilon}}$ and applying Chebyshev's inequality to Z, we have

$$\Pr_{\tilde{D}}[Z \geq \bar{Z} + \Delta_1] \leq \Pr[|Z-\bar{Z}| \geq \Delta_1] \leq \frac{\mathbf{Var}(Z)}{\Delta_1^2} \leq \epsilon/2.$$

Similarly, letting $\Delta_2 = \sqrt{\frac{2m}{\epsilon}}$ and applying Chebyshev's inequality to Z^+ yields

$$\Pr_{\tilde{D}_1}[Z^+ \leq \bar{Z}^+ - \Delta_2] \leq 1/4.$$

It is easily checked that $\Delta_1 + \Delta_2 < \frac{\epsilon^2\gamma m}{2k} < \Delta Z$. Therefore,

$$\begin{aligned}
\epsilon/2 &\geq \Pr_{\tilde{D}}[Z(X) \geq \bar{Z} + \Delta_1] \geq \Pr_{\tilde{D}}[Z(X) \geq \bar{Z}^+ - \Delta_2] \\
&\geq \Pr_{\tilde{D}}[Z(X) \geq \bar{Z}^+ - \Delta_2 \text{ and } X \text{ is a positive example}] \\
&= \Pr_{\tilde{D}_1}[Z^+(X^+) \geq \bar{Z}^+ - \Delta_2]\Pr_{\tilde{D}}[c(X)=1] \\
&\geq (1-\frac{1}{4})\Pr_{\tilde{D}}[c(X)=1],
\end{aligned}$$

and hence

$$\Pr_{\tilde{D}}[c(X) = 1] = \Pr_{D}[c(X) = 1] \le \frac{\epsilon/2}{1 - 1/4} = \frac{2}{3}\epsilon \le \epsilon,$$

which completes the proof. □

B.4 Putting Everything Together

Now we are ready to put everything together and prove the correctness of list-learning algorithm, i.e., Theorem B.1.

Proof of Theorem B.1. First of all, the claimed sample complexity of the learning algorithm follows directly from Lemma B.3, and the time complexity bound is due to the fact that we need to estimate, using the random examples, $\widehat{\mathbf{Cov}_{\tilde{D}_1}(X_i}, X_j)$ for every pair $1 \le i < j \le n$, and that at the end we may need to output a list of $\binom{m}{\le k}$ conjunctions.

Next, by Lemma B.3, every conjunction bit passes the Pairwise-Independence-Test and hence in S. Then, by Lemma B.2, filtering out low label-sensitive bits can cause at most an error of ϵ. That is, if we output all $\binom{m}{\le k}$ conjunctions of size at most k from bits in S, at least one of these is ϵ-close to the target concept $c(x)$.

Finally, Lemma B.7 ensures that when the size of S is large, we can simply output the $\mathbf{0}$ function which is ϵ-close to c. □

C THE TRIVIAL "BEST AGREEMENT" ALGORITHM (INFORMATION-THEORETIC-BOUND VERSION)

A naive algorithm for learning k-conjunctions with attribute noise is to try all $\sum_{i=0}^{k} 2^i \binom{n}{i} < (2n)^{k+1}$ conjunctions of size at most k and output the one that agrees with examples best.

Theorem C.1. *Given $0 < \epsilon < 1/2$ and assume the noise rate per coordinate is unknown and satisfies $\nu \le \frac{\epsilon}{2k}$, the naive algorithm that outputs the k-conjunction with maximum agreement with the observed distribution runs in time $O(n^k)$ and with probability $1 - \delta$ outputs a conjunction that is $(1-\epsilon)$-close to the conjunction labeling the noisy examples.*

Proof. Let D be the underlying distribution and let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ be the attribute noise vector with upper bound ν, i.e. $\nu_i \le \nu$ for every $1 \le i \le n$. For ease of exposition, assume that $f(x) = x_1 \wedge \cdots \wedge x_k$ is the target concept. For every $x \in \{0,1\}^n$, let $\tilde{x} = x \oplus \mu$ be the vector obtained from x by adding the attribute noise μ specified by $\boldsymbol{\nu}$. Lastly, let $\hat{X}$ denote the set of noisy examples output by the oracle $\{\tilde{x_1}, \tilde{x_2}, \ldots, \tilde{x_m}\}$. Define the *empirical disagreement* of a conjunction g on the sample by

$$\text{disagreement(g)}_{\hat{X}} = \frac{1}{m} \sum_{\tilde{x} \in \hat{X}} I_{g(\tilde{x}) \ne f(x)},$$

where $I_{g(\tilde{x}) \ne f(x)}$ is the indicator random variable of the event that $g(\tilde{x}) \ne f(x)$.

By a Hoeffding bound, it follows that

$$\Pr[|\text{disagreement(g)}_{\hat{X}} - \mathrm{E}_{x,\nu}[\text{disagreement(g)}_{\hat{X}}]| > t] \le e^{-2mt^2}.$$

Let us calculate $\mathrm{E}_{x,\nu}[\text{disagreement(g)}_{\hat{X}}]$ first when $g = f$, and then when $dist_D(f,g) > \epsilon$. We will upper bound this quantity when $f = g$ and lower bound it when f and g are ϵ-far. We will show that the minimum disagreement among all ϵ-far functions g is larger than the disagreement of f on the observed set $\hat{X}$, with high probability. Therefore we output an $\epsilon-$close conjunction with high probability $1 - \delta$.

Note that the example oracle generates an example in the following process: first draws a string x according to D, labels it as $f(x)$, then adds the attribute noise which transforms x into $\tilde{x}$. Therefore the example we see is $(\tilde{x}, f(x))$. But f will predict the label as $f(\tilde{x})$. Hence, the probability that f makes a mistake, i.e., the disagreement between f and the example oracle is

$$\mathrm{E}_{x,\boldsymbol{\nu}}[\text{disagreement(g)}_{\hat{X}}] = \Pr_{D,\boldsymbol{\nu}}[f(x) \neq f(\tilde{x})] \leq \max_{x} \Pr_{\boldsymbol{\nu}}[f(x) \neq f(\tilde{x})]. \tag{2}$$

Write $x|_{[k]}$ for the k-bit string obtained by projecting x onto index subset $[k]$. Clearly $f(x) = 1$ if and only if $x|_{[k]} = 1^k$. If $f(x) = 0$, then $\Pr_{\boldsymbol{\nu}}[f(x) \neq f(\tilde{x})] = \Pr_{\boldsymbol{\nu}}[\tilde{x}|_{[k]} = 1^k] = \prod_{i \in [k]: x_i = 1}(1 - \nu_i) \cdot \prod_{i \in [k]: x_i = 0} \nu_i \leq \prod_{i \in [k]} \nu_i \leq 1 - \prod_{i \in [k]}(1 - \nu_i)$, assuming $\nu < 1/2$.

On the other hand, when $f(x) = 1$, then

$$\Pr_{\boldsymbol{\nu}}[]f(x) \neq f(\tilde{x})] = \Pr_{\boldsymbol{\nu}}[\tilde{x}|_{[k]} \neq 1^k] = 1 - \prod_{i \in [k]}(1 - \nu_i) \leq 1 - (1 - \nu)^k \leq k\nu.$$

Therefore, $\mathrm{E}_{x,\boldsymbol{\nu}}[\text{disagreement(g)}_{\hat{X}}] \leq k\nu$.

Note that $\Pr_{\boldsymbol{\nu}}[g(x) \neq g(\tilde{x})] \leq k\nu$ holds for any conjunction g of size at most k. Now for any k-conjunction g which is at distance ϵ from f under D, i.e. $\mathrm{dist}_D(f, g) = \epsilon$, we have

$$\begin{aligned}
\mathrm{E}_{x,\boldsymbol{\nu}}[\text{disagreement(g)}_{\hat{X}}] &= \sum_{x} D(x) \Pr_{\boldsymbol{\nu}}[f(x) \neq g(\tilde{x})] \\
&= \sum_{x: f(x) = g(x)} D(x) \Pr_{\boldsymbol{\nu}}[g(x) \neq g(\tilde{x})] + \sum_{x: f(x) \neq g(x)} D(x) \Pr_{\boldsymbol{\nu}}[g(x) = g(\tilde{x})] \\
&\geq \sum_{x: f(x) \neq g(x)} D(x) \Pr_{\boldsymbol{\nu}}[g(x) = g(\tilde{x})] \geq (1 - k\nu)\mathrm{dist}_D(f, g) = (1 - k\nu)\epsilon.
\end{aligned}$$

By taking a union bound over all the $O(n^k)$ conjunctions that are ϵ-far from g, it follows that with probability $> 1 - n^k e^{-2mt^2}$ all these conjunctions g are such that

$$\text{disagreement(g)}_{\hat{X}} \geq (1 - k\nu)\epsilon - t.$$

By the above calculations it also follows that f itself satisfies

$$\text{disagreement(f)}_{\hat{X}} \leq k\nu + t.$$

It follows that if we assume that the maximum attribute noise is small enough, e.g. $\nu \leq \frac{\epsilon}{2k}$, $t = \epsilon/8$, $\epsilon < 1/2$ and $n^k e^{-2mt^2} < \delta/2$, then with probability $1 - \delta$ we output a conjunction that is ϵ-close to f, using $m = \Theta(\frac{1}{\epsilon^2}(\log \frac{1}{\delta} + k \log n))$ examples.

□

D PROOF OF CLAIM 3.5

Proof. First of all, for any $1 \leq i \leq n$, if we let $p_i := \Pr_D[X_i = 1]$ and $\tilde{p}_i := \Pr_{\tilde{D}}[X_i = 1]$, then

$$\begin{pmatrix} 1 - \tilde{p}_i \\ \tilde{p}_i \end{pmatrix} = \begin{pmatrix} 1 - \nu_i & \nu_i \\ \nu_i & 1 - \nu_i \end{pmatrix} \begin{pmatrix} 1 - p_i \\ p_i \end{pmatrix}.$$

More generally, for any subset of k indices $\{i_1, \ldots, i_k\} \subset [n]$,

$$\begin{pmatrix} \Pr_{\tilde{D}}[X_{i_1} \cdots X_{i_k} = 0^k] \\ \vdots \\ \Pr_{\tilde{D}}[X_{i_1} \cdots X_{i_k} = 1^k] \end{pmatrix} = \begin{pmatrix} 1 - \nu_{i_1} & \nu_{i_1} \\ \nu_{i_1} & 1 - \nu_{i_1} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 - \nu_{i_k} & \nu_{i_k} \\ \nu_{i_k} & 1 - \nu_{i_k} \end{pmatrix} \begin{pmatrix} \Pr_D[X_{i_1} \cdots X_{i_k} = 0^k] \\ \vdots \\ \Pr_D[X_{i_1} \cdots X_{i_k} = 1^k] \end{pmatrix},$$

where $\otimes$ stands for the Kronecker product of matrices. Now suppose that D is k-wise independent, then

$$\begin{pmatrix} \Pr_D[X_{i_1} \cdots X_{i_k} = 0^k] \\ \vdots \\ \Pr_D[X_{i_1} \cdots X_{i_k} = 1^k] \end{pmatrix} = \begin{pmatrix} 1 - p_{i_1} \\ p_{i_1} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 - p_{i_k} \\ p_{i_k} \end{pmatrix},$$

and it follows that

$$\begin{aligned} \begin{pmatrix} \Pr_{\tilde{D}}[X_{i_1} \cdots X_{i_k} = 0^k] \\ \vdots \\ \Pr_{\tilde{D}}[X_{i_1} \cdots X_{i_k} = 1^k] \end{pmatrix} &= \left(\begin{pmatrix} 1 - \nu_{i_1} & \nu_{i_1} \\ \nu_{i_1} & 1 - \nu_{i_1} \end{pmatrix} \begin{pmatrix} 1 - p_{i_1} \\ p_{i_1} \end{pmatrix} \right) \otimes \cdots \otimes \left(\begin{pmatrix} 1 - \nu_{i_k} & \nu_{i_k} \\ \nu_{i_k} & 1 - \nu_{i_k} \end{pmatrix} \begin{pmatrix} 1 - p_{i_k} \\ p_{i_k} \end{pmatrix} \right) \\ &= \begin{pmatrix} 1 - \tilde{p}_{i_1} \\ \tilde{p}_{i_1} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 - \tilde{p}_{i_k} \\ \tilde{p}_{i_k} \end{pmatrix}. \end{aligned}$$

That is, $\tilde{D}$ is also k-wise independent. The other direction follow from an identical argument by noting that matrix $\begin{pmatrix} 1 - \nu_i & \nu_i \\ \nu_i & 1 - \nu_i \end{pmatrix}$ is invertible — namely

$$\begin{pmatrix} 1 - \nu_i & \nu_i \\ \nu_i & 1 - \nu_i \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1-\nu_i}{1-2\nu_i} & -\frac{\nu_i}{1-2\nu_i} \\ -\frac{\nu_i}{1-2\nu_i} & \frac{1-\nu_i}{1-2\nu_i} \end{pmatrix},$$

for every $0 \leq \nu_i < 1/2$. □