## A Proof of Lemma 4

*Proof.* Proof of Lemma 4 Let $c \in \mathcal{C}$ be an optimal decision tree, i.e., a size $t$ decision tree that maximizes correlation for $\mathcal{U}_f$, i.e., $\mathrm{cor}(c) = \mathrm{cor}(\mathcal{C})$. The Fourier representation of $c$ is,

$$c = \sum_{S \subseteq [n]} \hat{c}(S)\chi_S(x), \tag{4}$$

where $[n]$ denotes $\{1, 2, \ldots, n\}$, $\hat{c}(S) \in [-1, 1]$, and parity classifier $\chi_S(x) = \prod_{i \in S} x[i]$ where $x[i]$ is the $i$th coordinate of $x$. Kushilevitz and Mansour show that if $c$ has at most $t$ leaves then $\sum_{S \subseteq [n]} |\hat{c}(S)| \leq t$. Now,

$$\mathrm{cor}(\mathcal{C}) = \mathrm{cor}(c) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{U}_f}[c(x)y] = \sum_{S \subseteq [n]} \hat{c}(S)\,\mathrm{cor}(\chi_S)$$

Hence, $\max_{S \subseteq [n]} |\,\mathrm{cor}(\chi_S)| \geq \mathrm{cor}(\mathcal{C})/t$ (otherwise the quantity displayed above on the left would be less than $\sum_{S \subseteq [n]} |\hat{c}(S)|\,\mathrm{cor}(\mathcal{C})/t \leq \mathrm{cor}(\mathcal{C})$, a contradiction). For any $\tau > 0$, the KM algorithm with $\mathrm{poly}(n, 1/\tau, \log(1/\delta))$ queries and runtime outputs estimates of the correlations $\mathrm{cor}(\chi_S)$ (these are exactly the estimated Fourier coefficients $\hat{f}(S)$) for each $S$ that are accurate to within an additive $\tau$, with probability $\geq 1 - \delta$ (there is a sparse polynomial-sized approximation using the fact that at most $1/\tau^2$ sets $S$ can have $|\,\mathrm{cor}(\chi_S)| \geq \tau$). Hence, if we take the set $S$ for which KM estimates $|\,\mathrm{cor}(\chi_S)|$ to be largest, it will have an correlation within $2\tau$ of that of the best $S$. Hence, setting $\tau = \epsilon_0/2$ suffices for the Lemma. Note that if $\mathrm{cor}(\chi_S) < 0$, one simply outputs the classifier $-\chi_S$. □

## B Proof of Theorem 3

We first prove Lemma 5.

*Proof of Lemma 5.* Let $c : \{-1, 1\}^n \to \{-1, 1\}$, and let $c_d$ be the degree-$d$ truncated Fourier approximation of $c$, which is the best degree-$d$ approximation to $c$ under the uniform distribution over $x \in \{-1, 1\}^n$. It is well-known that, in terms of the Fourier approximation equation 4, $c_d = \sum_{S:|S| \leq d} \hat{c}(S)\chi_S(x)$.

Klivans, O'Donnell, and Servedio [25] have shown that, for any $0 < \epsilon < \frac{1}{2}$, $d = \frac{20}{\epsilon^2}$, and any $n \geq 1$ and any halfspace $c(x) = \mathrm{sign}(w \cdot x - \theta)$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{U}}\left[(c(x) - c_d(x))^2\right] \leq \epsilon$$

In particular, let $c$ be the best halfspace approximation to $f$, i.e., one with maximum correlation, and let $c_d$ be its degree-$d$ truncation. Then,

$$\mathrm{cor}(c_d, \mathcal{U}_f) = \mathrm{cor}(c, \mathcal{U}_f) - \mathrm{cor}(c - c_d, \mathcal{U}_f) = \mathrm{cor}(\mathcal{C}, \mathcal{U}_f) - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{U}_f}[(c(x) - c_d(x))y]$$

Now, by Cauchy-Schwartz,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{U}_f}[(c(x) - c_d(x))y] \leq \sqrt{\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{U}_f}[(c(x) - c_d(x))^2] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{U}_f}[y^2]} \leq \sqrt{\epsilon \cdot 1}$$

Hence, $\mathrm{cor}(c_d) \geq \mathrm{cor}(\mathcal{C}) - \sqrt{\epsilon}$. Now,

$$\mathrm{cor}(c_d) = \sum_{S:|S| \leq d} \hat{c}(S)\,\mathrm{cor}(\chi_S) \leq \sum_{S:|S| \leq d} |\hat{c}(S)| \max_{S:|S| \leq d} |\,\mathrm{cor}(\chi_S)|$$

Finally, $\sum_{S:|S| \leq d} |\hat{c}(S)| \leq n^d$ (since each $\hat{c}(S) \in [-1, 1]$ and there are $\leq n^d$ of them), hence there must be some set $S$ of size $\leq d$ for which $|\,\mathrm{cor}(\chi_S)| \geq (\mathrm{cor}(\mathcal{C}) - \sqrt{\epsilon})/n^d$. Substituting $\epsilon = \epsilon_0^2$ proves the lemma. □

Theorem 3 now follows easily from the above lemma and our boosting theorem.

*Proof of Theorem 3.* Consider the weak learner simply finds the degree-$d$ term, $\chi_S(x)$ with $|S| \leq d$, with greatest empirical correlation $\frac{1}{m}\sum_{i=1}^{m} \chi_S(x_i)y_i$ on a data set $(x_1, y_1), \ldots, (x_m, y_m)$. Standard Chernoff-Hoeffding bounds guarantee that, for $m \geq poly(\log(1/\delta), n^d)$, with probability $\geq 1 - \delta$, the empirical correlation of each of the $\leq n^d$ different $\chi_S$'s will be within $\epsilon_0/4$ of their true correlation. $\square$