1 PROOF OF THEOREM 4

We will prove the theorem for the case when $\mathcal{H}$ contains probabilistic hypotheses. The proof can easily be transferred to the case where $\mathcal{H}$ is the labeling set by following the construction in (Cuong et al., 2013, sup.).

Let $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ with $n$ probabilistic hypotheses, and assume a uniform prior on them. A labeling is generated by first randomly drawing a hypothesis from the prior and then drawing a labeling from this hypothesis. This induces a distribution on all labelings.

We construct $k$ independent distractor instances $x_1, x_2, \ldots, x_k$ with identical output distributions for the $n$ probabilistic hypotheses. Our aim is to trick the greedy algorithm $\pi$ to select these $k$ instances. Since the hypotheses are identical on these instances, the greedy algorithm learns nothing when receiving each label.

Let $H(Y_1)$ be the Shannon entropy of the prior label distribution of any $x_i$ (this entropy is the same for all $x_i$). Since the greedy algorithm always selects the $k$ instances $x_1, x_2, \ldots, x_k$ and their labels are independent, we have

$$H_{\text{ent}}(\pi) = kH(Y_1).$$

Next, we construct an instance $x_0$ where its label will deterministically identify the probabilistic hypotheses. Specifically, $\mathbb{P}[h_i(x_0) = i | h_i] = 1$ for all $i$. Note that $H(Y_0) = \ln n$.

To make sure that the greedy algorithm $\pi$ selects the distractor instances instead of $x_0$, a constraint is that $H(Y_1) > H(Y_0) = \ln n$. This constraint can be satisfied by, for example, allowing $Y$ to have $n + 1$ labels and letting $\mathbb{P}[h(x_j)|h]$ be the uniform distribution for all $j \geq 1$ and $h \in \mathcal{H}$. In this case, $H(Y_1) = \ln(n + 1) > \ln n$.

We compare the greedy algorithm $\pi$ with an algorithm $\pi_A$ that selects $x_0$ first, and hence knows the true hypothesis after observing its label.

Finally, we construct $n(k-1)$ more instances, and the algorithm $\pi_A$ will select the appropriate $k-1$ instances from them after figuring out the true hypothesis. Let the instances be $\{x_{(i,j)} : 1 \leq i \leq n$ and $1 \leq j \leq k - 1\}$. Let $Y_{(i,j)}^h$ be the (random) label of $x_{(i,j)}$ according to the hypothesis $h$. For all $h \in \mathcal{H}$, $Y_{(i,j)}^h$ has identical distributions for $1 \leq j \leq k - 1$. Thus, we only need to specify $Y_{(i,1)}^h$.

We specify $Y_{(i,1)}^h$ as follows. If $h \neq h_i$, then let $\mathbb{P}[Y_{(i,1)}^h = 0] = 1$. Otherwise, let $\mathbb{P}[Y_{(i,1)}^h = 0] = 0$, and the distribution on other labels has entropy $H(Y_{(1,1)}^h)$, as all hypotheses are defined the same way.

When the true hypothesis is unknown, the distribution for $Y_{(1,1)}$ has entropy

$$H(Y_{(1,1)}) = (1 - \frac{1}{n}) + \frac{1}{n}H(Y_{(1,1)}^{h_1}),$$

where $H(1 - \frac{1}{n})$ is the entropy of the Bernoulli distribution $(1 - \frac{1}{n}, \frac{1}{n})$.

As we want the greedy algorithm to select the distractors, we also need $H(Y_1) > H(Y_{(1,1)})$, giving

$$H(Y_{(1,1)}^{h_1}) < n(H(Y_1) - H(1 - \frac{1}{n})).$$

Algorithm $\pi_A$ first selects $x_0$, identifies the true hypothesis exactly, and then selects $k-1$ instances with entropy $H(Y_{(1,1)}^{h_1})$. Thus,

$$H_{\text{ent}}(\pi_A) = \ln n + (k-1)H(Y_{(1,1)}^{h_1}).$$

Hence, we have

$$\frac{H_{\text{ent}}(\pi)}{H_{\text{ent}}(\pi_A)} = \frac{kH(Y_1)}{\ln n + (k-1)H(Y_{(1,1)}^{h_1})}.$$

Set $H(Y_{(1,1)}^{h_1})$ to $n(H(Y_1) - H(1 - \frac{1}{n})) - c$ for some small constant $c$. The above ratio becomes

$$\frac{kH(Y_1)}{\ln n + (k-1)(n(H(Y_1) - H(1 - \frac{1}{n})) - (k-1)c)}.$$

Since $H(1 - \frac{1}{n})$ approaches 0 as $n$ grows and $H(Y_1) = \ln(n+1)$, we can make the ratio $H_{\text{ent}}(\pi)/H_{\text{ent}}(\pi_A)$ as small as we like by increasing $n$. Furthermore, $H_{\text{ent}}(\pi)/H_{\text{ent}}(\pi_A) \geq H_{\text{ent}}(\pi)/H_{\text{ent}}(\pi^*)$. Thus, Theorem 4 holds.
2 PROOF OF THEOREM 5

It is clear that the version space reduction function \( f \) satisfies the minimal dependency property, is pointwise monotone and \( f(\emptyset, h) = 0 \) for all \( h \). Let \( x_D \equiv \text{dom}(D) \) and \( y_D \equiv D(x_D) \). From Equation (3), we have

\[
\arg \max_x \min_y \{f(\text{dom}(D) \cup \{x\}, D \cup \{(x, y)\}) - f(\text{dom}(D), D)\}
\]

\[
= \arg \max_x \min_y f(\text{dom}(D) \cup \{x\}, D \cup \{(x, y)\})
\]

\[
= \arg \max_x \min_y f(\text{dom}(D) \cup \{x\}, D \cup \{(x, y)\})
\]

\[
= \arg \max_x \min y \{f(y_D \cup \{y\}; x_D \cup \{x\})\}
\]

\[
= \arg \max_x \min y \{f(y_D \cup \{y\}; x_D \cup \{x\})\}
\]

Thus, Equation (6) is equivalent to Equation (3). To apply Theorem 3, what remains is to show that \( f \) is pointwise submodular.

Consider \( f_h(S) \equiv f(S, h) \) for any \( h \). Fix \( A \subseteq B \subseteq \mathcal{X} \) and \( x \in \mathcal{X} \setminus B \). We have

\[
f_h(A \cup \{x\}) - f_h(A)
\]

\[
= p_0[h(A)] - p_0[h(A \cup \{x\})]
\]

\[
= \sum_{h'(A) = h(A)} p_0[h'] - \sum_{h'(A) \neq h(A)} p_0[h']
\]

\[
= \sum_{h'} p_0[h'] \mathbb{1}(h'(A) = h(A)) \mathbb{1}(h'(x) \neq h(x)).
\]

Similarly, we have

\[
f_h(B \cup \{x\}) - f_h(B)
\]

\[
= \sum_{h'} p_0[h'] \mathbb{1}(h'(B) = h(B)) \mathbb{1}(h'(x) \neq h(x)).
\]

Since \( A \subseteq B \), all pairs \( h, h' \) such that \( h'(B) = h(B) \) also satisfy \( h'(A) = h(A) \).

Thus, \( f_h(A \cup \{x\}) - f_h(A) \geq f_h(B \cup \{x\}) - f_h(B) \) and \( f_h \) is submodular. Therefore, \( f \) is pointwise submodular.

3 PROOF OF THEOREM 7

Consider any prior \( p_0 \) such that \( p_0[h] > 0 \) for all \( h \). Fix any \( D \) and \( D' \) where \( D' = D \cup \mathcal{E} \) with \( \mathcal{E} \neq \emptyset \), and fix any \( x \in \mathcal{X} \setminus \text{dom}(D') \). For a partial labeling \( D \), let \( x_D \equiv \text{dom}(D) \) and \( y_D \equiv D(x_D) \). We have

\[
\Delta(x|D)
\]

\[
= \sum_{h_D} \sum_{h' \neq D} p_0[h'] p_0[h'] \mathbb{1}(h, h')
\]

\[
= \sum_{h_D} \sum_{h' \neq D} p_0[h'] p_0[h'] \mathbb{1}(h, h')
\]

\[
= \sum_{h_D} \sum_{h' \neq D} p_0[h'] p_0[h'] \mathbb{1}(h, h')
\]

\[
= \sum_{h_D} \sum_{h' \neq D} p_0[h'] p_0[h'] \mathbb{1}(h, h')
\]

where \( h \sim \mathcal{E} \) denotes that \( h \) is not consistent with \( \mathcal{E} \). Now we can construct the loss function \( L \) such that \( L(h, h') = 0 \) for all \( h, h' \) satisfying \( h \sim \mathcal{E} \) or \( h' \sim \mathcal{E} \), thus

\[
\Delta(x|D') = \sum_{h_D} \sum_{h' \neq D} p_0[h'] p_0[h'] \mathbb{1}(h, h')
\]

From the assumption \( p_0[h] > 0 \) for all \( h \), we have \( \sum_{h_D} p_0[h] < \sum_{h_D} p_0[h] \). Thus, \( \Delta(x|D') > \Delta(x|D) \) and \( f_L \) is not adaptive submodular.
4 SUFFICIENT CONDITION FOR ADAPTIVE SUBMODULARITY OF $f_L$

From the previous section, let

$$A \equiv \sum_{h \sim D} \sum_{h'(x) \neq h(x)} p_0[h] p_0[h'] L(h, h')$$

$$B \equiv \sum_{h \sim D} \sum_{h'(x) \neq h(x)} p_0[h] p_0[h'] L(h, h') 1(h \sim \mathcal{E} \text{ or } h' \sim \mathcal{E})$$

$$C \equiv \sum_{h \sim D} p_0[h] \quad \text{and} \quad D \equiv \sum_{h \sim D} p_0[h] 1(h \sim \mathcal{E}).$$

In this section, we allow $\mathcal{E}$ to be empty. Note that $\Delta(x|D) = \frac{A}{B}$ and $\Delta(x|D') = \frac{A-B}{B-D}$. A sufficient condition for $f_L$ to be adaptive submodular with respect to $p_0$ is that for all $D, D', x$, and $h$, we have $\frac{A}{B} \geq \frac{A-B}{B-D}$. This condition is equivalent to $\frac{A}{B} \leq \frac{B}{D}$. That means

$$\sum_{h \sim D} \sum_{h'(x) \neq h(x)} p_0[h] p_0[h'] L(h, h')$$

$$\sum_{h \sim D} p_0[h]$$

for all $D, D'$, and $x$. This condition holds if $L$ is the 0-1 loss. However, it remains open whether this condition is true for any interesting loss function other than 0-1 loss.

5 PROOF OF THEOREM 8

It is clear that $t_L$ satisfies the minimal dependency property and Equation (8) is equivalent to Equation (3). It is also clear that $t_L$ is pointwise monotone and $t_L(\emptyset, h) = 0$ for all $h$. Thus, to apply Theorem 3, what remains is to show that $t_L$ is pointwise submodular.

Consider $t_{L,h}(S) \equiv t_L(S, h)$ for any $h$. Fix $A \subseteq B \subseteq \mathcal{X}$ and $x \in \mathcal{X} \setminus B$. We have

$$t_{L,h}(A \cup \{x\}) - t_{L,h}(A) = \sum_{h'(A) = h(A)} \sum_{h''(A) = h(A)} p_0[h'] L(h', h'') p_0[h'']$$

$$- \sum_{h'(A) = h(A)} \sum_{h''(A) = h(A)} \sum_{h'(x) = h(x)} \sum_{h''(x) = h(x)} p_0[h'] L(h', h'') p_0[h'']$$

$$+ \sum_{h'} \sum_{h''} [p_0[h'] L(h', h'') p_0[h''] \cdot$$

$$1(h'(A) = h(A) \text{ and } h''(A) = h(A)) \cdot$$

$$1(h'(x) \neq h(x) \text{ or } h''(x) \neq h(x))].$$

Similarly, we have

$$t_{L,h}(B \cup \{x\}) - t_{L,h}(B) = \sum_{h' \neq h''} [p_0[h'] L(h', h'') p_0[h''] \cdot$$

$$1(h'(B) = h(B) \text{ and } h''(B) = h(B)) \cdot$$

$$1(h'(x) \neq h(x) \text{ or } h''(x) \neq h(x))].$$

Since $A \subseteq B$, all pairs $h, h'$ such that $1(h'(B) = h(B) \text{ and } h''(B) = h(B)) = 1$ also satisfy $1(h'(A) = h(A) \text{ and } h''(A) = h(A)) = 1$.

Thus, $t_{L,h}(A \cup \{x\}) - t_{L,h}(A) \geq t_{L,h}(B \cup \{x\}) - t_{L,h}(B)$ and $t_{L,h}$ is submodular. Therefore, $t_L$ is pointwise submodular.

6 POINTWISE SUBMODULARITY OF $f_L$

Consider $f_{L,h}(S) \equiv f_L(S, h)$ for any $h$. Fix $A \subseteq B \subseteq \mathcal{X}$ and $x \in \mathcal{X} \setminus B$. We have

$$f_{L,h}(A \cup \{x\}) - f_{L,h}(A) = \sum_{h'(A) = h(A)} \sum_{h''(A) = h(A)} p_0[h'] L(h, h') - \sum_{h'(A) = h(A)} \sum_{h''(A) = h(A)} p_0[h'] L(h, h')$$

$$= \sum_{h'} p_0[h'] L(h, h') 1(h'(A) = h(A)) 1(h'(x) \neq h(x)).$$

Similarly, we have

$$f_{L,h}(B \cup \{x\}) - f_{L,h}(B) = \sum_{h'} p_0[h'] L(h, h') 1(h'(B) = h(B)) 1(h'(x) \neq h(x)).$$

Since $A \subseteq B$, all pairs $h, h'$ such that $h'(B) = h(B)$ also satisfy $h'(A) = h(A)$.

Thus, $f_{L,h}(A \cup \{x\}) - f_{L,h}(A) \geq f_{L,h}(B \cup \{x\}) - f_{L,h}(B)$ and $f_{L,h}$ is submodular. Therefore, $f_L$ is pointwise submodular.

7 PROOF OF PROPOSITION 1

Let $x \in \text{dom}(D)$ and $y \in \mathcal{D}(x \mathcal{D})$. Using Equation (7) and the definition of $f_L$, we have
\[ x^* \]

\[ x^* = \arg \max_x E_{h \sim p_D}[L(x_D \cup \{x\}, h) - f_L(x_D, h)] \]

\[ x^* = \arg \max_x E_{h \sim p_D}[L(x_D \cup \{x\}, h)] \]

\[ x^* = \arg \max_x E_{h \sim p_D}(\sum_{h'} p_0[h']L(h, h')) \]

\[ x^* = \arg \min_x E_{h \sim p_D}(\sum_{h'} p_0[h']L(h, h')) \]

\[ x^* = \arg \min_x E_{h \sim p_D}(\sum_{h'} p_0[h']L(h, h')) \]

Note that if \( p_D[h'] > 0 \), then

\[ p_0[h'] = p_D[h']p_0[y_D; x_D] . \]

Hence, the last expression above is equal to

\[ \arg \min_x E_{h \sim p_D}(\sum_{h'} p_0[h']L(h, h')) \]

\[ \arg \min_x E_{h \sim p_D}(\sum_{h'} p_0[h']L(h, h')) \]

\[ \arg \min_x \sum_h p_D[h]\sum_{h'} p_D[h']L(h, h') \]

\[ \arg \min_x \sum_h \sum_{y} p_D[h]\sum_{h'} p_D[h']L(h, h') \]

\[ \arg \min_x \sum_y \sum_{h} p_D[h]\sum_{h'} p_D[h']L(h, h') \]

\[ \arg \min_x \sum_y \sum_{h'} p_D[h']L(h, h') \]

\[ \arg \min_x \sum_y E_{h,h' \sim p_D}[L(h, h')] \]

So, Proposition 1 holds.

8 PROOF OF PROPOSITION 2

Let \( x_D \subseteq \text{dom}(D) \) and \( y_D \subseteq D(x_D) \). Using Equation (8) and the definition of \( t_L \), we have

\[ \arg \max \sum_{h'} p_0[h']L(h, h') \]

\[ \arg \max \sum_{h'} p_0[h']L(h', h'')p_0[h''] \]

Using the same observation about \( p_0[h'] \) and \( p_0[h''] \) as in the previous section, we note that the last expression above is equal to

\[ \arg \max \sum_{h'} p_0[h']L(h, h') \]

\[ \arg \max \sum_{h'} p_0[h']L(h', h'')p_0[h''] \]

Thus, Proposition 2 holds.

References