Supplementary Material: Near-optimal Adaptive Pool-based Active Learning with General Loss

Nguyen Viet Cuong
Department of Computer Science
National University of Singapore
nvcuong@comp.nus.edu.sg

Wee Sun Lee
Department of Computer Science
National University of Singapore
leews@comp.nus.edu.sg

Nan Ye
Department of Computer Science
National University of Singapore
yenan@comp.nus.edu.sg

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

We specify $Y^h_{(i,1)}$ as follows. If $h \neq h_i$, then let $\mathbb{P}[Y^h_{(i,1)} = 0] = 1$. Otherwise, let $\mathbb{P}[Y^h_{(i,1)} = 0] = 0$, and the distribution on other labels has entropy $H(Y^{h_1}_{(1,1)})$, as all hypotheses are defined the same way.

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

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

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

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

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

Hence, we have

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

Set $H(Y^{h_1}_{(1,1)})$ 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 $x_{\mathcal{D}} \stackrel{\text{def}}{=} \text{dom}(\mathcal{D})$ and $y_{\mathcal{D}} \stackrel{\text{def}}{=} \mathcal{D}(x_{\mathcal{D}})$. From Equation (3), we have

$$
\begin{aligned}
& \arg\max_x \min_y \{f(\text{dom}(\mathcal{D}) \cup \{x\}, \mathcal{D} \cup \{(x,y)\}) \\
& \qquad - f(\text{dom}(\mathcal{D}), \mathcal{D})\} \\
= \;& \arg\max_x \min_y f(\text{dom}(\mathcal{D}) \cup \{x\}, \mathcal{D} \cup \{(x,y)\}) \\
= \;& \arg\max_x \min_y [1 - p_0[y_{\mathcal{D}} \cup \{y\}; x_{\mathcal{D}} \cup \{x\}]] \\
= \;& \arg\min_x \max_y p_0[y_{\mathcal{D}} \cup \{y\}; x_{\mathcal{D}} \cup \{x\}] \\
= \;& \arg\min_x \max_y \frac{p_0[y_{\mathcal{D}} \cup \{y\}; x_{\mathcal{D}} \cup \{x\}]}{p_0[y_{\mathcal{D}}; x_{\mathcal{D}}]} \\
= \;& \arg\min_x \max_y p_{\mathcal{D}}[y; x].
\end{aligned}
$$

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

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

$$
\begin{aligned}
& f_h(A \cup \{x\}) - f_h(A) \\
= \;& p_0[h(A); A] - p_0[h(A \cup \{x\}); A \cup \{x\}] \\
= \;& \sum_{h'(A)=h(A)} p_0[h'] - \sum_{\substack{h'(A)=h(A) \\ h'(x)=h(x)}} p_0[h'] \\
= \;& \sum_{h'} p_0[h'] \mathbf{1}(h'(A) = h(A)) \mathbf{1}(h'(x) \neq h(x)).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
& f_h(B \cup \{x\}) - f_h(B) \\
= \;& \sum_{h'} p_0[h'] \mathbf{1}(h'(B) = h(B)) \mathbf{1}(h'(x) \neq h(x)).
\end{aligned}
$$

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 $\mathcal{D}$ and $\mathcal{D}'$ where $\mathcal{D}' = \mathcal{D} \cup \mathcal{E}$ with $\mathcal{E} \neq \emptyset$, and fix any $x \in \mathcal{X} \setminus \text{dom}(\mathcal{D}')$. For a partial labeling $\mathcal{D}$, let $x_{\mathcal{D}} \stackrel{\text{def}}{=} \text{dom}(\mathcal{D})$ and $y_{\mathcal{D}} \stackrel{\text{def}}{=} \mathcal{D}(x_{\mathcal{D}})$. We have

$$
\begin{aligned}
& \Delta(x|\mathcal{D}) \\
= \;& \mathbb{E}_{h \sim p_{\mathcal{D}}}[f_L(\text{dom}(\mathcal{D}) \cup \{x\}, h) - f_L(\text{dom}(\mathcal{D}), h)] \\
= \;& \mathbb{E}_{h \sim p_{\mathcal{D}}}[\sum_{h'(x_{\mathcal{D}})=h(x_{\mathcal{D}})} p_0[h'] L(h, h') \\
& \qquad - \sum_{\substack{h'(x_{\mathcal{D}})=h(x_{\mathcal{D}}) \\ h'(x)=h(x)}} p_0[h'] L(h, h')] \\
= \;& \mathbb{E}_{h \sim p_{\mathcal{D}}} \sum_{\substack{h'(x_{\mathcal{D}})=h(x_{\mathcal{D}}) \\ h'(x) \neq h(x)}} p_0[h'] L(h, h') \\
= \;& \sum_{p_{\mathcal{D}}[h]>0} p_{\mathcal{D}}[h] \sum_{\substack{h'(x_{\mathcal{D}})=h(x_{\mathcal{D}}) \\ h'(x) \neq h(x)}} p_0[h'] L(h, h').
\end{aligned}
$$

Note that if $p_{\mathcal{D}}[h] > 0$, then

$$
p_{\mathcal{D}}[h] = \frac{p_0[h]}{p_0[y_{\mathcal{D}}; x_{\mathcal{D}}]} = \frac{p_0[h]}{\sum_{h(x_{\mathcal{D}})=y_{\mathcal{D}}} p_0[h]}.
$$

Thus, $\Delta(x|\mathcal{D}) =$

$$
\begin{aligned}
& \frac{\sum_{p_{\mathcal{D}}[h]>0} \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h')}{\sum_{h(x_{\mathcal{D}})=y_{\mathcal{D}}} p_0[h]} \\
= \;& \frac{\sum_{h \sim \mathcal{D}} \sum_{\substack{h' \sim \mathcal{D} \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h')}{\sum_{h \sim \mathcal{D}} p_0[h]}.
\end{aligned}
$$

Similarly, for $\mathcal{D}'$, we also have

$$
\begin{aligned}
& \Delta(x|\mathcal{D}') \\
= \;& \frac{\sum_{h \sim \mathcal{D}'} \sum_{\substack{h' \sim \mathcal{D}' \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h')}{\sum_{h \sim \mathcal{D}'} p_0[h]} \\
= \;& \frac{1}{\sum_{h \sim \mathcal{D}'} p_0[h]} [\sum_{h \sim \mathcal{D}} \sum_{\substack{h' \sim \mathcal{D} \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h') \\
& - \sum_{h \sim \mathcal{D}} \sum_{\substack{h' \sim \mathcal{D} \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h') \mathbf{1}(h \nsim \mathcal{E} \text{ or } h' \nsim \mathcal{E})]
\end{aligned}
$$

where $h \nsim \mathcal{E}$ 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 $h \nsim \mathcal{E}$ or $h' \nsim \mathcal{E}$. Thus,

$$
\Delta(x|\mathcal{D}') = \frac{\sum_{h \sim \mathcal{D}} \sum_{\substack{h' \sim \mathcal{D} \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h')}{\sum_{h \sim \mathcal{D}'} p_0[h]}.
$$

From the assumption $p_0[h] > 0$ for all h, we have $\sum_{h \sim \mathcal{D}'} p_0[h] < \sum_{h \sim \mathcal{D}} p_0[h]$. Thus, $\Delta(x|\mathcal{D}') > \Delta(x|\mathcal{D})$ and f_L is not adaptive submodular.

4 SUFFICIENT CONDITION FOR ADAPTIVE SUBMODULARITY OF f_L

From the previous section, let

$$A \stackrel{\text{def}}{=} \sum_{h \sim \mathcal{D}} \sum_{\substack{h' \sim \mathcal{D} \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h')$$

$$B \stackrel{\text{def}}{=} \sum_{h \sim \mathcal{D}} \sum_{\substack{h' \sim \mathcal{D} \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h') \mathbf{1}(h \nsim \mathcal{E} \text{ or } h' \nsim \mathcal{E})$$

$$C \stackrel{\text{def}}{=} \sum_{h \sim \mathcal{D}} p_0[h] \quad \text{and} \quad D \stackrel{\text{def}}{=} \sum_{h \sim \mathcal{D}} p_0[h] \mathbf{1}(h \nsim \mathcal{E}).$$

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

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

$$\leq \frac{\sum_{h \sim \mathcal{D}} \sum_{\substack{h' \sim \mathcal{D} \\ h'(x) \neq h(x)}} p_0[h] p_0[h'] L(h, h') \mathbf{1}(h \nsim \mathcal{E} \text{ or } h' \nsim \mathcal{E})}{\sum_{h \sim \mathcal{D}} p_0[h] \mathbf{1}(h \nsim \mathcal{E})}$$

for all $\mathcal{D}$, $\mathcal{D}'$, 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 $t_{L,h}(S) \stackrel{\text{def}}{=} t_L(S, h)$ for any h. Fix $A \subseteq B \subseteq \mathcal{X}$ and $x \in \mathcal{X} \setminus B$. We have

$$\begin{aligned}
& t_{L,h}(A \cup \{x\}) - t_{L,h}(A) \\
= & \sum_{h'(A)=h(A)} \sum_{h''(A)=h(A)} p_0[h'] \, L(h', h'') \, p_0[h''] \\
& - \sum_{\substack{h'(A)=h(A) \\ h'(x)=h(x)}} \sum_{\substack{h''(A)=h(A) \\ h''(x)=h(x)}} p_0[h'] \, L(h', h'') \, p_0[h''] \\
= & \sum_{h'} \sum_{h''} [p_0[h'] \, L(h', h'') \, p_0[h''] \cdot \\
& \mathbf{1}(h'(A) = h(A) \text{ and } h''(A) = h(A)) \cdot \\
& \mathbf{1}(h'(x) \neq h(x) \text{ or } h''(x) \neq h(x))].
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
& t_{L,h}(B \cup \{x\}) - t_{L,h}(B) \\
= & \sum_{h'} \sum_{h''} [p_0[h'] \, L(h', h'') \, p_0[h''] \cdot \\
& \mathbf{1}(h'(B) = h(B) \text{ and } h''(B) = h(B)) \cdot \\
& \mathbf{1}(h'(x) \neq h(x) \text{ or } h''(x) \neq h(x))].
\end{aligned}$$

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

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

$$\begin{aligned}
& f_{L,h}(A \cup \{x\}) - f_{L,h}(A) \\
= & \sum_{h'(A)=h(A)} p_0[h'] L(h, h') - \sum_{\substack{h'(A)=h(A) \\ h'(x)=h(x)}} p_0[h'] L(h, h') \\
= & \sum_{h'} p_0[h'] L(h, h') \mathbf{1}(h'(A) = h(A)) \mathbf{1}(h'(x) \neq h(x)).
\end{aligned}$$

Similarly, we have

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

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 $x_{\mathcal{D}} \stackrel{\text{def}}{=} \text{dom}(\mathcal{D})$ and $y_{\mathcal{D}} \stackrel{\text{def}}{=} \mathcal{D}(x_{\mathcal{D}})$. Using Equation (7) and the definition of f_L, we have

$$
\begin{aligned}
& x^* \\
= \ & \arg\max_x \mathbb{E}_{h\sim p_{\mathcal{D}}}[f_L(x_{\mathcal{D}} \cup \{x\}, h) - f_L(x_{\mathcal{D}}, h)] \\
= \ & \arg\max_x \mathbb{E}_{h\sim p_{\mathcal{D}}}[f_L(x_{\mathcal{D}} \cup \{x\}, h)] \\
= \ & \arg\max_x \mathbb{E}_{h\sim p_{\mathcal{D}}}(\sum_{h'} p_0[h']L(h,h') \\
& \qquad - \sum_{\substack{h(x_{\mathcal{D}})=h'(x_{\mathcal{D}}) \\ h(x)=h'(x)}} p_0[h']L(h,h')) \\
= \ & \arg\min_x \mathbb{E}_{h\sim p_{\mathcal{D}}} \sum_{\substack{h(x_{\mathcal{D}})=h'(x_{\mathcal{D}}) \\ h(x)=h'(x)}} p_0[h']L(h,h') \\
= \ & \arg\min_x \mathbb{E}_{h\sim p_{\mathcal{D}}} \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h(x)=h'(x)}} p_0[h']L(h,h').
\end{aligned}
$$

Note that if $p_{\mathcal{D}}[h'] > 0$, then

$$
p_0[h'] = p_{\mathcal{D}}[h']p_0[y_{\mathcal{D}}; x_{\mathcal{D}}].
$$

Hence, the last expression above is equal to

$$
\begin{aligned}
& \arg\min_x \mathbb{E}_{h\sim p_{\mathcal{D}}} \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h(x)=h'(x)}} p_{\mathcal{D}}[h']p_0[y_{\mathcal{D}}; x_{\mathcal{D}}]L(h,h') \\
= \ & \arg\min_x \mathbb{E}_{h\sim p_{\mathcal{D}}} \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h(x)=h'(x)}} p_{\mathcal{D}}[h']L(h,h') \\
= \ & \arg\min_x \sum_h p_{\mathcal{D}}[h] \sum_{h(x)=h'(x)} p_{\mathcal{D}}[h']L(h,h') \\
= \ & \arg\min_x \sum_y \sum_{h(x)=y} p_{\mathcal{D}}[h] \sum_{h'(x)=y} p_{\mathcal{D}}[h']L(h,h') \\
= \ & \arg\min_x \sum_y \sum_h p_{\mathcal{D}}[h] \sum_{h'} p_{\mathcal{D}}[h'](L(h,h') \cdot \\
& \qquad \mathbf{1}(h(x) = h'(x) = y)) \\
= \ & \arg\min_x \sum_y \mathbb{E}_{h,h'\sim p_{\mathcal{D}}}[L(h,h') \cdot \\
& \qquad \mathbf{1}(h(x) = h'(x) = y)].
\end{aligned}
$$

Thus, Proposition 1 holds.

8 PROOF OF PROPOSITION 2

Let $x_{\mathcal{D}} \stackrel{\text{def}}{=} \text{dom}(\mathcal{D})$ and $y_{\mathcal{D}} \stackrel{\text{def}}{=} \mathcal{D}(x_{\mathcal{D}})$. Using Equation (8) and the definition of t_L, we have

$$
\begin{aligned}
& x^* \\
= \ & \arg\max_x \min_y [t_L(x_{\mathcal{D}} \cup \{x\}, \mathcal{D} \cup \{(x,y)\}) - t_L(x_{\mathcal{D}}, \mathcal{D})] \\
= \ & \arg\max_x \min_y [t_L(x_{\mathcal{D}} \cup \{x\}, \mathcal{D} \cup \{(x,y)\})] \\
= \ & \arg\max_x \min_y [\sum_{h'} \sum_{h''} p_0[h']L(h',h'')p_0[h''] \\
& - \sum_{\substack{h'(x_{\mathcal{D}})=y_{\mathcal{D}} \\ h'(x)=y}} \sum_{\substack{h''(x_{\mathcal{D}})=y_{\mathcal{D}} \\ h''(x)=y}} p_0[h']L(h',h'')p_0[h'']] \\
= \ & \arg\min_x \max_y \sum_{\substack{h'(x_{\mathcal{D}})=y_{\mathcal{D}} \\ h'(x)=y}} \sum_{\substack{h''(x_{\mathcal{D}})=y_{\mathcal{D}} \\ h''(x)=y}} p_0[h']L(h',h'')p_0[h''] \\
= \ & \arg\min_x \max_y \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h'(x)=y}} \sum_{\substack{p_{\mathcal{D}}[h'']>0 \\ h''(x)=y}} p_0[h']L(h',h'')p_0[h''] \\
= \ & \arg\min_x \max_y \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h'(x)=y}} p_0[h'] \sum_{\substack{p_{\mathcal{D}}[h'']>0 \\ h''(x)=y}} L(h',h'')p_0[h''].
\end{aligned}
$$

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

$$
\begin{aligned}
& \arg\min_x \max_y \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h'(x)=y}} (p_{\mathcal{D}}[h']p_0[y_{\mathcal{D}}; x_{\mathcal{D}}] \cdot \\
& \qquad \sum_{\substack{p_{\mathcal{D}}[h'']>0 \\ h''(x)=y}} L(h',h'')p_{\mathcal{D}}[h'']p_0[y_{\mathcal{D}}; x_{\mathcal{D}}]) \\
= \ & \arg\min_x \max_y \sum_{\substack{p_{\mathcal{D}}[h']>0 \\ h'(x)=y}} p_{\mathcal{D}}[h'] \sum_{\substack{p_{\mathcal{D}}[h'']>0 \\ h''(x)=y}} L(h',h'')p_{\mathcal{D}}[h''] \\
= \ & \arg\min_x \max_y \sum_{h'(x)=y} p_{\mathcal{D}}[h'] \sum_{h''(x)=y} L(h',h'')p_{\mathcal{D}}[h''] \\
= \ & \arg\min_x \max_y \sum_{h'} p_{\mathcal{D}}[h'] \sum_{h''} p_{\mathcal{D}}[h''](L(h',h'') \cdot \\
& \qquad \mathbf{1}(h''(x) = h'(x) = y)) \\
= \ & \arg\min_x \max_y \mathbb{E}_{h',h''\sim p_{\mathcal{D}}}[L(h',h'') \cdot \\
& \qquad \mathbf{1}(h''(x) = h'(x) = y)].
\end{aligned}
$$

Thus, Proposition 2 holds.

References

Nguyen Viet Cuong, Wee Sun Lee, Nan Ye, Kian Ming A. Chai, and Hai Leong Chieu. Active learning for probabilistic hypotheses using the maximum Gibbs error criterion. In *Advances in Neural Information Processing Systems*, pages 1457–1465, 2013.