1 Rashomon Membership across Iterations

Consider binary[1] samples (x_i, y_i), $x_i \in \{0,1\}^d$, $y_i \in \{0,1\}$. Let $D_{tr}^{(t)} = \{(x_i, y_i)\}_{i \in I_t}$ be the training dataset at iteration t, where I_t is the set of indices of these training samples. Let $D_{cdd}^{(t)} = \{(x_k, y_k)\}_{k \in K_t}$ be the candidate dataset at iteration t, where K_t is the set of indices of these training samples. At each iteration, we move one sample from the candidate dataset to the training dataset, meaning $\forall t \in \{0, 1, \ldots, |K_0| - 1\}, |I_{t+1}| - |I_t| = |K_t| - |K_{t+1}| = 1$.
For any model $f : \{0,1\}^d \mapsto \{0,1\}$, Define

$$L_t(f) = \frac{1}{|I_t|} \sum_{i \in I_t} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$\text{Obj}_t(f) = L_t(f) + H(f)$$

where H is some non-negative function of the model independent of the data (such as a penalty on the number of leaves in a tree-based model) that evaluates to 0 on the simplest possible model f (such as predicting a single class for all samples). Then, for a given hypothesis space $\mathcal{F}$:

$$\hat{\mathcal{R}}_\epsilon^{(t)} = \{f \in \mathcal{F} : \text{Obj}_t(f) \leq \min_{f \in \mathcal{F}} \text{Obj}_t(f) + \epsilon\}$$

Theorem 1.1. *For $0 \leq t_1 \leq t_2$, $0 \leq \epsilon \leq \epsilon' \leq \frac{1}{2}$,*

$$\left(\epsilon' - \epsilon \geq 2\frac{t_2 - t_1}{|I_{t_2}|}\right) \implies \left(\hat{\mathbb{R}}_\epsilon^{(t_2)} \subseteq \hat{\mathbb{R}}_{\epsilon'}^{(t_1)}\right).$$

Proof. It is sufficient to show that $\hat{\mathbb{R}}_\epsilon^{(t_2)} \subseteq \hat{\mathbb{R}}_{\epsilon'}^{(t_1)}$ given $\epsilon' - \epsilon \geq 2\frac{t_2 - t_1}{|I_{t_2}|}$.

To show $\hat{\mathbb{R}}_\epsilon^{(t_2)} \subseteq \hat{\mathbb{R}}_{\epsilon'}^{(t_1)}$, it is sufficient to show that $f \in \hat{\mathbb{R}}_\epsilon^{(t_2)} \implies f \in \hat{\mathbb{R}}_{\epsilon'}^{(t_1)}$, so we will show this. Define

$$f_2^* = \text{argmin}_{f \in \mathcal{F}} \text{Obj}_{t_2}(f),$$

$$f_1^* = \text{argmin}_{f \in \mathcal{F}} \text{Obj}_{t_1}(f).$$

Note that for binary classification, $H(f_1^*) \leq \frac{1}{2}$, because otherwise f_1^* would have worse objective than just using a default rule predicting the majority class (recall that such a default rule incurs the minimal complexity penalty of 0).

Then for any $f \in \hat{\mathbb{R}}_\epsilon^{(t_2)}$, we have:

[1] The proof extends trivially to multiclass categorical labels or continuous/ordinal/categorical feature values, provided the function class can handle these cases. The factor in the theorem may adjust based on the number of categories, or require sharper restrictions on epsilon and the worst case complexity penalty.

$$\text{Obj}_{t_2}(f) \leq \text{Obj}_{t_2}(f_2^*) + \epsilon$$

Note that $\text{Obj}_{t_2}(f_2^*) \leq \text{Obj}_{t_2}(f_1^*)$:

$$\text{Obj}_{t_2}(f) \leq \text{Obj}_{t_2}(f_1^*) + \epsilon$$

$$L_{t_2}(f) + H(f) \leq L_{t_2}(f_1^*) + H(f_1^*) + \epsilon$$

Apply the lower bound from Lemma 1.1 :

$$L_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + H(f) \leq L_{t_2}(f_1^*) + H(f_1^*) + \epsilon$$

Apply the upper bound from Lemma 1.1 :

$$L_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + H(f) \leq L_{t_1}(f_1^*)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + \frac{t_2 - t_1}{|I_{t_2}|} + H(f_1^*) + \epsilon$$

$$\text{Obj}_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + \frac{t_2 - t_1}{|I_{t_2}|}H(f) \leq \text{Obj}_{t_1}(f_1^*)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + \frac{t_2 - t_1}{|I_{t_2}|} + \frac{t_2 - t_1}{|I_{t_2}|}H(f_1^*) + \epsilon$$

$$\text{Obj}_{t_1}(f) \leq \text{Obj}_{t_1}(f_1^*) + \frac{\frac{t_2-t_1}{|I_{t_2}|} + \frac{t_2-t_1}{|I_{t_2}|}\left(H(f_1^*) - H(f)\right) + \epsilon}{1 - \frac{t_2-t_1}{|I_{t_2}|}} \tag{1}$$

Separately, we can use our given bound on $\epsilon' - \epsilon$ to state:

$$\epsilon' - \epsilon \geq 2\frac{t_2 - t_1}{|I_{t_2}|}$$

$$\epsilon' - \epsilon \geq \frac{t_2 - t_1}{|I_{t_2}|}(1 + \frac{1}{2} + \frac{1}{2})$$

$$\epsilon' - \epsilon \geq \frac{t_2 - t_1}{|I_{t_2}|}(1 + \frac{1}{2} + \epsilon')$$

$$\epsilon' - \epsilon \geq \frac{t_2 - t_1}{|I_{t_2}|}(1 + H(f_1^*) + \epsilon')$$

$$\epsilon' - \epsilon \geq \frac{t_2 - t_1}{|I_{t_2}|} + \frac{t_2 - t_1}{|I_{t_2}|}H(f_1^*) + \frac{t_2 - t_1}{|I_{t_2}|}\epsilon'$$

$$\epsilon' - \frac{t_2 - t_1}{|I_{t_2}|}\epsilon' \geq \frac{t_2 - t_1}{|I_{t_2}|} + \frac{t_2 - t_1}{|I_{t_2}|}H(f_1^*) + \epsilon$$

$$\epsilon'(1 - \frac{t_2 - t_1}{|I_{t_2}|}) \geq \frac{t_2 - t_1}{|I_{t_2}|} + \frac{t_2 - t_1}{|I_{t_2}|}H(f_1^*) + \epsilon$$

$$\epsilon' \geq \frac{\frac{t_2-t_1}{|I_{t_2}|} + \frac{t_2-t_1}{|I_{t_2}|}H(f_1^*) + \epsilon}{1 - \frac{t_2-t_1}{|I_{t_2}|}}$$

$$\epsilon' \geq \frac{\frac{t_2-t_1}{|I_{t_2}|} + \frac{t_2-t_1}{|I_{t_2}|}\left(H(f_1^*) - H(f)\right) + \epsilon}{1 - \frac{t_2-t_1}{|I_{t_2}|}}$$

Using that bound, we now return to Equation 1 to say:

$$\text{Obj}_{t_1}(f) \leq \text{Obj}_{t_1}(f_1^*) + \epsilon'$$

Which guarantees

$$f \in \hat{\mathbb{R}}_{\epsilon'}^{(t_1)},$$

as required. □

Lemma 1.1. *For* $0 \leq t_1 \leq t_2$,

$$L_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) \leq L_{t_2}(f) \leq L_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + \frac{t_2 - t_1}{|I_{t_2}|}$$

Proof.

$$L_{t_2}(f) = \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]} + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = \left(\frac{1}{|I_{t_1}|} - \left(\frac{1}{|I_{t_1}|} - \frac{1}{|I_{t_2}|}\right)\right) \sum_{i \in I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]} + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = \frac{1}{|I_{t_1}|} \sum_{i \in I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]} - \left(\frac{1}{|I_{t_1}|} - \frac{1}{|I_{t_2}|}\right) \sum_{i \in I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]} + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = L_{t_1}(f) - \left(\frac{1}{|I_{t_1}|} - \frac{1}{|I_{t_2}|}\right) \sum_{i \in I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]} + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = L_{t_1}(f) - \left(\frac{|I_{t_2}| - |I_{t_1}|}{|I_{t_1}||I_{t_2}|}\right) \sum_{i \in I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]} + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = L_{t_1}(f) - \left(\frac{|I_{t_2}| - |I_{t_1}|}{|I_{t_2}|}\right) L_{t_1}(f) + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = L_{t_1}(f) - \frac{t_2 - t_1}{|I_{t_2}|} L_{t_1}(f) + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

$$L_{t_2}(f) = L_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + \frac{1}{|I_{t_2}|} \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]}$$

Recognizing that $0 \leq \sum_{i \in I_{t_2}, i \notin I_{t_1}} \mathbb{1}_{[f(x_i) \neq y_i]} \leq t_2 - t_1$, we have:

$$L_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) \leq L_{t_2}(f) \leq L_{t_1}(f)(1 - \frac{t_2 - t_1}{|I_{t_2}|}) + \frac{t_2 - t_1}{|I_{t_2}|}$$

□