8 Algorithm FRL

In this section, we present Algorithm FRL in detail. Given an instance $(D, A, w, C)$ of Program 2.9, the algorithm searches through the space of falling rule lists that are compatible with $D$ and outputs a compatible falling rule list that respects the constraints of Program 2.9, and whose objective value is the smallest among all the falling rule lists that the algorithm explores. It does so by iterating over $T$ steps, in each of which the algorithm constructs a compatible falling rule list $d$, while keeping track of the falling rule list $d^*$ that has the smallest objective value $L_{\text{best}} = L(d^*, D, 1/(1+w), w, C)$ among all the falling rule lists that the algorithm has constructed so far. At the end of $T$ iterations, the algorithm outputs the falling rule list that has the smallest objective value out of the $T$ lists it has constructed.

In the process of constructing a falling rule list $d$, the algorithm chooses the antecedents successively: first for the antecedent $a_0^{(d)}$ in the top rule, then for the antecedent $a_1^{(d)}$ in the next rule, and so forth. For each antecedent $a_j^{(d)}$ chosen, the algorithm also computes its empirical positive proportion $\alpha_j^{(d), D}$. After $p$ rules have been constructed so that $d$ currently holds the prefix $e = \{(a_0^{(d)}, \alpha_0^{(d), D}), (a_1^{(d)}, \alpha_1^{(d), D}), \ldots, (a_{p-1}^{(d)}, \alpha_{p-1}^{(d), D})\}$, the algorithm either: (1) terminates the construction of $d$ by computing the empirical positive proportion after $e$, $\alpha_{e, D}$, and then adding to $d$ the final else clause with probability estimate $\alpha_{e, D}$, or (2) randomly picks an antecedent from a candidate set $S$ of possible next antecedents, computes its empirical positive proportion, and uses these as the next rule $(a_p^{(d)}, \alpha_p^{(d), D})$ for $d$.

The algorithm uses various properties of Program 2.9, which are presented in Section 4, to prune the search space. More specifically, the algorithm terminates the construction of $d$ if Inequality (9) in Theorem 4.6 holds. Otherwise it either terminates the construction of $d$ with some probability, or proceeds to construct a candidate set $S$ of possible next antecedents, as follows. For every antecedent $A_l \in A$ that has not been chosen before, it constructs a candidate next rule $(a_p^{(d)}, \alpha_p^{(d), D})$ by setting $a_p^{(d)} = A_l$ and computing $\alpha_p^{(d), D}$ using Definition 2.5. The algorithm then checks if the monotonicity constraint $\alpha_p^{(d), D} \leq \alpha_{p-1}^{(d), D}$ and the necessary condition for optimality $\alpha_p^{(d), D} > 1/(1+w)$ (Corollary 4.5) are satisfied, if the prefix $e' = \{e, (a_p^{(d)}, \alpha_p^{(d), D})\}$ is feasible under Program 2.9 (i.e. whether there exists a compatible falling rule list that begins with the prefix $e'$) using Proposition 4.2, and if the best possible objective value $L^*(e', D, w, C)$ achievable by any falling rule list that begins with $e'$ and is compatible with $D$ (Theorem 4.6) is less than the current best objective value $L_{\text{best}} = L(d^*, D, 1/(1+w), w, C)$. If all of the above conditions are satisfied, the algorithm adds $A_l$ to $S$. Once the construction of $S$ is complete, the algorithm randomly chooses an antecedent $A_l \in S$ with probability $P(A_l|S, e, D)$ and uses this antecedent, together with its empirical positive proportion, as the next rule $(a_p^{(d)}, \alpha_p^{(d), D})$ for $d$. If $S$ is empty, the algorithm terminates the construction of $d$.

In practice, we define the probability $P(A_l|S, e, D)$ for $A_l \in S$ by first defining a curiosity function $f_{S, e, D} : S \rightarrow \mathbb{R}_{\geq 0}$ and then normalizing it:

$$P(A_l|S, e, D) = \frac{f_{S, e, D}(A_l)}{\sum_{A_l'} f_{S, e, D}(A_{l'})}.$$ 

A possible choice of the curiosity function $f_{S, e, D}$ for use in Algorithm FRL is given by

$$f_{S, e, D}(A_l) = \lambda \alpha(A_l, e, D) + (1 - \lambda) \frac{n^+(A_l, e, D)}{\hat{n}_{e, D}}.$$ 

(11)
where $\alpha(A_t, e, D)$ is the empirical positive proportion of $A_t$, and $n^+(A_t, e, D)$ is the number of positive training inputs captured by $A_t$, should $A_t$ be chosen as the next antecedent after the prefix $e$. The curiosity function $f_{S,e,D}$ given by (11) is a weighted sum of $\alpha(A_t, e, D)$ and $n^+(A_t, e, D)/\hat{n}_{e,D}^+$ for each $A_t \in S$: the former encourages the algorithm to choose antecedents that have large empirical positive proportions, and the latter encourages the algorithm to choose antecedents that have large positive supports in the training data not captured by $e$. We used this curiosity function for Algorithm FRL in our experiments.

The pseudocode of Algorithm FRL is shown in Algorithm 1.

### Algorithm 1: Algorithm FRL

**Input:** an instance $(D, A, w, C)$ of Program 2.9  
**Result:** a falling rule list $d^*$ that are compatible with $D$ and whose antecedents come from $A$  
**initialize** $d^* = \emptyset$, $L_{best} = \infty$;  
**for** $t = 1, \ldots, T$ **do**  
  **while** Inequality (9) in Theorem 4.6 does not hold **do**  
    set $p = -1$, $\alpha_p = 1$, $d = e = \emptyset$;  
    **for** every antecedent $A_t \in A$ that is not in $d$ **do**  
      set $a_p^{(d)} = A_t$, compute $\alpha_p^{(d,D)}$, and let $e' = \{e, (a_p^{(d)}, \alpha_p^{(d,D)})\}$;  
      if $\alpha_p^{(d,D)} \leq \alpha_{p-1}^{(d,D)}$, $\alpha_p^{(d,D)} > 1/(1 + w)$, and $e'$ is feasible under Program 2.9 then  
        compute $L(e', D, w, C)$ using Theorem 4.6;  
        if $L(e', D, w, C) < L(d^*, D, 1/(1 + w), w, C)$ then  
          add $A_t$ to $S$;  
        end  
      end  
    end  
  end  
  **if** $S \neq \emptyset$ **then**  
    choose an antecedent $A_t \in S$ with probability $P(A_t | S, e, D)$ according to a discrete probability distribution over $S$;  
    set $a_p^{(d)} = A_t$ and add $(a_p^{(d)}, \alpha_p^{(d,D)})$ to $d$;  
    set $e = d$;  
    // save the partially constructed list $d$ as the prefix $e$  
  **else**  
    go to Terminate  
  end  
  Terminate: terminate the construction of $d$, and compute $L(d, D, 1/(1 + w), w, C)$;  
  **if** $L(d, D, 1/(1 + w), w, C) < L_{best}$ **then**  
    set $d^* = d$, $L_{best} = L(d, D, 1/(1 + w), w, C)$;  
  end  
end

### 9 Algorithm softFRL

In this section, we present Algorithm softFRL in detail. Given an instance $(D, A, w, C, C_1)$ of Program 5.1, the algorithm searches through the space of rule lists that are compatible with $D$ and finds a compatible rule list whose antecedents come from $A$, and whose objective value is the smallest among all the rule lists that the algorithm explores. It does so by iterating over $T$ steps, in each of which the algorithm constructs a compatible rule list $d$, while keeping track of the rule list $d^*$ that has the smallest objective value $L_{best} = L(d^*, D, 1/(1 + w), w, C, C_1)$ among all the rule lists that the algorithm has constructed so far. At the end of $T$ iterations, the algorithm transforms the rule list $d^*$ that has the smallest objective value out of
Theorem 2.8. Given the training data $D$, a rule list $d$ that is compatible with $D$, and the weight $w$ for the positive class, we have

$$R(d, D, 1/(1 + w), w) \leq R(d, D, \tau, w)$$

10 Proofs of Theorem 2.8, Proposition 4.2, Lemma 4.4, Corollary 4.5, and Theorem 4.6
Input: an instance \((D, A, w, C, C_1)\) of Program 5.1
Result: a falling rule list \(d^*\) whose antecedents come from \(A\)
initialize \(d^* = \emptyset\), \(\hat{L}_{\text{best}} = \infty\);
for \(t = 1, ..., T\) do
  set \(p = -1\), \(\alpha_p = 1\), \(d = e = \emptyset\);
  while \(\hat{L}^*(e, D, w, C, C_1) < \hat{L}(\bar{e}, D, 1/(1 + w), w, C, C_1)\) do
    go to Terminate with some probability;
    set \(p = p + 1\), \(S = \emptyset\);
    for every antecedent \(A_l \in A\) that is not in \(d\) do
      set \(a_p^{(d)} = A_l\), compute \(\alpha_p^{(d,D)}\), and let \(e' = \{e, (a_p^{(d)}, \alpha_p^{(d,D)})\}\);
      compute \(\hat{L}^*(e', D, w, C, C_1)\) using Theorem 5.2;
      if \(\hat{L}^*(e', D, w, C, C_1) < \hat{L}(d^*, D, 1/(1 + w), w, C, C_1)\) then
        add \(A_l\) to \(S\);
      end
    end
    if \(S \neq \emptyset\) then
      choose an antecedent \(A_l \in S\) with probability \(P(A_l | S, e, D)\) according to a discrete
      probability distribution over \(S\);
      set \(a_p^{(d)} = A_l\) and add \((a_p^{(d)}, \alpha_p^{(d,D)})\) to \(d\);
      set \(e = d\);
      // save the partially constructed list \(d\) as the prefix \(e\)
    else
      go to Terminate
    end
  end
  Terminate: terminate the construction of \(d\), and compute \(\hat{L}(d, D, 1/(1 + w), w, C, C_1)\);
  if \(\hat{L}(d, D, 1/(1 + w), w, C, C_1) < \hat{L}_{\text{best}}\) then
    set \(d^* = d\), \(\hat{L}_{\text{best}} = \hat{L}(d, D, 1/(1 + w), w, C, C_1)\);
  end
end
transform \(d^*\) into a falling rule list by setting \(\hat{\alpha}_j^{(d^*)} = \min_{k \leq j} \alpha_k^{(d^*, D)}\);

**Algorithm 2**: Algorithm softFRL
for all $\tau \geq 0$.

Proof. Suppose $\tau > 1/(1 + w)$. Consider the $j$-th rule $(\alpha_j^{(d)}, \alpha_j^{(d,D)})$ in $d$, whose antecedent captures $\alpha_j^{(d,D)} n_{j,d,D}$ positive training inputs and $(1 - \alpha_j^{(d,D)}) n_{j,d,D}$ negative training inputs. Let $R_j(d, D, \tau, w)$ denote the contribution by the $j$-th rule to $R(d, D, \tau, w)$, i.e.

$$R_j(d, D, \tau, w) = \frac{1}{n} \sum_{i:y_i=1} \mathbb{1}[\alpha_j^{(d,D)} \leq \tau] + \sum_{i:y_i=-1} \mathbb{1}[\alpha_j^{(d,D)} > \tau] = \begin{cases} \frac{1}{n} n_{j,d,D}^- & \text{if } \alpha_j^{(d,D)} > \tau \\ \frac{w}{n} n_{j,d,D}^+ & \text{otherwise.} \end{cases}$$

(13)

Case 1. $1/(1 + w) < \alpha_j^{(d,D)} \leq \tau$. In this case, we have

$$R_j(d, D, 1/(1 + w), w) = \frac{1}{n} n_{j,d,D}^- \quad \text{(by the definition of } R_j \text{ in Equation (13))}$$

$$= \frac{1}{n} (n_{j,d,D} - n_{j,d,D}^+) \quad \text{(by the definition of } n_{j,d,D}, n_{j,d,D}^+, n_{j,d,D}^- \text{ in Definition 2.5)}$$

$$= \frac{1}{n} (n_{j,d,D} - \alpha_j^{(d,D)} n_{j,d,D}) \quad \text{(by the definition of } \alpha_j^{(d,D)} \text{ in Definition 2.5)}$$

$$= \frac{1}{n} (1 - \alpha_j^{(d,D)}) n_{j,d,D}$$

$$< \frac{w}{n} \frac{1}{1 + w} n_{j,d,D}$$

$$= \frac{w}{n} n_{j,d,D}^+ \quad \text{(by the definition of } \alpha_j^{(d,D)} \text{ in Definition 2.5)}$$

$$= R_j(d, D, \tau, w). \quad \text{(by the definition of } R_j \text{ in Equation (13))}$$

Case 2. $\alpha_j^{(d,D)} > \tau$. In this case, both $R_j(d, D, 1/(1 + w), w)$ and $R_j(d, D, \tau, w)$ are equal to $\frac{1}{n} n_{j,d,D}^-.$

Case 3. $\alpha_j^{(d,D)} \leq 1/(1 + w)$. In this case, both $R_j(d, D, 1/(1 + w), w)$ and $R_j(d, D, \tau, w)$ are equal to $\frac{w}{n} n_{j,d,D}^+.$

Hence, given $\tau > 1/(1 + w)$, we have

$$R(d, D, 1/(1 + w), w) = \sum_{j=0}^{d} R_j(d, D, 1/(1 + w), w) \leq \sum_{j=0}^{d} R_j(d, D, \tau, w) = R(d, D, \tau, w).$$

The proof for $R(d, D, 1/(1 + w), w) \leq R(d, D, \tau, w)$ given $\tau < 1/(1 + w)$ is similar. \qed

Proposition 4.2. Given the training data $D$, the set of antecedents $A$, and a prefix $e$ that is compatible with $D$ and satisfies $\alpha_j^{(e)} \in A$ for all $j \in \{0, 1, \ldots, |e| - 1\}$ and $\alpha_k^{(e,D)} \geq \alpha_k^{(e)}$ for all $k \in \{1, 2, \ldots, |e| - 1\}$, the following statements are equivalent: (1) $e$ is feasible for Program 2.9 under $D$ and $A$; (2) $\hat{\alpha}_{e,D} \leq \alpha_{|e|-1}^{(e,D)}$ holds; (3) $\tilde{n}_{e,D} \geq ((1/\alpha_{|e|-1}^{(e,D)}) - 1)\hat{n}_{e,D}$ holds.

Proof. (1) $\Rightarrow$ (3): Suppose that Statement (1) holds. Then there exists a falling rule list

$$d = \{e, (\alpha_{|e|}^{(d)}, \alpha_{|e|-1}^{(d,D)}), \ldots, (\alpha_{|d|-1}^{(d)}, \alpha_{|d|-2}^{(d,D)}), (\alpha_{|d|-1}^{(d,D)})\}$$

This is a valid falling rule list with $\alpha_{|d|-1}^{(d,D)}$ positive training inputs and $(1 - \alpha_{|d|-1}^{(d,D)}) n_{|d|-1,d,D}$ negative training inputs.
that is compatible with \( D \), and we have
\[
\tilde{n}_{e,D} = \frac{\tilde{n}_{e,D}}{\bar{n}_{e,D}} = \frac{\tilde{n}_{e,D}}{\tilde{n}_{e,D} + \bar{n}_{e,D}} \\
\geq \frac{1}{\alpha_{|\cdot|}} \tilde{n}_{e,D}^+ + \frac{1}{\alpha_{|\cdot| - 1}} \tilde{n}_{e,D}^- - \tilde{n}_{e,D}^+ \\
= \frac{1}{\alpha_{|\cdot|}} \tilde{n}_{e,D}^+ - \frac{1}{\alpha_{|\cdot| - 1}} \tilde{n}_{e,D}^- \\
= ((1/\alpha_{|\cdot|}) - 1)\tilde{n}_{e,D}^+ \\
= ((1/\alpha_{|\cdot| - 1}) - 1)\tilde{n}_{e,D}^+.
\]

(3) \( \Rightarrow \) (2): Suppose that Statement (3) holds. Then we have
\[
\tilde{\alpha}_{e,D} = \frac{\tilde{n}_{e,D}^+}{\bar{n}_{e,D}} = \frac{\tilde{n}_{e,D}^+}{\tilde{n}_{e,D}^+ + \bar{n}_{e,D}^-} \\
\leq \frac{\tilde{n}_{e,D}^+}{\tilde{n}_{e,D}^+ + ((1/\alpha_{|\cdot|}) - 1)\tilde{n}_{e,D}^-} \\
= \frac{1}{(1 + (1/\alpha_{|\cdot|}) - 1)\bar{n}_{e,D}^-} = \alpha_{|\cdot| - 1}.
\]

(2) \( \Rightarrow \) (1): Suppose that Statement (2) holds. Then the falling rule list \( d = \{ e, \tilde{\alpha}_{e,D} \} \) begins with \( e \) and is compatible with \( D \). By Definition 4.1, \( e \) is feasible for Program 2.9 under the training data \( D \).

Before we proceed with proving Lemma 4.4, we make the following observation.

**Observation 10.1** For any rule list
\[
d' = \{ e, (a_{|\cdot|}^{(d')}, \tilde{\alpha}_{|\cdot|}), \ldots, (a_{|d'| - 1}^{(d')}, \tilde{\alpha}_{|d'| - 1}), \tilde{\alpha}_{|d'|}^{(d')} \}
\]
that begins with a given prefix \( e \), we have
\[
\tilde{n}_{e,D}^+ = n_{|\cdot|,d',D}^+ + \cdots n_{|d'|,d',D}^+, \quad (14)
\]
\[
\tilde{n}_{e,D}^- = n_{|\cdot|,d',D}^- + \cdots n_{|d'|,d',D}^-, \quad (15)
\]
and
\[
\tilde{n}_{e,D} = n_{|\cdot|,d',D} + \cdots n_{|d'|,d',D}. \quad (16)
\]

**Proof.** Any positive training input \( x \), that is not captured by the prefix \( e \) must be captured by some antecedent \( a_j^{(d')} \) with \( |e| \leq j < |d'| \) in \( d' \), or the final else clause in \( d' \). Conversely, any positive training input \( x \), that is captured by some antecedent \( a_j^{(d')} \) with \( |e| \leq j < |d'| \) in \( d' \), or the final else clause in \( d' \), must not satisfy
We can establish Equations (15) and (16) using essentially the same argument. We now prove Lemma 4.4.

To see this, we observe that the training instances captured by Step 2.

\[\frac{L(d', D, 1/(1 + w), w, C)}{L(d, D, 1/(1 + w), w, C)}\]

Moreover, if either \(\alpha_j^{(d,D)} > 1/(1 + w)\) holds for all \(j \in \{|e|, |e| + 1, ..., |d|\}\), or \(\alpha_j^{(d,D)} \leq 1/(1 + w)\) holds for all \(j \in \{|e|, |e| + 1, ..., |d|\}\), then the falling rule list \(\bar{e} = \{\bar{e}, \bar{\alpha}_{e,D}\}\) (i.e. the falling rule list in which the final else clause immediately follows the prefix \(e\), and the contribution of the final else clause is \(\bar{\alpha}_{e,D}\)) is compatible with \(D\) and satisfies \(L(\bar{e}, D, 1/(1 + w), w, C) \leq L(d, D, 1/(1 + w), w, C)\).

**Proof. Case 1.** There exists some \(k \in \{|e| + 1, ..., |d|\}\) that satisfies \(\alpha_k^{(d,D)} > 1/(1 + w)\) but \(\alpha_k^{(d,D)} \leq 1/(1 + w)\).

For any \(j \in \{|e|, ..., k - 1\}\), we have \(\alpha_j^{(d,D)} > 1/(1 + w)\), and the contribution \(R_j(d, D, 1/(1 + w), w)\) by the \(j\)-th rule to \(R(d, D, 1/(1 + w), w)\), defined by Equation (13) with \(\tau = 1/(1 + w)\), is given by

\[R_j(d, D, 1/(1 + w), w) = \frac{1}{n} n_{j,d,D}^- \] (17)

For any \(j \in \{k, ..., |d|\}\), we have \(\alpha_j^{(d,D)} \leq 1/(1 + w)\), and the contribution \(R_j(d, D, 1/(1 + w), w)\) by the \(j\)-th rule to \(R(d, D, 1/(1 + w), w)\) is given by

\[R_j(d, D, 1/(1 + w), w) = \frac{w}{n} n_{j,d,D}^+ \] (18)

The rest of the proof for this case proceeds in three steps.

**Step 1.** Construct a hypothetical falling rule list \(d'\) that begins with \(e\), has exactly one more rule (excluding the final else clause) following \(e\), and is compatible with \(D\). In later steps, we shall show that the falling rule list \(d'\) constructed in this step satisfies \(L(d', D, 1/(1 + w), w, C) \leq L(d, D, 1/(1 + w), w, C)\).

Let \(d' = \{e, (a_{|e|}^{(d')}, \hat{\alpha}_{|e|}), (\hat{\alpha}_{|e|+1})\}\) be the falling rule list of size \(|d'| = |e| + 1\) that is compatible with \(D\), such that

\[a_{|e|}^{(d')} = a_{|e|}^{(d)} \lor \ldots \lor a_{k-1}^{(d)}\]

is the antecedent given by the logical or's of the antecedents \(a_{|e|}^{(d')}\) through \(a_{k-1}^{(d)}\) in \(d\).

**Step 2.** Show that the empirical risk of misclassification by the falling rule list \(d'\) is the same as that by the falling rule list \(d\).

To see this, we observe that the training instances captured by \(a_{|e|}^{(d')}\) in \(d'\) are exactly those captured by the antecedents \(a_{|e|}^{(d)}\) through \(a_{k-1}^{(d)}\) in \(d\), and the training instances captured by \(a_{|e|+1}^{(d')}\) (i.e. the final else clause) in \(d'\) are exactly those captured by the antecedents \(a_{k-1}^{(d)}\) through \(a_{|d|}^{(d)}\) in \(d\). This observation implies

\[n_{|e|,d,D}^+ + n_{k-1,d,D}^+ \]

\[n_{|e|,d',D}^+ = n_{|e|,d,D}^+ + \ldots + n_{k-1,d,D}^+ \] (19)

\[n_{|e|,d',D}^+ = n_{|e|,d,D}^+ + \ldots + n_{k-1,d,D}^+ \] (19)
This means that
\[ n_{|e|,d',D} = n_{|e|,d,D} + \ldots + n_{k-1,d,D}, \]  
\[ n_{|e|,d',D} = n_{|e|,d,D} + \ldots + n_{k-1,d,D}, \]  
\[ n_{|e|+1,d',D} = n_{k,d,d,D} + \ldots + n_{|d|,d,D}, \]
and
\[ n_{|e|+1,d',D} = n_{k,d,d,D} + \ldots + n_{|d|,d,D}. \]

Since \( d' \) is compatible with \( D \), using the definition of a compatible rule list in Definition 2.6 and the definition of the empirical positive proportion in Definition 2.5, together with (19), (21), (22), and (23), we must have
\[ \frac{n_{|e|,d',D}}{n_{|e|,d,D}} = \frac{n_{|e|,d,D}}{n_{|e|,d,D}} + \ldots + \frac{n_{k-1,d,D}}{n_{k-1,d,D}} > \frac{1}{1 + w}, \]
and
\[ \frac{n_{|e|+1,d',D}}{n_{|e|+1,d,D}} = \frac{n_{k,d,d,D}}{n_{k,d,d,D}} + \ldots + \frac{n_{|d|,d,D}}{n_{|d|,d,D}} \leq \frac{1}{1 + w}. \]

This means that the contribution \( R_{|e|}(d', D, 1/(1 + w), w) \) by the \(|e|\)-th rule to \( R(d', D, 1/(1 + w), w) \) is given by
\[ R_{|e|}(d', D, 1/(1 + w), w) = \frac{1}{n} n_{|e|,d',D} = \frac{1}{n} (n_{|e|,d,D} + \ldots + n_{k-1,d,D}), \]
where we have used (20), and the contribution \( R_{|e|+1}(d', D, 1/(1 + w), w) \) by the \((|e| + 1)\)-st “rule” (i.e. the final else clause) to \( R(d', D, 1/(1 + w), w) \) is given by
\[ R_{|e|+1}(d', D, 1/(1 + w), w) = \frac{w}{n} n_{|e|+1,d',D} = \frac{w}{n} (n_{k,d,d,D} + \ldots + n_{|d|,d,D}), \]
where we have used (22).

It then follows that the empirical risk of misclassification by the rule list \( d' \) is the same as that by the rule list \( d \):
\[ R(d', D, 1/(1 + w), w) = R(e, D, 1/(1 + w), w) + R_{|e|}(d', D, 1/(1 + w), w) + R_{|e|+1}(d', D, 1/(1 + w), w) \]
\[ = R(e, D, 1/(1 + w), w) + \frac{1}{n} (n_{|e|,d,D} + \ldots + n_{k-1,d,D}) + \frac{w}{n} (n_{k,d,d,D} + \ldots + n_{|d|,d,D}) \]
\[ = R(e, D, 1/(1 + w), w) + \sum_{j=|e|}^{|d|} R_j(d, D, 1/(1 + w), w) \]
\[ = R(d, D, 1/(1 + w), w). \]

Step 3. Put everything together.

Using (24), together with the observation \(|d'| = |e| + 1 \leq |d|\), we must also have
\[ L(d', D, 1/(1 + w), w, C) = R(d', D, 1/(1 + w), w) + C|d'| \leq R(d, D, 1/(1 + w), w) + C|d| = L(d, D, 1/(1 + w), w, C), \]
as desired.

**Case 2.** \( \alpha^{(d, D)}_{j} > 1/(1+w) \) holds for all \( j \in \{|e|, |e|+1, \ldots, |d|\} \). Then the contribution \( R_{j}(d, D, 1/(1+w), w) \) by the \( j \)-th rule to \( R(d, D, 1/(1+w), w) \), for all \( j \in \{|e|, |e|+1, \ldots, |d|\} \), is given by Equation (17). Let \( d' = \{e, \tilde{a}_{e}^{(d')}\} \) be the falling rule list of size \( |d'| = |e| \) that is compatible with \( D \). Then the instances captured by \( \tilde{a}_{e}^{(d')} \) (i.e. the final else clause) in \( d' \) are exactly those that are not captured by \( e \), or equivalently, those that are captured by \( a_{|e|}^{(d)} \) through \( a_{|d|}^{(d)} \). This implies

\[
\begin{align*}
n_{|e|, d', D}^{+} &= n_{|e|, d, D}^{+} + \ldots + n_{|d|, d, D}^{+} \\
n_{|e|, d', D}^{-} &= n_{|e|, d, D}^{-} + \ldots + n_{|d|, d, D}^{-} \\
n_{|e|, d', D} &= n_{|e|, d, D} + \ldots + n_{|d|, d, D}.
\end{align*}
\]

(25) and (26)

Since \( d' \) is compatible with \( D \), using the definition of a compatible rule list in Definition 2.6 and the definition of the empirical positive proportion in Definition 2.5, together with (25) and (27), we must have

\[
\alpha^{(d')}_{|e|} = \alpha^{(d', D)}_{|e|} = \frac{n_{|e|, d', D}^{+}}{n_{|e|, d', D}} = \frac{n_{|e|, d, D}^{+} + \ldots + n_{|d|, d, D}^{+}}{n_{|e|, d, D} + \ldots + n_{|d|, d, D}} = n_{|e|, d, D}^{-} + \ldots + n_{|d|, d, D}^{-} > \frac{1}{1+w}.
\]

(28) and (29)

Note that the right-hand side of Equality (28) is equal to \( n_{|e|, D}^{+}/\tilde{n}_{e, D} = \tilde{\alpha}_{e, D} \), by Equations (14) and (16) in Observation 10.1. Therefore, we also have \( \tilde{\alpha}_{|e|}^{(d')} = \tilde{\alpha}_{e, D} \).

Inequality (29) implies that the contribution \( R_{|e|}(d', D, 1/(1+w), w) \) by the \( |e| \)-th “rule” (i.e. the final else clause) to \( R(d', D, 1/(1+w), w) \) is given by

\[
R_{|e|}(d', D, 1/(1+w), w) = \frac{1}{n} n_{|e|, d', D}^{-} = \frac{1}{n} (n_{|e|, d, D}^{-} + \ldots + n_{|d|, d, D}^{-}),
\]

where we have used (26).

It then follows that the empirical risk of misclassification by the rule list \( d' \) is the same as that by the rule list \( d \):

\[
\begin{align*}
R(d', D, 1/(1+w), w) &= R(e, D, 1/(1+w), w) + R_{|e|}(d', D, 1/(1+w), w) \\
&= R(e, D, 1/(1+w), w) + \frac{1}{n} (n_{|e|, d, D}^{-} + \ldots + n_{|d|, d, D}^{-}) \\
&= R(e, D, 1/(1+w), w) + \sum_{j=|e|}^{|d|} R_{j}(d, D, 1/(1+w), w) \\
&= R(d, D, 1/(1+w), w).
\end{align*}
\]

Since we clearly have \( |d'| = |e| \leq |d| \), we must also have

\[
L(d', D, 1/(1+w), w, C) = R(d', D, 1/(1+w), w) + C|d'| \\
\leq R(d, D, 1/(1+w), w) + C|d| = L(d, D, 1/(1+w), w, C),
\]

(26)
as desired.

**Case 3.** \( \alpha_j^{(d,D)} \leq 1/(1 + w) \) holds for all \( j \in \{|e|, |e| + 1, \ldots, |d|\} \). The proof is similar to Case 2, with \( R_{|e|}(d, D, 1/(1 + w), w) \) for all \( j \in \{|e|, |e| + 1, \ldots, |d|\} \) given by Equation (18), the “greater than” in Inequality replaced by “less than or equal to”, and \( R_{|e|}(d', D, 1/(1 + w), w) \) given by

\[
R_{|e|}(d', D, 1/(1 + w), w) = \frac{w}{n} n^+_{|e|, |d'|, D} = \frac{w}{n} \left(n^+_{|e|, |d|, D} + \cdots + n^+_{|d|, |d|, D}\right).
\]

**Corollary 4.5.** If \( d^* \) is an optimal solution for a given instance \( (D, A, w, C) \) of Program 2.9, then we must have \( \alpha_j^{(d^*, D)} > 1/(1 + w) \) for all \( j \in \{0, 1, \ldots, |d^*| - 1\} \).

**Proof.** Suppose that \( d^* \) were an optimal solution for a given instance \( (D, A, w, C) \) of Program 2.9, such that \( \alpha_k^{(d^*, D)} \leq 1/(1 + w) \) for some \( k \in \{0, 1, \ldots, |d^*| - 1\} \). Let

\[
e = \{(\alpha_0^{(d^*)}, \alpha_0^{(d^*, D)}), \ldots, (\alpha_{k-1}^{(d^*)}, \alpha_{k-1}^{(d^*, D)})\}
\]

be a prefix consisting of the top \( k \) rules in \( d^* \). By Lemma 4.4, the falling rule list \( \tilde{e} = \{e, \alpha_{e,D}\} \) satisfies \( L(\tilde{e}, D, 1/(1 + w), w, C) \leq L(d^*, D, 1/(1 + w), w, C) \). In fact, the inequality is strict because the size of \( \tilde{e} \) is strictly less than that of \( d^* \). This contradicts the optimality of \( d^* \). \( \square \)

Before we proceed with proving Theorem 4.6, we make two other observations.

**Observation 10.2.** For any rule list \( d' \), we have

\[
n^+_{|e|, |d'|, D} = \left(\frac{1}{\alpha^+_{|e|}} - 1\right) n^+_{|e|, |d|, D}.
\]

**Proof.** By Definition 2.5, we have

\[
\alpha^+_{|e|} = n^+_{|e|, |d|, D}/n^+_{|e|, |d'|, D}.
\]

Since \( n^+_{|e|, |d|, D} \) denotes the total number of training inputs captured by the \( |e| \)-th antecedent in \( d' \), which is exactly the sum of the number of positive training inputs captured by that antecedent (denoted \( n^+_{|e|, |d|, D} \)), and the number of negative training inputs captured by the same antecedent (denoted \( n^-_{|e|, |d|, D} \)), we have

\[
\alpha^+_{|e|} = \frac{n^+_{|e|, |d|, D}}{n^+_{|e|, |d|, D} + n^-_{|e|, |d|, D}}.
\]

The desired equation follows from rearranging the terms. \( \square \)

**Observation 10.3.** For any rule list

\[
d' = \{e, (\alpha^-_{|e|}, \alpha^+_{|e|}, \alpha^+_{|e| + 1})\}
\]

that has exactly one rule (excluding the final else clause) following a given prefix \( e \), we have

\[
n^+_{|e| + 1, |d'|, D} = \tilde{n}^+_{e,D} - n^+_{|e|, |d'|, D},
\]

\[
n^-_{|e| + 1, |d'|, D} = \tilde{n}^-_{e,D} - n^-_{|e|, |d'|, D},
\]

and

\[
n_{|e| + 1, |d'|, D} = \tilde{n}_{e,D} - n_{|e|, |d'|, D}.
\]

Note that since \( n^+_{|e| + 1, |d'|, D}, n^-_{|e| + 1, |d'|, D} \) and \( n_{|e| + 1, |d'|, D} \) are non-negative, Equations (31), (32), and (33) imply \( n^+_{|e|, |d'|, D} \leq \tilde{n}^+_{e,D}, n^-_{|e|, |d'|, D} \leq \tilde{n}^-_{e,D}, \) and \( n_{|e|, |d'|, D} \leq \tilde{n}_{e,D}. \)
Proof. Applying Observation 10.1 with \(|d'| = |e| + 1\), we have
\[
\tilde{n}_{e,D}^+ = n_{|e|,d',D}^+ + n_{|e|+1,d',D},
\]
\[
\tilde{n}_{e,D}^- = n_{|e|,d',D}^- + n_{|e|+1,d',D},
\]
and
\[
\tilde{n}_{e,D} = n_{|e|,d',D} + n_{|e|+1,d',D}.
\]
Equations (31), (32), and (33) follow from rearranging the terms in the above equations.

We now prove Theorem 4.6.

**Theorem 4.6.** Suppose that we are given an instance \((D, A, w, C)\) of Program 2.9 and a prefix \(e\) that is feasible for Program 2.9 under the training data \(D\) and the set of antecedents \(A\). Then any falling rule list \(d\) that begins with \(e\) and is compatible with \(D\) satisfies
\[
L(d, D, 1/(1 + w), w, C) \geq L^*(e, D, w, C),
\]
where
\[
L^*(e, D, w, C) = L(e, D, 1/(1 + w), w, C) + \min \left( \frac{1}{n} \left( \frac{1}{\bar{\alpha}_{|e|-1}^{(e,D)}} - 1 \right) \tilde{n}_{e,D}^+, \frac{w}{n} \tilde{n}_{e,D}^+ - \frac{1}{n} \tilde{n}_{e,D}^- \right)
\]
is a lower bound on the objective value of any compatible falling rule list that begins with \(e\), which we call a prefix bound for \(e\), under the instance \((D, A, w, C)\) of Program 2.9. Furthermore, if
\[
C \geq \min \left( \frac{w}{n} \frac{n_{e,D}^+}{\tilde{n}_{e,D}^+}, \frac{1}{n} \frac{n_{e,D}^-}{\tilde{n}_{e,D}^-} \right) - \frac{1}{n} \left( \frac{1}{\bar{\alpha}_{|e|-1}^{(e,D)}} - 1 \right) \tilde{n}_{e,D}^+
\]
holds, then the falling rule list \(\bar{e} = \{e, \bar{\alpha}_{e,D}\}\) satisfies \(L(\bar{e}, D, 1/(1 + w), w, C) = L^*(e, D, w, C)\).

**Proof.** Let \(F(X, D, e)\) be the set of (hypothetical and non-hypothetical) falling rule lists that begin with \(e\) and are compatible with \(D\), and let \(F(X, D, e, k)\) be the subset of \(F(X, D, e)\), consisting of those falling rule lists in \(F(X, D, e)\) that have exactly \(k\) rules (excluding the final else clause) following the prefix \(e\).

Let \(d \in F(X, D, e)\).

**Case 1.** \(\alpha_{|e|-1}^{(e,D)} > 1/(1 + w)\).

In this case, Lemma 4.4 implies
\[
L(d, D, 1/(1 + w), w, C) \geq \inf_{d' \in F(X, D, e, 1) \cup F(X, D, e, 0)} L(d', D, 1/(1 + w), w, C).
\]

Note that we have \(F(X, D, e, 0) = \{\bar{e}\}\), where \(\bar{e} = \{e, \bar{\alpha}_{e,D}\}\) is the falling rule list in which the final else clause immediately follows the prefix \(e\), and the probability estimate of the final else clause is \(\bar{\alpha}_{e,D}\). To see this, we first observe \(\bar{e} \in F(X, D, e, 0)\). This is because:
(i) \(\bar{e}\) clearly begins with \(e\), and has no additional rules (excluding the final else clause) following the prefix \(e\);
(ii) the feasibility of \(e\) implies \(\alpha_{k-1}^{(e,D)} \geq \alpha_k^{(e,D)}\) for all \(k \in \{1, 2, ..., |e| - 1\}\) (otherwise we could not possibly have a falling rule list that begins with \(e\), and we would violate Definition 4.1), and \(\bar{\alpha}_{e,D} \leq \alpha_{|e|-1}^{(e,D)}\) (by Proposition 4.2), which together imply that \(\bar{e}\) is indeed a falling rule list; and
(iii) we have
\[
\bar{\alpha}_{e,D} = \frac{n_{e,D}^-}{n_{e,D}^+} \quad \text{(by the definition of \(\bar{\alpha}_{e,D}\) in Definition 2.5)}
\]
\[
= \frac{n_{|e|,\bar{e},D}^+}{n_{|e|,\bar{e},D}^-} \quad \text{(by Equations (14) and (16) in Observation 10.1, applied to \(\bar{e}\))}
\]
\[
= \alpha_{|e|}^{(e,D)} \quad \text{(by the definition of the empirical positive proportion in Definition 2.5)}
\]
which implies that $\bar{e}$ is indeed compatible with $D$.

Conversely, for any $d_0 = \{e, \alpha^{(d_0)}_{|e|}\} \in \mathcal{F}(\mathcal{X}, D, e, 0)$, we must have

$$\alpha^{(d_0)}_{|e|} = \alpha^{(d_0, D)}_{|e|}$$

(because $d_0$ must be compatible with $D$)

$$= \frac{n^+_{|e|, d_0}}{n_{|e|, d_0, D}}$$

(by the definition of the empirical positive proportion in Definition 2.5)

$$= \frac{\bar{\alpha}_{e, D}}{\hat{n}^+_{e, D}}$$

(by Equations (14) and (16) in Observation 10.1, applied to $d_0$ here)

$$= \bar{\alpha}_{e, D},$$

which implies $d_0 = \bar{e}$. This establishes $\mathcal{F}(\mathcal{X}, D, e, 0) = \{\bar{e}\}$.

Let $\mathcal{F}'(\mathcal{X}, D, e, 1)$ be the subset of $\mathcal{F}(\mathcal{X}, D, e, 1)$, consisting of those falling rule lists

$$d' = \{e, (\alpha^{(d')}_{|e|}, \alpha^{(d', D)}_{|e|}, \alpha^{(d', D)}_{|e|+1}) \in \mathcal{F}(\mathcal{X}, D, e, 1)$$

with $\alpha^{(d', D)}_{|e|} > 1/(1 + w)$ and $\alpha^{(d', D)}_{|e|+1} \leq 1/(1 + w)$. Note that for any $d_1 = \{e, (\alpha^{(d_1)}_{|e|}, \alpha^{(d_1, D)}_{|e|}, \alpha^{(d_1, D)}_{|e|+1}) \in \mathcal{F}(\mathcal{X}, D, e, 1) - \mathcal{F}'(\mathcal{X}, D, e, 1)$, we have either $\alpha^{(d_1)}_{|e|} \geq \alpha^{(d_1, D)}_{|e|} > 1/(1 + w)$ or $\alpha^{(d_1, D)}_{|e|} \leq \alpha^{(d_1, D)}_{|e|+1} \leq 1/(1 + w)$, and Lemma 4.4 implies $L(d_1, D, 1/(1 + w), w, C) \geq L(\bar{e}, D, 1/(1 + w), w, C)$. This means

$$\inf_{d' \in \mathcal{F}(\mathcal{X}, D, e, 1) - \mathcal{F}'(\mathcal{X}, D, e, 1)} L(d', D, 1/(1 + w), w, C) \geq L(\bar{e}, D, 1/(1 + w), w, C).$$

Using $\mathcal{F}(\mathcal{X}, D, e, 0) = \{\bar{e}\}$ and (36), we can write the right-hand side of (35) as

$$\inf_{d' \in \mathcal{F}(\mathcal{X}, D, e, 1)} L(d', D, 1/(1 + w), w, C)$$

$$= \inf_{d' \in \mathcal{F}(\mathcal{X}, D, e, 1) \cup \mathcal{F}(\mathcal{X}, D, e, 0)} L(d', D, 1/(1 + w), w, C)$$

$$= \min \left( \inf_{d' \in \mathcal{F}(\mathcal{X}, D, e, 1) \cup \mathcal{F}(\mathcal{X}, D, e, 0)} L(d', D, 1/(1 + w), w, C) \right)$$

$$= \min \left( \inf_{d' \in \mathcal{F}(\mathcal{X}, D, e, 1) \cup \mathcal{F}(\mathcal{X}, D, e, 0)} L(d', D, 1/(1 + w), w, C), L(\bar{e}, D, 1/(1 + w), w, C) \right).$$

The rest of the proof for this case proceeds in three steps.

**Step 1.** Compute $L(\bar{e}, D, 1/(1 + w), w, C)$.

Since the contribution by the final else clause to $L(\bar{e}, D, 1/(1 + w), w, C)$ is given by

$$R_{\bar{e}}(\bar{e}, D, 1/(1 + w), w) = \begin{cases} \frac{1}{n} n^-_{\bar{e}, D} & \text{if } \bar{\alpha}_{e, D} > 1/(1 + w) \\ \frac{w}{n} n^+_{\bar{e}, D} & \text{otherwise}, \end{cases}$$

where we have used Equation (13), and since Observation 10.1 implies $\bar{n}^+_{\bar{e}, D} = n^+_{\bar{e}, e, D}$ and $\bar{n}^-_{\bar{e}, D} = n^-_{\bar{e}, e, D}$, it is not difficult to see

$$L(\bar{e}, D, 1/(1 + w), w, C) = \begin{cases} L(e, D, 1/(1 + w), w, C) + \frac{1}{n} \bar{n}^-_{e, D} & \text{if } \bar{\alpha}_{e, D} > 1/(1 + w) \\ L(e, D, 1/(1 + w), w, C) + \frac{w}{n} \bar{n}^+_{e, D} & \text{otherwise}. \end{cases}$$
Since \( \tilde{\alpha}_{e,D} > 1/(1+w) \) is equivalent to \( \tilde{n}_{e,D}^+ / (\tilde{n}_{e,D}^- + \tilde{n}_{e,D}^-) > 1/(1+w) \), or \( w\tilde{n}_{e,D}^+ > \tilde{n}_{e,D}^- \), and similarly \( \tilde{\alpha}_{e,D} \leq 1/(1+w) \) is equivalent to \( w\tilde{n}_{e,D}^+ \leq \tilde{n}_{e,D}^- \), we can write

\[
L(\tilde{e}, D, 1/(1+w), w, C) = \min \left( L(e, D, 1/(1+w), w, C) + \frac{1}{n} \tilde{n}_{e,D}^-, \right.
\]
\[
\left. L(e, D, 1/(1+w), w, C) + \frac{w}{n} \tilde{n}_{e,D}^- \right).
\]

(38)

**Step 2.** Determine a lower bound of \( L(d', D, 1/(1+w), w, C) \) for all \( d' \in \mathcal{F}'(X, D, e, 1) \).

Let \( d' = \{ e, (\alpha'_{[e]}), \alpha'_{[e]+1} \} \in \mathcal{F}'(X, D, e, 1) \). Since the contribution by both the \([e]\)-th rule and the final else clause to \( L(d', D, 1/(1+w), w, C) \) is given by \( R_{[e]}(d', D, 1/(1+w), w) + R_{[e]+1}(d', D, 1/(1+w), w) + C \), where \( R_{[e]}(d', D, 1/(1+w), w) \) and \( R_{[e]+1}(d', D, 1/(1+w), w) \) are defined by Equation (39) and are given by

\[
R_{[e]}(d', D, 1/(1+w), w) = \frac{1}{n} n_{[e],d,D}^- \quad \text{and} \quad R_{[e]+1}(d', D, 1/(1+w), w) = \frac{w}{n} n_{[e]+1,d,D}^+ \]

(because we have \( \alpha'_{[e]} > 1/(1+w) \) and \( \alpha'_{[e]+1} \leq 1/(1+w) \) for \( d' \in \mathcal{F}'(X, D, e, 1) \)), it is not difficult to see

\[
L(d', D, 1/(1+w), w, C) = L(e, D, 1/(1+w), w, C) + \frac{1}{n} n_{[e],d,D}^- + \frac{w}{n} n_{[e]+1,d,D}^+ + C.
\]

(39)

Substituting (39) in Observation 10.2 and (31) in Observation 10.3 into Equation (39), we have

\[
L(d', D, 1/(1+w), w, C)
\]
\[
= L(e, D, 1/(1+w), w, C) + \frac{1}{n} \left( \frac{1}{\alpha'_{[e]}} - 1 \right) n_{[e],d,D}^+ + \frac{w}{n} (\tilde{n}_{e,D}^- - n_{[e],d,D}) + C
\]
\[
= L(e, D, 1/(1+w), w, C) + \frac{1}{n} \left( \left( \frac{1}{\alpha'_{[e]}} - 1 - w \right) n_{[e],d,D}^+ + w\tilde{n}_{e,D}^- \right) + C.
\]

(40)

Note that Equation (40) shows that given the prefix \( e \), \( L(d', D, 1/(1+w), w, C) \) is a function of \( \alpha'_{[e]} \) and of \( n_{[e],d,D}^+ \). Since we have

\[
\frac{\partial L(d', D, 1/(1+w), w, C)}{\partial n^+_{[e],d,D}} = \frac{1}{n} \left( \frac{1}{\alpha'_{[e]}} - 1 - w \right) < 0
\]

because \( \alpha'_{[e]} > 1/(1+w) \) holds for any \( d' \in \mathcal{F}'(X, D, e, 1) \), and

\[
\frac{\partial L(d', D, 1/(1+w), w, C)}{\partial \alpha'_{[e]}} = -\frac{n_{[e],d,D}^+}{n} \frac{1}{(\alpha'_{[e]})^2} \leq 0,
\]

we see that \( L(d', D, 1/(1+w), w, C) \) is indeed a monotonically decreasing function of both \( n_{[e],d,D}^+ \) and \( \alpha'_{[e]} \). Thus, we can obtain a lower bound of \( L(d', D, 1/(1+w), w, C) \) by substituting \( n_{[e],d,D}^+ \) and \( \alpha'_{[e]} \) with their respective upper bound. The inequality \( n_{[e],d,D}^+ \leq \tilde{n}_{e,D}^+ \) in Observation 10.3 gives an upper bound for \( n_{[e],d,D}^+ \), and the inequality \( \alpha'_{[e]} \leq \alpha'_{[e]+1} = \alpha_{[e]-1} \), from \( d' \) being a falling rule list gives an upper bound for \( \alpha'_{[e]} \). Substituting these upper bounds into (40), we obtain the following inequality, which gives...
This implies which means

\[ \alpha \min \left( \frac{1}{\alpha_{|e|-1}} - 1 \right) \tilde{n}_{e,D}^+ \]

This means

\[ \min \left( L(e, D, 1/(1+w), w, C) \right) \]

\[ \geq \min \left( L(d, D, 1/(1+w), w, C), L(\bar{e}, D, 1/(1+w), w, C) \right) \]

\[ \geq \min \left( L(e, D, 1/(1+w), w, C) + \frac{1}{n} \left( \frac{1}{\alpha_{|e|-1}} - 1 \right) \tilde{n}_{e,D}^+ \right) + C, \]

\[ \min \left( L(e, D, 1/(1+w), w, C) + \frac{1}{n} \tilde{n}_{e,D}^-, L(e, D, 1/(1+w), w, C) + \frac{w}{n} \tilde{n}_{e,D}^+ \right) \]

\[ = L(e, D, 1/(1+w), w, C) + \min \left( \frac{1}{n} \left( \frac{1}{\alpha_{|e|-1}} - 1 \right) \tilde{n}_{e,D}^+ + C, \frac{w}{n} \tilde{n}_{e,D}^+, \frac{1}{n} \tilde{n}_{e,D}^- \right), \]

as desired.

**Case 2.** \( \alpha_{|e|-1} \leq 1/(1+w) \).

This implies \( \alpha_{j}^{(d,D)} \leq 1/(1+w) \) for all \( j \in \{|e|, \ldots, |d|\} \). By Lemma 4.4, we have

\[ L(d, D, 1/(1+w), w, C) \geq L(\bar{e}, D, 1/(1+w), w, C). \]

Since \( L(\bar{e}, D, 1/(1+w), w, C) \) is given by Equation (38), we have

\[ L(d, D, 1/(1+w), w, C) \geq L(e, D, 1/(1+w), w, C) + \min \left( \frac{w}{n} \tilde{n}_{e,D}^+, \frac{1}{n} \tilde{n}_{e,D}^- \right). \]

(42)

Given \( \alpha_{|e|-1} \leq 1/(1+w) \), we must also have

\[ \frac{1}{n} \left( \frac{1}{\alpha_{|e|-1}} - 1 \right) \tilde{n}_{e,D}^+ \]

which means

\[ \min \left( \frac{w}{n} \tilde{n}_{e,D}^+, \frac{1}{n} \tilde{n}_{e,D}^- \right) = \min \left( \frac{w}{n} \tilde{n}_{e,D}^+, \frac{1}{n} \tilde{n}_{e,D}^- \right). \]

(43)
Theorem 5.2.

To prove Theorem 5.2, we need the following lemma:

Finally, if Inequality (34) holds, then we have

Substituting (43) into (42) completes the proof for Case 2.

\[ \frac{1}{n} \left( \frac{1}{\alpha^{(d',D)}} - 1 \right) \tilde{n}^+_{e,D} + C \geq \min \left( \frac{w}{n} \tilde{n}^+_{e,D}, \frac{1}{n} \tilde{n}^-_{e,D} \right), \]

which implies

\[ L^*(e, D, 1/(1 + w), w, C) = L(e, D, 1/(1 + w), w, C) + \min \left( \frac{w}{n} \tilde{n}^+_{e,D}, \frac{1}{n} \tilde{n}^-_{e,D} \right) = L(\bar{e}, D, 1/(1 + w), w, C). \]

\[ L^*(e, D, 1/(1 + w), w, C, C) = \bar{L}(e, D, 1/(1 + w), w, C, C) \]

11 Proof of Theorem 5.2

Theorem 5.2. Suppose that we are given an instance \((D, A, w, C, C_1)\) of Program 5.1 and a prefix \(e\) that is compatible with \(D\). Then any rule list \(d\) that begins with \(e\) and is compatible with \(D\) satisfies

\[ \bar{L}(d, D, 1/(1 + w), w, C, C_1) \geq \bar{L}^*(e, D, w, C, C_1), \]

where

\[ \bar{L}^*(e, D, w, C, C_1) = \bar{L}(e, D, 1/(1 + w), w, C, C_1) \]

\[ + \min \left( \frac{1}{n} \left( \frac{1}{\alpha^{(e,D)}} - 1 \right) \tilde{n}^+_{e,D} + C + C_1 | \tilde{\alpha}_{e,D} - \alpha^{(e,D)}_{\min} |, \right. \]

\[ \left. \frac{w}{n} \tilde{n}^+_{e,D} + C \tilde{\alpha}_{e,D} + C_1 | \tilde{\alpha}_{e,D} - \alpha^{(e,D)}_{\min} |, \right) \]

is a lower bound on the objective value of any compatible rule list that begins with \(e\), under the instance \((D, A, w, C, C_1)\) of Program 5.1. In Equation (44), \(\alpha^{(e,D)}_{\min}\), \(\zeta\), and \(g\) are defined by

\[ \alpha^{(e,D)}_{\min} = \min_{k \leq |e|} \alpha^{(e,D)}_{k}, \quad \zeta = \max(\alpha^{(e,D)}_{\min} \tilde{\alpha}_{e,D}, 1/(1 + w)), \]

\[ g(\beta) = \frac{1}{n} \left( \beta - 1 \right) \tilde{n}^+_{e,D} + C + C_1 (\beta - \alpha^{(e,D)}_{\min}). \]

Note that \(\inf_{\beta: \zeta \leq \beta \leq 1} g(\beta)\) can be computed analytically: \(\inf_{\beta: \zeta \leq \beta \leq 1} g(\beta) = g(\beta^*)\) if \(\beta^* = \sqrt{\tilde{n}^+_{e,D}/(C_1 n)}\) satisfies \(\zeta < \beta^* \leq 1\), and \(\inf_{\beta: \zeta \leq \beta \leq 1} g(\beta) = \min(g(\zeta), g(1))\) otherwise.

To prove Theorem 5.2, we need the following lemma:

Lemma. Suppose that we are given an instance \((D, A, w, C, C_1)\) of Program 5.1, a prefix \(e\) that is compatible with \(D\), and a (possibly hypothetical) rule list \(d\) that begins with \(e\) and is compatible with \(D\). Then there exists a rule list \(d'\), possibly hypothetical with respect to \(A\), such that \(d'\) begins with \(e\), has at most one more rule (excluding the final else clause) following \(e\), is compatible with \(D\), and satisfies

\[ \bar{L}(d', D, 1/(1 + w), w, C, C_1) \leq \bar{L}(d, D, 1/(1 + w), w, C, C_1). \]

Moreover, if either \(\alpha^{(d,D)}_j > 1/(1 + w)\) holds for all \(j \in \{|e|, |e| + 1, \ldots, |d|\}\), or \(\alpha^{(d,D)}_j < 1/(1 + w)\) holds for all \(j \in \{|e|, |e| + 1, \ldots, |d|\}\), then the rule list \(\bar{e} = \{e, \tilde{\alpha}_{e,D}\}\) (i.e. the rule list in which the final else clause follows immediately the prefix \(e\), and the probability estimate of the final else clause is \(\tilde{\alpha}_{e,D}\)) is compatible with \(D\) and satisfies \(\bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) \leq \bar{L}(d, D, 1/(1 + w), w, C).\)
Proof. Case 1. There exists some \( k \in \{ |e|, \ldots, |d| \} \) that satisfies \( \alpha_k^{(d,D)} > 1/(1+w) \) and some \( k' \in \{ |e|, \ldots, |d| \} \) that satisfies \( \alpha_{k'}^{(d,D)} \leq 1/(1+w) \). For any \( j \in \{ |e|, \ldots, |d| \} \) with \( \alpha_j^{(d,D)} > 1/(1+w) \), the contribution \( R_j(d,D,1/(1+w),w) \) by the \( j \)-th rule to \( R(d,D,1/(1+w),w) \), defined by the right-hand side of Equation (13) with \( \tau = 1/(1+w) \), is given by

\[
R_j(d,D,1/(1+w),w) = \frac{1}{n} n_{j,d,D}^-.
\]

For any \( j \in \{ |e|, \ldots, |d| \} \) with \( \alpha_j^{(d,D)} \leq 1/(1+w) \), the contribution \( R_j(d,D,1/(1+w),w) \) by the \( j \)-th rule to \( R(d,D,1/(1+w),w) \) is given by

\[
R_j(d,D,1/(1+w),w) = \frac{w}{n} n_{j,d,D}^+.
\]

The rest of the proof for this case proceeds in four steps.

Step 1. Construct a hypothetical rule list \( d' \) that begins with \( e \), has exactly one more rule (excluding the final else clause) following \( e \), and is compatible with \( D \). In later steps, we shall show that the rule list \( d' \) constructed in this step satisfies (15).

Let \( d' = \{ e, (a'_{|e|}, \alpha'_{|e|}), (a'_{|e|+1}, \alpha'_{|e|+1}) \} \) be the hypothetical rule list of size \( |d'| = |e| + 1 \) that is compatible with \( D \), and whose \( |e| \)-th antecedent \( a'_{|e|} \) is defined by

\[
a_{|e|}^{(d')} (x) = \mathbb{1} [\alpha^{(d,D)}_{\text{capt}(x,d)} > 1/(1+w)] \cdot \mathbb{1} [|e| \leq \text{capt}(x,d) \leq |d|].
\]

Step 2. Show that the empirical risk of misclassification by the rule list \( d' \) is the same as that by the rule list \( d \).

To see this, we observe that the training instances in \( D \) captured by \( a'_{|e|} \) in \( d' \) are exactly those captured by the antecedents \( a_j^{(d)}, |e| \leq j \leq |d| \), in \( d \) whose empirical positive proportion satisfies \( \alpha_j^{(d,D)} > 1/(1+w) \), and the training instances in \( D \) captured by \( a_{|e|+1}^{(d')} \) (i.e. the final else clause) in \( d' \) are exactly those captured by the antecedents \( a_j^{(d)}, |e| \leq j \leq |d| \), in \( d \) whose empirical positive proportion satisfies \( \alpha_j^{(d,D)} \leq 1/(1+w) \). This observation implies

\[
n_{|e|,d',D}^+ = \sum_{j: |e| \leq j \leq |d|, \alpha_j^{(d,D)} > 1/(1+w)} n_{j,d,D}^+ \tag{46}
\]

\[
n_{|e|,d',D}^- = \sum_{j: |e| \leq j \leq |d|, \alpha_j^{(d,D)} > 1/(1+w)} n_{j,d,D}^- \tag{47}
\]

\[
n_{|e|,d',D} = \sum_{j: |e| \leq j \leq |d|, \alpha_j^{(d,D)} > 1/(1+w)} n_{j,d,D} \tag{48}
\]

\[
n_{|e|+1,d',D}^+ = \sum_{j: |e| \leq j \leq |d|, \alpha_j^{(d,D)} \leq 1/(1+w)} n_{j,d,D}^+ \tag{49}
\]

and

\[
n_{|e|+1,d',D} = \sum_{j: |e| \leq j \leq |d|, \alpha_j^{(d,D)} \leq 1/(1+w)} n_{j,d,D}. \tag{50}
\]

Since \( d' \) is compatible with \( D \), using the definition of a compatible rule list in Definition 2.6 and the definition
of the empirical positive proportion in Definition 2.5, together with (46), (48), (49), and (50), we must have

\[ \hat{\alpha}_{|e|}^{(d')} = \alpha_{|e|}^{(d', D)} = \frac{n^+_{|e|, d', D}}{n_{|e|, d', D}} = \frac{\sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} > 1/(1+w) n^+_{j, d, D}}{\sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} > 1/(1+w) n_{j, d, D}} \]

and

\[ \hat{\alpha}_{|e|+1}^{(d')} = \alpha_{|e|+1}^{(d', D)} = \frac{n^+_{|e|+1, d', D}}{n_{|e|+1, d', D}} = \frac{\sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} \leq 1/(1+w) n^+_{j, d, D}}{\sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} \leq 1/(1+w) n_{j, d, D}} \leq \frac{1}{1+w}. \]

This means that the contribution \( R_{|e|}(d', D, 1/(1+w), w) \) by the \(|e|\)-th rule to \( R(d', D, 1/(1+w), w) \) is given by

\[ R_{|e|}(d', D, 1/(1+w), w) = \frac{1}{n} n^-_{|e|, d', D} = \frac{1}{n} \sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} > 1/(1+w) n^-_{j, d, D}, \]

where we have used (47), and the contribution \( R_{|e|+1}(d', D, 1/(1+w), w) \) by the \((|e|+1)\)-st “rule” (i.e. the final else clause) to \( R(d', D, 1/(1+w), w) \) is given by

\[ R_{|e|+1}(d', D, 1/(1+w), w) = \frac{w}{n} n^+_{|e|+1, d', D} = \frac{w}{n} \sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} \leq 1/(1+w) n^+_{j, d, D}, \]

where we have used (49).

It then follows that the empirical risk of misclassification by the rule list \( d' \) is the same as by the rule list \( d \):

\[
\begin{align*}
R(d', D, 1/(1+w), w) &= R(e, D, 1/(1+w), w) + R_{|e|}(d', D, 1/(1+w), w) + R_{|e|+1}(d', D, 1/(1+w), w) \\
&= R(e, D, 1/(1+w), w) \\
&\quad + \frac{1}{n} \sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} > 1/(1+w) n^-_{j, d, D} + \frac{w}{n} \sum j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} \leq 1/(1+w) n^+_{j, d, D} \\
&= R(e, D, 1/(1+w), w) + \sum_{j=|e|}^{|d|} R_j(d, D, 1/(1+w), w) \\
&= R(d, D, 1/(1+w), w).
\end{align*}
\]

(51)

**Step 3.** Show that the monotonicity penalty of the rule list \( d' \) is at most that of \( d \).

Let \( S(d, D) = \sum_{j=0}^{|d|} |\alpha_j^{(d, D)} - \min_{k<j} \alpha_k^{(d, D)}|_+ \) be the monotonicity penalty of the rule list \( d \). We now show \( S(d', D) \leq S(d, D) \). Let \( S_j(d, D) = |\alpha_j^{(d, D)} - \min_{k<j} \alpha_k^{(d, D)}|_+ \) be the monotonicity penalty for the \( j \)-th rule in \( d \).

Let \( l \in \{ |e|, \ldots, |d| \} \) be any integer with

\[
\alpha_l^{(d, D)} = \max_{j: |e| \leq j \leq |d| \land \alpha_j^{(d, D)} > 1/(1+w)} \alpha_j^{(d, D)}. \]

(52)
Then the total monotonicity penalty for all the rules \( (a_j^{(d)}, a_j^{(d,D)}) \) in \( d \) with \(|e| \leq j \leq |d| \) and \( a_j^{(d,D)} > 1/(1+w) \) satisfies

\[
S_j(d, D) \geq S_i(d, D) \quad \text{(because } S_i(d, D) \text{ is included in the sum on the left)}
\]

\[
= \left[ a_i^{(d,D)} - \min_{k < |e|} a_k^{(d,D)} \right]_+ \\
\geq \left[ a_i^{(d,D)} - \min_{k < |e|} a_k^{(d,D)} \right]_+.
\]

(53)

On the other hand, the monotonicity penalty for the \(|e|\)-th rule in \( d' \) satisfies

\[
S_{|e|}(d', D) = \left[ a_{|e|}^{(d', D)} - \min_{k < |e|} a_k^{(d', D)} \right]_+ \leq \left[ a_i^{(d,D)} - \min_{k < |e|} a_k^{(d,D)} \right]_+.
\]

(54)

because we have \( \min_{k < |e|} a_k^{(d', D)} = \min_{k < |e|} a_k^{(d,D)} \) (\( d \) and \( d' \) begin with the same prefix \( e \)), and

\[
\alpha_{|e|}^{(d', D)} = \frac{n_+^{(d', D)}}{n_0^{(d', D)}} \quad \text{(by the definition of the empirical positive proportion in Definition 2.5)}
\]

(55)

\[
= \frac{\sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} n_+^{(d,D)}}{\sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} n_0^{(d,D)}} \quad \text{(by Equations (46) and (48))}
\]

(56)

\[
= \frac{\sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} a_j^{(d,D)} n_{j,d,D}}{\sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} n_{j,d,D}} \quad \text{(by the definition of } a_j^{(d,D)} \text{ in Definition 2.5)}
\]

\[
\leq \frac{\sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} a_i^{(d,D)} n_{j,d,D}}{\sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} n_{j,d,D}} \quad \text{(by the definition of } l \text{ in (52))}
\]

\[
= a_i^{(d,D)}.
\]

Combining (53) and (54), we have

\[
S_{|e|}(d', D) \leq \sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} S_j(d, D).
\]

(55)

A similar argument will show

\[
S_{|e|+1}(d', D) \leq \sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} \leq 1/(1+w)} S_j(d, D).
\]

(56)

It then follows from (55) and (56) that the monotonicity penalty of \( d' \) is at most that of \( d \):

\[
S(d', D) = \left[ \sum_{j=0}^{(|e|-1)} S_j(d', D) \right] + S_{|e|}(d', D) + S_{|e|+1}(d', D)
\]

\[
\leq \left[ \sum_{j=0}^{(|e|-1)} S_j(d, D) \right] + \sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} > 1/(1+w)} S_j(d, D)
\]

\[
+ \sum_{j:|e| \leq j \leq |d| \land a_j^{(d,D)} \leq 1/(1+w)} S_j(d, D)
\]

(57)

\[
= S(d, D).
\]

(58)

Using \([51]\) and \([58]\), together with the observation \(|d'| = |e| + 1 \leq |d|\), we must also have

\[
\bar{L}(d', D, 1/(1 + w), w, C, C_1) = R(d', D, 1/(1 + w), w) + C|d'| + C_1 S(d', D) \\
\leq R(d, D, 1/(1 + w), w) + C|d| + C_1 S(d, D) \\
= \bar{L}(d, D, 1/(1 + w), w, C, C_1).
\]

Case 2. Either \(a_j^{(d, D)} > 1/(1 + w)\) holds for all \(j \in \{1, \ldots, |d|\}\), or \(a_j^{(d, D)} \leq 1/(1 + w)\) holds for all \(j \in \{1, \ldots, |d|\}\). The construction of \(d' = \bar{e}\) and the proof for \(R(d', D, 1/(1 + w), w) = R(d, D, 1/(1 + w), w)\) is similar to those given in the proof of Lemma 4.4. The proof for \(S(d', D) \leq S(d, D)\) is similar to that in Case 1. The desired inequality then follows from \(|d'| = |e| \leq |d|\).

Before we proceed with proving Theorem 5.2, we make the following four observations. Observations 11.1, 11.2, and 11.3 are the same as Observations 10.1, 10.2 and 10.3. They are repeated here for convenience.

Observation 11.1. For any rule list

\[
d' = \{e, (a_{|e|}^{(d')}, \hat{\alpha}_{|e|}^{(d')}), \ldots, (a_{|d'|-1}^{(d')}, \hat{\alpha}_{|d'|-1}^{(d')}), \hat{\alpha}_{|d'|}^{(d')}\}
\]

that begins with a given prefix \(e\), we have

\[
\hat{n}_{e, D}^+ = n_{|e|, d', D}^+ + \ldots + n_{|d'|-1, d', D}^+, \tag{59}
\]

\[
\hat{n}_{e, D}^- = n_{|e|, d', D}^- + \ldots + n_{|d'|-1, d', D}^- \tag{60}
\]

and

\[
\hat{n}_{e, D} = n_{|e|, d', D} + \ldots + n_{|d'|-1, d', D}. \tag{61}
\]

Proof. Same as Observation 10.1.

Observation 11.2. For any rule list \(d'\), we have

\[
n_{|e|, d', D}^+ = \left(1 - \frac{1}{\alpha_{|e|}^{(d')}}\right) n_{|e|, d', D}^+. \tag{62}
\]

Proof. Same as Observation 10.2.

Observation 11.3. For any rule list

\[
d' = \{e, (a_{|e|}^{(d')}, \hat{\alpha}_{|e|}^{(d')}, \hat{\alpha}_{|e|+1}^{(d')}\}
\]

that has exactly one rule (excluding the final else clause) following a given prefix \(e\), we have

\[
n_{|e|+1, d', D}^+ = \hat{n}_{e, D}^+ - n_{|e|, d', D}^+ \tag{63}
\]

\[
n_{|e|+1, d', D}^- = \hat{n}_{e, D}^- - n_{|e|, d', D}^- \tag{64}
\]

and

\[
n_{|e|+1, d', D} = \hat{n}_{e, D} - n_{|e|, d', D}. \tag{65}
\]

Note that since \(n_{|e|+1, d', D}^+, n_{|e|+1, d', D}^-, n_{|e|+1, d', D}\) are non-negative, Equations \([63]\), \([64]\), and \([65]\) imply \(n_{|e|, d', D} \leq \hat{n}_{e, D}^+, n_{|e|, d', D} \leq \hat{n}_{e, D}^-, n_{|e|, d', D} \leq \hat{n}_{e, D}^+\).
We are now ready to prove Theorem 5.2.

Applying Equation (62) in Observation 11.2, we have

Applying Equations (63) and (64) in Observation 11.3, we have

The lemma that we have proved in this section, along with its proof, implies

\[ \hat{L}(d, D, 1/(1 + w), w, C, C_1) \geq \inf_{d' \in S(X, D, e, 1) \cup D(X, D, e, 0)} \hat{L}(d', D, 1/(1 + w), w, C, C_1). \] (67)
This is because if \( d \) obeys Case 1 in the proof of the lemma, then using the same argument as in the proof of the lemma we can construct a rule list \( d_1 = \{ e, (\alpha^{(d_1,D)}_{|e|}, \alpha^{(d_1,D)}_{|e|+1}) \} \in S(\mathcal{X}, D, e, 1) \) that satisfies
\[
\bar{L}(d, D, 1/(1 + w), w, C, C_1) \geq \bar{L}(d_1, D, 1/(1 + w), w, C, C_1).
\] (68)

Since \( d_1 \) must also obey
\[
\bar{L}(d_1, D, 1/(1 + w), w, C, C_1) \geq \inf_{d' \in S(\mathcal{X}, D, e, 1)} \bar{L}(d', D, 1/(1 + w), w, C, C_1)
\[
\geq \inf_{d' \in S(\mathcal{X}, D, e, 1) \cup D(X, D, e, 0)} \bar{L}(d', D, 1/(1 + w), w, C, C_1),
\]
combining the inequalities in (68) and (69) gives us (67). On the other hand, if \( d \) obeys Case 2 in the proof of the lemma, then by the lemma itself we know
\[
\bar{L}(d, D, 1/(1 + w), w, C, C_1) \geq \bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1).
\] (70)

Since we have \( D(X, D, e, 0) = \{ \bar{e} \} \), it is straightforward to see
\[
\bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) = \inf_{d' \in D(X, D, e, 0)} \bar{L}(d', D, 1/(1 + w), w, C, C_1)
\[
\geq \inf_{d' \in S(\mathcal{X}, D, e, 1) \cup D(X, D, e, 0)} \bar{L}(d', D, 1/(1 + w), w, C, C_1).
\] (71)

Combining the inequalities in (70) and (71) again gives us (67).

Note that if \( S(X, D, e, 1) \) is not empty, then the right-hand side of (67) can be expressed as
\[
\inf_{d' \in S(\mathcal{X}, D, e, 1) \cup D(X, D, e, 0)} \bar{L}(d', D, 1/(1 + w), w, C, C_1)
\[
= \min \left( \inf_{d' \in S(\mathcal{X}, D, e, 1) \cup \{ \bar{e} \}} \bar{L}(d', D, 1/(1 + w), w, C, C_1), \bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) \right).
\] (72)

The rest of the proof proceeds in six steps.

**Step 1.** Compute \( \bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) \).

Since the contribution by the final else clause to \( \bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) \) is given by \( R_{|e|}(\bar{e}, D, 1/(1 + w), w) + [\alpha_{|e|, D} - \alpha_{\min}^{(e,D)}]_+ \), where \( R_{|e|}(\bar{e}, D, 1/(1 + w), w) \) is defined by Equation (13) and is given by
\[
R_{|e|}(\bar{e}, D, 1/(1 + w), w) = \begin{cases} 
\frac{1}{n} n_{|e|, D} & \text{if } \alpha_{|e|, D} > 1/(1 + w) \\
\frac{w}{n} n_{|e|, D} & \text{otherwise},
\end{cases}
\]
and since Observation 11.1 implies \( \bar{n}_{e,D}^+ = n_{|e|, e,D}^+ \) and \( \bar{n}_{e,D}^- = n_{|e|, e,D}^- \), it is not difficult to see
\[
\bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1)
\[
= \begin{cases} 
\bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \bar{n}_{e,D}^- + C_1 [\alpha_{e,D} - \alpha_{\min}^{(e,D)}]_+ & \text{if } \alpha_{e,D} > 1/(1 + w) \\
\bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) + \frac{w}{n} \bar{n}_{e,D}^+ + C_1 [\alpha_{e,D} - \alpha_{\min}^{(e,D)}]_+ & \text{otherwise},
\end{cases}
\]

Since \( \alpha_{e,D} > 1/(1 + w) \) is equivalent to \( \bar{n}_{e,D}^+ / (\bar{n}_{e,D}^+ + \bar{n}_{e,D}^-) > 1/(1 + w) \), or \( w \bar{n}_{e,D}^+ > \bar{n}_{e,D}^- \), and similarly \( \alpha_{e,D} \leq 1/(1 + w) \) is equivalent to \( w \bar{n}_{e,D}^+ \leq \bar{n}_{e,D}^- \), we can write
\[
\bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1)
\[
= \min \left( \bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \bar{n}_{e,D}^- + C_1 [\alpha_{e,D} - \alpha_{\min}^{(e,D)}]_+, \right.
\]
\[
\bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) + \frac{w}{n} \bar{n}_{e,D}^+ + C_1 [\alpha_{e,D} - \alpha_{\min}^{(e,D)}]_+
\]
\[
= \bar{L}(\bar{e}, D, 1/(1 + w), w, C, C_1) + \min \left( \frac{w}{n} \bar{n}_{e,D}^+, \frac{1}{n} \bar{n}_{e,D}^- \right) + C_1 [\alpha_{e,D} - \alpha_{\min}^{(e,D)}]_+.
\] (73)
Step 2. Partition the set $S(\mathcal{X}, D, e, 1)$ into three subsets based on how the softly falling objective is computed.

For any $d' = \{e, (a_{|e|}^{(d')}, a_{|e|+1}^{(d')}), \alpha_{|e|+1}^{(d', D)}\} \in S(\mathcal{X}, D, e, 1)$, the softly falling objective is given by

$$
\tilde{L}(d', D, 1/(1 + w), w, C, C_1)
= L(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} n_{|e|, d', D}^- + \frac{w}{n} n_{|e|+1, d', D}^+ + C
+ C_1[\alpha_{|e|}^{(d', D)} - \alpha_{\min}^{(e, D)}]_+ + C_1[\alpha_{|e|+1}^{(d', D)} - \alpha_{\min}^{(e, D)}]_+.
$$

(74)

This is because for any $d' \in S(\mathcal{X}, D, e, 1)$, the contribution by both the $|e|$-th rule and the final else clause to $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$ is given by

$$
R_{e|e|}(d', D, 1/(1 + w), w) + R_{e|e|+1}(d', D, 1/(1 + w), w) + C + C_1[\alpha_{|e|}^{(d', D)} - \alpha_{\min}^{(e, D)}]_+ + C_1[\alpha_{|e|+1}^{(d', D)} - \alpha_{\min}^{(e, D)}]_+,
$$

where $R_{e|e|}(d', D, 1/(1 + w), w)$ and $R_{e|e|+1}(d', D, 1/(1 + w), w)$ are defined by Equation (73) and are given by

$$
R_{e|e|}(d', D, 1/(1 + w), w) = \frac{1}{n} n_{|e|, d', D}^- \quad \text{and} \quad R_{e|e|+1}(d', D, 1/(1 + w), w) = \frac{w}{n} n_{|e|+1, d', D}^+.
$$

(because we have $\alpha_{|e|}^{(d', D)} > 1/(1 + w)$ and $\alpha_{|e|+1}^{(d', D)} \leq 1/(1 + w)$ for $d' \in S(\mathcal{X}, D, e, 1)$).

Let

$$
S_1(\mathcal{X}, D, e, 1) = \{d' = \{e, (a_{|e|}^{(d')}, a_{|e|+1}^{(d')}), \alpha_{|e|+1}^{(d', D)}\} \in S(\mathcal{X}, D, e, 1) : \alpha_{\min}^{(e, D)} \geq \alpha_{|e|}^{(d', D)} > \alpha_{|e|+1}^{(d', D)}\},
$$

$$
S_2(\mathcal{X}, D, e, 1) = \{d' = \{e, (a_{|e|}^{(d')}, a_{|e|+1}^{(d')}), \alpha_{|e|+1}^{(d', D)}\} \in S(\mathcal{X}, D, e, 1) : \alpha_{|e|}^{(d', D)} > \alpha_{\min}^{(e, D)} \geq \alpha_{|e|+1}^{(d', D)}\},
$$

and

$$
S_3(\mathcal{X}, D, e, 1) = \{d' = \{e, (a_{|e|}^{(d')}, a_{|e|+1}^{(d')}), \alpha_{|e|+1}^{(d', D)}\} \in S(\mathcal{X}, D, e, 1) : \alpha_{|e|}^{(d', D)} > \alpha_{|e|+1}^{(d', D)} > \alpha_{\min}^{(e, D)}\}.
$$

It is easy to see

$$
S(\mathcal{X}, D, e, 1) = S_3(\mathcal{X}, D, e, 1) \cup S_1(\mathcal{X}, D, e, 1) \cup S_2(\mathcal{X}, D, e, 1).
$$

We observe here that given the prefix $e$, we can write $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$ as a function of $n_{|e|, d', D}^-$ and $\alpha_{|e|}^{(d', D)}$, by substituting (62), (63), and (66) in Observations 11.1, 11.2, and 11.3 into (74).

Step 3. Determine a lower bound of $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$ for all $d' \in S_1(\mathcal{X}, D, e, 1)$.

Let $d' = \{e, (a_{|e|}^{(d')}, a_{|e|+1}^{(d')}), \alpha_{|e|+1}^{(d', D)}\} \in S_1(\mathcal{X}, D, e, 1)$. By the definition of $S_1(\mathcal{X}, D, e, 1)$, we have

$$
\alpha_{\min}^{(e, D)} \geq \alpha_{|e|}^{(d', D)} > \frac{1}{1 + w} \geq \alpha_{|e|+1}^{(d', D)}.
$$

(75)

We first prove the following inequality

$$
\alpha_{\min}^{(e, D)} \geq \alpha_{|e|}^{(d', D)} \geq \max(1/(1 + w), \tilde{\alpha}_{|e|, D}),
$$

(76)

which will be useful later.

To prove (76), we use Definition 2.5 as well as (63) and (65) in Observation 11.3 to obtain

$$
\tilde{\alpha}_{|e|, D} = \frac{n_{|e|, D}^+}{n_{|e|, D}^-} = \frac{n_{|e|, d', D}^- + n_{|e|+1, d', D}^+}{n_{|e|, d', D}^-} = \frac{\alpha_{|e|}^{(d', D)} n_{|e|, d', D}^- + \alpha_{|e|+1}^{(d', D)} n_{|e|+1, d', D}^+}{n_{|e|, d', D}^- + n_{|e|+1, d', D}^+}.
$$

(77)

Substituting $\alpha_{|e|+1}^{(d', D)} < \alpha_{|e|}^{(d', D)}$ from (75) into (77), we obtain $\tilde{\alpha}_{|e|, D} < \alpha_{|e|}^{(d', D)}$. Combining this inequality with $\alpha_{\min}^{(e, D)} \geq \alpha_{|e|}^{(d', D)} > \frac{1}{1 + w}$ from (75), we obtain (76), as desired.
We now show that given the prefix \( e \), if \( \alpha_{\min}^{(e,D)} \leq \max(1/(1 + w), \bar{\alpha}_{e,D}) \) is true for the given prefix \( e \), then \( \mathcal{S}_1(\mathcal{X}, D, e, 1) \) is empty.

We first prove the following inequality

To do so, we substitute (62) and (63) in Observations 11.1 and 11.2 into (74) to obtain

\[
\tilde{L}(d', D, 1/(1 + w), w, C, C_1) = \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \left( \frac{1}{\alpha_{e,D}} - 1 - w \right) n_{e,D}^+ + w \tilde{n}_{e,D}^+ + C.
\] (78)

Note that Equation (78) shows that given the prefix \( e \), \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) is a function of \( n_{e,D}^+ + \alpha_{e,D}^{(d',D)} \). Since we have

\[
\frac{\partial \tilde{L}(d', D, 1/(1 + w), w, C, C_1)}{\partial n_{e,D}^+} = \frac{1}{n} \left( \frac{1}{\alpha_{e,D}} - 1 - w \right) < 0
\]

and

\[
\frac{\partial \tilde{L}(d', D, 1/(1 + w), w, C, C_1)}{\partial \alpha_{e,D}^{(d',D)}} = \frac{n_{e,D}^+}{n (\alpha_{e,D}^{(d',D)})^2} \leq 0,
\]

we see that \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) is indeed a monotonically decreasing function of both \( n_{e,D}^+ + \alpha_{e,D}^{(d',D)} \). Thus, we can obtain a lower bound of \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) by substituting \( n_{e,D}^+ + \alpha_{e,D}^{(d',D)} \) with their respective upper bound. The inequality \( n_{e,D}^+ \leq \tilde{n}_{e,D}^+ \) in Observation 11.3 gives an upper bound for \( n_{e,D}^+ \) and the inequality \( \alpha_{e,D}^{(d',D)} \leq \alpha_{\min}^{(e,D)} \) from (76) gives an upper bound for \( \alpha_{e,D}^{(d',D)} \). Substituting these upper bounds into (78), we obtain the following inequality, which gives a lower bound of \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \):

\[
\tilde{L}(d', D, 1/(1 + w), w, C, C_1) \\
\geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \left( \frac{1}{\alpha_{\min}^{(e,D)}} - 1 - w \right) \tilde{n}_{e,D}^+ + w \tilde{n}_{e,D}^+ + C \\
= \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \left( \frac{1}{\alpha_{\min}^{(e,D)}} - 1 \right) \tilde{n}_{e,D}^+ + C.
\]

**Step 4.** Determine a lower bound of \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) for all \( d' \in \mathcal{S}_2(\mathcal{X}, D, e, 1) \).

Let \( d' = \{ e, (\alpha_{e,D}^{(d')}, \alpha_{e,D}^{(d',D)}), \alpha_{e,D}^{(d',D)}_{e,D+1} \} \in \mathcal{S}_2(\mathcal{X}, D, e, 1) \). By the definition of \( \mathcal{S}_2(\mathcal{X}, D, e, 1) \), we have

\[
\alpha_{e,D}^{(d',D)} > \alpha_{e,D}^{(d',D)} + \frac{1}{1 + w}
\]

and

\[
\alpha_{e,D}^{(d',D)} > \alpha_{e,D}^{(d',D)} \geq \alpha_{e,D}^{(d',D)}_{e,D+1}
\]

We first prove the following inequality

\[
1 \geq \alpha_{e,D}^{(d',D)} > \max(\alpha_{e,D}^{(d',D)}, \bar{\alpha}_{e,D}, 1/(1 + w)) = \zeta,
\]

(81)
which will be useful later.

To prove (81), we use Definition 2.5 as well as (63) and (65) in Observation 11.3 to obtain (77). Substituting \( \alpha_{[e]}^{(d',D)} < \alpha_{[e]}^{(d',D)} \) from (75) into (77), we obtain \( \tilde{\alpha}_{e,D} < \alpha_{[e]}^{(d',D)} \). Combining this inequality with (79) and \( \alpha_{[e]}^{(d',D)} > \alpha_{\text{min}}(\varepsilon, D) \) from (80), we obtain (81), as desired.

We now show that given the prefix \( e \) and a particular value of \( \alpha_{[e]}^{(d',D)} \) that obeys (81), the softly falling objective \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) for \( d' \) is a decreasing function of \( n_{[e],d',D}^+ \).

To do so, we substitute (62) and (63) in Observations 11.1 and 11.2 into (74) to obtain

\[
\tilde{L}(d', D, 1/(1 + w), w, C, C_1) = \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \left( \frac{1}{\alpha_{[e]}^{(d',D)}} - 1 - w \right) n_{[e],d',D}^+ + \frac{1}{\alpha_{[e]}^{(d',D)}} - \alpha_{\text{min}}(\varepsilon, D) + C
\]

(82)

Note that Equation (82) shows that given the prefix \( e \), \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) is a function of \( n_{[e],d',D}^+ \) and \( \alpha_{[e]}^{(d',D)} \). Differentiating \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) given in (82) with respect to \( n_{[e],d',D}^+ \), we obtain

\[
\frac{\partial \tilde{L}(d', D, 1/(1 + w), w, C, C_1)}{\partial n_{[e],d',D}^+} = \frac{1}{n} \left( \frac{1}{\alpha_{[e]}^{(d',D)}} - 1 - w \right).
\]

(83)

Since \( \alpha_{[e]}^{(d',D)} \) obeys (81), in particular, it obeys \( \alpha_{[e]}^{(d',D)} > 1/(1 + w) \), we have

\[
\frac{1}{\alpha_{[e]}^{(d',D)}} - 1 - w < 0,
\]

which then gives \( \partial \tilde{L}(d', D, 1/(1 + w), w, C, C_1) / \partial n_{[e],d',D}^+ < 0 \). This means that given the prefix \( e \) and a particular value of \( \alpha_{[e]}^{(d',D)} \) that obeys (81), \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) is a decreasing function of \( n_{[e],d',D}^+ \).

Thus, given the prefix \( e \) and a particular value of \( \alpha_{[e]}^{(d',D)} \) that obeys (81), we can obtain a lower bound of \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \) by substituting \( n_{[e],d',D}^+ \) with its upper bound. The inequality \( n_{[e],d',D}^+ \leq \tilde{n}_{e,D}^+ \) in Observation 11.3 gives an upper bound for \( n_{[e],d',D}^+ \). Substituting \( n_{[e],d',D}^+ \) with its upper bound \( \tilde{n}_{e,D}^+ \) into (82), we obtain a lower bound of \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \), denoted by \( \tilde{g}(\alpha_{[e]}^{(d',D)}) \), when \( \alpha_{[e]}^{(d',D)} \) is held constant:

\[
\tilde{g}(\alpha_{[e]}^{(d',D)}) = \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \left( \frac{1}{\alpha_{[e]}^{(d',D)}} - 1 \right) \tilde{n}_{e,D}^+ + C + \alpha_{\text{min}}(\varepsilon, D) - \alpha_{[e]}^{(d',D)}
\]

where \( g \) is defined in the statement of the theorem. In other words, given the prefix \( e \) and a particular value of \( \alpha_{[e]}^{(d',D)} \) that obeys (81), we have \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \geq \tilde{g}(\alpha_{[e]}^{(d',D)}) \). Since (81) is true for any \( d' \in S_2(X, D, e, 1) \), we always have \( \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \geq \tilde{g}(\alpha_{[e]}^{(d',D)}) \) for any \( d' \in S_2(X, D, e, 1) \). This implies

\[
\tilde{L}(d', D, 1/(1 + w), w, C, C_1) \geq \inf_{\alpha_{[e]}^{(d',D)} \leq \alpha_{[e]}^{(d',D)} \leq 1} \tilde{g}(\alpha_{[e]}^{(d',D)})
\]

\[
= \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \inf_{\alpha_{[e]}^{(d',D)} \leq \alpha_{[e]}^{(d',D)} \leq 1} g(\alpha_{[e]}^{(d',D)}).
\]
Step 5. Determine a lower bound of $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$ for all $d' \in S_3(\mathcal{X}, D, e, 1)$.

Let $d' = \{ e, (\alpha_{[e]}^{(d')}, \alpha_{[e]}^{(d'), D}), \alpha_{[e] + 1}^{(d'), D} \} \in S_3(\mathcal{X}, D, e, 1)$. By the definition of $S_3(\mathcal{X}, D, e, 1)$, we have

$$\alpha_{[e]}^{(d', D)} > \frac{1}{1 + w} \geq \alpha_{[e] + 1}^{(d', D)} > \alpha_{\min}^{(e, D)}.$$  \hfill (84)

We first prove the following inequality

$$1 \geq \alpha_{[e]}^{(d', D)} > \max(\alpha_{\min}^{(e, D)}, \tilde{\alpha}_{e, D}, 1/(1 + w)) = \zeta,$$  \hfill (85)

which will be useful later.

To prove (85), we use Definition 2.5 as well as (63) and (65) in Observation 11.3 to obtain (77). Substituting $\alpha_{[e] + 1}^{(d', D)} < \alpha_{[e]}^{(d', D)}$ from (84) into (77), we obtain $\tilde{\alpha}_{e, D} < \alpha_{[e]}^{(d', D)}$. Combining this inequality with $\alpha_{[e]}^{(d', D)} > \frac{1}{1 + w} > \alpha_{\min}^{(e, D)}$ from (84), we obtain (85), as desired.

To determine a lower bound of $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$, we observe

$$\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$$

$$\geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} n_{[e], d', D}^+ + \frac{w}{n} n_{[e] + 1, d', D}^+ + C + C_1 [\alpha_{[e]}^{(d', D)} - \alpha_{\min}^{(e, D)}] +$$

$$= \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \left( \frac{1}{\alpha_{[e]}^{(d', D)}} - 1 - w \right) n_{[e], d', D}^+ + w\tilde{n}_{e, D}^+ + C +$$

$$+ C_1 (\alpha_{[e]}^{(d', D)} - \alpha_{\min}^{(e, D)})$$  \hfill (87)

where the last equality follows by substituting (62) and (63) in Observations 11.1 and 11.2 into (86). Using (85) and applying the same argument as in Step 4, the quantity labeled (87) is also lower-bounded by

$$\tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \inf_{\alpha^{(d', D)}_{[e]}, \zeta < \alpha^{(d', D)}_{[e]} \leq 1} g(\alpha^{(d', D)}_{[e]}),$$

so that we again have

$$\tilde{L}(d', D, 1/(1 + w), w, C, C_1) \geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \inf_{\alpha^{(d', D)}_{[e]}, \zeta < \alpha^{(d', D)}_{[e]} \leq 1} g(\alpha^{(d', D)}_{[e]}).$$


Suppose, first, that $S(\mathcal{X}, D, e, 1)$ is not empty.

In the case where $S_1(\mathcal{X}, D, e, 1)$ is not empty, we observe the following inequality

$$\inf_{d' \in S_1(\mathcal{X}, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \frac{1}{n} \left( \frac{1}{\alpha_{\min}^{(e, D)}} - 1 \right) \tilde{n}_{e, D}^+ + C,$$  \hfill (88)

which follows from the definition of inf being the greatest lower bound, as well as the lower bound of $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$ for $d' \in S_1(\mathcal{X}, D, e, 1)$, which we have derived in Step 3.

In the case where $S_2(\mathcal{X}, D, e, 1) \cup S_3(\mathcal{X}, D, e, 1)$ is not empty, we observe the following inequality

$$\inf_{d' \in S_2(\mathcal{X}, D, e, 1) \cup S_3(\mathcal{X}, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \inf_{\beta, \zeta < \beta \leq 1} g(\beta),$$  \hfill (89)

which follows from the definition of inf being the greatest lower bound, as well as the lower bound of $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$ for $d' \in S_2(\mathcal{X}, D, e, 1)$, which we have derived in Step 4, and the lower bound of $\tilde{L}(d', D, 1/(1 + w), w, C, C_1)$ for $d' \in S_3(\mathcal{X}, D, e, 1)$, which we have derived in Step 5.
To derive a lower bound of \(\tilde{L}(d', D, 1/(1 + w), w, C, C_1)\) for \(d' \in S(X, D, e, 1)\), we further observe that if \(\alpha_{\min}^{(e,D)} \leq \max(1/(1 + w), \tilde{\alpha}_{\infty,D})\) holds, then by our remark in Step 3, \(S_1(X, D, e, 1)\) is empty, and consequently, using (89), we have

\[
\inf_{d' \in S(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) = \inf_{d' \in S_2(X, D, e, 1) \cup S_3(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \\
\geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \inf_{\beta, \xi < \beta \leq 1} g(\beta).
\] (90)

On the other hand, if \(\alpha_{\min}^{(e,D)} > \max(1/(1 + w), \tilde{\alpha}_{\infty,D})\) holds, then \(S_1(X, D, e, 1)\) may or may not be empty. If, in addition, both \(S_1(X, D, e, 1)\) and \(S_2(X, D, e, 1) \cup S_3(X, D, e, 1)\) are not empty, then using (88) and (89), we have

\[
\inf_{d' \in S(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \\
= \min \left( \inf_{d' \in S_1(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1), \inf_{d' \in S_2(X, D, e, 1) \cup S_3(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \right) \\
\geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) + \min \left( \frac{1}{n} \left( \frac{1}{\alpha_{\min}^{(e,D)}} - 1 \right) \tilde{n}_{e,D}^+ + C, \inf_{\beta, \xi < \beta \leq 1} g(\beta) \right).
\] (91)

If either \(S_1(X, D, e, 1)\) or \(S_2(X, D, e, 1) \cup S_3(X, D, e, 1)\) is empty, then \(\inf_{d' \in S(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1)\) is given by either

\[
\inf_{d' \in S_2(X, D, e, 1) \cup S_3(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \quad \text{or} \quad \inf_{d' \in S_1(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1),
\]

both of which are lower-bounded by the quantity labeled (91) because of (89) and (88).

Putting these cases together, we have

\[
\inf_{d' \in S(X, D, e, 1)} \tilde{L}(d', D, 1/(1 + w), w, C, C_1) \\
\geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) \\
+ \left\{ \begin{array}{ll}
\min \left( \frac{1}{n} \left( \frac{1}{\alpha_{\min}^{(e,D)}} - 1 \right) \tilde{n}_{e,D}^+ + C, \inf_{\beta, \xi < \beta \leq 1} g(\beta) \right) & \text{if } \alpha_{\min}^{(e,D)} > \max(1/(1 + w), \tilde{\alpha}_{\infty,D}), \\
\inf_{\beta, \xi < \beta \leq 1} g(\beta) & \text{otherwise}.
\end{array} \right.
\] (92)

Combining (67), (72), (73), and (92), we have

\[
\tilde{L}(d, D, 1/(1 + w), w, C, C_1) \\
\geq \tilde{L}(e, D, 1/(1 + w), w, C, C_1) \\
+ \left\{ \begin{array}{ll}
\min \left( \frac{1}{n} \left( \frac{1}{\alpha_{\min}^{(e,D)}} - 1 \right) \tilde{n}_{e,D}^+ + C, \inf_{\beta, \xi < \beta \leq 1} g(\beta), \frac{w}{n} \tilde{n}_{e,D}^+ + C_1 [\tilde{\alpha}_{e,D} - \alpha_{\min}^{(e,D)}]_+ \right) & \text{if } \alpha_{\min}^{(e,D)} > \max(1/(1 + w), \tilde{\alpha}_{\infty,D}), \\
\frac{1}{n} \tilde{n}_{e,D}^- + C_1 [\tilde{\alpha}_{e,D} - \alpha_{\min}^{(e,D)}]_+ & \text{otherwise}.
\end{array} \right.
\] (93)

Note that the quantity labeled (93) is precisely equal to \(\tilde{L}^*(e, D, w, C, C_1)\) given by Equation (44) in the statement of the theorem, because:

(i) if \(\alpha_{\min}^{(e,D)} > \max(1/(1 + w), \tilde{\alpha}_{\infty,D})\) holds, then the first term in the minimum on the right-hand side of Equation (44) is precisely \(\frac{1}{n} \left( \frac{1}{\alpha_{\min}^{(e,D)}} - 1 \right) \tilde{n}_{e,D}^+ + C\);
(ii) if \( \alpha_{\text{min}}^{(e,D)} > \max(1/(1+w), \tilde{\alpha}_{e,D}) \) does not hold, then we have \( \alpha_{\text{min}}^{(e,D)} \leq 1/(1+w) \) or \( \alpha_{\text{min}}^{(e,D)} \leq \tilde{\alpha}_{e,D} \) - in the former case where \( \alpha_{\text{min}}^{(e,D)} \leq 1/(1+w) \) holds, we have

\[
\frac{1}{n} \left( \frac{1}{\alpha_{\text{min}}} - 1 \right) \tilde{n}_{e,D}^+ \geq \frac{w}{n} \tilde{n}_{e,D}^-
\]

which implies that the first term in the minimum on the right-hand side of Equation (44) is bounded below by \( \frac{w}{n} \tilde{n}_{e,D}^+ + C_1 [\tilde{\alpha}_{e,D} - \alpha_{\text{min}}^{(e,D)}]_+ \), and thus has no influence over the computation of the minimum; in the latter case where \( \alpha_{\text{min}}^{(e,D)} \leq \tilde{\alpha}_{e,D} \) holds, the first term in the minimum on the right-hand side of Equation (44) is clearly bounded below by \( \frac{w}{n} \tilde{n}_{e,D}^+ + C_1 [\tilde{\alpha}_{e,D} - \alpha_{\text{min}}^{(e,D)}]_+ \), and again has no influence over the computation of the minimum.

This proves that \( \tilde{L}^*(e, D, w, C, C_1) \) given by Equation (44) is indeed a lower bound of \( \tilde{L}(d, D, 1/(1+w), w, C, C_1) \) for \( d \in \mathcal{D}(\mathcal{X}, D, e) \), in the case where \( \mathcal{S}(\mathcal{X}, D, e, 1) \) is not empty. In the case where \( \mathcal{S}(\mathcal{X}, D, e, 1) \) is empty, using (67) and (73), along with the fact \( \mathcal{D}(\mathcal{X}, D, e, 0) = \{\tilde{e}\} \), we have

\[
\tilde{L}(d, D, 1/(1+w), w, C, C_1) \geq \inf_{d' \in \mathcal{D}(\mathcal{X}, D, e, 0)} \tilde{L}(d', D, 1/(1+w), w, C, C_1)
= \tilde{L}(\tilde{e}, D, 1/(1+w), w, C, C_1)
= \tilde{L}(e, D, 1/(1+w), w, C, C_1) + \min \left( \frac{w}{n} \tilde{n}_{e,D}^+, \frac{1}{n} \tilde{n}_{e,D}^- \right) + C_1 [\tilde{\alpha}_{e,D} - \alpha_{\text{min}}^{(e,D)}]_+,
\]

where the last quantity is clearly lower-bounded by \( \tilde{L}^*(e, D, w, C, C_1) \) defined in Equation (44). We have now proven that \( \tilde{L}^*(e, D, w, C, C_1) \) given by Equation (44) is a lower bound of \( \tilde{L}(d, D, 1/(1+w), w, C, C_1) \) for \( d \in \mathcal{D}(\mathcal{X}, D, e) \).

Finally, we compute \( \inf_{\beta: \zeta < \beta \leq 1} g(\beta) \) analytically. Since the derivative of \( g \) is given by

\[
g'(\beta) = -\frac{\tilde{n}_{e,D}}{n\beta^2} + C_1,
\]

and \( \beta \) must be positive, the only stationary point \( \beta^* \) of \( g \) that could satisfy the constraint \( \zeta < \beta^* \leq 1 \) is given by

\[
\beta^* = \sqrt{\frac{\tilde{n}_{e,D}}{C_1n}},
\]

and the second derivative test confirms that \( \beta^* \) is a local minimum of \( g \). It then follows that \( \inf_{\beta: \zeta < \beta \leq 1} g(\beta) \) is given by

\[
\inf_{\beta: \zeta < \beta \leq 1} g(\beta) = \begin{cases} g(\beta^*) & \text{if } \zeta < \beta^* \leq 1 \\ \min(g(\zeta), g(1)) & \text{otherwise.} \end{cases}
\]

\[\square\]

12 Additional Rule Lists Demonstrating the Effect of Varying Parameter Values

In this section, we include some additional rule lists created using Algorithm FRL and Algorithm softFRL with varying parameter values. The default parameter values we used in creating these rule lists are \( w = 7 \), \( C = 0.000001 \), and \( C_1 = 0.5 \). In each of the following subsections, the rule lists were created with default parameter values, other than the parameter that was being varied.

12.1 Effect of Varying \( w \) on Algorithm FRL

Running Algorithm FRL with \( w = 1 \) on the bank-full dataset produces the following falling rule list:
Running Algorithm FRL with $w = 1$ on the bank-full dataset produces the following falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>THEN success prob. is</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND loan=no</td>
<td>0.65</td>
<td>934</td>
<td>495</td>
<td></td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND marital=married</td>
<td>THEN success prob. is</td>
<td>0.62</td>
<td>31</td>
<td>19</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND campaign=1</td>
<td>THEN success prob. is</td>
<td>0.56</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>ELSE</td>
<td>success prob. is</td>
<td>0.10</td>
<td>4315</td>
<td>39401</td>
</tr>
</tbody>
</table>

Table 2: Falling rule list for bank-full dataset, created using Algorithm FRL with $w = 1$

Running Algorithm FRL with $w = 3$ on the bank-full dataset produces the following falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>THEN success prob. is</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND previous $\geq 2$</td>
<td>0.65</td>
<td>677</td>
<td>361</td>
<td></td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND campaign=1</td>
<td>THEN success prob. is</td>
<td>0.65</td>
<td>185</td>
<td>99</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND loan=no</td>
<td>THEN success prob. is</td>
<td>0.63</td>
<td>111</td>
<td>65</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND marital=married</td>
<td>THEN success prob. is</td>
<td>0.56</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>ELSE IF 60 $\leq$ age $&lt;$ 100 AND housing=no</td>
<td>THEN success prob. is</td>
<td>0.30</td>
<td>390</td>
<td>919</td>
</tr>
<tr>
<td>ELSE</td>
<td>success prob. is</td>
<td>0.09</td>
<td>3921</td>
<td>38474</td>
</tr>
</tbody>
</table>

Table 3: Falling rule list for bank-full dataset, created using Algorithm FRL with $w = 3$

Running Algorithm FRL with $w = 5$ on the bank-full dataset produces the following falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>THEN success prob. is</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND default=no</td>
<td>THEN success prob. is</td>
<td>0.65</td>
<td>978</td>
<td>531</td>
</tr>
<tr>
<td>ELSE IF 60 $\leq$ age $&lt;$ 100 AND loan=no</td>
<td>THEN success prob. is</td>
<td>0.29</td>
<td>426</td>
<td>1030</td>
</tr>
<tr>
<td>ELSE IF 17 $\leq$ age $&lt;$ 30 AND housing=no</td>
<td>THEN success prob. is</td>
<td>0.25</td>
<td>504</td>
<td>1539</td>
</tr>
<tr>
<td>ELSE IF previous $\geq 2$ AND housing=no</td>
<td>THEN success prob. is</td>
<td>0.23</td>
<td>242</td>
<td>796</td>
</tr>
<tr>
<td>ELSE</td>
<td>success prob. is</td>
<td>0.08</td>
<td>3139</td>
<td>36026</td>
</tr>
</tbody>
</table>

Table 4: Falling rule list for bank-full dataset, created using Algorithm FRL with $w = 5$

Running Algorithm FRL with $w = 7$ on the bank-full dataset produces the following falling rule list:
As the positive class weight $w$ increases, the falling rule list created using Algorithm FRL tends to have rules whose probability estimates are smaller. This is not surprising – a larger value of $w$ means a smaller threshold $\tau = 1/(1 + w)$, and by including rules whose probability estimates are not much larger than the threshold, the falling rule list produced by the algorithm will more likely predict positive, thereby reducing the (weighted) empirical risk of misclassification. Note that Algorithm FRL will never include rules whose probability estimates are less than the threshold (see Corollary 4.5).

### 12.2 Effect of Varying $w$ on Algorithm softFRL

Running Algorithm softFRL with $w = 1$ on the bank-full dataset produces the following softly falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND campaign=1</td>
<td>THEN prob. is 0.67</td>
<td>0.67</td>
<td>557</td>
<td>280</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND marital=married</td>
<td>THEN prob. is 0.65</td>
<td>0.65</td>
<td>263</td>
<td>143</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND loan=no</td>
<td>THEN prob. is 0.61</td>
<td>0.61</td>
<td>154</td>
<td>98</td>
</tr>
<tr>
<td>ELSE</td>
<td>prob. is 0.10</td>
<td>0.10</td>
<td>4315</td>
<td>39401</td>
</tr>
</tbody>
</table>

Table 6: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $w = 1$
<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND marital=married</td>
<td>THEN prob. is 0.65</td>
<td>0.65</td>
<td>547</td>
<td>289</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND loan=no</td>
<td>THEN prob. is 0.65</td>
<td>0.65</td>
<td>418</td>
<td>225</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND campaign=1</td>
<td>THEN prob. is 0.56</td>
<td>0.56</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND previous ≥ 2</td>
<td>THEN prob. is 0.33</td>
<td>0.33</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND housing=no</td>
<td>THEN prob. is 0.30</td>
<td>0.30</td>
<td>390</td>
<td>919</td>
</tr>
<tr>
<td>ELSE IF previous ≥ 2 AND campaign=1</td>
<td>THEN prob. is 0.15</td>
<td>0.15</td>
<td>281</td>
<td>1559</td>
</tr>
<tr>
<td>ELSE prob. is</td>
<td>0.09</td>
<td>0.09</td>
<td>3640</td>
<td>36915</td>
</tr>
</tbody>
</table>

Table 7: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with \( w = 3 \)

Running Algorithm softFRL with \( w = 5 \) on the bank-full dataset produces the following softly falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success</td>
<td>THEN prob. is 0.65</td>
<td>0.65</td>
<td>978</td>
<td>533</td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND loan=no</td>
<td>THEN prob. is 0.29</td>
<td>0.29</td>
<td>426</td>
<td>1030</td>
</tr>
<tr>
<td>ELSE IF poutcome=unknown AND contact=cellular</td>
<td>THEN prob. is 0.11</td>
<td>0.11</td>
<td>2380</td>
<td>18659</td>
</tr>
<tr>
<td>ELSE prob. is</td>
<td>0.07</td>
<td>0.07</td>
<td>1505</td>
<td>19700</td>
</tr>
</tbody>
</table>

Table 8: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with \( w = 5 \)

Running Algorithm softFRL with \( w = 7 \) on the bank-full dataset produces the following softly falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success</td>
<td>THEN prob. is 0.65</td>
<td>0.65</td>
<td>978</td>
<td>533</td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND marital=single</td>
<td>THEN prob. is 0.28</td>
<td>0.28</td>
<td>435</td>
<td>1120</td>
</tr>
<tr>
<td>ELSE IF contact=cellular AND default=no</td>
<td>THEN prob. is 0.18</td>
<td>0.18</td>
<td>970</td>
<td>4504</td>
</tr>
<tr>
<td>ELSE IF contact=cellular AND housing=no</td>
<td>THEN prob. is 0.10</td>
<td>0.10</td>
<td>2255</td>
<td>19970</td>
</tr>
<tr>
<td>ELSE prob. is</td>
<td>0.05</td>
<td>0.05</td>
<td>651</td>
<td>13795</td>
</tr>
</tbody>
</table>

Table 9: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with \( w = 7 \)

As the positive class weight \( w \) increases, the softly falling rule list created using Algorithm softFRL also tends to have rules whose probability estimates are smaller. This is again not surprising – a larger value of \( w \) means a smaller threshold \( \tau = 1/(1 + w) \), and by including rules whose probability estimates are not much larger than the threshold, the softly falling rule list produced by the algorithm will more likely predict positive, thereby reducing the (weighted) empirical risk of misclassification.
12.3 Effect of Varying $C$ on Algorithm FRL

Running Algorithm FRL with $C = 0.000001$ on the bank-full dataset produces the following falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND default=no THEN success prob. is 0.65 978 531</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND default=no THEN success prob. is 0.28 434 1113</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF 17 ≤ age &lt; 30 AND housing=no THEN success prob. is 0.25 504 1539</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF previous ≥ 2 AND housing=no THEN success prob. is 0.23 242 794</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF campaign=1 AND housing=no THEN success prob. is 0.14 658 4092</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF previous ≥ 2 AND education=tertiary THEN success prob. is 0.13 108 707</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE success prob. is 0.07 2365 31146</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10: Falling rule list for bank-full dataset, created using Algorithm FRL with $C = 0.000001$

Running Algorithm FRL with $C = 0.01$ on the bank-full dataset produces the following falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND default=no THEN success prob. is 0.65 978 531</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND loan=no THEN success prob. is 0.29 426 1030</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF 17 ≤ age &lt; 30 AND contact=cellular THEN success prob. is 0.20 653 2621</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE IF campaign=1 AND housing=no THEN success prob. is 0.15 803 4634</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE success prob. is 0.07 2429 31106</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Falling rule list for bank-full dataset, created using Algorithm FRL with $C = 0.01$

Running Algorithm FRL with $C = 0.1$ on the bank-full dataset produces the following falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF housing=no AND contact=cellular THEN success prob. is 0.20 2883 11799</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSE success prob. is 0.08 2406 28123</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Falling rule list for bank-full dataset, created using Algorithm FRL with $C = 0.1$

As the cost $C$ of adding a rule increases, the size of the falling rule list created by Algorithm FRL decreases, as expected.
12.4 Effect of Varying $C$ on Algorithm softFRL

Running Algorithm softFRL with $C = 0.000001$ on the bank-full dataset produces the following softly falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success THEN prob. is 0.65</td>
<td>0.65</td>
<td>978</td>
<td>533</td>
<td></td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 THEN prob. is 0.28</td>
<td>0.28</td>
<td>435</td>
<td>1120</td>
<td></td>
</tr>
<tr>
<td>ELSE IF marital=single AND housing=no THEN prob. is 0.18</td>
<td>0.18</td>
<td>970</td>
<td>4504</td>
<td></td>
</tr>
<tr>
<td>ELSE IF contact=cellular AND default=no THEN prob. is 0.10</td>
<td>0.10</td>
<td>2255</td>
<td>19970</td>
<td></td>
</tr>
<tr>
<td>ELSE prob. is</td>
<td>0.05</td>
<td>651</td>
<td>13795</td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $C = 0.000001$.

Running Algorithm softFRL with $C = 0.01$ on the bank-full dataset produces the following softly falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND loan=no THEN prob. is 0.65</td>
<td>0.65</td>
<td>934</td>
<td>495</td>
<td></td>
</tr>
<tr>
<td>ELSE IF housing=no AND contact=cellular THEN prob. is 0.16</td>
<td>0.16</td>
<td>2245</td>
<td>11535</td>
<td></td>
</tr>
<tr>
<td>ELSE IF housing=yes AND default=no THEN prob. is 0.07</td>
<td>0.07</td>
<td>1677</td>
<td>22591</td>
<td></td>
</tr>
<tr>
<td>ELSE prob. is</td>
<td>0.07</td>
<td>433</td>
<td>5301</td>
<td></td>
</tr>
</tbody>
</table>

Table 14: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $C = 0.01$.

Running Algorithm softFRL with $C = 0.1$ on the bank-full dataset produces the following softly falling rule list:

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF housing=no AND contact=cellular THEN prob. is 0.20</td>
<td>0.20</td>
<td>2883</td>
<td>11799</td>
<td></td>
</tr>
<tr>
<td>ELSE prob. is</td>
<td>0.08</td>
<td>2406</td>
<td>28123</td>
<td></td>
</tr>
</tbody>
</table>

Table 15: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $C = 0.1$.

As the cost $C$ of adding a rule increases, the size of the softly falling rule list created by Algorithm softFRL decreases, as expected.

12.5 Effect of Varying $C_1$ on Algorithm softFRL

Running Algorithm softFRL with $C_1 \in \{0.005, 0.05, 0.5\}$ on the bank-full dataset produces the softly falling rule lists shown in Tables 16, 17, and 18.

When the monotonicity penalty $C_1$ is small, the softly falling rule list created by Algorithm softFRL exhibits the “pulling down” of the empirical positive proportion for a substantial number of rules, because with little monotonicity penalty the algorithm will more likely choose a rule list that frequently violates monotonicity.
<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive proportion</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success THEN prob. is 0.65</td>
<td>0.65</td>
<td>978</td>
<td>533</td>
<td></td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND housing=no THEN prob. is 0.30</td>
<td>0.30</td>
<td>599</td>
<td>1177</td>
<td></td>
</tr>
<tr>
<td>ELSE IF marital=single AND housing=no THEN prob. is 0.18</td>
<td>0.18</td>
<td>970</td>
<td>4504</td>
<td></td>
</tr>
<tr>
<td>ELSE IF marital=single AND previous=0 THEN prob. is 0.08</td>
<td>0.08</td>
<td>456</td>
<td>4936</td>
<td></td>
</tr>
<tr>
<td>ELSE IF campaign ≥ 3 AND education=secondary THEN prob. is 0.06</td>
<td>0.06</td>
<td>323</td>
<td>5294</td>
<td></td>
</tr>
<tr>
<td>ELSE IF 30 ≤ age &lt; 40 AND previous=0 THEN prob. is 0.06</td>
<td>0.08</td>
<td>568</td>
<td>6849</td>
<td></td>
</tr>
<tr>
<td>ELSE IF education=tertiary AND housing=no THEN prob. is 0.06</td>
<td>0.14</td>
<td>361</td>
<td>2237</td>
<td></td>
</tr>
<tr>
<td>ELSE IF loan=yes AND previous=0 THEN prob. is 0.05</td>
<td>0.05</td>
<td>106</td>
<td>1972</td>
<td></td>
</tr>
<tr>
<td>ELSE IF education=secondary AND default=no THEN prob. is 0.05</td>
<td>0.09</td>
<td>595</td>
<td>5779</td>
<td></td>
</tr>
<tr>
<td>ELSE IF campaign=1 THEN prob. is 0.05</td>
<td>0.08</td>
<td>233</td>
<td>2564</td>
<td></td>
</tr>
<tr>
<td>ELSE IF housing=no AND previous=0 THEN prob. is 0.05</td>
<td>0.05</td>
<td>68</td>
<td>1176</td>
<td></td>
</tr>
<tr>
<td>ELSE IF job=management AND contact=cellular THEN prob. is 0.05</td>
<td>0.10</td>
<td>75</td>
<td>693</td>
<td></td>
</tr>
<tr>
<td>ELSE IF job=technician AND poutcome=unknown THEN prob. is 0.05</td>
<td>0.07</td>
<td>10</td>
<td>143</td>
<td></td>
</tr>
<tr>
<td>ELSE IF marital=married THEN prob. is 0.05</td>
<td>0.06</td>
<td>110</td>
<td>1841</td>
<td></td>
</tr>
<tr>
<td>ELSE IF campaign ≥ 3 AND housing=yes THEN prob. is 0.05</td>
<td>0.06</td>
<td>16</td>
<td>238</td>
<td></td>
</tr>
<tr>
<td>ELSE IF marital=single AND housing=yes THEN prob. is 0.05</td>
<td>0.13</td>
<td>13</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>ELSE IF housing=yes AND contact=cellular THEN prob. is 0.05</td>
<td>0.10</td>
<td>8</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>ELSE IF job=blue-collar AND loan=no THEN prob. is 0.05</td>
<td>0.16</td>
<td>4</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>ELSE prob. is</td>
<td>0.05</td>
<td>5</td>
<td>63</td>
<td></td>
</tr>
</tbody>
</table>

Table 16: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $C_1 = 0.005$
but that has a small empirical risk on the training set, in the hope of getting more of the training instances “right”. This is also why the softly falling rule list tends to be longer when \( C_1 \) is small: in minimizing the empirical risk on the training set with little regularization (the default \( C_1 = 0.000001 \) is very small), the algorithm tends to overfit the training data.

When \( C_1 \) becomes larger, the softly falling rule list created by Algorithm softFRL exhibits less “pulling down” of the empirical positive proportion. This is consistent with our expectation that when \( C_1 \) is larger, the penalty for violating monotonicity is higher and the algorithm will less likely choose a rule list that frequently violates monotonicity.

### 13 Additional Experiments Comparing Algorithm FRL and Algorithm softFRL to Other Classification Algorithms

Figure 3 shows the ROC curves on the test set using different values of \( w \), for four additional training-test splits. As we can see, the curves in Figure 3 lie close to each other, again demonstrating the effectiveness of our algorithms in producing falling rule lists that, when used as classifiers, are comparable with classifiers produced by other widely used classification algorithms, in a cost-sensitive setting.
We conducted a set of experiments comparing the Bayesian approach to our optimization approach. We trained falling rule lists on the entire bank-full dataset using both the Bayesian approach and our optimization approach (Algorithm FRL), and plotted the weighted training loss over real runtime. In particular, for each positive class weight $w \in \{1, 3, 5, 7\}$, we set the threshold to $1/(1 + w)$ (By Theorem 2.8, this is the threshold with the least weighted training loss for any given rule list), and computed the weighted training loss using this threshold. For the Bayesian approach, we recorded the runtime and computed the weighted training loss for every $100$ iterations of Markov chain Monte-Carlo sampling with simulated annealing, up to $6000$ iterations. For our optimization approach, we ran Algorithm FRL for $3000$ iterations and recorded the runtime and the weighted training loss whenever the algorithm finds a falling rule list with a smaller (regularized) weighted training loss. Since we want to focus our experiments on the efficiency of searching the model space, the runtimes recorded do not include the time for mining the antecedents. Due to the random nature of both approaches, the experiments were repeated several times.

Figures 4 to 7 show the plots of the weighted training loss over real runtime for the Bayesian approach and our optimization approach (Algorithm FRL), for four additional runs of the same algorithms. Due to the random nature of both approaches, it is sometimes possible that our approach (Algorithm FRL) may find in $3000$ iterations a falling rule list with a slightly larger weighted training loss, compared to the Bayesian
approach with 6000 iterations (see Figure 6d). However, in general, our approach tends to find a falling rule list with a smaller weighted training loss faster, due to aggressive pruning of the search space.

It is worth pointing out that both the Bayesian approach and our optimization approach produce similar falling rule lists. Table 19 shows a falling rule list for the bank-full dataset, obtained in a particular run of the Bayesian approach with 6000 iterations. Table 20 shows a falling rule list for the same dataset, obtained in a particular run of Algorithm FRL with 3000 iterations and the positive class weight \( w = 7 \). As we can see, the top four rules in both falling rule lists are identical. Tables 21 and 22 show another pair of falling rule lists obtained using both approaches in different runs, and in this case, both approaches have identified some common rules for a high chance of marketing success. This means that both the Bayesian approach and our optimization approach tend to identify similar conditions that are significant, but our approach has the added advantage of faster training convergence over the Bayesian approach in general.
Figure 5: Plots of the weighted training loss over real runtime for the Bayesian approach and our optimization approach (Algorithm FRL): second additional run

Table 19: Falling rule list for bank-full dataset, trained using the Bayesian approach with 6000 iterations.
Figure 6: Plots of the weighted training loss over real runtime for the Bayesian approach and our optimization approach (Algorithm FRL): third additional run

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND default=no</td>
<td>0.65</td>
<td>978</td>
<td>531</td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND loan=no</td>
<td>0.29</td>
<td>426</td>
<td>1030</td>
</tr>
<tr>
<td>ELSE IF 17 ≤ age &lt; 30 AND housing=no</td>
<td>0.25</td>
<td>504</td>
<td>1539</td>
</tr>
<tr>
<td>ELSE IF campaign=1 AND housing=no</td>
<td>0.15</td>
<td>787</td>
<td>4471</td>
</tr>
<tr>
<td>ELSE success prob. is</td>
<td>0.07</td>
<td>2594</td>
<td>32351</td>
</tr>
</tbody>
</table>

Table 20: Falling rule list for bank-full dataset, trained using the optimization approach (Algorithm FRL) with 3000 iterations and the positive class weight $w = 7$. 
Figure 7: Plots of the weighted training loss over real runtime for the Bayesian approach and our optimization approach (Algorithm FRL): fourth additional run
<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND housing=no</td>
<td>success prob. is 0.70</td>
<td>729</td>
<td>311</td>
</tr>
<tr>
<td>ELSE IF poutcome=success</td>
<td>success prob. is 0.53</td>
<td>249</td>
<td>222</td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND loan=no</td>
<td>success prob. is 0.29</td>
<td>426</td>
<td>1030</td>
</tr>
<tr>
<td>ELSE IF 17 ≤ age &lt; 30 AND housing=no</td>
<td>success prob. is 0.25</td>
<td>504</td>
<td>1538</td>
</tr>
<tr>
<td>ELSE IF education=tertiary AND housing=no</td>
<td>success prob. is 0.14</td>
<td>790</td>
<td>4750</td>
</tr>
<tr>
<td>ELSE IF marital=single AND contact=cellular</td>
<td>success prob. is 0.12</td>
<td>648</td>
<td>4754</td>
</tr>
<tr>
<td>ELSE IF 1000 ≤ balance &lt; 2000 AND housing=no</td>
<td>success prob. is 0.11</td>
<td>135</td>
<td>1061</td>
</tr>
<tr>
<td>ELSE IF campaign=1 AND contact=cellular</td>
<td>success prob. is 0.10</td>
<td>571</td>
<td>4904</td>
</tr>
<tr>
<td>ELSE IF contact=cellular AND loan=no</td>
<td>success prob. is 0.08</td>
<td>587</td>
<td>6800</td>
</tr>
<tr>
<td>ELSE success prob. is</td>
<td>success prob. is 0.04</td>
<td>650</td>
<td>14552</td>
</tr>
</tbody>
</table>

Table 21: Falling rule list for bank-full dataset, trained using the Bayesian approach with 6000 iterations.

<table>
<thead>
<tr>
<th>antecedent</th>
<th>probability</th>
<th>positive support</th>
<th>negative support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF poutcome=success AND housing=no</td>
<td>success prob. is 0.70</td>
<td>729</td>
<td>311</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND previous ≥ 2</td>
<td>success prob. is 0.55</td>
<td>185</td>
<td>154</td>
</tr>
<tr>
<td>ELSE IF poutcome=success AND default=no</td>
<td>success prob. is 0.48</td>
<td>64</td>
<td>68</td>
</tr>
<tr>
<td>ELSE IF 60 ≤ age &lt; 100 AND loan=no</td>
<td>success prob. is 0.29</td>
<td>426</td>
<td>1030</td>
</tr>
<tr>
<td>ELSE IF previous ≥ 2 AND loan=no</td>
<td>success prob. is 0.25</td>
<td>302</td>
<td>921</td>
</tr>
<tr>
<td>ELSE IF 17 ≤ age &lt; 30 AND housing=no</td>
<td>success prob. is 0.24</td>
<td>444</td>
<td>1413</td>
</tr>
<tr>
<td>ELSE IF education=tertiary AND housing=no</td>
<td>success prob. is 0.13</td>
<td>671</td>
<td>4435</td>
</tr>
<tr>
<td>ELSE success prob. is</td>
<td>success prob. is 0.07</td>
<td>2468</td>
<td>31590</td>
</tr>
</tbody>
</table>

Table 22: Falling rule list for bank-full dataset, trained using the optimization approach (Algorithm FRL) with 3000 iterations and the positive class weight \( w = 7 \).