An Optimization Approach to Learning Falling Rule Lists (Supplementary Material)

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 $e = \{(a_0^{(d)}, \alpha_0^{(d,D)}), (a_1^{(d)}, \alpha_1^{(d,D)}), ..., (a_{p-1}^{(d)}, \alpha_{p-1}^{(d,D)})\}$, the algorithm either: (1) terminates the construction of d by computing the empirical positive proportion after e, $\tilde{\alpha}_{e,D}$, and then adding to d the final else clause with probability estimate $\tilde{\alpha}_{e,D}$, 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 $f_{S,e,D}: S \to \mathbb{R}_{\geq 0}$ 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

$$f_{S,e,D}(A_l) = \lambda\alpha(A_l, e, D) + (1-\lambda)\frac{n^+(A_l, e, D)}{\tilde{n}^+_{e,D}}, \tag{11}$$

where $\alpha(A_l, e, D)$ is the empirical positive proportion of A_l, and $n^+(A_l, e, D)$ is the number of positive training inputs captured by A_l, should A_l 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 $\alpha(A_l, e, D)$ and $n^+(A_l, e, D)/\tilde{n}^+_{e,D}$ for each $A_l \in S$: 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.

```
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* = ∅, L_best = ∞;
for t = 1, ..., T do
    set p = −1, α_p = 1, d = e = ∅;
    while Inequality (9) in Theorem 4.6 does not hold do
        go to Terminate with some probability;
        set p = p + 1, S = ∅;
        for every antecedent A_l ∈ A that is not in d do
            set a_p^(d) = A_l, compute α_p^(d,D), and let e' = {e, (a_p^(d), α_p^(d,D))};
            if α_p^(d,D) ≤ α_{p−1}^(d,D), α_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_l to S;
                end
            end
        end
        if S ≠ ∅ then
            choose an antecedent A_l ∈ 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), α_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 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
```

Algorithm 1: Algorithm FRL

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 $\tilde{L}_{\text{best}} = \tilde{L}(d^*, D, 1/(1+w), w, C, C_1)$ 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 the T lists it has constructed, into a falling rule list by setting $\hat{\alpha}_j^{(d^*)} = \min_{k \le j} \alpha_k^{(d^*,D)}$.

In the process of constructing a 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 $e = \{(a_0^{(d)}, \alpha_0^{(d,D)}), (a_1^{(d)}, \alpha_1^{(d,D)}), ..., (a_{p-1}^{(d)}, \alpha_{p-1}^{(d,D)})\}$, the algorithm either: (1) terminates the construction of d by computing the empirical positive proportion after e, $\tilde{\alpha}_{e,D}$, and then adding to d the final else clause with probability estimate $\tilde{\alpha}_{e,D}$, or (2) randomly picks an antecedent from a candidate set S of possible next antecedents, computes its empirical positive proportion, and use these as the next rule $(a_p^{(d)}, \alpha_p^{(d,D)})$ for d.

The algorithm uses Theorem 5.2 to prune the search space. More specifically, the algorithm terminates the construction of d if $\tilde{L}^*(e, D, w, C, C_1)$ defined by Equation (10) in Theorem 5.2 is equal to $\tilde{L}(\bar{e}, D, 1/(1 + w), w, C, C_1)$, where $\bar{e} = \{e, \tilde{\alpha}_{e,D}\}$ is the compatible rule list in which the prefix e is followed directly by the final else clause. The condition $\tilde{L}^*(e, D, w, C, C_1) = \tilde{L}(\bar{e}, D, 1/(1 + w), w, C, C_1)$ implies that $\bar{e}$ is an optimal compatible rule list that begins with e. If we have $\tilde{L}^*(e, D, w, C, C_1) < \tilde{L}(\bar{e}, D, 1/(1 + w), w, C, C_1)$ instead, the algorithm either terminates the construction of d with some probability, or it 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 best possible objective value $\tilde{L}^*(e', D, w, C, C_1)$ achievable by any rule list that begins with $e' = \{e, (a_p^{(d)}, \alpha_p^{(d,D)})\}$ and is compatible with D (Theorem 5.2) is less than the current best objective value $\tilde{L}_{\text{best}} = \tilde{L}(d^*, D, 1/(1 + w), w, C, C_1)$. If so, 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 $f_{S,e,D} : S \to \mathbb{R}_{\ge 0}$ 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 softFRL is given by

$$f_{S,e,D}(A_l) = \lambda \lfloor \min(\alpha(A_l, e, D), \frac{1.01}{0.01} \alpha_{\min}^{(e,D)} - \frac{1}{0.01} \alpha(A_l, e, D)) \rfloor_+ + (1 - \lambda) \frac{n^+(A_l, e, D)}{\tilde{n}^+_{e,D}}, \tag{12}$$

where $\alpha_{\min}^{(e,D)} = \min_{k<|e|} \alpha_k^{(e,D)}$ is the minimum empirical positive proportion of the antecedents in the prefix e, $\alpha(A_l, e, D)$ is the empirical positive proportion of A_l, and $n^+(A_l, e, D)$ is the number of positive training inputs captured by A_l, should A_l be chosen as the next antecedent after the prefix e. The curiosity function $f_{S,e,D}$ given by (12) is a weighted sum of $\lfloor \min(\alpha(A_l, e, D), (1.01/0.01)\alpha_{|e|-1}^{(e,D)} - (1/0.01)\alpha(A_l, e, D)) \rfloor_+$ and $n^+(A_l, e, D)/\tilde{n}^+_{e,D}$ for each $A_l \in S$: the former encourages the algorithm to choose antecedents that have large empirical positive proportions but do not violate the monotonicity constraint $\alpha(A_l, e, D) \le \alpha_{\min}^{(e,D)}$ by more than 1%, 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 softFRL in our experiments.

The pseudocode of Algorithm softFRL is shown in Algorithm 2.

10 Proofs of Theorem 2.8, Proposition 4.2, Lemma 4.4, Corollary 4.5, and Theorem 4.6

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

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

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 $d^* = \emptyset$, $\tilde{L}_{\text{best}} = \infty$;
for $t = 1, ..., T$ **do**
    set $p = -1$, $\alpha_p = 1$, $d = e = \emptyset$;
    while $\tilde{L}^*(e, D, w, C, C_1) < \tilde{L}(\bar{e}, D, 1/(1+w), w, C, C_1)$ **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 $\tilde{L}^*(e', D, w, C, C_1)$ using Theorem 5.2;
            if $\tilde{L}^*(e', D, w, C, C_1) < \tilde{L}(d^*, D, 1/(1+w), w, C, C_1)$ **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 $\tilde{L}(d, D, 1/(1+w), w, C, C_1)$;
    if $\tilde{L}(d, D, 1/(1+w), w, C, C_1) < \tilde{L}_{best}$ **then**
        set $d^* = d$, $\tilde{L}_{\text{best}} = \tilde{L}(d, D, 1/(1+w), w, C, C_1)$;
    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 $(a_j^{(d)}, \alpha_j^{(d,D)})$ 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.

$$R_j(d,D,\tau,w) = \frac{1}{n}\left(w \sum_{\substack{i:y_i=1\wedge\\ \text{capt}(\mathbf{x}_i,d)=j}} [\alpha_j^{(d,D)} \leq \tau] + \sum_{\substack{i:y_i=-1\wedge\\ \text{capt}(\mathbf{x}_i,d)=j}} [\alpha_j^{(d,D)} > \tau] \right) = \begin{cases} \frac{1}{n} n_{j,d,D}^- & \text{if } \alpha_j^{(d,D)} > \tau \\ \frac{w}{n} n_{j,d,D}^+ & \text{otherwise.} \end{cases} \tag{13}$$

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

$$\begin{aligned}
R_j(d,D,1/(1+w),w) &= \frac{1}{n} n_{j,d,D}^- \quad \text{(by the definition of } R_j \text{ in Equation (13))} \\
&= \frac{1}{n}(n_{j,d,D} - n_{j,d,D}^+) \quad \text{(by the definition of } n_{j,d,D}^+, n_{j,d,D}^-, n_{j,d,D} \text{ in Definition 2.5)} \\
&= \frac{1}{n}(n_{j,d,D} - \alpha_j^{(d,D)} n_{j,d,D}) \quad \text{(by the definition of } \alpha_j^{(d,D)} \text{ in Definition 2.5)} \\
&= \frac{1}{n}(1 - \alpha_j^{(d,D)}) n_{j,d,D} \\
&< \frac{1}{n}\left(1 - \frac{1}{1+w}\right) n_{j,d,D} \\
&= \frac{w}{n}\frac{1}{1+w} n_{j,d,D} \\
&< \frac{w}{n}\alpha_j^{(d,D)} n_{j,d,D} \\
&= \frac{w}{n} n_{j,d,D}^+ \quad \text{(by the definition of } \alpha_j^{(d,D)} \text{ in Definition 2.5)} \\
&= R_j(d,D,\tau,w). \quad \text{(by the definition of } R_j \text{ in Equation (13))}
\end{aligned}$$

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. □

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

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

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

that is compatible with D, and we have

$$\begin{aligned}
\tilde{n}^-_{e,D} &= \tilde{n}_{e,D} - \tilde{n}^+_{e,D} \\
&= n_{|e|,d,D} + ... + n_{|d|,d,D} - \tilde{n}^+_{e,D} \\
&= \frac{1}{\alpha^{(d,D)}_{|e|}} n^+_{|e|,d,D} + ... + \frac{1}{\alpha^{(d,D)}_{|d|}} n^+_{|d|,d,D} - \tilde{n}^+_{e,D} \quad \text{(by Definition 2.5)} \\
&\geq \frac{1}{\alpha^{(d,D)}_{|e|-1}} n^+_{|e|,d,D} + ... + \frac{1}{\alpha^{(d,D)}_{|e|-1}} n^+_{|d|,d,D} - \tilde{n}^+_{e,D} \quad \text{(by the monotonicity constraint)} \\
&= \frac{1}{\alpha^{(d,D)}_{|e|-1}} (n^+_{|e|,d,D} + ... + n^+_{|d|,d,D}) - \tilde{n}^+_{e,D} \\
&= \frac{1}{\alpha^{(d,D)}_{|e|-1}} \tilde{n}^+_{e,D} - \tilde{n}^+_{e,D} \\
&= ((1/\alpha^{(d,D)}_{|e|-1}) - 1)\tilde{n}^+_{e,D} \\
&= ((1/\alpha^{(e,D)}_{|e|-1}) - 1)\tilde{n}^+_{e,D}.
\end{aligned}$$

(3) ⇒ (2): Suppose that Statement (3) holds. Then we have

$$\begin{aligned}
\tilde{\alpha}_{e,D} &= \frac{\tilde{n}^+_{e,D}}{\tilde{n}_{e,D}} \quad \text{(by Definition 2.5)} \\
&= \frac{\tilde{n}^+_{e,D}}{\tilde{n}^+_{e,D} + \tilde{n}^-_{e,D}} \\
&\leq \frac{\tilde{n}^+_{e,D}}{\tilde{n}^+_{e,D} + ((1/\alpha^{(d,D)}_{|e|-1}) - 1)\tilde{n}^+_{e,D}} \quad \text{(by Statement (3))} \\
&= \frac{\tilde{n}^+_{e,D}}{(1 + (1/\alpha^{(d,D)}_{|e|-1}) - 1)\tilde{n}^+_{e,D}} = \alpha^{(d,D)}_{|e|-1}.
\end{aligned}$$

(2) ⇒ (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

$$d' = \{e, (a^{(d')}_{|e|}, \hat{\alpha}^{(d')}_{|e|}), ..., (a^{(d')}_{|d'|-1}, \hat{\alpha}^{(d')}_{|d'|-1}), \hat{\alpha}^{(d')}_{|d'|}\}$$

that begins with a given prefix e, we have

$$\tilde{n}^+_{e,D} = n^+_{|e|,d',D} + ...n^+_{|d'|,d',D}, \tag{14}$$

$$\tilde{n}^-_{e,D} = n^-_{|e|,d',D} + ...n^-_{|d'|,d',D}, \tag{15}$$

and

$$\tilde{n}_{e,D} = n_{|e|,d',D} + ...n_{|d'|,d',D}. \tag{16}$$

Proof. Any positive training input $\mathbf{x}_i$ that is not captured by the prefix e must be captured by some antecedent $a^{(d')}_j$ with $|e| \leq j < |d'|$ in d', or the final else clause in d'. Conversely, any positive training input $\mathbf{x}_i$ that is captured by some antecedent $a^{(d')}_j$ with $|e| \leq j < |d'|$ in d', or the final else clause in d', must not satisfy any antecedent in the prefix e and is consequently not captured by the prefix e. This means that the set of positive training inputs that are not captured by e is exactly the set of positive training inputs that are captured by some antecedent $a_j^{(d')}$ with $|e| \leq j < |d'|$ in d', or the final else clause in d'. It then follows that these two sets have the same number of elements. The former set has $\tilde{n}^+_{e,D}$ number of elements, and the latter has $n^+_{|e|,d',D} + ... + n^+_{|d'|,d',D}$ number of elements. This establishes Equation (14).

We can establish Equations (15) and (16) using essentially the same argument. □

We now prove Lemma 4.4.

Lemma 4.4. Suppose that we are given an instance (D, A, w, C) of Program 2.9, a prefix e that is feasible for Program 2.9 under the training data D and the set of antecedents A, and a (possibly hypothetical) falling rule list d that begins with e and is compatible with D. Then there exists a falling 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

$$L(d', D, 1/(1 + w), w, C) \leq L(d, D, 1/(1 + w), w, C).$$

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 $\bar{e} = \{e, \tilde{\alpha}_{e,D}\}$ (i.e. 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 $\tilde{\alpha}_{e,D}$) 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 $\alpha_{k-1}^{(d,D)} > 1/(1+w)$ 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

$$R_j(d, D, 1/(1 + w), w) = \frac{1}{n} n^-_{j,d,D}. \tag{17}$$

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

$$R_j(d, D, 1/(1 + w), w) = \frac{w}{n} n^+_{j,d,D}. \tag{18}$$

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

$$a_{|e|}^{(d')} = a_{|e|}^{(d)} \vee ... \vee a_{k-1}^{(d)}$$

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 $a_k^{(d)}$ through $a_{|d|}^{(d)}$ in d. This observation implies

$$n^+_{|e|,d',D} = n^+_{|e|,d,D} + ... + n^+_{k-1,d,D}, \tag{19}$$

$$n^-_{|e|,d',D} = n^-_{|e|,d,D} + ... + n^-_{k-1,d,D}, \tag{20}$$

$$n_{|e|,d',D} = n_{|e|,d,D} + ... + n_{k-1,d,D}, \tag{21}$$

$$n^+_{|e|+1,d',D} = n^+_{k,d,D} + ... + n^+_{|d|,d,D}, \tag{22}$$

and

$$n_{|e|+1,d',D} = n_{k,d,D} + ... + n_{|d|,d,D}. \tag{23}$$

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

$$\begin{aligned}\hat{\alpha}^{(d')}_{|e|} = \alpha^{(d',D)}_{|e|} = \frac{n^+_{|e|,d',D}}{n_{|e|,d',D}} &= \frac{n^+_{|e|,d,D} + ... + n^+_{k-1,d,D}}{n_{|e|,d,D} + ... + n_{k-1,d,D}} \\ &= \frac{\alpha^{(d,D)}_{|e|} n_{|e|,d,D} + ... + \alpha^{(d,D)}_{k-1} n_{k-1,d,D}}{n_{|e|,d,D} + ... + n_{k-1,d,D}} > \frac{1}{1+w},\end{aligned}$$

and

$$\begin{aligned}\hat{\alpha}^{(d')}_{|e|+1} = \alpha^{(d',D)}_{|e|+1} = \frac{n^+_{|e|+1,d',D}}{n_{|e|+1,d',D}} &= \frac{n^+_{k,d,D} + ... + n^+_{|d|,d,D}}{n_{k,d,D} + ... + n_{|d|,d,D}} \\ &= \frac{\alpha^{(d,D)}_{k} n_{k,d,D} + ... + \alpha^{(d,D)}_{|d|} n_{|d|,d,D}}{n_{k,d,D} + ... + n_{|d|,d,D}} \leq \frac{1}{1+w}.\end{aligned}$$

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

$$R_{|e|}(d', D, 1/(1+w), w) = \frac{1}{n} n^-_{|e|,d',D} = \frac{1}{n}(n^-_{|e|,d,D} + ... + n^-_{k-1,d,D}),$$

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

$$R_{|e|+1}(d', D, 1/(1+w), w) = \frac{w}{n} n^+_{|e|+1,d',D} = \frac{w}{n}(n^+_{k,d,D} + ... + n^+_{|d|,d,D}),$$

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:

$$\begin{aligned} &R(d', D, 1/(1+w), w) \\ &= R(e, D, 1/(1+w), w) + R_{|e|}(d', D, 1/(1+w), w) + R_{|e|+1}(d', D, 1/(1+w), w) \\ &= R(e, D, 1/(1+w), w) + \frac{1}{n}(n^-_{|e|,d,D} + ... + n^-_{k-1,d,D}) + \frac{w}{n}(n^+_{k,d,D} + ... + n^+_{|d|,d,D}) \\ &= R(e, D, 1/(1+w), w) + \sum_{j=|e|}^{|d|} R_j(d, D, 1/(1+w), w) \\ &= R(d, D, 1/(1+w), w). \end{aligned} \tag{24}$$

Step 3. Put everything together.

Using (24), together with the observation $|d'| = |e| + 1 \leq |d|$, we must also have

$$\begin{aligned} L(d', D, 1/(1+w), w, C) &= R(d', D, 1/(1+w), w) + C|d'| \\ &\leq R(d, D, 1/(1+w), w) + C|d| = L(d, D, 1/(1+w), w, C), \end{aligned}$$

as desired.

Case 2. $\alpha_j^{(d,D)} > 1/(1+w)$ holds for all $j \in \{|e|, |e|+1, ..., |d|\}$. 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 $j \in \{|e|, |e|+1, ..., |d|\}$, is given by Equation (17). Let $d' = \{e, \hat{\alpha}_{|e|}^{(d')}\}$ be the falling rule list of size $|d'| = |e|$ that is compatible with D. Then the instances captured by $a_{|e|}^{(d')}$ (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

$$n^+_{|e|,d',D} = n^+_{|e|,d,D} + ... + n^+_{|d|,d,D}, \tag{25}$$

$$n^-_{|e|,d',D} = n^-_{|e|,d,D} + ... + n^-_{|d|,d,D}, \tag{26}$$

and

$$n_{|e|,d',D} = n_{|e|,d,D} + ... + n_{|d|,d,D}. \tag{27}$$

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

$$\begin{aligned}
\hat{\alpha}_{|e|}^{(d')} = \alpha_{|e|}^{(d',D)} &= \frac{n^+_{|e|,d',D}}{n_{|e|,d',D}} \\
&= \frac{n^+_{|e|,d,D} + ... + n^+_{|d|,d,D}}{n_{|e|,d,D} + ... + n_{|d|,d,D}} \qquad (28)\\
&= \frac{\alpha_{|e|}^{(d,D)} n_{|e|,d,D} + ... + \alpha_{|d|}^{(d,D)} n_{|d|,d,D}}{n_{|e|,d,D} + ... + n_{|d|,d,D}} > \frac{1}{1+w}. \qquad (29)
\end{aligned}$$

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

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

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

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:

$$\begin{aligned}
&R(d', D, 1/(1+w), w) \\
&= R(e, D, 1/(1+w), w) + R_{|e|}(d', D, 1/(1+w), w) \\
&= R(e, D, 1/(1+w), w) + \frac{1}{n}(n^-_{|e|,d,D} + ... + n^-_{|d|,d,D}) \\
&= R(e, D, 1/(1+w), w) + \sum_{j=|e|}^{|d|} R_j(d, D, 1/(1+w), w) \\
&= R(d, D, 1/(1+w), w).
\end{aligned}$$

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

$$\begin{aligned}
L(d', D, 1/(1+w), w, C) &= R(d', D, 1/(1+w), w) + C|d'| \\
&\leq R(d, D, 1/(1+w), w) + C|d| = L(d, D, 1/(1+w), w, C),
\end{aligned}$$

as desired.

Case 3. $\alpha_j^{(d,D)} \leq 1/(1+w)$ holds for all $j \in \{|e|, |e|+1, ..., |d|\}$. The proof is similar to Case 2, with $R_j(d, D, 1/(1+w), w)$ for all $j \in \{|e|, |e|+1, ..., |d|\}$ given by Equation (18), the "greater than" in Inequality 29 replaced by "less than or equal to", and $R_{|e|}(d', D, 1/(1+w), w)$ given by

$$R_{|e|}(d', D, 1/(1+w), w) = \frac{w}{n} n^+_{|e|,d',D} = \frac{w}{n}(n^+_{|e|,d,D} + ... + n^+_{|d|,d,D}).$$

□

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 $j \in \{0, 1, ..., |d^*| - 1\}$.

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

$$e = \{(a_0^{(d^*)}, \alpha_0^{(d^*,D)}), ..., (a_{k-1}^{(d^*)}, \alpha_{k-1}^{(d^*,D)})\}$$

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

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

Observation 10.2. For any rule list d', we have

$$n^-_{|e|,d',D} = \left(\frac{1}{\alpha_{|e|}^{(d',D)}} - 1\right) n^+_{|e|,d',D}, \tag{30}$$

Proof. By Definition 2.5, we have

$$\alpha_{|e|}^{(d',D)} = n^+_{|e|,d',D} / n_{|e|,d',D}.$$

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

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

The desired equation follows from rearranging the terms. □

Observation 10.3. 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

$$n^+_{|e|+1,d',D} = \tilde{n}^+_{e,D} - n^+_{|e|,d',D}, \tag{31}$$

$$n^-_{|e|+1,d',D} = \tilde{n}^-_{e,D} - n^-_{|e|,d',D}, \tag{32}$$

and

$$n_{|e|+1,d',D} = \tilde{n}_{e,D} - n_{|e|,d',D}. \tag{33}$$

Note that since $n^+_{|e|+1,d',D}$, $n^-_{|e|+1,d',D}$, and $n_{|e|+1,d',D}$ are non-negative, Equations (63), (64), and (65) 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

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

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

$$C \geq \min\left(\frac{w}{n}\tilde{n}^+_{e,D}, \frac{1}{n}\tilde{n}^-_{e,D}\right) - \frac{1}{n}\left(\frac{1}{\alpha^{(e,D)}_{|e|-1}} - 1\right)\tilde{n}^+_{e,D} \tag{34}$$

holds, then the falling rule list $\bar{e} = \{e, \tilde{\alpha}_{e,D}\}$ satisfies $L(\bar{e}, D, 1/(1 + w), w, C) = L^*(e, D, w, C)$.

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

Let $d \in \mathcal{F}(\mathcal{X}, D, e)$.

Case 1. $\alpha^{(e,D)}_{|e|-1} > 1/(1 + w)$.

In this case, Lemma 4.4 implies

$$L(d, D, 1/(1 + w), w, C) \geq \inf_{d' \in \mathcal{F}(\mathcal{X},D,e,1) \cup \mathcal{F}(\mathcal{X},D,e,0)} L(d', D, 1/(1 + w), w, C). \tag{35}$$

Note that we have $\mathcal{F}(\mathcal{X}, D, e, 0) = \{\bar{e}\}$, where $\bar{e} = \{e, \tilde{\alpha}_{e,D}\}$ 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 $\tilde{\alpha}_{e,D}$. To see this, we first observe $\bar{e} \in \mathcal{F}(\mathcal{X}, D, e, 0)$. 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 $\alpha^{(e,D)}_{k-1} \geq \alpha^{(e,D)}_k$ 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 $\tilde{\alpha}_{e,D} \leq \alpha^{(e,D)}_{|e|-1}$ (by Proposition 4.2), which together imply that $\bar{e}$ is indeed a falling rule list; and
(iii) we have

$$\begin{aligned}
\tilde{\alpha}_{e,D} &= \frac{\tilde{n}^+_{e,D}}{\tilde{n}_{e,D}} \quad \text{(by the definition of } \tilde{\alpha}_{e,D} \text{ in Definition 2.5)}\\
&= \frac{n^+_{|\bar{e}|,\bar{e},D}}{n_{|\bar{e}|,\bar{e},D}} \quad \text{(by Equations (14) and (16) in Observation 10.1, applied to } \bar{e})\\
&= \alpha^{(\bar{e},D)}_{|\bar{e}|}, \quad \text{(by the definition of the empirical positive proportion in Definition 2.5)}
\end{aligned}$$

which implies that $\bar{e}$ is indeed compatible with D.

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

$$\begin{aligned}
\hat{\alpha}_{|e|}^{(d_0)} &= \alpha_{|e|}^{(d_0,D)} \quad \text{(because } d_0 \text{ must be compatible with } D\text{)} \\
&= \frac{n_{|e|,d_0,D}^{+}}{n_{|e|,d_0,D}} \quad \text{(by the definition of the empirical positive proportion in Definition 2.5)} \\
&= \frac{\tilde{n}_{e,D}^{+}}{\tilde{n}_{e,D}} \quad \text{(by Equations (14) and (16) in Observation 10.1, applied to } d_0 \text{ here)} \\
&= \tilde{\alpha}_{e,D},
\end{aligned}$$

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

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

with $\alpha_{|e|}^{(d',D)} > 1/(1+w)$ and $\alpha_{|e|+1}^{(d',D)} \leq 1/(1+w)$. Note that for any $d_1 = \{e, (a_{|e|}^{(d_1)}, \alpha_{|e|}^{(d_1,D)}), \alpha_{|e|+1}^{(d_1,D)}\} \in \mathcal{F}(\mathcal{X}, D, e, 1) - \mathcal{F}'(\mathcal{X}, D, e, 1)$, we have either $\alpha_{|e|}^{(d_1,D)} \geq \alpha_{|e|+1}^{(d_1,D)} > 1/(1+w)$ or $\alpha_{|e|+1}^{(d_1,D)} \leq \alpha_{|e|}^{(d_1,D)} \leq 1/(1+w)$, 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

$$\inf_{d' \in \mathcal{F}(\mathcal{X},D,e,1) - \mathcal{F}'(\mathcal{X},D,e,1)} L(d', D, 1/(1+w), w, C) \geq L(\bar{e}, D, 1/(1+w), w, C). \tag{36}$$

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

$$\begin{aligned}
&\inf_{d' \in \mathcal{F}(\mathcal{X},D,e,1) \cup \mathcal{F}(\mathcal{X},D,e,0)} L(d', D, 1/(1+w), w, C) \\
&= \inf_{d' \in \mathcal{F}'(\mathcal{X},D,e,1) \cup (\mathcal{F}(\mathcal{X},D,e,1) - \mathcal{F}'(\mathcal{X},D,e,1)) \cup \{\bar{e}\}} L(d', D, 1/(1+w), w, C) \\
&= \min \Bigg(\inf_{d' \in \mathcal{F}'(\mathcal{X},D,e,1)} L(d', D, 1/(1+w), w, C), \\
&\qquad \inf_{d' \in \mathcal{F}(\mathcal{X},D,e,1) - \mathcal{F}'(\mathcal{X},D,e,1)} L(d', D, 1/(1+w), w, C), L(\bar{e}, D, 1/(1+w), w, C) \Bigg) \\
&= \min \left(\inf_{d' \in \mathcal{F}'(\mathcal{X},D,e,1)} L(d', D, 1/(1+w), w, C), L(\bar{e}, D, 1/(1+w), w, C) \right).
\end{aligned} \tag{37}$$

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

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

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

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

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

$$L(\bar{e}, D, 1/(1+w), w, C) = \min\Bigg(L(e, D, 1/(1+w), w, C) + \frac{1}{n}\tilde{n}^-_{e,D},\\ L(e, D, 1/(1+w), w, C) + \frac{w}{n}\tilde{n}^+_{e,D}\Bigg). \tag{38}$$

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

Let $d' = \{e, (a^{(d')}_{|e|}, \alpha^{(d',D)}_{|e|}), \alpha^{(d',D)}_{|e|+1}\} \in \mathcal{F}'(\mathcal{X}, D, e, 1)$. 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 (13) 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 $\alpha^{(d',D)}_{|e|} > 1/(1+w)$ and $\alpha^{(d',D)}_{|e|+1} \leq 1/(1+w)$ for $d' \in \mathcal{F}'(\mathcal{X}, D, e, 1)$), it is not difficult to see

$$L(d', D, 1/(1+w), w, C) = L(e, D, 1/(1+w), w, C) + \frac{1}{n}n^-_{|e|,d',D} + \frac{w}{n}n^+_{|e|+1,d',D} + C. \tag{39}$$

Substituting (30) in Observation 10.2 and (31) in Observation 10.3 into Equation (39), we have

$$\begin{aligned} &L(d', D, 1/(1+w), w, C) \\ &= L(e, D, 1/(1+w), w, C) + \frac{1}{n}\left(\frac{1}{\alpha^{(d',D)}_{|e|}} - 1\right)n^+_{|e|,d',D} + \frac{w}{n}(\tilde{n}^+_{e,D} - n^+_{|e|,d',D}) + C \\ &= L(e, D, 1/(1+w), w, C) + \frac{1}{n}\left(\left(\frac{1}{\alpha^{(d',D)}_{|e|}} - 1 - w\right)n^+_{|e|,d',D} + w\tilde{n}^+_{e,D}\right) + C. \end{aligned} \tag{40}$$

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

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

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

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

we see that $L(d', D, 1/(1+w), w, C)$ is indeed a monotonically decreasing function of both $n^+_{|e|,d',D}$ and $\alpha^{(d',D)}_{|e|}$. Thus, we can obtain a lower bound of $L(d', D, 1/(1+w), w, C)$ by substituting $n^+_{|e|,d',D}$ and $\alpha^{(d',D)}_{|e|}$ 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 $\alpha^{(d',D)}_{|e|} \leq \alpha^{(d',D)}_{|e|-1} = \alpha^{(e,D)}_{|e|-1}$ from d' being a falling rule list gives an upper bound for $\alpha^{(d',D)}_{|e|}$. Substituting these upper bounds into (40), we obtain the following inequality, which gives a lower bound of $L(d', D, 1/(1+w), w, C)$:

$$\begin{aligned}
&L(d', D, 1/(1+w), w, C) \\
&\geq L(e, D, 1/(1+w), w, C) + \frac{1}{n}\left(\left(\frac{1}{\alpha_{|e|-1}^{(e,D)}} - 1 - w\right)\tilde{n}_{e,D}^{+} + w\tilde{n}_{e,D}^{+}\right) + C \\
&= L(e, D, 1/(1+w), w, C) + \frac{1}{n}\left(\left(\frac{1}{\alpha_{|e|-1}^{(e,D)}} - 1\right)\tilde{n}_{e,D}^{+}\right) + C.
\end{aligned}$$

This means

$$\inf_{d' \in \mathcal{F}'(\mathcal{X}, D, e, 1)} L(d', D, 1/(1+w), w, C) \geq L(e, D, 1/(1+w), w, C) + \frac{1}{n}\left(\left(\frac{1}{\alpha_{|e|-1}^{(e,D)}} - 1\right)\tilde{n}_{e,D}^{+}\right) + C. \tag{41}$$

Step 3. Put everything together.

Using (35), (37), (38), and (41), we have

$$\begin{aligned}
&L(d, D, 1/(1+w), w, C) \\
&\geq \min\left(\inf_{d' \in \mathcal{F}'(\mathcal{X}, D, e, 1)} L(d', D, 1/(1+w), w, C), L(\bar{e}, D, 1/(1+w), w, C)\right) \\
&\geq \min\Bigg(L(e, D, 1/(1+w), w, C) + \frac{1}{n}\left(\left(\frac{1}{\alpha_{|e|-1}^{(e,D)}} - 1\right)\tilde{n}_{e,D}^{+}\right) + C, \\
&\qquad \min\left(L(e, D, 1/(1+w), w, C) + \frac{1}{n}\tilde{n}_{e,D}^{-}, L(e, D, 1/(1+w), w, C) + \frac{w}{n}\tilde{n}_{e,D}^{+}\right)\Bigg) \\
&= L(e, D, 1/(1+w), w, C) + \min\left(\frac{1}{n}\left(\left(\frac{1}{\alpha_{|e|-1}^{(e,D)}} - 1\right)\tilde{n}_{e,D}^{+}\right) + C, \frac{w}{n}\tilde{n}_{e,D}^{+}, \frac{1}{n}\tilde{n}_{e,D}^{-}\right),
\end{aligned}$$

as desired.

Case 2. $\alpha_{|e|-1}^{(e,D)} \leq 1/(1+w)$.

This implies $\alpha_j^{(d,D)} \leq 1/(1+w)$ for all $j \in \{|e|, ..., |d|\}$. 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

$$L(d, D, 1/(1+w), w, C) \geq L(e, D, 1/(1+w), w, C) + \min\left(\frac{w}{n}\tilde{n}_{e,D}^{+}, \frac{1}{n}\tilde{n}_{e,D}^{-}\right). \tag{42}$$

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

$$\frac{1}{n}\left(\left(\frac{1}{\alpha_{|e|-1}^{(d',D)}} - 1\right)\tilde{n}_{e,D}^{+}\right) + C \geq \frac{w}{n}\tilde{n}_{e,D}^{+} + C \geq \frac{w}{n}\tilde{n}_{e,D}^{+},$$

which means

$$\min\left(\frac{w}{n}\tilde{n}_{e,D}^{+}, \frac{1}{n}\tilde{n}_{e,D}^{-}\right) = \min\left(\frac{1}{n}\left(\left(\frac{1}{\alpha_{|e|-1}^{(d',D)}} - 1\right)\tilde{n}_{e,D}^{+}\right) + C, \frac{w}{n}\tilde{n}_{e,D}^{+}, \frac{1}{n}\tilde{n}_{e,D}^{-}\right). \tag{43}$$

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

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

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

which implies

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

□

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

$$\tilde{L}(d,D,1/(1+w),w,C,C_1)\geq \tilde{L}^*(e,D,w,C,C_1),$$

where

$$\begin{aligned}\tilde{L}^*(e,D,&w,C,C_1)=\tilde{L}(e,D,1/(1+w),w,C,C_1)\\ &+\min\Bigg(\frac{1}{n}\left(\frac{1}{\alpha_{\min}^{(e,D)}}-1\right)\tilde{n}_{e,D}^{+}+C+C_1\lfloor\tilde{\alpha}_{e,D}-\alpha_{\min}^{(e,D)}\rfloor_{+}+\frac{w}{n}\tilde{n}_{e,D}^{+}\;\;[\tilde{\alpha}_{e,D}\geq\alpha_{\min}^{(e,D)}],\\ &\inf_{\beta:\zeta<\beta\leq 1}g(\beta),\frac{w}{n}\tilde{n}_{e,D}^{+}+C_1\lfloor\tilde{\alpha}_{e,D}-\alpha_{\min}^{(e,D)}\rfloor_{+},\frac{1}{n}\tilde{n}_{e,D}^{-}+C_1\lfloor\tilde{\alpha}_{e,D}-\alpha_{\min}^{(e,D)}\rfloor_{+}\Bigg)\end{aligned}\tag{44}$$

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), $\alpha_{\min}^{(e,D)}$, ζ, and g are defined by

$$\alpha_{\min}^{(e,D)}=\min_{k<|e|}\alpha_k^{(e,D)},\quad \zeta=\max(\alpha_{\min}^{(e,D)},\tilde{\alpha}_{e,D},1/(1+w)),$$

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

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

$$\tilde{L}(d',D,1/(1+w),w,C,C_1)\leq\tilde{L}(d,D,1/(1+w),w,C,C_1).\tag{45}$$

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 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 $\tilde{L}(\bar{e},D,1/(1+w),w,C,C_1)\leq\tilde{L}(d,D,1/(1+w),w,C)$.

Proof. Case 1. There exists some $k \in \{|e|, ..., |d|\}$ that satisfies $\alpha_k^{(d,D)} > 1/(1+w)$ and some $k' \in \{|e|, ..., |d|\}$ that satisfies $\alpha_{k'}^{(d,D)} \leq 1/(1+w)$. For any $j \in \{|e|, ..., |d|\}$ 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 $j \in \{|e|, ..., |d|\}$ 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 (45).

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

$$a_{|e|}^{(d')}(\mathbf{x}) = \quad [\alpha_{\mathrm{capt}(\mathbf{x},d)}^{(d,D)} > 1/(1+w)] \cdot \quad [|e| \leq \mathrm{capt}(\mathbf{x}, d) \leq |d|].$$

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 $a_{|e|}^{(d')}$ 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

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

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

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

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

and

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

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

$$\hat{\alpha}^{(d')}_{|e|} = \alpha^{(d',D)}_{|e|} = \frac{n^{+}_{|e|,d',D}}{n_{|e|,d',D}} = \frac{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j > 1/(1+w)} n^{+}_{j,d,D}}{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j > 1/(1+w)} n_{j,d,D}}$$

$$= \frac{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j > 1/(1+w)} \alpha^{(d,D)}_j n_{j,d,D}}{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j > 1/(1+w)} n_{j,d,D}} > \frac{1}{1+w},$$

and

$$\hat{\alpha}^{(d')}_{|e|+1} = \alpha^{(d',D)}_{|e|+1} = \frac{n^{+}_{|e|+1,d',D}}{n_{|e|+1,d',D}} = \frac{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j \leq 1/(1+w)} n^{+}_{j,d,D}}{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j \leq 1/(1+w)} n_{j,d,D}}$$

$$= \frac{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j \leq 1/(1+w)} \alpha^{(d,D)}_j n_{j,d,D}}{\sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j \leq 1/(1+w)} n_{j,d,D}} \leq \frac{1}{1+w}.$$

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

$$R_{|e|}(d', D, 1/(1+w), w) = \frac{1}{n} n^{-}_{|e|,d',D} = \frac{1}{n} \sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j > 1/(1+w)} n^{-}_{j,d,D},$$

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

$$R_{|e|+1}(d', D, 1/(1+w), w) = \frac{w}{n} n^{+}_{|e|+1,d',D} = \frac{w}{n} \sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j \leq 1/(1+w)} n^{+}_{j,d,D},$$

where we have used (49).

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

$$\begin{aligned}
&R(d', D, 1/(1+w), w) \\
&= R(e, D, 1/(1+w), w) + R_{|e|}(d', D, 1/(1+w), w) + R_{|e|+1}(d', D, 1/(1+w), w) \\
&= R(e, D, 1/(1+w), w) \\
&\quad + \frac{1}{n} \sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j > 1/(1+w)} n^{-}_{j,d,D} + \frac{w}{n} \sum_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j \leq 1/(1+w)} n^{+}_{j,d,D} \\
&= R(e, D, 1/(1+w), w) + \sum_{j=|e|}^{|d|} R_j(d, D, 1/(1+w), w) \\
&= R(d, D, 1/(1+w), w).
\end{aligned} \tag{51}$$

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

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

Let $l \in \{|e|, ..., |d|\}$ be any integer with

$$\alpha^{(d,D)}_l = \max_{j:|e|\leq j\leq |d| \wedge \alpha^{(d,D)}_j > 1/(1+w)} \alpha^{(d,D)}_j. \tag{52}$$

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

$$\sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} S_j(d,D) \quad S_l(d,D) \quad \text{(because } S_l(d,D) \text{ is included in the sum on the left)}$$

$$= b \ _l^{(d,D)} \ \min_{k<l} \ _k^{(d,D)} c_+$$

$$b \ _l^{(d,D)} \ \min_{k<|e|} \ _k^{(d,D)} c_+ . \tag{53}$$

On the other hand, the monotonicity penalty for the $|e|$-th rule in d^0 satisfies

$$S_{|e|}(d^0,D) = b \ _{|e|}^{(d^0,D)} \ \min_{k<|e|} \ _k^{(d^0,D)} c_+ \quad b \ _l^{(d,D)} \ \min_{k<|e|} \ _k^{(d,D)} c_+ , \tag{54}$$

because we have $\min_{k<|e|} \ _k^{(d^0,D)} = \min_{k<|e|} \ _k^{(d,D)}$ (d and d^0 begin with the same prefix e), and

$$\begin{aligned}
\ _{|e|}^{(d^0,D)} &= \frac{n^+_{|e|,d^0,D}}{n_{|e|,d^0,D}} \quad \text{(by the definition of the empirical positive proportion in Definition 2.5)} \\
&= \frac{\sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} n^+_{j,d,D}}{\sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} n_{j,d,D}} \quad \text{(by Equations (46) and (48))} \\
&= \frac{\sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} \ _j^{(d,D)} n_{j,d,D}}{\sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} n_{j,d,D}} \quad \text{(by the definition of } \ _j^{(d,D)} \text{ in Definition 2.5)} \\
& \frac{\sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} \ _l^{(d,D)} n_{j,d,D}}{\sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} n_{j,d,D}} \quad \text{(by the definition of } l \text{ in (52))} \\
&= \ _l^{(d,D)} .
\end{aligned}$$

Combining (53) and (54), we have

$$S_{|e|}(d^0,D) \quad \sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} S_j(d,D). \tag{55}$$

A similar argument will show

$$S_{|e|+1}(d^0,D) \quad \sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} \ 1/(1+w)} S_j(d,D). \tag{56}$$

It then follows from (55) and (56) that the monotonicity penalty of d^0 is at most that of d:

$$\begin{aligned}
S(d^0,D) &= \left(\sum_{j=0}^{|e| \ 1} S_j(d^0,D) \right) + S_{|e|}(d^0,D) + S_{|e|+1}(d^0,D) \\
& \left(\sum_{j=0}^{|e| \ 1} S_j(d,D) \right) + \sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} > 1/(1+w)} S_j(d,D) \\
& + \sum_{j:|e| \ j \ |d| \wedge \ _j^{(d,D)} \ 1/(1+w)} S_j(d,D) \qquad (57) \\
&= S(d,D). \qquad (58)
\end{aligned}$$

Step 4. Put everything together.

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

$$\begin{aligned}\tilde{L}(d', D; 1/(1+w); w; C; C_1) &= R(d', D; 1/(1+w); w) + C|d'| + C_1 S(d', D) \\ &\leq R(d; D; 1/(1+w); w) + C|d| + C_1 S(d; D) \\ &= \tilde{L}(d; D; 1/(1+w); w; C; C_1).\end{aligned}$$

Case 2. Either $\theta_j^{(d;D)} > 1/(1+w)$ holds for all $j \in \{|e|, \ldots, |d|\}$, or $\theta_j^{(d;D)} \leq 1/(1+w)$ holds for all $j \in \{|e|, \ldots, |d|\}$. The construction of $d' = e$ and the proof for $R(d', D; 1/(1+w); w) \leq R(d; D; 1/(1+w); w)$ 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 $|d'| = |e| \leq |d|$. $\square$

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

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

that begins with a given prefix e, we have

$$\tilde{n}^+_{e;D} = n^+_{|e|;d';D} + \ldots n^+_{|d'|;d';D}; \tag{59}$$

$$\tilde{n}_{e;D} = n_{|e|;d';D} + \ldots n_{|d'|;d';D}; \tag{60}$$

and

$$\tilde{n}_{e;D} = n_{|e|;d';D} + \ldots n_{|d'|;d';D}. \tag{61}$$

Proof. Same as Observation 10.1. $\square$

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

$$n_{|e|;d';D} = \left(\frac{1}{\theta_{|e|}^{(d';D)}} - 1\right) n^+_{|e|;d';D}; \tag{62}$$

Proof. Same as Observation 10.2. $\square$

Observation 11.3. For any rule list

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

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

$$n^+_{|e|+1;d';D} = \tilde{n}^+_{e;D} - n^+_{|e|;d';D}; \tag{63}$$

$$n_{|e|+1;d';D} = \tilde{n}_{e;D} - n_{|e|;d';D}; \tag{64}$$

and

$$n_{|e|+1;d';D} = \tilde{n}_{e;D} - n_{|e|;d';D}. \tag{65}$$

Note that since $n^+_{|e|+1;d';D}$, $n_{|e|+1;d';D}$, and $n_{|e|+1;d';D}$ are non-negative, Equations (63), (64), and (65) 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. Same as Observation 10.3. □

Observation 11.4. For any rule list

$$d' = \{e, (a_{|e|}^{(d')}, \hat{\theta}_{|e|}^{(d')}), \hat{\theta}_{|e|+1}^{(d')}\}$$

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

$$\theta_{|e|+1}^{(d',D)} = \frac{\tilde{n}^{+}_{e,D} - n^{+}_{|e|,d',D}}{\tilde{n}^{+}_{e,D} + \tilde{n}^{-}_{e,D} - \frac{1}{\theta_{|e|}^{(d',D)}} n^{+}_{|e|,d',D}}. \tag{66}$$

Proof. By Definition 2.5, we have

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

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

$$\begin{aligned}
\theta_{|e|+1}^{(d',D)} &= \frac{\tilde{n}^{+}_{e,D} - n^{+}_{|e|,d',D}}{(\tilde{n}^{+}_{e,D} - n^{+}_{|e|,d',D}) + (\tilde{n}^{-}_{e,D} - n^{-}_{|e|,d',D})} \\
&= \frac{\tilde{n}^{+}_{e,D} - n^{+}_{|e|,d',D}}{\tilde{n}^{+}_{e,D} + \tilde{n}^{-}_{e,D} - n^{+}_{|e|,d',D} - n^{-}_{|e|,d',D}}.
\end{aligned}$$

Applying Equation (62) in Observation 11.2, we have

$$\begin{aligned}
\theta_{|e|+1}^{(d',D)} &= \frac{\tilde{n}^{+}_{e,D} - n^{+}_{|e|,d',D}}{\tilde{n}^{+}_{e,D} + \tilde{n}^{-}_{e,D} - n^{+}_{|e|,d',D} - \left(\frac{1}{\theta_{|e|}^{(d',D)}} - 1\right) n^{+}_{|e|,d',D}} \\
&= \frac{\tilde{n}^{+}_{e,D} - n^{+}_{|e|,d',D}}{\tilde{n}^{+}_{e,D} + \tilde{n}^{-}_{e,D} - \frac{1}{\theta_{|e|}^{(d',D)}} n^{+}_{|e|,d',D}}.
\end{aligned}$$

□

We are now ready to prove Theorem 5.2.

Proof of Theorem 5.2. Let $\mathcal{D}(X, D, e)$ be the set of (hypothetical and non-hypothetical) rule lists that begin with e and are compatible with D, and let $\mathcal{D}(X, D, e, k)$ be the subset of $\mathcal{D}(X, D, e)$, consisting of those rule lists in $\mathcal{D}(X, D, e)$ that have exactly k rules (excluding the final else clause) following the prefix e. Let $\mathcal{S}(X, D, e, 1)$ be the subset of $\mathcal{D}(X, D, e, 1)$, consisting of those rule lists

$$d' = \{e, (a_{|e|}^{(d')}, \theta_{|e|}^{(d',D)}), \theta_{|e|+1}^{(d',D)}\} \in \mathcal{D}(X, D, e, 1)$$

with $\theta_{|e|}^{(d',D)} > 1/(1+w)$ and $\theta_{|e|+1}^{(d',D)} \leq 1/(1+w)$.

Note that we have $\mathcal{D}(X, D, e, 0) = \{\bar{e}\}$, where $\bar{e} = \{e, \tilde{\theta}_{e,D}\}$ is the rule list in which the final else clause immediately follows the prefix e, and the probability estimate of the final else clause is $\tilde{\theta}_{e,D}$, by a similar argument as that given in the proof of Theorem 4.6 for $\mathcal{F}(X, D, e, 0) = \{\bar{e}\}$.

Let $d \in \mathcal{D}(X, D, e)$.

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

$$\tilde{L}(d, D, 1/(1+w), w, C, C_1) \geq \inf_{d' \in \mathcal{S}(X, D, e, 1) \cup \mathcal{D}(X, D, e, 0)} \tilde{L}(d', D, 1/(1+w), w, C, C_1). \tag{67}$$

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 $d_1 = \{e; (a^{(d_1)}_{|e|}; \;^{(d_1;D)}_{|e|}); \;^{(d_1;D)}_{|e|+1}\} \in S(X; D; e; 1)$ that satisfies

$$\bar{L}(d; D; 1/(1+w); w; C; C_1) \geq \bar{L}(d_1; D; 1/(1+w); w; C; C_1). \tag{68}$$

Since d_1 must also obey

$$\begin{aligned}\bar{L}(d_1; D; 1/(1+w); w; C; C_1) &\geq \inf_{d^0 \in S(X;D;e;1)} \bar{L}(d^0; D; 1/(1+w); w; C; C_1) \\ &\geq \inf_{d^0 \in S(X;D;e;1) \bigcup D(X;D;e;0)} \bar{L}(d^0; D; 1/(1+w); w; C; C_1);\end{aligned} \tag{69}$$

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

$$\bar{L}(d; D; 1/(1+w); w; C; C_1) \geq \bar{L}(e; D; 1/(1+w); w; C; C_1). \tag{70}$$

Since we have $D(X; D; e; 0) = \{e\}$, it is straightforward to see

$$\begin{aligned}\bar{L}(e; D; 1/(1+w); w; C; C_1) &= \inf_{d^0 \in D(X;D;e;0)} \bar{L}(d^0; D; 1/(1+w); w; C; C_1) \\ &\geq \inf_{d^0 \in S(X;D;e;1) \bigcup D(X;D;e;0)} \bar{L}(d^0; D; 1/(1+w); w; C; C_1).\end{aligned} \tag{71}$$

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

$$\begin{aligned}&\inf_{d^0 \in S(X;D;e;1) \bigcup D(X;D;e;0)} \bar{L}(d^0; D; 1/(1+w); w; C; C_1) \\ &= \inf_{d^0 \in S(X;D;e;1) \bigcup \{e\}} \bar{L}(d^0; D; 1/(1+w); w; C; C_1) \\ &= \min\left(\inf_{d^0 \in S(X;D;e;1)} \bar{L}(d^0; D; 1/(1+w); w; C; C_1); \bar{L}(e; D; 1/(1+w); w; C; C_1) \right).\end{aligned} \tag{72}$$

The rest of the proof proceeds in six steps.

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

Since the contribution by the final else clause to $\bar{L}(e; D; 1/(1+w); w; C; C_1)$ is given by $R_{|e|}(e; D; 1/(1+w); w) + b_{\tilde{\ }e;D}\; ^{(e;D)}_{\min} c_+$, where $R_{|e|}(e; D; 1/(1+w); w)$ is defined by Equation (13) and is given by

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

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

$$\begin{aligned}&\bar{L}(e; D; 1/(1+w); w; C; C_1) \\ &= \begin{cases} \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{1}{n}\bar{n}^-_{e;D} + C_1 b_{\tilde{\ }e;D}\; ^{(e;D)}_{\min} c_+ & \text{if } \tilde{\ }_{e;D} > 1/(1+w) \\ \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{w}{n}\bar{n}^+_{e;D} + C_1 b_{\tilde{\ }e;D}\; ^{(e;D)}_{\min} c_+ & \text{otherwise.}\end{cases}\end{aligned}$$

Since $\tilde{\ }_{e;D} > 1/(1+w)$ 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 $\tilde{\ }_{e;D} \leq 1/(1+w)$ is equivalent to $w\bar{n}^+_{e;D} \leq \bar{n}^-_{e;D}$, we can write

$$\begin{aligned}&\bar{L}(e; D; 1/(1+w); w; C; C_1) \\ &= \min\left(\bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{1}{n}\bar{n}^-_{e;D} + C_1 b_{\tilde{\ }e;D}\; ^{(e;D)}_{\min} c_+ ;\right. \\ &\qquad \left. \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{w}{n}\bar{n}^+_{e;D} + C_1 b_{\tilde{\ }e;D}\; ^{(e;D)}_{\min} c_+ \right) \\ &= \bar{L}(e; D; 1/(1+w); w; C; C_1) + \min\left(\frac{w}{n}\bar{n}^+_{e;D}; \frac{1}{n}\bar{n}^-_{e;D} \right) + C_1 b_{\tilde{\ }e;D}\; ^{(e;D)}_{\min} c_+ .\end{aligned} \tag{73}$$

Step 2. Partition the set $S(X;D;e;1)$ into three subsets based on how the softly falling objective is computed. For any $d^0 = \{e;(a^{(d^0)}_{jej};\ ^{(d^0;D)}_{jej});\ ^{(d^0;D)}_{jej+1}\} \in S(X;D;e;1)$, the softly falling objective is given by

$$
\begin{aligned}
&\tilde{L}(d^0;D;1=(1+w);w;C;C_1)\\
&= \tilde{L}(e;D;1=(1+w);w;C;C_1) + \frac{1}{n} n_{jej;d^0;D} + \frac{w}{n} n^+_{jej+1;d^0;D} + C\\
&\quad + C_1 b\ ^{(d^0;D)}_{jej} \quad ^{(e;D)}_{min} c_+ + C_1 b\ ^{(d^0;D)}_{jej+1} \quad ^{(e;D)}_{min} c_+ :
\end{aligned}
\tag{74}
$$

This is because for any $d^0 \in S(X;D;e;1)$, the contribution by both the jej-th rule and the final else clause to $\tilde{L}(d^0;D;1=(1+w);w;C;C_1)$ is given by

$$
R_{jej}(d^0;D;1=(1+w);w) + R_{jej+1}(d^0;D;1=(1+w);w) + C + C_1 b\ ^{(d^0;D)}_{jej} \quad ^{(e;D)}_{min} c_+ + C_1 b\ ^{(d^0;D)}_{jej+1} \quad ^{(e;D)}_{min} c_+ ;
$$

where $R_{jej}(d^0;D;1=(1+w);w)$ and $R_{jej+1}(d^0;D;1=(1+w);w)$ are defined by Equation (13) and are given by

$$
R_{jej}(d^0;D;1=(1+w);w) = \frac{1}{n} n_{jej;d^0;D} \quad \text{and} \quad R_{jej+1}(d^0;D;1=(1+w);w) = \frac{w}{n} n^+_{jej+1;d^0;D}
$$

(because we have $^{(d^0;D)}_{jej} > 1=(1+w)$ and $^{(d^0;D)}_{jej+1} \quad 1=(1+w)$ for $d^0 \in S(X;D;e;1)$).

Let

$$
S_1(X;D;e;1) = \{d^0 = \{e;(a^{(d^0)}_{jej};\ ^{(d^0;D)}_{jej});\ ^{(d^0;D)}_{jej+1}\} \in S(X;D;e;1) :\ ^{(e;D)}_{min} \quad ^{(d^0;D)}_{jej} > \ ^{(d^0;D)}_{jej+1}\};
$$

$$
S_2(X;D;e;1) = \{d^0 = \{e;(a^{(d^0)}_{jej};\ ^{(d^0;D)}_{jej});\ ^{(d^0;D)}_{jej+1}\} \in S(X;D;e;1) :\ ^{(d^0;D)}_{jej} > \ ^{(e;D)}_{min} \quad ^{(d^0;D)}_{jej+1}\};
$$

and

$$
S_3(X;D;e;1) = \{d^0 = \{e;(a^{(d^0)}_{jej};\ ^{(d^0;D)}_{jej});\ ^{(d^0;D)}_{jej+1}\} \in S(X;D;e;1) :\ ^{(d^0;D)}_{jej} > \ ^{(d^0;D)}_{jej+1} > \ ^{(e;D)}_{min}\};
$$

It is easy to see

$$
S(X;D;e;1) = S_3(X;D;e;1) \cup S_1(X;D;e;1) \cup S_2(X;D;e;1):
$$

We observe here that given the prefix e, we can write $\tilde{L}(d^0;D;1=(1+w);w;C;C_1)$ as a function of $n^+_{jej;d^0;D}$ and $^{(d^0;D)}_{jej}$, by substituting (62), (63), and (66) in Observations 11.1, 11.2, and 11.3 into (74).

Step 3. Determine a lower bound of $\tilde{L}(d^0;D;1=(1+w);w;C;C_1)$ for all $d^0 \in S_1(X;D;e;1)$.

Let $d^0 = \{e;(a^{(d^0)}_{jej};\ ^{(d^0;D)}_{jej});\ ^{(d^0;D)}_{jej+1}\} \in S_1(X;D;e;1)$. By the definition of $S_1(X;D;e;1)$, we have

$$
^{(e;D)}_{min} \quad ^{(d^0;D)}_{jej} > \frac{1}{1+w} \quad ^{(d^0;D)}_{jej+1}
\tag{75}
$$

We first prove the following inequality

$$
^{(e;D)}_{min} \quad ^{(d^0;D)}_{jej} > \max(1=(1+w);\tilde{}_{e;D});
\tag{76}
$$

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

$$
\tilde{}_{e;D} = \frac{\tilde{n}^+_{e;D}}{\tilde{n}_{e;D}} = \frac{n^+_{jej;d^0;D} + n^+_{jej+1;d^0;D}}{n_{jej;d^0;D} + n_{jej+1;d^0;D}} = \frac{^{(d^0;D)}_{jej} n_{jej;d^0;D} + \ ^{(d^0;D)}_{jej+1} n_{jej+1;d^0;D}}{n_{jej;d^0;D} + n_{jej+1;d^0;D}}:
\tag{77}
$$

Substituting $^{(d^0;D)}_{jej+1} < \ ^{(d^0;D)}_{jej}$ from (75) into (77), we obtain $\tilde{}_{e;D} < \ ^{(d^0;D)}_{jej}$. Combining this inequality with $^{(e;D)}_{min} \quad ^{(d^0;D)}_{jej} > \frac{1}{1+w}$ from (75), we obtain (76), as desired.

Note that since (76) has to hold for any $d' \in S_1(X, D, e, 1)$, if ${}^{(e,D)}_{\min} \leq \max(1/(1+w), \tilde{\ }_{e,D})$ is true for the given prefix e, then $S_1(X, D, e, 1)$ is empty.

We now show that given the prefix e, the softly falling objective $\tilde{L}(d', D, 1/(1+w), w, C, C_1)$ for d' is a monotonically decreasing function of both $n^+_{|e|,d',D}$ and ${}^{(d',D)}_{|e|}$.

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

$$\begin{aligned}&\tilde{L}(d', D, 1/(1+w), w, C, C_1)\\ &= \tilde{L}(e, D, 1/(1+w), w, C, C_1) + \frac{1}{n}\left(\left(\frac{1}{{}^{(d',D)}_{|e|}} - 1 - w\right) n^+_{|e|,d',D} + w\bar{n}^+_{e,D}\right) + C.\end{aligned} \tag{78}$$

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 $n^+_{|e|,d',D}$ and ${}^{(d',D)}_{|e|}$. Since we have

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

because ${}^{(d',D)}_{|e|} > 1/(1+w)$ holds for any $d' \in S_1(X, D, e, 1)$, and

$$\frac{\partial \tilde{L}(d', D, 1/(1+w), w, C, C_1)}{\partial {}^{(d',D)}_{|e|}} = -\frac{n^+_{|e|,d',D}}{n}\frac{1}{({}^{(d',D)}_{|e|})^2} \leq 0,$$

we see that $\tilde{L}(d', D, 1/(1+w), w, C, C_1)$ is indeed a monotonically decreasing function of both $n^+_{|e|,d',D}$ and ${}^{(d',D)}_{|e|}$. Thus, we can obtain a lower bound of $\tilde{L}(d', D, 1/(1+w), w, C, C_1)$ by substituting $n^+_{|e|,d',D}$ and ${}^{(d',D)}_{|e|}$ with their respective upper bound. The inequality $n^+_{|e|,d',D} \leq \bar{n}^+_{e,D}$ in Observation 11.3 gives an upper bound for $n^+_{|e|,d',D}$, and the inequality ${}^{(d',D)}_{|e|} \leq {}^{(e,D)}_{\min}$ from (75) gives an upper bound for ${}^{(d',D)}_{|e|}$. Substituting these upper bounds into (78), we obtain the following inequality, which gives a lower bound of $\tilde{L}(d', D, 1/(1+w), w, C, C_1)$:

$$\begin{aligned}&\tilde{L}(d', D, 1/(1+w), w, C, C_1)\\ &\geq \tilde{L}(e, D, 1/(1+w), w, C, C_1) + \frac{1}{n}\left(\left(\frac{1}{{}^{(e,D)}_{\min}} - 1 - w\right)\bar{n}^+_{e,D} + w\bar{n}^+_{e,D}\right) + C\\ &= \tilde{L}(e, D, 1/(1+w), w, C, C_1) + \frac{1}{n}\left(\frac{1}{{}^{(e,D)}_{\min}} - 1\right)\bar{n}^+_{e,D} + C.\end{aligned}$$

Step 4. Determine a lower bound of $\tilde{L}(d', D, 1/(1+w), w, C, C_1)$ for all $d' \in S_2(X, D, e, 1)$.

Let $d' = \{e, (a^{(d')}_{|e|}, {}^{(d',D)}_{|e|}), {}^{(d',D)}_{|e|+1}\} \in S_2(X, D, e, 1)$. By the definition of $S_2(X, D, e, 1)$, we have

$${}^{(d',D)}_{|e|} > \frac{1}{1+w} \tag{79}$$

and

$${}^{(d',D)}_{|e|} \geq {}^{(e,D)}_{\min} \geq {}^{(d',D)}_{|e|+1} \tag{80}$$

We first prove the following inequality

$$1 \geq {}^{(d',D)}_{|e|} > \max({}^{(e,D)}_{\min}, \tilde{\ }_{e,D}, 1/(1+w)) = \ , \tag{81}$$

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 $\rho^{(d',D)}_{|e|+1} < \rho^{(d',D)}_{|e|}$ from (75) into (77), we obtain $\tilde{\rho}_{e;D} < \rho^{(d',D)}_{|e|}$. Combining this inequality with (79) and $\rho^{(d',D)}_{|e|} > \rho^{(e;D)}_{\min}$ from (80), we obtain (81), as desired.

We now show that given the prefix e and a particular value of $\rho^{(d',D)}_{|e|}$ that obeys (81), the softly falling objective $\bar{L}(d', D; 1/(1+w); w; C; C_1)$ 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

$$
\begin{aligned}
&\bar{L}(d', D; 1/(1+w); w; C; C_1)\\
&= \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{1}{n}\left(\left(\frac{1}{\rho^{(d',D)}_{|e|}} - 1 - w\right) n^+_{|e|;d';D} + w\bar{n}^+_{e;D}\right) + C\\
&\quad + C_1(\rho^{(d',D)}_{|e|} - \rho^{(e;D)}_{\min}).
\end{aligned} \tag{82}
$$

Note that Equation (82) shows that given the prefix e, $\bar{L}(d', D; 1/(1+w); w; C; C_1)$ is a function of $n^+_{|e|;d';D}$ and $\rho^{(d',D)}_{|e|}$. Differentiating $\bar{L}(d', D; 1/(1+w); w; C; C_1)$ given in (82) with respect to $n^+_{|e|;d';D}$, we obtain

$$
\frac{\partial \bar{L}(d', D; 1/(1+w); w; C; C_1)}{\partial n^+_{|e|;d';D}} = \frac{1}{n}\left(\frac{1}{\rho^{(d',D)}_{|e|}} - 1 - w\right). \tag{83}
$$

Since $\rho^{(d',D)}_{|e|}$ obeys (81), in particular, it obeys $\rho^{(d',D)}_{|e|} > 1/(1+w)$, we have

$$
\frac{1}{\rho^{(d',D)}_{|e|}} - 1 - w < 0;
$$

which then gives $\partial \bar{L}(d', D; 1/(1+w); w; C; C_1)/\partial n^+_{|e|;d';D} < 0$. This means that given the prefix e and a particular value of $\rho^{(d',D)}_{|e|}$ that obeys (81), $\bar{L}(d', D; 1/(1+w); w; C; C_1)$ is a decreasing function of $n^+_{|e|;d';D}$.

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

$$
\begin{aligned}
\tilde{g}(\rho^{(d',D)}_{|e|}) &= \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{1}{n}\left(\frac{1}{\rho^{(d',D)}_{|e|}} - 1\right)\bar{n}^+_{e;D} + C + C_1(\rho^{(d',D)}_{|e|} - \rho^{(e;D)}_{\min})\\
&= \bar{L}(e; D; 1/(1+w); w; C; C_1) + g(\rho^{(d',D)}_{|e|})
\end{aligned}
$$

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

$$
\begin{aligned}
\bar{L}(d', D; 1/(1+w); w; C; C_1) &\ge \inf_{\rho^{(d',D)}_{|e|} :\; < \rho^{(d',D)}_{|e|} \le 1} \tilde{g}(\rho^{(d',D)}_{|e|})\\
&= \bar{L}(e; D; 1/(1+w); w; C; C_1) + \inf_{\rho^{(d',D)}_{|e|} :\; < \rho^{(d',D)}_{|e|} \le 1} g(\rho^{(d',D)}_{|e|}):
\end{aligned}
$$

Step 5. Determine a lower bound of $\bar{L}(d', D; 1/(1+w); w; C; C_1)$ for all $d' \in S_3(X; D; e; 1)$.

Let $d' = \{e; (a^{(d')}_{|e|};\ {}^{(d';D)}_{|e|});\ {}^{(d';D)}_{|e|+1}\} \in S_3(X; D; e; 1)$. By the definition of $S_3(X; D; e; 1)$, we have

$$ {}^{(d';D)}_{|e|} > \frac{1}{1+w} \qquad {}^{(d';D)}_{|e|+1} > {}^{(e;D)}_{\min}. \tag{84}$$

We first prove the following inequality

$$ 1 \qquad {}^{(d';D)}_{|e|} > \max({}^{(e;D)}_{\min}; \tilde{}_{e;D}; 1/(1+w)) = \quad ; \tag{85}$$

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 ${}^{(d';D)}_{|e|+1} < {}^{(d';D)}_{|e|}$ from (84) into (77), we obtain $\tilde{}_{e;D} < {}^{(d';D)}_{|e|}$. Combining this inequality with ${}^{(d';D)}_{|e|} > \frac{1}{1+w} > {}^{(e;D)}_{\min}$ from (84), we obtain (85), as desired.

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

$$\begin{aligned}
&\bar{L}(d', D; 1/(1+w); w; C; C_1) \\
&\quad \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{1}{n} n_{|e|;d';D} + \frac{w}{n} n^{+}_{|e|+1;d';D} + C + C_1 b\ {}^{(d';D)}_{|e|} \qquad {}^{(e;D)}_{\min} c_{+} \qquad (86)\\
&= \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{1}{n} \left(\left(\frac{1}{{}^{(d';D)}_{|e|}} \quad 1 \right) w \right) n^{+}_{|e|;d';D} + w n^{+}_{e;D} \Bigg) + C \\
&\quad + C_1 ({}^{(d';D)}_{|e|} \qquad {}^{(e;D)}_{\min}) \qquad (87)
\end{aligned}$$

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

$$\bar{L}(e; D; 1/(1+w); w; C; C_1) + \inf_{{}^{(d';D)}_{|e|}: < {}^{(d';D)}_{|e|} \quad 1} g({}^{(d';D)}_{|e|});$$

so that we again have

$$\bar{L}(d', D; 1/(1+w); w; C; C_1) \qquad \bar{L}(e; D; 1/(1+w); w; C; C_1) + \inf_{{}^{(d';D)}_{|e|}: < {}^{(d';D)}_{|e|} \quad 1} g({}^{(d';D)}_{|e|}):$$

Step 6. Put everything together.

Suppose, first, that $S(X; D; e; 1)$ is not empty.

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

$$\inf_{d' \in S_1(X;D;e;1)} \bar{L}(d', D; 1/(1+w); w; C; C_1) \qquad \bar{L}(e; D; 1/(1+w); w; C; C_1) + \frac{1}{n} \left(\frac{1}{{}^{(e;D)}_{\min}} \quad 1 \right) n^{+}_{e;D} + C; \tag{88}$$

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

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

$$\inf_{d' \in S_2(X;D;e;1) \cup S_3(X;D;e;1)} \bar{L}(d', D; 1/(1+w); w; C; C_1) \qquad \bar{L}(e; D; 1/(1+w); w; C; C_1) + \inf_{: < \quad 1} g(\ ); \tag{89}$$

which follows from the definition of inf being the greatest lower bound, as well as the lower bound of $\bar{L}(d', D; 1/(1+w); w; C; C_1)$ for $d' \in S_2(X; D; e; 1)$, which we have derived in Step 4, and the lower bound of $\bar{L}(d', D; 1/(1+w); w; C; C_1)$ for $d' \in S_3(X; D; e; 1)$, which we have derived in Step 5.

To derive a lower bound of $\bar{L}(d'; D; 1/(1+w); w; C; C_1)$ for $d' \in S(X; D; e; 1)$, we further observe that if $^{(e;D)}_{\min} \; \max(1/(1+w); \tilde{}_{e;D})$ holds, then by our remark in Step 3, $S_1(X; D; e; 1)$ is empty, and consequently, using (89), we have

$$\inf_{d' \in S(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1) = \inf_{d' \in S_2(X;D;e;1) \cup S_3(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1)$$
$$\bar{L}(e; D; 1/(1+w); w; C; C_1) + \inf_{: < 1} g(\;): \tag{90}$$

On the other hand, if $^{(e;D)}_{\min} > \max(1/(1+w); \tilde{}_{e;D})$ 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

$$\inf_{d' \in S(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1)$$
$$= \min\left(\inf_{d' \in S_1(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1); \inf_{d' \in S_2(X;D;e;1) \cup S_3(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1) \right)$$
$$\bar{L}(e; D; 1/(1+w); w; C; C_1) + \min\left(\frac{1}{n} \quad \frac{1}{^{(e;D)}_{\min}} \quad 1 \quad \tilde{n}^{+}_{e;D} + C; \inf_{: < 1} g(\;) \right): \tag{91}$$

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 $\inf_{d' \in S(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1)$ is given by either

$$\inf_{d' \in S_2(X;D;e;1) \cup S_3(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1) \quad \text{or} \quad \inf_{d' \in S_1(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1);$$

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

Putting these cases together, we have

$$\inf_{d' \in S(X;D;e;1)} \bar{L}(d'; D; 1/(1+w); w; C; C_1)$$
$$\bar{L}(e; D; 1/(1+w); w; C; C_1)$$
$$+ \begin{cases} \min\left(\frac{1}{n} \quad \frac{1}{^{(e;D)}_{\min}} \quad 1 \quad \tilde{n}^{+}_{e;D} + C; \inf_{: < 1} g(\;) \right) & \text{if } ^{(e;D)}_{\min} > \max(1/(1+w); \tilde{}_{e;D}); \\ \inf_{: < 1} g(\;) & \text{otherwise.} \end{cases} \tag{92}$$

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

$$\bar{L}(d; D; 1/(1+w); w; C; C_1)$$
$$\bar{L}(e; D; 1/(1+w); w; C; C_1)$$
$$+ \begin{cases} \min\left(\frac{1}{n} \quad \frac{1}{^{(e;D)}_{\min}} \quad 1 \quad \tilde{n}^{+}_{e;D} + C; \inf_{: < 1} g(\;); \frac{w}{n} \tilde{n}^{+}_{e;D} + C_1 b \tilde{}_{e;D} \quad ^{(e;D)}_{\min} c_+; \right. \\ \left. \quad \frac{1}{n} \tilde{n}_{e;D} + C_1 b \tilde{}_{e;D} \quad ^{(e;D)}_{\min} c_+ \right) & \text{if } ^{(e;D)}_{\min} > \max(1/(1+w); \tilde{}_{e;D}); \\ \min\left(\inf_{: < 1} g(\;); \frac{w}{n} \tilde{n}^{+}_{e;D} + C_1 b \tilde{}_{e;D} \quad ^{(e;D)}_{\min} c_+; \right. \\ \left. \quad \frac{1}{n} \tilde{n}_{e;D} + C_1 b \tilde{}_{e;D} \quad ^{(e;D)}_{\min} c_+ \right) & \text{otherwise.} \end{cases} \tag{93}$$

Note that the quantity labeled (93) is precisely equal to $\bar{L}(e; D; w; C; C_1)$ given by Equation (44) in the statement of the theorem, because:
(i) if $^{(e;D)}_{\min} > \max(1/(1+w); \tilde{}_{e;D})$ holds, then the rst term in the minimum on the right-hand side of Equation (44) is precisely $\frac{1}{n} \quad \frac{1}{^{(e;D)}_{\min}} \quad 1 \quad \tilde{n}^{+}_{e;D} + C$;

(ii) if $^{(e;D)}_{\min} > \max(1=(1+w);\tilde{}_{e;D})$ does not hold, then we have $^{(e;D)}_{\min}$ $1=(1+w)$ or $^{(e;D)}_{\min}$ $\tilde{}_{e;D})$ in the former case where $^{(e;D)}_{\min}$ $1=(1+w)$ holds, we have

$$\frac{1}{n}\left(\frac{1}{^{(e;D)}_{\min}} \quad 1\right) \hbar^{+}_{e;D} \quad \frac{w}{n}\hbar^{+}_{e;D};$$

which implies that the rst term in the minimum on the right-hand side of Equation (44) is bounded below by $\frac{w}{n}\hbar^{+}_{e;D} + C_1 b\tilde{}_{e;D} \quad ^{(e;D)}_{\min} c_{+}$, and thus has no in uence over the computation of the minimum; in the latter case where $^{(e;D)}_{\min}$ $\tilde{}_{e;D})$ holds, the rst term in the minimum on the right-hand side of Equation (44) is clearly bounded below by $\frac{w}{n}\hbar^{+}_{e;D} + C_1 b\tilde{}_{e;D} \quad ^{(e;D)}_{\min} c_{+}$, and again has no in uence over the computation of the minimum.

This proves that $L(e;D;w;C;C_1)$ given by Equation (44) is indeed a lower bound of $L(d;D;1=(1+w);w;C;C_1)$ for $d \; 2 \; D(X;D;e)$, in the case where $S(X;D;e;1)$ is not empty. In the case where $S(X;D;e;1)$ is empty, using (67) and (73), along with the fact $D(X;D;e;0) = f e g$, we have

$$\begin{aligned} L(d;D;1=(1+w);w;C;C_1) \quad & \inf_{d^0 2 D(X;D;e;0)} L(d^0;D;1=(1+w);w;C;C_1) \\ &= L(e;D;1=(1+w);w;C;C_1) \\ &= L(e;D;1=(1+w);w;C;C_1) + \min\left(\frac{w}{n}\hbar^{+}_{e;D};\frac{1}{n}\hbar_{e;D}\right) + C_1 b\tilde{}_{e;D} \quad ^{(e;D)}_{\min} c_{+}; \end{aligned}$$

where the last quantity is clearly lower-bounded by $L(e;D;w;C;C_1)$ de ned in Equation (44). We have now proven that $L(e;D;w;C;C_1)$ given by Equation (44) is a lower bound of $L(d;D;1=(1+w);w;C;C_1)$ for $d \; 2 \; D(X;D;e)$.

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

$$g^0(\ ) = \frac{\hbar^{+}_{e;D}}{n\ ^2} + C_1;$$

and must be positive, the only stationary point of g that could satisfy the constraint $<$ 1 is given by

$$= \sqrt{\frac{\hbar^{+}_{e;D}}{C_1 n}};$$

and the second derivative test con rms that is a local minimum of g. It then follows that $\inf_{:<\ 1} g(\ )$ is given by

$$\inf_{:<\ 1} g(\ ) = \begin{cases} g(\ ) & \text{if } < \quad 1 \\ \min(g(\ );g(1)) & \text{otherwise:} \end{cases}$$

□

12 Additional Rule Lists Demonstrating the E ect 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 E ect of Varying w on Algorithm FRL

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

	antecedent		probability	positive support	negative support
IF	poutcome=success AND loan=no	THEN success prob. is	0.65	934	495
ELSE IF	poutcome=success AND marital=married	THEN success prob. is	0.62	31	19
ELSE IF	poutcome=success AND campaign=1	THEN success prob. is	0.56	9	7
ELSE		success prob. is	0.10	4315	39401

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:

	antecedent		probability	positive support	negative support
IF	poutcome=success AND previous 2	THEN success prob. is	0.65	677	361
ELSE IF	poutcome=success AND campaign=1	THEN success prob. is	0.65	185	99
ELSE IF	poutcome=success AND loan=no	THEN success prob. is	0.63	111	65
ELSE IF	poutcome=success AND marital=married	THEN success prob. is	0.56	5	4
ELSE IF	60 age< 100 AND housing=no	THEN success prob. is	0.30	390	919
ELSE		success prob. is	0.09	3921	38474

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:

	antecedent		probability	positive support	negative support
IF	poutcome=success AND default=no	THEN success prob. is	0.65	978	531
ELSE IF	60 age< 100 AND loan=no	THEN success prob. is	0.29	426	1030
ELSE IF	17 age< 30 AND housing=no	THEN success prob. is	0.25	504	1539
ELSE IF	previous 2 AND housing=no	THEN success prob. is	0.23	242	796
ELSE		success prob. is	0.08	3139	36026

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:

	antecedent		probability	positive support	negative support
IF	poutcome=success AND default=no	THEN success prob. is	0.65	978	531
ELSE IF	60 age < 100 AND default=no	THEN success prob. is	0.28	434	1113
ELSE IF	17 age < 30 AND housing=no	THEN success prob. is	0.25	504	1539
ELSE IF	previous 2 AND housing=no	THEN success prob. is	0.23	242	794
ELSE IF	campaign=1 AND housing=no	THEN success prob. is	0.14	658	4092
ELSE IF	previous 2 AND education=tertiary	THEN success prob. is	0.13	108	707
ELSE		success prob. is	0.07	2365	31146

Table 5: Falling rule list for bank-full dataset, created using Algorithm FRL with $w = 7$

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 $= 1=(1 + w)$, 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:

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success AND campaign=1	THEN prob. is	0.67	0.67	557	280
ELSE IF	poutcome=success AND marital=married	THEN prob. is	0.65	0.65	263	143
ELSE IF	poutcome=success AND loan=no	THEN prob. is	0.61	0.61	154	98
ELSE		prob. is	0.10	0.10	4315	39401

Table 6: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $w = 1$

Note that there is an extra column positive proportion in a table showing a softly falling rule list. This column gives the empirical positive proportion of each antecedent in the softly falling rule list. When the probability estimate of a rule is less than the positive proportion of the antecedent in the same rule, we know that the softly falling rule list has been transformed from a non-falling compatible rule list, and that the monotonicity penalty has been incurred in the process of running Algorithm softFRL.

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

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success AND marital=married	THEN prob. is	0.65	0.65	547	289
ELSE IF	poutcome=success AND loan=no	THEN prob. is	0.65	0.65	418	225
ELSE IF	poutcome=success AND campaign=1	THEN prob. is	0.56	0.56	9	7
ELSE IF	poutcome=success AND previous 2	THEN prob. is	0.33	0.33	4	8
ELSE IF	60 age < 100 AND housing=no	THEN prob. is	0.30	0.30	390	919
ELSE IF	previous 2 AND campaign=1	THEN prob. is	0.15	0.15	281	1559
ELSE		prob. is	0.09	0.09	3640	36915

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:

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success	THEN prob. is	0.65	0.65	978	533
ELSE IF	60 age < 100 AND loan=no	THEN prob. is	0.29	0.29	426	1030
ELSE IF	poutcome=unknown AND contact=cellular	THEN prob. is	0.11	0.11	2380	18659
ELSE		prob. is	0.07	0.07	1505	19700

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:

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success	THEN prob. is	0.65	0.65	978	533
ELSE IF	60 age < 100	THEN prob. is	0.28	0.28	435	1120
ELSE IF	marital=single AND housing=no	THEN prob. is	0.18	0.18	970	4504
ELSE IF	contact=cellular AND default=no	THEN prob. is	0.10	0.10	2255	19970
ELSE		prob. is	0.05	0.05	651	13795

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 $= 1/(1 + w)$, 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 misclassi cation.

12.3 E ect of Varying C on Algorithm FRL

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

	antecedent		probability	positive support	negative support
IF	poutcome=success AND default=no	THEN success prob. is	0.65	978	531
ELSE IF	60 age < 100 AND default=no	THEN success prob. is	0.28	434	1113
ELSE IF	17 age < 30 AND housing=no	THEN success prob. is	0.25	504	1539
ELSE IF	previous 2 AND housing=no	THEN success prob. is	0.23	242	794
ELSE IF	campaign=1 AND housing=no	THEN success prob. is	0.14	658	4092
ELSE IF	previous 2 AND education=tertiary	THEN success prob. is	0.13	108	707
ELSE		success prob. is	0.07	2365	31146

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:

	antecedent		probability	positive support	negative support
IF	poutcome=success AND default=no	THEN success prob. is	0.65	978	531
ELSE IF	60 age < 100 AND loan=no	THEN success prob. is	0.29	426	1030
ELSE IF	17 age < 30 AND contact=cellular	THEN success prob. is	0.20	653	2621
ELSE IF	campaign=1 AND housing=no	THEN success prob. is	0.15	803	4634
ELSE		success prob. is	0.07	2429	31106

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:

	antecedent		probability	positive support	negative support
IF	housing=no AND contact=cellular	THEN success prob. is	0.20	2883	11799
ELSE		success prob. is	0.08	2406	28123

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:

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success	THEN prob. is	0.65	0.65	978	533
ELSE IF	60 age < 100	THEN prob. is	0.28	0.28	435	1120
ELSE IF	marital=single AND housing=no	THEN prob. is	0.18	0.18	970	4504
ELSE IF	contact=cellular AND default=no	THEN prob. is	0.10	0.10	2255	19970
ELSE		prob. is	0.05	0.05	651	13795

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:

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success AND loan=no	THEN prob. is	0.65	0.65	934	495
ELSE IF	housing=no AND contact=cellular	THEN prob. is	0.16	0.16	2245	11535
ELSE IF	housing=yes AND default=no	THEN prob. is	0.07	0.07	1677	22591
ELSE		prob. is	0.07	0.08	433	5301

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:

	antecedent		probability	positive proportion	positive support	negative support
IF	housing=no AND contact=cellular	THEN prob. is	0.20	0.20	2883	11799
ELSE		prob. is	0.08	0.08	2406	28123

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

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success	THEN prob. is	0.65	0.65	978	533
ELSE IF	$60 \leq$ age < 100 AND housing=no	THEN prob. is	0.30	0.30	599	1177
ELSE IF	marital=single AND housing=no	THEN prob. is	0.18	0.18	970	4504
ELSE IF	marital=single AND previous=0	THEN prob. is	0.08	0.08	456	4936
ELSE IF	campaign ≥ 3 AND education=secondary	THEN prob. is	0.06	0.06	323	5294
ELSE IF	$30 \leq$ age < 40 AND previous=0	THEN prob. is	0.06	0.08	568	6849
ELSE IF	education=tertiary AND housing=no	THEN prob. is	0.06	0.14	361	2237
ELSE IF	loan=yes AND previous=0	THEN prob. is	0.05	0.05	106	1972
ELSE IF	education=secondary AND default=no	THEN prob. is	0.05	0.09	595	5779
ELSE IF	campaign=1	THEN prob. is	0.05	0.08	233	2564
ELSE IF	housing=no AND previous=0	THEN prob. is	0.05	0.05	68	1176
ELSE IF	job=management AND contact=cellular	THEN prob. is	0.05	0.10	75	693
ELSE IF	job=technician AND poutcome=unknown	THEN prob. is	0.05	0.07	10	143
ELSE IF	marital=married	THEN prob. is	0.05	0.06	110	1841
ELSE IF	campaign ≥ 3 AND housing=yes	THEN prob. is	0.05	0.06	16	238
ELSE IF	marital=single AND housing=yes	THEN prob. is	0.05	0.13	13	91
ELSE IF	housing=yes AND contact=cellular	THEN prob. is	0.05	0.10	8	69
ELSE IF	job=blue-collar AND loan=no	THEN prob. is	0.05	0.16	4	21
ELSE		prob. is	0.05	0.07	5	63

Table 16: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $C_1 = 0.005$

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success AND default=no	THEN prob. is	0.65	0.65	978	531
ELSE IF	housing=yes	THEN prob. is	0.07	0.07	1686	22974
ELSE IF	$50 \leq$ age < 60 AND poutcome=unknown	THEN prob. is	0.07	0.09	367	3806
ELSE IF	contact=cellular AND default=no	THEN prob. is	0.07	0.18	1927	8961
ELSE IF	campaign=1 AND poutcome=unknown	THEN prob. is	0.07	0.08	126	1374
ELSE IF	campaign ≥ 3 AND loan=no	THEN prob. is	0.07	0.08	93	1110
ELSE IF	campaign=2 AND education=tertiary	THEN prob. is	0.07	0.09	18	192
ELSE IF	loan=no	THEN prob. is	0.07	0.10	72	648
ELSE		prob. is	0.06	0.06	22	326

Table 17: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $C_1 = 0.05$

	antecedent		probability	positive proportion	positive support	negative support
IF	poutcome=success	THEN prob. is	0.65	0.65	978	533
ELSE IF	$60 \leq$ age < 100	THEN prob. is	0.28	0.28	435	1120
ELSE IF	marital=single AND housing=no	THEN prob. is	0.18	0.18	970	4504
ELSE IF	contact=cellular AND default=no	THEN prob. is	0.10	0.10	2255	19970
ELSE		prob. is	0.05	0.05	651	13795

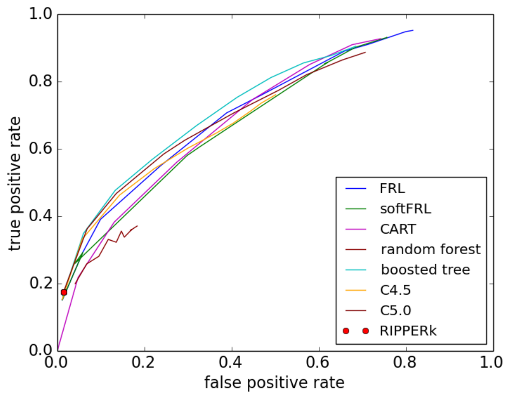
Table 18: Softly falling rule list for bank-full dataset, created using Algorithm softFRL with $C_1 = 0.5$

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 $C = 0.000001$ is very small), the algorithm tends to overfit the training data.

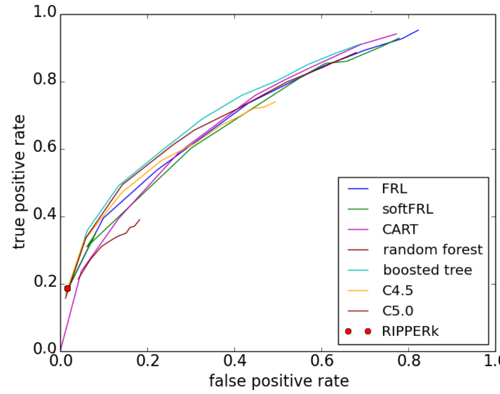
When C_1 becomes larger, the softly falling rule list created by Algorithm softFRL exhibits less "pulling down" of the empirical positive proportion. This is consistent with our expectation that when C_1 is larger, the penalty for violating monotonicity is higher and the algorithm will less likely choose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

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

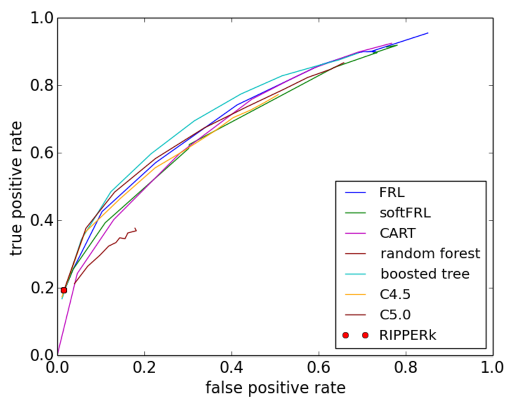
Figure 3 shows the ROC curves on the test set using different values of w, for four additional training-test splits. As we can see, the curves in Figure 3 lie close to each other, again demonstrating the effectiveness of our algorithms in producing falling rule lists that, when used as classifiers, are comparable with classifiers produced by other widely used classification algorithms, in a cost-sensitive setting.



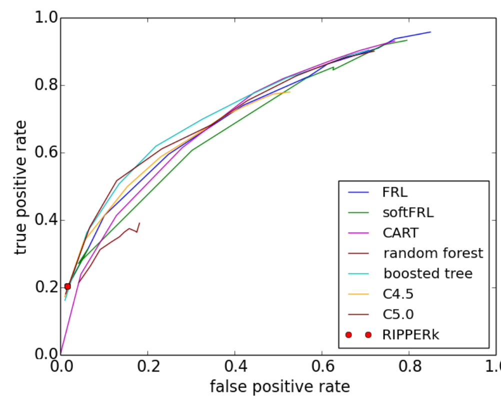
(a) ROC curves on the test set using different w values for the first additional training-test split



(b) ROC curves on the test set using different w values for the second additional training-test split



(c) ROC curves on the test set using different w values for the third additional training-test split



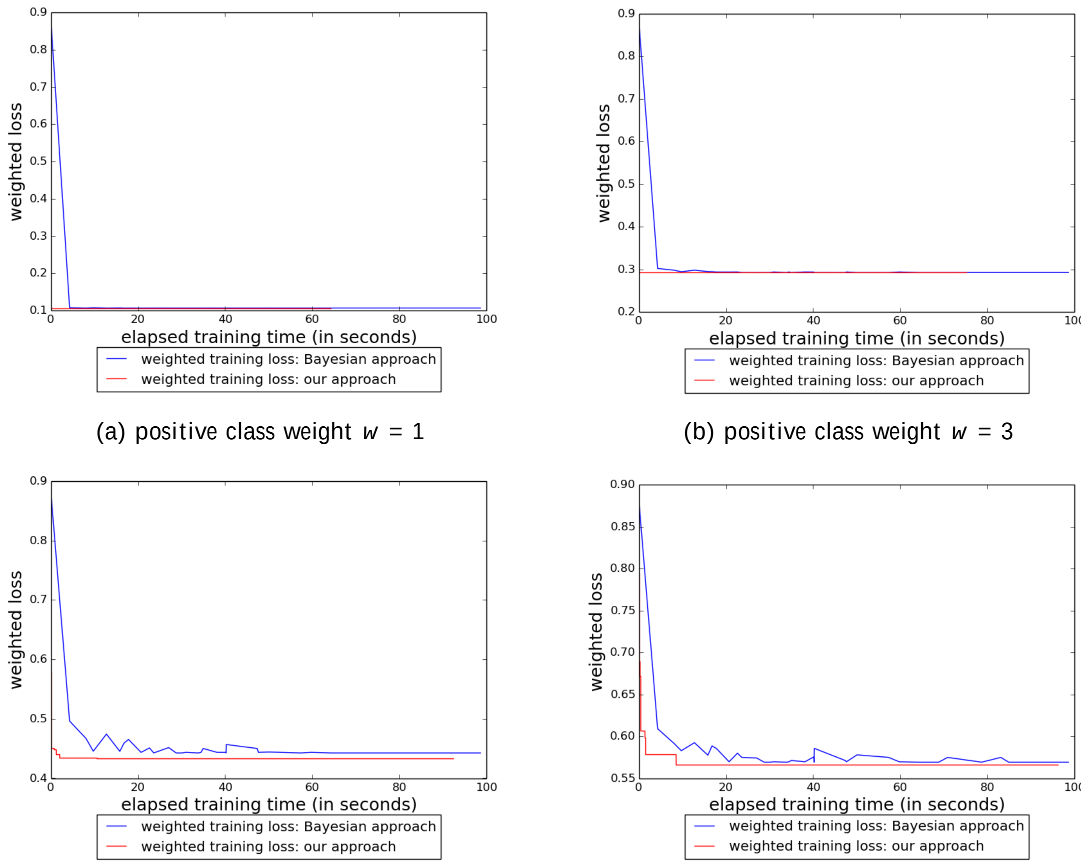
(d) ROC curves on the test set using different w values for the fourth additional training-test split

Figure 3: ROC curves on the test set using different w values for four additional training-test splits

14 Additional Experiments Comparing Bayesian Approach to Our Optimization Approach

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.



(a) positive class weight w = 1

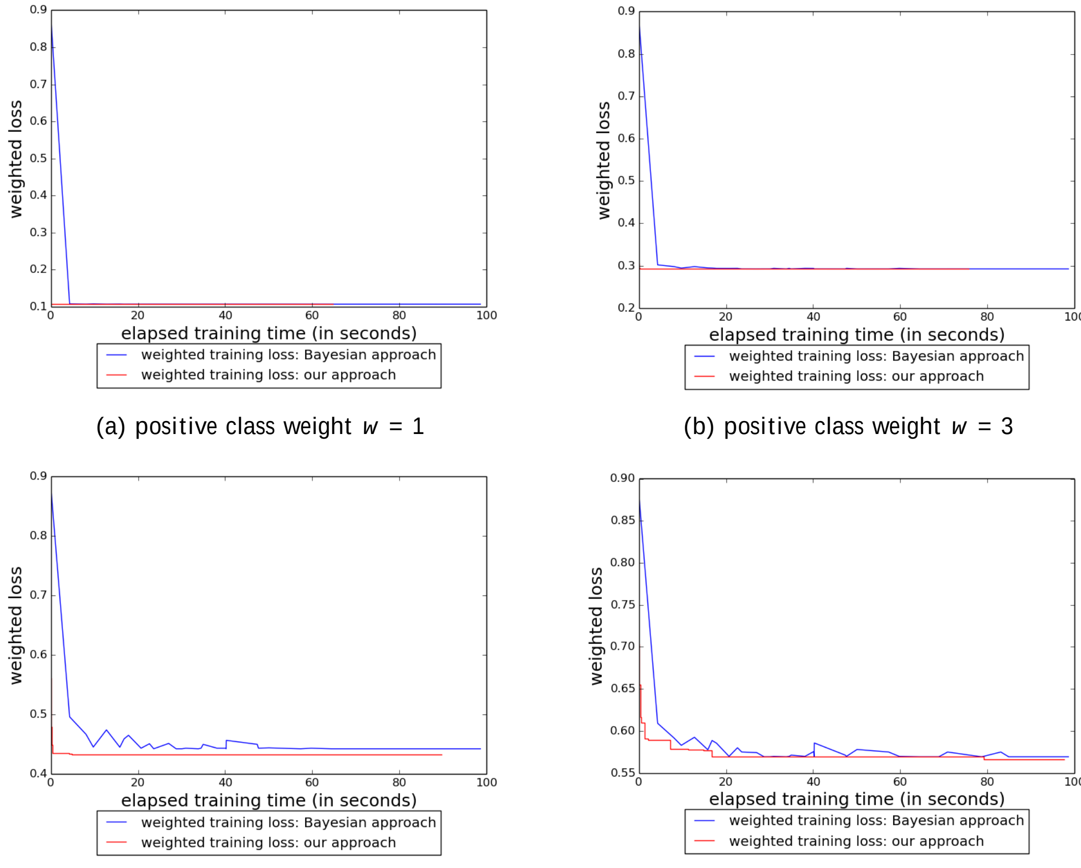
(b) positive class weight w = 3

(c) positive class weight w = 5

(d) positive class weight w = 7

Figure 4: Plots of the weighted training loss over real runtime for the Bayesian approach and our optimization approach (Algorithm FRL): first additional run

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.



(a) positive class weight $w = 1$

(b) positive class weight $w = 3$

(c) positive class weight $w = 5$

(d) positive class weight $w = 7$

Figure 5: Plots of the weighted training loss over real runtime for the Bayesian approach and our optimization approach (Algorithm FRL): second additional run

	antecedent		probability	positive support	negative support
IF	poutcome=success AND default=no	THEN success prob. is	0.65	978	531
ELSE IF	$60 \leq$ age < 100 AND loan=no	THEN success prob. is	0.29	426	1030
ELSE IF	$17 \leq$ age < 30 AND housing=no	THEN success prob. is	0.25	504	1539
ELSE IF	campaign=1 AND housing=no	THEN success prob. is	0.15	787	4471
ELSE IF	education=tertiary AND housing=no	THEN success prob. is	0.12	460	3313
ELSE IF	marital=single AND contact=cellular	THEN success prob. is	0.11	550	4331
ELSE IF	contact=cellular	THEN success prob. is	0.08	1080	12709
ELSE		success prob. is	0.04	504	11998

Table 19: Falling rule list for bank-full dataset, trained using the Bayesian approach with 6000 iterations.

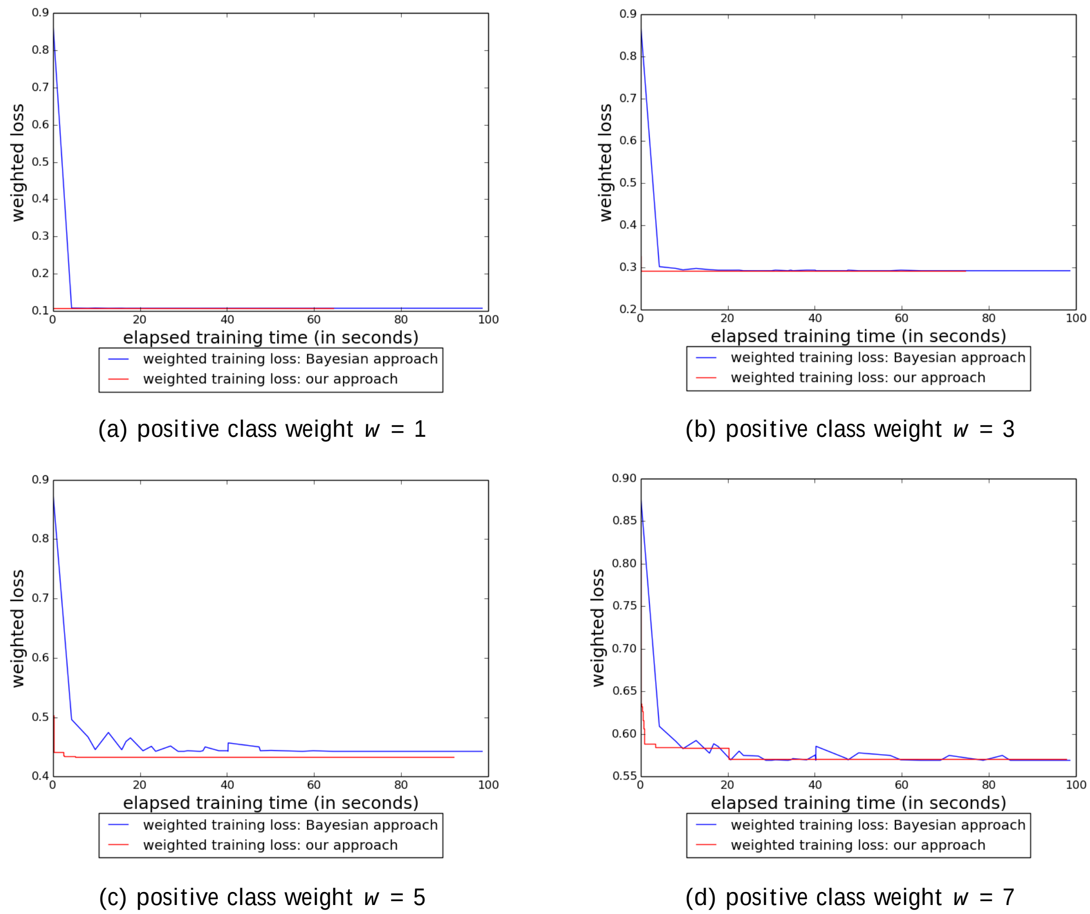
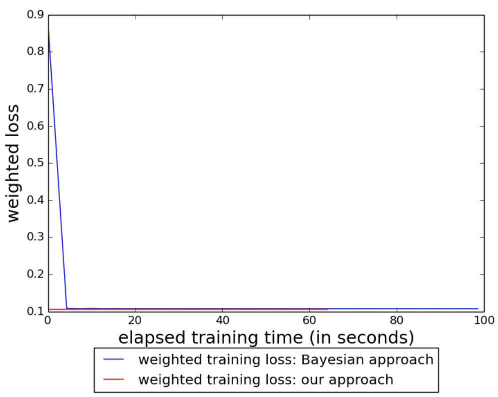


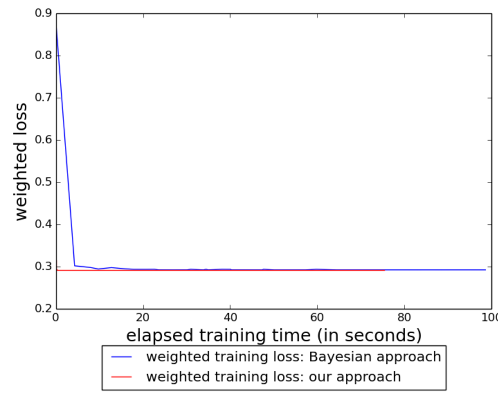
Figure 6: Plots of the weighted training loss over real runtime for the Bayesian approach and our optimization approach (Algorithm FRL): third additional run

	antecedent		probability	positive support	negative support
IF	poutcome=success AND default=no	THEN success prob. is	0.65	978	531
ELSE IF	$60 \leq$ age < 100 AND loan=no	THEN success prob. is	0.29	426	1030
ELSE IF	$17 \leq$ age < 30 AND housing=no	THEN success prob. is	0.25	504	1539
ELSE IF	campaign=1 AND housing=no	THEN success prob. is	0.15	787	4471
ELSE		success prob. is	0.07	2594	32351

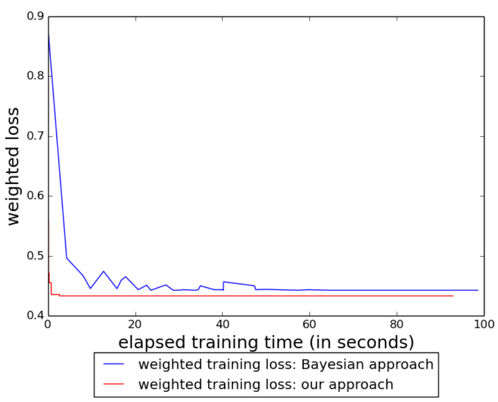
Table 20: Falling rule list for bank-full dataset, trained using the optimization approach (Algorithm FRL) with 3000 iterations and the positive class weight $w = 7$.



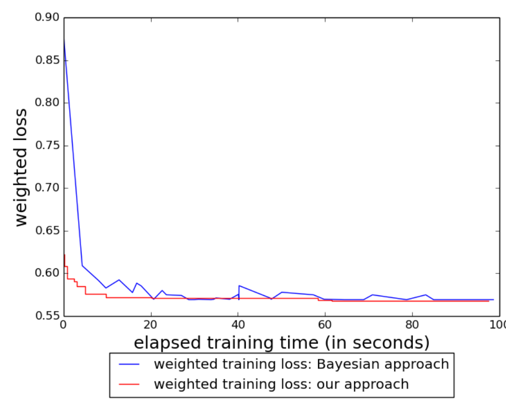
(a) positive class weight $w = 1$



(b) positive class weight $w = 3$



(c) positive class weight $w = 5$



(d) 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

	antecedent		probability	positive support	negative support
IF	poutcome=success AND housing=no	THEN success prob. is	0.70	729	311
ELSE IF	poutcome=success	THEN success prob. is	0.53	249	222
ELSE IF	$60 \leq$ age < 100 AND loan=no	THEN success prob. is	0.29	426	1030
ELSE IF	$17 \leq$ age < 30 AND housing=no	THEN success prob. is	0.25	504	1538
ELSE IF	education=tertiary AND housing=no	THEN success prob. is	0.14	790	4750
ELSE IF	marital=single AND contact=cellular	THEN success prob. is	0.12	648	4754
ELSE IF	$1000 \leq$ balance < 2000 AND housing=no	THEN success prob. is	0.11	135	1061
ELSE IF	campaign=1 AND contact=cellular	THEN success prob. is	0.10	571	4904
ELSE IF	contact=cellular AND loan=no	THEN success prob. is	0.08	587	6800
ELSE		success prob. is	0.04	650	14552

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

	antecedent		probability	positive support	negative support
IF	poutcome=success AND housing=no	THEN success prob. is	0.70	729	311
ELSE IF	poutcome=success AND previous ≥ 2	THEN success prob. is	0.55	185	154
ELSE IF	poutcome=success AND default=no	THEN success prob. is	0.48	64	68
ELSE IF	$60 \leq$ age < 100 AND loan=no	THEN success prob. is	0.29	426	1030
ELSE IF	previous ≥ 2 AND housing=no	THEN success prob. is	0.25	302	921
ELSE IF	$17 \leq$ age < 30 AND housing=no	THEN success prob. is	0.24	444	1413
ELSE IF	education=tertiary AND housing=no	THEN success prob. is	0.13	671	4435
ELSE		success prob. is	0.07	2468	31590

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