

## Proofs

For every  $f, g \in \mathcal{F}_n$ , we note  $f \models g$  when for every  $\mathbf{x} \in \{0, 1\}^n$ ,  $f(\mathbf{x}) = 1$  implies that  $g(\mathbf{x}) = 1$ .

### Proof of Proposition 1

**Proof** One first needs the following lemma that gives a recursive characterization of the set  $sr(\mathbf{x}, f)$  of sufficient reasons for an instance  $\mathbf{x}$  given a Boolean classifier  $f$  when  $\mathbf{x}$  is a positive instance of  $f$  (in the case when  $\mathbf{x}$  is a negative instance of  $f$ , just replace  $f$  by  $\neg f$ ).

**Lemma 1** For any Boolean function  $f \in \mathcal{F}_n$  and any instance  $\mathbf{x} \in \{0, 1\}^n$ , the following inductive characterization of  $sr(\mathbf{x}, f)$  holds:

$$\begin{aligned} sr(\mathbf{x}, 1) &= \{1\} \\ sr(\mathbf{x}, 0) &= \{\} \\ sr(\mathbf{x}, f) &= sr(\mathbf{x}, (f \mid \ell) \wedge (f \mid \bar{\ell})) \cup \{\ell \wedge t_\ell : t_\ell \in sr(\mathbf{x}, f \mid \ell) \text{ s.t. } t_\ell \not\models f \mid \bar{\ell}\} \\ &\text{where } Var(\ell) \subseteq Var(f) \text{ and } t_{\mathbf{x}} \models \ell \end{aligned}$$

and

$$sr(\mathbf{x}, (f \mid \ell) \wedge (f \mid \bar{\ell})) = \max(\{t_\ell \wedge t_{\bar{\ell}} : t_\ell \in sr(\mathbf{x}, f \mid \ell), t_{\bar{\ell}} \in sr(\mathbf{x}, f \mid \bar{\ell})\}, \models).$$

**Proof** Let us recall first the following inductive characterization of  $pi(f)$ , the set of prime implicants of  $f \in \mathcal{F}_n$ , based on the Shannon decomposition of  $f$  over any of its variables  $x$  (see e.g., [Brayton *et al.*, 1984]):

$$\begin{aligned} pi(1) &= \{1\} \\ pi(0) &= \{\} \\ pi(f) &= pi((f \mid \bar{x}) \wedge (f \mid x)) \\ &\cup \{\bar{x} \wedge t_{\bar{x}} : t_{\bar{x}} \in pi(f \mid \bar{x}) \text{ s.t. } \nexists t \in pi((f \mid \bar{x}) \wedge (f \mid x)), t_{\bar{x}} \models t\} \\ &\cup \{x \wedge t_x : t_x \in pi(f \mid x) \text{ s.t. } \nexists t \in pi((f \mid \bar{x}) \wedge (f \mid x)), t_x \models t\} \\ &\text{where } x \in Var(f) \end{aligned}$$

and

$$pi((f \mid \bar{x}) \wedge (f \mid x)) = \max(\{t_{\bar{x}} \wedge t_x : t_{\bar{x}} \in pi(f \mid \bar{x}), t_x \in pi(f \mid x)\}, \models).$$

For the base cases  $sr(\mathbf{x}, 1) = \{1\}$  and  $sr(\mathbf{x}, 0) = \{\}$ , the result is obvious. For the general case, taking  $x \in Var(\ell)$ , we have:

$$\begin{aligned} sr(\mathbf{x}, f) &= \{t \in pi(f) : t_{\mathbf{x}} \models t\} \\ &= \{t \in pi((f \mid \bar{x}) \wedge (f \mid x)) : t_{\mathbf{x}} \models t\} \\ &\cup \{t \in \{\bar{x} \wedge t_{\bar{x}} : t_{\bar{x}} \in pi(f \mid \bar{x}) \text{ s.t. } \nexists t' \in pi((f \mid \bar{x}) \wedge (f \mid x)), t_{\bar{x}} \models t'\}, \text{ and } t_{\mathbf{x}} \models t\} \\ &\cup \{t \in \{x \wedge t_x : t_x \in pi(f \mid x) \text{ s.t. } \nexists t' \in pi((f \mid \bar{x}) \wedge (f \mid x)), t_x \models t'\}, \text{ and } t_{\mathbf{x}} \models t\} \\ &= sr(\mathbf{x}, (f \mid \bar{x}) \wedge (f \mid x)) \\ &\cup \{t \in \{\bar{x} \wedge t_{\bar{x}} : t_{\bar{x}} \in pi(f \mid \bar{x}) \text{ s.t. } \nexists t' \in pi((f \mid \bar{x}) \wedge (f \mid x)), t_{\bar{x}} \models t'\}, \text{ and } t_{\mathbf{x}} \models t\} \\ &\cup \{t \in \{x \wedge t_x : t_x \in pi(f \mid x) \text{ s.t. } \nexists t' \in pi((f \mid \bar{x}) \wedge (f \mid x)), t_x \models t'\}, \text{ and } t_{\mathbf{x}} \models t\} \end{aligned}$$

Now, since  $\mathbf{x}$  is an instance, whatever  $\ell$ , it cannot be the case that  $t_{\mathbf{x}} \models \ell$  and  $t_{\mathbf{x}} \models \bar{\ell}$ . Suppose that  $\ell = x$  (the case  $\ell = \bar{x}$  is similar). In this situation, no element of  $\{\bar{x} \wedge t_{\bar{x}} : t_{\bar{x}} \in pi(f \mid \bar{x}) \text{ s.t. } \nexists t \in pi((f \mid \bar{x}) \wedge (f \mid x)), t_{\bar{x}} \models t\}$ , and  $t_{\mathbf{x}} \models t\}$  can belong to  $sr(\mathbf{x}, f)$ . As a consequence, we get that:

$$\begin{aligned} sr(\mathbf{x}, f) &= sr(\mathbf{x}, (f \mid \bar{x}) \wedge (f \mid x)) \\ &\cup \{t \in \{\ell \wedge t_\ell : t_\ell \in pi(f \mid \ell) \text{ s.t. } \nexists t' \in pi((f \mid \bar{\ell}) \wedge (f \mid \ell)), t_\ell \models t'\}, \text{ and } t_{\mathbf{x}} \models t\} \\ &\text{where } Var(\ell) \subseteq Var(f) \text{ and } t_{\mathbf{x}} \models \ell \end{aligned}$$

If  $t = \ell \wedge t_\ell$  is such that  $t_{\mathbf{x}} \models t$  holds, then we have  $t_{\mathbf{x}} \models t_\ell$ . Hence, we have:

$$\begin{aligned} sr(\mathbf{x}, f) &= sr(\mathbf{x}, (f \mid \bar{x}) \wedge (f \mid x)) \\ &\cup \{\ell \wedge t_\ell : t_\ell \in sr(\mathbf{x}, f \mid \ell) \text{ s.t. } \nexists t' \in pi((f \mid \bar{\ell}) \wedge (f \mid \ell)), t_\ell \models t'\} \\ &\text{where } Var(\ell) \subseteq Var(f) \text{ and } t_{\mathbf{x}} \models \ell \end{aligned}$$

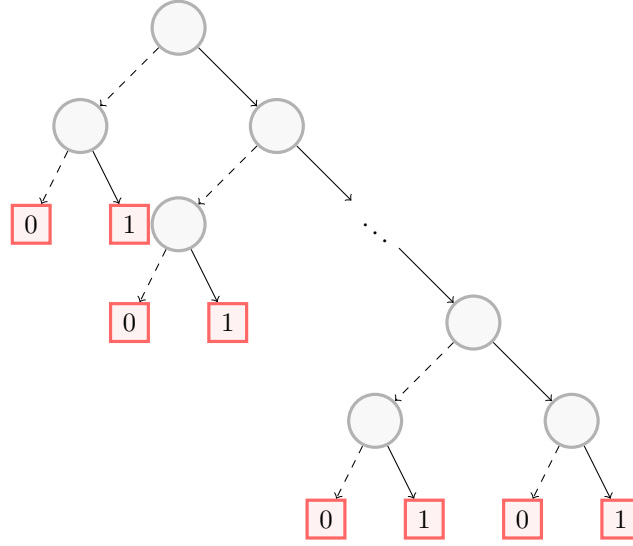
Consider now the condition  $\exists t' \in pi((f \mid \bar{\ell}) \wedge (f \mid \ell)), t_\ell \models t'$  and suppose that it is satisfied. Since  $pi((f \mid \bar{\ell}) \wedge (f \mid \ell)) = \max(\{t'_\ell \wedge t''_\ell : t'_\ell \in pi(f \mid \bar{\ell}), t''_\ell \in pi(f \mid \ell)\}, \models)$ , there exist  $t'_\ell \in pi(f \mid \bar{\ell})$  and  $t''_\ell \in pi(f \mid \ell)$  such that  $t' = t'_\ell \wedge t''_\ell$ . Thus, we have  $t_\ell \models t'_\ell \wedge t''_\ell$ , and in particular  $t_\ell \models t'_\ell$  holds. But since  $t_\ell$  and  $t'_\ell$  are prime implicants of  $f \mid \bar{\ell}$ , this implies that  $t_\ell \equiv t'_\ell$  holds. Furthermore, from  $t_\ell \models t'_\ell \wedge t''_\ell$  we get that  $t_\ell \models t''_\ell$ . In addition, a prime implicant  $t'_\ell$  of  $f \mid \bar{\ell}$  such that  $t_\ell \models t'_\ell$  exists if and only if  $t_\ell \models f \mid \bar{\ell}$ . Altogether, the condition  $\exists t' \in pi((f \mid \bar{\ell}) \wedge (f \mid \ell)), t_\ell \models t'$  is equivalent to  $t_\ell \models f \mid \bar{\ell}$ . Thus, we get that:

$$\begin{aligned}
sr(\mathbf{x}, f) &= sr(\mathbf{x}, (f \mid \ell) \wedge (f \mid \bar{\ell})) \\
&\cup \{\ell \wedge t_\ell : t_\ell \in sr(\mathbf{x}, f \mid \ell) \text{ s.t. } t_\ell \not\models f \mid \bar{\ell}\} \\
&\text{where } Var(\ell) \subseteq Var(f) \text{ and } t_{\mathbf{x}} \models \ell
\end{aligned}$$

Finally, if  $t \in \max(\{t_{\bar{x}} \wedge t_x : t_{\bar{x}} \in pi(f \mid \bar{x}), t_x \in pi(f \mid x)\}, \models)$ , then by construction  $t$  is such that there exist  $t_{\bar{x}} \in pi(f \mid \bar{x})$  and  $t_x \in pi(f \mid x)$  satisfying  $t = t_{\bar{x}} \wedge t_x$ . If  $t_{\mathbf{x}} \models t$  holds, then  $t_{\mathbf{x}} \models t_{\bar{x}}$  and  $t_{\mathbf{x}} \models t_x$  hold. Hence  $t_{\bar{x}} \in sr(\mathbf{x}, f \mid \bar{x})$  and  $t_x \in sr(\mathbf{x}, f \mid x)$ . Consequently,  $t \in \max(\{t_{\bar{x}} \wedge t_x \mid t_{\bar{x}} \in sr(\mathbf{x}, f \mid \bar{x}), t_x \in sr(\mathbf{x}, f \mid x)\}, \models)$ .  $\square$

From the inductive characterization of  $sr(\mathbf{x}, f)$  given by the previous proposition, we can easily derive a bottom-up algorithm allowing to derive  $sr(\mathbf{x}, f)$  when  $f$  is represented by a decision tree.

Consider now a decision tree  $T$  of depth  $k \geq 1$  having the following form:



$T$  has  $2k - 1$  decision nodes and  $2k$  leaves. Suppose that the variables associated with the decision nodes are in one-to-one correspondence with the decision nodes (i.e., they are all distinct). The number of variables occurring in  $T$  is thus  $n = 2k - 1$ , therefore  $T$  has  $2n + 1$  nodes. Consider now the instance  $\mathbf{x} \in \{0, 1\}^n$  such that  $x_i = 1$  for every  $i \in [n]$ . We are going to prove by induction on the depth  $k$  of such a tree  $T$  that  $\mathbf{x}$  has  $2^{k-1}$  minimal reasons given  $T$ , each of them containing  $k$  literals. The proof takes advantage of the recursive characterization of the set of all sufficient reasons for an instance given a decision tree, as made precise by Lemma 1.

- Base case  $k = 1$ . We have  $n = 1$ .  $T$  consists of a decision node labelled by the single variable of  $X_n$ , say  $x$ , a left child that is a 0-leaf and a right child that is a 1-leaf.  $T$  is equivalent to  $x$  and  $x$  is implied by  $t_{\mathbf{x}}$ . Hence,  $x$  is the unique sufficient reason for  $\mathbf{x}$  given  $T$ , so it is also the unique minimal reason for  $\mathbf{x}$  given  $T$ . As expected, the number of minimal reasons for  $\mathbf{x}$  given  $T$  is equal to  $2^{k-1}$ . The size of the unique minimal reason is  $k = 1$ .
- Inductive step  $k > 1$ . Let  $x$  be the variable of  $X_n$  labelling the root node of  $T$ . By construction, the left child  $T_l$  of  $T$  is equivalent to a single variable, say  $x_l$ , that is the unique minimal reason for  $\mathbf{x}$  given  $T_l$ . The right child  $T_r$  of  $T$  has the same form as  $T$ , but with depth  $k - 1$ . By induction hypothesis, we know that  $\mathbf{x}$  has  $2^{k-2}$  minimal reasons given  $T_r$ , each of them containing  $k - 1$  literals. As shown by Lemma 1, provided that the variables labelling the decision nodes are pairwise distinct, the minimal reasons for  $\mathbf{x}$  given  $T$  are obtained by extending every minimal reason for  $\mathbf{x}$  given  $T_r$  with  $x_l$  and by extending every minimal reason for  $\mathbf{x}$  given  $T_r$  with  $x$ . Accordingly,  $\mathbf{x}$  has  $2 \times (2^{k-2}) = 2^{k-1}$  minimal reasons given  $T$  and each of them contains  $k - 1 + 1 = k$  literals.

Finally, since  $n = 2k - 1$ , we have  $k = \frac{n+1}{2}$  and the number of minimal reasons for  $\mathbf{x}$  given  $T$  is equal to  $2^{k-1} = 2^{\sqrt{n-1}}$ .  $\square$

## Proof of Proposition 2

**Proof** The decision boils down to checking whether the term  $t_{\mathbf{x}}^S$  defined as the subset of  $t_{\mathbf{x}}$  over  $S$  is an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees of  $F$ , and this can be tested in polynomial time since the language of decision trees offers the implicant query (see e.g., [Audemard *et al.*, 2020]). If the test is positive, the greedy algorithm presented in the paper can be used to extract a majoritary reason for  $\mathbf{x}$  given  $F$  by starting with  $t_{\mathbf{x}}^S$  instead of starting with  $t_{\mathbf{x}}$ .  $\square$

### Proof of Proposition 3

**Proof** A majoritary reason for  $x$  given  $F$  that satisfies  $C$  exists precisely when  $t_x \models C$ , which can be tested in time  $O(n + |F|)$ . If the test is positive and  $C$  belongs to propositional fragment offering a polynomial-time implicant test (e.g.,  $C$  is a CNF formula), deriving such a reason is tractable as well: the greedy algorithm can be leveraged (at each step, it is enough to test whether the current term is an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees of  $F$ , and is still an implicant of  $C$ ).  $\square$

### Proof of Proposition 4

**Proof** Let  $<$  be any linear ordering over  $X_n$  that extends  $\leq$ .  $X_n$  ordered by  $<$  can be obtained by topologically sorting the graph over  $X_n$  associated with  $\leq$ . Let  $F = \{T_1, \dots, T_m\}$  and  $x$  be such that  $F(x) = 1$  (the case when  $F(x) = 0$  can be handled in the same way by considering a random forest equivalent to  $\neg F$  instead of  $F$ ; such a random forest equivalent to  $\neg F$  can be computed in linear-time in the size of  $F$ , see Proposition 1 in [Audemard *et al.*, 2021]). Let us run the greedy algorithm (described in the paper) on  $x$  and  $F = \{T_1, \dots, T_m\}$ , where instead of trying to eliminate the literals associated with the characteristics of  $x$  in any fixed, yet arbitrary way, sort them first in ascending order w.r.t.  $<$  (i.e., the characteristics associated with the less prioritized features are considered first). Let  $t$  be the resulting term. Then  $t$  is an inclusion-preferred majoritary reason for  $x$  given  $F$ . Indeed, towards a contradiction, suppose that this is not the case. This means that there exists a term  $t'$  such that  $t_x \models t'$  and  $t' \sqsubset t$ , i.e.,  $\exists i \in \{1, \dots, p\} \forall j \in \{1, \dots, i-1\}, t'[S_j] = t[S_j]$  and  $t'[S_i] \subset t[S_i]$ . Thus there exists a literal  $\ell \in t$  such that  $\text{var}(\ell) = \{x\}$ ,  $x \in S_i$ ,  $\ell \notin t'$ , and for every literal based on a variable that precedes  $x$  in the enumeration given by  $<$ ,  $t$  and  $t'$  coincide.

Now, let  $t_{<x}$  (resp.  $t'_{<x}$ ) be the conjunction of literals of  $t$  (resp.  $t'$ ) based on variables that precede  $x$  in the enumeration given by  $<$  and let  $t''$  be the conjunction of all the literals  $\ell''$  based on variables that are after  $x$  in the enumeration given by  $<$  and such that  $x$  satisfies  $\ell''$ . Since  $t'$  is an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees of  $F$  and satisfies  $t_x \models t'$ , and since  $t'_{<x} \wedge t'' \models t'$  holds,  $t'_{<x} \wedge t''$  is an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees of  $F$  such that  $t_x \models t'_{<x} \wedge t''$ . But  $t'_x = t_{<x}$ , hence  $t_{<x} \wedge t''$  is an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees of  $F$  such that  $x \models t_{<x} \wedge t''$ . This conflicts with the fact that  $\ell \in t$  since  $\ell$  being kept by the greedy algorithm implies that  $t_{<x} \wedge t''$  is not an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees of  $F$ .  $\square$

### Proof of Proposition 5

**Proof** First of all, one can easily check that, by construction, (1) an assignment  $v$  over  $X_n \cup Y$  satisfies  $C_{\text{hard}}$  if and only if  $t_v \cap t_x$  is an implicant of more than  $\frac{m}{2}$  trees of  $F$  and  $t_v[Y]$  is an implicant of  $\text{CNF}(\sum_{i=1}^m y_i > \frac{m}{2})$ . Here,  $Y$  contains  $\{y_1, \dots, y_m\}$  as a subset, plus additional variables used for the CNF encoding of the cardinality constraint  $\sum_{i=1}^m y_i > \frac{m}{2}$ .

Now, we prove that (2) if  $v$  satisfies  $C_{\text{hard}}$  and  $v$  maximizes the sum of the weights of clauses from  $C_{\text{soft}}$  that are satisfied, then  $t_v \cap t_x$  is an implicant of more than  $\frac{m}{2}$  trees of  $F$  and  $\forall \ell \in t_v \cap t_x, (t_v \cap t_x) \setminus \{\ell\}$  does not satisfy this property. From (1) one knows that  $t_v \cap t_x$  is an implicant of more than  $\frac{m}{2}$  trees of  $F$ . Towards a contradiction, suppose that there exists  $\ell \in t_v \cap t_x$ , such that  $(t_v \cap t_x) \setminus \{\ell\}$  is an implicant of more than  $\frac{m}{2}$  trees of  $F$ . Then let  $v'$  be the assignment over  $X_n \cup Y$  such that  $v$  and  $v'$  coincide on every variable except the one of  $\ell$ , and  $\ell \in t_v$  while  $\bar{\ell} \in t_{v'}$ . Then if  $(t_v \cap t_x) \setminus \{\ell\}$  is an implicant of more than  $\frac{m}{2}$  trees of  $F$ , this is also the case for  $(t_{v'} \cap t_x) \setminus \{\ell\}$  because  $(t_{v'} \cap t_x) \setminus \{\ell\} = (t_v \cap t_x) \setminus \{\ell\}$ . Now, since  $t_{v'}[Y] = t_v[Y]$ , using (1) we get that  $v'$  satisfies  $C_{\text{hard}}$ . Moreover, by construction,  $w(v') = w(v) + w(\text{var}(\bar{\ell}))$ . Since  $w(\text{var}(\bar{\ell})) \geq 1$ ,  $v$  is not an assignment that maximizes the sum of the weights of clauses from  $C_{\text{soft}}$  that are satisfied, among those satisfying  $C_{\text{hard}}$ .

Finally, it remains to show that (3) if  $v$  is a solution of the WEIGHTED PARTIAL MAXSAT instance given by  $(C_{\text{soft}}, C_{\text{hard}})$ , then  $t_v \cap t_x$  is a majoritary reason for  $x$  given  $F$  that is of minimal weight. From (2) we know that  $t_v \cap t_x$  is a majoritary reason for  $x$  given  $F$ , hence it remains to show that it is of minimal weight. Towards a contradiction, suppose that there exists a majoritary reason  $t$  for  $x$  given  $F$  such that  $w(t) < w(t_v \cap t_x)$ . Let  $v'$  be any assignment  $v$  over  $X_n \cup Y$  that satisfies  $C_{\text{hard}}$  and is such that  $t \subseteq t_{v'}$ ,  $\bar{\ell} \in t_{v'}$  whenever  $\text{var}(\ell) \in X_n$  and  $\ell \notin t$ ,  $y_i \in t_{v'}$  whenever  $t$  satisfies every  $c[x]$  where  $c \in \text{CNF}(T_i)$  and  $i \in [m]$ , and finally  $\bar{y}_i \in t_{v'}$  whenever  $t$  does not satisfy every  $c[x]$  where  $c \in \text{CNF}(T_i)$  and  $i \in [m]$ . Such a  $v'$  exists since  $t = t_{v'} \cap t_x$  satisfies more than  $\frac{m}{2}$  trees of  $F$  (as a consequence,  $t_{v'}[\{y_1, \dots, y_m\}]$  satisfies  $\sum_{i=1}^m y_i > \frac{m}{2}$ ). It remains to compare  $w(v')$  to  $w(v)$ . Let  $t'$  be any term such that  $t' \subseteq t_x$ . We have  $w(t') = \sum_{i=1}^n w(x_i) - \sum_{x_i \notin \text{var}(t')} w(x_i)$ .  $\sum_{i=1}^n w(x_i)$  is a constant (independent of  $t'$ ). Hence, if  $w(t) < w(t_v \cap t_x)$  holds, then we have  $\sum_{x_i \notin \text{var}(t)} w(x_i) > \sum_{x_i \notin \text{var}(t_v \cap t_x)} w(x_i)$ . This conflicts with the fact that  $v$  is a solution of  $(C_{\text{soft}}, C_{\text{hard}})$ .  $\square$