Appendix: Proofs

Proof of Proposition 1

Proof. It is easy to verify that each classification rule $R = \varphi_X \Rightarrow y$ (resp. $R = \varphi_X \Rightarrow \overline{y}$) satisfies $R(y) \equiv \top$ and $R(\overline{y}) \equiv \neg\varphi_X$ (resp. $R(\overline{y}) \equiv \top$ and $R(y) \equiv \neg\varphi_X$). It is then sufficient to replace $R(y)$ and $R(\overline{y})$ by the equivalent expressions above in the definition of $\varSigma \star T$ and to simplify the result obtained to obtain the characterization given in the proposition.

Proof of Proposition 2

Proof. The proof is based on three lemmas. The first lemma shows that the classification rules that can be deduced from a classification circuit on $X \cup \{y\}$ are never conflicting (two classification rules are conflicting whenever they have consistent premises and distinct conclusions).

Lemma 1. *Let $\varSigma = \varSigma_X \Leftrightarrow y$ be a classification circuit on $X \cup \{y\}$. Two classification rules $R_1 = \varphi^1_X \Rightarrow y$ and $R_2 = \varphi^2_X \Rightarrow \overline{y}$ that can be deduced from $\varSigma$ are never conflicting.*

Proof. $R_1 = \varphi^1_X \Rightarrow y$ can be deduced from $\varSigma$ iff $\varSigma \wedge \varphi^1_X \wedge \overline{y}$ is contradictory iff $(\varSigma_X \Leftrightarrow y) \wedge \varphi^1_X \wedge \overline{y}$ is contradictory iff $\overline{\varSigma_X} \wedge \varphi^1_X$ is contradictory iff we have $\varphi^1_X \models \varSigma_X$. Similarly, $R_2 = \varphi^2_X \Rightarrow \overline{y}$ can be deduced from $\varSigma$ iff $\varSigma \wedge \varphi^2_X \wedge y$ is contradictory iff $(\varSigma_X \Leftrightarrow y) \wedge \varphi^2_X \wedge y$ is contradictory iff $\varSigma_X \wedge \varphi^2_X$ is contradictory iff we have $\varphi^2_X \models \overline{\varSigma_X}$. Consequently, $\varphi^1_X \wedge \varphi^2_X$ is necessarily contradictory.

The second lemma shows that if R_1 and R_2 are two classification rules over y (resp. two classification rules over $\overline{y}$), then rectifying a classifier $\varSigma$ by R_1 first and by R_2 then is equivalent to rectifying $\varSigma$ by the conjunction $R_1 \wedge R_2$ of the two rules.

Lemma 2. *Let $\varSigma = \varSigma_X \Leftrightarrow y$ be a classification circuit on $X \cup \{y\}$. Let R_1 and R_2 be two classification rules over y or two classification rules over $\overline{y}$. We have $\varSigma \star (R_1 \wedge R_2) \equiv (\varSigma \star R_1) \star R_2$.*

Proof. Let us first assume that $R_1 = \varphi^1_X \Rightarrow y$ and $R_2 = \varphi^2_X \Rightarrow y$ are two classification rules over y. We have $R_1 \wedge R_2 \equiv (\varphi^1_X \vee \varphi^2_X) \Rightarrow y$, showing that $R_1 \wedge R_2$ is equivalent to the classification rule $\varphi_X \Rightarrow y$ over y, with $\varphi_X \equiv (\varphi^1_X \vee \varphi^2_X)$. Next, we exploit the fact that $\varSigma^{R_1 \wedge R_2}_X$ can be simplified to $\varSigma_X \vee \varphi^1_X \vee \varphi^2_X$. Furthermore, we have $\varSigma^{R_1}_X \equiv \varSigma_X \vee \varphi^1_X$ and $\varSigma^{R_1,R_2}_X \equiv \varSigma^{R_1}_X \vee \varphi^2_X$. Since $\varSigma^{R_1}_X \equiv \varSigma_X \vee \varphi^1_X$, we have $\varSigma^{R_1 \wedge R_2}_X \equiv \varSigma^{R_1,R_2}_X$, which concludes the proof.

Similarly, suppose that $R_1 = \varphi^1_X \Rightarrow \overline{y}$ and $R_2 = \varphi^2_X \Rightarrow \overline{y}$ are two classification rules over $\overline{y}$. We have $R_1 \wedge R_2 \equiv (\varphi^1_X \vee \varphi^2_X) \Rightarrow \overline{y}$, showing that $R_1 \wedge R_2$ is equivalent to the classification rule $\varphi_X \Rightarrow \overline{y}$ over $\overline{y}$, with $\varphi_X \equiv (\varphi^1_X \vee \varphi^2_X)$. Next, we exploit the fact that $\varSigma^{R_1 \wedge R_2}_X$ can be simplified to $\varSigma_X \wedge \neg(\varphi^1_X \vee \varphi^2_X)$, which is equivalent to $\varSigma_X \wedge \neg\varphi^1_X \wedge \neg\varphi^2_X$. On the other hand, we have $\varSigma^{R_1}_X \equiv \varSigma_X \wedge \neg\varphi^1_X$ and $\varSigma^{R_1,R_2}_X \equiv \varSigma^{R_1}_X \wedge \neg\varphi^2_X$. Since $\varSigma^{R_1}_X \equiv \varSigma_X \wedge \neg\varphi^1_X$, we have $\varSigma^{R_1 \wedge R_2}_X \equiv \varSigma^{R_1,R_2}_X$, which concludes the proof.

Since conjunction is commutative, the resulting circuit is equivalent to Σ rectified by R_2 first and by R_1 then, that is, we have $\Sigma \star (R_1 \wedge R_2) \equiv (\Sigma \star R_2) \star R_1$. In other words, the rectification order does not matter.

The third lemma concerns classification rules having contradictory premises and contradictory conclusions (y and $\overline{y}$). For such rules R_1 and R_2, here again, rectifying a classification circuit Σ by the conjunction $R_1 \wedge R_2$ amounts to rectify Σ by R_1 first and by R_2 then. And since conjunction is commutative, the rectification order actually does not matter.

Lemma 3. *Let $\Sigma = \Sigma_X \Leftrightarrow y$ be a classification circuit on $X \cup \{y\}$. Let $R_1 = \varphi^1_X \Rightarrow y$ and $R_2 = \varphi^2_X \Rightarrow \overline{y}$ be two classification rules such that $\varphi^1_X \wedge \varphi^2_X$ is contradictory. We have $\Sigma \star (R_1 \wedge R_2) \equiv (\Sigma \star R_1) \star R_2$.*

Proof. On the one hand, we have

$$\Sigma_X^{R_1 \wedge R_2} \equiv (\Sigma_X \wedge \neg((R_1 \wedge R_2)(\overline{y}) \wedge \neg(R_1 \wedge R_2)(y))) \vee ((R_1 \wedge R_2)(y) \wedge \neg(R_1 \wedge R_2)(\overline{y})).$$

Since $(R_1 \wedge R_2)(\overline{y})$ is equivalent to $R_1(\overline{y}) \wedge R_2(\overline{y})$ and $(R_1 \wedge R_2)(y)$ is equivalent to $R_1(y) \wedge R_2(y)$, $\Sigma_X^{R_1 \wedge R_2}$ is equivalent to

$$(\Sigma_X \wedge \neg(R_1(\overline{y}) \wedge R_2(\overline{y}) \wedge \neg(R_1(y) \wedge R_2(y)))).$$

Now, we exploit the facts that $R_1(y) \equiv R_2(\overline{y}) \equiv \top$, that $R_1(\overline{y}) \equiv \neg\varphi^1_X$, and that $R_2(y) \equiv \neg\varphi^2_X$ to simplify the previous formula. We obtain that

$$\Sigma_X^{R_1 \wedge R_2} \equiv (\Sigma_X \wedge \neg(\neg\varphi^1_X \wedge \varphi^2_X)) \vee (\varphi^1_X \wedge \neg\varphi^2_X).$$

Since $\varphi^1_X \wedge \varphi^2_X$ is contradictory, we have $\neg\varphi^1_X \wedge \varphi^2_X \equiv \varphi^2_X$ and $\varphi^1_X \wedge \neg\varphi^2_X \equiv \varphi^1_X$. Thus, $\Sigma_X^{R_1 \wedge R_2}$ is equivalent to $(\Sigma_X \wedge \neg\varphi^2_X) \vee \varphi^1_X$.

On the other hand, we have $\Sigma_X^{R_1} \equiv \Sigma_X \vee \varphi^1_X$ since R_1 is a classification rule over y. So, given that R_2 is a classification rule over $\overline{y}$, we have $\Sigma_X^{R_1,R_2} \equiv \Sigma_X^{R_1} \wedge \neg\varphi^2_X \equiv (\Sigma_X \vee \varphi^1_X) \wedge \neg\varphi^2_X \equiv (\Sigma_X \wedge \neg\varphi^2_X) \vee (\varphi^1_X \wedge \neg\varphi^2_X)$. Since $\varphi^1_X \wedge \neg\varphi^2_X \equiv \varphi^1_X$, this last formula is equivalent to $(\Sigma_X \wedge \neg\varphi^2_X) \vee \varphi^1_X$. Therefore, $\Sigma_X^{R_1 \wedge R_2}$ is equivalent to $\Sigma_X^{R_1,R_2}$, and this concludes the proof.

Lemma 1 shows that we can take advantage of Lemma 3 whenever two rules $R_1 = \varphi^1_X \Rightarrow y$ and $R_2 = \varphi^2_X \Rightarrow \overline{y}$ are deduced from a classification circuit $\Phi_X \Leftrightarrow y$ where Φ_X is a binary classifier. Indeed, Lemma 1 ensures that for such rules, $\varphi^1_X \wedge \varphi^2_X$ is contradictory (otherwise, the two rules would conflict).

Finally, a simple induction on k can be used to obtain the desired result from Lemma 2 and Lemma 3.

Proof of Proposition 3

Proof. Let $R = t \Rightarrow y$. If t is an abductive explanation for $\boldsymbol{x}$ given C, then we have $t \models C_X$. Equivalently, $\neg C_X \models \neg t$ holds. Since $y \notin X$, this is equivalent to $\neg C_X \vee y \models \neg t \vee y$, i.e., $C_X \Rightarrow y \models t \Rightarrow y$. Since $\Phi \models C_X \Rightarrow y$, we have $\Phi \models R$. The case when $R = t \Rightarrow \overline{y}$ is similar.

Proof of Proposition 4

Proof. By construction, $(I_X \Leftrightarrow y) \star R$ is a classification circuit that classifies every instance $\boldsymbol{x}' \in \boldsymbol{X}$ as the classification circuit $I_X \Leftrightarrow y$, except those instances $\boldsymbol{x}'$ such that $I(\boldsymbol{x}') \neq R(\boldsymbol{x}')$, which are classified by $(I_X \Leftrightarrow y) \star R$ in the same way as they are classified by R [16]. Let us consider any instance $\boldsymbol{x}' \in \boldsymbol{X}^{\pm}_{I^R}$. Then $\boldsymbol{x}'$ is not covered by R, otherwise we would have $I^R(\boldsymbol{x}') = P(\boldsymbol{x}')$ given that R is implied by the classification circuit $P_X \Leftrightarrow y$. Therefore, $\boldsymbol{x}'$ is classified by I^R in the same way as it is classified by I, so that we also have $\boldsymbol{x}' \in \boldsymbol{X}^{\pm}_I$. To prove that the inclusion $X^{\pm}_{I^R} \subset X^{\pm}_I$ is strict, it is enough to observe that R covers at least one instance that belongs to $X^{\pm}_I$, namely the instance $\boldsymbol{x}$ that triggered the correction step. This instance is classified by I^R in the same way as it is classified by R, thus in the same way as it is classified by P since R is implied by the classification circuit $P_X \Leftrightarrow y$. Accordingly, $\boldsymbol{x} \notin X^{\pm}_{I^R}$, which completes the proof.