Appendix A. Proof of convergence

Hence, in what follows we show that solving the optimization problem in Eq. 2 equals solving the same problem in Eq. 5 in the limit where i approaches infinity. To do this, we first prove that the sequence of functions F_i converges to $F(\theta)$ for increasing values of i, *i.e.* that $\lim_{i\to\infty} F_i(\theta) = F(\theta)$. Then, we also prove that from this result it derives that the sequence of θ_i^* converges to θ^* as i approaches infinity. We do so under several assumptions.

Uniform convergence for F_i

Let us introduce the following proposition on the uniform convergence of functions.

Proposition 3 *A sequence of functions $f_i : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$ is uniformly convergent to a limit function $f : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$, if and only if*

$$||f - f_i||_\infty \xrightarrow{i\to\infty} 0$$

where $||\cdot||_\infty$ denotes the supremum norm of the functions $f_i - f$ on D [1].

On this basis, we propose the following theorem.

Theorem 4 *A sequence of functions $\big\{F_i\big\}_i$ defined as $F_i(\theta) = \sum_{z\in S_i} \mathcal{P}(\theta|f(z), g)$, uniformly converges to $F(\theta) = \sum_{z\in S} \mathcal{P}(\theta|f(z), g)$.*

Proof Let us start by taking the supremum norm

$$\begin{aligned} ||F(\theta) - F_i(\theta)||_\infty &= \sup_{\theta\in\Theta} \Big\{ \Big| F(\theta) - F_i(\theta) \Big| \Big\} \\ &= \sup_{\theta\in\Theta} \Big\{ \Big| \sum_{z\in S} \mathcal{P}(\theta|f(z), g) - \sum_{z\in S_i} \mathcal{P}(\theta|f(z), g) \Big| \Big\} \\ &= \sup_{\theta\in\Theta} \Big\{ \Big| \sum_{z\in S\setminus S_i} \mathcal{P}(\theta|f(z), g) \Big| \Big\} \end{aligned}$$

Since $\mathcal{P}(\theta|f(z), g)$ is a probability function, it holds that $0 \leq \mathcal{P}(\theta|f(z), g) \leq 1$. Hence, the norm is bounded by the equation below

$$\begin{aligned} ||F(\theta) - F_i(\theta)||_\infty &= \sup_{\theta\in\Theta} \Big\{ \Big| \sum_{z\in S\setminus S_i} \mathcal{P}(\theta|f(z), g) \Big| \Big\} \\ &\leq \sup_{\theta\in\Theta} \Big\{ \sum_{z\in S\setminus S_i} \Big| \mathcal{P}(\theta|f(z), g) \Big| \Big\} \leq \Big| S \setminus S_i \Big| \end{aligned}$$

for $\Big| S \setminus S_i \Big|$ the cardinality of the set $S \setminus S_i$. As discussed before, by definition S_i converges to S for large values of i. According to the proposition above, therefore, we can prove that

$$||F(\theta) - F_i(\theta)||_\infty \xrightarrow{i\to\infty} 0$$

[1] https://www.bookofproofs.org/branches/supremum-norm-and-uniform-convergence/proof/

From this proof, it follows that when i approaches infinity the function $F_i(\theta)$ uniformly converges to $F(\theta)$

$$||F(\theta) - F_i(\theta)||_\infty \xrightarrow{i\to\infty} 0 \implies F_i(\theta) \rightrightarrows F(\theta) \tag{9}$$

■

As a consequence of this uniform convergence, two properties of the function F naturally arise. Firstly, $F_i(\theta)$ converges point-wise to $F(\theta)$. Secondly, $F(\theta)$ is a continuous function on Θ.

Parameter convergence

Let us now focus on the convergence of the copy parameters θ. As previously introduced, we can define the optimal copy parameters for a given value of i, and its corresponding function F_i, according to the following equation

$$\theta_i^* = \arg\max_{\theta\in\Theta} F_i(\theta) \tag{10}$$

for Θ the complete parameter set. We assume that this set is only well defined if F_i and F have a unique global maximum. This is a commonly made assumption in the literature. In addition, we also assume that Θ is compact.

From the definition of θ_i^* above it follows that

$$F_i(\theta_i^*) \geq F_i(\theta),\ \ \forall\theta \in \Theta,\ \forall i \in \mathbb{N}$$

and using the point-wise convergence of F_i we obtain

$$F(\hat{\theta}^*) \geq F(\theta),\ \ \forall\theta \in \Theta \tag{11}$$

where $\hat{\theta}^*$ is the limit of the sequence $\left(\theta_i^*\right)_i$. As a consequence of the compactness of Θ and the continuity of F, we can conclude that $\hat{\theta}^*$ must exist. Moreover, given our assumption that Θ is *well defined*, we can conclude that

$$\hat{\theta}^* = \theta^* = \arg\max_{\theta} F(\theta). \tag{12}$$

The proof above shows that we can approximate the true optimal parameters by sequentially estimating their value using a sequence of subsets S_i that uniformly converge to S. In other words, it demonstrates the feasibility of the sequential approach. This is a theoretical feasibility. Now, we discuss how this approach can be implemented in practice, by considering different optimizations that ensure that the process can be conducted within reasonable computational and memory requirements.

Appendix B. Proof of convergence with weighted data

Let us introduce confidence values for each data point at each step i as follows

$$z_j^k \longrightarrow \beta_i(z_j^k, \theta_{i-1}^*) = 1 - \rho_i(z_j^k, \theta_{i-1}^*) \tag{13}$$

Note that the weights β are introduced as a complementary value of the confidence ρ for each data point. This definition ensures that we give a higher weight to those samples that have been better learned by the copy. Introducing these weights for each data point we can rewrite, Eq. 4 as

$$F_i(\theta) = \sum_{z \in S_i} \beta_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g). \tag{14}$$

In what follows we study the convergence of F under the new form above. Fortunately, we know that $\beta_i(z)$ must converge to 1 as $i \to \infty$, since the copy converges to the original model. Then, the same limit function can be used to study the convergence of F_i

$$F(\theta) = \sum_{z \in S} \mathcal{P}(\theta | f(z), g)$$

and the convergence $F_i(\theta) \rightrightarrows F(\theta)$ can be proved by a simple modification of theorem 1

Theorem 5 *A sequence of functions $\{F_i\}_i$ defined as $F_i(\theta) = \sum_{z \in S_i} \beta_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g)$, uniformly converges to $F(\theta) = \sum_{z \in S} \mathcal{P}(\theta | f(z), g)$.*

Proof Let us start by taking the supremum norm

$$\begin{aligned} ||F(\theta) - F_i(\theta)||_\infty &= \sup_{\theta \in \Theta} \Big\{ \Big| F(\theta) - F_i(\theta) \Big| \Big\} \\ &= \sup_{\theta \in \Theta} \Big\{ \Big| \sum_{z \in S} \mathcal{P}(\theta | f(z), g) - \sum_{z \in S_i} \beta_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g) \Big| \Big\} \\ &= \sup_{\theta \in \Theta} \Big\{ \Big| \sum_{z \in S \setminus S_i} \mathcal{P}(\theta | f(z), g) + \sum_{z \in S_i} \rho_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g) \Big| \Big\} \end{aligned}$$

Using the triangular inequality we obtain

$$\begin{aligned} ||F(\theta) - F_i(\theta)||_\infty &= \sup_{\theta \in \Theta} \Big\{ \Big| \sum_{z \in S \setminus S_i} \mathcal{P}(\theta | f(z), g) + \sum_{z \in S_i} \rho_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g) \Big| \Big\} \\ &\leq \sup_{\theta \in \Theta} \Big\{ \Big| \sum_{z \in S \setminus S_i} \mathcal{P}(\theta | f(z), g) \Big| + \Big| \sum_{z \in S_i} \rho_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g) \Big| \Big\} \\ &\leq \sup_{\theta \in \Theta} \Big\{ \Big| \sum_{z \in S \setminus S_i} \mathcal{P}(\theta | f(z), g) \Big| \Big\} + \sup_{\theta \in \Theta} \Big\{ \Big| \sum_{z \in S_i} \rho_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g) \Big| \Big\} \end{aligned}$$

Proceeding as we did in the original proof, we can bound the supremum norm by

$$||F(\theta) - F_i(\theta)||_\infty \leq \Big| S \setminus S_i \Big| + \sup_{\theta \in \Theta} \Big\{ \Big| \sum_{z \in S_i} \rho_i(z, \theta^*_{i-1}) \mathcal{P}(\theta | f(z), g) \Big| \Big\}$$

for $\Big| S \setminus S_i \Big|$ the cardinality of the set $S \setminus S_i$. By definition, S_i converges to S for large values of i while $\rho_i \to 0$. According to the proposition we previously prove, as we have show that

$$||F(\theta) - F_i(\theta)||_\infty \xrightarrow{i \to \infty} 0$$

it follows that when i approaches infinity the function $F_i(\theta)$ uniformly converges to $F(\theta)$

$$||F(\theta) - F_i(\theta)||_\infty \xrightarrow{i\to\infty} 0 \implies F_i(\theta) \rightrightarrows F(\theta) \tag{15}$$

■

Having proved the converged of our sequence in the presence of weighted data, we now provide an intuitive interpretation for the parameter β in Eq. 14.

To do so, let us assume that we are at step i and that $g(\cdot, \theta^*_{i-1})$ is the best approximation to the original model considering the points in S_{i-1} in the previous iteration. At the current step i, we feed the copy a new set of N points through S_i. Let us further assume that the copy has a total confidence on its prediction for all the data points in this set except for one, for which it fails to correctly predict the class label. In this situation we can write

$$\rho(z^k_j, \theta^*_{i-1}) = \begin{cases} \rho_{fail} & \text{if } z^k_j = z_{fail} \\ 0 & \text{other} \end{cases} \implies \beta(z^k_j, \theta^*_{i-1}) = \begin{cases} \beta_{fail} & \text{if } z^k_j = z_{fail} \\ 1 & \text{other} \end{cases} \tag{16}$$

which, at step i gives us

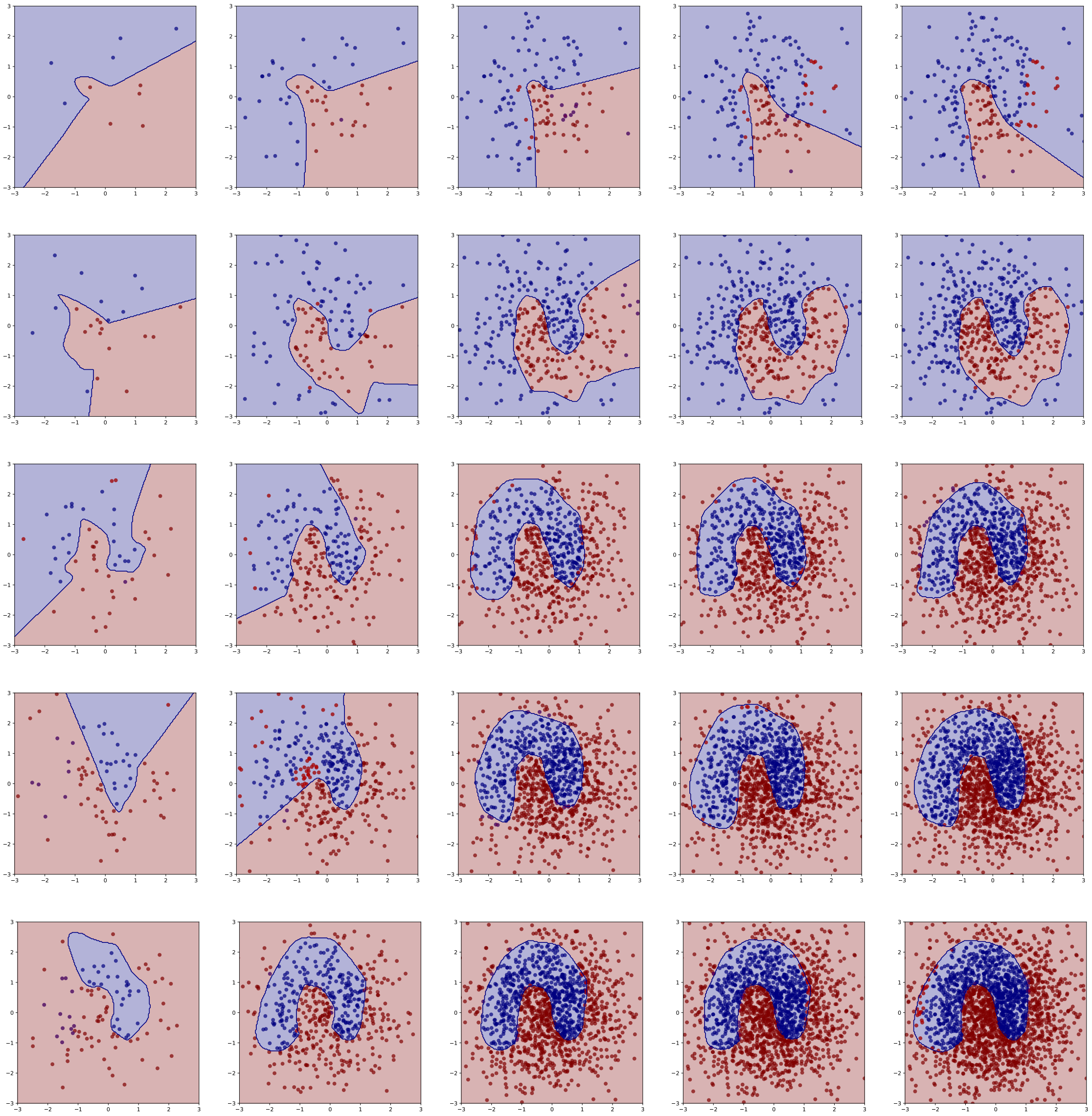
$$F_i(\theta) = \sum_{z \neq z_{fail}} 1 \cdot \mathcal{P}(\theta | f(z), g) + \beta_{fail} \cdot \mathcal{P}(\theta | f(z_{fail}), g) \tag{17}$$

which is the function that needs to be maximized to obtain parameters θ. However, since the model has learned almost all the points with a perfect confidence, the value $\mathcal{P}(\theta|f(z), g)$ in the first term turns to be equals to 1

$$\mathcal{P}(\theta^*_{i-1} | f(z), g) = 1$$

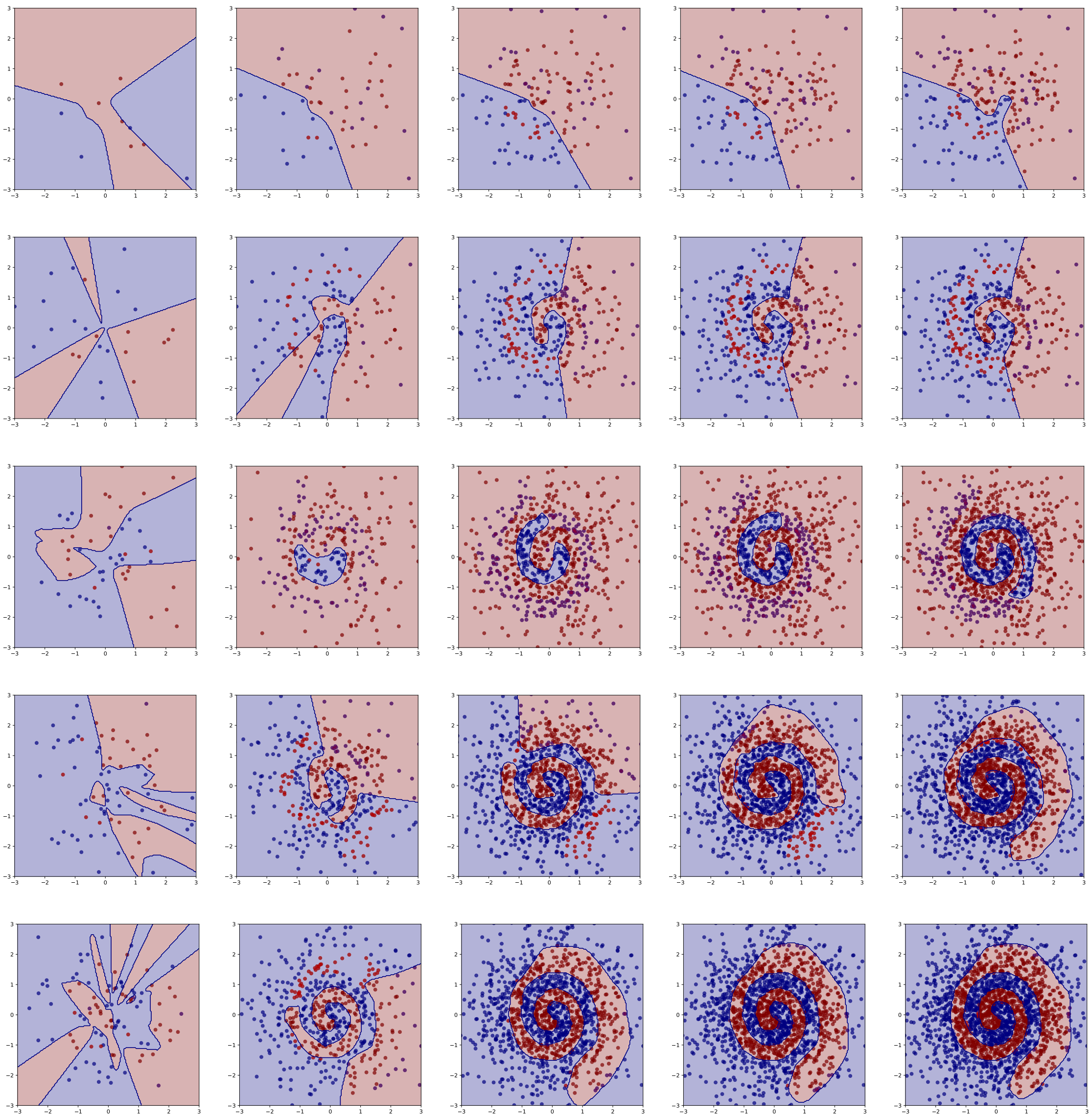
for the optimal parameter in the previous step, so the only way to improve the result, this is, to get an even higher value for the sum, is to find a new optimal parameter that conserves the probabilities of 1 for all the learned points. This is because those points with a larger weight and a slightly decrease on their probability can dramatically affect the value of the total sum. In consequence, the copy can learn new data points but never at the expense of loss knowledge on the points it has already learned.

Appendix C. Toy problems instantaneous decision boundaries



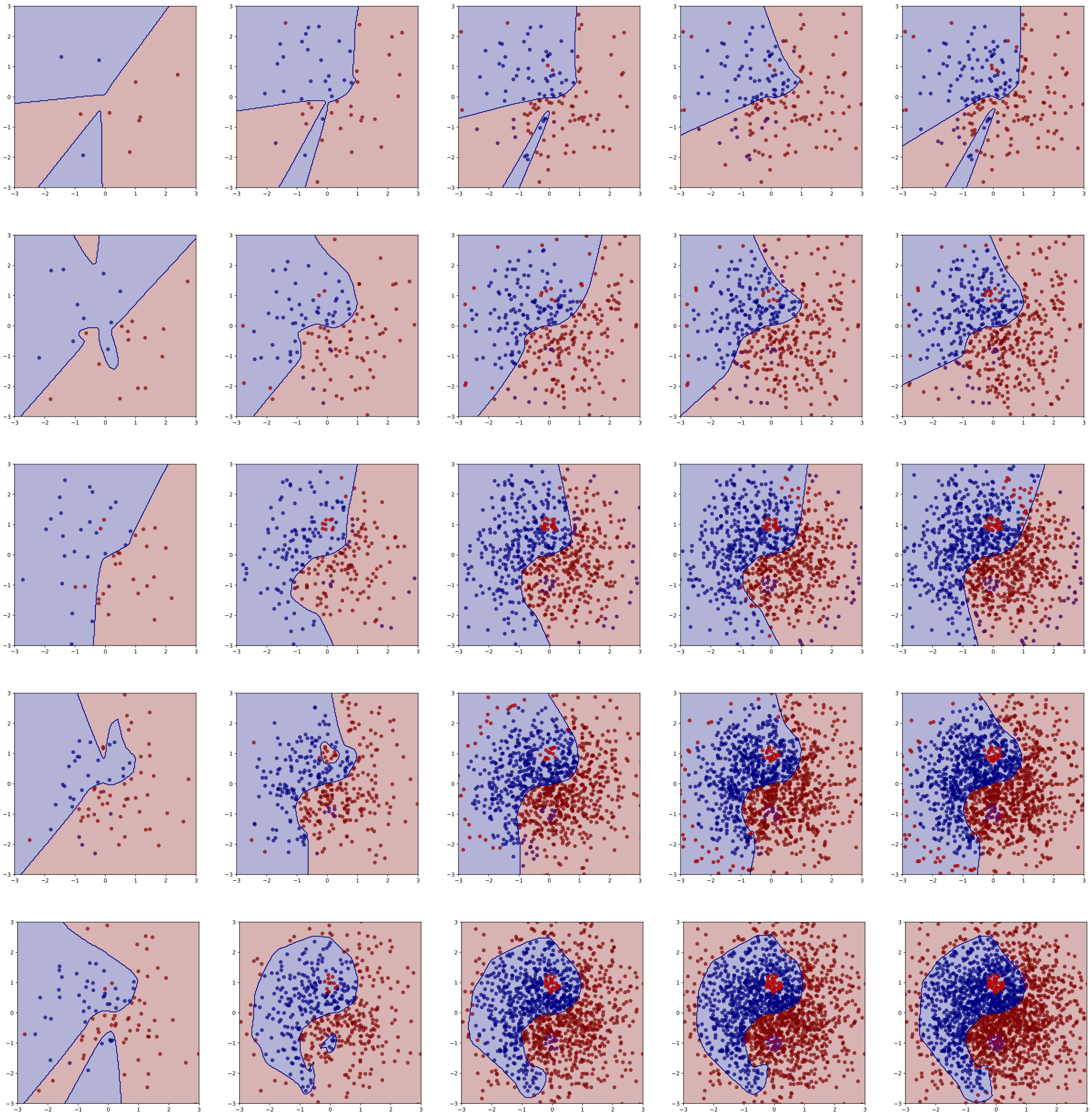
(u) (v) (w) (x) (y)

Figure 10: From top to bottom, instantaneous decision boundaries for $N = 10$, $N = 25$, $N = 50$, $N = 75$ and $N = 100$ at steps (a) 1, (b) 5, (c) 10, (d) 15 and (e) 20 for the *moons* dataset.



(u) (v) (w) (x) (y)

Figure 11: From top to bottom, instantaneous decision boundaries for $N = 10$, $N = 25$, $N = 50$, $N = 75$ and $N = 100$ at steps (a) 1, (b) 5, (c) 10, (d) 15 and (e) 20 for the *spirals* dataset.



(u) (v) (w) (x) (y)

Figure 12: From top to bottom, instantaneous decision boundaries for $N = 10$, $N = 25$, $N = 50$, $N = 75$ and $N = 100$ at steps (a) 1, (b) 5, (c) 10, (d) 15 and (e) 20 for the *yin-yang* dataset.