Proof

We provide proofs for Theorem 1 and Theorem 2 respectively.

Proof 1. For the data point x_u to unlearn, $c_S(x_u)$ is calculated by Equation 8. For any $x \in S(x_u)$, we have $n_S = 1$ according to Lemma 2 and $c_S(x_u) = \mathrm{ReLU}(m_S(x_u)) - \mathrm{ReLU}(-m_S(x_u)) = m_S(x_u)$. Lemma 3 guarantees that there exists $\hat{y}_u \neq y_u$ such that

$$M_U(x_u) = M_p(x_u) \oplus (M_c(x_u) + m_S(x_u)) \to \hat{y}_u$$

Hence, M_U obtained by adding the patch network c_S for x_u satisfies 1) $M_U(x_u) \neq M_{\mathcal{D}}(x_u)$ in Theorem 1. That is, the soundness is proved for a change in the output domain of the model on x_u.

For $x_r \in \mathcal{D}/x_u$, $Q(x_r) = \{a_j x_r \leq b_j\}_{j=1,2,\ldots,N}$ is the linear region where x_r lies. If $Q \cap S = \emptyset$, We can make $n_S(x_r, \lambda) = 0$ by taking a large enough λ. Then, $c_S(x_r) = \mathrm{ReLU}(m_S(x_r) - H) - \mathrm{ReLU}(-m_S(x_r) - H) = 0$. $M_U(x_r) = M_{\mathcal{D}}(x_r)$ in Theorem 1 is satisfied.

Proof 2. For the to-be-forgotten $\mathcal{D}_U$ we cluster into $\mathcal{D}_U^k$ each time, and optimizing m_k for the centroid x_c^k in $\mathcal{D}_U^k$ will result in at least an incorrect prediction of M_U on x_c^k according to Theorem 1. Therefore, in each iteration, we determine that there is

$$\left\{k \in K | x_c^k \in \mathcal{D}_U \wedge x_c^k \notin \mathcal{D}_{UR}\right\}$$

The number of data points in set $\mathcal{D}_{UR}$ is monotonically decreasing. The rate of convergence of multipoint unlearning algorithm can be expressed as

$$\lim_{n \to \infty} \frac{||\mathcal{D}_{UR}^{(n+1)} - \mathcal{D}_{UR}^{*}||}{||\mathcal{D}_{UR}^{(n)} - \mathcal{D}_{UR}^{*}||} < 1$$

where $\mathcal{D}_{UR}^{*} = (1-\delta)\mathcal{D}_U$.

Based on the above convergence analysis, $\{Pr(M_U(x) \neq M_{\mathcal{D}}(x)) \geq \delta | x \in \mathcal{D}\}$ in Theorem 2 can be satisfied.