A The Proof of Theorem 2

Theorem 2. *To satisfy the conditions shown in Eq.* (5)*,* $\boldsymbol{T}(\boldsymbol{z}, y)$ *is derived as:*

$$\boldsymbol{T}_i(\boldsymbol{z}, y) = \begin{cases} z_i, & i \neq y \\ \ln(\frac{1}{K-1}\sum_{j \neq y} e^{z_j}), & i = y \end{cases}$$

where $T_i(\boldsymbol{z}, y)$ *presents the* i*-th element of* $T(\boldsymbol{z}, y)$*.*

Proof. According to Eq. (5), we start with KL-Divergence:

$$\begin{aligned} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) &= \sum_{i=1}^{K} \frac{1}{K} \ln \frac{\frac{1}{K}}{\sigma_i(\boldsymbol{t})} \\ &= \sum_{i=1}^{K} \frac{1}{K} \ln \frac{\frac{1}{K}}{\frac{\boldsymbol{e^{t_i}}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{e^{t_j}}}}. \end{aligned} \tag{14}$$

To simplify the formula, let $a_i = e^{t_i}$, so we can obtain

$$\begin{aligned} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) &= \sum_{i=1}^{K} \frac{1}{K} \ln \frac{\frac{1}{K}}{\sigma_i(\boldsymbol{t})} \\ &= \sum_{i=1}^{K} \frac{1}{K} \ln \frac{\frac{1}{K}}{\frac{\boldsymbol{a_i}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{a_j}}} \\ &= -\ln K - \sum_{i=1}^{K} \frac{1}{K} \ln \frac{a_i}{\sum_{j=1}^{K} a_j} \\ &= -\ln K - \sum_{i \neq y} \frac{1}{K} \ln \frac{a_i}{\sum_{j=1}^{K} a_j} - \frac{1}{K} \ln \frac{a_y}{\sum_{j=1}^{K} a_j}. \end{aligned} \tag{15}$$

Proceed with the derivative of D_{KL} with respect to a_y:

$$\begin{aligned} \frac{\partial}{\partial a_y} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) &= \frac{1}{K} \frac{K-1}{\sum_{j=1}^{K} a_j} - \frac{1}{K} \frac{1}{a_y} \frac{\sum_{j=1}^{K} a_j - a_y}{\sum_{j=1}^{K} a_j} \\ &= \frac{1}{K} \frac{K-1}{\sum_{j=1}^{K} a_j} - \frac{1}{K} \frac{1}{a_y} \frac{\sum_{j \neq y} a_j}{\sum_{j=1}^{K} a_j}. \end{aligned} \tag{16}$$

Now, let $\frac{\partial}{\partial a_y} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) = 0$, we can have $a_y^* = \frac{1}{K-1} \sum_{j \neq y} a_j$. When $a_y < a_y^*$, $\frac{\partial}{\partial a_y} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) < 0$. When $a_y > a_y^*$, $\frac{\partial}{\partial a_y} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) > 0$. So, when $a_y = a_y^*$, $D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p})$ reaches the minimal divergence. Additionally, it satisfies all requirements in Eq. (5). Then we can get:

$$\boldsymbol{T}_y(\boldsymbol{z}, y) = \ln(\frac{1}{K-1} \sum_{j \neq y} e^{z_j}) \tag{17}$$

□

B The Proof of Theorem 3

Theorem 3. *Let $\boldsymbol{z} = \boldsymbol{f}_{PL}(\boldsymbol{x}, \boldsymbol{\theta})$ and z_i denotes the i-th element of $\boldsymbol{z}$. The i-th element of the unlearning target $\boldsymbol{T}(\boldsymbol{z}, s)$ in PLL is expressed as:*

$$\boldsymbol{T}_i(\boldsymbol{z}, s) = \begin{cases} z_i, & i \notin s \\ \ln(\frac{1}{K-|s|}\sum_{j \notin s} e^{z_j}), & i \in s \end{cases}$$

Proof. According to Eq. (15):

$$D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) = -\ln K - \sum_{i \notin s} \frac{1}{K} \ln \frac{a_i}{\sum_{j=1}^{K} a_j} - \sum_{i \in s} \frac{1}{K} \ln \frac{a_i}{\sum_{j=1}^{K} a_j}. \tag{18}$$

Calculate partial derivatives of D_{KL} with respect to a_i, for $i \in s$:

$$\begin{aligned} \frac{\partial}{\partial a_i} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) &= \frac{1}{K} \frac{K-1}{\sum_{j=1}^{K} a_j} - \frac{1}{K} \frac{1}{a_i} \frac{\sum_{j=1}^{K} a_j - a_i}{\sum_{j=1}^{K} a_j} \\ &= \frac{1}{K} \frac{K-1}{\sum_{j=1}^{K} a_j} - \frac{1}{K} \frac{1}{a_i} \frac{\sum_{j \neq y} a_j}{\sum_{j=1}^{K} a_j}. \end{aligned} \tag{19}$$

If we have confidence that the model has learned comprehensively from the dataset and can consistently make correct predictions on the training set, we can directly use Theorem 2. However, in most situations, it is difficult to determine whether the prediction is correct. To be on the safe side, we should consider all labels in the candidate set s. To achieve this, we define:

$$a_i = a_p, \ i \in s. \tag{20}$$

Then with Eq. (19) and Eq. (20), we can derive

$$\frac{1}{K} \frac{K-1}{\sum_{j=1}^{K} a_j} - \frac{1}{K} \frac{1}{a_i} \frac{\sum_{j \neq y} a_j}{\sum_{j=1}^{K} a_j} = \frac{1}{K} \frac{K-1}{\sum_{j=1}^{K} a_j} - \frac{1}{K} \frac{1}{a_p} \frac{(|s|-1)a_p + \sum_{j \notin s} a_j}{\sum_{j=1}^{K} a_j} = 0, \tag{21}$$

$$a_p^* = \frac{1}{K-|s|} \sum_{j \notin s} a_j. \tag{22}$$

When $a_p < a_p^*$, $\frac{\partial}{\partial a_y} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) < 0$. When $a_p > a_p^*$, $\frac{\partial}{\partial a_y} D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p}) > 0$. So, when $a_p = a_p^*$, $D_{KL}(\bar{\boldsymbol{p}} \| \boldsymbol{p})$ reaches the minimal divergence and requirements. Therefore, we can formulate the target as:

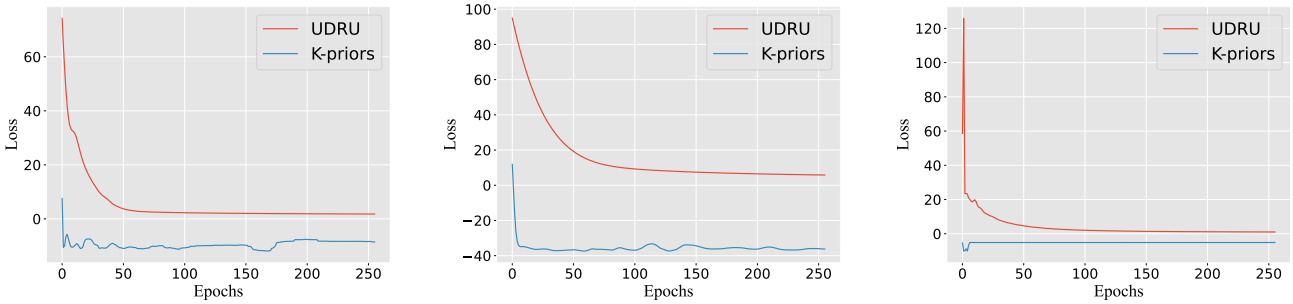
$$\boldsymbol{T}_i(\boldsymbol{z}, s) = \ln(\frac{1}{K-|s|} \sum_{j \notin s} e^{z_j}), \ i \in s. \tag{23}$$

□

C Experiments

The Experiments of Convergence. We conducted experiments to compare the convergence progress of UDRU with K-priors in SL. Both approaches only rely on loss calculated from unlearning data, while the loss of retraining models is based on all remaining data. Amnesiac ML, on the other hand, does not produce loss. Figure 1 illustrates the results of loss convergence on three datasets. The findings indicate that UDRU exhibits smoother convergence, which is faster than K-priors to reach a stable value. In CIFAR10, the starting increase in UDRU's loss curve is attributed to the larger parameter scale of ResNet50, contributing to a sharp rise in regularization at the beginning.

Effect of Different δ. To assess the parameter sensitivity of δ, we assign different orders of magnitude to δ, ranging from 1 to 10^{-4}, to evaluate its effect in SL. Observing the results in Table 6, it is evident that smaller datasets are more sensitive to changes in δ, with a significant drop in accuracy on remaining data when δ is set as 10^{-2} on MNIST. Meanwhile, in Fashion and CIFAR10, accuracy on unlearning data and remaining data needs to be balanced to consider both unlearning and preserving performance. Therefore, we set δ as 1 for MNIST and as 10^{-2} for Fashion and CIFAR10 to achieve more comprehensive performance.



(a) Unlearning loss on MNIST of 256 epochs (b) Unlearning loss on Fashion of 256 epochs (c) Unlearning loss on CIFAR10 of 256 epochs

Figure 1: Comparison of loss convergence between UDRU and K-priors on 3 datasets when models are required to unlearn 20% data of one class.

Table 6: Experimental results of different δ ranging from 1 to 10^{-4} on 3 datasets in the SL paradigm.

Dataset	Original	δ				
		1	10^{-1}	10^{-2}	10^{-3}	10^{-4}
		unlearning data↓				
MNIST	98.46	0.00	0.00	0.00	0.00	0.00
Fashion	97.98	0.87	0.84	0.54	0.33	0.21
CIFAR10	99.61	1.90	0.09	0.02	0.02	0.00
		remaining data↑				
MNIST	98.44	97.63	97.09	89.97	89.29	89.38
Fashion	98.01	97.78	97.47	96.72	96.59	96.33
CIFAR10	99.55	99.21	99.14	98.52	96.38	94.53