How Far Are We from True Unlearnability?

Ye Kai[1†], Liangcai Su[1†], Chenxiong Qian[1] ✉

[1]The University of Hong Kong. † Contributed Equally. ✉Corresponding Author.

Background

High-quality data plays an indispensable role in the era of large models, but the use of unauthorized data for model training greatly damages the interests of data owners. Unlearnable Examples (UEs), which actively disrupt data usability in training by adding imperceptible perturbations, have been proposed to decrease model performance on clean data.

Motivation

Due to **unknown** training purposes and the powerful representation learning capabilities of existing models, these data are expected to be unlearnable for models across multiple tasks, i.e., they will not help improve the model's performance. However, we find that on multi-task dataset (e.g., Taskonomy) current UEs still perform well in tasks such as semantic segmentation, failing to exhibit *multi/cross-task unlearnability*. This leads us to question: *How far are we from attaining truly unlearnable examples?*

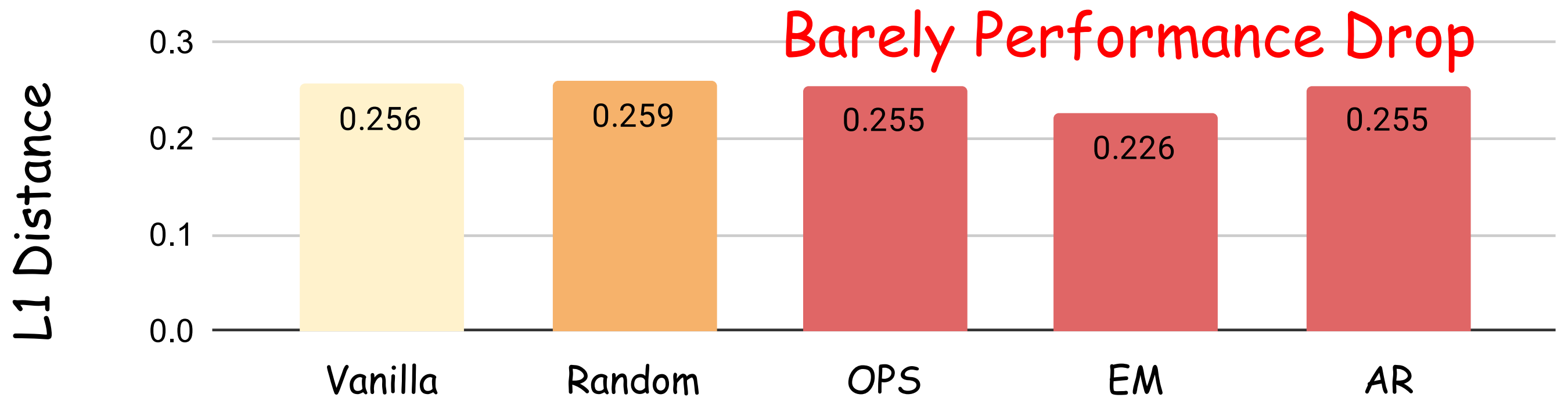

Figure 1. Current UEs for classification task (Scene Classification) fail in other task (Segment Unsup2d) on Taskonomy dataset.

Contributions

1. First to uncover that existing unlearnable methods fail to maintain unlearnability under multi/cross-task scenarios, thus offering new research directions to enhance the practicality of unlearnable examples.
2. Put forward an explanation for the effectiveness of unlearnable examples by analyzing the loss landscape. Further introduce Sharpness-Aware Learnability (SAL) and Unlearnable Distance (UD) as metrics for measuring the unlearnability of model parameters and data, respectively.
3. Benchmark existing unlearnable methods and provide a more intrinsic tool for evaluating UEs. Our approach ascertains the gap between existing research efforts and the truly UEs while encouraging the development of more practical unlearnable methods from a novel perspective.

Exploration of the Model Optimization Process

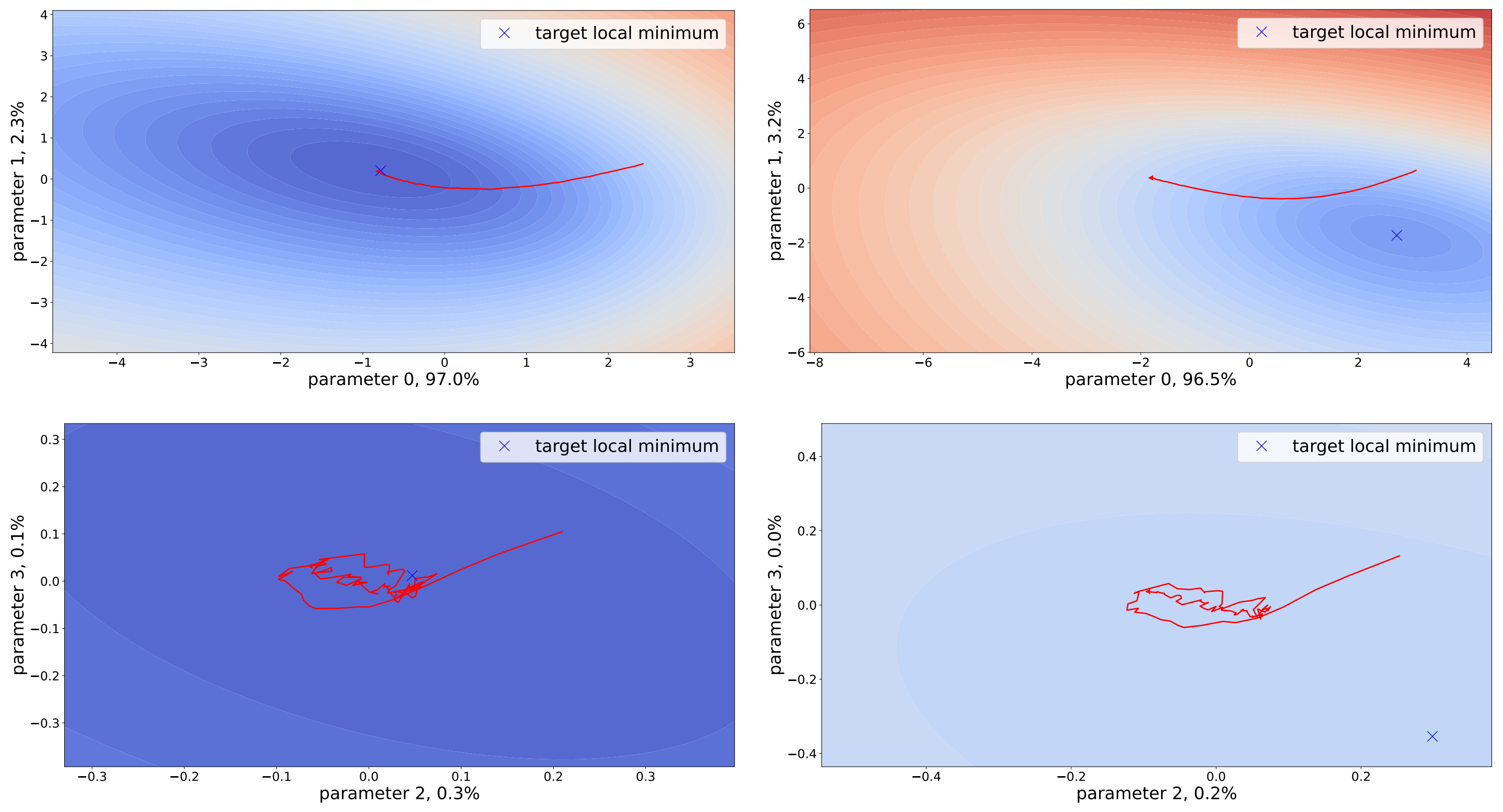

Figure 2. Optimization in loss landscape of training process on Toy Classification task. The x, y-axis are selected **Top 2** parameters. The training campaign of the model on UEs tends to take detours rather than shortcuts to the target minimum. However, only very few parameters converge in line with the model performance. The fewer contour lines in the bottom two figures are due to the loss fluctuations being significantly lower.

Unveiling the Relationship between Loss Landscape and Unlearnability

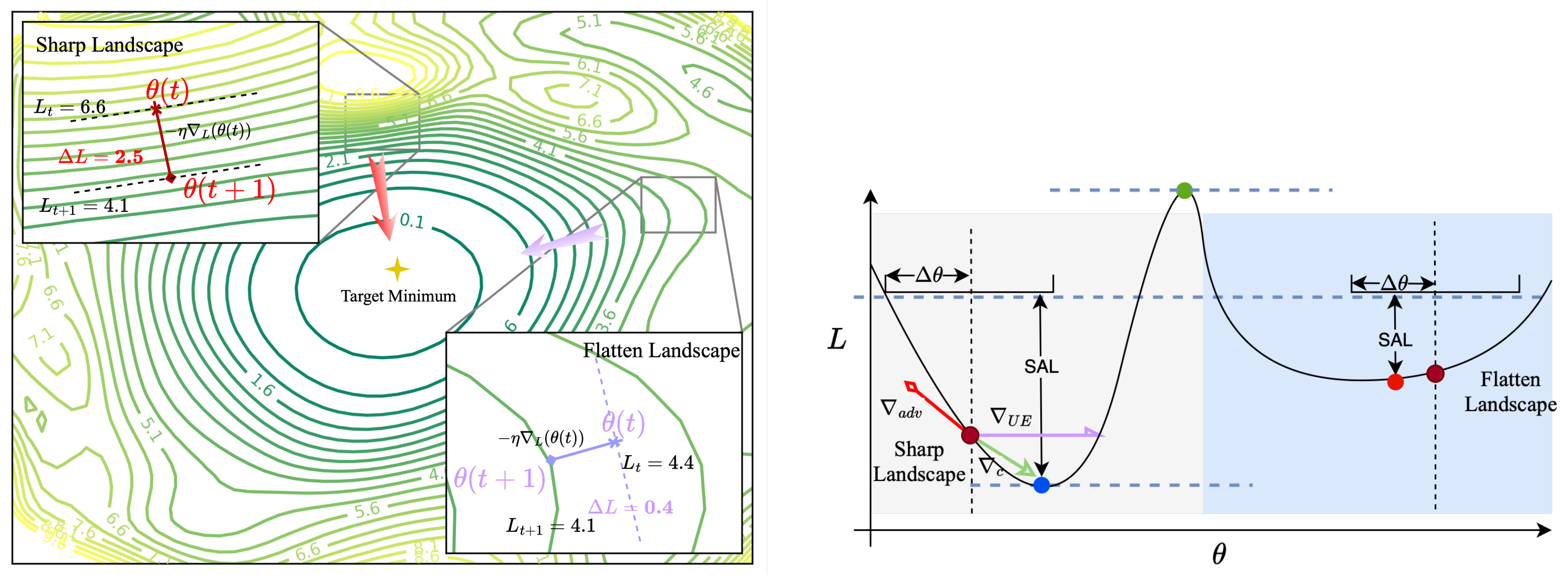

Figure 3. (a) Training loss landscape. (b) When the neighborhood is flatter, the smaller the SAL of the model parameters is, meaning that it is difficult for the model to escape from the local minima because this requires more steps compared to when it is in a sharp neighborhood.

Unlearnable training needs to satisfy one of the following two conditions: (1) the parameter update direction moves along the contour lines; (2) the parameter update magnitude is small enough so that the training loss barely changes.

Table 1. UD of ResNet-18 trained on UEs constructed on CIFAR-10, CIFAR-100 and ImageNet subset.

Unlearnable Method	CIFAR-10			CIFAR-100			ImageNet-100		
	Test Acc	#LP	UD	Test Acc	#LP	UD	Test Acc	#LP	UD
Vanilla	94.11	3.32	\	75.23	2.25	\	69.13	1.86	\
EM	26.52	0.62	0.187	12.34	0.25	0.111	1.20	0.02	**0.011**
REM	30.26	1.54	0.464	20.32	2.25	1.000	4.90	2.38	1.280
DC	18.51	1.00	0.301	54.66	2.24	0.996	5.00	3.40	1.828
TAP	29.85	5.44	1.639	33.75	2.73	1.213	1.20	4.22	2.269
LSP	10.23	1.14	0.343	2.15	0.84	0.373	4.40	3.34	1.796
OPS	11.98	0.52	**0.157**	10.09	0.02	**0.009**	3.30	2.66	1.430

Sharpness-Aware Learnability

We define the **S**harpness-**A**ware **L**earnability (SAL) of layer parameters of model trained on particular dataset as SAL:

$$SAL\left(\boldsymbol{\theta}_l, \epsilon, t\right) = \max_{\|\boldsymbol{v}\|_p \leq \epsilon} \left|\mathcal{L}(\boldsymbol{\theta}_l + \boldsymbol{v}; \mathcal{D}_{tr}) - \mathcal{L}\left(\boldsymbol{\theta}_l; \mathcal{D}_{tr}\right)\right|, \tag{1}$$

where $\boldsymbol{\theta}$ is a l-layer DNN and $\boldsymbol{\theta}_l$ is the l-th layer parameters. $\boldsymbol{v}$ is the perturbation of $\boldsymbol{\theta}_l$, and the parameters of the remaining layers are temporarily frozen. The target training loss ($e.g.$ cross-entropy) is denoted by $\mathcal{L}$. $\|\cdot\|_p \leq \epsilon$ is denoted as the ℓ_p norm. $\mathcal{D}_{tr}$ denotes the training dataset.

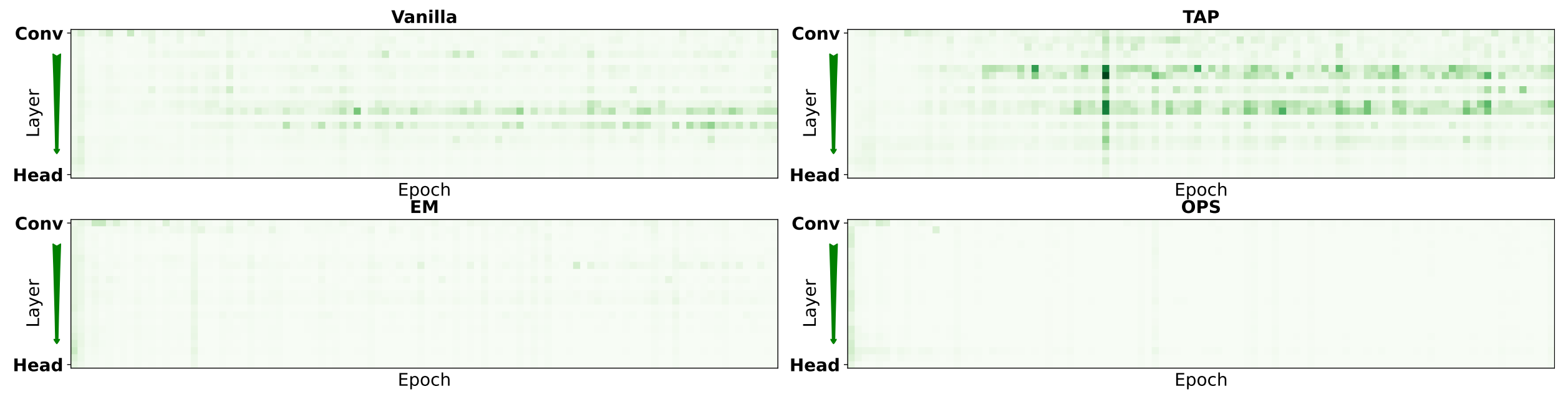

Figure 4. Most parameters in the model trained on UEs exhibit a low SAL (except for TAP, which is considered as adversarial examples), while a small number of parameters on clean training maintain a relatively high SAL (darker green and larger SAL). This suggests that UEs exert unlearnability by reducing the SAL of the parameters.

We define the **L**earnable **T**hreshold (LT) to distinguish between learnable and unlearnable parameters:

$$\beta(T) = \frac{\sum_{t=1}^{T} \sum_{i}^{2} \kappa_i(SAL(\boldsymbol{\theta}^c, \epsilon, t))}{2 \times T}, \tag{2}$$

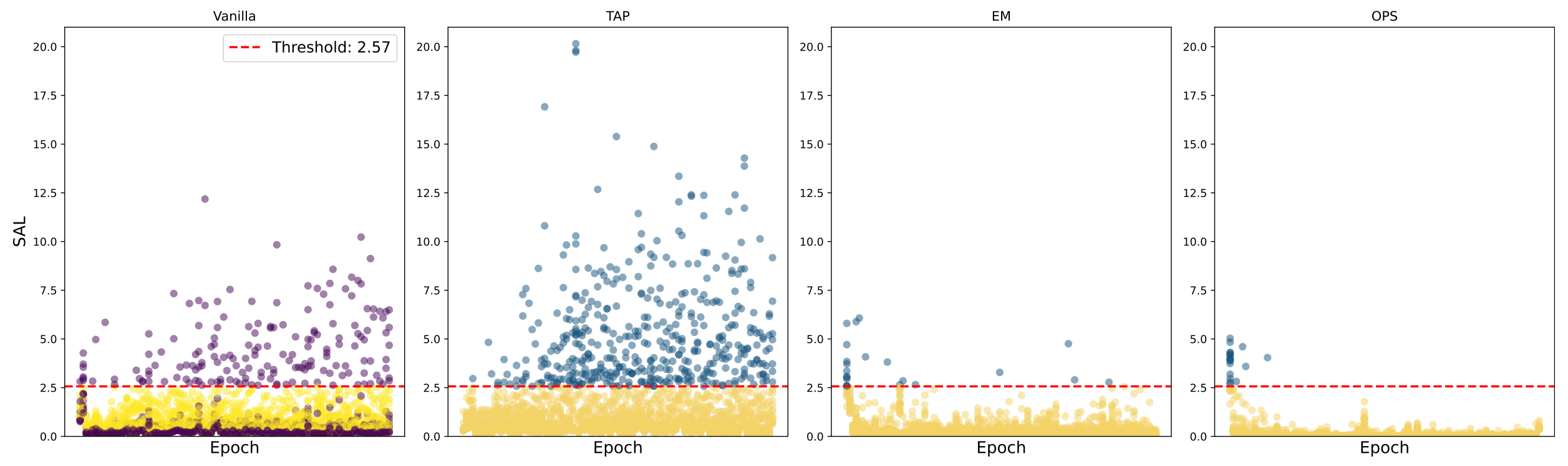

Figure 5. SAL threshold $\beta(T)$ to distinguish learnable and unlearnable parameters in model trained on the vanilla dataset.

T denotes the number of epochs; κ_i represents mean value of the i-th cluster center; $SAL(\boldsymbol{\theta}^c, \epsilon, t)$ refers to the Sharpness-Aware Learnability of the clean model $\boldsymbol{\theta}^c$ at the t-th epoch.

Data Unlearnability Metric: Unlearnable Distance

$$UD(\boldsymbol{\theta}^p) = \frac{\frac{1}{T^p}\sum_{t}^{T^p} \lambda(SAL(\boldsymbol{\theta}^p, \epsilon, t), \beta(T^c))}{\frac{1}{T^c}\sum_{t}^{T^c} \lambda(SAL(\boldsymbol{\theta}^c, \epsilon, t), \beta(T^c))}. \tag{3}$$

A higher UD indicates that the model contains more learnable parameters compared to the clean training process, which suggests weaker data unlearnability for the target task.