A WHY WE CALCULATE θ USING HELD-OUT DATA

In Section 3.2, we estimate θ_i for each training sample using the output of an auxiliary network $f^{\text{aux}}(x_i)$ that is trained on a held-out dataset. In fact, this adaptive flood level θ_i can be considered as the sample difficulty when training the main network. Hence, it is reasonable to consider existing difficulty measurements based on learning dynamics, like C-score (Jiang et al., 2021) or forgetting score (Maini et al., 2022). However, we find these methods are not robust when wrong labels exist in the training data, because the network will learn to remember the wrong label of x_i, and hence provide a low θ_i for the wrong sample, which is harmful to our method. That is why we propose to split the whole training set into n parts and train $f^{\text{aux}}(x_i)$ for n times (each with different $n-1$ parts).

Dataset and implementation To verify this, we conduct experiments on a toy Gaussian dataset, as illustrated in the first panel in Figure 6. Assume we have N samples, each sample in 2-tuple (x, y). To draw a sample, we first select the label $y = k$ following a uniform distribution over all K classes. After that, we sample the input signal $x \mid_{y=k} \sim \mathcal{N}(\mu_k, \sigma^2 I)$, where σ is the noisy level for all the samples. μ_k is the mean vector for all the samples in class k. Each μ_k is a 10-dim vector, in which each dimension is randomly selected from $\{-\delta_\mu, 0, \delta_\mu\}$. Such a process is similar to selecting 10 different features for each class. We consider 3 types of samples for each class: regular sample, which is the typical or easy sample in our training set, has a small σ. Irregular sample, which is generated by using a larger σ. Wrong label sample, which is generated using small σ, but with a flipped label. We generate two datasets following the same procedure (call them datasets A and B).

The, we randomly initialize a 2-layer MLP with ReLU layers and train it on dataset A. At the end of every epoch, we record the loss of each sample in dataset A.

Result The learning paths are illustrated in the second panel in Figure 6. Obviously, the model is capable enough to remember all the wrong labels, as all the curves converge to a small value. If we calculate θ_i in this way, all θ_i would have similar values. However, if we instead train the model using dataset B, which comes from the same distribution but is different from dataset A, the learning curves of samples in dataset A will behave like the last panel in Figure 6. Obviously, the samples with wrong labels and some of the irregular samples can be clearly identified from the figure. Calculating θ_i in this way gives different samples more distinct flood values, which makes our method more robust to sample noise, as our experiments on various scenarios show.

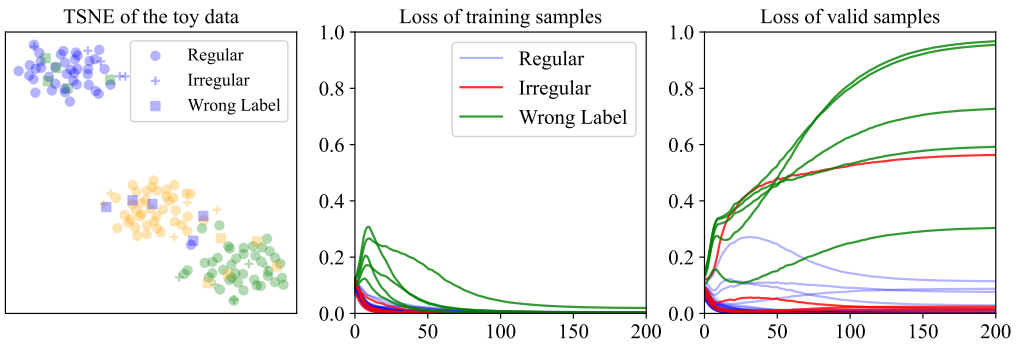


Figure 6: Left: the t-SNE (Van der Maaten & Hinton, 2008) of toy Gaussian example; middle: loss of different samples in the training set; right: loss of different samples in the validation set.

B DETAILS ABOUT DATASETS

Stack Overflow It contains 6,633 sequences with 480,414 events where an event is the acquisition of badges received by users. The maximum number of sequence length is 736 and the number of marks is 22. The dataset is provided by Du et al. (2016) and we use the first folder following Shchur et al. (2020); Bae et al. (2023).

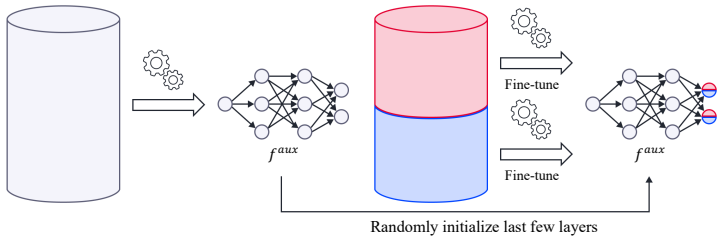
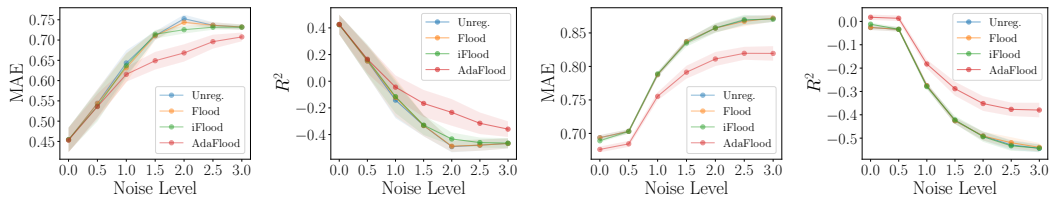
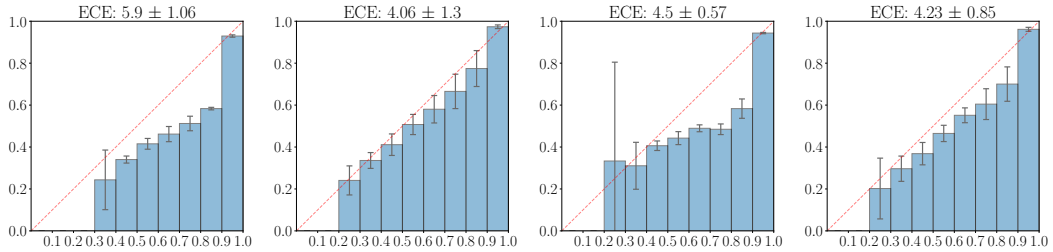


Figure 7: Efficient fine-tuning method for training a auxiliary network when held-out split is $n = 2$.



(a) Brazilian House (MAE) (b) Brazilian House (R^2) (c) Wine Quality (MAE) (d) Wine Quality (R^2)

Figure 8: Additional results in various metrics on tabular datasets with noise and bias



(a) Unregularized (b) Flood (c) iFlood (d) AdaFlood

Figure 9: Calibration results of flooding methods with 10 bins on CIFAR10.

Reddit It contains 10,000 sequences with 532,026 events where an event is posting in Reddit. The maximum number of sequence length is 736 and the number of marks is 22. Marks represent sub-reddit categories.

Uber It contains 791 sequences with 701,579 events where an event is pick-up of customers. The maximum number of sequence length is 2,977 and there is no marks. It is processed and provided by Bae et al. (2023).

Brazilian Houses It contains information of 10,962 houses to rent in Brazil in 2020 with 13 features. The target is the rent price for each house in Brazilian Real. According to OpenML (Vanschoren et al., 2013) where we obtained this dataset, since the data is web-scrapped, there are some values in the dataset that can be considered outliers.

Wine Quality It contains 6,497 samples with 11 features and the quality of wine is numerically labeled as targets. This dataset is also obtained from OpenML (Vanschoren et al., 2013).

SST-2 The Stanford Sentiment Treebank (SST-2) is a dataset containing fully annotated parse trees, enabling a comprehensive exploration of how sentiment influences language composition. Comprising 11,855 individual sentences extracted from film reviews, this dataset underwent parsing using the Stanford parser, resulting in a collection of 215,154 distinct phrases.