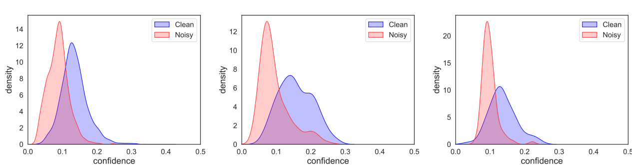
Stochastic Feature Averaging for Learning with Long-Tailed Noisy Labels Supplementary Material

Hao-Tian Li[1,2,3] , **Tong Wei**[1,2*] , **Hao Yang**[3] , **Kun Hu**[3] , **Chong Peng**[3] ,
Li-Bo Sun[3] , **Xun-Liang Cai**[3] , **Min-Ling Zhang**[1,2]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Lab. of Computer Network and Information Integration (Southeast University), MOE, China
[3]Meituan, Shanghai, China
{liht, weit, zhangml}@seu.edu.cn, {yanghao52, hukun05, pengchong, sunlibo03, caixunliang}@meituan.com

1 Further Experimental Results

1.1 Analyses of Instant Centroid Estimation

The Instant Centroid Estimation (ICE), which utilizes the predicted confidence of the auxiliary balanced classifier, allows for a rough estimation of the class centroids with a single threshold. This is based on the observation that samples of head, medium, and tail classes typically exhibit similar ranges in the confidence distribution, and the confidence of most clean samples is generally higher than noisy samples within each class, as shown in Figure 1. ICE can be considered as a primary sample selection strategy that aims to remove as many noisy samples as possible in order to minimize the impact of label noise on the estimation of the class centroids. To further verify the effectiveness of ICE, we illustrate the curve of the precision of sample selection as a function of training iterations in Figure 2. It can be seen that ICE consistently achieves high precision on both CIFAR-10 and CIFAR-100 datasets, which ultimately leads to a decent estimation of class centroid.



(a) Head Class (b) Medium Class (c) Tail Class

Figure 1: Confidence distributions in head, medium and tail classes output by balanced classifier in CIFAR-10 dataset after warm-up.

1.2 Additional Ablation Studies on CIFAR

We conduct additional ablation studies on key components of the proposed SFA framework on the CIFAR dataset. Table 1 reports the test accuracy on CIFAR-100 dataset with varying levels of noise and imbalance factors. It is obvious that the performance of the final model is the result of the collective contributions from each key component, with the auxiliary balanced classifier having the most significant impact. This can be attributed to the extremely small number of samples in tail classes under certain data settings (i.e., $N_k = 5$ under 100 imbalance factor). This makes the model highly dependent on the balanced softmax function, while hindering the accurate estimation of class centroids.

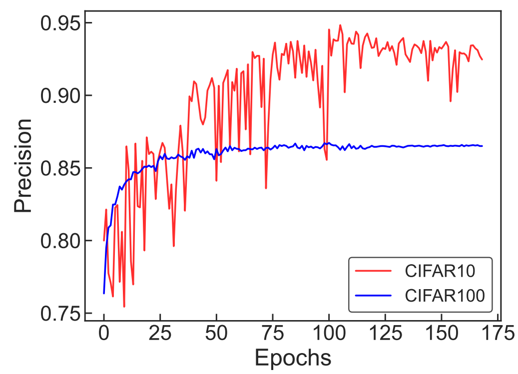


Figure 2: Precision of sample selection based on confidences output by the balanced classifier in CIFAR-10 and CIFAR-100 datasets.

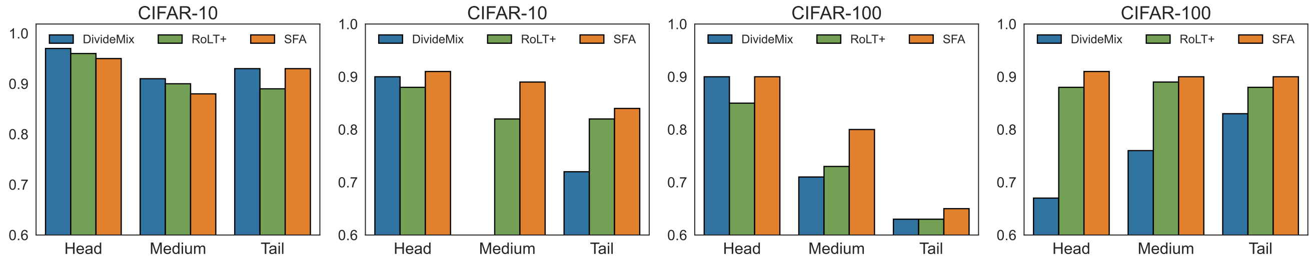
Noise Level		0.2			0.5		
Imbalance Ratio		10	50	100	10	50	100
SFA	Best	66.32	54.29	48.51	57.41	44.37	39.73
	Last	65.65	53.10	47.73	57.28	43.41	39.73
w/o ICE	Best	66.28	53.23	48.39	56.15	43.43	38.52
	Last	65.92	53.23	47.02	56.00	43.34	38.01
w/o SCC	Best	65.92	52.95	48.19	55.96	42.30	38.39
	Last	65.92	52.85	48.00	55.58	42.26	38.16
w/o ABC	Best	64.62	50.25	45.27	54.12	39.79	34.63
	Last	64.11	50.17	45.12	53.92	39.56	33.33

Table 1: Ablation studies on key components of our proposed SFA framework. Test accuracy on CIFAR-100 dataset is reported.

1.3 Additional Analyses of Sample Selection

The effectiveness of sample selection by the SFA framework is further illustrated in Figure 3, showcasing the precision and recall on CIFAR-10 and CIFAR-100 datasets under challenging conditions of 50% noise level and 100 imbalance factor. The results indicate that our method demonstrates superior

*Corresponding author



(a) CIFAR-10 Precision (b) CIFAR-10 Recall (c) CIFAR-100 Precision (d) CIFAR-100 Recall

Figure 3: Precision and recall of the head, medium and tail classes on CIFAR-10 and CIFAR-100 datasets under $\gamma = 0.5$ and $\rho = 100$.

		CIFAR-10						CIFAR-100					
Noise Level		0.2			0.5			0.2			0.5		
Imbalance Ratio		10	50	100	10	50	100	10	50	100	10	50	100
$S = 1$	Best	92.37	84.95	79.04	90.42	79.21	74.65	**66.67**	54.20	48.47	**57.59**	44.32	39.05
	Last	92.09	84.34	78.15	90.02	**79.08**	73.69	66.17	53.52	47.41	57.22	43.39	39.05
$S = 3$	Best	92.21	85.01	**80.64**	90.27	79.47	**75.17**	66.48	**54.46**	48.34	57.42	**44.51**	39.19
	Last	91.60	84.18	78.34	90.05	78.65	**74.30**	**66.29**	**53.74**	47.39	56.90	**43.63**	38.49
$S = 5$	Best	**92.53**	**85.96**	80.26	**90.57**	**79.89**	**75.17**	66.32	54.29	**48.51**	57.41	44.37	**39.73**
	Last	**92.13**	**84.80**	**79.22**	**90.08**	78.93	74.06	65.65	53.10	**47.73**	**57.28**	43.41	**39.73**

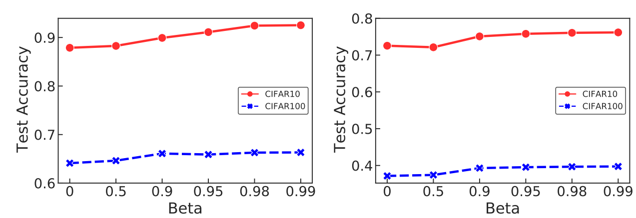
Table 2: Test accuracy (%) on simulated CIFAR datasets with different sampling rate.

precision on CIFAR-100 and improved recall on both CIFAR-10 and CIFAR-100 datasets. While the precision of the head and medium classes on CIFAR-10 is relatively lower, SFA significantly increases the recall of these classes, resulting in an overall improvement in performance. Therefore, sample selection using stochastic feature averaging is effective in identifying clean samples for model training.

1.4 Analyses of Parameter Sensitivity

Smoothing Factor We conducted an extensive evaluation of the impact of the smoothing factor β in exponential moving average by considering various values ranging from 0 to 0.99. It is worth noting that higher β values are less responsive to recent data, while lower values place greater emphasis on recent data. Given the uncertainties arising from label noise and data scarcity, a higher β is recommended in practice, which is consistent with previous literature such as Mean-Teacher *[NeurIPS'17]*. Our results, shown in Figure 4, demonstrate that our method consistently achieves good performance when β is set to a high value ($\beta > 0.9$), indicating its robustness across a wide range of β values.

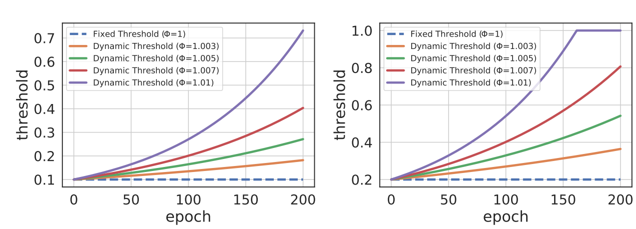
Dynamic Confidence Threshold The impact of the two parameters for the dynamic threshold can be analyzed from two perspectives: how they determine the dynamic threshold (Figure 5), and how the dynamic threshold affects the final performance (Table 3). Based on our results, we have made the following observations: (1) using a low fixed threshold ($\phi = 1$, $\hat{\tau} = 1/K$) can lead to a decrease in performance because it results in more noisy samples being selected in the class centroid estimation; (2) when the dynamic threshold increases too rapidly ($\phi = 1.007$, $\hat{\tau} = 2/K$ or $\phi = 1.01$), the



(a) $\gamma = 0.2$, $\rho = 10$ (b) $\gamma = 0.5$, $\rho = 100$

Figure 4: Test accuracy on CIFAR-10 and CIFAR-100 with varying smoothing factor (β) under different noise and imbalance ratios.

performance is also compromised because the number of selected samples is too limited to obtain an accurate centroid. However, if the changes in the dynamic threshold are appropriate ($\phi = 1.003$ and 1.005), the model can consistently achieve high performance.



(a) $\hat{\tau} = 1/K$ (b) $\hat{\tau} = 2/K$

Figure 5: Fixed and dynamic thresholds with different $\hat{\tau}$ and ϕ on CIFAR-10. K is the number of classes.

ϕ	1	1.003	1.005	1.007	1.01
$\hat{\tau} = 1/K$	38.12	39.10	39.20	38.77	37.99
$\hat{\tau} = 2/K$	38.75	39.18	39.05	38.00	37.51

Table 3: Test accuracy on CIFAR-100 under 50% noise level and 100 imbalacne ratio.

Sampling Rate An analysis of the effect of varying sampling rates of the stochastic class centroids is also present in Table 2. The results show that increasing the sampling rate can improve the classification accuracy especially in case of severe class imbalance, but the impact is not significant. To achieve a trade-off between performance and efficiency in real-world applications, it is advisable to choose a small value of S for large-scale datasets.

1.5 Results on Class-balanced Datasets

To evaluate the performance of our proposed framework on class-balanced datasets, we conducted experiments on the CIFAR datasets and compared our method with the vanilla Cross-entropy (CE), DivideMix, and ELR+. The results are summarized in Table 4. Although our method is primarily designed to address the challenge of long-tailed noisy labels, we were pleased to observe that it also achieves comparable performance to state-of-the-art methods in the field of label-noise learning.

	CIFAR-10			CIFAR-100		
Noise Level	0.2	0.5	0.8	0.2	0.5	0.8
CE	86.8	79.4	62.9	62.0	46.7	19.9
DivideMix	**96.1**	94.6	93.2	77.3	**74.6**	60.2
ELR+	95.8	94.8	**93.3**	77.6	73.6	**60.8**
Ours	95.8	**94.9**	93.2	**78.4**	73.8	60.0

Table 4: Test accuracy (%) on balanced CIFAR datasets. Results except for ours are taken from ELR+.