ProSelfLC: Progressive Self Label Correction for Training Robust Deep Neural Networks

Xinshao Wang[1, 2], Yang Hua[3], Elyor Kodirov[1],
David A. Clifton[2], Neil M. Robertson[1, 3]

[1]Zenith Ai, UK
[2]Institute of Biomedical Engineering, University of Oxford, UK
[3]Institute of Electronics, Communications and Information Technology, Queen's University Belfast, UK

{xinshao, elyor}@zenithai.co.uk, {y.hua, n.robertson}@qub.ac.uk,
{xinshao.wang, david.clifton}@eng.ox.ac.uk

Code:https://github.com/XinshaoAmosWang/ProSelfLC-CVPR2021

19-25/06/2021, CVPR VIRTUAL

Overlook: core research questions we study

1. In Self LC, how much should we trust a learner to leverage its knowledge?
   - The trust score is fixed or updated stage-by-stage in prior work.
   - ProSelfLC modifies the target progressively, is end-to-end trainable, and requires negligible extra cost.
2. Should we penalise a low-entropy status or reward it?
   - OR methods penalise low entropy while LC rewards it.
   - ProSelfLC redirects and promotes entropy minimisation, which is in marked contrast to recent practices of confidence penalty [6, 4, 1].

Zenith Ai UNIVERSITY OF OXFORD QUEEN'S UNIVERSITY BELFAST

BEYOND SEMANTIC CLASS

THE SIMILARITY STRUCTURE IN A LABEL DISTRIBUTION

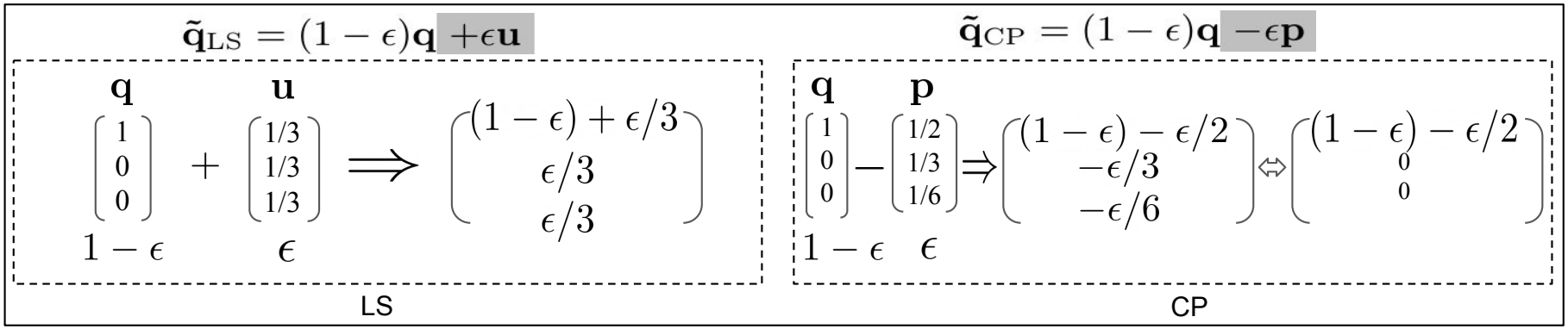
A label distribution defines what to learn:

- **Definition 1** (*Semantic Class*). Given a target label distribution $\tilde{q}(x) \in \mathbb{R}^C$, the semantic class is defined by $\arg\max_j \tilde{q}(j|x)$, i.e., the class whose probability is the largest.

- **Definition 2** (*Similarity Structure*). In CCE, LS and CP, a data point has an identical probability of belonging to other classes except for the semantic class. Instead, in LC, a target label distribution captures the probability difference of an example being predicted to every class. We define it to be the similarity structure of one example versus all training classes.

An overview of label (target) modification

OR(LS and CP) + LC(Self LC and Non-self LC)



OR includes LS [6] and CP [4], which smoothes similarity structure:

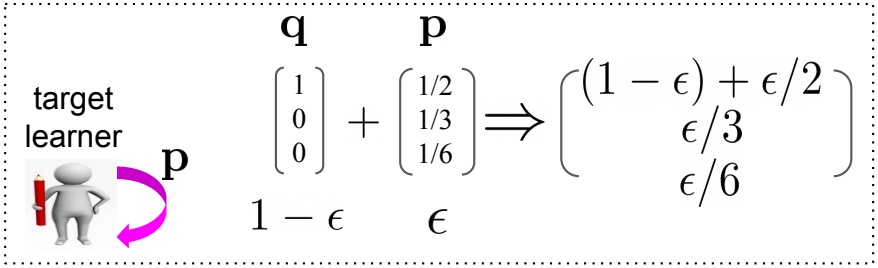
- LS softens a target by adding a uniform label distribution.
- CP changes the probability 1 to a smaller value $1-\epsilon$ in the one-hot target.
  The double-ended arrow means factual equivalence, because an output is definitely non-negative after a softmax layer.

Zenith Ai UNIVERSITY OF OXFORD QUEEN'S UNIVERSITY BELFAST

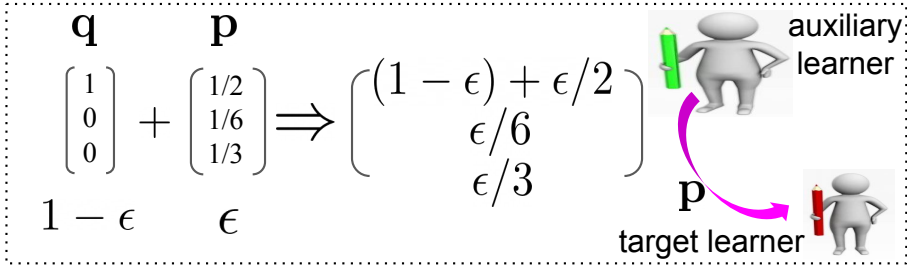
An overview of label (target) modification

OR(LS and CP) + LC(Self LC and Non-self LC)

$$\tilde{\mathbf{q}}_{\mathrm{LC}} = (1-\epsilon)\mathbf{q} + \epsilon\mathbf{p}$$



Self LC: **p** is the output of a learner itself.



Non-self LC: **p** is the output of an auxiliary learner.

- LC contains Self LC [3, 5, 7] and Non-self LC [2].
- The convex combination parameter ϵ defines how much a predicted label distribution is trusted.

Zenith Ai UNIVERSITY OF OXFORD QUEEN'S UNIVERSITY BELFAST

Drawbacks of existing target modification

Why Self LC to exploit a model's self knowledge?

1. OR methods naively penalise confident outputs without leveraging easily accessible knowledge from other learners or itself.
2. Non-self LC relies on accurate auxiliary models.
3. Self LC:
   - It exploits its own knowledge;
   - It requires no extra learners;
   - However, **how much should we trust a learner to leverage its knowledge?**

Zenith Ai UNIVERSITY OF OXFORD QUEEN'S UNIVERSITY BELFAST

Overview of existing variants of Self LC

Without considering a model's knowledge grows as time goes

1. In bootstrapping, ϵ is fixed throughout the training process.
2. Joint Optimisation fully trusts a learner by setting $\epsilon = 1$, and uses stage-wise training to gradually train the model.
   - Stage-wise training requires a significant human intervention and is time-consuming in practice.
3. **Requirements of improving Self LC**
   - End-to-end trainable.
   - Negligible extra cost.
   - Modifies the target progressively and adaptively as training goes.

To penalise or reward a low-entropy status?

The 2nd core research question we studied

- OR methods penalise low entropy $\Rightarrow$ OR is against entropy minimisation principle.

- LC rewards a low-entropy status $\Rightarrow$ LC defends entropy minimisation principle.
  - LC has the same principle as the widely used expectation–maximization (EM) algorithm.

Zenith Ai
UNIVERSITY OF OXFORD
QUEEN'S UNIVERSITY BELFAST

ProSelfLC

Self Trust according to Training Time and Confidence

ϵ indicates how much a predicted label distribution is trusted.
For any x, we summarise the loss and modified label:

$$
\begin{aligned}
&L_(\tilde{\mathrm{q}}_{\mathrm{ProSelfLC}}, \mathrm{p}; \epsilon_{\mathrm{ProSelfLC}}) = \mathrm{H}(\tilde{\mathrm{q}}_{\mathrm{ProSelfLC}}, \mathrm{p}) = \mathrm{E}_{\tilde{\mathrm{q}}_{\mathrm{ProSelfLC}}}(-\log\ \mathrm{p}),\\
&\tilde{\mathrm{q}}_{\mathrm{ProSelfLC}} = (1-\epsilon_{\mathrm{ProSelfLC}})\mathrm{q} + \epsilon_{\mathrm{ProSelfLC}}\mathrm{p},\\
&\epsilon_{\mathrm{ProSelfLC}} = g(t) \times l(\mathrm{p}),\\
&g(t) = h(t/\Gamma - 0.5, B) \in (0,1), \Rightarrow \text{Trusting learning time}\\
&l(\mathrm{p}) = 1 - \mathrm{H}(\mathrm{p})/\mathrm{H}(\mathrm{u}) \in (0,1). \Rightarrow \text{Trusting sample confidence}
\end{aligned}
\tag{1}
$$

t and Γ are the iteration (time) counter and the number of total iterations, respectively.
$h(\cdot)$ is a logistic function where B controls its smoothness.

Zenith Ai
UNIVERSITY OF OXFORD
QUEEN'S UNIVERSITY BELFAST

ProSelfLC

Self Trust according to Training Time and Confidence

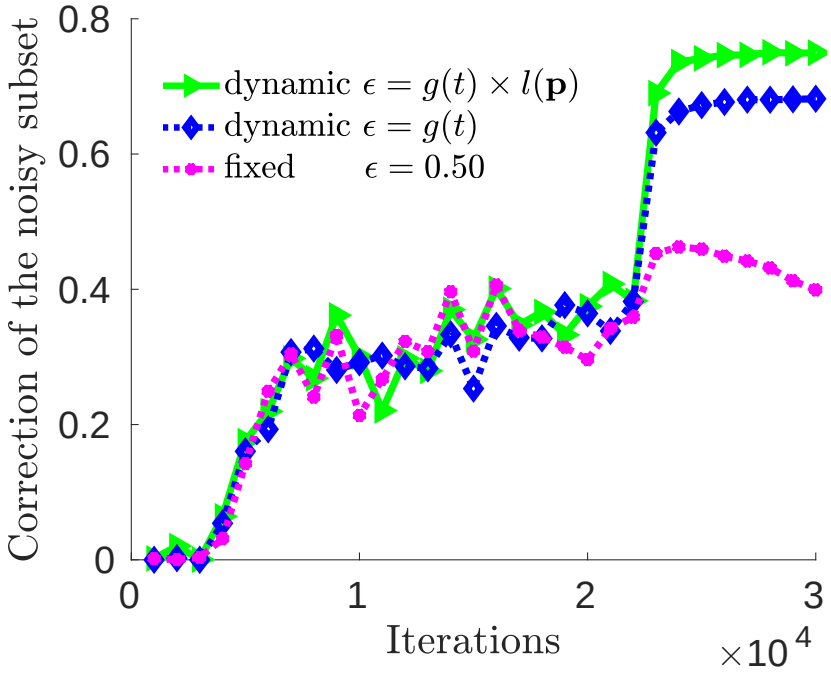
Table: Case analysis of ProSelfLC. Consistency is determined merely by the semantic class.

		l(p)		
	$g(t)$	0.1(non-confident)	0.9 (confidently consistent)	0.9 (confidently inconsistent)
The earlier phase	0.1	0.01	0.09	0.09
The later phase	0.9	0.09	0.81	**0.81**(correct the semantic class)

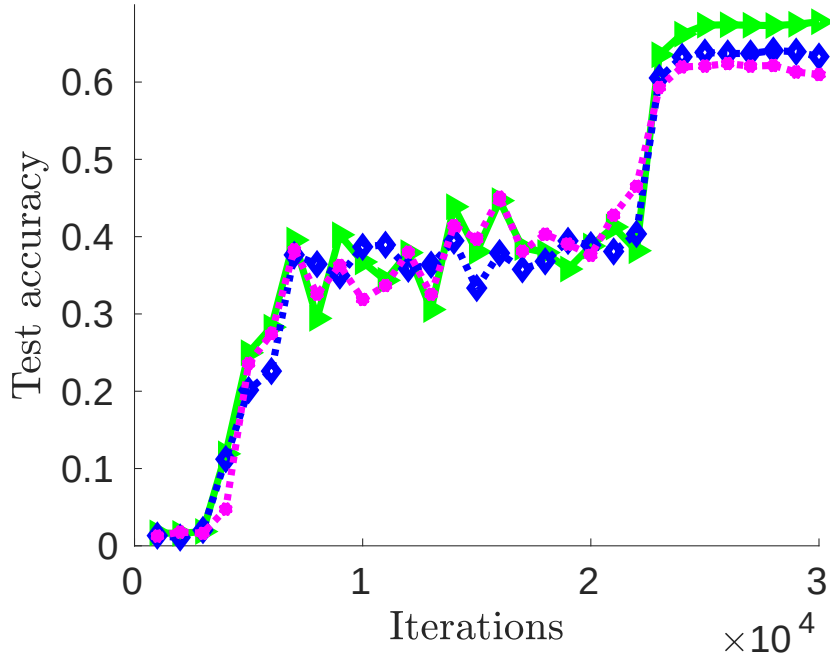
- We use concrete values, e.g., 0.1 and 0.9, for more concise interpretation.
- We bold the special case, where an output distribution p is confident but inconsistent with q.

ProSelfLC

Experiments on CIFAR-100 with 40% asymmetric label noise



(a) Semantic class correction



(b) Generalisation

Figure: Study of setting ϵ using three schemes: global trust and local trust, merely global trust, and fixed ϵ.

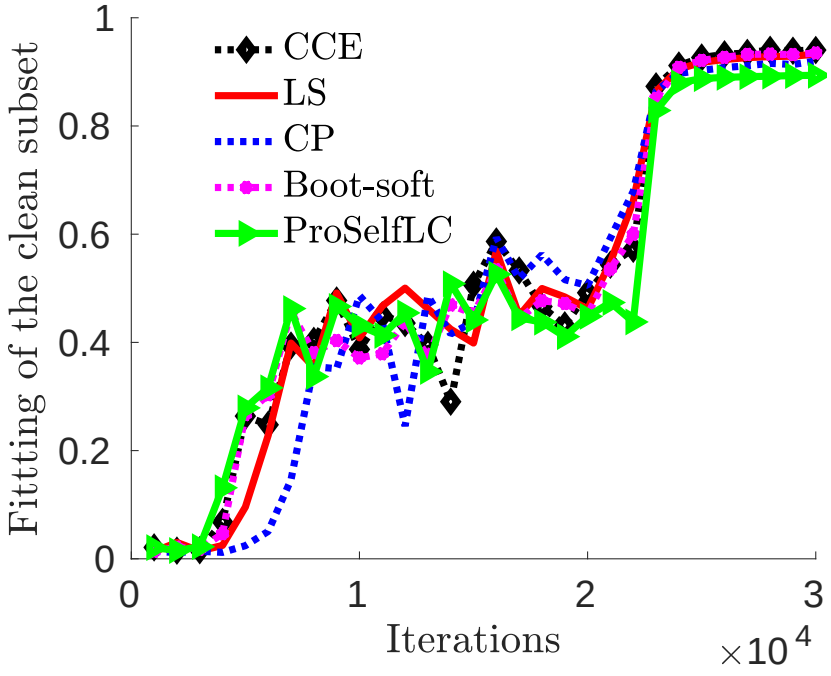
Zenith Ai

UNIVERSITY OF OXFORD

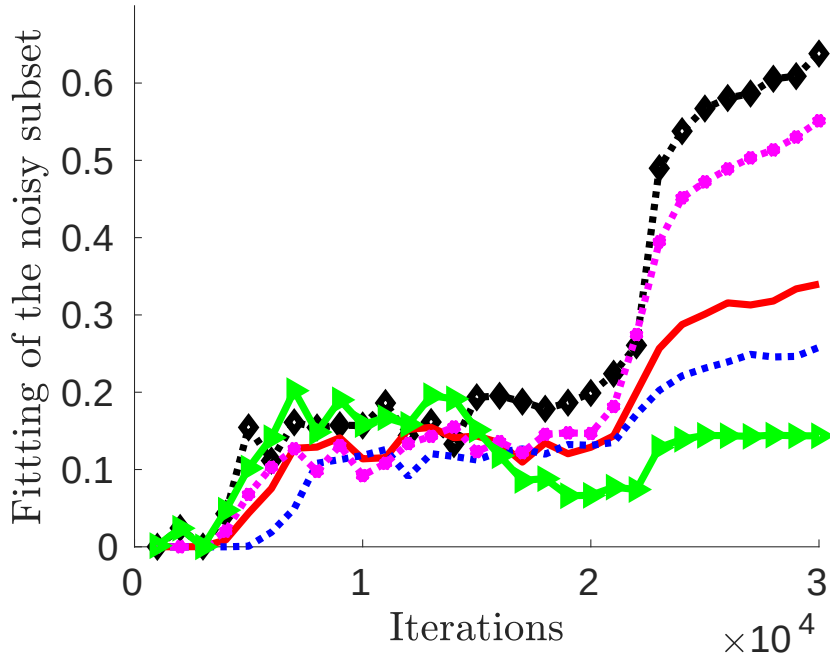
QUEEN'S UNIVERSITY BELFAST

ProSelfLC

Training Dynamics On CIFAR-100 with asymmetric label noise $r = 0.4$.



(a) Correct fitting.



(b) Wrong fitting.

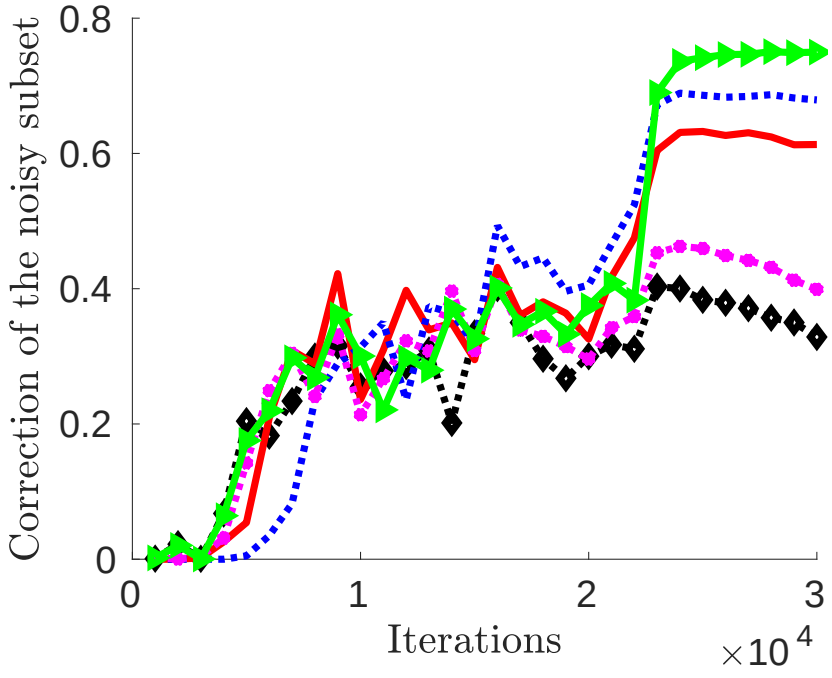
Zenith Ai

UNIVERSITY OF OXFORD

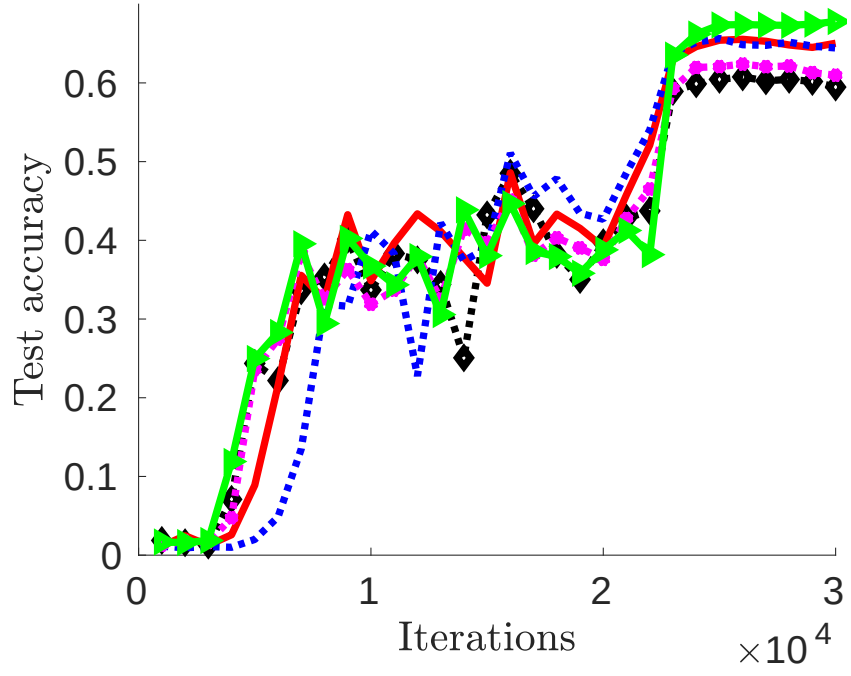
QUEEN'S UNIVERSITY BELFAST

ProSelfLC

Training Dynamics On CIFAR-100 with asymmetric label noise $r = 0.4$.



(a) Semantic class correction



(b) Generalisation.

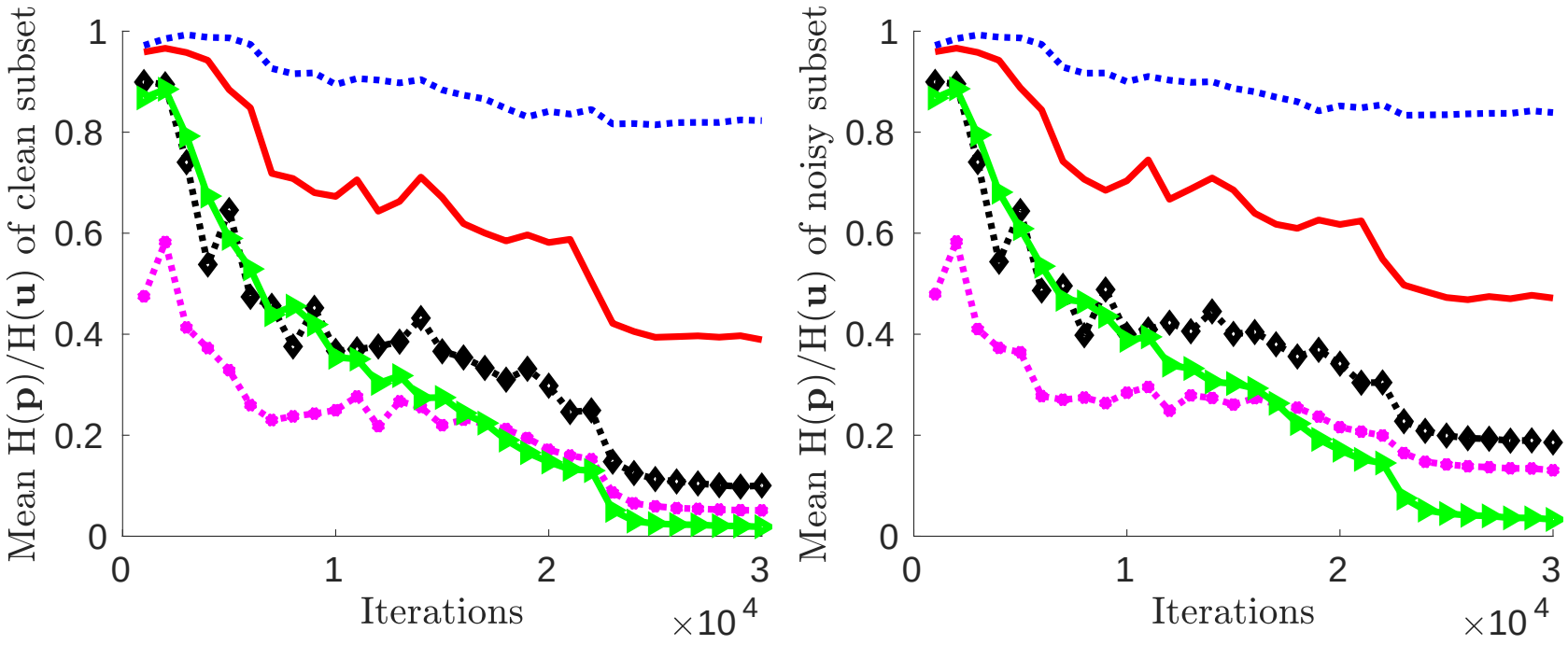
Zenith Ai

UNIVERSITY OF OXFORD

QUEEN'S UNIVERSITY BELFAST

ProSelfLC

Training Dynamics On CIFAR-100 with asymmetric label noise $r = 0.4$.



(a) Entropy of clean subset. (b) Entropy of noisy subset.

Figure: **Should we penalise a low-entropy status or reward it?**

Zenith Ai UNIVERSITY OF OXFORD QUEEN'S UNIVERSITY BELFAST

Summary/Conclusion

1. ProSelfLC:
   - enhance the similarity structure information over training classes.
   - correct the semantic classes of noisy label distributions.
   - is the first method to trust self knowledge progressively and adaptively.
2. Our extensive experiments:
   - defend the entropy minimisation principle;
   - demonstrate the effectiveness of ProSelfLC in clean and noisy settings.
3. Code:
   https://github.com/XinshaoAmosWang/ProSelfLC-CVPR2021

Zenith Ai
UNIVERSITY OF OXFORD
QUEEN'S UNIVERSITY BELFAST

Thanks for your attention !

Questions are greatly welcome!

References

[1] Dubey, A., Gupta, O., Raskar, R., and Naik, N. Maximum-entropy fine grained classification. In *NeurIPS*, 2018.

[2] Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[3] Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 2013.

[4] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. In *ICLR Workshop*, 2017.

[5] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. In *ICLR Workshop*, 2015.

[6] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[7] Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. Joint optimization framework for learning with noisy labels. In *CVPR*, 2018.

Zenith Ai

UNIVERSITY OF OXFORD

QUEEN'S UNIVERSITY BELFAST