How does unlabeled data improve generalization in self-training? A one-hidden-layer theoretical analysis

Shuai Zhang [*], Meng Wang [*], Sijia Liu [†], Pin-Yu Chen [§], Jinjun Xiong [¶]

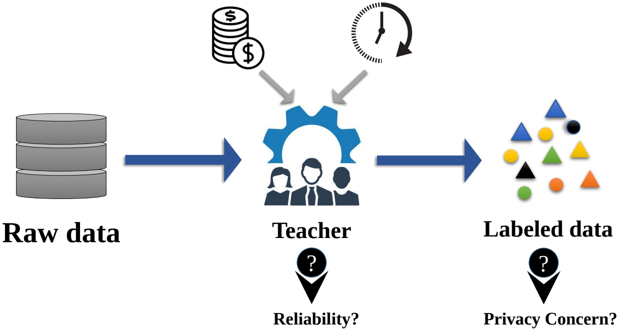
[*]Rensselaer Polytechnic Institute
[†]Michigan State University
[§]IBM Research
[¶]University at Buffalo

March, 2022

Semi-Supervised Learning (Semi-SL)

- Semi-Supervised Learning: Few labeled data & Plenty of unlabeled data;
- Why unlabeled data? Problems of labeled data:

☐ Expensive ☐ Time-consuming ☐ Lack of Quality ☐ Privacy Concern



Self-training Algorithm

labeled data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ & unlabeled data $\widetilde{\mathcal{D}} = \{\tilde{\boldsymbol{x}}_m\}_{m=1}^{M}$:

- Generate pseudo-labels:

$$\tilde{y}_m = g(\boldsymbol{W}^{(\ell)}; \tilde{\boldsymbol{x}}_m);$$

- Objective function in (S3):

$$\hat{f}_{\mathcal{D},\tilde{\mathcal{D}}}(\boldsymbol{W}) = \frac{\lambda}{2N} \sum_{n=1}^{N} \Big(y_n - g(\boldsymbol{W}; \boldsymbol{x}_n)\Big)^2 + \frac{1-\lambda}{2M} \sum_{m=1}^{M} \Big(\tilde{y}_m - g(\boldsymbol{W}; \tilde{\boldsymbol{x}}_m)\Big)^2;$$

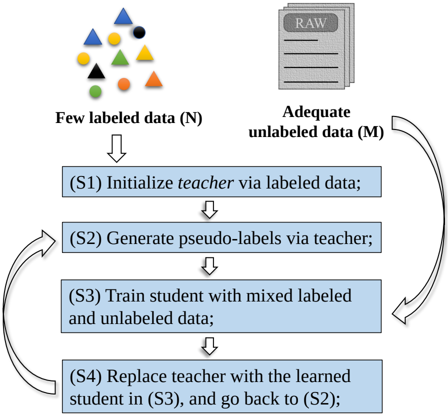


Figure 1: Illustration of iterative self-training method

Formal theoretical results

Theorem 1 (Convergence analysis in low labeled data regime, [ICLR'2022])

If the following conditions hold:

$$1/2 \leq \lambda \leq \sqrt{N/N^*} \quad \textit{and} \quad M \geq \Theta((1-\lambda)^2 K^3 d \log q).$$

Then, the iterations $\{\boldsymbol{W}^{(\ell)}\}_{\ell=0}^{L}$ *converges to* $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$ *as*

$$\|\boldsymbol{W}^{(L)} - \boldsymbol{W}^{[\lambda]}\|_F \leq \Big(\big(1+\Theta(\frac{1}{\sqrt{M}})\big)\cdot\frac{1}{K}\Big)^L \cdot \|\boldsymbol{W}^{(0)} - \boldsymbol{W}^{[\lambda]}\|_2 + \big(1+\Theta(\frac{1}{\sqrt{M}})\big)\cdot\frac{1}{K}\cdot\|\boldsymbol{W}^* - \boldsymbol{W}^{[\lambda]}\|_F.$$

Theorem 2 (Zero generalization error, [ICLR'2022])

If the following conditions hold:

$$(1-\Theta(1/\sqrt{K}))^2 \cdot N^* \leq N \leq N^* \qquad , M \geq \Theta((1-\lambda)^2 K^3 d \log q)$$

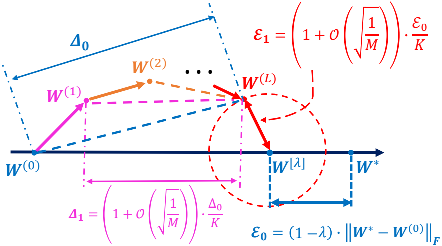
Then, the iterations converge to the ground truth $\boldsymbol{W}^*$ *as follows,*

$$\|\boldsymbol{W}^{(L)} - \boldsymbol{W}^*\|_F \leq \Big[(1+\frac{\lambda}{\sqrt{N}}+\frac{1-\lambda}{\sqrt{M}})\cdot\sqrt{K}(1-\lambda)\Big]^L \cdot \|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_F.$$

Insights from the theoretical results

The roles of unlabeled data amount:

- The convergence rate is a linear function of $1/\sqrt{M}$;
- The distance between the convergent point $\boldsymbol{W}^{(L)}$ and $\boldsymbol{W}^{[\lambda]}$ is a linear function of $1/\sqrt{M}$.



The selections of λ (weighted sum factor of the labeled data's loss function):

- Large λ requires less number of unlabeled data, and the convergent point move towards the desired point $\boldsymbol{W}^*$;
- The upper bound of λ in convergence analysis is controlled by the initialization and labeled data amount; large labeled data and better initialization indicates a high upper bound of λ;

Simulation Results: Real data

- Image classification via the Wide ResNet 28-10 with augmented Cifar-10 dataset;

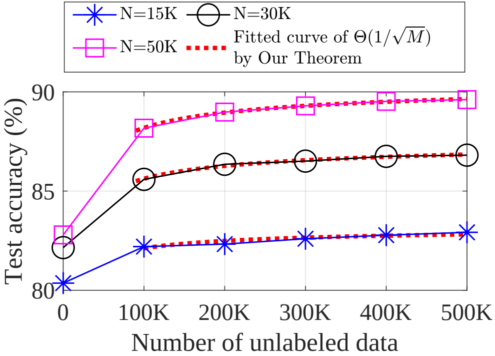


Figure 2: The test accuracy against the number of unlabeled data

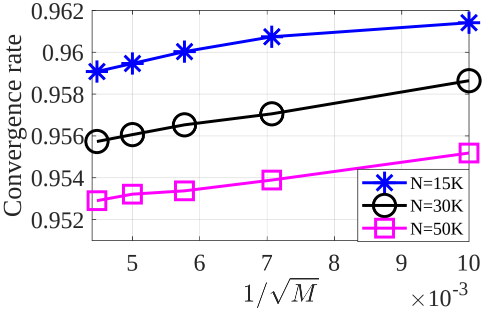


Figure 3: The convergence rate against the number of unlabeled data

Allen-Zhu, Z., Li, Y., and Liang, Y. (2019).
Learning and generalization in overparameterized neural networks, going beyond two layers.
In *Advances in Neural Information Processing Systems 32*, pages 6158–6169.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. (2018).
Stronger generalization bounds for deep nets via a compression approach.
In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 254–263. PMLR.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017).
Spectrally-normalized margin bounds for neural networks.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc.

Baykal, C., Liebenwein, L., Gilitschenski, I., Feldman, D., and Rus, D. (2018).

Data-dependent coresets for compressing neural networks with applications to generalization bounds.
International Conference on Learning Representations.

Ben, M., Osadchy, M., Braverman, V., Zhou, S., and Feldman, D. (2020).
Data-independent neural pruning via coresets.
In *International Conference on Learning Representations (ICLR)*.

Chen, Y., Wei, C., Kumar, A., and Ma, T. (2020a).
Self-training avoids using spurious features under domain shift.
Advances in Neural Information Processing Systems, 33.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. (2020b).
A generalized neural tangent kernel analysis for two-layer neural networks.
Advances in Neural Information Processing Systems, 33.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2019).
Gradient descent provably optimizes over-parameterized neural networks.
In *International Conference on Learning Representations*.

Frankle, J. and Carbin, M. (2019).
The lottery ticket hypothesis: Finding sparse, trainable neural networks.
In *International Conference on Learning Representations*.

Fu, H., Chi, Y., and Liang, Y. (2020).
Guaranteed recovery of one-hidden-layer neural networks via cross entropy.
IEEE Transactions on Signal Processing, 68:3225–3235.

Jacot, A., Gabriel, F., and Hongler, C. (2018).
Neural tangent kernel: Convergence and generalization in neural networks.
In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Malach, E., Yehudai, G., Shalev-Shwartz, S., and Shamir, O. (2020).
Proving the lottery ticket hypothesis: Pruning is all you need.
arXiv preprint arXiv:2002.00585.

Oymak, S. and Gulcu, T. C. (2020).
Statistical and algorithmic insights for semi-supervised learning with self-training.

arXiv preprint arXiv:2006.11006.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. (2020).
Understanding and mitigating the tradeoff between robustness and accuracy.
In *International Conference on Machine Learning*, pages 7909–7919. PMLR.

Wang, C., Zhang, G., and Grosse, R. (2019).
Picking winning tickets before training by preserving gradient flow.
In *International Conference on Learning Representations*.

Wei, C., Shen, K., Chen, Y., and Ma, T. (2020).
Theoretical analysis of self-training with deep networks on unlabeled data.
In *International Conference on Learning Representations*.

Zhang, S., Hao, Y., Wang, M., and Chow, J. H. (2018).
Multi-channel Hankel matrix completion through nonconvex optimization.
IEEE J. Sel. Topics Signal Process., Special Issue on Signal and Information Processing for Critical Infrastructures, 12(4):617–632.

Zhang, S. and Wang, M. (2019).

Correction of corrupted columns through fast robust hankel matrix completion.
IEEE Transactions on Signal Processing, 67(10):2580–2594.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. (2020a).
Fast learning of graph neural networks with guaranteed generalizability:one-hidden-layer case.
In *2020 International Conference on Machine Learning (ICML)*.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. (2021).
Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks.
Proceedings of the 35th International Conference on Neural Information Processing Systems.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. (2022).
How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis.
International Conference on Learning Representations (ICLR), 2022.

Zhang, S., Wang, M., Xiong, J., Liu, S., and Chen, P.-Y. (2020b).
Improved linear convergence of training cnns with generalizability guarantees: A one-hidden-layer case.
IEEE Transactions on Neural Networks and Learning Systems, 32(6):2622–2635.

Zhang, X., Yu, Y., Wang, L., and Gu, Q. (2019).
Learning one-hidden-layer relu networks via gradient descent.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1524–1534.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. (2017).
Recovery guarantees for one-hidden-layer neural networks.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149. JMLR. org, https://arxiv.org/abs/1706.03175.