How does semi-supervised learning with pseudo-labelers work? A case study

Anonymous Author(s)
Affiliation
Address
email

Abstract

Semi-supervised learning is a popular machine learning paradigm that utilizes a large amount of unlabeled data as well as a small amount of labeled data to facilitate learning tasks. While semi-supervised learning has achieved great success in training neural networks, its theoretical understanding remains largely open. In this paper, we aim to theoretically understand a semi-supervised learning approach based on pre-training and linear probing. We prove that, under a certain data generation model and two-layer convolutional neural network, the semi-supervised learning approach can achieve nearly zero test loss, while a neural network directly trained by supervised learning on the same amount of labeled data can only achieve constant test loss. Through this case study, we demonstrate a separation between semi-supervised learning and supervised learning in terms of test loss provided the same amount of labeled data.

1 Introduction

Semi-supervised learning (Scudder, 1965; Fralick, 1967; Agrawala, 1970), which leverages both a small amount of labeled data and a large amount of unlabeled data to improve learning performance, is one of the most widely used approaches. It has been shown to achieve promising performance for a wide variety of tasks, including image classification (Rasmus et al., 2015; Springenberg, 2015; Laine and Aila, 2016), image generation (Kingma et al., 2014; Odena, 2016; Salimans et al., 2016), domain adaptation (Saito et al., 2017; Shu et al., 2018; Lee et al., 2019), and word embedding (Turian et al., 2010; Peters et al., 2017). One of the popular semi-supervised learning approaches is *pseudo-labeling* (Lee et al., 2013; Xie et al., 2020; Pham et al., 2021b; Rizve et al., 2021), which generates pseudo-labels of unlabeled data for pre-training. This approach has been remarkably successful in improving performance on many tasks. In this paper, we attempt to theoretically explain the success of semi-supervised learning with pseudo-labelers in training neural networks. The contributions of our work are summarized as follows.

- We theoretically show that with the help of pseudo-labelers, CNN can learn the feature representation during the pre-training stage. Moreover, the learned feature is highly correlated with the true labels of the data, even though the true labels are not used during the pre-training stage.
- Based on our analysis of the pre-training process, we further show that when linear-probing the pre-trained model in the downstream task, the final classifier can achieve near-zero test loss and test error. Notably, these guarantees of small test loss and error only require a very small number of labeled training data.
- As a comparison, we show that standard supervised learning cannot learn a good classifier under the same setting. Specifically, we show that, even when the training process converges to a global minimum of the training loss, the learned two-layer CNN can only achieve constant level test loss. This, together with the aforementioned results for semi-supervised learning, demonstrates the advantage of semi-supervised learning over standard supervised learning.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

2 Problem Setup and Preliminaries

In this section, we will introduce our data model, the convolutional neural network, and the details of the training algorithms considered in this paper. Inspired by recent work (Allen-Zhu and Li, 2020b; Zou et al., 2021; Shen et al., 2022; Cao et al., 2022), we consider a data model where each data input $\mathbf{x}$ consists of two patches $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, where each patch has d dimensions. We focus on the binary classification task and present our data distribution $\mathcal{D}$ as follows.

Data distribution. Each data point $(\mathbf{x}, y)$ with $\mathbf{x} = [\mathbf{x}^{(1)\top}, \mathbf{x}^{(2)\top}]^\top \in \mathbb{R}^{2d}$ and $y \in \{-1, +1\}$ is generated as follows: the label y is generated as a Rademacher random variable; one of $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ is given by the feature vector $y \cdot \mathbf{v}$, the other is given by a noise vector $\boldsymbol{\xi}$ that is generated from a dd-dimensional Gaussian distribution $\mathcal{N}\big(\mathbf{0}, \sigma_p^2(\mathbf{I} - \mathbf{v}\mathbf{v}^\top/\|\mathbf{v}\|_2^2)\big)$. We denote by $\mathcal{D}$ the joint distribution of $(\mathbf{x}, y)$, and denote by $\mathcal{D}_\mathbf{x}$ the marginal distribution of $\mathbf{x}$.

2.1 Supervised Learning Models

For supervised learning, we consider a two-layer CNN whose filters are applied to the patches $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ respectively and parameters in the second layers are set to be ± 1. Then the CNN can be written as $f_\mathbf{W}(\mathbf{x}) = f_\mathbf{W}^{+1}(\mathbf{x}) - f_\mathbf{W}^{-1}(\mathbf{x})$ where $f_\mathbf{W}(\mathbf{x})^{+1}, f_\mathbf{W}(\mathbf{x})^{-1}$ are formulated as

$$f_\mathbf{W}^{+1}(\mathbf{x}) = \sum_{j=1}^{m} \Big[\sigma\big(\langle \mathbf{w}_j, \mathbf{x}^{(1)}\rangle\big) + \sigma\big(\langle \mathbf{w}_j, \mathbf{x}^{(2)}\rangle\big)\Big], f_\mathbf{W}^{-1}(\mathbf{x}) = \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \mathbf{w}_j, \mathbf{x}^{(1)}\rangle\big) + \sigma\big(\langle \mathbf{w}_j, \mathbf{x}^{(2)}\rangle\big)\Big]. \tag{2.1}$$

Here σ is activation function $\mathrm{ReLU}^q(\cdot) = [\,\cdot\,]_+^q (q > 2)$, m is the width of the network, $\mathbf{w}_j \in \mathbb{R}^d$ denotes the j-th filter, and $\mathbf{W}$ is the collection of all filters $\{\mathbf{w}_j\}_{j=1}^{2m}$. Given labeled training dataset $S' = \{(\mathbf{x}_i', y_i')\}_{i=1}^{n_l}$, we train the CNN model by minimizing the empirical cross-entropy loss

$$L_{S'}(\mathbf{W}) = \frac{1}{n_l} \sum_{i=1}^{n_l} L_i(\mathbf{W}),$$

where $L_i(\mathbf{W}) = \ell\big(y_i' \cdot f_\mathbf{W}(\mathbf{x}_i')\big)$ with $\ell(z) = \log(1 + \exp(-z))$ denotes the individual loss for the training example $(\mathbf{x}_i, y_i)$. We minimize the empirical function $L_{S'}(\mathbf{W})$ with gradient descent as follows

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} - \eta \cdot \nabla_{\mathbf{w}_j} L_{S'}(\mathbf{W}^{(t)}), \quad \mathbf{w}_j^{(0)} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}), \quad j \in [2m],$$

where $\eta > 0$ is the learning rate and σ_0 defines the scale of random initialization.

2.2 Semi-supervised Learning Models

For semi-supervised pre-training, we assume that we have access to K pseudo-labelers $\{f_k^{\mathrm{w}}\}_{k=1}^K$. The accuracy of k-th pseudo-labeler is $p_k \in (1/2, 1)$. Then we use K pseudo-labelers to generate K pseudo-labeled dataset $\{S_k\}_{k=1}^K$, where $S_k := \big\{(\mathbf{x}_i, \widehat{y}_{k,i}) \,\big|\, \widehat{y}_{k,i} = f_k^{\mathrm{w}}(\mathbf{x}_i)\big\}_{i=1}^{n_u}$. Next we solve K pre-training tasks with two-layer CNN models $\{f_{\mathbf{W}_k}\}_{k=1}^K$ defined in (2.1) using $\{S_k\}_{k=1}^K$ respectively. We consider learning the model parameter $\mathbf{W}_k$ by optimizing the empirical loss of both pseudo-labeled dataset S_k and labeled dataset $S' = \{(\mathbf{x}_i', y_i')\}_{i=1}^{n_l}$ with weight decay regularization

$$L_{S_k \cup S'}(\mathbf{W}_k) = \frac{1}{n_u + n_l}\bigg(\sum_{i=1}^{n_u} L_i(\mathbf{W}_k) + \sum_{i'=1}^{n_l} L_{i'}(\mathbf{W}_k)\bigg) + \frac{\lambda}{2}\|\mathbf{W}_k\|_F^2,$$

where $\lambda \geq 0$ is the regularization parameter, $L_i(\mathbf{W}_k) = \ell\big(\widehat{y}_{k,i} \cdot f_{\mathbf{W}_k}(\mathbf{x}_i)\big)$ denotes the individual loss for the pseudo-labeled data $L_{i'}(\mathbf{W}_k) = \ell(y_i' \cdot f_{\mathbf{W}_k}(\mathbf{x}_i'))$ denotes the individual loss for the labeled data $(\mathbf{x}_i', y_i')$. We also use gradient descent to minimize the regularized loss function $L_{S_k \cup S'}(\mathbf{W}_k)$ starting from $\mathbf{w}_{k,j}^{(0)} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_d)$.

Downstream Task: Linear Model. The semi-supervised pre-training gives us K CNN models with parameters $\{\mathbf{W}_k^*\}_{k=1}^K$. Based on them, for the downstream task, we consider a linear model

$$g_\mathbf{a}(\mathbf{x}) = \sum_{k=1}^{K} a_k f_{\mathbf{W}_k^*}(\mathbf{x}),$$

where $a_k \in \mathbb{R}$ denotes the trainable weight for the k-th pre-trained model. Then, given $\{f_{\mathbf{W}_k^*}\}_{k=1}^K$ and labeled training data $S' = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^n$, we consider learning the downstream linear model parameter $\mathbf{a}$ by optimizing the following empirical loss

$$L_{S'}(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(y'_i \cdot g_{\mathbf{a}}(\mathbf{x}'_i)\big).$$

We initialize $\mathbf{a}$ as an all-zero vector and optimize the empirical loss by gradient descent with learning rate η, i.e.,

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \eta \cdot \nabla_{\mathbf{a}} L_{S'}(\mathbf{a}^{(t)}), \ \mathbf{a}^{(0)} = \mathbf{0}.$$

3 Main Results

In this section, we start with a condition that is required by our analysis.

Condition 3.1. The strength of the signal is $\|\mathbf{v}\|_2^2 = \Theta(d)$, the noise variance is $\sigma_p = \Theta(d^{\epsilon})$, where $0 < \epsilon < 1/8$ is a small constant, and the width of the network satisfies $m = \mathrm{polylog}(d)$. We also assume that the size of the unlabeled dataset $n_{\mathrm{u}} = \Omega(d^{4\epsilon})$, and labeled data $n_{\mathrm{l}} = \widetilde{\Theta}(1)$. For both supervised learning and semi-supervised learning settings, we initialize the weight with $\sigma_0 = \Theta(d^{-3/4})$. For semi-supervised learning, we require $\lambda = o(d^{3/4})$ and assume that there exists a constant C such that for all pseudo-labelers, their test accuracy $p_k > 1/2 + C$.

Next, we present the main theoretical results in this paper.

Theorem 3.2 (Semi-supervised Learning: Pre-training)**.** Let $k \in [K]$ and consider the semi-supervised pre-training of $f_{\mathbf{W}_k}(\mathbf{x})$. For any test data point $(\mathbf{x}, y)$, denote $\widehat{y} = f_k^{\mathrm{w}}(\mathbf{x})$. Then under Condition 3.1, after $T_0 = \widetilde{\Theta}(d^{q/4-3/2}\eta^{-1})$ training iterations with learning rate $\eta = O(d^{-1.1})$, the trained neural network $f_{\mathbf{W}_k^{(T_0)}}(\mathbf{x})$ can achieve nearly 0 test error on the distribution $\mathcal{D}$.

Theorem 3.2 characterizes the prediction power of the feature representation learned in the pre-trained models using unlabeled data. For any test data point $(\mathbf{x}, y)$, the sign of y can be predicted based on $f_{\mathbf{W}^{(T_0)}}(\mathbf{x})$ with high probability.

Theorem 3.3 (Semi-supervised Learning: Downstream)**.** Let $\big\{f_{\mathbf{W}_k^{(T_0^k)}}\big\}_{k=1}^d$ be the neural networks trained according to the K pre-training tasks, and consider the learning of the downstream task based in $\big\{f_{\mathbf{W}_k^{(T_0^k)}}\big\}_{k=1}^d$. Under Condition 3.1, after $T' = \Theta(d^{0.1}/\eta)$ iterations with learning rate $\eta = \Theta(1)$, with probability $1 - o(1)$, the obtained $\mathbf{a}^{(T')}$ satisfies:

- Training error is 0: $\frac{1}{n}\sum_{i=1}^n \mathbb{1}[y_i \cdot g_{\mathbf{a}^{(T')}}(\mathbf{x}_i) \le 0] = 0$.
- Test error and loss are nearly 0: $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot g_{\mathbf{a}^{(T')}}(\mathbf{x}) \le 0] = o(1), L_{\mathcal{D}}\big(\mathbf{a}^{(T')}\big) = o(1)$.

Theorem 3.3 shows that the feature representation learned based on the semi-supervised pre-training can ensure small training and test errors for the supervised downstream task. Notably, this result holds even though we assume that there are only a constant number of labeled data. This shows that semi-supervised learning can significantly reduce the need for a large labeled training dataset. For comparison, we also have the following guarantees on the performance of standard supervised learning of CNNs.

Theorem 3.4 (Supervised Learning)**.** Under supervised learning setting, after gradient descent for $T = \widetilde{\Theta}(d^{(1/4-\epsilon)q-3/2}\eta^{-1})$ iterations with learning rate $\eta = O(d^{-1-2\epsilon})$, then there exists $t \le T$ such that with probability $1 - o(1)$ the CNNs defined in (2.1) with parameter $\mathbf{W}^{(t)}$ satisfies:

- Training loss is nearly zero: $L_{S'}\big(\mathbf{W}^{(t)}\big) = o(1)$.
- Test loss is high: $L_{\mathcal{D}}\big(\mathbf{W}^{(t)}\big) = \Theta(1)$.

Theorem 3.4 shows that although standard supervised learning can train a CNN model with nearly zero training loss, the obtained CNN model generalizes poorly to test data. Comparing Theorem 3.4 with Theorem 3.3 shows that the generalization of semi-supervised learning and supervised learning are largely different. The reason behind this difference is that the pre-training, with a relatively large number of unlabeled training data, helps learn a feature representation that captures the feature in

Table 1: Training error and loss, test error and loss for semi-supervised and supervised learning.

	Semi-supervised		Supervised
	Pre-train	Downstream	
Training error	0.1753 ± 0.0259	0	0
Test error	0	0	0.4982 ± 0.0208
Training loss	0.4155 ± 0.0418	0.0150 ± 0.0022	$(6.473 \pm 5.031) \times 10^{-7}$
Test loss	0.2200 ± 0.0886	0.0182 ± 0.0021	0.6931 ± 0.0005

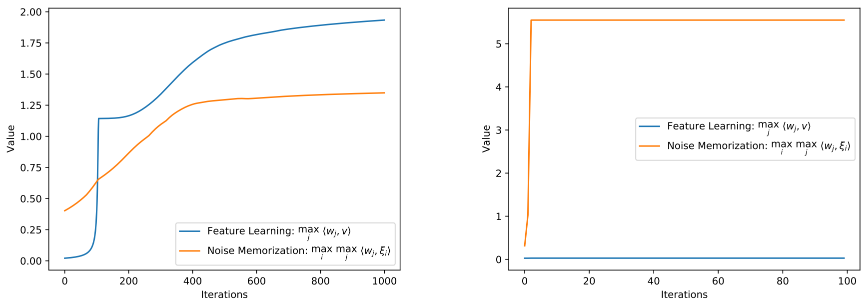

Figure 1: Visualization of the feature learning and noise memorization in the training process. (Left: Semi-supervised, Right: Supervised)

our data model, while direct application of supervised learning can only memorize the noises in the training dataset, which is independent of the labels of the data.

4 Experiments

In this section, we perform numerical experiments on synthetic datasets, generated according to the data distribution in Section 2, to verify our main theoretical results. The detailed experiment setting can be seen from Appendix B.

For semi-supervised learning, we first use a plain classifier to generate n_u pseudo-labels for unlabeled samples in order to help semi-supervised learning. After that, for pre-training, we use these pseudo-labeled samples and n_l labeled samples together to train a CNN. After 200 iterations, we can obtain a CNN model with a training error close to the error of pseudo-labeler and zero test error, according to Table 1. For the downstream task, we use n_l labeled samples to train a linear probe. After 100 iterations, we can obtain a final model with low training and test loss as well as 100% training accuracy and test accuracy. For supervised learning, we directly use n_l labeled data to train the same CNN model. After 200 iterations, we obtain a CNN with 0 training error and small training loss, about 0.5 test error, and high test loss, which indicates supervised learning will give a model that behaves badly and even no better than a random guess.

Moreover, we also calculate the inner products representing feature learning and noise memorization respectively, to verify our key lemmas. The results are reported in Figure 1. It can be seen from Figure 1 that under semi-supervised learning setting the algorithm will the feature learning will dominate the noise memorization though the noise patch has a larger norm than the signal patch, while under the supervised learning setting, the algorithm will entirely forget the feature but fit noise.

5 Conclusion

In this paper, we study semi-supervised learning with pseudo-labelers and provide a theoretical understanding of the success of semi-supervised learning. We show the advantage of semi-supervised learning over supervised learning through a case study. By considering a simple data model and two-layer CNN, we present a comprehensive analysis of the training procedure from a beyond-NTK feature learning perspective. We prove that the final classifier of a semi-supervised learning scenario can achieve near-zero test loss and error with only a small number of labeled training data, while its supervised-learned counterpart fails to achieve the same performance with the same data complexity.

References

Agrawala, A. (1970). Learning with a probabilistic teacher. *IEEE Transactions on Information Theory* **16** 373–379.

Allen-Zhu, Z. and Li, Y. (2020a). Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190* .

Allen-Zhu, Z. and Li, Y. (2020b). Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816* .

Balcan, M.-F. and Blum, A. (2010). A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)* **57** 1–46.

Belkin, M., Niyogi, P. and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research* **7**.

Bennett, K. and Demiriz, A. (1998). Semi-supervised support vector machines. *Advances in Neural Information processing systems* **11**.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*.

Brock, A., Donahue, J. and Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* .

Cao, Y., Chen, Z., Belkin, M. and Gu, Q. (2022). Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526* .

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems* **33** 9912–9924.

Castelli, V. and Cover, T. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory* **42** 2102–2117.

Castelli, V. and Cover, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters* **16** 105–111.

Chapelle, O., Scholkopf, B. and Zien, A. (2010). *Semi-supervised learning*. The MIT Press.

Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR.

Darnstädt, M. (2015). *An investigation on the power of unlabeled data*. Ph.D. thesis, Bochum, Ruhr-Universität Bochum, Diss., 2015.

Fralick, S. (1967). Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory* **13** 57–64.

Frei, S., Chatterji, N. S. and Bartlett, P. L. (2022a). Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *arXiv preprint arXiv:2202.05928* .

Frei, S., Zou, D., Chen, Z. and Gu, Q. (2022b). Self-training converts weak learners to strong learners in mixture models. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Gidaris, S., Singh, P. and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* .

Globerson, A., Livni, R. and Shalev-Shwartz, S. (2017). Effective semisupervised learning on manifolds. In *Conference on Learning Theory*. PMLR.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems* **27**.

HaoChen, J. Z., Wei, C., Gaidon, A. and Ma, T. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems* **34**.

He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Jing, L., Yang, X., Liu, J. and Tian, Y. (2018). Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387* .

Karras, T., Laine, S. and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kingma, D. P., Mohamed, S., Jimenez Rezende, D. and Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in neural information processing systems* **27**.

Laine, S. and Aila, T. (2016). Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242* .

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, vol. 3.

Lee, H.-Y., Huang, J.-B., Singh, M. and Yang, M.-H. (2017). Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*.

Lee, J. D., Lei, Q., Saunshi, N. and Zhuo, J. (2020). Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064* .

Lee, S., Kim, D., Kim, N. and Jeong, S.-G. (2019). Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, Y., Pan, Q., Wang, S., Peng, H., Yang, T. and Cambria, E. (2019). Disentangled variational auto-encoder for semi-supervised learning. *Information Sciences* **482** 73–85.

Miller, D. J. and Uyar, H. (1996). A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in Neural Information Processing Systems* (M. Mozer, M. Jordan and T. Petsche, eds.), vol. 9. MIT Press.

Misra, I., Zitnick, C. L. and Hebert, M. (2016). Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*. Springer.

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L. and Blundell, C. (2020). Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922* .

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning* **39** 103–134.

Niyogi, P. (2013). Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research* **14**.

NOROOZI, M. and FAVARO, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*. Springer.

ODENA, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583* .

OYMAK, S. and GULCU, T. C. (2021). Statistical and algorithmic insights for semi-supervised learning with self-training. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

PATHAK, D., KRAHENBUHL, P., DONAHUE, J., DARRELL, T. and EFROS, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

PETERS, M. E., AMMAR, W., BHAGAVATULA, C. and POWER, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108* .

PHAM, H., DAI, Z., XIE, Q. and LE, Q. V. (2021a). Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

PHAM, H., DAI, Z., XIE, Q., LUONG, M.-T. and LE, Q. V. (2021b). Meta pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

RASMUS, A., BERGLUND, M., HONKALA, M., VALPOLA, H. and RAIKO, T. (2015). Semi-supervised learning with ladder networks. *Advances in neural information processing systems* **28**.

RIZVE, M. N., DUARTE, K., RAWAT, Y. S. and SHAH, M. (2021). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

SAITO, K., USHIKU, Y. and HARADA, T. (2017). Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*. PMLR.

SAJJADI, M., JAVANMARDI, M. and TASDIZEN, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems* **29**.

SALIMANS, T., GOODFELLOW, I., ZAREMBA, W., CHEUNG, V., RADFORD, A. and CHEN, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems* **29**.

SAUNSHI, N., ASH, J., GOEL, S., MISRA, D., ZHANG, C., ARORA, S., KAKADE, S. and KRISHNAMURTHY, A. (2022). Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037* .

SAUNSHI, N., PLEVRAKIS, O., ARORA, S., KHODAK, M. and KHANDEPARKAR, H. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*. PMLR.

SCUDDER, H. (1965). Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* **11** 363–371.

SHEN, R., BUBECK, S. and GUNASEKAR, S. (2022). Data augmentation as feature manipulation. In *International Conference on Machine Learning*. PMLR.

SHU, R., BUI, H. H., NARUI, H. and ERMON, S. (2018). A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735* .

SINGH, A., NOWAK, R. and ZHU, J. (2008). Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.), vol. 21. Curran Associates, Inc.

SPRINGENBERG, J. T. (2015). Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390* .

TARVAINEN, A. and VALPOLA, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* **30**.

TIAN, Y., YU, L., CHEN, X. and GANGULI, S. (2020). Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578* .

TOSH, C., KRISHNAMURTHY, A. and HSU, D. (2021a). Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research* **22** 1–31.

TOSH, C., KRISHNAMURTHY, A. and HSU, D. (2021b). Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research* **22** 1–31.

TSAI, Y.-H. H., WU, Y., SALAKHUTDINOV, R. and MORENCY, L.-P. (2020). Demystifying self-supervised learning: An information-theoretical framework. *arXiv preprint arXiv:2006.05576* .

TULYAKOV, S., LIU, M.-Y., YANG, X. and KAUTZ, J. (2018). Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

TURIAN, J., RATINOV, L. and BENGIO, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*.

VONDRICK, C., PIRSIAVASH, H. and TORRALBA, A. (2016). Generating videos with scene dynamics. *Advances in neural information processing systems* **29** 613–621.

WANG, T. and ISOLA, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*. PMLR.

WEI, C., SHEN, K., CHEN, Y. and MA, T. (2020). Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622* .

WEI, D., LIM, J. J., ZISSERMAN, A. and FREEMAN, W. T. (2018). Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

WEN, Z. and LI, Y. (2021). Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*. PMLR.

XIE, Q., LUONG, M.-T., HOVY, E. and LE, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

XU, Z., JIN, R., ZHU, J., KING, I. and LYU, M. (2007). Efficient convex relaxation for transductive support vector machine. *Advances in neural information processing systems* **20**.

XU, Z., JIN, R., ZHU, J., KING, I., LYU, M. and YANG, Z. (2009). Adaptive regularization for transductive support vector machine. *Advances in Neural Information Processing Systems* **22**.

ZHAI, X., OLIVER, A., KOLESNIKOV, A. and BEYER, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

ZHANG, R., ISOLA, P. and EFROS, A. A. (2016). Colorful image colorization. In *European conference on computer vision*. Springer.

ZHOU, D., BOUSQUET, O., LAL, T., WESTON, J. and SCHÖLKOPF, B. (2003). Learning with local and global consistency. *Advances in neural information processing systems* **16**.

Zhu, X., Ghahramani, Z. and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*.

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* **3** 1–130.

Zou, D., Cao, Y., Li, Y. and Gu, Q. (2021). Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371* .

A Related Work

Semi-supervised learning methods in practice. Since the invention of semi-supervised learning in Scudder (1965); Fralick (1967); Agrawala (1970), a wide range of semi-supervised learning approaches have been proposed, including generative models (Miller and Uyar, 1996; Nigam et al., 2000), semi-supervised support vector machines (Bennett and Demiriz, 1998; Xu et al., 2007, 2009), graph-based methods (Zhu et al., 2003; Belkin et al., 2006; Zhou et al., 2003), and co-training (Blum and Mitchell, 1998), etc. For a comprehensive review of classical semi-supervised learning methods, please refer to Chapelle et al. (2010); Zhu and Goldberg (2009). In the past years, a number of deep semi-supervised learning approaches have been proposed, such as generative methods (Odena, 2016; Li et al., 2019), consistency regularization methods (Sajjadi et al., 2016; Laine and Aila, 2016; Rasmus et al., 2015; Tarvainen and Valpola, 2017) and pseudo-labeling methods (Lee et al., 2013; Zhai et al., 2019; Xie et al., 2020; Pham et al., 2021a). In this work, we will focus on pseudo-labeling methods.

Theory of semi-supervised learning. To understand semi-supervised learning, Castelli and Cover (1995, 1996) studied the relative value of labeled data over unlabeled data under a parametric assumption on the marginal distribution of input features. Later, a series of works proved that semi-supervised learning can possess better sample complexity or generalization performance than supervised learning under certain assumptions on the marginal distribution (Niyogi, 2013; Globerson et al., 2017) or the ratio of labeled and unlabeled samples (Singh et al., 2008; Darnstädt, 2015), while Balcan and Blum (2010) provided a unified PAC framework able to analyze both sample-complexity and algorithmic issues. Oymak and Gulcu (2021); Frei et al. (2022b) considered semi-supervised learning with pseudo-labers by learning a linear classifier for mixture models and convergence to Bayes-optimal predictor.

Self-supervised learning in practice. A closely related learning paradigm to semi-supervised learning is called self-supervised learning, which creates human-designed supervised learning problems to leverage natural structures and learn representations from unlabeled data. Representative self-supervised learning approaches include *contrastive learning* and *pretext-based self-supervised learning*. Contrastive learning (Caron et al., 2020; He et al., 2020; Chen et al., 2020) aims to group similar examples closer and dissimilar examples far from each other by utilizing a similarity metric, while pretext-based self-supervised tries to learn a good representation from *pretext tasks* generated from the unlabeled data to facilitate *downstream learning tasks*. In practice, various pretext tasks have been proposed, which include (1) generation-based ones such as colorizing grayscale images (Zhang et al., 2016), image inpainting (Pathak et al., 2016), image and video generation with GAN (Goodfellow et al., 2014; Brock et al., 2018; Karras et al., 2019; Vondrick et al., 2016; Tulyakov et al., 2018); and (2) context-based ones such as image jigsaw puzzle (Noroozi and Favaro, 2016), geometric transformation (Gidaris et al., 2018; Jing et al., 2018), frame order verification and recognition (Lee et al., 2017; Misra et al., 2016; Wei et al., 2018). The semi-supervised learning approach with pseudo-labelers studied in this paper is related to pretext-based self-supervised learning because the unlabeled data with pseudo-labels can be seen as a particular pretext task.

Theory of self-supervised learning. In order to understand self-supervised learning, there is a line of work towards understanding *contrastive learning* (Saunshi et al., 2019; Tsai et al., 2020; Mitrovic et al., 2020; Tian et al., 2020; Wang and Isola, 2020; Tosh et al., 2021a,b; HaoChen et al., 2021; Wen and Li, 2021; Saunshi et al., 2022), which is one of the most used self-supervised learning approaches based on data augmentation. Unlike contrastive learning, the theoretical understanding of pretext-based self-supervised learning is still rather limited. The only notable works are Lee et al. (2020) and Wei et al. (2020). Lee et al. (2020) proved generalization guarantees for self-supervised algorithms using empirical risk minimization on the pretext task under certain conditional independence assumptions. Wei et al. (2020) proved that under an "expansion" assumption, the minimizer of the population loss based on self-training and input-consistency regularization will achieve high prediction accuracy. Since semi-supervised learning with pseudo-labelers can be seen as a special case of pretext-based self-supervised learning (the pretext task is generated by the pseudo-labelers), we believe the case study in the current paper and its theoretical understanding can shed light on pretext-based self-supervised learning as well.

Feature learning by neural networks. Our work is also closely related to several recent works that study how neural networks learn the features. Allen-Zhu and Li (2020a) showed that adversarial training purifies the learned features by removing certain "dense mixtures" in the hidden layer weights of the network. Allen-Zhu and Li (2020b) studied how ensemble and knowledge distillation work in deep learning when the data have "multi-view" features. Zou et al. (2021) studied an aspect of feature learning by Adam and GD and showed that GD can learn the sparse features while Adam may fail even with proper regularization. Notably, there are two concurrent works studying the benign overfitting phenomenon in learning neural networks: Frei et al. (2022a) established theoretical guarantees for benign overfitting of two-layer fully connected neural networks with zero training error and test error close to the Bayes-optimal error, while Cao et al. (2022) studied the benign overfitting phenomenon in training a two-layer convolutional neural network (CNN), achieving arbitrarily small training and test loss. Our work studies a different aspect of feature learning afforded by semi-supervised learning versus supervised learning: given a small amount of labeled data, semi-supervised learning can learn the features with the help of pseudo-labelers, while supervised learning fails to learn the features and tends to overfit the noise in the training data.

Comparison with related work. A recent line of work (Oymak and Gulcu, 2021; Frei et al., 2022b) studies the semi-supervised learning methods with pseudo-labelers. Our results are different from theirs in several aspects: (i) we are considering learning with CNNs rather than a linear model, so the problem is highly non-convex with various local minima, which makes the optimization analysis more challenging; (ii) the Bayesian optimal predictor is no longer unique for CNNs. Therefore, we measure the quality of the learned features via downstream task instead of making a comparison with the Bayesian optimal predictor; (iii) They can only deal with the case where the teacher network (pseudo-labeler) is the same as the student network (Frei et al., 2022b) or the case where the teacher network (pseudo-labeler) is at least as complex as the student network (Oymak and Gulcu, 2021). However, our teacher network (pseudo-labeler) is not specified and can be any structure, such as a linear network. Therefore we can handle the case where the student network is more complex than the teacher network, one of the most natural settings for semi-supervised learning with pseudo-labeler (Xie et al., 2020).

B Experiment Setting

In particular, we set the problem dimension $d = 10000$, labeled training sample size $n_{\mathrm{l}} = 20$ (10 positive samples and 10 negative samples), pseudo-labeled training sample size $n_{\mathrm{u}} = 20000$ (10000 positive samples and 10000 negative samples), feature vector $\mathbf{v}$ sampled from distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and noise vector sampled from distribution $\mathcal{N}(\mathbf{0}, \sigma_p^2 \boldsymbol{I})$ where $\sigma_p = 10d^{0.01}$.

For semi-supervised learning tasks, we have a linear pseudo-labeler with test error 0.196 ± 0.044. Then, we use this classifier to generate pseudo-labels for $n_u = 20000$ unlabeled samples in order to help semi-supervised learning. After that, for pre-training, we use these pseudo-labeled samples and n_l labeled samples together to train a CNN with network width $m = 20$, activation function $\sigma(z) = [z]_+^3$, regularization parameter $\lambda = 0.1$ and learning rate $\eta = 1 \times 10^{-4}$. Besides, we initialize CNN parameters from $\mathcal{N}(0, \sigma_0^2)$, where $\sigma_0 = 0.1 \times d^{-3/4}$. After 200 iterations, we can obtain a CNN model with a training error close to the error of pseudo-labeler and zero test error, according to Table 1. For a downstream task, we use n_{l} labeled samples to train a linear probe. By applying learning rate $\eta = 0.1$ and after $T = 100$ iterations, we can obtain a final model with low training and test loss as well as 100% training accuracy and test accuracy.

For supervised learning task, we directly use n_{l} labeled data to train a CNN with network width $m = 20$, activation function $\sigma(z) = [z]_+^3$, learning rate $\eta = 1 \times 10^{-4}$. After 200 iterations, we obtain a CNN with 0 training error and small training loss, about 0.5 test error, and high test loss, which indicates supervised learning will give a model that behaves badly and even no better than a random guess.

C Proof for Semi-supervised Learning Setting

We consider learning K functions $f_{\mathbf{W}_k}(\mathbf{x}), k \in [K]$ based on the pre-training. Since the learning process of these K functions can be analyzed in exactly the same way, here we only focus on the learning of one of these functions. For simplicity of notation, we drop the subscript k in the following proof for Sections C.2, C.3, C.4, C.5, C.6, C.7 and C.8. We start with a condition that is required by our analysis.

Condition C.1. The strength of the signal is $\|\mathbf{v}\|_2^2 = \Theta(d)$, the noise variance is $\sigma_p = \Theta(d^\epsilon)$, where $0 < \epsilon < 1/8$ is a small constant, and the width of the network satisfies $m = \mathrm{polylog}(d)$. We also assume that the size of the unlabeled dataset $n_\mathrm{u} = \Omega(d^{4\epsilon})$, and labeled data $n_\mathrm{l} = \widetilde{\Theta}(1)$. For both supervise learning and semi-supervised learning settings, we initialize the weight with $\sigma_0 = \Theta(d^{-3/4})$. For semi-supervised learning, we require $\lambda = o(d^{3/4})$ and assume that there exists a constant C such that for all pseudo-labelers, their test accuracy $p_k > 1/2 + C$.

Since we generate the noise patch from the Gaussian distribution, the strength of the noise patch is $\|\boldsymbol{\xi}\|_2^2 \approx d^{1+\epsilon}$ by standard concentration inequalities, which is larger than the strength of the signal patch $\|\mathbf{v}\|_2^2 = \Theta(d)$. Therefore, Condition 3.1 defines a setting with large noises. The condition of $d \gg n_u \gg n_l$ further ensures that learning is in a sufficiently over-parameterized setting. Here we only require the neural network width m to be polylogarithmic in the dimension d and require the pseudo-lablers to perform better than a random guess.

Notation. We use lower case letters, lower case bold face letters, and upper case bold face letters to denote scalars, vectors, and matrices respectively. For a scalar x, we use $[x]_+$ to denote $\max\{x, 0\}$. For a vector $\mathbf{v} = (v_1, \cdots, v_d)^\top$, we denote by $\|\mathbf{v}\|_2 := \big(\sum_{i=1}^d v_i^2\big)^{\frac{1}{2}}$ its ℓ_2 norm, and use $\mathrm{supp}(\mathbf{v}) := \{j : v_j \neq 0\}$ to denote its support. For two sequences $\{a_k\}$ and $\{b_k\}$, we denote $a_k = O(b_k)$ if $|a_k| \leq C|b_k|$ for some absolute constant C, denote $a_k = \Omega(b_k)$ if $b_k = O(a_k)$, and denote $a_k = \Theta(b_k)$ if $|a_k| \leq C|b_k|$ and $a_k = \Omega(b_k)$. We also denote $a_k = o(b_k)$ if $\lim |a_k/b_k| = 0$. Finally, we use $\widetilde{\Theta}(\cdot)$, $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to omit logarithmic terms in the notations.

C.1 Proof Sketch

In this section, we present the proof sketch for the semi-supervised learning setting.

Semi-supervised Pre-training. We consider learning K functions $f_{\mathbf{W}_k}(\mathbf{x})$, $k \in [K]$ based on the pre-training. Since the learning process of these K functions can be analyzed in exactly the same way, here we only focus on the learning of one of these functions. For simplicity of notation, we drop the subscript k in the following proof sketch.

Our study of the pre-training focuses on two aspects of the training process: *feature learning* and *noise memorization*. Specifically, we aim to monitor how the filters in the CNN model learn the feature vector $\mathbf{v}$ and the noise vectors $\boldsymbol{\xi}_i$'s. Therefore, we introduce the following notations.

$$
\begin{aligned}
&\widehat{\Lambda}_1^{(t)} := \max_{1\leq j\leq m} \langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle,\ \bar{\Lambda}_1^{(t)} := \max_{1\leq j\leq m} -\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle,\\
&\widehat{\Lambda}_{-1}^{(t)} := \max_{m+1\leq j\leq 2m} -\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle,\ \bar{\Lambda}_{-1}^{(t)} := \max_{m+1\leq j\leq 2m} \langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle,\\
&\Gamma_i^{(t)} := \max_{1\leq j\leq 2m} \langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i\rangle,\ \Gamma_i'^{(t)} := \max_{1\leq j\leq 2m} \langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i'\rangle,\ \Gamma^{(t)} = \max\Big\{\max_{i\in[n_\mathrm{u}]} \Gamma_i^{(t)}, \max_{i\in[n_\mathrm{l}]} \Gamma_i'^{(t)},\Big\}.
\end{aligned}
\tag{C.1}
$$

Based on the above definitions for $r \in \{\pm 1\}$, a larger $\widehat{\Lambda}_r^{(t)}$ implies better feature learning along the positive feature direction $\mathbf{v}$, while a larger $\bar{\Lambda}_r^{(t)}$ implies better feature learning along the negative feature direction $-\mathbf{v}$. Moreover, a larger $\Gamma^{(t)}$ implies a higher level of noise memorization.

Based on the update rule of gradient descent, for the inner products $\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle$ and $\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l\rangle$, for $j \in [2m]$, $l \in [n_\mathrm{u}]$, we can obtain iterative equations in (C.2).

With the help of the iterative equations and definitions in (C.1), we can further show the following lemma.

Lemma C.2. Assume we use both unlabeled data with pseudo-labels generated by the pseudo-labeler and labeled data for the training of our CNN model. Then for $r \in \{\pm 1\}$, let T_r be the first iteration that $r\widehat{\Lambda}_r^{(t)}$ reaches $\Theta(1/m)$, then for $t \in [0, T_r]$, we have

$$
\begin{aligned}
&\widehat{\Lambda}_r^{(t+1)} \geq (1-\eta\lambda)\cdot \widehat{\Lambda}_r^{(t)} + \eta\cdot C\cdot \Theta(d)\cdot (\widehat{\Lambda}_r^{(t)})^{q-1}, r \in \{\pm 1\},\\
&\bar{\Lambda}_r^{(t+1)} \leq (1-\eta\lambda)\cdot \bar{\Lambda}_r^{(t)}, r \in \{\pm 1\},\\
&\Gamma^{(t+1)} \leq (1-\eta\lambda)\cdot \Gamma^{(t)} + \eta\cdot \widetilde{\Theta}(d^{1-2\epsilon})\cdot (\Gamma^{(t)})^{q-1},
\end{aligned}
$$

where C is defined in Condition 3.1.

Lemma C.3. Assume we use only labeled data for the training of our CNN model. Then for $i \in [n_l]$, let T_i' be the first iteration that $\Gamma_i'^{(t)}$ reaches $\Theta(1/m)$, then we have

$$\widehat{\Lambda}_r^{(t+1)} \leq (1-\eta\lambda)\cdot\widehat{\Lambda}_r^{(t)} + \eta\cdot\Theta(d)\cdot\big((\widehat{\Lambda}_r^{(t)})^{q-1} + (\bar{\Lambda}_r^{(t)})^{q-1}\big), r \in \{\pm 1\},$$
$$\bar{\Lambda}_r^{(t+1)} \leq (1-\eta\lambda)\cdot\bar{\Lambda}_r^{(t)}, r \in \{\pm 1\},$$
$$\Gamma_i'^{(t+1)} \geq (1-\eta\lambda)\cdot\Gamma_i'^{(t)} + \eta\cdot\widetilde{\Theta}(d^{1+2\epsilon})\cdot(\Gamma_i'^{(t)})^{q-1}, i \in [n_l], \text{ for } t \in [0, T_i'].$$

Based on the results in Lemma C.2, we can observe that if both pseudo-labeled and labeled data are used for training, the CNN will learn the positive direction of the feature vector $\mathbf{v}$, while barely tending to fit the negative direction of the feature vector or memorize the noise. And if only labeled data are used, the CNN will fit noise faster than a feature, which can be seen from Lemma C.3. Leveraging Lemmas C.2 and C.3, we can obtain the following Lemmas C.4 and C.5, which characterize the magnitude of feature learning and noise memorization.

Lemma C.4. If both pseudo-labeled and labeled data are used to train CNN, for $r \in \{\pm 1\}$, let T_r be the first iteration that $\widehat{\Lambda}_r^{(t)}$ reaches $\Theta(1/m)$ respectively. Let $T_0 = \max_{r\in\{\pm 1\}}\{T_r\}$. Then, it holds that $\widehat{\Lambda}_r^{(T_0)} = \widetilde{\Theta}(1)$, $\bar{\Lambda}_r^{(t)} = \widetilde{O}(d^{-\frac{1}{4}})$ and $\Gamma^{(t)} = \widetilde{O}(d^{-\frac{1}{4}+\epsilon})$ for all $t \in [0, T_0]$.

Lemma C.5. If only labeled data are used to train CNN, for $i \in [n_l]$, let T_i' be the first iteration that $\Gamma_i'^{(t)}$ reaches $\Theta(1/m)$. Let $T_0' = \max_{i\in[n_l]} T_i'$. Then, it holds that $\widehat{\Lambda}_r = \widetilde{O}(d^{-\frac{1}{4}})$, $\bar{\Lambda}_r = \widetilde{O}(d^{-\frac{1}{4}})$ for $r \in \{\pm 1\}$ and $\Gamma_i'^{(t)} = \widetilde{\Theta}(1)$ for $i \in [n_l]$.

The above results indicate the deviation between the two settings. The reason is that assume we consider a sequence $\{x_t\}$ with iterative equation $x_{t+1} = x_t + \eta\cdot C_t x_t^{q-1}$. If we only use labeled data, as shown in Lemma C.3, $\Gamma_i'^{(t)}$ has $C_t = \widetilde{\Theta}(d^{1+2\epsilon})$ while $\widehat{\Lambda}_r^{(t)}$ has $C_t = \Theta(d)$, therefore $\Gamma_i'^{(t)}$ increases faster than $\widehat{\Lambda}_r^{(t)}$. In contrast, if we use both labeled data and pseudo-labeled data, C_t will be $\widetilde{\Theta}(d^{1-2\epsilon})$ for $\Gamma_i'^{(t)}$ and $\Theta(d)$ for $\widehat{\Lambda}_r^{(t)}$, leading to a slower increasing speed of $\Gamma_i'^{(t)}$.

Downstream task. After the pre-training, we have obtained K CNN classifiers $\big\{f_{\mathbf{W}_k^{(T_0^k)}}\big\}_{k=1}^{K}$. Now we train the second-layer parameters $\mathbf{a}$ with the training data whose true labels are available. The following lemma shows that the l_1-norm of $\mathbf{a}$ will increase with a logarithmic order.

Lemma C.6. For any learning rate $\eta = \Theta(1)$, we have $\|\mathbf{a}^{(t)}\|_1 = \log(t)/\widetilde{\Theta}(1)$. For any labeled data $(\mathbf{x}_i', y_i') \in S'$, we have with high probability that $y_i' \cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}_i') = \|\mathbf{a}^{(t)}\|_1 \cdot \widetilde{\Theta}(1)$. For any newly generated data $(\mathbf{x}, y) \sim \mathcal{D}$, we also have with high probability that $y \cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}) = \|\mathbf{a}^{(t)}\|_1 \cdot \widetilde{\Theta}(1)$.

With the help of the above lemma and note that training error and test error are related to $y \cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x})$ and test loss is related to $\|\mathbf{a}^{(T_0)}\|_1$, we can prove that after $T = \Theta(d^{0.1}/\eta)$ iterations with learning rate $\eta = \Theta(1)$, the model can achieve nearly zero training error, test error, training loss and test loss.

C.2 Gradient Calculation

Lemma C.7 (Gradient Calculation)**.** The gradient of loss function $L_S(\mathbf{W})$ with respect to weight parameters $\mathbf{w}_j$ is

$$\nabla_{\mathbf{w}_j} L_{S\cup S'}(\mathbf{W}) = -\frac{q}{n_l + n_u}\Bigg(\sum_{i=1}^{n_u} c_i\widehat{y}_i\big([\langle\mathbf{w}_j, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\cdot y_i\cdot\mathbf{v} + [\langle\mathbf{w}_j, \boldsymbol{\xi}_i\rangle]_+^{q-1}\cdot\boldsymbol{\xi}_i\big)$$
$$+ \sum_{i=1}^{n_l} b_i y_i'\big([\langle\mathbf{w}_j, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\cdot y_i'\cdot\mathbf{v} + [\langle\mathbf{w}_j, \boldsymbol{\xi}_i'\rangle]_+^{q-1}\cdot\boldsymbol{\xi}_i'\big)\Bigg) + \lambda\cdot\mathbf{w}_j,$$

for $1 \leq j \leq m$; and

$$\nabla_{\mathbf{w}_j} L_{S\cup S'}(\mathbf{W}) = \frac{q}{n_l + n_u}\Bigg(\sum_{i=1}^{n_u} c_i\widehat{y}_i\big([\langle\mathbf{w}_j, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\cdot y_i\cdot\mathbf{v} + [\langle\mathbf{w}_j, \boldsymbol{\xi}_i\rangle]_+^{q-1}\cdot\boldsymbol{\xi}_i\big)$$

$$+\sum_{i=1}^{n_{\mathrm{l}}} b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i'\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i'\big)\Bigg) + \lambda \cdot \mathbf{w}_j,$$

for $m+1 \le j \le 2m$, where $-\ell'\big(\widehat{y}_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)\big) = \exp[-\widehat{y}_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)]/(1+\exp[-\widehat{y}_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)])$ is denoted by c_i and $-\ell'(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')) = \exp[-y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')]/(1+\exp[-y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')])$ is denoted by b_i.

Proof of Lemma C.7. When $1 \le j \le m$,

$$\begin{aligned}
\nabla_{\mathbf{w}_j} \ell\big(\widehat{y}_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)\big) &= \ell'\big(\widehat{y}_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)\big) \cdot \widehat{y}_i \cdot \nabla_{\mathbf{w}_j} f_{\mathbf{W}}(\mathbf{x}_i) \\
&= -c_i \cdot \widehat{y}_i \cdot \nabla_{\mathbf{w}_j} f_{\mathbf{W}}(\mathbf{x}_i) \\
&= -c_i \widehat{y}_i \cdot \big(\sigma'\big(\langle \mathbf{w}_j, y_i \cdot \mathbf{v}\rangle\big) \cdot y_i \cdot \mathbf{v} + \sigma'\big(\langle \mathbf{w}_j, \boldsymbol{\xi}_i\rangle\big) \cdot \boldsymbol{\xi}_i\big) \\
&= -q c_i \widehat{y}_i \big([\langle \mathbf{w}_j, y_i \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i\big)
\end{aligned}$$

$$\begin{aligned}
\nabla_{\mathbf{w}_j} \ell\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) &= \ell'\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) \cdot y_i' \cdot \nabla_{\mathbf{w}_j} f_{\mathbf{W}}(\mathbf{x}_i') \\
&= -b_i \cdot y_i' \cdot \nabla_{\mathbf{w}_j} f_{\mathbf{W}}(\mathbf{x}_i') \\
&= -b_i y_i' \cdot \big(\sigma'(\langle \mathbf{w}_j, y_i' \cdot \mathbf{v}\rangle) \cdot y_i' \cdot \mathbf{v} + \sigma'(\langle \mathbf{w}_j, \boldsymbol{\xi}_i'\rangle) \cdot \boldsymbol{\xi}_i'\big) \\
&= -q b_i y_i' \cdot \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i'\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i'\big)
\end{aligned}$$

and when $m+1 \le j \le 2m$,

$$\begin{aligned}
\nabla_{\mathbf{w}_j} \ell\big(\widehat{y}_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)\big) &= q c_i \widehat{y}_i \big([\langle \mathbf{w}_j, y_i \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i\big) \\
\nabla_{\mathbf{w}_j} \ell\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) &= q b_i y_i' \cdot \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i'\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i'\big)
\end{aligned}$$

Note that $\nabla_{\mathbf{w}_j} L_{S\cup S'}(\mathbf{W}) = \big(\sum_{i=1}^{n_{\mathrm{u}}} \nabla_{\mathbf{w}_j} \ell\big(\widehat{y}_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)\big) + \sum_{i=1}^{n_{\mathrm{l}}} \nabla_{\mathbf{w}_j} \ell\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big)\big)/(n_{\mathrm{l}}+n_{\mathrm{u}}) + \lambda \cdot \mathbf{w}_j$, we have proved the lemma. $\square$

C.3 Inner Product Update Rule Calculation

When the model is trained by gradient descent, the update rule can be formulated by

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} - \eta \cdot \nabla_{\mathbf{w}_j} L_S(\mathbf{W}^{(t)}), \quad j \in [2m]. \tag{C.2}$$

We study the performance of entire training process from two perspective: feature learning and noise memorization. Mathematically, we will focus on two quantities: $\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle$ and $\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l\rangle$. And then we have following lemma for the inner product update rule.

Lemma C.8 (Inner Product Update Rule)**.** The feature learning and noise memorization performance of gradient descent can be formulated by

$$\begin{aligned}
\langle \mathbf{w}_j^{(t+1)}, \mathbf{v}\rangle = (1-\eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle + \frac{q\eta u_j}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(&\sum_{i=1}^{n_{\mathrm{u}}} y_i \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_j^{(t)}, y_i \cdot \mathbf{v}\rangle]_+^{q-1} \|\mathbf{v}\|_2^2 \\
&+ \sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)} [\langle \mathbf{w}_j^{(t)}, y_i' \cdot \mathbf{v}\rangle]_+^{q-1} \|\mathbf{v}\|_2^2\Bigg),
\end{aligned}$$

$$\begin{aligned}
\langle \mathbf{w}_j^{(t+1)}, \boldsymbol{\xi}_l\rangle = (1-\eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l\rangle + \frac{q\eta u_j}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(&\sum_{i=1}^{n_{\mathrm{u}}} \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i\rangle]_+^{q-1} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l\rangle \\
&+ \sum_{i=1}^{n_{\mathrm{l}}} y_i' b_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1} \langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l\rangle\Bigg),
\end{aligned}$$

$$\langle \mathbf{w}_j^{(t+1)}, \boldsymbol{\xi}_l'\rangle = (1-\eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l'\rangle + \frac{q\eta u_j}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(\sum_{i=1}^{n_{\mathrm{u}}} \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i\rangle]_+^{q-1} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l'\rangle$$

$$+ \sum_{i=1}^{n_\mathrm{l}} y_i' b_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l' \rangle \Bigg),$$

where $j \in [2m]$, $l \in [n_\mathrm{u}]$ and $u_j := \mathbb{1}_{[1 \le j \le m]} - \mathbb{1}_{[m+1 \le j \le 2m]}$.

Proof of Lemma C.8. According to Lemma C.7 and gradient descent update rule (C.2), we have

$$\mathbf{w}_j^{(t+1)} = (1 - \eta\lambda) \cdot \mathbf{w}_j^{(t)} + \frac{q\eta u_j}{n_\mathrm{l} + n_\mathrm{u}} \cdot \Bigg(\sum_{i=1}^{n_\mathrm{u}} c_i \widehat{y}_i \big([\langle \mathbf{w}_j, y_i \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i \big)$$
$$+ \sum_{i=1}^{n_\mathrm{l}} b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i' \big) \Bigg)$$

Taking inner product with feature vector $\mathbf{v}$ and noise patch $\boldsymbol{\xi}_l$ and note that $\mathbf{v}$ is orthogonal to $\boldsymbol{\xi}_l$ according to the data model, we have

$$\langle \mathbf{w}_j^{(t+1)}, \mathbf{v} \rangle = (1 - \eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \mathbf{v} \rangle + \frac{q\eta u_j}{n_\mathrm{l} + n_\mathrm{u}} \Bigg(\sum_{i=1}^{n_\mathrm{u}} c_i^{(t)} \widehat{y}_i \big([\langle \mathbf{w}_j, y_i \cdot \mathbf{v} \rangle]_+^{q-1} y_i \|\mathbf{v}\|_2^2 + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i, \mathbf{v} \rangle \big)$$
$$+ \sum_{i=1}^{n_\mathrm{l}} b_i^{(t)} y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} y_i' \|\mathbf{v}\|_2^2 + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i', \mathbf{v} \rangle \big) \Bigg)$$
$$= (1 - \eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \mathbf{v} \rangle + \frac{q\eta u_j}{n_\mathrm{l} + n_\mathrm{u}} \Bigg(\sum_{i=1}^{n_\mathrm{u}} y_i \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_j^{(t)}, y_i \cdot \mathbf{v} \rangle]_+^{q-1} \|\mathbf{v}\|_2^2$$
$$+ \sum_{i=1}^{n_\mathrm{l}} b_i^{(t)} [\langle \mathbf{w}_j^{(t)}, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \|\mathbf{v}\|_2^2 \Bigg),$$

$$\langle \mathbf{w}_j^{(t+1)}, \boldsymbol{\xi}_l \rangle = (1 - \eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l \rangle + \frac{q\eta u_j}{n_\mathrm{l} + n_\mathrm{u}} \Bigg(\sum_{i=1}^{n_\mathrm{u}} c_i^{(t)} \widehat{y}_i \big([\langle \mathbf{w}_j, y_i \cdot \mathbf{v} \rangle]_+^{q-1} y_i \langle \mathbf{v}, \boldsymbol{\xi}_l \rangle + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l \rangle \big)$$
$$+ \sum_{i=1}^{n_\mathrm{l}} b_i^{(t)} y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} y_i' \langle \mathbf{v}, \boldsymbol{\xi}_l \rangle + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l \rangle \big) \Bigg)$$
$$= (1 - \eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l \rangle + \frac{q\eta u_j}{n_\mathrm{l} + n_\mathrm{u}} \Bigg(\sum_{i=1}^{n_\mathrm{u}} \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l \rangle$$
$$+ \sum_{i=1}^{n_\mathrm{l}} y_i' b_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l \rangle \Bigg),$$

and

$$\langle \mathbf{w}_j^{(t+1)}, \boldsymbol{\xi}_l' \rangle = (1 - \eta\lambda) \cdot \langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l' \rangle + \frac{q\eta u_j}{n_\mathrm{l} + n_\mathrm{u}} \Bigg(\sum_{i=1}^{n_\mathrm{u}} \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l' \rangle$$
$$+ \sum_{i=1}^{n_\mathrm{l}} y_i' b_i^{(t)} [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l' \rangle \Bigg),$$

which completes the proof. □

C.4 Estimate $\widehat{\Lambda}_r^{(0)}, \bar{\Lambda}_r^{(0)}, \Gamma_i^{(0)}, \Gamma_i'^{(0)}$

Let $\widehat{\Lambda}_1^{(t)} = \max_{1 \le j \le m} \langle \mathbf{w}_j^{(t)}, \mathbf{v} \rangle$, $\widehat{\Lambda}_{-1}^{(t)} = \max_{m+1 \le j \le 2m} -\langle \mathbf{w}_j^{(t)}, \mathbf{v} \rangle$, $\bar{\Lambda}_1^{(t)} = \max_{m+1 \le j \le 2m} \langle \mathbf{w}_j^{(t)}, \mathbf{v} \rangle$, $\bar{\Lambda}_{-1}^{(t)} = \max_{1 \le j \le m} -\langle \mathbf{w}_j^{(t)}, \mathbf{v} \rangle$, which characterize the *feature learning* aspect of training process. An easy way to distinguish between $\widehat{\Lambda}_r^{(t)}$ and $\bar{\Lambda}_r^{(t)}$ is that $\widehat{\Lambda}_r^{(t)}$ should be large while $\bar{\Lambda}_r^{(t)}$ should be small.

Let $\Gamma_i^{(t)} = \max_{1\le j\le 2m}\langle \mathbf{w}_j, \boldsymbol{\xi}_i\rangle, i\in[n_{\mathrm{u}}], \Gamma_i'^{(t)} = \max_{1\le j\le 2m}\langle \mathbf{w}_j, \boldsymbol{\xi}_i'\rangle, i\in[n_{\mathrm{l}}]$, which characterize the *noise memorization* aspect of training process with respect to a particular sample.

Let $\Gamma^{(t)} = \max\big\{\max_{i\in[n_{\mathrm{u}}]}\Gamma_i^{(t)}, \max_{i\in[n_{\mathrm{l}}]}\Gamma_i'^{(t)}\big\}$, which characterize the *noise memorization* aspect of training process regardless of the sample index.

We first provide the concentration inequality for $\widehat{\Lambda}_r^{(0)}$ and $\bar{\Lambda}_r^{(0)}$ in the following lemma.

Lemma C.9. With probability at least $1-4\delta$ with respect to the randomness of initialization of $\mathbf{w}$, we have

$$\big|\widehat{\Lambda}_r^{(0)} - \mathbb{E}[\widehat{\Lambda}_r^{(0)}]\big| < \sqrt{8\log\Big(\frac{1}{\delta}\Big)}\sigma_0\|\mathbf{v}\|_2,$$
$$\big|\bar{\Lambda}_r^{(0)} - \mathbb{E}[\bar{\Lambda}_r^{(0)}]\big| < \sqrt{8\log\Big(\frac{1}{\delta}\Big)}\sigma_0\|\mathbf{v}\|_2,$$

and

$$\mathbb{E}[\widehat{\Lambda}_r^{(0)}] \asymp \sqrt{\log(m)}\sigma_0\|\mathbf{v}\|_2, \mathbb{E}[\bar{\Lambda}_r^{(0)}] \asymp \sqrt{\log(m)}\sigma_0\|\mathbf{v}\|_2, r\in\{\pm 1\}.$$

Proof of Lemma C.9. Note that $\widehat{\Lambda}_1^{(0)} = \max_{1\le j\le m}\langle \mathbf{w}_j^{(0)}, \mathbf{v}\rangle, \widehat{\Lambda}_{-1}^{(0)} = \max_{m+1\le j\le 2m} -\langle \mathbf{w}_j^{(0)}, \mathbf{v}\rangle$, $\bar{\Lambda}_1^{(0)} = \max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(0)}, \mathbf{v}\rangle$ and $\bar{\Lambda}_{-1}^{(0)} = \max_{m+1\le j\le 2m} -\langle \mathbf{w}_j^{(0)}, \mathbf{v}\rangle$, $\mathbf{w}_j^{(0)} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I})$ and $\mathbf{v}$ is a fixed vector. Therefore, $\langle \mathbf{w}_j^{(0)}, \mathbf{v}\rangle \sim \mathcal{N}(0, \sigma_0^2\|\mathbf{v}\|_2^2)$, $-\langle \mathbf{w}_j^{(0)}, \mathbf{v}\rangle \sim \mathcal{N}(0, \sigma_0^2\|\mathbf{v}\|_2^2)$ for all $1\le j\le 2m$ and $\widehat{\Lambda}_r^{(0)}, \bar{\Lambda}_r^{(0)}, r\in\{\pm1\}$ are identically distributed. Therefore, without loss of generality, we only need to discuss the concentration of $\widehat{\Lambda}_1^{(0)}$. By applying Lemma E.1, we have

$$\mathbb{P}\Big(\big|\widehat{\Lambda}_1^{(0)} - \mathbb{E}[\widehat{\Lambda}_1^{(0)}]\big| > t\Big) \le 2e^{-\frac{t^2}{2\sigma_0^2\|\mathbf{v}\|_2^2}}.$$

By applying Lemma E.2, we have

$$\mathbb{E}[\widehat{\Lambda}_1^{(0)}] \asymp \sqrt{\log(m)}\sigma_0\|\mathbf{v}\|_2,$$

which completes the proof. $\square$

Then we provide concentration inequality for $\Gamma_i^{(0)}$ in the following lemma.

Lemma C.10. Suppose that $d \ge \Omega(\log(m(n_{\mathrm{u}}+n_{\mathrm{l}})/\delta))$, $m = \Omega(\log(1/\delta))$. Then with probability at least $1-\delta$,

$$\frac{\sigma_0\sigma_p\sqrt{d}}{4} \le \Gamma_i^{(0)} \le 2\sqrt{\log(16m(n_{\mathrm{u}}+n_{\mathrm{l}})/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}, \text{ for all } i\in[n_{\mathrm{u}}],$$
$$\frac{\sigma_0\sigma_p\sqrt{d}}{4} \le \Gamma_i'^{(0)} \le 2\sqrt{\log(16m(n_{\mathrm{u}}+n_{\mathrm{l}})/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}, \text{ for all } i\in[n_{\mathrm{l}}].$$

Proof of Lemma C.10. By Lemma E.3, with probability at least $1-\delta/4$,

$$\begin{aligned}\sigma_p\sqrt{d}/\sqrt{2} &\le \|\boldsymbol{\xi}_i\|_2 \le \sqrt{3/2}\cdot\sigma_p\sqrt{d}, \text{ for } i\in[n_{\mathrm{u}}],\\ \sigma_p\sqrt{d}/\sqrt{2} &\le \|\boldsymbol{\xi}_i'\|_2 \le \sqrt{3/2}\cdot\sigma_p\sqrt{d}, \text{ for } i\in[n_{\mathrm{l}}].\end{aligned} \tag{C.3}$$

Therefore, by Gaussian tail bound and union bound, with probability at least $1-\delta/4$,

$$\begin{aligned}\langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i\rangle &\le |\langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i\rangle| \le \sqrt{2\log(8m/\delta)}\cdot\sigma_0\|\boldsymbol{\xi}_i\|_2, \text{ for } i\in[n_{\mathrm{u}}],\\ \langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i'\rangle &\le |\langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i'\rangle| \le \sqrt{2\log(8m/\delta)}\cdot\sigma_0\|\boldsymbol{\xi}_i'\|_2, \text{ for } i\in[n_{\mathrm{l}}].\end{aligned} \tag{C.4}$$

Note that $\mathbb{P}\big(\sigma_0\sigma_p\sqrt{d}/4 > \langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i\rangle\big)$ is an absolute constant and therefore by the condition on m, we have

$$\begin{aligned}\mathbb{P}\bigg(\frac{\sigma_0\sigma_p\sqrt{d}}{4} \le \Gamma_i^{(t)}\bigg) &= \mathbb{P}\bigg(\frac{\sigma_0\sigma_p\sqrt{d}}{4} \le \max_{j\in[2m]}\langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i\rangle\bigg)\\ &= 1 - \mathbb{P}\bigg(\frac{\sigma_0\sigma_p\sqrt{d}}{4} > \max_{j\in[2m]}\langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i\rangle\bigg)\end{aligned}$$

$$= 1 - \left(\mathbb{P}\left(\frac{\sigma_0\sigma_p\sqrt{d}}{4} > \langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i\rangle\right)\right)^{2m}$$
$$\geq 1 - \frac{\delta}{4},$$

and

$$\mathbb{P}\left(\frac{\sigma_0\sigma_p\sqrt{d}}{4} \leq \Gamma_i'^{(t)}\right) \geq 1 - \frac{\delta}{4}.$$

On the other hand, according to (C.3) and (C.4), we have

$$\mathbb{P}\big(\Gamma_i^{(t)} \leq 2\sqrt{\log(16m(n_{\mathrm{u}} + n_{\mathrm{l}})/\delta)} \cdot \sigma_0\sigma_p\sqrt{d}\big)$$
$$= \mathbb{P}\Big(\max_{j\in[2m]} \langle \mathbf{w}_j^{(0)}, \boldsymbol{\xi}_i\rangle \leq 2\sqrt{\log(16m(n_{\mathrm{u}} + n_{\mathrm{l}})/\delta)} \cdot \sigma_0\sigma_p\sqrt{d}\Big)$$
$$\geq 1 - \frac{\delta}{4},$$

and

$$\mathbb{P}\big(\Gamma_i'^{(t)} \leq 2\sqrt{\log(16m(n_{\mathrm{u}} + n_{\mathrm{l}})/\delta)} \cdot \sigma_0\sigma_p\sqrt{d}\big) \geq 1 - \frac{\delta}{4},$$

which completes the proof. □

C.5 Stage I of GD: On-diagonal feature learning

In this stage, $\widehat{\Lambda}_1^{(t)}$ and $\widehat{\Lambda}_{-1}^{(t)}$ respectively increase to magnitude $\Theta(1/m)$ and $\bar{\Lambda}_1^{(t)}$, $\bar{\Lambda}_{-1}^{(t)}$ and $\Gamma_j^{(t)}$ remain small, the same magnitude as initialization. In order to characterize the behaviour of feature learning and noise memorization during Stage I, we decompose the analysis into following three parts:

1. First, in Lemma C.15, we provide a lower bound of the update rules of on-diagonal feature learning term of $\widehat{\Lambda}_1^{(t)}, \widehat{\Lambda}_{-1}^{(t)}$ to lower-bound their increasing speed, and an upper bound of off-diagonal feature learning term $\bar{\Lambda}_1^{(t)}, \bar{\Lambda}_{-1}^{(t)}$ to indicate their decrease.
2. Second, in Lemma C.17, we provide a upper bound of the update rules of noise memorization term $\Gamma^{(t)}$ to upper-bound its increasing speed.
3. Third, we provide a useful lemma, which is a derivation of Claim C.20 in Allen-Zhu and Li (2020b), which is called tensor power method. By applying tensor power method, we will prove that:
   - When $\widehat{\Lambda}_1^{(t)}$ reaches $\Theta(1/m)$ at T_1, $\bar{\Lambda}_1^{(t)}$ and $\Gamma^{(t)}$ remain a magnitude no more than initialization.
   - When $\widehat{\Lambda}_{-1}^{(t)}$ reaches $\Theta(1/m)$ at T_{-1}, $\bar{\Lambda}_{-1}$ and $\Gamma^{(t)}$ remain a magnitude no more than initialization.

C.5.1 Upper bound and lower bound for $\widehat{\Lambda}_1^{(t)}, \widehat{\Lambda}_{-1}^{(t)}$ and $\bar{\Lambda}_1^{(t)}, \bar{\Lambda}_{-1}^{(t)}$

We first consider Stage I of GD when $\max_{r\in\{\pm1\}}\big\{\widehat{\Lambda}_r^{(t)}, \bar{\Lambda}_r^{(t)}\big\} \leq \Theta(m^{-1})$.
In this stage, we first prove following lemma:

Lemma C.11. As long as $\max_{r\in\{\pm1\}}\big\{\widehat{\Lambda}_r^{(t)}, \bar{\Lambda}_r^{(t)}\big\} \leq \Theta(m^{-1})$, we have $c_i^{(t)} := -\ell'\big(\widehat{y}_i \cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\big)$ and $b_i^{(t)} := -\ell'\big(y_i' \cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}_i')\big)$ remains $1/2 \pm o(1)$.

Proof of Lemma C.11. Note that $\ell(z) = \log(1+\exp(-z))$ and $-\ell'(z) = \exp(-z)/\big(1+\exp(-z)\big)$, and without loss of generality assuming $\widehat{y}_i = y_i = 1$, we can express $c_i^{(t)}$ as follow:

$$c_i^{(t)} = -\ell'(f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)) = \frac{e^{\sum_{j=m+1}^{2m}[\sigma(\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i\rangle)]}}{e^{\sum_{j=1}^{m}[\sigma(\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i\rangle)]} + e^{\sum_{j=m+1}^{2m}[\sigma(\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i\rangle)]}},$$

Since $\sigma(\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle)$ dominates $\sigma(\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}\rangle)$ for $j \in [m]$, which will be proved later by using *tensor power method*, we have

$$c_i^{(t)} = \frac{e^{\sum_{j=m+1}^{2m}[\sigma(\langle \mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\sigma(\langle \mathbf{w}_j^{(t)},\boldsymbol{\xi}_i\rangle)]}}{e^{\sum_{j=1}^{m}\sigma(\langle \mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\{\text{lower order term}\}}+e^{\sum_{j=m+1}^{2m}[\sigma(\langle \mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\sigma(\langle \mathbf{w}_j^{(t)},\boldsymbol{\xi}_i\rangle)]}}.$$

On the one side,

$$c_i^{(t)} \geq \frac{1}{e^{\sum_{j=1}^{m}\sigma(\langle \mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\{\text{lower order term}\}}+1} \geq \frac{1}{e^{m(\widehat{\Lambda}_1^{(t)})^{q-1}}+1} \geq \frac{1}{e^{\Theta(m^{-(q-1)})}+1} = \frac{1}{2+o(1)} = \frac{1}{2} - o(1).$$

On the other side, according to Lemma C.4, we have $\bar{\Lambda}_1^{(t)} = \widetilde{O}(d^{-\frac{1}{4}})$ and $\Gamma^{(t)} = \widetilde{O}(d^{-\frac{1}{4}+\epsilon})$, it follows that

$$\begin{aligned}
c_i^{(t)} &\leq \frac{e^{m(\bar{\Lambda}_1^{(t)})^{q-1}+m(\Gamma^{(t)})^{q-1}}}{e^{\sum_{j=1}^{m}\sigma(\langle \mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\{\text{lower order term}\}}+e^{m(\bar{\Lambda}_1^{(t)})^{q-1}+m(\Gamma^{(t)})^{q-1}}} \\
&= \frac{1+o(1)}{e^{\sum_{j=1}^{m}\sigma(\langle \mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\{\text{lower order term}\}}+1+o(1)} \\
&\leq \frac{1+o(1)}{1+1+o(1)} = \frac{1}{2}+o(1).
\end{aligned}$$

Therefore, we have $c_i^{(t)} = 1/2 \pm o(1)$ if $\widehat{y}_i = y_i = 1$ and other cases ($\widehat{y}_i = y_i = 1$, $\widehat{y}_i = -y_i$, $b_i^{(t)}$) can be proved in a similar way. $\square$

By applying above lemma, we can obtain following lemma:

Lemma C.12. For any $\delta < 1/2$, with probability at least $1 - 2\delta$ over pseudo-labels generated by the pseudo-labeler, we have

$$\left|\frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\widehat{y}_i y_i c_i^{(t)} - \left(p - \frac{1}{2}\right)\right| < \sqrt{\frac{1}{8n_{\mathrm{u}}}\log\frac{1}{\delta}} + o(1),$$

where $o(1)$ is with respect to d.

If we denote $\{(\mathbf{x}_i, y_i)|y_i = 1, i \in [n_{\mathrm{u}}]\}$ as S_1, $\{(\mathbf{x}_i, y_i)|y_i = -1, i \in [n_{\mathrm{u}}]\}$ as S_{-1}, $|S_1|$ as n_1 and $|S_{-1}|$ as n_{-1}, we have with probability at least $1 - 4\delta$ that

$$\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{y}_i y_i c_i^{(t)} - \left(p - \frac{1}{2}\right)\right| < \sqrt{\frac{1}{8n_1}\log\frac{1}{\delta}} + o(1),$$

and

$$\left|\frac{1}{n_{-1}}\sum_{i=1}^{n_{-1}}\widehat{y}_i y_i c_i^{(t)} - \left(p - \frac{1}{2}\right)\right| < \sqrt{\frac{1}{8n_{-1}}\log\frac{1}{\delta}} + o(1).$$

Proof of Lemma C.12. First, according to Lemma C.11, we have

$$\frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\widehat{y}_i y_i c_i^{(t)} = \frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\widehat{y}_i y_i\left(c_i^{(t)} - \frac{1}{2}\right) + \frac{1}{2n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\widehat{y}_i y_i = \frac{1}{2n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\widehat{y}_i y_i \pm o(1) \tag{C.5}$$

Then, according to Hoeffding's inequality when $a_i = -1, b_i = 1$, we have

$$\mathbb{P}\left(\left|\frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\widehat{y}_i y_i - \mathbb{E}\Big[\frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\widehat{y}_i y_i\Big]\right| \geq t\right) \leq 2\exp\left(-\frac{2n_{\mathrm{u}}^2 t^2}{\sum_{i=1}^{n_{\mathrm{u}}}(a_i - b_i)^2}\right) = 2\exp\left(-2n_{\mathrm{u}}t^2\right).$$

Note that the pseudo-label $\widehat{y}_i$ generated by the pseudo-labeler takes y_i with probability p and $-y_i$ with probability $1-p$, we have $\mathbb{E}\big[\frac{1}{n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\widehat{y}_i y_i\big] = \frac{1}{n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\mathbb{E}\big[\widehat{y}_i y_i\big] = 2p-1$. It follows that

$$\mathbb{P}\bigg(\bigg|\frac{1}{2n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\widehat{y}_i y_i - \Big(p-\frac{1}{2}\Big)\bigg| \geq t\bigg) \leq 2\exp\big(-8n_\mathrm{u}t^2\big),$$

and therefore

$$\bigg|\frac{1}{2n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\widehat{y}_i y_i - \Big(p-\frac{1}{2}\Big)\bigg| < \sqrt{\frac{1}{8n_\mathrm{u}}\log\frac{1}{\delta}} \tag{C.6}$$

holds with probability at least $1-2\delta$. According to (C.5) and (C.6), we have

$$\bigg|\frac{1}{2n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\widehat{y}_i y_i - \Big(p-\frac{1}{2}\Big)\bigg| < \sqrt{\frac{1}{8n_\mathrm{u}}\log\frac{1}{\delta}} + o(1),$$

which verifies the first statement of the lemma. And the other part of the lemma can be proved in a similar way. $\square$

According to above lemma and note that $n_\mathrm{u}, n_1, n_{-1} = \omega(1)$, we have further that

$$\bigg|\frac{1}{n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\widehat{y}_i y_i c_i^{(t)} - \Big(p-\frac{1}{2}\Big)\bigg| = o(1),\ \bigg|\frac{1}{n_r}\sum_{i=1}^{n_r}\widehat{y}_i y_i c_i^{(t)} - \Big(p-\frac{1}{2}\Big)\bigg| = o(1), r\in\{\pm 1\}, \tag{C.7}$$

with high probability.

Besides, we also need an approximation about n_1 and n_{-1}, which is given as the following lemma:

Lemma C.13. For $r \in \{\pm 1\}$, it holds with probability at least $1-2\delta$ that

$$\Big|n_r - \frac{n_\mathrm{u}}{2}\Big| < \sqrt{\frac{n_\mathrm{u}}{2}\log\frac{1}{\delta}},$$

where $n_r := |\{(\mathbf{x}_i, y_i)|y_i = r, i\in[n_\mathrm{u}]\}|$.

Proof of Lemma C.13. Note that $n_r = \sum_{i=1}^{n_\mathrm{u}}\mathbb{1}[X_i = r], r\in\{\pm 1\}$ where X_i takes label $+1$ or -1 with equal probability $1/2$, according to Hoeffding's inequality, we have

$$\mathbb{P}\bigg(\bigg|\sum_{i=1}^{n_\mathrm{u}}\mathbb{1}[X_i = r] - \mathbb{E}\Big[\sum_{i=1}^{n_\mathrm{u}}\mathbb{1}[X_i = r]\Big]\bigg| \geq t\bigg) \leq 2\exp\Big(-\frac{2t^2}{n_\mathrm{u}}\Big), r\in\{\pm 1\},$$

and it follows that

$$\mathbb{P}\Big(\Big|n_r - \frac{n_\mathrm{u}}{2}\Big| \geq t\Big) \leq 2\exp\Big(-\frac{2t^2}{n_\mathrm{u}}\Big), r\in\{\pm 1\},$$

leading to

$$\Big|n_r - \frac{n_\mathrm{u}}{2}\Big| < \sqrt{\frac{n_\mathrm{u}}{2}\log\frac{1}{\delta}},$$

with probability at least $1-2\delta$. $\square$

For labeled dataset $S' = \{(\mathbf{x}_i', y_i')\}_{i=1}^{n_\mathrm{l}}$, we also have

Lemma C.14. For $r \in \{\pm 1\}$, it holds with probability at least $1-2\delta$ that

$$\Big|n_r' - \frac{n_\mathrm{l}}{2}\Big| < \sqrt{\frac{n_\mathrm{l}}{2}\log\frac{1}{\delta}},$$

where $n_r' := |\{(\mathbf{x}_i', y_i')|y_i' = r, i\in[n_\mathrm{l}]\}|$.

Then we are prepared to estimate a lower bound of increasing speed of $\widehat{\Lambda}^{(t)}$ and an upper bound of decreasing speed of $\overline{\Lambda}^{(t)}$ in the following lemma.

Lemma C.15. For $\widehat{\Lambda}_1^{(t)} := \max_{1\le j\le m}\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle$ and $\widehat{\Lambda}_{-1}^{(t)} := \max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t)}, -\mathbf{v}\rangle$, we have with high probability that

$$\widehat{\Lambda}_r^{(t+1)} \ge (1-\eta\lambda)\cdot\widehat{\Lambda}_r^{(t)} + \eta\cdot\Big(p-\frac{1}{2}\Big)\cdot\Theta(d)\cdot(\widehat{\Lambda}_r^{(t)})^{q-1}, r\in\{\pm1\}.$$

For $\bar{\Lambda}_1^{(t)} := \max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle$ and $\bar{\Lambda}_1^{(t)} := \max_{1\le j\le m}\langle \mathbf{w}_j^{(t)}, -\mathbf{v}\rangle$, we have with high probability that

$$\bar{\Lambda}_r^{(t+1)} \le (1-\eta\lambda)\cdot\bar{\Lambda}_r^{(t)}, r\in\{\pm1\}.$$

Proof of Lemma C.15. We first prove the former inequality. Let $j^* = \arg\max_{1\le j\le m}\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle$ and note that $u_{j^*} = \mathbb{1}_{[1\le j\le m]} - \mathbb{1}_{[m+1\le j\le 2m]} = 1$, then we have

$$\begin{aligned}
&\widehat{\Lambda}_1^{(t+1)} \ge \langle \mathbf{w}_{j^*}^{(t+1)}, \mathbf{v}\rangle\\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle + \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} + \underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar}\Bigg)
\end{aligned}$$

Then we respectively estimate terms $\clubsuit$ and $\bigstar$.

For $\clubsuit$, note the definition of j^* that $\widehat{\Lambda}_1^{(t)} = \langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle$ and note the increasing property of $\widehat{\Lambda}_1^{(t)}$ and $\widehat{\Lambda}_1^{(0)} > 0$ with high probability, we have $\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle > 0$. It follows that

$$\begin{aligned}
\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} &= \sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \sum_{i\in S_{-1}} y_i\widehat{y}_i c_i^{(t)}[-\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\\
&= \sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\\
&= \Bigg(\sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}\Bigg)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\\
&= n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}, \qquad\text{(C.8)}
\end{aligned}$$

where $S_1 := \{(\mathbf{x}_i, y_i)|y_i = 1, i\in[n_{\mathrm{u}}]\}$, $S_{-1} := \{(\mathbf{x}_i, y_i)|y_i = -1, i\in[n_{\mathrm{u}}]\}$, $n_1 = |S_1|$ and the last equality is due to (C.7).

For $\bigstar$, similarly we have

$$\begin{aligned}
\underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar} &= \sum_{i\in S_1'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \sum_{i\in S_{-1}'} b_i^{(t)}[-\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\\
&= \sum_{i\in S_1'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\\
&= \Bigg(\sum_{i\in S_1'} b_i^{(t)}\Bigg)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\\
&= n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}, \qquad\text{(C.9)}
\end{aligned}$$

where $S_1' = \{(\mathbf{x}_i', y_i')|y_i' = 1, i\in[n_{\mathrm{l}}]\}$, $S_{-1}' = \{(\mathbf{x}_i', y_i')|y_i' = -1, i\in[n_{\mathrm{l}}]\}$, $n_1' = |S_1'|$ and the last equality is due to Lemma C.11.

According to (C.8) and (C.9), we have

$$\widehat{\Lambda}_1^{(t+1)}$$

$$
\begin{aligned}
&\geq (1-\eta\lambda)\cdot \widehat{\Lambda}_1^{(t)} + \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Big(n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1} + n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\Big)\\
&= (1-\eta\lambda)\cdot \widehat{\Lambda}_1^{(t)} + \frac{q\eta n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1} + \frac{q\eta n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\\
&= (1-\eta\lambda)\cdot \widehat{\Lambda}_1^{(t)} + q\eta\cdot\Big(\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big) + \frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(\frac{1}{2}\pm o(1)\Big)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\\
&= (1-\eta\lambda)\cdot \widehat{\Lambda}_1^{(t)} + q\eta\cdot\Big(\underbrace{\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big) + \frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}}_{\spadesuit}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}.
\end{aligned}
\tag{C.10}
$$

According to Lemma C.13 and Lemma C.14, and note that $n_{\mathrm{l}} = \widetilde{\Theta}(1)$, $n_{\mathrm{u}} = \omega(d^{4\epsilon})$, we have for $\spadesuit$ that with probability at least $1-4\delta$

$$
\begin{aligned}
&\Big|\underbrace{\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big) + \frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}}_{\spadesuit} - \frac{n_{\mathrm{u}}}{2(n_{\mathrm{l}}+n_{\mathrm{u}})}\cdot\Big(p-\frac{1}{2}\Big) - \frac{n_{\mathrm{l}}}{2(n_{\mathrm{l}}+n_{\mathrm{u}})}\cdot\frac{1}{2}\Big|\\
&\leq \frac{|n_1-\frac{n_{\mathrm{u}}}{2}|}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big) + \frac{|n_1'-\frac{n_{\mathrm{l}}}{2}|}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}\\
&\leq \frac{\sqrt{\frac{n_{\mathrm{u}}}{2}\log\frac{1}{\delta}}}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big) + \frac{\sqrt{\frac{n_{\mathrm{l}}}{2}\log\frac{1}{\delta}}}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}\\
&= \Theta\Big(\frac{1}{\sqrt{n_{\mathrm{u}}}}\Big)\\
&= o(1)
\end{aligned}
$$

Therefore, note that $n_{\mathrm{u}} = \omega(n_{\mathrm{l}})$ and $n_{\mathrm{u}} = \omega(1)$, we have

$$
\begin{aligned}
\underbrace{\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big) + \frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}}_{\spadesuit} &= \frac{n_{\mathrm{u}}}{2(n_{\mathrm{l}}+n_{\mathrm{u}})}\cdot\Big(p-\frac{1}{2}\Big) + \frac{n_{\mathrm{l}}}{2(n_{\mathrm{l}}+n_{\mathrm{u}})}\cdot\frac{1}{2}\pm o(1)\\
&= \frac{1}{2}\cdot\Big(p-\frac{1}{2}\Big)\pm o(1)
\end{aligned}
\tag{C.11}
$$

Plugging (C.11) into (C.10), we have

$$
\begin{aligned}
\widehat{\Lambda}_1^{(t+1)} &\geq (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + q\eta\cdot\Big(\frac{1}{2}\cdot\Big(p-\frac{1}{2}\Big)\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\\
&= (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + \eta\cdot\Big(p-\frac{1}{2}\Big)\cdot\Theta(d)\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1},
\end{aligned}
\tag{C.12}
$$

which verifies the first inequality of case $r=1$ in the lemma.

Let $j^{**} = \mathrm{argmax}_{m+1\leq j\leq 2m}\langle\mathbf{w}_j^{(t)}, -\mathbf{v}\rangle$ and note that $u_{j^{**}} = \mathbb{1}_{[1\leq j\leq m]} - \mathbb{1}_{[m+1\leq j\leq 2m]} = -1$, we have

$$
\begin{aligned}
\widehat{\Lambda}_{-1}^{(t+1)} &\geq \langle\mathbf{w}_{j^*}^{(t+1)}, -\mathbf{v}\rangle\\
&= (1-\eta\lambda)\cdot\langle\mathbf{w}_{j^{**}}^{(t)}, -\mathbf{v}\rangle + \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Big(\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle\mathbf{w}_{j^{**}}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit}\\
&\quad + \underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle\mathbf{w}_{j^{**}}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\star}\Big)
\end{aligned}
$$

For $\clubsuit$, note the definition of j^{**} that $\widehat{\Lambda}_{-1}^{(t)} = \langle \mathbf{w}_{j^{**}}^{(t)}, -\mathbf{v} \rangle$ and note the increasing property of $\widehat{\Lambda}_{-1}^{(t)}$ and $\widehat{\Lambda}_{-1}^{(0)} > 0$ with high probability, we have $\langle \mathbf{w}_{j^{**}}^{(t)}, -\mathbf{v} \rangle > 0$. According to (C.7), it follows that

$$\begin{aligned}\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_{j^{**}}^{(t)}, y_i \cdot \mathbf{v} \rangle]_+^{q-1} \|\mathbf{v}\|_2^2}_{\clubsuit} &= \sum_{i \in S_{-1}} y_i \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_{j^{**}}^{(t)}, -\mathbf{v} \rangle]_+^{q-1} \|\mathbf{v}\|_2^2 \\ &= n_{-1} \cdot \Big(p - \frac{1}{2} \pm o(1)\Big) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\widehat{\Lambda}_{-1}^{(t)}\big)^{q-1}, \end{aligned} \tag{C.13}$$

where $S_{-1} := \{(\mathbf{x}_i, y_i) | y_i = -1, i \in [n_{\mathrm{u}}]\}$, $n_{-1} = |S_{-1}|$.

For $\bigstar$, according to Lemma C.11, similarly we have

$$\underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)} [\langle \mathbf{w}_{j^{**}}^{(t)}, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \|\mathbf{v}\|_2^2}_{\bigstar} = \sum_{i \in S'_{-1}} b_i^{(t)} [\langle \mathbf{w}_{j^{**}}^{(t)}, -\mathbf{v} \rangle]_+^{q-1} \|\mathbf{v}\|_2^2 = n'_{-1} \cdot \Big(\frac{1}{2} \pm o(1)\Big) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\widehat{\Lambda}_{-1}^{(t)}\big)^{q-1}, \tag{C.14}$$

where $S'_{-1} = \{(\mathbf{x}'_i, y'_i) | y'_i = -1, i \in [n_{\mathrm{l}}]\}$ and $n'_{-1} = |S'_{-1}|$.

According to (C.13) and (C.14), we have

$$\widehat{\Lambda}_{-1}^{(t+1)} \geq (1 - \eta\lambda) \cdot \widehat{\Lambda}_{-1}^{(t)} + q\eta \cdot \Bigg(\underbrace{\frac{n_{-1}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \Big(p - \frac{1}{2}\Big) + \frac{n'_{-1}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \frac{1}{2}}_{\spadesuit} \pm o(1) \Bigg) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\widehat{\Lambda}_{-1}^{(t)}\big)^{q-1}. \tag{C.15}$$

According to Lemma C.13 and Lemma C.14, and note that $n_{\mathrm{l}} = \widetilde{\Theta}(1)$, $n_{\mathrm{u}} = \omega(d^{4\epsilon})$, we have for $\spadesuit$ that with probability at least $1 - 4\delta$

$$\begin{aligned} &\Bigg| \underbrace{\frac{n_{-1}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \Big(p - \frac{1}{2}\Big) + \frac{n'_{-1}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \frac{1}{2}}_{\spadesuit} - \frac{n_{\mathrm{u}}}{2(n_{\mathrm{l}} + n_{\mathrm{u}})} \cdot \Big(p - \frac{1}{2}\Big) - \frac{n_{\mathrm{l}}}{2(n_{\mathrm{l}} + n_{\mathrm{u}})} \cdot \frac{1}{2} \Bigg| \\ &\leq \frac{|n_{-1} - \frac{n_{\mathrm{u}}}{2}|}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \Big(p - \frac{1}{2}\Big) + \frac{|n'_{-1} - \frac{n_{\mathrm{l}}}{2}|}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \frac{1}{2} \\ &\leq \frac{\sqrt{\frac{n_{\mathrm{u}}}{2} \log \frac{1}{\delta}}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \Big(p - \frac{1}{2}\Big) + \frac{\sqrt{\frac{n_{\mathrm{l}}}{2} \log \frac{1}{\delta}}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \frac{1}{2} \\ &= \Theta\Big(\frac{1}{\sqrt{n_{\mathrm{u}}}}\Big) \\ &= o(1). \end{aligned}$$

Therefore, note that $n_{\mathrm{u}} = \omega(n_{\mathrm{l}})$ and $n_{\mathrm{u}} = \omega(1)$, we have

$$\begin{aligned} \underbrace{\frac{n_{-1}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \Big(p - \frac{1}{2}\Big) + \frac{n'_{-1}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \frac{1}{2}}_{\spadesuit} &= \frac{n_{\mathrm{u}}}{2(n_{\mathrm{l}} + n_{\mathrm{u}})} \cdot \Big(p - \frac{1}{2}\Big) + \frac{n_{\mathrm{l}}}{2(n_{\mathrm{l}} + n_{\mathrm{u}})} \cdot \frac{1}{2} \pm o(1) \\ &= \frac{1}{2} \cdot \Big(p - \frac{1}{2}\Big) \pm o(1) \end{aligned} \tag{C.16}$$

Plugging (C.16) into (C.15), we have

$$\begin{aligned} \widehat{\Lambda}_{-1}^{(t+1)} &\geq (1 - \eta\lambda) \cdot \widehat{\Lambda}_{-1}^{(t)} + q\eta \cdot \Bigg(\frac{1}{2} \cdot \Big(p - \frac{1}{2}\Big) \pm o(1) \Bigg) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\widehat{\Lambda}_{-1}^{(t)}\big)^{q-1} \\ &= (1 - \eta\lambda) \cdot \widehat{\Lambda}_{-1}^{(t)} + \eta \cdot \Big(p - \frac{1}{2}\Big) \cdot \Theta(d) \cdot \big(\widehat{\Lambda}_{-1}^{(t)}\big)^{q-1}, \end{aligned} \tag{C.17}$$

which verifies the first inequality of case $r = -1$ in the lemma.

Next, we prove the latter part of the lemma. Let $j^{\natural} = \arg\max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t+1)}, \mathbf{v}\rangle$, then we have:

$$\begin{aligned}
\bar{\Lambda}_1^{(t+1)} &= \langle \mathbf{w}_{j^{\natural}}^{(t+1)}, \mathbf{v}\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^{\natural}}^{(t)}, \mathbf{v}\rangle - \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^{\natural}}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} \\
&\qquad + \underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^{\natural}}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar}\Bigg).
\end{aligned}$$

For $\clubsuit$, according to (C.7), we have

$$\begin{aligned}
&\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^{\natural}}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} \\
&= \sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^{\natural}}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \sum_{i\in S_{-1}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^{\natural}}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&= \Bigg(\sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}\Bigg)\cdot[\langle \mathbf{w}_{j^{\natural}}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \Bigg(\sum_{i\in S_{-1}} y_i\widehat{y}_i c_i^{(t)}\Bigg)\cdot[\langle \mathbf{w}_{j^{\natural}}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&= n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot[\langle \mathbf{w}_{j^{\natural}}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + n_{-1}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot[\langle \mathbf{w}_{j^{\natural}}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \ge 0,
\end{aligned}$$

and for $\bigstar$ it's obvious that

$$\underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^{\natural}}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar} \ge 0.$$

Therefore, it follows that

$$\bar{\Lambda}_1^{(t+1)} \le (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^{\natural}}^{(t)}, \mathbf{v}\rangle \le (1-\eta\lambda)\bar{\Lambda}_1^{(t)}.$$

Let $j^{\natural\natural} = \arg\max_{1\le j\le m}\langle \mathbf{w}_j^{(t+1)}, -\mathbf{v}\rangle$, then we have:

$$\begin{aligned}
\bar{\Lambda}_{-1}^{(t+1)} &= \langle \mathbf{w}_{j^{\natural\natural}}^{(t+1)}, -\mathbf{v}\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^{\natural\natural}}^{(t)}, -\mathbf{v}\rangle - \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^{\natural\natural}}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&\qquad + \sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^{\natural\natural}}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\Bigg) \\
&\le (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^{\natural\natural}}^{(t)}, -\mathbf{v}\rangle \\
&\le (1-\eta\lambda)\cdot\bar{\Lambda}_{-1}^{(t)},
\end{aligned}$$

which verifies the second part of the lemma. $\square$

Although the accuracy of pseudo-labeler is larger than $1/2$, which is used as an assumption in the previous proof, we can also analyse the model with high label flipping probability and the accuracy of pseudo-labeler p is smaller than $1/2$. In this case, the neural network for pre-training will turn to fit the opposite direction of feature vector, $\bar{\Lambda}_r^{(t)}$ will increase and $\widehat{\Lambda}_r^{(t)}$ will decrease, which is formulated as the following lemma.

Lemma C.16. For $\widehat{\Lambda}_1^{(t)} := \max_{1\le j\le m}\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle$ and $\widehat{\Lambda}_{-1}^{(t)} := \max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t)}, -\mathbf{v}\rangle$, we have with high probability that

$$\widehat{\Lambda}_r^{(t+1)} \le (1-\eta\lambda)\cdot \widehat{\Lambda}_r^{(t)}, r \in \{\pm 1\}.$$

For $\bar{\Lambda}_1^{(t)} := \max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t)}, \mathbf{v}\rangle$ and $\bar{\Lambda}_1^{(t)} := \max_{1\le j\le m}\langle \mathbf{w}_j^{(t)}, -\mathbf{v}\rangle$, we have with high probability that

$$\bar{\Lambda}_r^{(t+1)} \ge (1-\eta\lambda)\cdot \bar{\Lambda}_r^{(t)} + \eta\cdot\left(\frac{1}{2}-p\right)\cdot \Theta(d)\cdot (\bar{\Lambda}_r^{(t)})^{q-1}, r\in\{\pm 1\}.$$

Proof of Lemma C.16. First, we prove the former part of this lemma. Let $j^* = \arg\max_{1\le j\le m}\langle \mathbf{w}_j^{(t+1)}, \mathbf{v}\rangle$ and note that $u_{j^*} = \mathbb{1}_{[1\le j\le m]} - \mathbb{1}_{[m+1\le j\le 2m]} = 1$, then we have

$$\begin{aligned}
\widehat{\Lambda}_1^{(t+1)} &= \langle \mathbf{w}_{j^*}^{(t+1)}, \mathbf{v}\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle + \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} \\
&\qquad + \underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar}\Bigg).
\end{aligned}$$

For $\clubsuit$, according to (C.7), we have

$$\begin{aligned}
&\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} \\
&= \sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \sum_{i\in S_{-1}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&= \left(\sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \left(\sum_{i\in S_{-1}} y_i\widehat{y}_i c_i^{(t)}\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&= n_1\cdot\left(p-\frac{1}{2}\pm o(1)\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + n_{-1}\cdot\left(p-\frac{1}{2}\pm o(1)\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2,
\end{aligned}$$

For $\bigstar$, according to (C.7), we have

$$\begin{aligned}
&\underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar} \\
&= \sum_{i\in S_1'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \sum_{i\in S_{-1}'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&= n_1'\cdot\left(\frac{1}{2}\pm o(1)\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + n_{-1}'\cdot\left(\frac{1}{2}\pm o(1)\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2,
\end{aligned}$$

It follows that

$$\begin{aligned}
&\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} + \underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar} \\
&= \left(n_1\cdot\left(p-\frac{1}{2}\pm o(1)\right) + n_1'\cdot\left(\frac{1}{2}\pm o(1)\right)\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2
\end{aligned}$$

$$+\left(n_{-1}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)+n'_{-1}\cdot\Big(\frac{1}{2}\pm o(1)\Big)\right)\cdot[\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2.$$

676 According to Lemma C.13 and note that $n_{\mathrm{u}}=\omega(n_{\mathrm{l}})$, it holds with probability at least $1-8\delta$ that

$$\begin{aligned}n'_1\cdot\Big(\frac{1}{2}\pm o(1)\Big)&\le\bigg(\frac{n_{\mathrm{l}}}{2}+\sqrt{\frac{n_{\mathrm{l}}}{2}\log\frac{1}{\delta}}\bigg)\cdot\Big(\frac{1}{2}\pm o(1)\Big)=\Theta(n_{\mathrm{l}})=o(n_{\mathrm{u}})\\&\le\bigg(\frac{n_{\mathrm{u}}}{2}+\sqrt{\frac{n_{\mathrm{u}}}{2}\log\frac{1}{\delta}}\bigg)\cdot\Big(\frac{1}{2}-p\pm o(1)\Big)\le n_1\cdot\Big(\frac{1}{2}-p\pm o(1)\Big),\end{aligned}$$

677

$$\begin{aligned}n'_{-1}\cdot\Big(\frac{1}{2}\pm o(1)\Big)&\le\bigg(\frac{n_{\mathrm{l}}}{2}+\sqrt{\frac{n_{\mathrm{l}}}{2}\log\frac{1}{\delta}}\bigg)\cdot\Big(\frac{1}{2}\pm o(1)\Big)=\Theta(n_{\mathrm{l}})=o(n_{\mathrm{u}})\\&\le\bigg(\frac{n_{\mathrm{u}}}{2}+\sqrt{\frac{n_{\mathrm{u}}}{2}\log\frac{1}{\delta}}\bigg)\cdot\Big(\frac{1}{2}-p\pm o(1)\Big)\le n_{-1}\cdot\Big(\frac{1}{2}-p\pm o(1)\Big),\end{aligned}$$

678 leading to $\clubsuit+\bigstar\le 0$. Therefore,

$$\widehat{\Lambda}_1^{(t+1)}\le(1-\eta\lambda)\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle\le(1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)}.$$

679 And we can prove in a similar way that $\widehat{\Lambda}_{-1}^{(t+1)}\le(1-\eta\lambda)\cdot\widehat{\Lambda}_{-1}^{(t)}$.

680 Next, we prove the second part of the lemma. Let $j^\natural=\arg\max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t)},\mathbf{v}\rangle$ and note that
681 $u_{j^\natural}=\mathbb{1}_{[1\le j\le m]}-\mathbb{1}_{[m+1\le j\le 2m]}=-1$, then we have

$$\begin{aligned}\bar{\Lambda}_1^{(t+1)}&\ge\langle \mathbf{w}_{j^\natural}^{(t+1)},\mathbf{v}\rangle\\&=(1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle-\frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Bigg(\underbrace{\sum_{i=1}^{n_{\mathrm{u}}}y_i\widehat{y}_ic_i^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)},y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit}\\&\qquad+\underbrace{\sum_{i=1}^{n_{\mathrm{l}}}b_i^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)},y'_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar}\Bigg).\end{aligned}$$

682 For $\clubsuit$, note the definition of $j^\natural$ that $\bar{\Lambda}_1^{(t)}=\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle$ and note the increasing property of $\bar{\Lambda}_1^{(t)}$ in this
683 case and $\bar{\Lambda}_1^{(0)}>0$ with high probability, we have $\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle>0$. It follows that

$$\begin{aligned}\underbrace{\sum_{i=1}^{n_{\mathrm{u}}}y_i\widehat{y}_ic_i^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)},y_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit}&=\sum_{i\in S_1}y_i\widehat{y}_ic_i^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2+\sum_{i\in S_{-1}}y_i\widehat{y}_ic_i^{(t)}[-\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\\&=\sum_{i\in S_1}y_i\widehat{y}_ic_i^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\\&=\bigg(\sum_{i\in S_1}y_i\widehat{y}_ic_i^{(t)}\bigg)\cdot\|\mathbf{v}\|_2^2\cdot\big(\bar{\Lambda}_1^{(t)}\big)^{q-1}\\&=n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\bar{\Lambda}_1^{(t)}\big)^{q-1},\end{aligned}\tag{C.18}$$

684 For $\bigstar$, similarly we have

$$\underbrace{\sum_{i=1}^{n_{\mathrm{l}}}b_i^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)},y'_i\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar}=\sum_{i\in S'_1}b_i^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2+\sum_{i\in S'_{-1}}b_i^{(t)}[-\langle \mathbf{w}_{j^\natural}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2$$

$$
\begin{aligned}
&= \sum_{i\in S_1'} b_i^{(t)} [\langle \mathbf{w}_{j^\natural}^{(t)}, \mathbf{v}\rangle]_+^{q-1} \|\mathbf{v}\|_2^2 \\
&= \Big(\sum_{i\in S_1'} b_i^{(t)}\Big) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\bar{\Lambda}_1^{(t)}\big)^{q-1} \\
&= n_1' \cdot \Big(\frac{1}{2} \pm o(1)\Big) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\bar{\Lambda}_1^{(t)}\big)^{q-1}. \qquad \text{(C.19)}
\end{aligned}
$$

According to Lemma C.13, (C.18) and (C.19), we have $n_1' = o(n_1)$ with high probability, therefore

$$
\clubsuit + \bigstar = n_1 \cdot \Big(p - \frac{1}{2} \pm o(1)\Big) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\bar{\Lambda}_1^{(t)}\big)^{q-1},
$$

leading to

$$
\begin{aligned}
\bar{\Lambda}_1^{(t+1)} &\geq (1-\eta\lambda)\cdot \langle \mathbf{w}_{j^\natural}^{(t)}, \mathbf{v}\rangle - \frac{q\eta n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}} \cdot \Big(p - \frac{1}{2} \pm o(1)\Big) \cdot \|\mathbf{v}\|_2^2 \cdot \big(\bar{\Lambda}_1^{(t)}\big)^{q-1} \\
&= (1-\eta\lambda)\cdot \bar{\Lambda}_1^{(t)} + \eta \cdot \Big(\frac{1}{2} - p\Big) \cdot \Theta(d) \cdot \big(\bar{\Lambda}_1^{(t)}\big)^{q-1}.
\end{aligned}
$$

And we can prove in a similar way that

$$
\bar{\Lambda}_1^{(t+1)} \geq (1-\eta\lambda)\cdot \bar{\Lambda}_1^{(t)} + \eta \cdot \Big(\frac{1}{2} - p\Big) \cdot \Theta(d) \cdot \big(\bar{\Lambda}_1^{(t)}\big)^{q-1}.
$$

□

In this case ($p < 1/2$), given a small amount of labeled data, downstream task parameter $\mathbf{a}$ will learn the negative direction and the main theorems still hold.

C.5.2 Uniform upper bound for $\Gamma^{(t)}$

The following lemma provides an upper bound for the increasing rate of $\Gamma^{(t)}$.

Lemma C.17. For $\Gamma_i^{(t)} := \max_{j\in[2m]}\langle \mathbf{w}_j, \boldsymbol{\xi}_i\rangle, i \in [n_{\mathrm{u}}]$, $\Gamma_i'^{(t)} := \max_{j\in[2m]}\langle \mathbf{w}_j, \boldsymbol{\xi}_i'\rangle, i \in [n_{\mathrm{l}}]$, $\Gamma^{(t)} := \max\{\max_{i\in[n_{\mathrm{u}}]}\Gamma_i^{(t)}, \max_{i\in[n_{\mathrm{l}}]}\Gamma_i'^{(t)}\}$, we have with high probability that

$$
\begin{aligned}
\Gamma_i^{(t+1)} &\leq (1-\eta\lambda)\cdot \Gamma_i^{(t)} + \eta \cdot \max\Big\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\Big\} \cdot \big(\Gamma^{(t)}\big)^{q-1}, i \in [n_{\mathrm{l}}], \\
\Gamma_i'^{(t+1)} &\leq (1-\eta\lambda)\cdot \Gamma_i'^{(t)} + \eta \cdot \max\Big\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\Big\} \cdot \big(\Gamma^{(t)}\big)^{q-1}, i \in [n_{\mathrm{l}}],
\end{aligned}
$$

and

$$
\Gamma^{(t+1)} \leq (1-\eta\lambda)\cdot \Gamma^{(t)} + \eta \cdot \max\Big\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\Big\} \cdot \big(\Gamma^{(t)}\big)^{q-1},
$$

where $\epsilon < 1/8$.

Proof of Lemma C.17. We first prove the former inequality. Let $j^\star = \arg\max_{1\leq j\leq 2m}\langle \mathbf{w}_j^{(t+1)}, \boldsymbol{\xi}_l\rangle$, where $l \in [n_{\mathrm{u}}]$ is fixed. According to Lemma C.8, we have

$$
\begin{aligned}
&\Gamma_l^{(t+1)} = \langle \mathbf{w}_{j^\star}^{(t+1)}, \boldsymbol{\xi}_l\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_l\rangle + \frac{q\eta u_{j^\star}}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Big(\sum_{i=1}^{n_{\mathrm{u}}} \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i\rangle]_+^{q-1}\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l\rangle + \sum_{i=1}^{n_{\mathrm{l}}} y_i' b_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1}\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l\rangle\Big) \\
&\leq (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_l\rangle + \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\Big(\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} c_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i\rangle]_+^{q-1}|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l\rangle|}_{\clubsuit} + \underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1}|\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l\rangle|}_{\bigstar}\Big),
\end{aligned}
\qquad \text{(C.20)}
$$

where the last inequality is due to triangle inequality.

For $\clubsuit$, note that $l \in [n_{\mathrm{u}}]$ and there exists an $i \in [n_{\mathrm{u}}]$ equivalent to l, it follows that

$$\begin{aligned}
&\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} c_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l \rangle|}_{\clubsuit} \\
&= \sum_{i\in[n_{\mathrm{u}}], i\neq l} c_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l \rangle| + c_l^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_l \rangle]_+^{q-1} \|\boldsymbol{\xi}_l\|_2^2 \qquad \text{(C.21)} \\
&\leq (n_{\mathrm{u}} - 1) \cdot \Big(\frac{1}{2} + o(1)\Big) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1} + \Big(\frac{1}{2} + o(1)\Big) \cdot \widetilde{\Theta}(d^{1+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1} \\
&= (n_{\mathrm{u}} - 1) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1} + \widetilde{\Theta}(d^{1+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1},
\end{aligned}$$

where the inequality is due to Lemma C.11, $\|\boldsymbol{\xi}_l\|_2^2 = \widetilde{\Theta}(d\sigma_p^2) = \widetilde{\Theta}(d^{1+2\epsilon})$, $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l \rangle| = \widetilde{\Theta}(d^{\frac{1}{2}}\sigma_p^2) = \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon})$ according to Lemma E.3 and the definition of $\Gamma^{(t)}$.

For $\bigstar$, we have

$$\underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l \rangle|}_{\bigstar} \leq n_{\mathrm{l}} \cdot \Big(\frac{1}{2} + o(1)\Big) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1} = n_{\mathrm{l}} \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1}, \qquad \text{(C.22)}$$

Plugging (C.21) and (C.22) into (C.20), we have

$$\begin{aligned}
\Gamma_l^{(t+1)} &\leq (1 - \eta\lambda) \cdot \Gamma_l^{(t)} + \eta \cdot \bigg(\frac{q}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \Big((n_{\mathrm{u}} + n_{\mathrm{l}} - 1) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) + \widetilde{\Theta}(d^{1+2\epsilon})\Big)\bigg) \cdot \big(\Gamma^{(t)}\big)^{q-1} \\
&\leq (1 - \eta\lambda) \cdot \Gamma_l^{(t)} + \eta \cdot \max\bigg\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\bigg\} \cdot \big(\Gamma^{(t)}\big)^{q-1},
\end{aligned}$$

which is the first part of this lemma.

Let $j^\star = \mathrm{argmax}_{1\leq j\leq 2m} \langle \mathbf{w}_j^{(t+1)}, \boldsymbol{\xi}_l' \rangle$, where $l \in [n_{\mathrm{l}}]$ is fixed. According to Lemma C.8, we have

$$\begin{aligned}
&\Gamma_l'^{(t+1)} = \langle \mathbf{w}_{j^\star}^{(t+1)}, \boldsymbol{\xi}_l' \rangle \\
&= (1 - \eta\lambda) \cdot \langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_l' \rangle + \frac{q\eta u_{j^\star}}{n_{\mathrm{l}} + n_{\mathrm{u}}} \bigg(\sum_{i=1}^{n_{\mathrm{u}}} \widehat{y}_i c_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l' \rangle + \sum_{i=1}^{n_{\mathrm{l}}} y_i' b_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l' \rangle\bigg) \\
&\leq (1 - \eta\lambda) \cdot \langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_l' \rangle + \frac{q\eta}{n_{\mathrm{l}} + n_{\mathrm{u}}} \bigg(\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} c_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l' \rangle|}_{\clubsuit} + \underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l' \rangle|}_{\bigstar}\bigg), \qquad \text{(C.23)}
\end{aligned}$$

For $\clubsuit$, we have

$$\underbrace{\sum_{i=1}^{n_{\mathrm{u}}} c_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l \rangle|}_{\clubsuit} \leq \sum_{i=1}^{n_{\mathrm{u}}} \Big(\frac{1}{2} \pm o(1)\Big) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1} = n_{\mathrm{u}} \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1}, \qquad \text{(C.24)}$$

where the inequality is due to Lemma C.11, $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l \rangle| = \widetilde{\Theta}(d^{\frac{1}{2}}\sigma_p^2) = \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon})$ and the definition of $\Gamma^{(t)}$.

For $\bigstar$, note that $l \in [n_{\mathrm{l}}]$ and there exists an $i \in [n_{\mathrm{l}}]$ equivalent to l, it follows that

$$
\begin{aligned}
&\underbrace{\sum_{i=1}^{n_{\mathrm{l}}} b_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l' \rangle|}_{\bigstar} \\
&= \sum_{i\in[n_{\mathrm{l}}], i\neq l} b_i^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_i' \rangle]_+^{q-1} |\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l' \rangle| + b_l^{(t)} [\langle \mathbf{w}_{j^\star}^{(t)}, \boldsymbol{\xi}_l' \rangle]_+^{q-1} \|\boldsymbol{\xi}_l'\|_2^2 \qquad \text{(C.25)} \\
&\le (n_{\mathrm{l}} - 1) \cdot \Big(\frac{1}{2} + o(1)\Big) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1} + \Big(\frac{1}{2} + o(1)\Big) \cdot \widetilde{\Theta}(d^{1+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1} \\
&= (n_{\mathrm{l}} - 1) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) + \widetilde{\Theta}(d^{1+2\epsilon}) \cdot \big(\Gamma^{(t)}\big)^{q-1}
\end{aligned}
$$

Plugging (C.24) and (C.25) into (C.23), we have

$$
\begin{aligned}
\Gamma_l'^{(t+1)} &\le (1 - \eta\lambda) \cdot \Gamma_l'^{(t+1)} + \eta \cdot \bigg(\frac{q}{n_{\mathrm{l}} + n_{\mathrm{u}}} \cdot \Big((n_{\mathrm{u}} + n_{\mathrm{l}} - 1) \cdot \widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}) + \widetilde{\Theta}(d^{1+2\epsilon})\Big)\bigg) \cdot \big(\Gamma^{(t)}\big)^{q-1} \\
&\le (1 - \eta\lambda) \cdot \Gamma_l'^{(t+1)} + \eta \cdot \max\bigg\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\bigg\} \cdot \big(\Gamma^{(t)}\big)^{q-1},
\end{aligned}
$$

which verifies the second inequality in this lemma.

Note that $\Gamma^{(t)} = \max\{\max_{l\in[n_{\mathrm{u}}]} \Gamma_l^{(t)}, \max_{l\in[n_{\mathrm{l}}]} \Gamma_l'^{(t)}\}$, without loss of generality, we assume $\Gamma^{(t)} = \max_{l\in[n_{\mathrm{u}}]} \Gamma_l^{(t)}$ and assume $l^* = \operatorname{argmax}_{l\in[n_{\mathrm{u}}]} \Gamma_l^{(t+1)}$, we have

$$
\begin{aligned}
\Gamma^{(t+1)} = \Gamma_{l^*}^{(t+1)} &\le (1 - \eta\lambda) \cdot \Gamma_{l^*}^{(t)} + \eta \cdot \max\bigg\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\bigg\} \cdot \big(\Gamma^{(t)}\big)^{q-1} \\
&\le (1 - \eta\lambda) \cdot \Gamma^{(t)} + \eta \cdot \max\bigg\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\bigg\} \cdot \big(\Gamma^{(t)}\big)^{q-1},
\end{aligned}
$$

which verifies the third inequality in this lemma.

□

C.5.3 Tensor Power Method: Proving $\Gamma^{(t)} = O(\Gamma^{(0)})$ during $[0, T_r]$ and computing the magnitude of T_r

In this section, we first show that off-diagonal correlation ($\bar{\Lambda}_r^{(t)}$ for $p > 1/2$ and $\widehat{\Lambda}_r^{(t)}$ for $p < 1/2$) remains initialization magnitude during $[0, T_r]$. If the accuracy of pseudo-labeler $p > 1/2$, we have off-diagonal correlation $\bar{\Lambda}_r^{(t+1)} \le (1-\eta\lambda)\cdot\bar{\Lambda}_r^{(t)}$ for $r \in \{\pm 1\}$, therefore, $\bar{\Lambda}_r^{(t)} = O(\bar{\Lambda}_r^{(0)}) = \widetilde{O}(d^{-\frac{1}{4}})$. If $p < 1/2$, we have off-diagonal correlation $\widehat{\Lambda}_r^{(t+1)} \le (1 - \eta\lambda) \cdot \widehat{\Lambda}_r^{(t)}$ for $r \in \{\pm 1\}$, therefore, $\widehat{\Lambda}_r^{(t)} = O(\widehat{\Lambda}_r^{(0)}) = \widetilde{O}(d^{-\frac{1}{4}})$. In this paper, we mainly focus on $p > 1/2$.

According to Sections C.5.1 and C.5.2, we have obtained following upper bounds and lower bounds for *feature learning* term $\widehat{\Lambda}_r^{(t)}, \bar{\Lambda}_r^{(t)}, r \in \{\pm 1\}$ and *noise memorization* term $\Gamma^{(t)}$: When $t \in [0, T_r]$, we have

$$
\begin{aligned}
&\widehat{\Lambda}_r^{(t+1)} \ge \widehat{\Lambda}_r^{(t)} + \eta \cdot (2p - 1) \cdot \Theta(d) \cdot (\widehat{\Lambda}_r^{(t)})^{q-1} \text{ and } \bar{\Lambda}_r^{(t+1)} \le (1 - \eta\lambda) \cdot \bar{\Lambda}_r^{(t)}, \text{ for } r \in \{\pm 1\}; \\
&\Gamma^{(t+1)} \le (1 - \eta\lambda) \cdot \Gamma^{(t)} + \eta \cdot \max\bigg\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\bigg\} \cdot (\Gamma^{(t)})^{q-1}.
\end{aligned}
\tag{C.26}
$$

According to Condition 3.1, assume $n_{\mathrm{u}} = \Omega(d^{4\epsilon})$ and note that $\epsilon < 1/8$, we have

$$
\max\bigg\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{\Theta}\Big(\frac{d^{1+2\epsilon}}{n_{\mathrm{u}}}\Big)\bigg\} = \max\big\{\widetilde{\Theta}(d^{\frac{1}{2}+2\epsilon}), \widetilde{O}(d^{1-2\epsilon})\big\} = \widetilde{O}(d^{1-2\epsilon}),
$$

leading to

$$
\Gamma^{(t+1)} \le (1 - \eta\lambda) \cdot \Gamma^{(t)} + \eta \cdot \widetilde{\Theta}(d^{1-2\epsilon}) \cdot (\Gamma^{(t)})^{q-1}.
$$

By leveraging tensor power method introduced in Lemma E.4, we can prove following lemma about the magnitude of $\Gamma^{(t)}$:

Lemma C.18. $\Gamma^{(t)}$ remains initialization magnitude during $[0, \max_{r\in\{\pm1\}}\{T_r\}]$.

Proof of Lemma C.18. Let T_r^* be the first iteration t in which $\widehat{\Lambda}_r^{(t)} \geq A$ for $r \in \{\pm1\}$, let T^* be the first iteration t in which $\Gamma^{(t)} \geq A'$, then according to Lemma E.4, we know

$$\sum_{t\geq 0, x_t \leq A} \eta \leq \frac{\delta}{(1-(1+\delta)^{-(q-2)})x_0 C_1} + \eta \cdot \frac{C_2}{C_1}(1+\delta)^{q-1}\bigg(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\bigg),$$

$$\sum_{t\geq 0, x_t \leq A} \eta \geq \frac{\delta\big(1-(x_0/A)^{q-2}\big)}{(1+\delta)^{q-1}\big(1-(1+\delta)^{-(q-2)}\big)x_0 C_2} - \eta \cdot (1+\delta)^{-(q-1)}\bigg(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\bigg).$$

And it follows that

$$\eta \cdot T_r^* \leq \frac{\delta}{(1-(1+\delta)^{-(q-2)})\widehat{\Lambda}_r^{(0)} C_1} + \eta \cdot \frac{C_2}{C_1}(1+\delta)^{q-1}\bigg(1 + \frac{\log(A/\widehat{\Lambda}_r^{(0)})}{\log(1+\delta)}\bigg),$$

$$\eta \cdot T^* \geq \frac{\delta'\big(1-(x_0/A')^{q-2}\big)}{(1+\delta)^{q-1}\big(1-(1+\delta)^{-(q-2)}\big)\Gamma^{(0)} C_2'} - \eta \cdot (1+\delta')^{-(q-1)}\bigg(1 + \frac{\log(A'/\Gamma^{(0)})}{\log(1+\delta')}\bigg),$$

where $C_1, C_2 = (2p-1)\cdot\widetilde{\Theta}(d)$ and $C_1', C_2' = \widetilde{\Theta}(d^{1-2\epsilon})$ according to (C.26).

Taking $A = \Theta(1/m), A' = C \cdot \Gamma^{(t)}$ where C is a large constant and $C = \Theta(1)$, $\delta = \delta' = \frac{1}{2}$ and note that $\widehat{\Lambda}_r^{(0)} = \widetilde{\Theta}(\sigma_0 d^{\frac{1}{2}}) = \widetilde{\Theta}(d^{-\frac{1}{4}}), \Gamma^{(0)} = \widetilde{\Theta}(\sigma_0\sigma_p d^{\frac{1}{2}}) = \widetilde{\Theta}(d^{-\frac{1}{4}+\epsilon})$, we have

$$\eta \cdot T_r^* \leq \widetilde{\Theta}(d^{-\frac{3}{4}}) + \eta \cdot \widetilde{\Theta}(1) = \widetilde{\Theta}(d^{-\frac{3}{4}}), \tag{C.27}$$

and

$$\eta \cdot T^* \geq \widetilde{\Theta}(d^{-\frac{3}{4}+\epsilon}) - \eta \cdot \widetilde{\Theta}(1) = \widetilde{\Theta}(d^{-\frac{3}{4}+\epsilon}). \tag{C.28}$$

Therefore, combining (C.27) and (C.28), we have $\eta \cdot T^* \geq \widetilde{\Theta}(d^{-\frac{3}{4}+\epsilon}) > \widetilde{\Theta}(d^{-\frac{3}{4}}) \geq \eta \cdot T_r^*$, leading to $T^* > T_r^*$ for both $r \in \{-1. +1\}$. This indicates that when $\widehat{\Lambda}_1^{(t)}, \widehat{\Lambda}_{-1}^{(t)}$ reach $\Theta(1/m)$, $\Gamma^{(t)}$ remain the same magnitude as initialization. $\square$

By leveraging tensor power method, we can also estimate the length of Stage I, i.e. T_1, T_{-1}, by applying tensor power method. To use tensor power method, we need to upper-bound the increasing speed of $\widehat{\Lambda}_r^{(t)}$. We have the following lemma:

Lemma C.19. For $r \in \{\pm1\}$, we have with high probability that

$$\widehat{\Lambda}_r^{(t+1)} \geq (1-\eta\lambda)\cdot\widehat{\Lambda}_r^{(t)} + \eta\cdot q\Big(p - \frac{1}{2} - o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_r^{(t)}\big)^{q-1},$$

$$\widehat{\Lambda}_r^{(t+1)} \leq (1-\eta\lambda)\cdot\widehat{\Lambda}_r^{(t)} + \eta\cdot q\Big(p - \frac{1}{2} + o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_r^{(t)}\big)^{q-1}.$$

Proof of Lemma C.19. Let $j^* = \arg\max_{1\leq j\leq m}\langle \mathbf{w}_j^{(t+1)}, \mathbf{v}\rangle$ and note that $u_{j^*} = \mathbb{1}_{[1\leq j\leq m]} = \mathbb{1}_{[m+1\leq j\leq 2m]} = 1$, then we have

$$\begin{aligned}
\widehat{\Lambda}_1^{(t+1)} &= \langle \mathbf{w}_{j^*}^{(t+1)}, \mathbf{v}\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle + \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\underbrace{\bigg(\sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \sum_{i\in S_{-1}} y_i\widehat{y}_i c_i^{(t)}[-\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{2}\|\mathbf{v}\|_2^{q-1}\bigg)}_{\clubsuit} \\
&\quad + \frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\underbrace{\bigg(\sum_{i\in S_1'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \sum_{i\in S_1'} b_i^{(t)}[-\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\bigg)}_{\star}.
\end{aligned} \tag{C.29}$$

For ♣, according to Lemma C.12, we have

$$\begin{aligned}
&\underbrace{\sum_{i\in S_1} y_i\widehat{y}_i c_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2+\sum_{i\in S_{-1}} y_i\widehat{y}_i c_i^{(t)}[-\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit}\\
&= n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot[\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2+n_{-1}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot[-\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2\\
&\le n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}+n_{-1}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\bar{\Lambda}_{-1}^{(t)}\big)^{q-1}\\
&= n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1},
\end{aligned}\tag{C.30}$$

where the last equality is due to $\widehat{\Lambda}_1^{(t)}=\omega(\bar{\Lambda}_{-1}^{(t)})$.

For ★, according to Lemma C.11, we have

$$\begin{aligned}
&\underbrace{\sum_{i\in S_1'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2+\sum_{i\in S_1'} b_i^{(t)}[-\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar}\\
&= n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot[\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2+n_{-1}'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot[-\langle \mathbf{w}_{j^*}^{(t)},\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \qquad \text{(C.31)}\\
&\le n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}+n_{-1}'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\bar{\Lambda}_{-1}^{(t)}\big)^{q-1}\\
&= n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1},
\end{aligned}$$

where the last equality is due to $\widehat{\Lambda}_1^{(t)}=\omega(\bar{\Lambda}_{-1}^{(t)})$.

Plugging (C.30) and (C.31) into (C.29), we have

$$\begin{aligned}
&\widehat{\Lambda}_1^{(t+1)}\\
&\le(1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)}+\frac{q\eta}{n_{\mathrm{l}}+n_{\mathrm{u}}}\bigg(n_1\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}+n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\bigg)\\
&=(1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)}+\frac{q\eta n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}+\frac{q\eta n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\\
&=(1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)}+q\eta\cdot\bigg(\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\pm o(1)\Big)+\frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(\frac{1}{2}\pm o(1)\Big)\bigg)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}\\
&=(1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)}+q\eta\cdot\bigg(\underbrace{\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big)+\frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}}_{\spadesuit}\pm o(1)\bigg)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}.
\end{aligned}\tag{C.32}$$

Note that we have already proved in (C.10) that

$$\widehat{\Lambda}_1^{(t+1)}\le(1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)}+q\eta\cdot\bigg(\underbrace{\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big)+\frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}}_{\spadesuit}\pm o(1)\bigg)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}.\tag{C.33}$$

Note we have already prove in (C.11) that

$$\underbrace{\frac{n_1}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\Big(p-\frac{1}{2}\Big)+\frac{n_1'}{n_{\mathrm{l}}+n_{\mathrm{u}}}\cdot\frac{1}{2}}_{\spadesuit}=\frac{1}{2}\cdot\Big(p-\frac{1}{2}\Big)\pm o(1)$$

Therefore, we have

$$\widehat{\Lambda}_1^{(t+1)} \geq (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + q\eta\cdot\Big(p-\frac{1}{2}-o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1},$$
$$\widehat{\Lambda}_1^{(t+1)} \leq (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + q\eta\cdot\Big(p-\frac{1}{2}+o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_1^{(t)}\big)^{q-1}.$$

In a similar way, we can prove that

$$\widehat{\Lambda}_{-1}^{(t+1)} \geq (1-\eta\lambda)\cdot\widehat{\Lambda}_{-1}^{(t)} + q\eta\cdot\Big(p-\frac{1}{2}-o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_{-1}^{(t)}\big)^{q-1},$$
$$\widehat{\Lambda}_{-1}^{(t+1)} \leq (1-\eta\lambda)\cdot\widehat{\Lambda}_{-1}^{(t)} + q\eta\cdot\Big(p-\frac{1}{2}+o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot\big(\widehat{\Lambda}_{-1}^{(t)}\big)^{q-1},$$

which completes the proof of this lemma. □

Lemma C.20 (Length of pre-training)**.** For $r\in\{\pm1\}$, let T_r be the first iteration that $\widehat{\Lambda}_r^{(t)}$ reaches $\Theta(1/m)$ respectively. Then $T_r=\widetilde{\Theta}(d^{\frac{q}{4}-\frac{3}{2}})/\eta$ for all $r\in\{\pm1\}$.

Proof of Lemma C.20. By leveraging tensor power method given in Lemma E.4,

$$\sum_{t\geq 0, x_t\leq A}\eta \leq \frac{\delta}{(1-(1+\delta)^{-(q-2)})x_0^{q-2}C_1} + \eta\cdot\frac{C_2}{C_1}(1+\delta)^{q-1}\bigg(1+\frac{\log(A/x_0)}{\log(1+\delta)}\bigg),$$
$$\sum_{t\geq 0, x_t\leq A}\eta \geq \frac{\delta\big(1-(x_0/A)^{q-2}\big)}{(1+\delta)^{q-1}\big(1-(1+\delta)^{-(q-2)}\big)x_0^{q-2}C_2} - \eta\cdot(1+\delta)^{-(q-1)}\bigg(1+\frac{\log(A/x_0)}{\log(1+\delta)}\bigg),$$

we have for $r\in\{\pm1\}$ that

$$\eta\cdot T_r^* = \sum_{t\geq 0,\widehat{\Lambda}_r^{(t)}\leq A}\eta \leq \underbrace{\frac{\delta}{(1-(1+\delta)^{-(q-2)})(\widehat{\Lambda}_r^{(0)})^{q-2}C_1}}_{(i)} + \underbrace{\eta\cdot\frac{C_2}{C_1}(1+\delta)^{q-1}\bigg(1+\frac{\log(A/\widehat{\Lambda}_r^{(0)})}{\log(1+\delta)}\bigg)}_{(ii)},$$
$$\eta\cdot T_r^* = \sum_{t\geq 0,\widehat{\Lambda}_r^{(t)}\leq A}\eta \geq \underbrace{\frac{\delta\big(1-(x_0/A)^{q-2}\big)}{(1+\delta)^{q-1}\big(1-(1+\delta)^{-(q-2)}\big)(\widehat{\Lambda}_r^{(0)})^{q-2}C_2}}_{(iii)} - \underbrace{\eta\cdot(1+\delta)^{-(q-1)}\bigg(1+\frac{\log(A/\widehat{\Lambda}_r^{(0)})}{\log(1+\delta)}\bigg)}_{(iv)},$$

where C_1 is taken as $q\big(p-\frac{1}{2}-o(1)\big)\cdot\|\mathbf{v}\|_2^2$ and C_2 is taken as $q\big(p-\frac{1}{2}+o(1)\big)\cdot\|\mathbf{v}\|_2^2$ according to Lemma C.19. Taking $\delta=\frac{1}{k}, A=\Theta(1/m)$ and note that terms $(ii),(iv)$ are respectively dominated by terms $(i),(iii)$ when η is sufficiently small and letting $k\to\infty$, we have

$$\frac{1}{(\widehat{\Lambda}_r^{(0)})^{q-2}C_2} - \{\text{lower order terms}\} \leq \eta\cdot T_r^* \leq \frac{1}{(\widehat{\Lambda}_r^{(0)})^{q-2}C_1} + \{\text{lower order terms}\},$$

for $r\in\{\pm1\}$. It follows that

$$\eta\cdot T_r^* = \frac{1}{q\big(p-\frac{1}{2}\big)\|\mathbf{v}\|_2^2\cdot(\widehat{\Lambda}_r^{(0)})^{q-2}} \pm \{\text{lower order terms}\}. \tag{C.34}$$

And by Lemma C.9, we have $\eta\cdot T_r^*=\Theta(1/q\big(p-\frac{1}{2}\big)\|\mathbf{v}\|_2^2\cdot(\sqrt{\log(m)}\sigma_0\|\mathbf{v}\|_2)^{q-2})=\widetilde{\Theta}(d^{q/4-3/2})$, which completes the proof. □

The discussion in this section verifies Lemma C.4 and provides a clear understanding about how $\widehat{\Lambda}_r^{(t)},\bar{\Lambda}_r^{(t)}$ varies within the iteration range $[0,T_r]$ for $r\in\{\pm1\}$. Note that the iteration numbers when $\widehat{\Lambda}_1^{(t)}$ and $\widehat{\Lambda}_{-1}^{(t)}$ reaches $\Theta(1/m)$ (T_1 and T_{-1}) are different, however, since T_{-1} and T_1 have the same magnitude, it remains clear that although $T_1\neq T_{-1}$ (wlog, assume $T_1<T_{-1}$), we still have $\widehat{\Lambda}_1^{(t)}=\widetilde{\Theta}(1)$ and $\bar{\Lambda}_1^{(t)}=\widetilde{O}(d^{-\frac{1}{4}})$ within the iteration range $[T_1,T_{-1}]$, since off-diagonal feature learning also costs time no less than order $\Theta(1/\eta\sigma_0\|\mathbf{v}\|_2^q(\log m)^{(q-2)/2})$, which is higher order than

$|T_1 - T_{-1}| = \Theta(1/\eta\sigma_0\|\mathbf{v}\|_2^q(\log m)^{(q-1)/2})$, according to (C.34) and Lemma C.9. Therefore, at time $T_0 := \max\{T_1, T_{-1}\}$, off-diagonal $\bar{\Lambda}_1^{(t)}, \bar{\Lambda}_{-1}^{(t)}$ still remain initialization magnitude $\widetilde{O}(d^{-\frac{1}{4}})$, $\Gamma_1^{(t)}, \Gamma_{-1}^{(t)}$ remain initialization magnitude $\widetilde{O}(d^{-\frac{1}{4}+\epsilon})$, while on-diagonal $\widehat{\Lambda}_1^{(t)}, \widehat{\Lambda}_{-1}^{(t)}$ reach and then remain $\widetilde{\Theta}(1)$.

C.6 Proof of Lemma C.3

If we only use labeled data S' for the optimization of CNN, according to Lemma D.1, we have

$$\begin{aligned}\mathbf{w}_j^{(t+1)} &= \mathbf{w}_j^{(t)} - \nabla_{\mathbf{w}_j} L_{S'}(\mathbf{W})\\ &= (1-\eta\lambda)\cdot\mathbf{w}_j^{(t)} + \frac{q\eta u_j}{n_l}\sum_{i=1}^{n_l} b_i^{(t)} y_i'\big([\langle\mathbf{w}_j, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\cdot y_i'\cdot\mathbf{v} + [\langle\mathbf{w}_j,\boldsymbol{\xi}_i'\rangle]_+^{q-1}\cdot\boldsymbol{\xi}_i'\big),\end{aligned}$$

where $u_j := \mathbb{1}_{[1\le j\le m]} - \mathbb{1}_{[m+1\le 2m]}$, $b_i^{(t)} = -\ell'(y_i'\cdot f_{\mathbf{W}}(\mathbf{x}_i')) = \exp[-y_i'\cdot f_{\mathbf{W}}(\mathbf{x}_i')]/(1+\exp[-y_i'\cdot f_{\mathbf{W}}(\mathbf{x}_i')])$.

Notice that $\mathbf{v}$ and $\boldsymbol{\xi}_i'$ are orthogonal to each other, we have

$$\begin{aligned}\langle\mathbf{w}_j^{(t+1)},\mathbf{v}\rangle &= (1-\eta\lambda)\cdot\langle\mathbf{w}_j^{(t)},\mathbf{v}\rangle + \frac{q\eta u_j}{n_l}\sum_{i=1}^{n_l} b_i^{(t)}\cdot[\langle\mathbf{w}_j^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\cdot\|\mathbf{v}\|_2^2,\\ \langle\mathbf{w}_j^{(t+1)},\boldsymbol{\xi}_l'\rangle &= (1-\eta\lambda)\cdot\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i\rangle + \frac{q\eta u_j}{n_l}\sum_{i=1}^{n_l} b_i^{(t)} y_i'\cdot[\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle]_+^{q-1}\cdot\langle\boldsymbol{\xi}_i',\boldsymbol{\xi}_l'\rangle, i\in[n_l].\end{aligned}$$

Let T_i' be the first iteration that $\Gamma_i'^{(t)}$ reaches $\Theta(1/m)$, then we have following lemma:

Lemma C.21. As long as $\Gamma_i'^{(t)} \le \Theta(1/m)$, $b_i^{(t)} := -\ell'(y_i'\cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}_i'))$ will remain $1/2\pm o(1)$.

Proof of Lemma C.21. Note that $\ell(z) = \log(1+\exp(-z))$ and $-\ell'(z) = \exp(-z)/\big(1+\exp(-z)\big)$, and without loss of generality assuming $y_i' = 1$, we can express $b_i^{(t)}$ as follow:

$$b_i^{(t)} = -\ell'(f_{\mathbf{W}^{(t)}}(\mathbf{x}_i')) = \frac{e^{\sum_{j=m+1}^{2m}[\sigma(\langle\mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)]}}{e^{\sum_{j=1}^{m}[\sigma(\langle\mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)]} + e^{\sum_{j=m+1}^{2m}[\sigma(\langle\mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)]}},$$

Since $\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}\rangle)$ will dominate $\sigma(\langle\mathbf{w}_j^{(t)},\mathbf{v}\rangle)$, which will be proved later by using *tensor power method*, we have

$$b_i^{(t)} = -\ell'(f_{\mathbf{W}^{(t)}}(\mathbf{x}_i')) = \frac{e^{\sum_{j=m+1}^{2m}[\sigma(\langle\mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)]}}{e^{\sum_{j=1}^{m}\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)\{+\text{lower order term}\}} + e^{\sum_{j=m+1}^{2m}[\sigma(\langle\mathbf{w}_j^{(t)},\mathbf{v}\rangle)+\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)]}},$$

On the one side,

$$\begin{aligned}b_i^{(t)} &\ge \frac{1}{e^{\sum_{j=1}^{m}\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)\{+\text{lower order term}\}}+1}\\ &\ge \frac{1}{e^{m(\Gamma_i'^{(t)})^q\{+\text{lower order term}\}}+1}\\ &\ge \frac{1}{e^{\Theta(m^{-(q-1)})}+1} = \frac{1}{2+o(1)} = \frac{1}{2}-o(1).\end{aligned}$$

On the other side, according to Lemma C.5, we have $\bar{\Lambda}_1^{(t)} = \widetilde{O}(d^{-\frac{1}{4}})$, it follows that

$$\begin{aligned}b_i^{(t)} &\le \frac{e^{m(\bar{\Lambda}_1^{(t)})^q+o(1)}}{e^{\sum_{j=1}^{m}\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)+\{\text{lower order term}\}} + e^{m(\bar{\Lambda}_1^{(t)})^q+o(1)}}\\ &= \frac{1+o(1)}{e^{\sum_{j=1}^{m}\sigma(\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle)+\{\text{lower order term}\}}+1+o(1)}\end{aligned}$$

$$\le \frac{1+o(1)}{1+1+o(1)} = \frac{1}{2} + o(1).$$

Therefore, we have $b_i^{(t)} = 1/2 \pm o(1)$ and the other case of $y_i = -1$ can be proved in a similar way. $\square$

With the help of above lemma, we are now ready to prove Lemma C.3.

Proof of Lemma C.3. Let $j^* = \arg\max_{1\le j\le m}\langle \mathbf{w}_j^{(t+1)}, \mathbf{v}\rangle$ and note that $u_j = 1$, according to Lemma C.21, we have

$$\begin{aligned}
\widehat{\Lambda}_1^{(t+1)} &= \langle \mathbf{w}_{j^*}^{(t+1)}, \mathbf{v}\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle + \frac{q\eta}{n_1}\sum_{i=1}^{n_1} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle + \frac{q\eta}{n_1}\sum_{i\in S_1'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 + \frac{q\eta}{n_1}\sum_{i\in S_{-1}'} b_i^{(t)}[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle + \frac{q\eta}{n_1}\underbrace{\sum_{i\in S_1'}\Big(\frac{1}{2}\pm o(1)\Big)[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} \\
&\quad + \frac{q\eta}{n_1}\underbrace{\sum_{i\in S_{-1}'}\Big(\frac{1}{2}\pm o(1)\Big)[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar}
\end{aligned} \tag{C.35}$$

For ♣, we have

$$\begin{aligned}
\underbrace{\sum_{i\in S_1'}\Big(\frac{1}{2}\pm o(1)\Big)[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\clubsuit} &= n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, \mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&\le n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\widehat{\Lambda}_1^{(t)})^{q-1}.
\end{aligned} \tag{C.36}$$

For ★, we have

$$\begin{aligned}
\underbrace{\sum_{i\in S_{-1}'}\Big(\frac{1}{2}\pm o(1)\Big)[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2}_{\bigstar} &= n_{-1}'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot[\langle \mathbf{w}_{j^*}^{(t)}, -\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&\le n_{-1}'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\bar{\Lambda}_{-1}^{(t)})^{q-1}.
\end{aligned} \tag{C.37}$$

By plugging (C.36) and (C.37) in (C.35), and according to Lemma C.14, we have with probability at least $1-4\delta$ that

$$\begin{aligned}
\widehat{\Lambda}_1^{(t+1)} &\le (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + \frac{q\eta}{n_1}\bigg(n_1'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\widehat{\Lambda}_1^{(t)})^{q-1} + n_{-1}'\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\bar{\Lambda}_{-1}^{(t)})^{q-1}\bigg) \\
&\le (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + \frac{q\eta}{n_1}\Bigg(\bigg(\frac{n_1}{2}+\sqrt{\frac{n_1}{2}\log\frac{1}{\delta}}\bigg)\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\widehat{\Lambda}_1^{(t)})^{q-1} \\
&\qquad + \bigg(\frac{n_1}{2}+\sqrt{\frac{n_1}{2}\log\frac{1}{\delta}}\bigg)\cdot\Big(\frac{1}{2}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\bar{\Lambda}_{-1}^{(t)})^{q-1}\Bigg) \\
&= (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + q\eta\bigg(\Big(\frac{1}{4}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\widehat{\Lambda}_1^{(t)})^{q-1} + \Big(\frac{1}{4}\pm o(1)\Big)\cdot\|\mathbf{v}\|_2^2\cdot(\bar{\Lambda}_{-1}^{(t)})^{q-1}\bigg)
\end{aligned}$$

$$= (1-\eta\lambda)\cdot\widehat{\Lambda}_1^{(t)} + \eta\cdot\Theta(d)\cdot\Big((\widehat{\Lambda}_1^{(t)})^2 + (\bar{\Lambda}_{-1}^{(t)})^{q-1}\Big).$$

And we can prove in the same way that with probability at least $1-4\delta$ we have

$$\widehat{\Lambda}_{-1}^{(t+1)} \le (1-\eta\lambda)\cdot\widehat{\Lambda}_{-1}^{(t)} + \eta\cdot\Theta(d)\cdot\Big((\widehat{\Lambda}_{-1}^{(t)})^{q-1} + (\bar{\Lambda}_1^{(t)})^{q-1}\Big).$$

Let $j^\star = \arg\max_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t+1)}, \mathbf{v}\rangle$ and note that $u_j=-1$, we have

$$\begin{aligned}
\bar{\Lambda}_1^{(t+1)} &= \langle \mathbf{w}_{j^\star}^{(t+1)}, \mathbf{v}\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\star}^{(t)}, \mathbf{v}\rangle - \frac{q\eta}{n_1}\sum_{i=1}^{n_1} b_i^{(t)}[\langle \mathbf{w}_{j^\star}^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\|\mathbf{v}\|_2^2 \\
&\le (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\star}^{(t)}, \mathbf{v}\rangle \\
&\le (1-\eta\lambda)\cdot\bar{\Lambda}_1^{(t)}.
\end{aligned} \tag{C.38}$$

And we can prove in the same way that $\bar{\Lambda}_{-1}^{(t+1)} \le (1-\eta\lambda)\cdot\bar{\Lambda}_{-1}^{(t)}$.

Next, we consider the increasing rate of $\Gamma_l'^{(t)}$ where $l\in[n_1]$ is fixed. If $y_l=1$, let $j^\natural = \mathrm{argmax}_{1\le j\le m}\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l'\rangle$ and note that $u_j=1$, we have

$$\begin{aligned}
\Gamma_l'^{(t+1)} &\ge \langle \mathbf{w}_{j^\natural}^{(t+1)}, \boldsymbol{\xi}_l'\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle + \frac{q\eta}{n_1}\sum_{i=1}^{n_1} b_i^{(t)} y_i'\cdot[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1}\cdot\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l'\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle + \frac{q\eta}{n_1} b_l^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle]_+^{q-1}\|\boldsymbol{\xi}_l'\|_2^2 + \frac{q\eta}{n_1}\sum_{i\in[n_1], i\ne l} b_i^{(t)} y_i'[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1}\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l'\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle + \frac{q\eta}{n_1} b_l^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle]_+^{q-1}\|\boldsymbol{\xi}_l'\|_2^2\{\pm \text{ lower order terms}\} \\
&\ge (1-\eta\lambda)\cdot\Gamma_l'^{(t)} + \frac{q\eta}{n_1}\cdot\Big(\frac{1}{2} - o(1)\Big)\cdot\|\boldsymbol{\xi}_l'\|_2^2\cdot(\Gamma_l'^{(t)})^{q-1} \\
&= (1-\eta\lambda)\cdot\Gamma_l'^{(t)} + \eta\cdot\widetilde{\Theta}(d^{1+2\epsilon})\cdot(\Gamma_l'^{(t)})^{q-1},
\end{aligned} \tag{C.39}$$

where the third equality holds if we properly choose the order of λ.

If $y_l=-1$, let $j^\sharp = \mathrm{argmax}_{m+1\le j\le 2m}\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_l'\rangle$ and note that $u_j=-1$, we have

$$\begin{aligned}
\Gamma_l'^{(t+1)} &\ge \langle \mathbf{w}_{j^\natural}^{(t+1)}, \boldsymbol{\xi}_l'\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle - \frac{q\eta}{n_1}\sum_{i=1}^{n_1} b_i^{(t)} y_i'\cdot[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1}\cdot\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l'\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle + \frac{q\eta}{n_1} b_l^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle]_+^{q-1}\|\boldsymbol{\xi}_l'\|_2^2 - \frac{q\eta}{n_1}\sum_{i\in[n_1], i\ne l} b_i^{(t)} y_i'[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1}\langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_l'\rangle \\
&= (1-\eta\lambda)\cdot\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle + \frac{q\eta}{n_1} b_l^{(t)}[\langle \mathbf{w}_{j^\natural}^{(t)}, \boldsymbol{\xi}_l'\rangle]_+^{q-1}\|\boldsymbol{\xi}_l'\|_2^2\{\pm \text{ lower order terms}\} \\
&\ge (1-\eta\lambda)\cdot\Gamma_l'^{(t)} + \frac{q\eta}{n_1}\cdot\Big(\frac{1}{2} - o(1)\Big)\cdot\|\boldsymbol{\xi}_l'\|_2^2\cdot(\Gamma_l'^{(t)})^{q-1} \\
&= (1-\eta\lambda)\cdot\Gamma_l'^{(t)} + \eta\cdot\widetilde{\Theta}(d^{1+2\epsilon})\cdot(\Gamma_l'^{(t)})^{q-1},
\end{aligned} \tag{C.40}$$

where the third equality holds if we properly choose the order of λ.

According to (C.39) and (C.40), we always have

$$\Gamma_l'^{(t+1)} \ge (1-\eta\lambda)\cdot\Gamma_l'^{(t)} + \eta\cdot\widetilde{\Theta}(d^{1+2\epsilon})\cdot(\Gamma_l'^{(t)})^{q-1}.$$

$\square$

C.7 Proof of Lemma C.5

By applying Lemma E.4 to $\Gamma_i^{(t)}$ and taking $C_1 = \widetilde{\Theta}(d^{1+2\epsilon})$, $\delta = 1/2$, $A = \Theta(1/m)$, we have

$$\sum_{t \geq 0, \Gamma_i^{(t)} \leq A} \eta \leq \Theta(1/C_1(\Gamma_i^{(0)})^{q-2}) = \widetilde{\Theta}(d^{(\frac{1}{4}-\epsilon)q-\frac{3}{2}}).$$

And note the definition of T_i', we have

$$\eta \cdot T_i' = \widetilde{\Theta}(d^{(\frac{1}{4}-\epsilon)q-\frac{3}{2}}). \tag{C.41}$$

In Lemma C.3, we have already prove that

$$\begin{aligned} &\widehat{\Lambda}_r^{(t+1)} \leq (1-\eta\lambda) \cdot \widehat{\Lambda}_r^{(t)} + \eta \cdot \Theta(d) \cdot \Big((\widehat{\Lambda}_r^{(t)})^{q-1} + (\bar{\Lambda}_{-r}^{(t)})^{q-1} \Big), \\ &\widehat{\Lambda}_r^{(t+1)} \leq (1-\eta\lambda) \cdot \widehat{\Lambda}_r^{(t+1)}, r \in \{\pm 1\}. \end{aligned} \tag{C.42}$$

Define $\Lambda^{(t)} := \max_{r \in \{\pm 1\}} \{\widehat{\Lambda}_r^{(t)}, \bar{\Lambda}_r^{(t)}\}$, according to (C.42), we have

$$\Lambda^{(t+1)} \leq (1-\eta\lambda) \cdot \Lambda^{(t)} + \eta \cdot \Theta(d) \cdot (\Lambda^{(t)})^{q-1}.$$

By applying Lemma E.4 to $\Lambda^{(t)}$, and taking $C_1 = \Theta(d)$, $\delta = 1/2$, $A = C \cdot \Lambda^{(0)}$, where A is a large constant, we have

$$\sum_{t \geq 0, \Lambda^{(t)} \leq A} \eta \geq \Theta(1/C_1(\Lambda^{(0)})^{q-2}) = \widetilde{\Theta}(d^{\frac{q}{4}-\frac{3}{2}}).$$

Let T' be the first iteration that $\Lambda^{(t)}$ reaches $C \cdot \Lambda^{(0)}$, then we have

$$\eta \cdot T' = \widetilde{\Theta}(d^{\frac{q}{4}-\frac{3}{2}}). \tag{C.43}$$

According to (C.41) and (C.43), we have $T' = \omega(T_i')$, which indicates that when $\Gamma_i^{(t)}$ reaches $\Theta(1/m)$, $\Lambda^{(t)}$ remains initialization magnitude $\widetilde{\Theta}(d^{-\frac{1}{4}})$.

C.8 Empirical, test error and loss for early stopped classifier

Assume the accuracy of pseudo-labeler p is larger than $1/2$. We first estimate the empirical loss for early stopped classifier $f_{\mathbf{W}^{(T_0)}}$, where $T_0 = \max_{r \in \{\pm 1\}} \{T_r\}$ and T_r is defined as the first iteration that $\widehat{\Lambda}_r^{(t)}$ reaches $\Theta(1/m)$. According to Section C.5.3 and Lemma C.18, we have $\widehat{\Lambda}_r^{(T_0)} = \widetilde{\Theta}(1), \bar{\Lambda}_r^{(T_0)} = \widetilde{O}(d^{-\frac{1}{4}}), \Gamma^{(t)} = \widetilde{O}(d^{-\frac{1}{4}+\epsilon})$, for $r \in \{\pm 1\}$. We have the following lemma:

Lemma C.22. Early stopped classifier $f_{\mathbf{W}^{(T_0)}}(\mathbf{x})$ possesses following properties:

1. Training error of early stopped classifier $f_{\mathbf{W}^{(T_0)}}(\mathbf{x})$ is asymptotically $1-p$: $\frac{1}{n_{\mathrm{u}}+n_{\mathrm{l}}} \big(\sum_{i=1}^{n_{\mathrm{u}}} \mathbb{1}[\widehat{y}_i \cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i) \leq 0] + \sum_{i=1}^{n_{\mathrm{l}}} \mathbb{1}[y_i' \cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i') \leq 0] \big) = 1 - p \pm o(1)$.
2. Test error is nearly $1-p$, if we use pseudo-label $\widehat{y}$ generated by pseudo-labeler as target: $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}, \widehat{y}\sim y\cdot\mathcal{B}(p)}[\widehat{y} \cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x}) \leq 0] = 1 - p \pm o(1)$.
3. Test error is nearly 0, if we use true label y as target: $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x}) \leq 0] = o(1)$ and hence sign $f_{\mathbf{W}^{(T_0)}}(\mathbf{x}) = \mathrm{sign}(y)$ with high probability,

where p is the accuracy of the pseudo-labeler. We can regard p as the probability that $\mathbf{x}_i$ is paired with true label y_i, $1-p$ is the probability that $\mathbf{x}_i$ is paired with wrong label $-y_i$.

Proof of Lemma C.22. Recall the definition of $f_{\mathbf{W}}$ in (2.1) that

$$\begin{aligned} f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i) = &\sum_{j=1}^{m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0)}, y_i \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i \rangle\big) \Big] \\ &- \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0)}, y_i \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i \rangle\big) \Big]. \end{aligned}$$

According to Section C.5.3 and Lemma C.18, we have $\widehat{\Lambda}_r^{(T_0)} = \widetilde{\Theta}(1), \bar{\Lambda}_r^{(T_0)} = \widetilde{O}(d^{-\frac{1}{4}}), \Gamma^{(t)} = \max\big\{\max_{i\in[n_\mathrm{u}]}\Gamma_i^{(t)}, \max_{i\in[n_\mathrm{l}]}\Gamma_i'^{(t)}\big\} = \widetilde{O}(d^{-\frac{1}{4}+\epsilon})$, for $r \in \{\pm 1\}$. If $y_i = 1$, we have following lower bound for $f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i)$

$$
\begin{aligned}
f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i) &= \sum_{j=1}^{m}\Big[\sigma\big(\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] - \sum_{j=m+1}^{2m}\Big[\sigma\big(\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] \\
&\geq \big(\widehat{\Lambda}_1^{(T_0)}\big)^q + \big(\Gamma_i^{(T_0)}\big)^q - m\big(\bar{\Lambda}_1^{(T_0)}\big)^q - m\big(\Gamma_i^{(T_0)}\big)^q \\
&\geq \big(\widehat{\Lambda}_1^{(T_0)}\big)^q\{-\text{ lower order terms}\},
\end{aligned}
$$

and following upper bound for $f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i)$:

$$
\begin{aligned}
f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i) &= \sum_{j=1}^{m}\Big[\sigma\big(\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] - \sum_{j=m+1}^{2m}\Big[\sigma\big(\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] \\
&\leq m(\widehat{\Lambda}_1^{(T_0)})^q + m(\Gamma_i^{(T_0)})^q - \big(\bar{\Lambda}_1^{(T_0)}\big)^q - \big(\Gamma_i^{(T_0)}\big)^q \\
&\leq (\widehat{\Lambda}_1^{(T_0)})^q\{+\text{ lower order terms}\}.
\end{aligned}
$$

If $y_i = -1$, we have following upper bound for $f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i)$:

$$
\begin{aligned}
f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i) &= \sum_{j=1}^{m}\Big[\sigma\big(-\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] - \sum_{j=m+1}^{2m}\Big[\sigma\big(-\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] \\
&\leq m\big(\bar{\Lambda}_{-1}^{(T_0)}\big)^q + m\big(\Gamma_i^{(T_0)}\big)^q - \big(\widehat{\Lambda}_{-1}^{(T_0)}\big)^q - \big(\Gamma_i^{(T_0)}\big)^q \\
&\leq -\big(\widehat{\Lambda}_{-1}^{(T_0)}\big)^q\{+\text{ lower order terms}\},
\end{aligned}
$$

and following lower bound for $f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i)$:

$$
\begin{aligned}
f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i) &= \sum_{j=1}^{m}\Big[\sigma\big(-\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] - \sum_{j=m+1}^{2m}\Big[\sigma\big(-\langle \mathbf{w}_j^{(T_0)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0)}, \boldsymbol{\xi}_i\rangle\big)\Big] \\
&\geq \big(\bar{\Lambda}_{-1}^{(T_0)}\big)^q + \big(\Gamma_i^{(T_0)}\big)^q - m\big(\widehat{\Lambda}_{-1}^{(T_0)}\big)^q - m\big(\Gamma_i^{(T_0)}\big)^q \\
&\geq -m\big(\bar{\Lambda}_{-1}^{(T_0)}\big)^q\{-\text{ lower order terms}\}.
\end{aligned}
$$

Therefore, for unlabeled data, we have $y_i \cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i) \in \big[\big(1-o(1)\big)\cdot(\widehat{\Lambda}_{y_i}^{(T_0)})^q, \big(m+o(1)\big)\cdot(\widehat{\Lambda}_{y_i}^{(T_0)})^q\big]$ and hence $\operatorname{sign}\big(f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i)\big) = \operatorname{sign}(y_i)$ holds with high probability. We can also prove for labeled data $(\mathbf{x}_i', y_i')$ that $y_i' \cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i') \in \big[\big(1-o(1)\big)\cdot(\widehat{\Lambda}_{y_i'}^{(T_0)})^q, \big(m+o(1)\big)\cdot(\widehat{\Lambda}_{y_i'}^{(T_0)})^q\big]$, $\operatorname{sign}\big(f_{\mathbf{W}^{(T_0)}}(\mathbf{x}_i')\big) = \operatorname{sign}(y_i')$ in the same way.

Note that $\widehat{y}_i$ takes y_i with probability p, $-y_i$ with probability p and $n_\mathrm{l} = o(n_\mathrm{u})$, the first statement in this lemma follows obviously.

To prove the other two statement, we need to give an upper bound for the norm of $\mathbf{w}_j$. According to the update rule of $\mathbf{w}_j^{(t)}$, we have

$$
\begin{aligned}
\mathbf{w}_j^{(t+1)} = (1-\eta\lambda)\cdot\mathbf{w}_j^{(t)} + \frac{q\eta u_j}{n_\mathrm{l}+n_\mathrm{u}}\cdot\Bigg(&\sum_{i=1}^{n_\mathrm{u}} c_i\widehat{y}_i\big([\langle \mathbf{w}_j^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\cdot y_i\cdot\mathbf{v} + [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i\rangle]_+^{q-1}\cdot\boldsymbol{\xi}_i\big) \\
&+ \sum_{i=1}^{n_\mathrm{l}} b_i y_i'\big([\langle \mathbf{w}_j^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\cdot y_i'\cdot\mathbf{v} + [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1}\cdot\boldsymbol{\xi}_i'\big)\Bigg),
\end{aligned}
$$

leading to

$$
\begin{aligned}
\|\mathbf{w}_j^{(t+1)}\|_2 &\le (1-\eta\lambda)\cdot\|\mathbf{w}_j^{(t)}\|_2 + \frac{q\eta}{n_\mathrm{l}+n_\mathrm{u}}\cdot\Bigg(\sum_{i=1}^{n_\mathrm{u}}\big([\langle\mathbf{w}_j^{(t)}, y_i\cdot\mathbf{v}\rangle]_+^{q-1}\cdot\|\mathbf{v}\|_2 + [\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i\rangle]_+^{q-1}\cdot\|\boldsymbol{\xi}_i\|_2\big) \\
&\quad + \sum_{i=1}^{n_\mathrm{l}}\big([\langle\mathbf{w}_j^{(t)}, y_i'\cdot\mathbf{v}\rangle]_+^{q-1}\cdot\|\mathbf{v}\|_2 + [\langle\mathbf{w}_j^{(t)},\boldsymbol{\xi}_i'\rangle]_+^{q-1}\cdot\|\boldsymbol{\xi}_i'\|_2\big)\Bigg) \\
&\le (1-\eta\lambda)\cdot\|\mathbf{w}_j^{(t)}\|_2 + \frac{q\eta}{n_\mathrm{l}+n_\mathrm{u}}\cdot\Bigg((n_\mathrm{l}+n_\mathrm{u})\cdot\|\mathbf{v}\|_2\cdot\big(\max_{r\in\{\pm1\}}\{\widehat{\Lambda}_r^{(t)},\bar{\Lambda}_r^{(t)}\}\big)^{q-1} \\
&\quad + \Big(\sum_{i\in[n_\mathrm{u}]}\|\boldsymbol{\xi}_i\|_2 + \sum_{i\in[n_\mathrm{l}]}\|\boldsymbol{\xi}_i'\|_2\Big)\cdot\big(\Gamma^{(t)}\big)^{q-1}\Bigg) \\
&\le \|\mathbf{w}_j^{(t)}\|_2 + \eta\cdot\Big(\Theta(d^{\frac{1}{2}})\cdot\widetilde{\Theta}(1) + \Theta(d^{\frac{1}{2}+\epsilon})\cdot\widetilde{O}(d^{(q-1)(-\frac{1}{4}+\epsilon)})\Big) \\
&= \|\mathbf{w}_j^{(t)}\|_2 + \eta\cdot\widetilde{\Theta}(d^{\frac{1}{2}}),
\end{aligned}
\tag{C.44}
$$

where the first inequality is by triangle inequality; the second inequality is due to the definition of $\widehat{\Lambda}_r^{(t)}, \bar{\Lambda}_r^{(t)}, \Gamma^{(t)}$, the last inequality is due to Lemma C.4.

According to Lemma C.20, we know that $T_r\cdot\eta = \widetilde{\Theta}(d^{-\frac{3}{4}}), r\in\{\pm1\}$ and $T_0\cdot\eta = \max_{r\in\{\pm1\}}\{T_r\cdot\eta\} = \widetilde{\Theta}(d^{-\frac{3}{4}})$. Note that $\mathbf{w}_j^{(0)}\sim\mathcal{N}(\mathbf{0},\sigma_0^2\mathbf{I}_d), \sigma_0 = \Theta(d^{-\frac{3}{4}})$ and hence $\|\mathbf{w}_j^{(0)}\|_2 = \widetilde{\Theta}(d^{-\frac{1}{4}})$, we know that

$$
\|\mathbf{w}_j^{(T_0)}\|_2 \le \|\mathbf{w}_j^{(0)}\|_2 + \eta\cdot T_0\cdot\widetilde{\Theta}(d^{-\frac{1}{4}}) = \widetilde{\Theta}(d^{-\frac{1}{4}}) + \widetilde{\Theta}(d^{-\frac{1}{4}}) = \widetilde{\Theta}(d^{-\frac{1}{4}}).
$$

Therefore, for any $(\mathbf{x}, y)$ sampled from distribution $\mathcal{D}$ where $\mathbf{x} = [y\cdot\mathbf{v}^\top,\boldsymbol{\xi}^\top]^\top$ and $\boldsymbol{\xi}\sim\mathcal{N}(0,\sigma_p^2)$, we have

$$
\langle\mathbf{w}_j^{(T_0)},\boldsymbol{\xi}\rangle \sim \mathcal{N}(0,\sigma_p^2\|\mathbf{w}_j^{(T_0)}\|_2^2), |\langle\mathbf{w}_j^{(T_0)},\boldsymbol{\xi}\rangle| = \Theta(\sigma_p\|\mathbf{w}_j^{(T_0)}\|_2) = \widetilde{O}(d^{-\frac{1}{4}+\epsilon}). \tag{C.45}
$$

And this indicates that $\langle\mathbf{w}_j^{(T_0)},\boldsymbol{\xi}\rangle$ will still be dominated by $\langle\mathbf{w}_j^{(T_0)},\mathbf{v}\rangle$, therefore it holds for newly sampled $(\mathbf{x}, y)$ that

$$
y\cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x}) \in \big[\big(1-o(1)\big)\cdot(\widehat{\Lambda}_{y_i}^{(T_0)})^q, \big(m+o(1)\big)\cdot(\widehat{\Lambda}_{y_i}^{(T_0)})^q\big],
$$

which means that

$$
\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y\cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x})\le 0] = o(1).
$$

This verifies the third statement that test error is nearly zero.

For the second statement, note that

$$
\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D},\widehat{y}\sim y\cdot\mathcal{B}(p)}[\widehat{y}\cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x})\le 0] \\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[\widehat{y}\cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x})\le 0|\widehat{y}=y]\cdot\mathbb{P}_{\widehat{y}\sim y\cdot\mathcal{B}(p)}(\widehat{y}=y) \\
&\quad + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[\widehat{y}\cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x})\le 0|\widehat{y}=-y]\cdot\mathbb{P}_{\widehat{y}\sim y\cdot\mathcal{B}(p)}(\widehat{y}=-y) \\
&= p\cdot\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y\cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x})\le 0] + (1-p)\cdot\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y\cdot f_{\mathbf{W}^{(T_0)}}(\mathbf{x})\ge 0] \\
&= p\cdot o(1) + (1-p)\cdot(1-o(1)) \\
&= 1-p\pm o(1),
\end{aligned}
$$

which verifies the second statement. □

C.9 Downstream task

For downstream tasks, we use early stopped classifiers, which are stopped when on-diagonal feature $\widehat{\Lambda}_r^{(t)}$ are learned while off-diagonal feature $\bar{\Lambda}_r^{(t)}$ and noise $\Gamma^{(t)}$ are not memorized. Assume we have learned K early stopped classifiers $f_{\mathbf{W}_1^{(T_0^1)}}(\mathbf{x}),\cdots,f_{\mathbf{W}_K^{(T_0^K)}}(\mathbf{x})$ by using n_u pseudo-labeled data generated by pseudo-labeler $f_1^\mathrm{w},\cdots,f_K^\mathrm{w}$ and n_l labeled data.

Then, we want to design a classifier on the learned representation $f_{\mathbf{W}_1^{(T_0^1)}}(\mathbf{x}), \cdots, f_{\mathbf{W}_K^{(T_0^K)}}(\mathbf{x})$ to fit y. Here we consider training a downstream linear model

$$g_{\mathbf{a}}(\mathbf{x}) = \sum_{k=1}^{K} a_k f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}),$$

where $a_k \in \mathbb{R}$ denotes the weight as the k-th pre-trained model. Given labeled training data $S' = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{n_l}$, we want to optimize the empirical loss function

$$L_{S'}(\mathbf{a}) = \frac{1}{n_l}\sum_{i=1}^{n_l} \ell\big(y'_i \cdot g_{\mathbf{a}}(\mathbf{x}'_i)\big),$$

where $\ell(z) = \log(1+\exp(-z))$ denotes the cross entropy loss. We initialize $\mathbf{a}$ as zero and optimize empirical loss function by gradient descent, i.e.

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \eta \cdot \nabla_{\mathbf{a}} L_{S'}(\mathbf{a}^{(t)}), \mathbf{a}^{(0)} = \mathbf{0}.$$

In order to estimate the training error and test error for downstream task, we first introduce following lemma about the increasing rate of $\|\mathbf{a}^{(t)}\|_1$.

Lemma C.23 (Logarithmic increasing rate)**.** For any learning rate $\eta > 0$, $a_k^{(t)}$ will always increase for any $k \in [K]$ and hence $\|\mathbf{a}^{(t)}\|_1 = \sum_{k=1}^K a_k^{(t)}$. And it holds that $\|\mathbf{a}^{(t)}\|_1 = \Theta(\log(t))$.

In order to give the increasing rate of $\|\mathbf{a}^{(t)}\|_1$, we introduce and prove the following lemma:

Lemma C.24. Consider following sequence $\{x_t\}_{t=1}^{\infty}$ with

$$x_{t+1} = x_t + C \cdot a^{-x_t}, x_0 = 0,$$

where $a > 1$ and $C > 0$ are constants, and it follows that

$$\log_a\big(\ln a \cdot C \cdot t + 1\big) \le x_t \le \log_a\big(\ln a \cdot C \cdot t + 1\big) + C,$$

and

$$x_{t+1} - x_t \le \frac{C}{C \cdot \ln a \cdot t + 1}.$$

Proof of Lemma C.24. Note that

$$x_{i+1} - x_i = C \cdot a^{-x_i} \iff a^{x_i}(x_{i+1} - x_i) = C,$$

by adding up above equation from $i = 0$ to $i = t-1$, we have

$$\begin{aligned} &\sum_{i=0}^{t-1} a^{x_i}(x_{i+1} - x_i) = C \cdot t \\ \Longrightarrow &\int_{x_0}^{x_t} a^x dx \ge C \cdot t \\ \Longrightarrow &\frac{a^{x_t} - a^{x_0}}{\ln a} \ge C \cdot t \\ \Longrightarrow &a^{x_t} \ge C \cdot \ln a \cdot t + 1 \\ \Longrightarrow &\begin{cases} x_t \ge \log_a\big(C \cdot \ln a \cdot t + 1\big), \\ x_{t+1} - x_t = C \cdot a^{-x_t} \le \frac{C}{C \cdot \ln a \cdot t + 1}, \end{cases} \end{aligned} \tag{C.46}$$

where the first arrow is due to a^x is monotone increasing.

On the other hand,

$$a^{x_{i+1}} = a^{x_i + C \cdot a^{-x_i}} = a^{x_i} \cdot a^{C \cdot a^{-x_i}} \le a^{x_i} \cdot a^{C/(C \cdot \ln a \cdot i + 1)} \le a^{x_i} \cdot a^{C},$$

which implies

$$\begin{aligned}
&\sum_{i=0}^{t-1} a^{x_{i+1}} \cdot (x_{i+1} - x_t) \leq a^C \sum_{i=0}^{t-1} a^{x_i} \cdot (x_{i+1} - x_i) \\
\Longrightarrow &\sum_{i=0}^{t-1} a^{x_{i+1}} \cdot (x_{i+1} - x_i) \leq a^C \cdot Ct \\
\Longrightarrow &\int_{x_0}^{x_t} a^x dx \leq a^C \cdot Ct,
\end{aligned}$$

where the first arrow is due to (C.46) and the last arrow is due to a^x is monotone increasing. This leads to

$$\begin{aligned}
x_t &\leq \log_a \big(\ln a \cdot C \cdot a^C \cdot n + 1\big) \\
&\leq \log_a \big(\ln a \cdot C \cdot a^C \cdot n + a^C\big) \\
&= \log_a \big(\ln a \cdot C \cdot t + 1\big) + C
\end{aligned}$$

Therefore, we have

$$\log_a \big(\ln a \cdot C \cdot t + 1\big) \leq x_t \leq \log_a \big(\ln a \cdot C \cdot t + 1\big) + C,$$

and

$$x_{t+1} - x_t \leq \frac{C}{\ln a \cdot C \cdot t + 1}.$$

□

Now we are ready to prove Lemma C.23.

Proof of Lemma C.23. Note that we take downstream task linear model $g_{\mathbf{a}}(\mathbf{x})$ as

$$\begin{aligned}
g_{\mathbf{a}}(\mathbf{x}) = \sum_{k=1}^{d} a_k \Bigg\{ &\sum_{j=1}^{m} \Big[\sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, y \cdot \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, \boldsymbol{\xi}\rangle\big)\Big] \\
&- \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, y \cdot \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, \boldsymbol{\xi}\rangle\big)\Big] \Bigg\} \\
= \sum_{k=1}^{d} a_k f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}). &
\end{aligned}$$

Then, we have following update rule for model parameter $\mathbf{a}$:

$$a_k^{(t+1)} = a_k^{(t)} - \eta \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} \ell'\big(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big) \cdot y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i'),$$

where we initialize $a_k^{(0)}$ as zero for all $k \in [K]$.
Next, we prove following statement by using induction method: when $t \geq 1$,

- $a_k^{(t)}, \forall k \in [K]$ is non-negative and increasing.
- $\big\|\mathbf{a}^{(t)}\big\|_1 = \sum_{i=1}^{K} a_k^{(t)}$.
- $a_k^{(t+1)} = a_k^{(t)} + \eta \cdot \widetilde{\Theta}(1) \cdot \Big(\exp\big(- \|\mathbf{a}^{(1)}\|_1 \cdot \widetilde{\Theta}(1)\big)\Big), \forall k \in [K].$

Note that $a_k^{(0)} = 0$ for all $k \in [d]$ and therefore $g_{\mathbf{a}^{(0)}}(\mathbf{x}_i') = 0$, $\ell'\big(y_i' \cdot g_{\mathbf{a}^{(0)}}(\mathbf{x}_i')\big) = \ell'(0) = -1/2$,

$$a_k^{(1)} = a_k^{(0)} - \eta \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} \ell'\big(y_i' \cdot g_{\mathbf{a}^{(0)}}(\mathbf{x}_i')\big) \cdot y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i')$$

$$= a_k^{(0)} + \eta \cdot \frac{1}{2n_l}\sum_{i=1}^{n_l} y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i') = \eta \cdot \frac{1}{2n_l}\sum_{i=1}^{n_l} y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i') \text{ for all } k \in [K].$$

Note that the accuracy of the k-th pseudo-labeler $p_k > 1/2$, accoring to the proof of Lemma C.22, we have

$$\begin{aligned} f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i') &= \sum_{j=1}^{m}\Big[\sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, y_i' \cdot \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, \boldsymbol{\xi}_i'\rangle\big)\Big] \\ &\qquad - \sum_{j=m+1}^{2m}\Big[\sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, y_i' \cdot \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_{k,j}^{(T_0^k)}, \boldsymbol{\xi}_i'\rangle\big)\Big] \\ &= y_i' \cdot \widetilde{\Theta}\big((\widehat{\Lambda}_{y_i'}^{(T_0^k)})^q\big), \end{aligned}$$

for all $k \in [K]$. Therefore

$$a_k^{(1)} = \eta \cdot \frac{1}{2n_l}\sum_{i=1}^{n_l} y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i') \geq \frac{\eta}{2} \cdot \widetilde{\Theta}\big((\widehat{\Lambda}_{y_i'}^{(T_0^k)})^q\big) > 0, \forall k \in [K].$$

It follows that

$$\big\|\mathbf{a}^{(t)}\big\|_1 = \sum_{i=1}^{K} |a_k^{(t)}| = \sum_{i=1}^{K} a_k^{(t)}.$$

Note that

$$\begin{aligned} y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i') &= y_i' \cdot \sum_{k=1}^{K} a_k^{(1)} f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i') \\ &= \sum_{k=1}^{K} a_k^{(1)} \cdot \Big(y_i' \cdot f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i')\Big) \\ &= \sum_{k=1}^{K} a_k^{(1)} \cdot \widetilde{\Theta}\big((\widehat{\Lambda}_{y_i'}^{(T_0^k)})^q\big) \\ &= \sum_{k=1}^{K} a_k^{(1)} \cdot \widetilde{\Theta}(1) \\ &= \|\mathbf{a}^{(1)}\|_1 \cdot \widetilde{\Theta}(1). \end{aligned} \tag{C.47}$$

This leads to

$$\begin{aligned} \ell'\big(y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i')\big) &= -\frac{\exp\left(-y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i')\right)}{1 + \exp\left(-y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i')\right)} \\ &= -c \cdot \Big(\exp\big(-y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i')\big)\Big) \\ &= -c \cdot \Big(\exp\big(-\|\mathbf{a}^{(1)}\|_1 \cdot \widetilde{\Theta}(1)\big)\Big), \end{aligned}$$

where the second equality is due to $y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i') > 0$, $\exp\left(-y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i')\right) < 1$ and $c \in (1/2, 1)$; the last equality is due to (C.47). It follows that

$$\begin{aligned} a_k^{(2)} &= a_k^{(1)} - \eta \cdot \frac{1}{n_l}\sum_{i=1}^{n_l} \ell'\big(y_i' \cdot g_{\mathbf{a}^{(1)}}(\mathbf{x}_i')\big) \cdot y_i' f_{\mathbf{W}_k^{(T_0)}}(\mathbf{x}_i') \\ &= a_k^{(1)} + \eta \cdot c \cdot \widetilde{\Theta}(1) \cdot \Big(\exp\big(-\|\mathbf{a}^{(1)}\|_1 \cdot \widetilde{\Theta}(1)\big)\Big), \forall k \in [K] \end{aligned}$$

where $c \in (1/2, 1)$. By then, we have already proved the induction hypothesis of $t = 1$.

Next, assume the induction hypotheses hold for t. For $t+1$, we have

$$a_k^{(t+1)} = a_k^{(t)} - \eta \cdot \frac{1}{n_1} \sum_{i=1}^{n_1} \underbrace{\ell'\big(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big)}_{<0} \cdot \underbrace{y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i')}_{>0} > a_k^{(t)} > 0.$$

And it follows that

$$\|\mathbf{a}^{(t+1)}\|_1 = \sum_{i=1}^{K} a_k^{(t+1)} \text{ and } y_i' \cdot g_{\mathbf{a}^{(t+1)}}(\mathbf{x}_i') = \|\mathbf{a}^{(t+1)}\|_1 \cdot \widetilde{\Theta}(1), \tag{C.48}$$

leading to

$$\ell'\big(y_i' \cdot g_{\mathbf{a}^{(t+1)}}(\mathbf{x}_i')\big) = -c \cdot \Big(\exp\big(- \|\mathbf{a}^{(t+1)}\|_1 \cdot \widetilde{\Theta}(1)\big)\Big), c \in (1/2, 1),$$

and

$$a_k^{(t+2)} = a_k^{(t+1)} + \eta \cdot \widetilde{\Theta}(1) \cdot \Big(\exp\big(- \|\mathbf{a}^{(t+1)}\|_1 \cdot \widetilde{\Theta}(1)\big)\Big), \forall k \in [K].$$

This indicates that if induction hypotheses hold for t, then they holds for $t+1$.
Adding up $k \in [K]$, we can obtain

$$\|\mathbf{a}^{(t+1)}\|_1 = \big\|\mathbf{a}^{(t)}\big\|_1 + \eta \cdot \widetilde{\Theta}(1) \cdot \exp\Big(- \widetilde{\Theta}(1) \cdot \big\|\mathbf{a}^{(t)}\big\|_1\Big) \tag{C.49}$$

According to Lemma C.24, we know that $\big\|\mathbf{a}^{(t)}\big\|_1 = \log t / \widetilde{\Theta}(1)\{\pm \text{ lower order terms w.r.t. t}\}$. □

The following lemma gives the convergence guarantee of downstream task:

Lemma C.25. (Convergence Guarantee) For any learning rate $\eta > 0$,

$$\|\nabla_{\mathbf{a}} L_{S'}(\mathbf{a}^{(t)})\|_1 \le \frac{\widetilde{\Theta}(1)}{\eta \cdot \widetilde{\Theta}(1) \cdot t + 1} \text{ and } \nabla_{\mathbf{a}}^2 L_S(\mathbf{a}) \succeq 0 \text{ for any } \mathbf{a} \in \mathbb{R}^d,$$

which means within polynomial steps, gradient descent is guaranteed to find a point with small gradient.

Proof of Lemma C.25. Note that

$$\begin{aligned}
\|\nabla_{\mathbf{a}} L_{S'}(\mathbf{a}^{(t)})\|_1 &= \sum_{k=1}^{K} |\partial_{a_k} L_{S'}(\mathbf{a}^{(t)})| \\
&= -\sum_{k=1}^{K} \partial_{a_k} L_{S'}(\mathbf{a}^{(t)}) \\
&= \sum_{k=1}^{K} \frac{a_k^{(t+1)} - a_k^{(t)}}{\eta} \\
&= \frac{\|\mathbf{a}^{(t+1)}\|_1 - \|\mathbf{a}^{(t)}\|_1}{\eta},
\end{aligned}$$

then according to Lemma C.24 and (C.49), we know

$$\|\mathbf{a}^{(t+1)}\|_1 - \|\mathbf{a}^{(t)}\|_1 \le \frac{\eta \cdot \widetilde{\Theta}(1)}{\eta \cdot \widetilde{\Theta}(1) \cdot t + 1}. \tag{C.50}$$

And it follows that

$$\|\nabla_{\mathbf{a}} L_{S'}(\mathbf{a}^{(t)})\|_1 \le \frac{\widetilde{\Theta}(1)}{\eta \cdot \widetilde{\Theta}(1) \cdot t + 1},$$

which shows that within polynomial steps, gradient descent is guaranteed to find a point with small gradient.

Note that

$$\partial_{a_k} L_{S'}(\mathbf{a}) = \frac{1}{n_l}\sum_{i=1}^{n_l} \ell'\big(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big) \cdot y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i'),$$

$$\partial_{a_k}\partial_{a_j} L_{S'}(\mathbf{a}) = \frac{1}{n_l}\sum_{i=1}^{n_l} \ell''\big(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big) \cdot \Big(f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i) \cdot f_{\mathbf{W}_j^{(T_0^j)}}(\mathbf{x}_i)\Big) \text{ for all } k, j \in [K],$$

Denote $\Big[f_{\mathbf{W}_1^{(T_0^1)}}(\mathbf{x}_i'), \cdots, f_{\mathbf{W}_K^{(T_0^K)}}(\mathbf{x}_i')\Big]^\top$ as $\mathbf{f}_{\mathbf{W}^*}(\mathbf{x}_i')$, then

$$\nabla^2_{\mathbf{a}} L_S(\mathbf{a}) = \frac{1}{n_l}\sum_{i=1}^{n_l} \ell''\big(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big) \cdot \Big(\mathbf{f}_{\mathbf{W}^*}(\mathbf{x}_i') \cdot \mathbf{f}_{\mathbf{W}^*}(\mathbf{x}_i')^\top\Big).$$

Note that $\mathbf{f}_{\mathbf{W}^*}(\mathbf{x}_i') \cdot \mathbf{f}_{\mathbf{W}^*}(\mathbf{x}_i')^\top$ is a non-negative definite matrix, $\ell''(z) = \exp(-z)/\big(1 + \exp(-z)\big)^2 > 0$ and the fact that sum of non-negative definite matrices is still a non-negative definite matrix, it follows that $\nabla^2_{\mathbf{a}} L_S(\mathbf{a}) \succeq 0$. $\square$

Theorem C.26 (Restatement of Theorem 3.3)**.** Under semi-supervised learning setting, for downstream task, suppose K early stopped classifiers $\{f_{\mathbf{W}_k^*}\}_{k=1}^K$ are obtained after the pre-training of KK CNN models finished, and after $T_{\mathrm{dt}} = \Theta(d^{0.1}/\eta)$ iterations with learning rate $\eta = \Theta(1)$, then we can find a linear model $\mathbf{a}^{(T_{\mathrm{dt}})}$, which satisfies: Both test error and loss are nearly 0, i.e. $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot g_{\mathbf{a}^{(T_{\mathrm{dt}})}}(\mathbf{x}) \le 0] = o(1), L_{\mathcal{D}}(\ell(y \cdot g_{\mathbf{a}^{(T_{\mathrm{dt}})}}(\mathbf{x}))) = o(1)$.

Proof of Theorem C.26. For test error, we have

$$\begin{aligned}\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot g_{a^{(T_{\mathrm{dt}})}}(\mathbf{x}) \le 0] &= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\Bigg[\sum_{k=1}^K a_k^{(T_{\mathrm{dt}})} \cdot \big(y \cdot f_{\mathbf{W}_k^*}(\mathbf{x})\big) \le 0\Bigg] \\ &= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\Bigg[\sum_{k=1}^K a_k^{(T_{\mathrm{dt}})} \cdot \widetilde{\Theta}(1) \le 0\Bigg] = o(1)\end{aligned}$$

where the last equality is due to $a_k^{(T_{\mathrm{dt}})} > 0$ according to Lemma C.23.

For test loss, we have

$$L_{\mathcal{D}}(\ell(y \cdot g_{\mathbf{a}^{(T_{\mathrm{dt}})}}(\mathbf{x}))) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(y \cdot g_{\mathbf{a}^{(T_{\mathrm{dt}})}}(\mathbf{x}))],$$

i.e., we estimate for newly generated data $(\mathbf{x}, y)$ the magnitude of $\ell(y \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}))$. In order to do so, we will first estimate $\ell(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i))$. Then, we will show that $\ell(y \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}))$ and $\ell(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i))$ nearly equal to each other.

According to the update rule of $a_k^{(t)}$, we have

$$a_k^{(t+1)} = a_k^{(t)} - \eta \cdot \frac{1}{n_l}\sum_{i=1}^{n_l} \ell'\big(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big) \cdot y_i' f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i').$$

Adding up the above equation for $k \in [K]$, we obtain

$$\|\mathbf{a}^{(t+1)}\|_1 = \|\mathbf{a}^{(t)}\|_1 - \eta \cdot \frac{1}{n_l}\sum_{i=1}^{n_l} \ell'\big(y_i' \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big) \cdot y_i' \sum_{k=1}^K f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i').$$

And according to (C.50), we have

$$\|\mathbf{a}^{(t+1)}\|_1 - \|\mathbf{a}^{(t)}\|_1 \le \frac{\eta \cdot \widetilde{\Theta}(1)}{\eta \cdot \widetilde{\Theta}(1) \cdot t + 1},$$

therefore it follows that

$$-\frac{1}{n_l}\sum_{i=1}^{n_l}\ell'\big(y_i'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big)\cdot y_i'\sum_{k=1}^{K} f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i') \le \frac{\widetilde{\Theta}(1)}{\eta\cdot\widetilde{\Theta}(1)\cdot t+1}.$$

Note that $K=\Theta(1)$ and for all $k\in[K]$ we have $y_i'\cdot f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_i')=\widetilde{\Theta}(1)$, it follows that

$$-\frac{1}{n_l}\sum_{i=1}^{n_l}\ell'\big(y_i'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_i')\big) \le \frac{\widetilde{\Theta}(1)}{\eta\cdot\widetilde{\Theta}(1)\cdot t+1}.$$

Note that $n_l=\widetilde{\Theta}(1)$ and according to Lemma C.14, there exists a positive sample $(\mathbf{x}_{i_1},y_{i_1})$ and a negative sample $(\mathbf{x}_{i_2},y_{i_2})$ with the property that

$$-\ell'\big(y_{i_1}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_1}')\big) \le \frac{\widetilde{\Theta}(1)}{\eta\cdot\widetilde{\Theta}(1)\cdot t+1},\ -\ell'\big(y_{i_2}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_2}')\big) \le \frac{\widetilde{\Theta}(1)}{\eta\cdot\widetilde{\Theta}(1)\cdot t+1}.$$

Note that $\ell(z)=\log(1+\exp(-z))$ and $\ell'(z)=-\exp(-z)/\big(1+\exp(-z)\big)$, we know that for $z>0$,

$$\begin{aligned}-\ell'(z)&=c\cdot\exp(-z),\\ \ell(z)&<\exp(-z)=-\ell'(z)/c, c\in(1/2,1).\end{aligned}$$

It follows that

$$\ell\big(y_{i_1}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_1}')\big) \le \frac{\widetilde{\Theta}(1)}{\eta\cdot\widetilde{\Theta}(1)\cdot t+1},\ \ell\big(y_{i_2}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_2}')\big) \le \frac{\widetilde{\Theta}(1)}{\eta\cdot\widetilde{\Theta}(1)\cdot t+1}.$$

Note that $\ell(z)$ is 1-Lipschitz, we have

$$\begin{aligned}\big|\ell\big(y\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x})\big)-\ell\big(y_{i_1}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_1}')\big)\big| &\le \big|y\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x})-y_{i_1}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_1}')\big|,\\ \big|\ell\big(y\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x})\big)-\ell\big(y_{i_2}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_2}')\big)\big| &\le \big|y\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x})-y_{i_2}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_2}')\big|.\end{aligned} \tag{C.51}$$

If $y=1$, we have

$$\begin{aligned}\big|y\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x})-y_{i_1}'\cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_1}')\big| &= \big|g_{\mathbf{a}^{(t)}}(\mathbf{x})-g_{\mathbf{a}^{(t)}}(\mathbf{x}_{i_1}')\big|\\ &= \left|\sum_{k=1}^{K}a_k^{(t)}f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x})-\sum_{k=1}^{K}a_k^{(t)}f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_{i_1}')\right|\\ &= \left|\sum_{k=1}^{K}a_k^{(t)}\Big(f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x})-f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}_{i_1}')\Big)\right|,\end{aligned} \tag{C.52}$$

and

$$
\begin{aligned}
f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}) - f_{\mathbf{W}_k^{(T_0^k)}}(\mathbf{x}'_{i_1}) = &\sum_{j=1}^{m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}\rangle\big)\Big] \\
&- \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}\rangle\big)\Big] \\
&- \sum_{j=1}^{m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}'_{i_1}\rangle\big)\Big] \\
&+ \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \mathbf{v}\rangle\big) + \sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}'_{i_1}\rangle\big)\Big] \\
= &\sum_{j=1}^{m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}\rangle\big) - \sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}'_{i_1}\rangle\big)\Big] \\
&+ \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}'_{i_1}\rangle\big) - \sigma\big(\langle \mathbf{w}_j^{(T_0^k)}, \boldsymbol{\xi}\rangle\big)\Big] \\
= &\widetilde{O}(d^{-\frac{1}{4}+\epsilon}),
\end{aligned} \tag{C.53}
$$

where the last equality is due to (C.45) and Lemma C.4.
Plugging (C.53) into (C.52), we have

$$
\big|y \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}) - y'_{i_1} \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}'_{i_1})\big| = \widetilde{O}(d^{-\frac{1}{4}+\epsilon}) \cdot \|\mathbf{a}^{(t)}\|_1. \tag{C.54}
$$

If $y = -1$, we can prove in a similar way that

$$
\big|y \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}) - y'_{i_2} \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}'_{i_2})\big| = \widetilde{O}(d^{-\frac{1}{4}+\epsilon}) \cdot \|\mathbf{a}^{(t)}\|_1. \tag{C.55}
$$

Plugging (C.54) and (C.55) into (C.51), we have

$$
\ell\big(y \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x})\big) \le \max\big\{y'_{i_1} \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}'_{i_1}), y'_{i_2} \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x}'_{i_2})\big\} + \widetilde{O}(d^{-\frac{1}{4}+\epsilon}) \cdot \|\mathbf{a}^{(t)}\|_1
$$

According to Lemma C.24 and (C.49), we have $\|\mathbf{a}^{(t)}\|_1 = \log t/\widetilde{\Theta}(1)\{\pm \text{ lower order terms w.r.t. } t\}$, therefore

$$
\ell\big(y \cdot g_{\mathbf{a}^{(t)}}(\mathbf{x})\big) \le \frac{\widetilde{\Theta}(1)}{\eta \cdot \widetilde{\Theta}(1) \cdot t + 1} + \widetilde{O}(d^{-\frac{1}{4}+\epsilon}) \cdot \log t \,\{\pm \text{ lower order terms w.r.t. } t\}
$$

Taking $\eta = \Theta(1)$ and $T_{dt} = \Theta(d^{\alpha}/\eta)$ where $\alpha > 0$ is a sufficiently small constant, we know that

$$
\begin{aligned}
&L_{\mathcal{D}}(\ell(y \cdot g_{\mathbf{a}^{(T_{\mathrm{dt}})}}(\mathbf{x}))) \\
=&\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(y \cdot g_{\mathbf{a}^{(T_{\mathrm{dt}})}}(\mathbf{x}))] \\
\le&\frac{\widetilde{\Theta}(1)}{\eta \cdot \widetilde{\Theta}(1) \cdot T_{dt} + 1} + \widetilde{O}(d^{-\frac{1}{4}+\epsilon}) \cdot \log T_{dt} \,\{\pm \text{ lower order terms w.r.t. } T_{dt}\} + o(1) \\
=&o(1),
\end{aligned}
$$

which completes the proof. □

D Proof of supervised learning setting

Here we prove Theorem 3.4. First, we give following lemma to facilitate the proof.

Lemma D.1 (Gradient Calculation)**.** The gradient of loss function $L_S(\mathbf{W})$ with respect to weight parameter $\mathbf{w}_j$ is

$$\nabla_{\mathbf{w}_j} L_{S'}(\mathbf{W}) = -\frac{qu_j}{n_{\mathrm{l}}} \cdot \sum_{i=1}^{n_{\mathrm{l}}} b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i' \big),$$

where $u_j := \big(\mathbb{1}_{[1 \leq j \leq m]} - \mathbb{1}_{[m+1 \leq j \leq 2m]} \big)$ and $-\ell'\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) = \exp[-y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')]/(1 + \exp[-y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')])$ is denoted as b_i.

Proof of Lemma D.1. When $1 \leq j \leq m$,

$$\begin{aligned} \nabla_{\mathbf{w}_j} \ell\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) &= \ell'\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) \cdot y_i' \cdot \nabla_{\mathbf{w}_j} f_{\mathbf{W}}(\mathbf{x}_i') \\ &= -b_i \cdot y_i' \cdot \nabla_{\mathbf{w}_j} f_{\mathbf{W}}(\mathbf{x}_i') \\ &= -b_i y_i' \cdot \big(\sigma'\big(\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle\big) \cdot y_i' \cdot \mathbf{v} + \sigma'\big(\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle\big) \cdot \boldsymbol{\xi}_i' \big) \\ &= -q b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i' \big) \end{aligned}$$

and when $m + 1 \leq j \leq 2m$,

$$\nabla_{\mathbf{w}_j} \ell\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) = q b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i' \big)$$

Combining above two cases, we have

$$\begin{aligned} \nabla_{\mathbf{w}_j} \ell\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) &= -q\big(\mathbb{1}_{[1 \leq j \leq m]} - \mathbb{1}_{[m+1 \leq j \leq 2m]} \big) b_i y_i' \big([\langle \mathbf{w}_j, y_i \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i' \big) \\ &= -q u_j b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i' \big) \end{aligned}$$

and therefore

$$\begin{aligned} \nabla_{\mathbf{w}_j} L_{S'}(\mathbf{W}) &= \frac{1}{n_{\mathrm{l}}} \sum_{i=1}^{n_{\mathrm{l}}} \nabla_{\mathbf{w}_j} L_i(\mathbf{W}) = \frac{1}{n_{\mathrm{l}}} \sum_{i=1}^{n_{\mathrm{l}}} \nabla_{\mathbf{w}_j} \ell\big(y_i' \cdot f_{\mathbf{W}}(\mathbf{x}_i')\big) \\ &= -\frac{qu_j}{n_{\mathrm{l}}} \sum_{i=1}^{n_{\mathrm{l}}} b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v} \rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i' \rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i' \big). \end{aligned}$$

□

Proof of Theorem 3.4. Recall the definition of $f_{\mathbf{W}}$ in (2.1) that

$$f_{\mathbf{W}}(\mathbf{x}) = \sum_{j=1}^{m} \Big[\sigma\big(\langle \mathbf{w}_j, y \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle \mathbf{w}_j, \boldsymbol{\xi} \rangle\big) \Big] - \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \mathbf{w}_j, y \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle \mathbf{w}_j, \boldsymbol{\xi} \rangle\big) \Big].$$

Define $\widetilde{\mathbf{w}}_j := m^{1/q} \cdot \mathbf{w}_j$, we have

$$\begin{aligned} f_{\mathbf{W}}(\mathbf{x}) = & \sum_{j=1}^{m} \Big[\sigma\big(\langle m^{-1/q} \cdot \widetilde{\mathbf{w}}_j, y \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle m^{-1/q} \cdot \widetilde{\mathbf{w}}_j, \boldsymbol{\xi} \rangle\big) \Big] \\ & - \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle m^{-1/q} \cdot \widetilde{\mathbf{w}}_j, y \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle m^{-1/q} \cdot \widetilde{\mathbf{w}}_j, \boldsymbol{\xi} \rangle\big) \Big] \\ = & \frac{1}{m} \sum_{j=1}^{m} \Big[\sigma\big(\langle \widetilde{\mathbf{w}}_j, y \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle \widetilde{\mathbf{w}}_j, \boldsymbol{\xi} \rangle\big) \Big] - \frac{1}{m} \sum_{j=m+1}^{2m} \Big[\sigma\big(\langle \widetilde{\mathbf{w}}_j, y \cdot \mathbf{v} \rangle\big) + \sigma\big(\langle \widetilde{\mathbf{w}}_j, \boldsymbol{\xi} \rangle\big) \Big] \\ := & f_{\widetilde{\mathbf{W}}}(\mathbf{x}). \end{aligned}$$

Since the standard deviation of Gaussian initialization of $\mathbf{w}_j$ is σ_0 and note that $\widetilde{\mathbf{w}}_j := m^{1/q} \cdot \mathbf{w}_j$, the standard deviation of Gaussian initialization of $\widetilde{\mathbf{w}}_j$ is $m^{1/q}\sigma_0 := \widetilde{\sigma}_0$.

On the other hand, note that the update rule of $\mathbf{w}_j^{(t)}$ is $\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} - \eta \cdot \nabla_{\mathbf{w}_j} L_{S'}(\mathbf{W}^{(t)})$, and in Lemma D.1, we have

$$\nabla_{\mathbf{w}_j} L_{S'}(\mathbf{W}) = -\frac{qu_j}{n_l} \cdot \sum_{i=1}^{n_l} b_i y_i' \big([\langle \mathbf{w}_j, y_i' \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j, \boldsymbol{\xi}_i'\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i'\big).$$

It follows that

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} + \frac{q\eta u_j}{n_l} \cdot \sum_{i=1}^{n_l} b_i^{(t)} y_i' \big([\langle \mathbf{w}_j^{(t)}, y_i' \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \mathbf{w}_j^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i'\big). \tag{D.1}$$

By plugging $\mathbf{w}_j = m^{-1/q} \cdot \widetilde{\mathbf{w}}_j$ into (D.1), we have

$$\widetilde{\mathbf{w}}_j^{(t+1)} = \widetilde{\mathbf{w}}_j^{(t)} + \frac{q\eta m^{-\frac{1}{q}} u_j}{n_l} \cdot \sum_{i=1}^{n_l} b_i^{(t)} y_i' \big([\langle \widetilde{\mathbf{w}}_j^{(t)}, y_i' \cdot \mathbf{v}\rangle]_+^{q-1} \cdot y_i' \cdot \mathbf{v} + [\langle \widetilde{\mathbf{w}}_j^{(t)}, \boldsymbol{\xi}_i'\rangle]_+^{q-1} \cdot \boldsymbol{\xi}_i'\big)$$

Assume $\widetilde{\eta} = m^{-\frac{1}{q}}\eta$, we have $\widetilde{\mathbf{w}}_j^{(t+1)} = \widetilde{\mathbf{w}}_j^{(t)} - \widetilde{\eta} \cdot \nabla_{\widetilde{\mathbf{w}}_j} L_{S'}(\widetilde{\mathbf{W}}^{(t)})$. Therefore, our data model and training algorithm is equivalent to the model and algorithm below:

$$\begin{aligned}
f_{\widetilde{\mathbf{W}}^{+1}}(\mathbf{x}) &= \frac{1}{m}\sum_{j=1}^{m}\Big[\sigma\big(\langle \widetilde{\mathbf{w}}_j, y \cdot \mathbf{v}\rangle\big) + \sigma\big(\langle \widetilde{\mathbf{w}}_j, \boldsymbol{\xi}\rangle\big)\Big],\\
f_{\widetilde{\mathbf{W}}^{-1}}(\mathbf{x}) &= \frac{1}{m}\sum_{j=m+1}^{2m}\Big[\sigma\big(\langle \widetilde{\mathbf{w}}_j, y \cdot \mathbf{v}\rangle\big) + \sigma\big(\langle \widetilde{\mathbf{w}}_j, \boldsymbol{\xi}\rangle\big)\Big],\\
f_{\widetilde{\mathbf{W}}}(\mathbf{x}) &= f_{\widetilde{\mathbf{W}}^{+1}}(\mathbf{x}) - f_{\widetilde{\mathbf{W}}^{-1}}(\mathbf{x}),
\end{aligned}$$

and we use gradient decent with learning rate $\widetilde{\eta}$ and cross-entropy loss to optimize such a data model, i.e.

$$\widetilde{\mathbf{w}}_0^{(t)} \sim \mathcal{N}(\mathbf{0}, \widetilde{\sigma}_0^2 \boldsymbol{I}_d), \widetilde{\mathbf{w}}_j^{(t+1)} = \widetilde{\mathbf{w}}_j^{(t)} - \widetilde{\eta} \cdot \nabla_{\widetilde{\mathbf{w}}_j} L_{S'}(\widetilde{\mathbf{W}}^{(t)}),\ L_{S'}(\widetilde{\mathbf{W}}^{(t)}) = \sum_{i=1}^{n_l} \ell(y_i' \cdot f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i')),$$

where $\ell(z) = \log(1 + \exp(-z))$, $\widetilde{\sigma}_0 = m^{1/q}\sigma_0$. Note that the new model meets the one used in Cao et al. (2022). To leverage their result, we introduce condition 4.3 from Cao et al. (2022) and verify that the new model meets the new condition.

Condition D.2 (Condition 4.2 in Cao et al. (2022))**.** Dimension d is sufficiently large that $d = \widetilde{\Omega}(m^{2\vee[4/(q-2)]} n^{4\vee[(2q-2)/(q-2)]})$. Training sample size n and neural network width m satisfy $n, m = \Omega(\mathrm{polylog}(d))$. Learning rate η satisfies $\eta \le \widetilde{O}(\min\{\|\mathbf{v}\|_2^{-2}, \sigma_p^{-2} d^{-1}\})$. The standard deviation of Gaussian initialization σ_0 is approximately chosen such that $\widetilde{O}(nd^{-\frac{1}{2}}) \cdot \min\{(\sigma_p\sqrt{d})^{-1}, \|\mathbf{v}\|_2^{-1}\} \le \sigma_0 \le \widetilde{O}(m^{-2/(q-2)} n^{-[1/(q-2)]\vee 1}) \cdot \min\{(\sigma_p\sqrt{d})^{-1}, \|\mathbf{v}\|_2^{-1}\}$.

Theorem D.3 (Theorem 4.4 in Cao et al. (2022))**.** For any $\epsilon > 0$, let $T = \widetilde{\Theta}(\eta^{-1} m \cdot n(\sigma_p\sqrt{d})^{-q} \cdot \sigma_0^{-(q-2)} + \eta^{-1}\epsilon^{-1} n m^3 d^{-1} \sigma_p^{-2})$. Under Condition D.2, if $n^{-1} \cdot \mathrm{SNR}^{-q} = \widetilde{\Omega}(1), \mathrm{SNR} = \|\mathbf{v}\|_2/\sigma_p\sqrt{d}$, then with probability at least $1 - d^{-1}$, there exists $0 \le t \le T$ such that:

1. The training loss converges to δ, i.e., $L_S(\mathbf{W}^{(t)}) \le \delta$.
2. The trained CNN has a constant order test loss: $L_{\mathcal{D}}(\mathbf{W}^{(t)}) = \Theta(1)$.

Note that in our setting, $m = \Theta(\mathrm{polylog}(d))$, $n_l = \widetilde{\Theta}(1)$, $\|\mathbf{v}\|_2 = \Theta(d^{\frac{1}{2}})$, $\widetilde{\sigma}_0 = m^{1/q}\sigma_0$, $\sigma_0 = \Theta(d^{-\frac{3}{4}})$ $\sigma_p = \Theta(d^{0.01})$, $\widetilde{\eta} = m^{-\frac{1}{q}}\eta$ and $\eta = O(d^{-1-2\epsilon})$, it's not difficult to verify that Condition D.2 holds. Besides, $\mathrm{SNR} = d^{-0.01}, n^{-1} \cdot \mathrm{SNR}^{-q} = \widetilde{\Theta}(d^{q\epsilon}) = \widetilde{\Omega}(1)$. Therefore, the conclusion of Theorem D.3 holds for

$$\begin{aligned}
T &= \widetilde{\Theta}(\widetilde{\eta}^{-1} m \cdot n(\sigma_p\sqrt{d})^{-q} \cdot \sigma_0^{-(q-2)} + \widetilde{\eta}^{-1}\epsilon^{-1} n m^3 d^{-1}\sigma_p^{-2})\\
&= \widetilde{\Theta}(\widetilde{\eta}^{-1} \cdot (d^{1/2+\epsilon})^{-q} \cdot (d^{-3/4})^{-(q-2)} + \widetilde{\eta}^{-1}\epsilon^{-1} d^{-1} d^{-2\epsilon})
\end{aligned}$$

$$
\begin{aligned}
&= \widetilde{\Theta}(\widetilde{\eta}^{-1} \cdot d^{(1/4-\epsilon)q-3/2} + \widetilde{\eta}^{-1}\epsilon^{-1}d^{-1-2\epsilon}) \\
&= \widetilde{\Theta}(\eta^{-1} \cdot d^{(1/4-\epsilon)q-3/2}).
\end{aligned}
$$

$\square$

E Auxiliary Lemmas

For the estimation of $\bar{\Lambda}^{(0)}$ and $\widehat{\Lambda}^{(0)}$, we introduce the following lemma.

Lemma E.1 (Borell-TIS inequality)**.** Let X be a centered Gaussian on $\mathbb{R}^m$ and set $\sigma_X^2 := \max_{i\in[m]} \mathbb{E}(X_i^2)$. Then for each $t > 0$,

$$
\mathbb{P}\Big(\Big|\max_{i\in[m]} X_i - \mathbb{E}\big(\max_{i\in[m]} X_i\big)\Big| > t\Big) \leq 2e^{-\frac{t^2}{2\sigma_X^2}}.
$$

For the expectation of $\widehat{\Lambda}_r^{(0)}$ and $\bar{\Lambda}_r^{(0)}$, we give the following lemma.

Lemma E.2. Let $Y = \max_{1\leq i\leq m} X_i$, where $X_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. random variables. Then

$$
\frac{1}{\sqrt{\pi \log 2}}\sigma\sqrt{\log m} \leq \mathbb{E}[Y] \leq \sqrt{2}\sigma\sqrt{\log m}.
$$

For the estimation of $\|\boldsymbol{\xi}_i\|_2^2$ and $\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l\rangle$, we introduce following lemma.

Lemma E.3 (Lemma B.2 in Cao et al. (2022))**.** Suppose that $\delta > 0$ and $d = \Omega(\log(4n/\delta))$. Then with probability at least $1-\delta$,

$$
\begin{aligned}
&\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}_i\|_2^2 \leq 3\sigma_p^2 d/2, \\
&|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_l\rangle| \leq 2\sigma_p^2 \cdot \sqrt{d\log(4n^2/\delta)},
\end{aligned}
$$

for all $i, l \in [n]$, $i \neq l$.

Besides, we introduce the following lemma about the tensor power method.

Lemma E.4. Consider an increasing sequence $x_t \geq 0$ defined as $x_{t+1} = x_t + \eta \cdot C_t x_t^{q-1}$, and $C_1 \leq C_t \leq C_2$ for all $t > 0$, then we have for $A > x_0$, every $\delta > 0$, and every $\eta > 0$:

$$
\sum_{t\geq 0, x_t \leq A} \eta \leq \frac{\delta}{(1-(1+\delta)^{-(q-2)})x_0^{q-2}C_1} + \eta \cdot \frac{C_2}{C_1}(1+\delta)^{q-1}\bigg(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\bigg),
$$

$$
\sum_{t\geq 0, x_t \leq A} \eta \geq \frac{\delta\big(1-(x_0/A)^{q-2}\big)}{(1+\delta)^{q-1}\big(1-(1+\delta)^{-(q-2)}\big)x_0^{q-2}C_2} - \eta \cdot (1+\delta)^{-(q-1)}\bigg(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\bigg).
$$

Proof of Lemma E.4. For every $g = 0, 1, 2, \cdots$, let τ_g be the first iteration such that $x_t \geq (1+\delta)^g x_0$. Let b be the smallest integer such that $(1+\delta)^b x_0 \geq A$. By the definition of τ_g, we have $x_t \in [(1+\delta)^g x_0, (1+\delta)^{g+1}x_0)$ for all $t \in [\tau_g, \tau_{g+1})$ and $x_{\tau_{g+1}} \geq (1+\delta)^{g+1}x_0$, $x_{\tau_g - 1} < (1+\delta)^g x_0$, leading to

$$
\begin{aligned}
\sum_{t\in[\tau_g,\tau_{g+1})} \eta \cdot C_t[(1+\delta)^g x_0]^{q-1} \leq x_{\tau_{g+1}} - x_{\tau_g} &= \sum_{t\in[\tau_g,\tau_{g+1})} (x_{t+1} - x_t) \\
&= \sum_{t\in[\tau_g,\tau_{g+1})} \eta \cdot C_t x_t^{q-1} \leq \sum_{t\in[\tau_g,\tau_{g+1})} \eta \cdot C_t[(1+\delta)^{g+1}x_0]^{q-1},
\end{aligned}
$$

following lower bound for $x_{\tau_{g+1}} - x_{\tau_g}$:

$$
\begin{aligned}
x_{\tau_{g+1}} - x_{\tau_g} &= x_{\tau_{g+1}} - x_{\tau_g - 1} - \eta \cdot C_{\tau_g - 1} x_{\tau_g - 1}^{q-1} \\
&\geq (1+\delta)^{g+1}x_0 - (1+\delta)^g x_0 - \eta \cdot C_{\tau_g - 1}[(1+\delta)^g x_0]^{q-1} \\
&= \delta(1+\delta)^g x_0 - \eta \cdot C_{\tau_g - 1}(1+\delta)^{(q-1)g}x_0^{q-1},
\end{aligned}
$$

and following upper bound for $x_{\tau_{g+1}} - x_{\tau_g}$:

$$\begin{aligned} x_{\tau_{g+1}} - x_{\tau_g} &= x_{\tau_{g+1}-1} + \eta \cdot C_{\tau_{g+1}-1} x_{\tau_{g+1}-1}^{q-1} - x_{\tau_g} \\ &\le (1+\delta)^{g+1} x_0 + \eta \cdot C_{\tau_{g+1}-1} [(1+\delta)^{(g+1)} x_0]^{q-1} - (1+\delta)^g x_0 \\ &= \delta (1+\delta)^g x_0 + \eta \cdot C_{\tau_{g+1}-1} (1+\delta)^{(q-1)(g+1)} x_0^{q-1}. \end{aligned}$$

Therefore,

$$\sum_{t \in [\tau_g, \tau_{g+1})} \eta \cdot C_t [(1+\delta)^g x_0]^{q-1} \le \delta (1+\delta)^g x_0 + \eta \cdot C_{\tau_{g+1}-1} (1+\delta)^{(q-1)(g+1)} x_0^{q-1},$$

$$\sum_{t \in [\tau_g, \tau_{g+1})} \eta \cdot C_t [(1+\delta)^{g+1} x_0]^{q-1} \ge \delta (1+\delta)^g x_0 - \eta \cdot C_{\tau_g - 1} (1+\delta)^{(q-1)g} x_0^{q-1}.$$

These imply that

$$\sum_{t \in [\tau_g, \tau_{g+1})} \eta \cdot C_t \le \frac{\delta}{(1+\delta)^{(q-2)g} x_0^{q-2}} + \eta \cdot C_{\tau_{g+1}-1} (1+\delta)^{q-1} \le \frac{\delta}{(1+\delta)^{(q-2)g} x_0^{q-2}} + \eta \cdot C_2 (1+\delta)^{q-1},$$

$$\begin{aligned} \sum_{t \in [\tau_g, \tau_{g+1})} \eta \cdot C_t &\ge \frac{\delta}{(1+\delta)^{(q-2)g + (q-1)} x_0^{q-2}} - \eta \cdot C_{\tau_g - 1} (1+\delta)^{-(q-1)} \\ &\ge \frac{\delta}{(1+\delta)^{(q-2)g + (q-1)} x_0^{q-2}} - \eta \cdot C_2 (1+\delta)^{-(q-1)}. \end{aligned}$$

Recall b is the smallest integer such that $(1+\delta)^b x_0 \ge A$, so we can calculate that

$$\begin{aligned} \sum_{t \ge 0, x_t \le A} \eta \cdot C_t &\le \sum_{g=0}^{b-1} \frac{\delta}{(1+\delta)^{(q-2)g} x_0^{q-2}} + \eta \cdot C_2 (1+\delta)^{q-1} b \\ &= \frac{\delta\big(1 - (1+\delta)^{-(q-2)b}\big)}{\big(1 - (1+\delta)^{-(q-2)}\big) x_0^{q-2}} + \eta \cdot C_2 (1+\delta)^{q-1} b \\ &\le \frac{\delta}{(1 - (1+\delta)^{-(q-2)}) x_0^{q-2}} + \eta \cdot C_2 (1+\delta)^{q-1} b, \end{aligned}$$

and

$$\begin{aligned} \sum_{t \ge 0, x_t \le A} \eta \cdot C_t &\ge \sum_{g=0}^{b-1} \frac{\delta}{(1+\delta)^{(q-2)g + (q-1)} x_0^{q-2}} - \eta \cdot C_2 (1+\delta)^{-(q-1)} b \\ &= \frac{\delta\big(1 - (1+\delta)^{-(q-2)b}\big)}{(1+\delta)^{q-1}\big(1 - (1+\delta)^{-(q-2)}\big) x_0^{q-2}} - \eta \cdot C_2 (1+\delta)^{-(q-1)} b \\ &\ge \frac{\delta\big(1 - (x_0/A)^{q-2}\big)}{(1+\delta)^{q-1}\big(1 - (1+\delta)^{-(q-2)}\big) x_0^{q-2}} - \eta \cdot C_2 (1+\delta)^{-(q-1)} b, \end{aligned}$$

where the last inequality is due to $(1+\delta)^b x_0 \ge A$.

Note that $(1+\delta)^{b-1} x_0 < A$, i.e. $b \le 1 + \frac{\log(A/x_0)}{\log(1+\delta)}$, therefore

$$\sum_{t \ge 0, x_t \le A} \eta \cdot C_t \le \frac{\delta}{(1 - (1+\delta)^{-(q-2)}) x_0^{q-2}} + \eta \cdot C_2 (1+\delta)^{q-1} \left(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\right),$$

$$\sum_{t \ge 0, x_t \le A} \eta \cdot C_t \ge \frac{\delta\big(1 - x_0/A\big)}{(1+\delta)^{q-1}\big(1 - (1+\delta)^{-(q-2)}\big) x_0^{q-2}} - \eta \cdot C_2 (1+\delta)^{-(q-1)} \left(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\right),$$

Note that $C_1 \leq C_t \leq C_2$, we have

$$\sum_{t \geq 0, x_t \leq A} \eta \leq \frac{\delta}{(1-(1+\delta)^{-(q-2)})x_0^{q-2}C_1} + \eta \cdot \frac{C_2}{C_1}(1+\delta)^{q-1}\left(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\right),$$

$$\sum_{t \geq 0, x_t \leq A} \eta \geq \frac{\delta\big(1-(x_0/A)^{q-2}\big)}{(1+\delta)^{q-1}\big(1-(1+\delta)^{-(q-2)}\big)x_0^{q-2}C_2} - \eta \cdot (1+\delta)^{-(q-1)}\left(1 + \frac{\log(A/x_0)}{\log(1+\delta)}\right).$$

$\square$