Gradient Descent Robustly Learns the Intrinsic Dimension of Data in Training Convolutional Neural Networks

Chenyang Zhang CHYZHANG@CONNECT.HKU.HK
Department of Statistics & Actuarial Science, The University of Hong Kong

Peifeng Gao GPF17860971528@163.COM
Department of Computer Science, The University of Hong Kong

Difan Zou DZOU@CS.HKU.HK
Department of Computer Science & Institute of Data Science, The University of Hong Kong

Yuan Cao YUANCAO@HKU.HK
Department of Statistics & Actuarial Science and Department of Mathematics, The University of Hong Kong

Abstract

Modern neural networks are usually highly over-parameterized. Behind the wide usage of over-parameterized networks is the belief that, if the data are simple, then the trained network will be automatically equivalent to a simple predictor. Following this intuition, many existing works have studied different notions of "ranks" of neural networks and their relation to the rank of data. In this work, we study the rank of convolutional neural networks (CNNs) trained by gradient descent, with a specific focus on the robustness of the rank to noises in data. Specifically, we point out that, when adding noises to data inputs, the rank of the CNN trained with gradient descent is affected far less compared with the rank of the data, and even when a significant amount of noises have been added, the CNN filters can still effectively recover the intrinsic dimension of the clean data. We back up our claim with a theoretical case study, where we consider data points consisting of "signals" and "noises" and we rigorously prove that CNNs trained by gradient descent can learn the intrinsic dimension of the data signals.

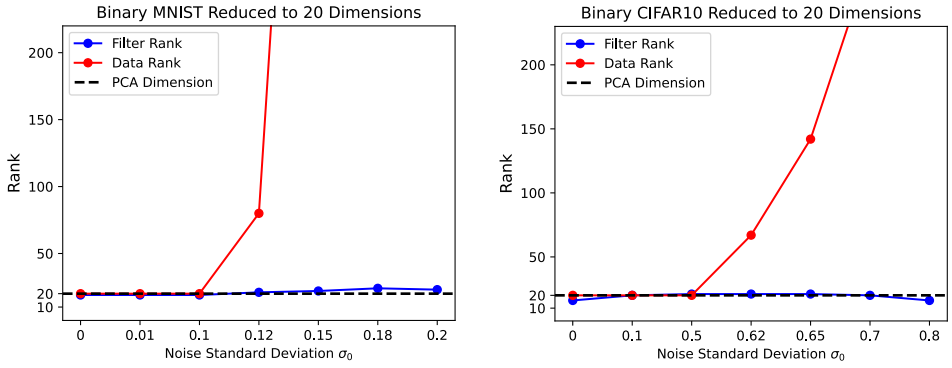
1. Introduction

Neural networks have become a cornerstone in modern machine learning, demonstrating remarkable performance across various domains. A common characteristic of modern networks is their tendency to be highly over-parameterized. Interestingly, it has been demonstrated that over-parameterized models trained by standard optimization algorithms exhibit a preference for simplicity [3, 11, 15, 16, 21, 26, 33–36]: if the training data can be fitted well by a simple predictor, then after training, an over-parameterized model may effectively reduce to this simple predictor.

A notable line of recent works have considered notions of "ranks" to characterize how simple the over-parameterized neural network after training is [3, 11, 15, 18, 26, 41]. Specifically for nonlinear networks, [11] showed that the effective hidden layer neurons in a two-layer neural network is sparse. [41] empirically demonstrated that the hidden neural weight vectors condense on isolated orientations when learning easy tasks, and provided explanations of this phenomenon with theoretical case studies. [18] further formulated the Jacobian and Bottleneck ranks for vector-valued neural networks, and demonstrated that over-parameterized networks tend to achieve small ranks.

In this work, we aim to study the "ranks" of two-layer convolutional neural networks (CNN) when learning from low-rank data sets from a new perspective: we examine the robustness of the neural network rank when noises of increasing levels are added to the low-rank data. Interestingly, we can draw the following conclusion:

The rank of the CNN is much more robust to noises, compared with the rank of the data.



(*a*) MNIST (*b*) CIFAR-10

Figure 1: Ranks of data and filters under different noise levels. In (a), we perform a principal component analysis (PCA) to a subset of MNIST images to reduce the intrinsic dimension of each image to 20. We then add noises around the obtained low-rank image, and train a two-layer CMM until convergence. We then calculate the ranks of the noisy images and the matrix consisting of all the convolutional filters of the CNN. When calculating ranks, eigenvalues smaller than $1/100$ of the largest eigenvalue are ignored. The curves of filter rank and data rank with respect to the noise level are plotted. In (b), we conduct a similar set of experiments on the CIFAR-10 data set.

An illustration of this claim is given in Figure 1 (A more comprehensive set of experiments are presented in Appendix B). While the empirical observation is clear, the explanation of this phenomenon requires more careful analysis. In order to theoretically understand this phenomenon, we consider a specific type of learning problems which have been considered in recent studies of the "benign overiftting" phenomenon [7, 23], where the data inputs consist of "signal patches" and "noise patches". Notably, this type of data model is particularly suitable for our study of the rank of neural network and its robustness with respect to noises – the signal patches can represent the clean (low-rank) data, while the noise patches naturally motivates the study of robustness. By studying this type of data, we are able to theoretically demonstrate our claim that the rank of CNN filters are robust to noises.

The major contributions of this paper are as follows:

- We reveal the "rank robustness" phenomenon in training convolutional neural networks. In particular, we add different levels of noises to low-rank data and then use CNNs to fit these noisy data. We observe that, even if a significant amount of noises have been added which causes the rank of the data to explode, the rank of the CNN filters can still remain around the rank of the clean data. This suggests that the rank of CNN is more robust to the noise compared to the rank of data.
- We theoretically prove that the observed phenomenon happens when training a two-layer CNN on a data model with multiple signals and noise. More specifically, we show that under a wide range of noise levels, the CNN model will learn the intrinsic dimension of the training data, i.e.,

the number of signal vectors. In comparison, we also show that under the same noise levels, the data rank can provably explode.

- Our theoretical analysis is inspired by [7] where the authors proposed a data model for the study the "benign overfitting" phenomenon. Compared to [7] where only one signal patch and one noise patch is considered, our analysis handles the more general model with arbitrarily many patches. More importantly, as the purpose of [7] is only to study the test loss, their analysis on the optimization process is not the most accurate. In comparison, this work establishes a more refined analysis that accurately characterizes the behaviour of all the CNN filters throughout training, which enables the study of the rank of the trained CNN. The conclusions of [7] can also be directly implied based on our theoretical results. Therefore, we believe that the theoretical tools developed in this paper may be of independent interest.

2. Problem Setup

In this section we introduce the theoretical setting considered in this paper. We first give the following definition on the distribution of data.

Definition 2.1 *Let $\mathcal{U} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K\} \subseteq \mathbb{R}^d$ be a set of K fixed vectors representing different signals. Based on these signal vectors, each data point $(\mathbf{x}, y)$ with $\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_P^\top]^\top \in \mathbb{R}^{Pd}$ denoting the P patches and $y \in \{-1, 1\}$ is generated from the following distribution $\mathcal{D}$:*

1. *The label y is generated as a Rademacher random variable.*
2. *An integer s is drawn from a distribution π over $\{1, 2, \ldots, K\}$. This distribution π takes value 1 with a positive probability π_1.*
3. *A set of s vectors $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_s$ are randomly and uniformly drawn from $\mathcal{U}$ without replacement. s patches among $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}$ are then randomly chosen and are assigned as $y \cdot \boldsymbol{\nu}_i$, $i \in [s]$.*
4. *The rest $P-s$ patches among $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}$ are assigned as Gaussian random vectors $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{P-s}$ that are independently drawn from $N(\mathbf{0}, \sigma_{\text{noise}}^2 \cdot (\mathbf{I}_d - \sum_{k=1}^K \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top \cdot \|\boldsymbol{\mu}_k\|_2^{-2}))$.*

We consider the training of CNNs based on a trained dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ for binary classification, where the training data points $(\mathbf{x}_i, y_i)$ are generated independently from the distribution given in Definition 2.1. For these training data points, we adopt the notations in Definition 2.1 and denote by s_i the number of signal vectors in $\mathbf{x}_i$. Similarly, we also denote by $\boldsymbol{\nu}_{i,1}, \ldots, \boldsymbol{\nu}_{i,s_i}$ the signal vectors contained in $\mathbf{x}_i$, and by $\boldsymbol{\xi}_{i,1}, \ldots, \boldsymbol{\xi}_{i,P-s_i}$ the noise vectors contained in $\mathbf{x}_i$. Similar training samples have been considered for a variety of different topics [2, 7, 19, 23, 29, 42].

Two-layer CNNs. We consider a two-layer convolutional neural network whose filters are applied to the P patches $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(P)}$ separately, and the second layer parameters of the network are fixed as $+1/m$ and $-1/m$ respectively. Then the network can be written as $f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$, where $F_{+1}(\mathbf{W}_{+1}, \mathbf{x})$, $F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$ are defined as:

$$F_j(\mathbf{W}_j, \mathbf{x}) = \frac{1}{m}\sum_{r=1}^{m}\sum_{p=1}^{P}\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_p\rangle) = \frac{1}{m}\sum_{r=1}^{m}\bigg[\sum_{k=1}^{s}\sigma(\langle \mathbf{w}_{j,r}, y\cdot\boldsymbol{\nu}_k\rangle) + \sum_{k'=1}^{P-s}\sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_{k'}\rangle)\bigg],$$

for $j \in \{+1, -1\}$, m is the number of convolutional filters in F_{+1} and F_{-1}. We consider using Huberized ReLU activation function $\sigma(\cdot)$ defined as $\sigma(z) = q^{-1}\kappa^{1-q}z^q \cdot \mathbb{1}_{\{z\in[0,\kappa]\}} + (z - \kappa - \kappa/q)\cdot \mathbb{1}_{\{z>\kappa\}}$, where κ is a small constant and $q \geq 3$. We use $\mathbf{w}_{j,r} \in \mathbb{R}^d$ to denote the weight for the r-th filter (i.e., neuron), and $\mathbf{W}_j$ is the collection of model weights associated with F_j. We also use $\mathbf{W}$ to denote the collection of all model weights.

Training Algorithm. We train the above CNN model by minimizing the empirical cross-entropy loss function $L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell[y_i \cdot f(\mathbf{W}, \mathbf{x}_i)]$, where $\ell(z) = \log(1+\exp(-z))$, and $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is the training data set. We consider gradient descent starting from Gaussian initialization, where each entry of $\mathbf{W}_{+1}$ and $\mathbf{W}_{-1}$ is sampled from a Gaussian distribution $N(0, \sigma_0^2)$, and σ_0^2 is the variance. The gradient descent update of the filters in the CNN can be written as

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \eta \cdot \nabla_{\mathbf{w}_{j,r}} L_S(\mathbf{W}^{(t)}) \tag{2.1}$$

for $j \in \{\pm 1\}$ and $r \in [m]$, where we introduce a shorthand notation $\ell_i'^{(t)} = \ell'[y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)]$.

3. Main Results

Before we demonstrate our results, we first present conditions on the dimension d, sample size n, neural network width m (number of filters), learning rate η, initialization scale σ_0, signal level $\|\boldsymbol{\mu}_k\|_2$, and noise level σ_{noise}.

Condition 3.1 *Suppose that*

1. *Dimension d is sufficiently large: $d = \widetilde{\Omega}(m^4 \vee n^4)$*
2. *Training sample size n and neural network width m satisfy $n, m = \Omega(\text{polylog}(d))$.*
3. *Signals are perpendicular to each other and at the same level, i.e. $\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'} \rangle = 0$ and $\frac{\|\boldsymbol{\mu}_k\|_2}{\|\boldsymbol{\mu}_{k'}\|_2} = \Theta(1)$ for all $k \neq k'$. W.L.O.G, we assume $\|\boldsymbol{\mu}_1\|_2 \geq \|\boldsymbol{\mu}_2\|_2 \geq \cdots \|\boldsymbol{\mu}_K\|_2$.*
4. *The level of signals is larger or equal to the noises: $\frac{n\|\boldsymbol{\mu}_K\|_2^q}{\sigma_{\text{noise}}^q d^{q/2}} \geq \widetilde{\Omega}(1)$*
5. *The learning rate η satisfies $\eta \leq \widetilde{O}(\min\{\|\boldsymbol{\mu}_1\|_2^{-2}, \sigma_{\text{noise}}^{-2} d^{-1}\})$.*
6. *The standard deviation of Gaussian initialization σ_0 is sufficiently small: $\sigma_0 \leq \widetilde{O}(d^{-1/2}) \cdot \min\{(\sigma_{\text{noise}}\sqrt{d})^{-1}, \|\boldsymbol{\mu}_1\|_2^{-1}\}$.*

These assumptions are widely made in a series of recent works on the benign overfitting phenomena of gradient descent in learning over-parameterized CNN models [7, 10, 23]. We remark that, although the condition on the ratio between signal vectors is established for a clear presentation, it can certainly be relaxed to include other quantities beyond a constant order. The condition on the levels of signal and noise, firstly proposed in [7], is to ensure that the signal learning will not be overridden by the noise. The condition on the learning rate is to ensure the convergence of gradient descent. The condition on the initialization scaling is to guarantee that gradient descent is performing feature learning rather than learning random kernels.

Now we are ready to deliver our main theorem, which characterizes the critical properties of the learned convolutional filters.

Theorem 3.2 *Under Condition 3.1, for our signal sets $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K\}$, there exist K distinct filters $\{\mathbf{w}_{1,r_{1,1}}, \mathbf{w}_{1,r_{1,2}}, \cdots, \mathbf{w}_{1,r_{1,K}}\}$ in $\mathbf{W}_{+1}$ and K distinct filters $\{\mathbf{w}_{-1,r_{-1,1}}, \mathbf{w}_{-1,r_{-1,2}}, \cdots, \mathbf{w}_{1,r_{-1,K}}\}$ in $\mathbf{W}_{-1}$ such that at any iteration $T = \eta^{-1}\text{poly}(\|\boldsymbol{\mu}_1\|_2^{-1}, \cdots, \|\boldsymbol{\mu}_K\|_2^{-1}, d^{-1}\sigma_{\text{noise}}^{-2}, \sigma_0^{-1}, n, m, d) \geq \widetilde{\Omega}\big(\frac{m}{\eta\sigma_0^{q-2}\|\boldsymbol{\mu}_K\|_2^q}\big)$, with probability at least $1 - O(m^{-1})$, it holds that*

$$\Big\| \mathbf{w}_{1,r_{1,k}}^{(T)} - (m \log T) \cdot \frac{\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_k\|_2^2} \Big\|_2 \leq O\Big(\frac{m}{\|\boldsymbol{\mu}_k\|_2}\Big);$$

$$\Big\|\mathbf{w}_{-1,r_{-1,k}}^{(T)} + (m\log T)\cdot\frac{\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_k\|_2^2}\Big\|_2 \le O\Big(\frac{m}{\|\boldsymbol{\mu}_k\|_2}\Big),$$
$$\big\|\mathbf{w}_{j,r}^{(T)}\big\|_2 \le O\Big(\sum_{k=1}^{K}\frac{1}{m\|\boldsymbol{\mu}_k\|_2}\Big) + O(\sigma_0 d^{1/2}),$$

with $r \neq r_{1,k}, r_{-1,k}$ for all $k \in [K]$. Moreover, denote $\mathbf{P}_{\mathcal{U}^\perp}$ the projection matrix of the compliment space $\mathrm{span}(\mathcal{U})^\perp$, then $\|\mathbf{P}_{\mathcal{U}^\perp}\cdot(\mathbf{w}_{j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)})\|_2 \le O(\sigma_0 n^{1/2})$ for all $j \in \{-1,+1\}$ and $r \in [m]$.

The results of Theorem 3.2 demonstrate that for each signal $\boldsymbol{\mu}_k$, only the two filters $\mathbf{w}_{1,r_{1,k}}$ and $\mathbf{w}_{-1,r_{-1,k}}$ can significantly learn $\boldsymbol{\mu}_k$ and $-\boldsymbol{\mu}_k$. The projections of filters into the complement subspace of signal sets are small values by our Condition 3.1, implying that the noise data are rarely learned in any filter when the magnitude of signals is large compared to noise. As demonstrated in the previous work of benign over-fitting in two-layer neural networks [7, 23], the neural network can achieve both lower empirical training loss and test population loss if it rarely learns noises. However, their result is not sufficient to conjecture the potential structure of the neural networks. In comparison, we propose a more refined analysis as we carefully check the optimization trajectory of each filter during the whole training. Our results implies that we can approximately regard $\mathbf{w}_{j,r_{j,k}}^{(T)} \approx m\log(T)\frac{\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_k\|_2^2}$ and $\mathbf{w}_{j,r}^{(T)} \approx \mathbf{w}_{j,r}^{(0)}$ if $r \neq r_{j,k}$. The significant distinctions among the filters intuitively show that the neural network can exhibit an inherent prioritizing towards a low-rank structure that aligns with the intrinsic complexity of the data set.

Corollary 3.3 *Let $\mathbf{X} = [\mathbf{x}_1^{(1)}, \cdots, \mathbf{x}_1^P, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_2^P, \ldots, \mathbf{x}_n^{(1)}, \cdots, \mathbf{x}_n^P]$ be the matrix consisting of all the training input patches, and denote by $\omega_1, \ldots, \omega_{nP}$ and $\lambda_1, \ldots, \lambda_{2m}$ the singular values (in descending order) of $\mathbf{X}$ and $\mathbf{W}$ respectively. Then under the same conditions as Theorem 3.2, with probability at least $1 - O(m^{-1})$, it holds that*

$$\frac{\omega_1}{\omega_{nP}} \le \frac{2\|\boldsymbol{\mu}_1\|_2}{\sigma_{\mathrm{noise}}\sqrt{d}}, \quad \textit{and} \quad \frac{\lambda_K}{\lambda_{K+1}} \ge \Omega\bigg(\frac{m\log(T)\|\boldsymbol{\mu}_1\|_2^{-1}}{\sigma_0\sqrt{d}}\bigg).$$

Corollary 3.3 clearly demonstrates the different patterns of data ranks and learned filters under varying noise levels. Specifically, as the noise becomes stronger (i.e., σ_{noise} increases), the condition number of the data matrix (i.e., ω_1/ω_{nP}) decreases and gradually approaches 1, indicating that the rank of the training data matrix increases and approaches nP. In contrast, the ratio between the K-th largest and $K+1$-th largest eigenvalues of the learned filters is independent of the noise strength, and thus remains largely unchanged as the noise increases. Furthermore, when using a small initialization scaling σ_0, we observe that the gap between λ_K and λ_{K+1} becomes significantly large, suggesting that the rank of the learned filters is approximately K. This clearly explains the "rank robustness" phenomenon of the CNN model trained by gradient descent.

4. Conclusions and Future Work

In this paper, we point out an interesting phenomenon on the robustness of CNNs trained by gradient descent in learning the intrinsic dimension of data. For a specific type of data, we theoretically show that the two-layer CNN will converge to a low-rank structure when learning from noisy data, and even if the ranks of the data have exploded due to the added noises, the CNN rank still remains robust.

Experiments on MNIST, CIFAR10 data sets also support our findings. Exploring more accurate definitions of "ranks" and extending our results to more complicated data and networks are some interesting future work directions.

References

[1] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[3] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[6] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[7] Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

[8] Yuan Cao, Difan Zou, Yuanzhi Li, and Quanquan Gu. The implicit bias of batch normalization in linear models and two-layer linear convolutional neural networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5699–5753. PMLR, 2023.

[9] Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

[10] Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

[11] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020.

[12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[13] Spencer Frei, Gal Vardi, Peter Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. In *The Eleventh International Conference on Learning Representations*, 2022.

[14] Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3173–3228. PMLR, 2023.

[15] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[16] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

[17] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[18] Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2022.

[19] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

[20] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[21] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/ji19a.html.

[22] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

[23] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2023.

[24] Yiwen Kou, Zixiang Chen, and Quanquan Gu. Implicit bias of gradient descent for two-layer relu and leaky relu networks on nearly-orthogonal data. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[26] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

[27] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020.

[28] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

[29] Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rate. In *The Twelfth International Conference on Learning Representations*, 2023.

[30] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[31] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

[32] Xuran Meng, Jianfeng Yao, and Yuan Cao. Multiple descent in the multiple random feature model. *Journal of Machine Learning Research*, 25(44):1–49, 2024.

[33] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 3051–3059. PMLR, 2019. URL http://proceedings.mlr.press/v89/nacson19a.html.

[34] Mary Phuong and Christoph H Lampert. The inductive bias of relu networks on orthogonally separable data. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[35] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

[36] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19:70:1–70:57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

[37] Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In *International Conference on Algorithmic Learning Theory*, pages 1429–1459. PMLR, 2023.

[38] Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66 (6):86–93, 2023.

[39] Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR, 2021.

[40] Denny Wu and Ji Xu. On the optimal weighted l2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

[41] Hanxu Zhou, Zhou Qixuan, Tao Luo, Yaoyu Zhang, and Zhi-Qin Xu. Towards understanding the condensation of neural networks at initial training. *Advances in Neural Information Processing Systems*, 35:2184–2196, 2022.

[42] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. In *International Conference on Learning Representations (ICLR)*, 2023.

Appendix A. Additional Related Works

Implicit bias. There emerges a line of works studying the concept of 'implicit bias', the inherent property of learning algorithms prioritizing a solution with some specific structures, especially some 'simple' structures. For the implicit bias study on neural networks, [22, 30] demonstrated that q-homogeneous neural networks trained by gradient descent converge in direction to a KKT point of the maximum ℓ_2-margin problem. [31] proposed a stronger result base on symmetric data assumption and [39] extend the results to adaptive methods. [20, 22] showed the each layer of deep linear neural networks converges to a rank 1 matrix. [27] establish the equivalence between the gradient flow of depth-2 matrix factorization and a heuristic rank minimization algorithm. [13] showed that on nearly orthogonal data, gradient flow in leaky ReLU networks will achieve a linear boundary, and the stable rank of the neural networks is always bounded by a constant. [24] extends this result to gradient descent on similar data structures. [37] study the rank minimization on non-linear networks and provide several counter-examples. Besides, [38] provides a literature review of the existing works of implicit bias on deep neural networks.

Benign over-fitting. [5, 6] demonstrated the "double descent" population risk curve for many models, containing decision tree and Gaussian and random Fourier feature model. [4] showed that the benign overfitting in linear regression is correlated with the effective rank of the data covariance, and provided a theoretical bound for over-parameterized minimum norm interpolator. [9] study the benign overfitting in linear classification for a sub-Gaussian mixture model with noise flipping. [17, 40] study the implicit bias under the regime that dimension and sample increase at a fixed ratio. [1, 28, 32] explored the multiple descent under different settings. Besides, [7, 14, 23] study the benign overfitting on two-layer neural networks.

Appendix B. Experiments

In this section, we present our experimental results to backup our theoretical results and show a two-layer CNN is robust to noise in data.

We generate training data from the MNIST [12] and CIFAR10 [25] datasets according to Definition 2.1. We use images from two selected classes as the source of signal patches for the $y = -1$ class and the $y = 1$ class, respectively. To control the rank of the training data, we reshape each image into one vector and stack them into a matrix, then use PCA to reduce its rank. After that, we generate P patches with noise as one data point $\mathbf{x}$ from each column of the matrix. For noise generation, we use entry-wise Gaussian noise $N(0, \sigma_{\text{noise}}^2)$, where we set σ_{noise} to different values to verify how our model behaves under varying levels of noise. We consider a CNN model as defined in Section 2. To ensure the rank of the initialized parameters of the model does not affect the observed rank after training, we multiply the initialized weights by a small coefficient. For different data, we use different setups to generate the noise data and run the full batch gradient decent to train the CNN:

MNIST. The MNIST images firstly undergo dimensionality reduction to three levels of rank: 10, 20, and 30. Then each column of the matrix, corresponding to an image, is reshaped to its original size and padded with a 14-pixel wide circle of noise (An example is shown in Figure 2). Finally, each noise-padded image is reshaped into P patches. The padded pixels are entry-wise Gaussian noise $N(0, \sigma_{\text{noise}}^2)$, where σ_{noise} is set to 0, 0.01, 0.1, 0.12, 0.15, 0.18, and 0.2. For the model, the model width m is set to 128. For each rank level, the initialization coefficients are set to 1e-3, 1e-3, and 1e-2, respectively.

(*a*) $\sigma_{\text{noise}} = 0.01$ (*b*) $\sigma_{\text{noise}} = 0.1$ (*c*) $\sigma_{\text{noise}} = 0.2$

Figure 2: Illustration of a training image from the MNIST dataset, reduced to rank 10 and padded with a circle of noise.

CIFAR10. To reduce the complexity of CIFAR10 data and more easily reveal the phenomenon, we transform the original CIFAR10 iamge into embeddings using ResNet-18. All embeddings are stacked into a matrix, which is then subjected to dimensionality reduction to ranks 15, 20, 25. After reducing the dimension of the embedding matrix, each embedding is concatenated with a noise vector and reshaped into P patches. Here, the standard deviation of the noise σ_{noise} is set to 0, 0.1, 0.5, 0.62, 0.65, 0.7 and 0.8. The model width m is set to 256, 512, 128 and the initialization coefficients are set to 1e-6, 3e-7, 3e-7 for each rank.

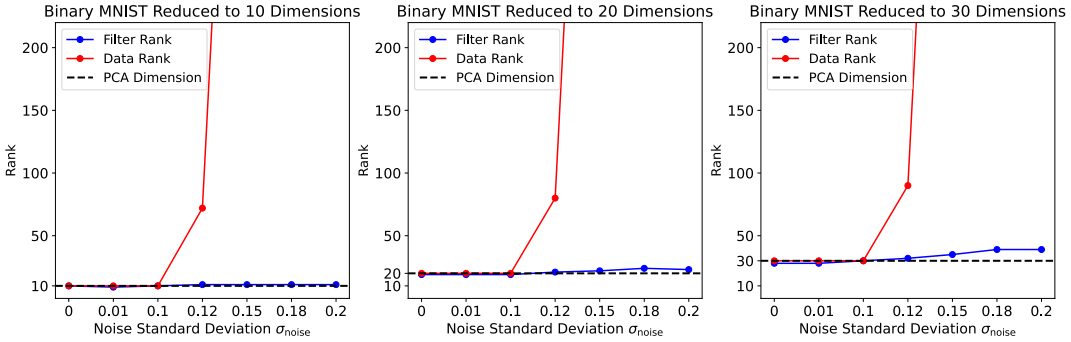
Synthetic Data. In addition to using two real-world datasets, we also conduct experiments on synthetic data. We strictly follow Definition 2.1 to generate the synthetic data. For the signal patches, we set $K = 10, 20$, and 30, and choose one-hot vectors as signals. Then, we set $\pi_1 = 1$ and $P = 3$, which means each data instance contains one signal patch and two Gaussian noise patches. And σ_{noise} is set to 0, 0.001, 0.0065, 0.009, 0.01, 0.012, 0.015. For the CNN model, The initialization coefficients are set to 1×10^{-4}, and the model width m is set to 128 for each K.

Result. According to Theorem 3.2, the rank of the filter is approximately equal to the number of signals. In this experiment, we report three different ranks: the dimension of PCA, *i.e.* the rank of pure signals wihtout noise, the rank of model weights, and the rank of the data with noise. Here, we verify whether the dimension of PCA is roughly equal to the rank of the filter after training. To evaluate the rank of the model weights and the noise data, we denote the number of their singular values larger than $\lambda_{\max}/100$ as the rank, where $\lambda_{\max}$ is the maximal singular value of the corresponding matrix. In all experiments, the rank results are presented when models have been trained to achieve a very small training error, in the range of 1e-1 to 1e-2. As shown in Figure 3, it is evident that as the noise and rank of data increase, the rank of the CNN filter remains approximately the same as the PCA dimension. This indicates that CNNs tend to learn the signals even when the data is exposed to significant noise.

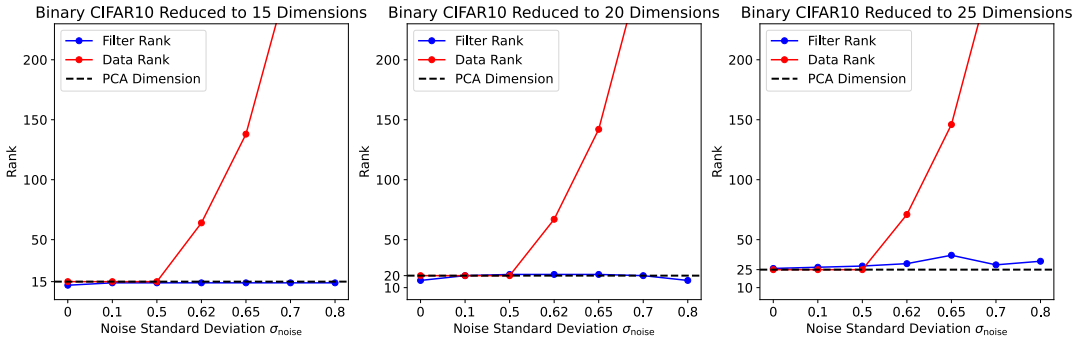
Appendix C. Overview of Proof Technique

In this section, we explain how we establish our main theoretical results. Similar to the definition of $\mathcal{U}$, we define the linear subspace spanned by the noise vectors as

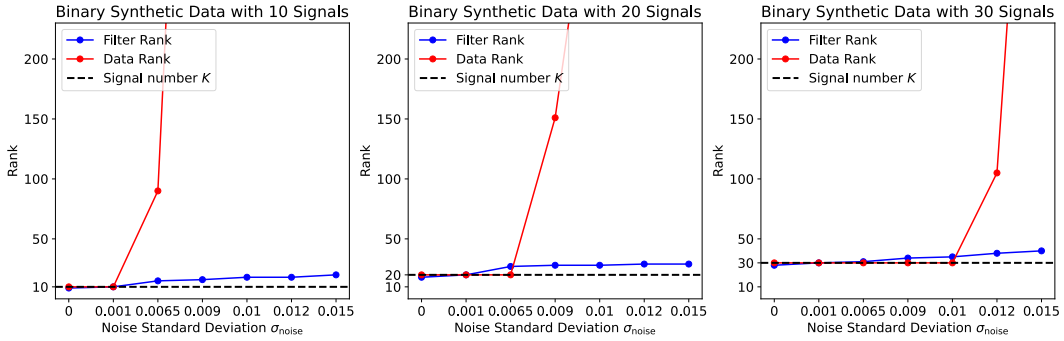
$$\mathcal{N} = \text{span}\{\boldsymbol{\xi}_{1,1}, \cdots, \boldsymbol{\xi}_{1,P-s_1}, \cdots, \boldsymbol{\xi}_{n,1}, \cdots, \boldsymbol{\xi}_{n,P-s_n}\} \subset \mathcal{U}^{\perp}.$$



(a) Rank visualization of data and CNN filters on MNIST dataset.



(b) Rank visualization of data and CNN filters on CIFAR10 dataset.



(c) Rank visualization of data and CNN filters on Synthetic Data.

Figure 3: Rank of the data and learned filters under different noise levels. Here x-axis represents the value of the standard deviation of noise σ_{noise}, and y-axis is the rank. From the figures, it can be clearly observed that the data rank increases rapidly as the noise becomes stronger, while the rank of the CNN filters remains robust against the noise, while keeps being the same as the intrinsic dimension of the data features.

According to the gradient iterative rule for filter (2.1), we could observe that at each iteration, the update of filter is always in the subspace spanned by the signal vectors and noise vectors i.e., $\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)} \in \mathcal{U} \cup \mathcal{N}$. Since the signal vectors are pairly orthogonal and also orthogonal to the noise vectors, we propose a decomposition of $\mathbf{w}_{j,r}^{(t)}$ as

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + \sum_{k=1}^{K} \frac{\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle}{\|\boldsymbol{\mu}_k\|_2^2} \boldsymbol{\mu}_k + \boldsymbol{\Xi}_{j,r}^{(t)} \tag{C.1}$$

where $\Xi_{j,r}^{(t)}$ is the linear projection of $\mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}$ into $\mathcal{N}$. By this decomposition, we can split the projection $\mathbf{P}_{\mathcal{U}^\perp} \cdot \mathbf{w}_{j,r}^{(t)} = \mathbf{P}_{\mathcal{U}^\perp} \cdot \mathbf{w}_{j,r}^{(0)} + \boldsymbol{\Xi}_{j,r}^{(t)}$. Since the initialization of neural networks is sufficiently small under our setting, we can treat $\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_k \rangle$. This inner product efficiently characterizes the level of different signals learned by each filter. On the other hand, $\Xi_{j,r}^{(t)}$ reflects the noise from different data points learned by each filter. In the following, we use a two-stage decoupling technique to characterize the increase of these values.

C.1. Training Phase I

Our training filters $\mathbf{w}_{j,r}$'s are initialized at a sufficiently small level, therefore we could treat $-\ell_i' \approx 1/2$ at the beginning. Until the output of neural networks surpasses a constant level, we can always treat $-\ell_i'$ as a constant and there is no significant difference among all training data. Therefore, the dominating factor in the iterative rules of our filters is the output of the activation function. Since the Huberized Relu activation function exhibits a power increase with order $q \geq 3$ at the beginning stage. This power increase significantly distinguishes the signal level among $\mathbf{w}_{j,r}$'s by the end of Phase I.

Lemma C.1 *Under the Condition 3.1, we can find a time $T_1 = \widetilde{\Theta}\big(\frac{m}{\eta \sigma_0^{q-2} \|\boldsymbol{\mu}_K\|_2^q}\big)$, then for our signal sets $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K\}$, there exist K distinct filters $\{\mathbf{w}_{1,r_{1,1}}, \mathbf{w}_{1,r_{1,2}}, \cdots, \mathbf{w}_{1,r_{1,K}}\}$ in $\mathbf{W}_{+1}$ and K distinct filters $\{\mathbf{w}_{-1,r_{-1,1}}, \mathbf{w}_{-1,r_{-1,2}}, \cdots, \mathbf{w}_{1,r_{-1,K}}\}$ in $\mathbf{W}_{-1}$ such that for all $j \in \{-1, +1\}$, $k \in [K]$, it holds that $\langle \mathbf{w}_{j,r_{j,k}}^{(T_1)}, j\boldsymbol{\mu}_k \rangle \geq \kappa$ and $\langle \mathbf{w}_{j,r}^{(T_1)}, j\boldsymbol{\mu}_k \rangle \leq \frac{1}{4Km}$ with all $r \neq r_{j,k}$. Moreover, it holds that $\|\boldsymbol{\Xi}_{j,r}^{(t)}\|_2^2 \leq \sigma_0^2 nP/2$ for all $j \in \{\pm 1\}$, $r \in [m]$ and $0 \leq t \leq T_1$.*

Lemma C.1 show that at a time $T_1 = \widetilde{\Theta}\big(\frac{m}{\eta \sigma_0^{q-2} \|\boldsymbol{\mu}_K\|_2^q}\big)$, for each signal $\boldsymbol{\mu}_k$, only one filter $\mathbf{w}_{1,r_{1,k}}$ can learn $+\boldsymbol{\mu}_k$ and only one filter $\mathbf{w}_{-1,r_{-1,k}}$ can learn $-\boldsymbol{\mu}_k$, and the level of $\langle \mathbf{w}_{1,r_{1,k}}^{(T_1)}, \boldsymbol{\mu}_k \rangle$ and $\langle \mathbf{w}_{-1,r_{-1,k}}^{(T_1)}, -\boldsymbol{\mu}_k \rangle$ will attain κ, the critical point of the Huberized Relu activation function. By definition of the Huberized Relu activation function, we can easily obtain that the power term will vanish when the input of the activation function, i.e., $\langle \mathbf{w}_{j,r_{j,k}}, j\boldsymbol{\mu}_k \rangle$ attains κ. Moreover, we can not treat $-\ell_i' = \Theta(1)$ like Phase I, since some input terms of the loss function also attain constant level. To better illustrate our results, we denote by $\mathcal{J}_k$ the set of data points containing only one signal vector $\boldsymbol{\mu}_k$ in their signal patches, i.e., $\mathcal{J}_k = \big\{i | s_i = 1,\ \boldsymbol{\nu}_{i,1} = \boldsymbol{\mu}_k,\ \text{and } i \in [n]\big\}$. One good property for $i \in \mathcal{J}_k$ is that we have $-\ell_i' = \Theta(\exp(-\langle \mathbf{w}_{y_i, r_{y_i,k}}, y_i \boldsymbol{\mu}_k \rangle / m))$ if the noise always remains at a small level. As we will show in the next lemma, this property guarantees that after the power term vanishes, $\langle \mathbf{w}_{j,r_{j,k}}, j\boldsymbol{\mu}_k \rangle$ will increase logarithmically.

C.2. Training Phase II

Lemma C.2 *Under the Condition 3.1, for our signal sets $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K\}$, there exist K distinct filters $\{\mathbf{w}_{1,r_{1,1}}, \mathbf{w}_{1,r_{1,2}}, \cdots, \mathbf{w}_{1,r_{1,K}}\}$ in $\mathbf{W}_{+1}$ and K distinct filters $\{\mathbf{w}_{-1,r_{-1,1}}, \mathbf{w}_{-1,r_{-1,2}}, \cdots, \mathbf{w}_{1,r_{-1,K}}\}$ in $\mathbf{W}_{-1}$ such that at any time $T^* = T_1 + \eta^{-1}\mathrm{poly}(\|\boldsymbol{\mu}_1\|_2^{-1}, \cdots, \|\boldsymbol{\mu}_K\|_2^{-1}, d^{-1}\sigma_{\mathrm{noise}}^{-2}, \sigma_0^{-1}, n, m, d)$ and for all $j \in \{-1, +1\}$, $k \in [K]$, then*

$$m\log(T^* - T_1) - O(m) \leq \langle \mathbf{w}_{j,r_{j,k}}^{(T^*)}, j\boldsymbol{\mu}_k \rangle \leq m\log(T^* - T_1) + O(m)$$

and $\langle \mathbf{w}_{j,r}^{(T^*)}, j\boldsymbol{\mu}_k \rangle \leq \frac{1}{2Km}$ for all $r \neq r_{j,k}$. Moreover, it holds that $\|\boldsymbol{\Xi}_{j,r}^{(t)}\|_2^2 \leq 2\sigma_0^2 nP$ for all $j \in \{\pm 1\}$, $r \in [m]$ and $0 \leq t \leq T^*$.

In the Phase II training, we don't specify a particular time T^*, T^* could be any polynomials of our parameters and the only requirement for T^* is that T^* is larger than the T_1, then we can approximately claim that $\mathbf{w}_{j,r_{j,k}}^{(T^*)} \approx \log(T^* - T_1)\frac{j\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_k\|_2^2}$, and $\mathbf{w}_{j,r}^{(T^*)} \approx \mathbf{w}_{j,r}^{(0)}$ if $r \neq r_{j,k}$. These intuitive results clearly illustrate after a long time of training, the neural network will prioritize a 'low-rank' structure that aligns with the intrinsic data complexity.

Now, we are ready to prove our main Theorem 3.2.

Proof [Proof of Theorem 3.2] Let $T = \eta^{-1}\mathrm{poly}(\|\boldsymbol{\mu}_1\|_2^{-1}, \cdots, \|\boldsymbol{\mu}_K\|_2^{-1}, d^{-1}\sigma_{\mathrm{noise}}^{-2}, \sigma_0^{-1}, n, m, d)$ and $T \geq \Omega\big(\frac{m}{\eta\sigma_0^{q-2}\|\boldsymbol{\mu}_1\|_2^q}\big)$. We calculate the filter norm by (C.1)

$$\begin{aligned}
&\bigg\|\mathbf{w}_{j,r_{j,k}}^{(T)} - \big(m\log T\big)\cdot\frac{j\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_k\|_2^2}\bigg\|_2 \\
=&\bigg\|\mathbf{w}_{j,r_{j,k}}^{(0)} + \big(\langle \mathbf{w}_{j,r_{j,k}}^{(T)} - \mathbf{w}_{j,r_{j,k}}^{(0)}, \boldsymbol{\mu}_k\rangle - jm\log T\big)\frac{\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_k\|_2^2} + \sum_{k'\neq k}\langle \mathbf{w}_{j,r_{j,k}}^{(T)} - \mathbf{w}_{j,r_{j,k}}^{(0)}, \boldsymbol{\mu}_{k'}\rangle\frac{\boldsymbol{\mu}_{k'}}{\|\boldsymbol{\mu}_{k'}\|_2^2} + \boldsymbol{\Xi}_{j,r_{j,k}}^{(T)}\bigg\|_2 \\
\leq&\|\mathbf{w}_{j,r_{j,k}}^{(0)}\|_2 + \frac{\big|\langle \mathbf{w}_{j,r_{j,k}}^{(T)} - \mathbf{w}_{j,r_{j,k}}^{(0)}, j\boldsymbol{\mu}_k\rangle - m\log T\big|}{\|\boldsymbol{\mu}_k\|_2} + \sum_{k'\neq k}\frac{\big|\langle \mathbf{w}_{j,r_{j,k}}^{(T)} - \mathbf{w}_{j,r_{j,k}}^{(0)}, \boldsymbol{\mu}_{k'}\rangle\big|}{\|\boldsymbol{\mu}_{k'}\|_2} + \|\boldsymbol{\Xi}_{j,r_{j,k}}^{(T)}\|_2 \\
\leq&O(\sigma_0 d^{1/2}) + O\Big(\frac{m}{\|\boldsymbol{\mu}_k\|_2}\Big) + O\Big(\sum_{k'\neq k}\frac{1}{m\|\boldsymbol{\mu}_{k'}\|_2}\Big) + O(\sigma_0 n^{1/2}) \leq O\Big(\frac{m}{\|\boldsymbol{\mu}_k\|_2}\Big).
\end{aligned}$$

The first inequality is from triangle inequality, the second inequality is from Lemma C.2 and concentration results in Appendix E.1 guaranteeing that $\|\mathbf{w}_{j,r_{j,k}}^{(0)}\|_2 \leq O(\sigma_0 d^{1/2})$ and $|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle| \leq O(\sigma_0\|\boldsymbol{\mu}_k\|_2) \leq O(d^{-1/2})$, and the last inequality is by Condition 3.1. For $\mathbf{w}_{j,r}$ with $r \neq r_{j,k}$ for all $j \in \{-1, +1\}$ and $k \in [K]$, we have

$$\begin{aligned}
\big\|\mathbf{w}_{j,r}^{(T)}\big\|_2 &\leq \big\|\mathbf{w}_{j,r}^{(0)}\big\|_2 + \sum_{k=1}^{K}\frac{\big|\langle \mathbf{w}_{j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle\big|}{\|\boldsymbol{\mu}_k\|_2} + \|\boldsymbol{\Xi}_{j,r}^{(T)}\|_2 \\
&\leq O(\sigma_0 d^{1/2}) + O\Big(\sum_{k=1}^{K}\frac{1}{m\|\boldsymbol{\mu}_k\|_2}\Big) + O(\sigma_0 n^{1/2}) \\
&\leq O(\sigma_0 d^{1/2}) + O\Big(\sum_{k=1}^{K}\frac{1}{m\|\boldsymbol{\mu}_k\|_2}\Big)
\end{aligned}$$

Similarly, the first inequality is from triangle inequality, the second inequality is from Lemma C.2 and some concentration results in Appendix E.1, and the last inequality is by Condition 3.1. By definition of $\mathbf{P}_{\mathcal{U}^\perp}$, we can directly have $\mathbf{P}_{\mathcal{U}^\perp} \cdot \big(\mathbf{w}_{j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)}\big) = \boldsymbol{\Xi}_{j,r}^{(t)}$, which proves the last result in Theorem 3.2. ■

Appendix D. Proof in Section C

D.1. Decomposition

We introduce a more refined decomposition of $\mathbf{w}_{j,r}^{(t)}$ compared to (C.1), the most significant difference is that we define an exact agent to describe the learning of each noise $\boldsymbol{\xi}_{i,k'}$ on each filter $\mathbf{w}_{j,r}$.

Definition D.1 *Let $\mathbf{w}_{j,r}^{(t)}$ for $j \in \{\pm 1\}$, $r \in [m]$ be the convolution filters of the CNN at the t-th iteration of gradient descent. Then there exist unique coefficients $\gamma_{j,k,r}^{(t)} \geq 0$ and $\rho_{j,r,i,k'}^{(t)}$ such that*

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + \sum_{k=1}^{K} j \cdot \gamma_{j,k,r}^{(t)} \cdot \|\boldsymbol{\mu}_k\|_2^{-2} \cdot \boldsymbol{\mu}_k + \sum_{i=1}^{n} \sum_{k'=1}^{P-s_i} \rho_{j,r,i,k'}^{(t)} \cdot \|\boldsymbol{\xi}_{i,k'}\|_2^{-2} \cdot \boldsymbol{\xi}_{i,k'}$$

We further denote $\bar{\rho}_{j,r,i,k'}^{(t)} := \rho_{j,r,i,k'}^{(t)} \mathbb{1}(\rho_{j,r,i,k'}^{(t)} \geq 0)$, $\underline{\rho}_{j,r,i,k'}^{(t)} := \rho_{j,r,i,k'}^{(t)} \mathbb{1}(\rho_{j,r,i,k'}^{(t)} \leq 0)$. Then we have,

$$\begin{aligned}\mathbf{w}_{j,r}^{(t)} =& \mathbf{w}_{j,r}^{(0)} + \sum_{k=1}^{K} j \cdot \gamma_{j,k,r}^{(t)} \cdot \|\boldsymbol{\mu}_k\|_2^{-2} \cdot \boldsymbol{\mu}_k + \sum_{i=1}^{n} \sum_{k'=1}^{P-s_i} \bar{\rho}_{j,r,i,k'}^{(t)} \cdot \|\boldsymbol{\xi}_{i,k'}\|_2^{-2} \cdot \boldsymbol{\xi}_{i,k'} \\ &+ \sum_{i=1}^{n} \sum_{k'=1}^{P-s_i} \underline{\rho}_{j,r,i,k'}^{(t)} \cdot \|\boldsymbol{\xi}_{i,k'}\|_2^{-2} \cdot \boldsymbol{\xi}_{i,k'}\end{aligned}$$

Then, instead of directly analyzing $\boldsymbol{\Xi}_{j,k}^{(t)}$, we prove some result for $\rho_{j,r,i,k'}$ and extend the results of $\rho_{j,r,i,k'}$ to $\boldsymbol{\Xi}_{j,k}^{(t)}$. Besides, we define two set notations: $\mathcal{I}_k$ is the set of data points containing signal vector $\boldsymbol{\mu}_k$ in their signal patches, $\mathcal{J}_k$ is the set of data points containing only one signal vector $\boldsymbol{\mu}_k$ in their signal patches, i.e., $\mathcal{I}_k = \big\{i | \boldsymbol{\mu}_k \in \{\nu_{i,1}, \cdots, \nu_{i,s_i}\},\ \text{and } i \in [n]\big\}$, and $\mathcal{J}_k = \big\{i | s_i = 1,\ \boldsymbol{\mu}_k = \nu_{i,1},\ \text{and } i \in [n]\big\}$

D.2. Preliminary Lemmas

Before we prove the Lemma C.1, we first present and prove several lemmas that will be used for the proof of Lemma C.1. We define $r_{j,k,t} = \operatorname{argmax}_{r \in [m]} \langle \mathbf{w}_{j,r}^{(t)}, j \cdot \boldsymbol{\mu}_k \rangle$. Then we have the following lemma demonstrating that the filter with the largest inner product with some signal at initialization will always have the largest inner product with this signal during the whole training process.

Lemma D.2 *Under Condition 3.1, we have $r_{j,k,t} = r_{j,k,0}$ for all $j \in \{+1, -1\}, k \in [K]$ and $t > 0$. Moreover, if $\{j, k\} \neq \{j', k'\}$, then $r_{j,k,0} \neq r_{j',k',0}$ holds with probability at least $1 - O(1/m)$.*

Since the filter with the largest inner product with any specific signal is consistent during the whole training process, in the following paragraphs, we can denote $r_{j,k,t}$ by $r_{j,k}$ for simplicity and $r_{j,k}$'s are distinct for different pair of $\{j,k\}$. To prove Lemma D.2, we consider two cases with positive initialization and negative initialization respectively. Instead of directly proving Lemma D.2, we introduce and prove the following Lemma D.3 and Lemma D.4, then Lemma D.2 is a direct corollary of Lemma D.3 and Lemma D.4.

Lemma D.3 *For all $r, r' \in [m]$, if $\langle \mathbf{w}_{j,r}^{(0)}, j\boldsymbol{\mu}_k \rangle \geq \langle \mathbf{w}_{j,r'}^{(0)}, j\boldsymbol{\mu}_k \rangle \geq 0$, then it holds that $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle \geq \langle \mathbf{w}_{j,r'}^{(t)}, j\boldsymbol{\mu}_k \rangle$ for all t.*

Proof [Proof of Lemma D.3] By multiple by $j\boldsymbol{\mu}_k$ on both sides of (2.1) and the orthogonality between signals and noises, we obtain that

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t+1)}, j\boldsymbol{\mu}_k \rangle &= \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle - \frac{\eta ||\boldsymbol{\mu}_k||_2^2}{nm} \sum_{i \in \mathcal{I}_k} \ell'[y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)] \sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\mu}_k \rangle\big) \\
&= \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle - \frac{\eta ||\boldsymbol{\mu}_k||_2^2}{nm} \sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle\big) \sum_{i \in \mathcal{I}_k, y_i = j} \ell_i'^{(t)} \\
&\quad - \frac{\eta ||\boldsymbol{\mu}_k||_2^2}{nm} \sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, -j\boldsymbol{\mu}_k \rangle\big) \sum_{i \in \mathcal{I}_k, y_i = -j} \ell_i'^{(t)}. \qquad \text{(D.1)}
\end{aligned}$$

It is clear that $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle$ is always non-decreasing, therefore if $\langle \mathbf{w}_{j,r}^{(0)}, j\boldsymbol{\mu}_k \rangle \geq 0$, we have $\langle \mathbf{w}_{j,r}^{(t)}, -j\boldsymbol{\mu}_k \rangle \leq 0$ for all t, then (D.1) could be simplified as

$$\langle \mathbf{w}_{j,r}^{(t+1)}, j\boldsymbol{\mu}_k \rangle = \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle - \frac{\eta ||\boldsymbol{\mu}_k||_2^2}{nm} \sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle\big) \sum_{i \in \mathcal{I}_k, y_i = j} \ell_i'^{(t)} \qquad \text{(D.2)}$$

And we could notice that the only item specific to filter r in formula (D.2) is the inner product $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle$. In another word, if we let $x_r^{(t)} = \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle$, then the recursion (D.2) of the positive sequences $\{x_r^{(t)}\}_{t=0}^{\infty}$ could be simplified as,

$$x_r^{(t+1)} = x_r^{(t)} + \eta C_t \sigma'\big(x_r^{(t)}\big)$$

where $C_t = \frac{||\boldsymbol{\mu}_k||_2^2}{nm} \sum_{i \in \mathcal{I}_k, y_i = j} \ell_i'^{(t)}$ is independent of filter index r, and $\sigma'(\cdot)$ is a non-decreasing function. Therefore we conclude that a filter with a larger initialization will always have a larger increment in each iteration, which completes the proof. ■

To compare with the filters with a negative initialization, we define an idealized filter $\widetilde{\mathbf{w}}_{j,k}$ satisfying that $\big(1 + \Theta(\frac{\delta}{m^2})\big)\langle \widetilde{\mathbf{w}}_{j,k}^{(0)}, j\boldsymbol{\mu}_k \rangle = \langle \mathbf{w}_{j,r_{j,k,0}}^{(0)}, j\boldsymbol{\mu}_k \rangle$, and also following the iterative rule in (D.2). The reason for such a definition is that if $r \neq r_{j,k,0}$, we have $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle \leq \langle \widetilde{\mathbf{w}}_{j,k}^{(0)}, \boldsymbol{\mu}_k \rangle$ by Lemma E.5. Next, we introduce our Lemma D.4.

Lemma D.4 *Suppose that Condition 3.1 holds and define an idealized filter $\widetilde{\mathbf{w}}_{j,k}$ satisfying that $\big(1 + \Theta(\frac{\delta}{m^2})\big)\langle \widetilde{\mathbf{w}}_{j,k}^{(0)}, j\boldsymbol{\mu}_k \rangle = \langle \mathbf{w}_{j,r_{j,k,0}}^{(0)}, j\boldsymbol{\mu}_k \rangle$, and also following the iterative rule in (D.2). Then it holds that $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k \rangle < \langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k \rangle$ for all $r \in [m]$ and t if $\langle \mathbf{w}_{j,r}^{(0)}, j\boldsymbol{\mu}_k \rangle < 0$.*

Proof [Proof of Lemma D.4]

Obviously, if $\langle \mathbf{w}_{j,r}^{(0)}, j\boldsymbol{\mu}_k\rangle < 0$, then $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle < \langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle$ since $\langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle$ is always positive. We assume that t^* is the first time such that $\langle \boldsymbol{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle > 0$, which means $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle < \langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle$ for $t < t^*$. And similarly, we could re-write the iterative formula for $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle$ at t^* is

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t^*)}, j\boldsymbol{\mu}_k\rangle &= \langle \mathbf{w}_{j,r}^{(t^*-1)}, j\boldsymbol{\mu}_k\rangle - \frac{\eta||\boldsymbol{\mu}_k||_2^2}{nm}\sigma'\big(\langle \mathbf{w}_{j,r}^{(t^*)}, j\boldsymbol{\mu}_k\rangle\big) \sum_{i\in\mathcal{I}_k, y_i=-j} \ell_i'^{(t^*-1)} \\
&\leq -\frac{\eta||\boldsymbol{\mu}_k||_2^2}{nm}\sigma'\big(\langle \mathbf{w}_{j,r}^{(t^*)}, j\boldsymbol{\mu}_k\rangle\big) \sum_{i\in\mathcal{I}_k, y_i=-j} \ell_i'^{(t^*-1)} \leq \frac{\eta||\boldsymbol{\mu}_k||_2^2}{\kappa^{q-1}m}\widetilde{O}(\sigma_0\|\boldsymbol{\mu}_k\|_2)^{q-1} \\
&\leq \frac{\sigma_0||\boldsymbol{\mu}_k||_2}{m} \leq \langle \widetilde{\mathbf{w}}_{j,k}^{(0)}, j\boldsymbol{\mu}_k\rangle \leq \langle \widetilde{\mathbf{w}}_{j,k}^{(t^*)}, j\boldsymbol{\mu}_k\rangle
\end{aligned}$$

The first inequality holds since $\langle \mathbf{w}_{j,r}^{(t^*-1)}, j\boldsymbol{\mu}_k\rangle < 0$. The second inequality is from $-\sum_{i\in\mathcal{I}_k, y_i=-j} \ell_i'^{(t)} \leq n$ and $\langle \mathbf{w}_{j,r}^{(t^*-1)}, -j\boldsymbol{\mu}_k\rangle \leq \langle \mathbf{w}_{j,r}^{(0)}, -j\boldsymbol{\mu}_k\rangle = \widetilde{O}(\sigma_0\|\boldsymbol{\mu}_k\|_2)$ by Lemma E.5. The third inequality is from our Condition 3.1 and the fourth inequality is from Lemma E.5. Since $\langle \mathbf{w}_{j,r}^{(t^*)}, j\boldsymbol{\mu}_k\rangle, \langle \widetilde{\mathbf{w}}_{j,k}^{(t^*)}, j\boldsymbol{\mu}_k\rangle > 0$, by Lemma D.3, we have $\langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle > \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle$ for all $t \geq t^*$, which completes the proof. ■

Now, we are ready to prove Lemma D.2

Proof [Proof of Lemma D.2] By Lemma D.3, Lemma E.5 and Lemma D.4, we can conclude that if $r \neq r_{j,k,0}$, then $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle \leq \langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle \leq \langle \mathbf{w}_{j,r_{j,k,0}}^{(t)}, j\boldsymbol{\mu}_k\rangle$ for all t. Since the initialization of $\mathbf{w}_{j,r}^{(0)}$ is i.i.d. Gaussian random vectors, we conclude that $P(r_{j,k,0} = r_{j',k',0}) = \frac{1}{m}$ for different pair of $\{j,k\}$. ■

Next, we introduce and prove the following Lemma D.5 and Lemma D.6 which will be helpful. Lemma D.5 characterize the relationship between $-\ell_i'$ and the output of F_{y_i} when $|\rho_{j,r,i,k'}^{(t)}|$ is small. Lemma D.6 show that when $|\rho_{j,r,i,k'}^{(t)}|$ is small $\|\boldsymbol{\Xi}_{j,r}^{(t)}\|_2$ is also small.

Lemma D.5 *Suppose that Condition 3.1 holds and* $|\rho_{j,r,i,k'}^{(t)}| \leq O(\sigma_0\sigma_{\text{noise}}\sqrt{d})$ *for all* $j \in \{\pm 1\}, r \in [m]$, $i \in [n]$ *and* $k' \in [P - s_i]$, *then we have*

$$\begin{aligned}
&\langle \mathbf{w}_{j,r}^{(t)}, \xi_{i,k'}\rangle \leq \widetilde{O}(\sigma_0\sigma_{\text{noise}}\sqrt{d}); \\
&F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i) \leq 1; \\
&|\ell_i'^{(t)}| = \Theta\left(e^{-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)}\right).
\end{aligned}$$

Proof [Proof of Lemma D.5] The iterative rule for $\langle \mathbf{w}_{j,r}^{(t)}, \xi_{i,k'}\rangle$ can be derived by multiple by $\xi_{i,k'}$ on both sides of (2.1), then we obtain that

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle &= \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{i,k'}\rangle + \rho_{j,r,i,k'}^{(t)} + \sum_{i'\neq i}\sum_{k''=1}^{P-s_i} \rho_{j,r,i',k''}^{(t)} \frac{\langle \boldsymbol{\xi}_{i,k'}, \boldsymbol{\xi}_{i',k''}\rangle}{\|\boldsymbol{\xi}_{i',k''}\|_2^2} + \sum_{k''\neq k'} \rho_{j,r,i,k''}^{(t)} \frac{\langle \boldsymbol{\xi}_{i,k'}, \boldsymbol{\xi}_{i,k''}\rangle}{\|\boldsymbol{\xi}_{i,k''}\|_2^2} \\
&\leq \widetilde{O}(\sigma_0\sigma_{\text{noise}}\sqrt{d}).
\end{aligned}$$

The last inequality holds because the first term $\langle \mathbf{w}_{-y_i,r}^{(0)}, \boldsymbol{\xi}_{i,k'} \rangle = \widetilde{O}(\sigma_0 \sigma_{\text{noise}} \sqrt{d})$ by applying Lemma E.5, the second term $\underline{\rho}_{j,r,i,k'}^{(t)} \leq 0$. For the third term and the forth term, $\rho_{j,r,i',k''}^{(t)}, \rho_{j,r,i,k''}^{(t)} = \widetilde{O}(\sigma_0 \sigma_{\text{noise}} \sqrt{d})$ by our assumption and $\frac{\langle \boldsymbol{\xi}_{i,k'}, \boldsymbol{\xi}_{i',k''} \rangle}{\|\boldsymbol{\xi}_{i',k''}\|_2^2}, \frac{\langle \boldsymbol{\xi}_{i,k'}, \boldsymbol{\xi}_{i,k''} \rangle}{\|\boldsymbol{\xi}_{i,k''}\|_2^2} \leq \widetilde{O}(1/\sqrt{d})$ by Lemma E.4. Then based on our Condition 3.1 about n and d, the last two terms are also $\widetilde{O}(\sigma_0 \sigma_{\text{noise}} \sqrt{d})$. Now for $F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i)$, its value is determined by $\langle \mathbf{w}_{-y_i,r}^{(t)}, y_i \boldsymbol{\mu}_k \rangle$ and $\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_{i,k'} \rangle$, and we can easily get

$$\langle \mathbf{w}_{-y_i,r}^{(t)}, y_i \boldsymbol{\mu}_k \rangle = \langle \mathbf{w}_{-y_i,r}^{(0)}, y_i \boldsymbol{\mu}_k \rangle - \gamma_{-y_i,k,r} \leq \langle \mathbf{w}_{-y_i,r}^{(0)}, y_i \boldsymbol{\mu}_k \rangle \leq \widetilde{O}(\sigma_0 \|\boldsymbol{\mu}_k\|_2),$$

Combining with the previous result about $\langle \mathbf{w}_{j,r}^{(t)}, \xi_{i,k'} \rangle$, we can present a bound of $F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i)$ as

$$\begin{aligned} F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i) &\leq \frac{1}{m} \sum_{r=1}^{m} \sum_{k=1}^{K} \sigma(\langle \mathbf{w}_{-y_i,r}^{(t)}, y_i \boldsymbol{\mu}_k \rangle) + \frac{1}{m} \sum_{r=1}^{m} \sum_{k=1'}^{s_i} \sigma(\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_{i,k'} \rangle) \\ &\leq \frac{2P}{q\kappa^{q-1}} \cdot \max\{\widetilde{O}(\sigma_0 \|\boldsymbol{\mu}_k\|_2), \widetilde{O}(\sigma_0 \sigma_{\text{noise}} \sqrt{d})\}^q \leq 1. \end{aligned}$$

The last inequality is from our Condition 3.1 about σ_0. Finally by the definition of $\ell'(\cdot)$, it is clear that,

$$|\ell_i'^{(t)}| = \frac{1}{1 + e^{y_i[F_{+1}(\mathbf{W}_{+1}^{(t)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(t)}, \mathbf{x}_i)]}} = \frac{1}{1 + e^{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i)}}.$$

By the fact $F_{+1}(\cdot), F_{-1}(\cdot) \geq 0$, the lower bound is straightforward that

$$|\ell_i'^{(t)}| = \frac{1}{1 + e^{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i)}} \geq \frac{1}{2e^{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)}}$$

On the other side, since $F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i) \leq 1$, we obtain that

$$|\ell_i'^{(t)}| = \frac{1}{1 + e^{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t)}, \mathbf{x}_i)}} \leq e \cdot e^{-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)}$$

The upper and lower bound of $|\ell_i'^{(t)}|$ indicates that $|\ell_i'^{(t)}| = \Theta\Big(e^{-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)}\Big)$. ■

Lemma D.6 *Suppose that Condition 3.1 holds. Then we have* $\|\Xi_{j,r}^{(t)}\|_2^2 \leq 2nPa^2\sigma_{\text{noise}}^{-2} d^{-1}$ *for all* $j \in \{-1, +1\}$, $r \in [m]$, $i \in [n]$ *and* $k' \in [P - s_i]$, *if* $|\rho_{j,r,i,k'}^{(t)}| \leq a$. *Here* a *could be any positive number.*

Proof [Proof of Lemma D.6] By deposition defined in Definition D.1, we have

$$\begin{aligned} \|\boldsymbol{\Xi}_{j,r}^{(t)}\|_2^2 &= \sum_{i=1}^{n} \sum_{k'=1}^{P-s_i} [\rho_{j,r,i,k'}^{(t)}]^2 \|\boldsymbol{\xi}_{i,k'}\|_2^{-2} + \sum_{\{i,k\} \neq \{i',k'\}} \rho_{j,r,i,k}^{(t)} \rho_{j,r,i',k'}^{(t)} \frac{\langle \boldsymbol{\xi}_{i,k}, \boldsymbol{\xi}_{i',k'} \rangle}{\|\boldsymbol{\xi}_{i,k}\|_2^2 \|\boldsymbol{\xi}_{i',k'}\|_2^2} \\ &\leq 2n(P-1)a^2\sigma_{\text{noise}}^{-2} d^{-1} + a^2\sigma_{\text{noise}}^{-2} d^{-1} \leq 2nPa^2\sigma_{\text{noise}}^{-2} d^{-1} \end{aligned}$$

where the first inequality is from Lemma E.4 and our Condition 3.1. ■

D.3. Proof of Lemma C.1

Now, we are ready to prove Lemma C.1

Proof [Proof of Lemma C.1] By Lemma D.6, to show $\|\boldsymbol{\Xi}_{j,r}^{(t)}\|_2^2 \leq \sigma_0^2 nP/2$, it suffices to show that $\max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}| \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}/2$. We will show it by induction, and we assume it holds when we prove the first result. For each $j \in \{-1,+1\}$ and $k \in [K]$, we consider the filter $\mathbf{w}_{j,r_{j,k}}$ and the idealized filter $\widetilde{\mathbf{w}}_{j,k}$ defined in Lemma D.4, then $\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle$ and $\langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle$ follows the same iterative rule:

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{(t+1)}, j\boldsymbol{\mu}_k\rangle &= \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle - \frac{\eta\|\boldsymbol{\mu}_k\|_2^2}{nm}\sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle\big) \sum_{i\in\mathcal{I}_k, y_i=j} \ell_i'^{(t)} \\ &= \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle + \eta C_t \sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle\big),\end{aligned}$$

where $C_t = \frac{\|\boldsymbol{\mu}_k\|_2^2}{nm}\sum_{i\in\mathcal{I}_k, y_i=j} \ell_i'^{(t)} \leq \frac{\|\boldsymbol{\mu}_k\|_2^2}{m}$. Define $T_{1,j,k}$ be the first time such that $\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle \geq \kappa$ and T' be the first time such that $\langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle \geq \frac{1}{4Km}$. Since we have $\frac{\langle \mathbf{w}_{j,r_{j,k}}^{(0)}, j\boldsymbol{\mu}_k\rangle}{\langle \widetilde{\mathbf{w}}_{j,k}^{(0)}, j\boldsymbol{\mu}_k\rangle} = 1 + \Theta(\frac{\delta}{m^2})$ by definition of the idealized filter $\widetilde{\mathbf{w}}_{j,k}$. By checking the conditions in Lemma E.7, we can conclude $T_{1,j,k} < T'$, which implies that $\langle \widetilde{\mathbf{w}}_{j,k}^{(T_{1,j,k})}, j\boldsymbol{\mu}_k\rangle < \frac{1}{4Km}$. Since for all $r \neq r_{j,k}$, we have $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle \leq \langle \widetilde{\mathbf{w}}_{j,k}^{(t)}, j\boldsymbol{\mu}_k\rangle$ for all t by Lemma E.5, Lemma D.3 and Lemma D.4, we can finally obtain that $\langle \mathbf{w}_{j,r}^{(T_{1,j,k})}, j\boldsymbol{\mu}_k\rangle \leq \frac{1}{4Km}$ for all $r \neq r_{j,k}$. Next, we try to derive the bound for $T_{1,j,k}$. As we assume $|\rho_{j,r,i,k'}| \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}/2$, then for all $t \leq T_{1,j,k}$ and $i \in \mathcal{J}_k$, we apply Lemma D.5 and derive that

$$\begin{aligned}-\ell_i^{(t)} &\geq \frac{1}{2e}\exp\big(-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\big) \\ &\geq \frac{1}{2e}\exp\bigg(-\frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\boldsymbol{\mu}_k\rangle\big) + \sum_{k'=1}^{P-1}\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle\big)\Big]\bigg) \geq \frac{1}{2e^2}.\end{aligned}$$

This is because

$$\frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\boldsymbol{\mu}_k\rangle\big) + \sum_{k'=1}^{P-1}\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle\big)\Big] \leq 1$$

by $\langle \mathbf{w}_{y_i,r_{y_i,k}}^{(t)}, y_i\boldsymbol{\mu}_k\rangle \leq \kappa$, $\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\boldsymbol{\mu}_k\rangle \leq \frac{1}{4Km}$ and $\langle \mathbf{w}_{j,r}^{(t)}, \xi_{i,k'}\rangle \leq \widetilde{O}(\sigma_0\sigma_{\text{noise}}\sqrt{d})$. Therefore, we get a lower bound for C_t as

$$C_t = \frac{\|\boldsymbol{\mu}_k\|_2^2}{nm}\sum_{i\in\mathcal{I}_k, y_i=j} \ell_i'^{(t)} \geq \frac{\|\boldsymbol{\mu}_k\|_2^2}{2e^2 nm}|\{i \in \mathcal{J}_k, y_i = j\}| \geq \frac{\pi_1\|\boldsymbol{\mu}_k\|_2^2}{8e^2 Km}.$$

The last inequality is because Lemma E.1 and Lemma E.2. Therefore we have $C_t = \Theta\big(\frac{\|\boldsymbol{\mu}_k\|_2^2}{m}\big)$, then by Lemma E.8 and Lemma E.5, we can obtain that

$$T_{1,j,k} = \Theta\bigg(\frac{m}{\eta\|\boldsymbol{\mu}_k\|_2^2(\langle \mathbf{w}_{j,r_{j,k}}, j\boldsymbol{\mu}_k\rangle)^{q-2}}\bigg) = \widetilde{\Theta}\bigg(\frac{m}{\eta\sigma_0^{q-2}\|\boldsymbol{\mu}_k\|_2^q}\bigg).$$

Since for all $k \in [K]$, $\|\boldsymbol{\mu}_1\|_2/\|\boldsymbol{\mu}_k\|_2 = \Theta(1)$, we conclude we can find a time $T_1 = \Theta\big(\frac{m}{\eta\sigma_0^{q-2}\|\boldsymbol{\mu}_1\|_2^q}\big)$ such that the preceding results hold at T_1 for all $j \in \{-1, +1\}$ and $k \in [K]$. Finally we use induction to prove that $\max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}| \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}/2$. For simplicity we denote $\phi^{(t)} = \max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}|$. Obviously $\phi^{(0)} = 0$, and we suppose that exists $\widetilde{T} \leq T_1$ such that $\phi^{(t)} \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}/2$ holds for all $0 < t < \widetilde{T} - 1$. Then by the iterative rule for $\rho_{j,r,i,k'}^{(t)}$, we have

$$
\begin{aligned}
\phi^{(t+1)} &\leq \phi^{(t)} + \max_{j,r,i,k'} \frac{\eta\|\boldsymbol{\xi}_{i,k'}\|_2^2}{\kappa^{q-1}nm}\bigg|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{i,k'}\rangle + \phi^{(t)}\bigg(1 + \sum_{i'\neq i}\sum_{k''=1}^{P-s_i}\frac{\langle \boldsymbol{\xi}_{i,k'}, \boldsymbol{\xi}_{i',k''}\rangle}{\|\boldsymbol{\xi}_{i',k''}\|_2^2} + \sum_{k''\neq k'}\frac{\langle \boldsymbol{\xi}_{i,k'}, \boldsymbol{\xi}_{i,k''}\rangle}{\|\boldsymbol{\xi}_{i,k''}\|_2^2}\bigg)\bigg|^{q-1} \\
&\leq \phi^{(t)} + \widetilde{O}\Big(\frac{\eta\sigma_0^{q-1}\sigma_{\text{noise}}^{q+1}d^{(q+1)/2}}{nm}\Big)
\end{aligned}
$$

By taking the telescoping sum, we have $\phi^{(\widetilde{T})} \leq T_1 \cdot \widetilde{O}\Big(\frac{\eta\sigma_0^{q-1}\sigma_{\text{noise}}^{q+1}d^{(q+1)/2}}{nm}\Big) \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}/2$ by the formula for $T_1 = \widetilde{\Theta}(\frac{m}{\eta\sigma_0^{q-2}\|\boldsymbol{\mu}_1\|_2^q})$ and our SNR conditions. Since then, we have finished all the proof for Lemma C.1. ■

D.4. Proof of Lemma C.2

During the phase I, we always threat $-\ell_i' = \Theta(1)$, while in this phase, as the increasing of $\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle$, we can not regard $-\ell_i' = \Theta(1)$ since the training loss will eventually converge.

Proof [Proof of Lemma C.1] By Lemma D.6, to show $\|\boldsymbol{\Xi}_{j,r}^{(t)}\|_2^2 \leq 2\sigma_0^2 nP$, it suffices to show that $\max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}| \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}$. Similar to the proof of Phase I, we first prove the result for $\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle$ and then use induction to prove the result for $\max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}|$ and $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle$ with $r \neq r_{j,k}$. We assume the results for $\max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}|$ and $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle$ with $r \neq r_{j,k}$ hold when we prove the first result. From Lemma D.5, we can obtain that for all $i \in \mathcal{I}_k$ and $t > T_1$, it holds

$$
-\ell_i'^{(t)} \leq e \cdot \exp\big(-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\big) \leq e \cdot e^{-\frac{1}{m}\langle \mathbf{w}_{y_i,r_{y_i,k}}^{(t)}, y_i\boldsymbol{\mu}_k\rangle}, \tag{D.3}
$$

since the activation function $\sigma(\cdot)$ is always positive. Additionally, we can also obtain that for all $i \in \mathcal{J}_k$ and $t > T_1$, it holds

$$
\begin{aligned}
-\ell_i^{(t)} &\geq \frac{1}{2e}\exp\big(-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\big) \\
&\geq \frac{1}{2e}\exp\bigg(-\frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\boldsymbol{\mu}_k\rangle\big) + \sum_{k'=1}^{P-1}\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle\big)\Big]\bigg) \\
&= \frac{1}{2e}\exp\bigg(-\frac{1}{m}\langle \mathbf{w}_{y_i,r_{y_i,k}}^{(t)}, y_i\boldsymbol{\mu}_k\rangle - \frac{1}{m}\sum_{k'=1}^{P-1}\sigma\big(\langle \mathbf{w}_{y_i,r_{y_i,k}}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle\big)\bigg) \\
&\quad \cdot \exp\bigg(-\frac{1}{m}\sum_{r\neq r_{y_i,k}}\Big[\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\boldsymbol{\mu}_k\rangle\big) + \sum_{k'=1}^{P-1}\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle\big)\Big]\bigg) \geq \frac{1}{2e^2}e^{-\frac{1}{m}\langle \mathbf{w}_{y_i,r_{y_i,k}}^{(t)}, y_i\boldsymbol{\mu}_k\rangle}
\end{aligned}
$$

The last inequality is because

$$\frac{1}{m}\sum_{k'=1}^{P-1}\sigma\big(\langle \mathbf{w}_{y_i,r_{y_i,k}}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle\big) + \frac{1}{m}\sum_{r\neq r_{y_i,k}}\Big[\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\boldsymbol{\mu}_k\rangle\big) + \sum_{k'=1}^{P-1}\sigma\big(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_{i,k'}\rangle\big)\Big] \leq 1$$

by our assumption $\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\boldsymbol{\mu}_k\rangle \leq \frac{1}{2Km}$ and $\langle \mathbf{w}_{j,r}^{(t)}, \xi_{i,k'}\rangle \leq \widetilde{O}(\sigma_0\sigma_{\text{noise}}\sqrt{d})$. With such upper and lower bounds for $\ell_i'^{(t)}$ in hands, we can provide an upper and lower bound for the iterations of $\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle$ as

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t+1)}, j\boldsymbol{\mu}_k\rangle &= \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle - \frac{\eta||\boldsymbol{\mu}_k||_2^2}{nm}\sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle\big)\sum_{i\in\mathcal{I}_k, y_i=j}\ell_i'^{(t)} \\
&\leq \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle + \frac{e\eta||\boldsymbol{\mu}_k||_2^2}{nm}e^{-\frac{1}{m}\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle}\cdot\big|\{i\in\mathcal{I}_k, y_i=j\}\big| \\
&\leq \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle + \frac{e\eta||\boldsymbol{\mu}_k||_2^2}{m}e^{-\frac{1}{m}\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle},
\end{aligned}$$

since $\sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle\big) = 1$ and $\big|\{i\in\mathcal{I}_k, y_i=j\}\big| \leq n$, and

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t+1)}, j\boldsymbol{\mu}_k\rangle &\geq \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle - \frac{\eta||\boldsymbol{\mu}_k||_2^2}{nm}\sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle\big)\sum_{i\in\mathcal{J}_k, y_i=j}\ell_i'^{(t)} \\
&\geq \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle + \frac{\eta||\boldsymbol{\mu}_k||_2^2}{2e^2nm}e^{-\frac{1}{m}\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle}\cdot\big|\{i\in\mathcal{J}_k, y_i=j\}\big| \\
&\geq \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle + \frac{\pi_1\eta||\boldsymbol{\mu}_k||_2^2}{8e^2Km}e^{-\frac{1}{m}\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, j\boldsymbol{\mu}_k\rangle}.
\end{aligned}$$

since $\sigma'\big(\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle\big) = 1$ and $\big|\{i\in\mathcal{J}_k, y_i=j\}\big| \geq \frac{\pi_1}{4K}$ by Lemma E.1 and Lemma E.2. Applying these upper and lower bound on Lemma E.9, for all $t > T_1$ we obtain that

$$\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, \boldsymbol{\mu}_{j,k}\rangle \geq m\log\Big(\frac{\eta||\boldsymbol{\mu}_k||_2^2}{8e^2Km^2}(t-T_1) + e^{\frac{\kappa}{m}}\Big) \geq m\log(t-T_1) - O(m), \tag{D.4}$$

and

$$\langle \mathbf{w}_{j,r_{j,k}}^{(t)}, \boldsymbol{\mu}_{j,k}\rangle \leq \frac{e\eta||\boldsymbol{\mu}_k||_2^2}{m}e^{-\frac{\kappa}{m}} + m\log\Big(\frac{e\eta||\boldsymbol{\mu}_k||_2^2}{m^2}(t-T_1) + e^{\frac{\kappa}{m}}\Big) \leq m\log(t-T_1) + O(m). \tag{D.5}$$

This finishes the proof of the conclusion for $\langle \mathbf{w}_{j,r_{j,k}}^{(T^*)}, j\boldsymbol{\mu}_k\rangle$. Now, we use induction to prove that $\langle \mathbf{w}_{j,r}^{(T^*)}, j\boldsymbol{\mu}_k\rangle \leq \frac{1}{2Km}$ when $r \neq r_{j,k}$. We first derive a result that will be used for the following induction proof. Plugging (D.4) into (D.3), for all $i\in\mathcal{I}_k$, $y_i = j$ and $t > T_1$, we have

$$-\ell_i'^{(t)} \leq \frac{8e^3Km^2}{\eta||\boldsymbol{\mu}_k||_2^2(t-T_1+1)}$$

Taking the sum from $t = T_1$ to T^*, we have

$$-\sum_{t=T_1}^{T^*} \ell_i'^{(t)} \leq \frac{8e^3Km^2}{\eta||\boldsymbol{\mu}_k||_2^2}\log(T^* - T_1) \leq \widetilde{\Theta}\Big(\frac{m^2}{\eta||\boldsymbol{\mu}_k||_2^2}\Big) \tag{D.6}$$

Since at T_1, we have $\langle \mathbf{w}_{j,r}^{(T_1)}, j\boldsymbol{\mu}_k\rangle \leq \frac{1}{4Km}$ if $r \neq r_{j,k}$. Suppose that exists $T_1 < \widetilde{T} \leq T^*$ such that $\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, j\boldsymbol{\mu}_k\rangle \leq \frac{1}{2Km}$. Then by the iterative rule for $\langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu}_k\rangle$ and applying (D.6), we have

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(\widetilde{T})}, j\boldsymbol{\mu}_k\rangle &\leq \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, j\boldsymbol{\mu}_k\rangle - \frac{\eta||\boldsymbol{\mu}_k||_2^2}{\kappa^{q-1}nm}\Big(\frac{1}{2Km}\Big)^{q-1}\sum_{i\in\mathcal{I}_k, y_i=j} \ell_i'^{(\widetilde{T}-1)} \\
&\leq \langle \mathbf{w}_{j,r}^{(T_1)}, j\boldsymbol{\mu}_k\rangle - \frac{\eta||\boldsymbol{\mu}_k||_2^2}{\kappa^{q-1}nm}\Big(\frac{1}{2Km}\Big)^{q-1}\sum_{t=T_1}^{T^*}\sum_{i\in\mathcal{I}_k, y_i=j} \ell_i'^{(\widetilde{T}-1)} \\
&\leq \frac{1}{4Km} + \Theta\Big(\frac{\eta||\boldsymbol{\mu}_k||_2^2}{nm^{q-1}}\Big)\cdot\widetilde{\Theta}\Big(\frac{nm^2}{\eta||\boldsymbol{\mu}_k||_2^2}\Big)\cdot\frac{1}{4Km} \leq \frac{1}{2Km}.
\end{aligned}$$

This finishes the induction proof that $\langle \mathbf{w}_{j,r}^{(T^*)}, j\boldsymbol{\mu}_k\rangle \leq \frac{1}{2Km}$ for all $r \neq r_{j,k}$ and $t < T^*$. Next, we proof that $\max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}| \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}$ holds for all $t < T^*$. For simplicity we denote $\phi^{(t)} = \max_{j,r,i,k'} |\rho_{j,r,i,k'}^{(t)}|$. Obviously we have $\phi^{(T_1)} \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}/2$, and we suppose that exists $T_1 \leq \widetilde{T} \leq T_2$ such that $\phi^{(t)} \leq \sigma_0\sigma_{\text{noise}}\sqrt{d}$ holds for all $T_1 < t < \widetilde{T} - 1$. Then by the iterative rule for $\rho_{j,r,i,k'}^{(t)}$ and plugging (D.6), we have

$$\phi^{(\widetilde{T})} \leq \phi^{(T_1)} + \widetilde{O}\Big(\frac{\eta\sigma_0^{q-2}\sigma_{\text{noise}}^q d^{q/2}}{nm}\Big)\cdot\widetilde{\Theta}\Big(\frac{m^2}{\eta||\boldsymbol{\mu}_k||_2^2}\Big)\cdot\frac{\sigma_0\sigma_{\text{noise}}\sqrt{d}}{2} \leq \sigma_0\sigma_{\text{noise}}\sqrt{d},$$

where the last inequality holds by our SNR condition and Condition 3.1 that $m = O(\sigma_0^{2-q}\sigma_{\text{noise}}^{2-q}d^{(2-q)/2})$. ■

Appendix E. Technical Lemmas

E.1. Concentration Results

Lemma E.1 *Suppose that $\delta > 0$, then for any $\mathcal{I} \subseteq [n]$, with probability at least $1 - O(\delta)$,*

$$|\{i \in \mathcal{I} : y_i = 1\}|,\ |\{i \in \mathcal{I} : y_i = -1\}| = \frac{|\mathcal{I}|}{2} + O\Big(\sqrt{|\mathcal{I}|\log(1/\delta)}\Big).$$

Proof [Proof of Lemma E.1] By Hoeffding's inequality, with probability at least $1 - O(\delta)$, we have

$$\bigg|\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\mathbb{1}\{y_i = 1\} - \frac{1}{2}\bigg| \leq O\bigg(\sqrt{\frac{\log(1/\delta)}{|\mathcal{I}|}}\bigg).$$

Therefore,

$$|\{i \in \mathcal{I} : y_i = 1\}| = \sum_{i\in\mathcal{I}}\mathbb{1}\{y_i = 1\} = \frac{|\mathcal{I}|}{2} + O\Big(\sqrt{|\mathcal{I}|\log(1/\delta)}\Big).$$

This proves the result for $|\{i \in \mathcal{I} : y_i = 1\}|$. The proof for $|\{i \in \mathcal{I} : y_i = -1\}|$ is exactly the same, and we can conclude the proof by applying a union bound. ■

Lemma E.2 *Suppose that $\delta > 0$, then for $\mathcal{J}_k$ defined in Section D, with probability at least $1 - O(\delta)$, it holds that*

$$|\mathcal{J}_k| = \frac{\pi_1}{K} n + O\Big(\sqrt{n \log(1/\delta)}\Big).$$

Proof [Proof of Lemma E.1] By Hoeffding's inequality, with probability at least $1 - O(\delta)$, we have

$$\left| \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{i \in \mathcal{J}_k\} - \frac{\pi_1}{K} \right| \leq O\left(\sqrt{\frac{\log(1/\delta)}{n}} \right).$$

Therefore,

$$|\mathcal{J}_k| = \sum_{i \in [n]} \mathbb{1}\{i \in \mathcal{J}_k\} = \frac{\pi_1}{K} n + O\Big(\sqrt{n \log(1/\delta)}\Big),$$

which finishes the proof. ■

Lemma E.3 *Suppose that $z \sim \mathcal{N}(0, 1)$, then $\mathbb{P}(|z| \leq t) = O(t)$.*

Proof [Proof of Lemma E.3] We use $\phi(x)$ to denote the density function of the standard Gaussian random variable, and then we know that $\max \phi(x) = \phi(0)$. By this fact,

$$\mathbb{P}(|z| \leq t) = 2 \int_0^t \phi(x) dx \leq 2\phi(0) t = O(t)$$

■

Lemma E.4 *Suppose that $\delta > 0$ and $d = \Omega\big(\log(nm/\delta)\big)$. Then with probability at least $1 - O(\delta)$, it holds that*

$$\begin{aligned}
&\|\boldsymbol{\xi}_{i,k}\|_2^2 = \Theta(\sigma_{\text{noise}}^2 d); \\
&\|\mathbf{w}_{j,r}^{(0)}\|_2^2 = \Theta(\sigma_0^2 d); \\
&\big|\langle \boldsymbol{\xi}_{i,k}, \boldsymbol{\xi}_{i',k'} \rangle\big| \leq O\big(\sigma_{\text{noise}}^2 \cdot \sqrt{d \log(n^2/\delta)}\big)
\end{aligned}$$

for all $j \in \{+1, -1\}$, $r \in [m]$, and all $i, i' \in [n]$, $k \in [P - s_i], k' \in [P - s_{i'}]$ such that $\{i, k\} \neq \{i', k'\}$.

Proof [Proof of Lemma E.4] By Bernstein's inequality, with probability at least $1 - O(\delta/n)$ we have

$$\big|\|\boldsymbol{\xi}_{i,k}\|_2^2 - \sigma_{\text{noise}}^2 d\big| = O\big(\sigma_{\text{noise}}^2 \cdot \sqrt{d \log(n/\delta)}\big).$$

Therefore, as long as $d = \Omega\big(\log(n/\delta)\big)$, we have

$$\|\boldsymbol{\xi}_{i,k}\|_2^2 = \Theta(\sigma_{\text{noise}}^2 d).$$

Similarly, by Bernstein's inequality, with probability at least $1 - O(\delta/m)$ we have

$$\big|\|\mathbf{w}_{j,r}^{(0)}\|_2^2 - \sigma_0^2 d\big| = O\big(\sigma_{\text{noise}}^2 \cdot \sqrt{d\log(m/\delta)}\big).$$

Therefore, as long as $d = \Omega\big(\log(m/\delta)\big)$, we have

$$\|\mathbf{w}_{j,r}^{(0)}\|_2^2 = \Theta(\sigma_0^2 d).$$

Moreover, for any i, i', k, k' with $\{i,k\} \neq \{i',k'\}$, clearly $\langle \boldsymbol{\xi}_{i,k}, \boldsymbol{\xi}_{i',k'}\rangle$ has mean zero and by Bernstein's inequality, with probability at least $1 - O(\delta/n^2)$ we have

$$|\langle \boldsymbol{\xi}_{i,k}, \boldsymbol{\xi}_{i',k'}\rangle| \leq O\big(\sigma_{\text{noise}}^2 \cdot \sqrt{d\log(n^2/\delta)}\big).$$

Applying a union bound completes the proof. ■

Lemma E.5 *Suppose that $d \geq \Omega\big(\log(mn/\delta)\big)$, $m = \Omega\big(\log(1/\delta)\big)$. Then with probability at least $1 - O(\delta)$, it holds that*

$$\begin{aligned}
\big|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle\big| &= O\left(\sqrt{\log(m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}_k\|_2\right), \\
\big|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{i,k'}\rangle\big| &= O\left(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_{\text{noise}} \sqrt{d}\right)
\end{aligned}$$

for all $r \in [m]$, $j \in \{\pm 1\}$, $i \in [n]$, $k \in [K]$ and $k' \in [P - s_i]$. Besides,

$$\begin{aligned}
\max_{r\in[m]} j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle &= \Omega\left(\sigma_0 \|\boldsymbol{\mu}_k\|_2\right), \\
\max_{r\in[m]} j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{i,k'}\rangle &= \Omega\left(\sigma_0 \sigma_{\text{noise}} \sqrt{d}\right)
\end{aligned}$$

for all $j \in \{\pm 1\}$, $i \in [n]$, $k \in [K]$ and $k' \in [P - s_i]$. Moreover,

$$j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle \left(1 + \Theta\left(\frac{\delta}{m^2}\right)\right) \leq \max_{r\in[m]} j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle$$

for all $r \neq \operatorname{argmax}_{r\in[m]} j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle$, $j \in \{\pm 1\}$ and $k \in [K]$.

Proof [Proof of Lemma E.5] It is clear that for each $r \in [m]$, $j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle$ is a Gaussian random variable with mean zero and variance $\sigma_0^2 \|\boldsymbol{\mu}_k\|_2^2$. Therefore, by Gaussian tail bound and union bound, with probability at least $1 - O(\delta)$,

$$j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle \leq |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle| \leq O\big(\sqrt{\log(m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}_k\|_2\big).$$

Moreover, $\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}_k\|_2/2 > j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big)$ is an absolute constant, and therefore by the condition on m, we have

$$\begin{aligned}\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}_k\|_2/2 \le \max_{r\in[m]} j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big) &= 1-\mathbb{P}(\sigma_0\|\boldsymbol{\mu}_k\|_2/2 > \max_{r\in[m]} j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big)\\ &= 1-\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}_k\|_2/2 > j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big)^{2m}\\ &\ge 1-O(\delta).\end{aligned}$$

By Lemma E.4, with probability at least $1-O(\delta)$, $\|\boldsymbol{\xi}_{i,k'}\|_2 = \Theta\big(\sigma_{\text{noise}}\sqrt{d}\big)$ for all $i\in[n]$ and $k'\in[P-s_i]$. Therefore, the result for $\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_{i,k'}\rangle$ follows the same proof as $j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle$. Lastly, for different r,r' and $\forall t>0$,

$$\begin{aligned}\mathbb{P}\left(\frac{\big|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big|\vee\big|\langle\mathbf{w}_{j,r'}^{(0)},\boldsymbol{\mu}_k\rangle\big|}{\big|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big|\wedge\big|\langle\mathbf{w}_{j,r'}^{(0)},\boldsymbol{\mu}_k\rangle\big|}\le 1+t\right) &\le \mathbb{P}\left(1-t\le\frac{\big|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big|}{\big|\langle\mathbf{w}_{j,r'}^{(0)},\boldsymbol{\mu}_k\rangle\big|}\le 1+t\right)\\ &\le \mathbb{P}\Big(\big|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\big|\le 2t\big|\langle\mathbf{w}_{j,r'}^{(0)},\boldsymbol{\mu}_k\rangle\big|\Big)=O(t)\end{aligned}$$

where the last equality holds from Lemma E.3 and the fact that $\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle$ and $\langle\mathbf{w}_{j,r'}^{(0)},\boldsymbol{\mu}_k\rangle$ are independent Gaussian random variables with mean 0 and same variance. By this result, let $t=\Theta(\frac{\delta}{m^2})$ and use union bound, we could deduce that with probability at least $1-O(\delta)$,

$$j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle\left(1+\Theta\left(\frac{\delta}{m^2}\right)\right)\le\max_{r\in[m]} j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle$$

for all $r\neq\text{argmax}_{r\in[m]}\, j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_k\rangle$. ■

E.2. Tensor Power Methods

The following lemmas are inspired by [2, 8, 19]

Lemma E.6 *If a positive sequence $\{x_t\}_{t=0}^{\infty}$ satisfies the updating rules $x_{t+1}=x_t+\eta\cdot C_t\cdot x_t^{q-1}$, then $\forall k\in\mathbb{N},\zeta\in(0,1)$, we have*

$$\sum_{t>0,x_t\le(1+\zeta)^k x_0}\eta C_t\le\frac{\zeta}{x_0^{q-2}}\frac{1-\frac{1}{(1+\zeta)^{(q-2)k}}}{1-\frac{1}{(1+\zeta)^{(q-2)}}}+\eta\cdot\left[(1+\zeta)^{q-1}\sum_{g=0}^{k-1}C_{\mathcal{T}_{g+1}-1}+C_{\mathcal{T}_k}\right],\tag{E.1}$$

and

$$\sum_{t>0,x_t\le(1+\zeta)^k x_0}\eta C_t\ge\frac{\zeta}{x_0^{q-2}(1+\zeta)^{q-1}}\frac{1-\frac{1}{(1+\zeta)^{(q-2)k}}}{1-\frac{1}{(1+\zeta)^{(q-2)}}}-\frac{\eta}{(1+\zeta)^{q-1}}\sum_{g=1}^{k-1}C_{\mathcal{T}_g-1},\tag{E.2}$$

where $\mathcal{T}_g$ be the first iteration such that $x_t\ge(1+\zeta)^g x_0$

Proof [Proof of Lemma E.6] By the definition of $\mathcal{T}_g$, we have

$$x_{\mathcal{T}_{g+1}} - x_{\mathcal{T}_g} = \sum_{t\in[\mathcal{T}_g,\mathcal{T}_{g+1})} \eta C_t x_t^{q-1} \geq \sum_{t\in[\mathcal{T}_g,\mathcal{T}_{g+1})} \eta \cdot C_t \cdot \left[x_0(1+\zeta)^g\right]^{q-1}, \tag{E.3}$$

and

$$\begin{aligned} x_{\mathcal{T}_{g+1}} - x_{\mathcal{T}_g} &= x_{\mathcal{T}_{g+1}-1} - x_{\mathcal{T}_g} + \eta \cdot C_{\mathcal{T}_{g+1}-1} \cdot x_{\mathcal{T}_{g+1}-1}^{q-1} \\ &\leq \zeta(1+\zeta)^g x_0 + \eta \cdot C_{\mathcal{T}_{g+1}-1} \cdot \left[x_0(1+\zeta)^{g+1}\right]^{q-1}. \end{aligned} \tag{E.4}$$

By combining (E.3) and (E.4) in order and rearranging some items, we could deduce,

$$\sum_{t\in[\mathcal{T}_g,\mathcal{T}_{g+1})} \eta C_t \leq \frac{\zeta}{x_0^{q-2}[(1+\zeta)^{q-2}]^g} + \eta(1+\zeta)^{q-1} C_{\mathcal{T}_{g+1}-1}.$$

Take a telescoping sum of this result, and then we finish the proof of (E.1). For (E.2), considering the opposite direction of the inequalities (E.3) and (E.4), repeating the previous process will get the result. ■

Lemma E.7 *Suppose there are two positive sequence $\{x_t\}_{t=0}^{\infty}$ and $\{y_t\}_{t=0}^{\infty}$ satisfying the following updating rules:*

$$\begin{aligned} x_{t+1} &= x_t + \eta \cdot C_t \cdot {x_t}^{q-1}; \\ y_{t+1} &= y_t + \eta \cdot C_t \cdot {y_t}^{q-1}, \end{aligned}$$

with $q \geq 3$ and $\frac{x_0}{y_0} \geq 1 + c$, where c is a small positive number. For any two positive number A_x and A_y, let T_x, T_y are the first time s.t. $x_{T_x} \geq A_x$ and $y_{T_y} \geq A_y$ respectively. If we have $0 < C_t < \bar{C}$ and η and y_0 are both sufficiently small such that $\eta = \widetilde{O}\left(\frac{c}{\bar{C} y_0^{q-3} A_y}\right)$ and $\frac{y_0}{A_y} \leq O(c)$, then it holds that $T_x \leq T_y$.

Proof [Proof of Lemma E.7] For a positive $\zeta > 0$, let k_1, k_2 be the smallest integer s.t. $x_0(1+\zeta)^{k_1} \geq A_x$ and $y_0(1+\zeta)^{k_2} \geq A_y$. From these definitions, we have

$$\frac{\log(\frac{A_x}{x_0})}{\log(1+\zeta)} \leq k_1 < \frac{\log(\frac{A_x}{x_0})}{\log(1+\zeta)} + 1,$$

and

$$\frac{\log(\frac{A_y}{y_0})}{\log(1+\zeta)} \leq k_2 < \frac{\log(\frac{A_y}{y_0})}{\log(1+\zeta)} + 1.$$

By Lemma E.6, we further derive that

$$\sum_{t=0}^{T_x} \eta C_t \leq \frac{\zeta}{x_0^{q-2}} \frac{1 - \frac{1}{(1+\zeta)^{(q-2)k_1}}}{1 - \frac{1}{(1+\zeta)^{(q-2)}}} + \eta \cdot \left[(1+\zeta)^{q-1} \sum_{g=0}^{k_1-1} C_{\mathcal{T}_{g+1}-1} + C_{\mathcal{T}_{k_1}}\right]$$

$$\leq \frac{\zeta}{x_0^{q-2}} \frac{1}{1-\frac{1}{(1+\zeta)^{(q-2)}}} + \eta \cdot (1+\zeta)^{q-1}(k_1+1)\bar{C}, \tag{E.5}$$

and

$$\begin{aligned}\sum_{t=0}^{T_y} \eta C_t &\geq \frac{\zeta}{y_0^{q-2}(1+\zeta)^{q-1}} \frac{1-\frac{1}{(1+\zeta)^{(q-2)k_2}}}{1-\frac{1}{(1+\zeta)^{(q-2)}}} - \frac{\eta}{(1+\zeta)^{q-1}} \sum_{g=1}^{k_2-1} C_{\mathcal{T}_g-1} \\ &\geq \frac{\zeta}{y_0^{q-2}(1+\zeta)^{q-1}} \frac{1-\left(\frac{y_0}{A_y}\right)^{q-2}}{1-\frac{1}{(1+\zeta)^{(q-2)}}} - \frac{\eta}{(1+\zeta)^{q-1}}(k_2-1)\bar{C}. \end{aligned} \tag{E.6}$$

We use (E.6) minus (E.5) and get

$$\begin{aligned}\sum_{t=0}^{T_y} \eta C_t - \sum_{t=0}^{T_x} \eta C_t \geq & \underbrace{\frac{\zeta}{1-\frac{1}{(1+\zeta)^{(q-2)}}} \left\{ \frac{1-\left(\frac{y_0}{A_y}\right)^{q-2}}{y_0^{q-2}(1+\zeta)^{q-1}} - \frac{1}{x_0^{q-2}} \right\}}_{I_1} \\ & \underbrace{- \eta\bar{C} \left\{ \frac{k_2-1}{(1+\zeta)^{q-1}} + (1+\zeta)^{q-1}(k_1+1) \right\}}_{I_2}. \end{aligned}$$

We consider the value of I_1 and I_2 separately and carefully choose ζ such that

$$(1+\zeta)^{q-1} = \left(1-\frac{y_0}{A_y}\right)^2 (1+c)^{q-2} = 1 + \Theta(c).$$

The last equality is from our assumption $\frac{y_0}{A_y} = O(c)$, and we could also conclude $\zeta = \Theta(c)$. Then for I_1, we have,

$$I_1 \geq \frac{\zeta}{y_0^{q-2}(1-\frac{1}{(1+\zeta)^{(q-2)}})(1+c)^{q-2}} \left\{ \frac{1-\frac{y_0}{A_y}}{\left(1-\frac{y_0}{A_y}\right)^2} - 1 \right\} = \Omega\left(\frac{1}{y_0^{q-3}A_y}\right). \tag{E.7}$$

Because $\frac{1}{1-\frac{1}{(1+\zeta)^{(q-2)}}} = \Theta(\zeta)$. For I_2, we have,

$$I_2 \leq \eta\bar{C}\Theta(k_1 \vee k_2) = \eta\bar{C}\widetilde{\Theta}\left(\frac{1}{c}\right). \tag{E.8}$$

Now by combining (E.7) and (E.8), we could conclude that $\sum_{t=0}^{T_y} \eta C_t - \sum_{t=0}^{T_x} \eta C_t \geq 0$, which completes the proof. ■

Lemma E.8 *Suppose a positive sequence $\{x_t\}_{t=0}^{\infty}$ satisfies the following iterative rules:*

$$x_{t+1} \geq x_t + \eta \cdot C_1 \cdot {x_t}^{q-1};$$

$$x_{t+1} \leq x_t + \eta \cdot C_2 \cdot {x_t}^{q-1},$$

with $C_2 \geq C_1 > 0$. For any $v > x_0$, let T_v be the first time such that $x_t \geq v$, then for any constant $\zeta > 0$, we have

$$T_v \leq \frac{1+\zeta}{\eta C_1 x_0^{q-2}} + \frac{(1+\zeta)^{q-1} C_2 \log(\frac{v}{x_0})}{C_1}, \tag{E.9}$$

and

$$T_v \geq \frac{1}{(1+\zeta)^{q-1}\eta C_2 x_0^{q-2}} - \frac{\log(\frac{v}{x_0})}{(1+\zeta)^{q-2}}. \tag{E.10}$$

Proof [Proof of Lemma E.8] Similar to the definition in Lemma E.6, let $\mathcal{T}_g$ be the first iteration such that $x_t \geq (1+\zeta)^g x_0$. Moreover, let g^* be the smallest integer such that $(1+\zeta)^{g^*} x_0 \geq v$, resulting

$$\frac{\log(\frac{v}{x_0})}{\log(1+\zeta)} \leq g^* < \frac{\log(\frac{v}{x_0})}{\log(1+\zeta)} + 1.$$

For $t = \mathcal{T}_1$,

$$x_{\mathcal{T}_1} \geq x_0 + \sum_{t=0}^{\mathcal{T}_1 - 1} \eta C_1 x_t^{q-1} \geq x_0 + \mathcal{T}_1 \eta C_1 x_0^{q-1},$$

and we could obtain that

$$\mathcal{T}_1 \leq \frac{x_{\mathcal{T}_1} - x_0}{\eta C_1 x_0^{q-1}}. \tag{E.11}$$

Consider the upper-bound iteration of $x_{\mathcal{T}_1}$ and the fact that $x_{\mathcal{T}_1 - 1} \leq x_0(1+\zeta)$, we could get

$$x_{\mathcal{T}_1} \leq x_{\mathcal{T}_1 - 1} + \eta C_2 x_{\mathcal{T}_1 - 1}^{q-1} \leq x_0(1+\zeta) + \eta C_2 x_0^{q-1}(1+\zeta)^{q-1}. \tag{E.12}$$

Combining the results from (E.11) and (E.12), we obtain that,

$$\mathcal{T}_1 \leq \frac{\zeta}{\eta C_1 x_0^{q-2}} + \frac{(1+\zeta)^{q-1} C_2}{C_1}.$$

Similarly for $g > 1$,

$$x_{\mathcal{T}_g} \geq x_{\mathcal{T}_{g-1}} + \sum_{t=\mathcal{T}_{g-1}}^{\mathcal{T}_g - 1} \eta C_1 x_t^{q-1} \geq x_{\mathcal{T}_{g-1}} + \eta C_1 (\mathcal{T}_g - \mathcal{T}_{g-1}) x_0^{q-1} (1+\zeta)^{(g-1)(q-1)}, \tag{E.13}$$

and we could bound the difference $x_{\mathcal{T}_g} - x_{\mathcal{T}_{g-1}}$ by the following formula,

$$x_{\mathcal{T}_g} - x_{\mathcal{T}_{g-1}} \leq x_{\mathcal{T}_g - 1} + \eta C_2 x_{\mathcal{T}_g}^{q-1} - x_{\mathcal{T}_{g-1}} \leq \zeta(1+\zeta)^{g-1} x_0 + \eta C_2 x_0^{q-1} (1+\zeta)^{g(q-1)}. \tag{E.14}$$

Combining the results from (E.13) and (E.14), we obtain that,

$$\mathcal{T}_g \leq \mathcal{T}_{g-1} + \frac{\zeta}{\eta C_1 x_0^{q-2}(1+\zeta)^{(g-1)(q-2)}} + \frac{(1+\zeta)^{q-1}C_2}{C_1}. \tag{E.15}$$

Taking a telescoping sum of the results of (E.15) from $g = 1$ to $g = g^*$ and by the fact that $T_v \leq \mathcal{T}_{g^*}$, we finally get (E.9). Now consider another side, similarly for $t = \mathcal{T}_1$, we have

$$x_{\mathcal{T}_1} \leq x_0 + \sum_{t=0}^{\mathcal{T}_1 - 1} \eta C_2 x_t^{q-1} \leq x_0 + \mathcal{T}_1 \eta C_2 x_0^{q-1}(1+\zeta)^{q-1}.$$

Substitute that $x_{\mathcal{T}_1} - x_0 \geq \zeta x_0$, we get

$$\mathcal{T}_1 \geq \frac{\zeta}{\eta C_2 x_0^{q-2}(1+\zeta)^{q-1}}. \tag{E.16}$$

For $g > 1$, similarly we could derive,

$$x_{\mathcal{T}_g} \leq x_{\mathcal{T}_{g-1}} + \sum_{t=\mathcal{T}_{g-1}}^{\mathcal{T}_g - 1} \eta C_2 x_t^{q-1} \leq x_{\mathcal{T}_{g-1}} + \eta C_2 (\mathcal{T}_g - \mathcal{T}_{g-1}) x_0^{q-1}(1+\zeta)^{g(q-1)} \tag{E.17}$$

and we could also lower bound the difference $x_{\mathcal{T}_g} - x_{\mathcal{T}_{g-1}}$ by

$$x_{\mathcal{T}_g} - x_{\mathcal{T}_{g-1}} \geq x_{\mathcal{T}_g} - x_{\mathcal{T}_{g-1}-1} - \eta C_2 x_{\mathcal{T}_{g-1}-1}^{q-1} \geq \zeta(1+\zeta)^{g-1}x_0 - \eta C_2 x_0^{q-1}(1+\zeta)^{(g-1)(q-1)}. \tag{E.18}$$

Combining the results from (E.17) and (E.18), we obtain that,

$$\mathcal{T}_g \geq \mathcal{T}_{g-1} + \frac{\zeta}{\eta C_2 x_0^{q-2}(1+\zeta)^{g(q-2)+1}} - \frac{1}{(1+\zeta)^{q-1}}. \tag{E.19}$$

Taking a telescoping sum of the results of (E.19) from $g = 1$ to $g = g^* - 1$ and by the fact that $T_v \geq \mathcal{T}_{g^*-1}$, we finally get (E.10). ■

Lemma E.9 *Suppose that a positive sequence x_t, $t \geq 0$ follows the iterative formula*

$$x_{t+1} = x_t + c_1 e^{-c_2 x_t}$$

for some $c_1, c_2 > 0$. Then it holds that

$$\frac{1}{c_2}\log(c_1 c_2 t + e^{c_2 x_0}) \leq x_t \leq c_1 e^{-c_2 x_0} + \frac{1}{c_2}\log(c_1 c_2 t + e^{c_2 x_0})$$

for all $t \geq 0$.

Proof [Proof of Lemma E.9] We first show the lower bound of x_t. Consider a continuous-time sequence $\underline{x}_t$, $t \geq 0$ defined by the integral equation with the same initialization.

$$\underline{x}_t = \underline{x}_0 + c_1 \cdot \int_0^t e^{-c_2 \underline{x}_\tau} \mathrm{d}\tau, \quad \underline{x}_0 = x_0. \tag{E.20}$$

Note that $\underline{x}_t$ is obviously an increasing function of t. Therefore we have

$$\begin{aligned}
\underline{x}_{t+1} &= \underline{x}_t + c_1 \cdot \int_t^{t+1} e^{-c_2 \underline{x}_\tau} \mathrm{d}\tau \\
&\leq \underline{x}_t + c_1 \cdot \int_t^{t+1} e^{-c_2 \underline{x}_t} \mathrm{d}\tau \\
&= \underline{x}_t + c_1 \exp(-c_2 \underline{x}_t)
\end{aligned}$$

for all $t \in \mathbb{N}$. Comparing the above inequality with the iterative formula of $\{x_t\}$, we conclude by the comparison theorem that $x_t \geq \underline{x}_t$ for all $t \in \mathbb{N}$. Note that (E.20) has an exact solution

$$\underline{x}_t = \frac{1}{c_2} \log(c_1 c_2 t + e^{c_2 x_0})$$

Therefore we have

$$x_t \geq \frac{1}{c_2} \log(c_1 c_2 t + e^{c_2 x_0})$$

for all $t \in \mathbb{N}$, which completes the first part of the proof. Now for the upper bound of x_t, we have

$$\begin{aligned}
x_t &= x_0 + c_1 \cdot \sum_{\tau=0}^{t-1} e^{-c_2 x_\tau} \\
&\leq x_0 + c_1 \cdot \sum_{\tau=0}^{t} e^{-\log(c_1 c_2 \tau + e^{c_2 x_0})} \\
&= x_0 + c_1 \cdot \sum_{\tau=0}^{t} \frac{1}{c_1 c_2 \tau + e^{c_2 x_0}} \\
&= x_0 + \frac{c_1}{e^{c_2 x_0}} + c_1 \cdot \sum_{\tau=1}^{t} \frac{1}{c_1 c_2 \tau + e^{c_2 x_0}} \\
&\leq x_0 + \frac{c_1}{e^{c_2 x_0}} + c_1 \cdot \int_0^t \frac{1}{c_1 c_2 \tau + e^{c_2 x_0}} \mathrm{d}\tau,
\end{aligned}$$

where the second inequality follows by the lower bound of x_t as the first part of the result of this lemma. Therefore we have

$$\begin{aligned}
x_t &\leq x_0 + \frac{c_1}{e^{c_2 x_0}} + \frac{1}{c_2} \log(c_1 c_2 t + e^{c_2 x_0}) - \frac{1}{c_2} \log(e^{c_2 x_0}) \\
&= c_1 e^{-c_2 x_0} + \frac{1}{c_2} \log(c_1 c_2 t + e^{c_2 x_0})
\end{aligned}$$

This finishes the proof. ■