ON THE FEATURE LEARNING IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The predominant success of diffusion models in generative modeling has spurred significant interest in understanding their theoretical foundations. In this work, we propose a feature learning framework aimed at analyzing and comparing the training dynamics of diffusion models with those of traditional classification models. Our theoretical analysis demonstrates that, under identical settings, diffusion models, due to the denoising objective, are encouraged to learn more balanced and comprehensive representations of the data. In contrast, neural networks with a similar architecture trained for classification tend to prioritize learning specific patterns in the data, often focusing on easy-to-learn features. To support these theoretical insights, we conduct several experiments on both synthetic and real-world datasets, which empirically validate our findings and highlight the distinct feature learning dynamics in diffusion models compared to classification.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021) have emerged as a powerful class of generative models for content synthesis and have demonstrated state-of-the-art generative performance in a variety of domains, such as computer vision (Dhariwal & Nichol, 2021; Peebles & Xie, 2023), acoustic (Kong et al., 2021; Chen et al., 2021) and biochemical (Hoogeboom et al., 2022; Watson et al., 2023). Recently, many works have employed (pre-trained) diffusion models to extract useful representations for tasks other than generative modelling, and demonstrated surprising capabilities in classical tasks such as image classification with little-to-no tuning (Mukhopadhyay et al., 2023; Xiang et al., 2023; Li et al., 2023a; Clark & Jaini, 2024; Yang & Wang, 2023; Jaini et al., 2024). Compared to discriminative models trained with supervised learning, diffusion models not only are able to achieve comparable recognition performance (Li et al., 2023a), but also demonstrate exceptional out-of-distribution transferablity (Li et al., 2023a; Jaini et al., 2024) and improved classification robustness (Chen et al., 2024c).

The significant representation learning power suggests diffusion models are able to extract meaningful features from training data. Indeed, the core of diffusion models is to estimate the data distribution through progressively denoising noisy inputs over several iterative steps. This inherently views data distribution as a composition of multiple latent features and therefore learning the data distribution corresponds to learning the underlying features. Nevertheless, it remains unclear

how feature learning happens during the training of diffusion models and whether the feature learning process is different to supervised learning.

Regardless of the ground-breaking success of diffusion models, the theoretical understanding is still in its infancy. Existing analysis on diffusion models has mostly focused on theoretical guarantees in terms of distribution estimation and sampling convergence. Several works have derived statistical estimation errors between distribution generated by diffusion models to ground-truth distribution (Oko et al., 2023; Zhang et al., 2024; Chen et al., 2023a), showing that diffusion models achieve a minimax optimal rate under certain assumptions on the true density (Oko et al., 2023; Zhang et al., 2024). Algorithmically, Li et al. (2023c); Han et al. (2024) study the estimation error of diffusion models trained with gradient descent using kernel methods. Shah et al. (2023); Gatmiry et al. (2024); Chen et al. (2024d) introduce algorithms based on diffusion models for learning Gaussian mixture models. In addition, given access to sufficiently accurate score estimation, Lee et al. (2022; 2023); Chen et al. (2023b); Li et al. (2023b) prove the convergence guarantees of sampling in (score-based) diffusion models. Despite showing provable guarantees for diffusion models, existing theories are limited to the generative aspects of diffusion models, namely distribution learning and sampling. To the best of our knowledge, no theoretical analysis is performed to elucidate the feature learning process in diffusion models.

Notations. We make use of the following notations throughout the paper. We use $\|\cdot\|$ to denote L_2 norm for vectors and Frobenius norm for matrices, unless mentioning otherwise. We use $O(\cdot), \Omega(\cdot), \Theta(\cdot), o(\cdot)$, $\omega(\cdot)$ for the big-O, big-Omega, big-Theta, small-o, small-omega notations. We write $\widetilde{O}(\cdot)$ to hid (poly)logarithmic factors and similar notations hold for $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$. For a binary condition $\mathcal{C}$, we let $\mathbb{1}(\mathcal{C}) = 1$ if $\mathcal{C}$ is true and $\mathbb{1}(\mathcal{C}) = 0$ otherwise.

1.1 Our main results

In this work, we develop a theoretical framework that studies feature learning dynamics of diffusion model and compares with classification. Inspired by the image data structure, we employ a multi-patch data distribution $\mathbf{x} = [\boldsymbol{\mu}_y, \boldsymbol{\xi}]$ for both classification and diffusion model training. We consider a two-class data setup with $y = \pm 1$ as the data label and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1} \in \mathbb{R}^d$ are two fixed orthogonal vectors, i.e., $\boldsymbol{\mu}_1 \perp \boldsymbol{\mu}_{-1}$, representing the signal. On the other hand, $\boldsymbol{\xi}$ is the label-independent noise, which is randomly sampled from a Gaussian distribution with standard deviation σ_ξ.

In order to elucidate the difference of feature learning dynamics for the two tasks, we adopt a *two-layer convolutional neural network* with *quadratic activation*. For diffusion model, we consider a weight-sharing setting for the first and second layer, which is commonly considered for analyzing autoencoders (Nguyen, 2021; Cui & Zdeborová, 2024). For classification, we fix the second layer weights to be ± 1, following Cao et al. (2022); Kou et al. (2023). In other words, the classifier can be viewed as attaching a fixed linear head to the intermediate layer of the diffusion model. Given a training dataset of n samples from the multi-patch data distribution, we use *gradient descent* to minimize the empirical logistic loss for classification and the DDPM loss (Ho et al., 2020) with expectation over the diffusion noise.

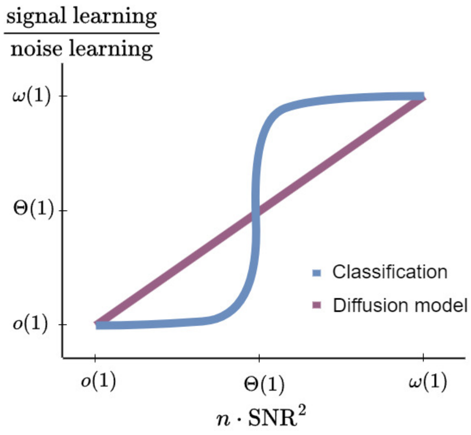


Figure 1: Illustration of the ratio of signal learning to noise learning when varying $n \cdot \mathrm{SNR}^2$, where $\mathrm{SNR} := \|\boldsymbol{\mu}\|/(\sigma_\xi \sqrt{d})$. We show diffusion model tends to study more balanced signal and noise while classification has a sharp phase transition and tends to focus on learning either signal or noise.

Under the above settings, we investigate the differences of feature learning dynamics between diffusion model and classification. We quantify the feature learning in terms of signal learning and noise learning, measured through the alignment between the network weights $\mathbf{w}$ to the directions of signal/noise respectively, i.e., $|\langle \mathbf{w}, \boldsymbol{\mu}_y \rangle|$, $|\langle \mathbf{w}, \boldsymbol{\xi} \rangle|$. We present the following (informal) results that compare the feature learning trajectories of the two learning paradigms.

Theorem 1.1 (Informal). *Let* $\mathrm{SNR} := \|\boldsymbol{\mu}\|/(\sigma_\xi \sqrt{d})$ *be the signal-to-noise ratio. We can show*

- *For* ***diffusion model****,* $|\langle \mathbf{w}, \boldsymbol{\mu}_y \rangle|, |\langle \mathbf{w}, \boldsymbol{\xi} \rangle|$ *exhibit linear growth initially and a neural network can find a stationary point that satisfies* $|\langle \mathbf{w}, \boldsymbol{\mu}_y \rangle| / |\langle \mathbf{w}, \boldsymbol{\xi} \rangle| = \Theta(n \cdot \mathrm{SNR}^2)$.

- *For* ***classification****,* $|\langle \mathbf{w}, \boldsymbol{\mu}_y \rangle|, |\langle \mathbf{w}, \boldsymbol{\xi} \rangle|$ *exhibit exponential growth initially and when* $n \cdot \mathrm{SNR}^2 \geq \beta$ *for some constant* $\beta > 1$, $|\langle \mathbf{w}, \boldsymbol{\mu}_y \rangle| / |\langle \mathbf{w}, \boldsymbol{\xi} \rangle| = \omega(1)$*, and when* $n \cdot \mathrm{SNR}^2 < 1/\beta$, $|\langle \mathbf{w}, \boldsymbol{\mu}_y \rangle| / |\langle \mathbf{w}, \boldsymbol{\xi} \rangle| = o(1)$.

Theorem 1.1 first highlights a difference in the learning speed during the early stage of training, where the growth rate is quadratic for classification and linear for diffusion model. As a result, the final learning outcomes can be largely different. Especially in the SNR region where $n \cdot \mathrm{SNR}^2 = \Theta(1)$, classification tends to be sensitive to the SNR value and will focus on learning either the signal $\boldsymbol{\mu}_y$ or the noise $\boldsymbol{\xi}$. In contrast, diffusion model learns both signal and noise with the same order. Such a claim is visualized in Figure 1.

We believe our framework represents the *first* attempt to systematically investigate feature learning within diffusion models, potentially uncovering novel insights into the less understood properties of diffusion models, such as the critical window (Sclocchi et al., 2024; Li & Chen, 2024), shape bias (Jaini et al., 2024), classification robustness (Chen et al., 2024c), among others.

2 Problem setting

This section introduces the problem settings for both diffusion model and classification, including the data model, neural network functions as well as training objectives and algorithm.

Definition 2.1 (Data distribution)**.** *Each data sample consists of two patches, as* $\mathbf{x} = [\mathbf{x}^{(1)\top}, \mathbf{x}^{(2)\top}]^\top$*, where each patch is generated as follows:*

- *Sample* $y \in \{-1, 1\}$ *uniformly with* $\mathbb{P}(y = -1) = \mathbb{P}(y = 1) = 1/2$*.*
- *Given two orthogonal signal vectors* $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1}$*, with* $\boldsymbol{\mu}_1 \perp \boldsymbol{\mu}_{-1}$*, we set* $\mathbf{x}^{(1)} = \boldsymbol{\mu}_y$*, i.e.,* $\mathbf{x}^{(1)} = \boldsymbol{\mu}_1$ *if* $y = 1$ *and* $\mathbf{x}^{(1)} = \boldsymbol{\mu}_{-1}$ *if* $y = -1$*. For simplicity, we assume* $\|\boldsymbol{\mu}_1\| = \|\boldsymbol{\mu}_{-1}\| = \|\boldsymbol{\mu}\|$*.*
- *Set* $\mathbf{x}^{(2)} = \boldsymbol{\xi}$ *where* $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2(\mathbf{I} - \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top\|\boldsymbol{\mu}_1\|^{-2} - \boldsymbol{\mu}_{-1}\boldsymbol{\mu}_{-1}^\top\|\boldsymbol{\mu}_{-1}\|^{-2}))$*.*

Such a multi-patch data model mimics the properties of image data where each image is composed of several patches. Only some patches are relevant to the class label while others become background noise. The noise patch $\boldsymbol{\xi}$ is generated from the Gaussian distribution such that it is orthogonal to the signal vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1}$ for simplicity of analysis. This data model has been employed in several existing works (Cao et al., 2022; Chen et al., 2022; Kou et al., 2023; Allen-Zhu & Li, 2023). A difference in our model is that we have two signal vectors that are orthogonal, instead of a single vector with signal patch being $y\boldsymbol{\mu}$ as in the previous studies. We also highlight that although we only consider two patches for simplicity, our analysis can be easily extended to multi-patch data.

Neural network functions. We consider two-layer convolutional-type neural networks for both diffusion model and classification. For **diffusion model**, we consider neural network with quadratic activation and shared first-layer and second-layer weights:

$$\boldsymbol{f}(\mathbf{W}, \mathbf{x}) = \Big[\boldsymbol{f}_1(\mathbf{W}, \mathbf{x}^{(1)})^\top, \boldsymbol{f}_2(\mathbf{W}, \mathbf{x}^{(2)})^\top\Big]^\top \in \mathbb{R}^{2d},$$
$$\text{where} \quad \boldsymbol{f}_p\big(\mathbf{W}, \mathbf{x}^{(p)}\big) = \frac{1}{\sqrt{m}}\sum_{r=1}^m \langle \mathbf{w}_r, \mathbf{x}^{(p)}\rangle^2 \mathbf{w}_r, \quad p = 1, 2$$

where m denotes the network width and r represents the neuron index. That is, we decouple the training of neural network at each diffusion time step with separate weight parameters, a strategy also adopted in (Shah et al., 2023) for simplicity of analysis.

For **classification**, we consider a similar neural network with quadratic activation where second-layer weights are fixed to be ± 1 (instead of $\mathbf{w}_r$):

$$f(\mathbf{W}, \mathbf{x}) = F_1(\mathbf{W}_1, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x}),$$
$$\text{where} \quad F_j(\mathbf{W}, \mathbf{x}) = \frac{1}{m}\sum_{r=1}^m \langle \mathbf{w}_{j,r}, \mathbf{x}^{(1)}\rangle^2 + \frac{1}{m}\sum_{r=1}^m \langle \mathbf{w}_{j,r}, \mathbf{x}^{(2)}\rangle^2.$$

We remark that the use of polynomial activation, such as quadratic, cubic and ReLU with polynomial smoothing is not uncommon in existing theoretical works (Cao et al., 2022; Jelassi & Li, 2022; Zou et al., 2023; Huang et al., 2023; Meng et al., 2023). The aim is to better elucidate the separation between signal and noise learning dynamics in the process of training.

Training objectives and algorithm. For **diffusion model**, the goal is to estimate the distribution of input images through the process of gradual denoising. In particular, we employ the objective of denoising diffusion probabilistic model (DDPM) (Ho et al., 2020). We let $\mathbf{x}_0 = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}]^\top \in \mathbb{R}^{2d}$

to denote input image. For a given diffusion time step $t \in [0, T]$, we sample $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\epsilon}_t$ for $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ and some noise schedule coefficients $\{\alpha_t, \beta_t\}_{t=0}^T$. In this work, we do not make any assumption over the noise schedule.

The aim is estimate the mean of the posterior distribution of the noise $\boldsymbol{\epsilon}_t$ conditioned on $\mathbf{x}_t$. This is achieved by training a neural network f to predict the noise added at each step t. The DDPM loss is given by $\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}_t, t} \| f(\mathbf{x}_t) - \boldsymbol{\epsilon}_t \|^2$ up to some re-scaling (Ho et al., 2020). We consider a finite-sample setup given by the training images $\{\mathbf{x}_i\}_{i=1}^n$ sampled according to Definition 2.1 and thus the empirical DDPM loss at time step t becomes

$$L_F(\mathbf{W}_t) = \frac{1}{2n} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\epsilon}_{t,i}} \| \boldsymbol{f}(\mathbf{W}_t, \mathbf{x}_{t,i}) - \boldsymbol{\epsilon}_{t,i} \|^2 = \frac{1}{2n} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\epsilon}_{t,i}} \| \boldsymbol{f}(\mathbf{W}_t, \alpha_t \mathbf{x}_{0,i} + \beta_t \boldsymbol{\epsilon}_{t,i}) - \boldsymbol{\epsilon}_{t,i} \|^2,$$

where we let $\mathbf{x}_{0,i} = \mathbf{x}_i$ and $\mathbf{x}_{t,i} = \alpha_t \mathbf{x}_{0,i} + \beta_t \boldsymbol{\epsilon}_{t,i}$. Unlike (Han et al., 2024), where each sample i is associated with a single noise $\boldsymbol{\epsilon}_{t,i} \sim \mathcal{N}(0, \mathbf{I})$, we here consider taking the expectation over the noise distribution, which aligns with the practical setting where multiple noises are sampled for each input data. We use gradient descent to train diffusion model starting from random Gaussian initialization $\mathbf{w}_{r,t}^0 \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$ as $\mathbf{w}_{r,t}^{k+1} = \mathbf{w}_{r,t}^k - \eta \nabla_{\mathbf{w}_{r,t}} L_F(\mathbf{W}_t^k)$, where we let superscript k to denote iteration index and subscript r, t to denote neuron index and diffusion timestep respectively.

For **classification**, we minimize the empirical logistic loss over the training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$,

$$L_S(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^n \ell\big(y_i f(\mathbf{W}, \mathbf{x}_i)\big), \quad \ell(z) = \log\big(1 + \exp(-z)\big).$$

The same as diffusion model, we use gradient descent to train the neural network starting from random Gaussian initialization $\mathbf{w}_{j,r}^0 \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$.

3 Main results

Our main results are based on the following conditions.

Condition 3.1. *Suppose the following holds.*

1. *Dimension d is sufficiently large with $d = \widetilde{\Omega}\big(\max\{\|\boldsymbol{\mu}\|^2, n^2 m \sigma_\xi^{-1} \|\boldsymbol{\mu}\|, n^4 m^2, n^4 m^8 \sigma_\xi^{-2}\}\big)$.*
2. *The sample size n and network width m satisfies $n, m = \widetilde{\Omega}(1)$.*
3. *The standard deviation of initialization σ_0 is chosen such that $\widetilde{O}(n^2 m \sigma_\xi^{-1} d^{-1}) \leq \sigma_0 \leq \widetilde{O}\big(\min\{\|\boldsymbol{\mu}\|^{-1}, \sigma_\xi^{-1} d^{-1/2}, m^{-6} d^{-1/2}\}\big)$.*
4. *The learning rate η satisfies $\eta \leq \widetilde{O}\big(\min\{m \|\boldsymbol{\mu}\|^{-2}, nm\sigma_0 \sigma_\xi^{-1} d^{-1/2}, nm\sigma_\xi^{-2} d^{-1}\}\big)$.*
5. *The noise coefficients for diffusion model satisfy $\alpha_t, \beta_t = \Theta(1)$.*

Condition 3.1 requires d to be large to ensure learning in an over-parameterized setting. Furthermore, we only require the network width and sample size to be lower bounded by some logarithmic factors, in order to achieve certain concentration properties of neurons and samples. The upper bound on the initialization σ_0 is to ensure random initialization does not significantly affect the signal and noise learning dynamics. The lower bound on σ_0 is required to bound the noise inner product at initialization for properly minimizing the training loss of classification. We can ensure the lower bound is valid by imposing further conditions on the dimension d. The learning rate η is chosen sufficiently small for the convergence analysis for the classification. Lastly for diffusion model, we consider the constant order for α_t, β_t. We believe that the constant order of α_t and β_t reflects the standard practice in diffusion models, such as (Ho et al., 2020). Similar assumptions are commonly employed to derive learning guarantees for classification (Chatterji & Long, 2021; Cao et al., 2022; Kou et al., 2023).

Based on Condition 3.1, we present the main results for diffusion model (Theorem 3.1) and classification (Theorem 3.2).

Theorem 3.1 (Diffusion model). *Under Condition 3.1, along the training trajectory of diffusion model, there exists a stationary point $\mathbf{W}_t^*$, i.e., $\nabla_{\mathbf{w}_{r,t}} L_F(\mathbf{W}_t^*) = 0$ that satisfies* (1) $\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle = \Theta(\langle \mathbf{w}_{r',t}^*, \boldsymbol{\mu}_{j'} \rangle)$, (2) $\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle = \Theta(\langle \mathbf{w}_{r',t}^*, \boldsymbol{\xi}_{i'} \rangle)$*, and* (3) *for all $j = \pm 1, r \in [m]$, $i \in [m]$,*

$$|\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle| / |\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle| = \Theta(n \cdot \mathrm{SNR}^2).$$

Theorem 3.1 states that the training of diffusion model encourages performing balanced signal and noise learning, i.e., the neurons are sharing the same order in the directions of signals and noise. Notably, the ratio between signal and noise learning is governed by the SNR, with a stationary magnitude as $n \cdot \mathrm{SNR}^2$.

Theorem 3.2 (Classification)**.** *Let* $T_\mu = \widetilde{\Theta}(\eta^{-1} m \|\boldsymbol{\mu}\|^{-2})$ *and* $T_\xi = \widetilde{\Theta}(\eta^{-1} n m \sigma_\xi^{-2} d^{-1})$ *and suppose* $\delta > 0$*. Under Condition 3.1, there exist two absolute constants* $\overline{C} > \underline{C} > 0$ *such that with probability at least* $1 - \delta$*, it satisfies that:*

- *When* $n \cdot \mathrm{SNR}^2 \geq \overline{C}$*, there exists* $0 \leq k \leq T_\mu$ *such that training loss converges with* $L_S(\mathbf{W}^k) \leq 0.1$ *and*
$$\max_r |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle| \geq 2, \, \forall j = \pm 1, \qquad \max_{j,r,i} |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i \rangle| = o(1).$$
- *When* $n \cdot \mathrm{SNR}^2 \leq \underline{C}$*, there exists* $0 \leq k \leq T_\xi$ *such that training loss converges with* $L_S(\mathbf{W}^k) \leq 0.1$ *and*
$$\max_r |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle| \geq 1, \, \forall i \in [n], \qquad \max_{j,r,y} |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_y \rangle = o(1).$$

Theorem 3.2 establishes a sharp phase transition between signal and noise learning for the case of classification. The transition is precisely determined by $n \cdot \mathrm{SNR}^2$. That is, when $n \cdot \mathrm{SNR}^2 \geq \overline{C}$ for some constant $\overline{C} > 0$, the neural network learns signal to achieve small training loss. On the contrary, when $n \cdot \mathrm{SNR}^2 \leq \underline{C}$ for some constant $\underline{C} \in (0, \overline{C})$, the neural network overfits noise in order to converge. Using standard techniques, such as in (Cao et al., 2022), we can show signal and noise learning corresponds to the regime of benign and harmful overfitting respectively. To the best of our knowledge, this is the first result that shows separation under the constant of $n \cdot \mathrm{SNR}^2$.

Diffusion model learns balanced features while classification learn dominant features. Comparing the learning outcomes of diffusion model and classification, we reveal a critical difference that *diffusion model learn more balanced features depending on the SNR conditions, while classification is prone to learning either signal or noise predominately*. This can be best understood in the case of $n \cdot \mathrm{SNR}^2 = \Theta(1)$. By Theorem 3.2, we have either signal learning or noise dominating the learning process in classification, while Theorem 3.1 suggests signal and noise learning are in the same order in diffusion models. The theoretical findings corroborate the empirical observations that the neural network trained for classification is prone to overly rely on learning a specific pattern that is easier to learn, a process known as shortcut learning (Geirhos et al., 2020). Meanwhile, diffusion models tend to learn low-frequency, global patterns (Jaini et al., 2024), which helps to improve the classification robustness (Chen et al., 2024b;c).

4 Proof overview

In summary, for diffusion model, both the mean-squared loss and the joint training of two layers impose significant challenges for the analysis. Thus, we decouple the training into two stages, and characterize the stationary points based on the derived results at the end of first-stage. For classification, the two-stage analysis is similar as in (Cao et al., 2022; Kou et al., 2023) where the first stage learns signal or noise vector sufficiently fast and the second stage shows convergence in the training loss where the learned scale difference in the first stage is maintained. However for classification analysis, we highlight two critical differences compared to existing works (Cao et al., 2022; Kou et al., 2023; Meng et al., 2024), i.e., a constant $n \cdot \mathrm{SNR}^2$ condition and quadratic activation.

4.1 Diffusion model

We first simplify the DDPM loss by taking the expectation with respect to the added diffusion noise:

$$L_F(\mathbf{W}_t) = d + \frac{1}{2n} \sum_{i=1}^n \sum_{p=1}^2 \Big(\underbrace{\frac{1}{m} \mathbb{E}_{\boldsymbol{\epsilon}_{t,i}} \Big\| \sum_{r=1}^m \langle \mathbf{w}_{r,t}, \mathbf{x}_{t,i}^{(p)} \rangle^2 \mathbf{w}_{r,t} \Big\|^2}_{I_1} - \underbrace{\frac{4\alpha_t \beta_t}{\sqrt{m}} \sum_{r=1}^m \|\mathbf{w}_{r,t}\|^2 \langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}^{(p)} \rangle}_{I_2} \Big),$$

where we recall for $p = 1, 2$, $\mathbf{x}_{t,i}^{(p)} = \alpha_t \mathbf{x}_{0,i}^{(p)} + \beta_t \boldsymbol{\epsilon}_{t,i}^{(p)}$, with $\mathbf{x}_{0,i}^{(1)} = \boldsymbol{\mu}_{y_i}$ and $\mathbf{x}_{0,i}^{(2)} = \boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_{t,i}^{(1)}, \boldsymbol{\epsilon}_{t,i}^{(2)} \sim \mathcal{N}(0, \mathbf{I})$. We further simplify I_1 in Lemma E.2 (in Appendix). We make several remarks in order. First, I_1 corresponds to a regularization term that regulates the magnitude of each neuron as well as the alignment among neurons. I_2 corresponds to the main learning term. Second, in the current setting, when either α_t or β_t vanishes, the loss is dominated by the regularization term such that neural network converges towards zero.

First stage. In the first stage, where all the key quantities, including signal and noise inner products, norm of the weights and cross-neuron inner product remain close to their respective initialization, we can show the growth of the signal and noise inner products is approximately linear:

$$\begin{aligned}\langle \mathbf{w}_{r,t}^{k+1}, \boldsymbol{\mu}_j \rangle &= \langle \mathbf{w}_{r,t}^{k}, \boldsymbol{\mu}_j \rangle + \tfrac{4\eta\alpha_t\beta_t|\mathcal{S}_j|}{n\sqrt{m}} \|\mathbf{w}_{r,t}^k\|^2 \|\boldsymbol{\mu}_j\|^2 + \widetilde{O}(\sigma_0^5 d^2 \|\boldsymbol{\mu}_j\|^3), \\ \langle \mathbf{w}_{r,t}^{k+1}, \boldsymbol{\xi}_i \rangle &= \langle \mathbf{w}_{r,t}^{k}, \boldsymbol{\xi}_i \rangle + \tfrac{4\eta\alpha_t\beta_t}{n\sqrt{m}} \|\mathbf{w}_{r,t}^k\|^2 \|\boldsymbol{\xi}_i\|^2 + \widetilde{O}(\sigma_0^5 \sigma_\xi^3 d^{7/2}),\end{aligned}$$

This allows to simplify the analysis for the initial iterations and we have the following scale at the end of the first stage.

Lemma 4.1. *Under Condition 3.1, there exists an iteration $T_1 = \max\{T_\mu, T_\xi\}$, where $T_\mu = \widetilde{\Theta}(\sqrt{m}\sigma_0^{-1} d^{-1} \|\boldsymbol{\mu}\|^{-1} \eta^{-1})$ and $T_\xi = \widetilde{\Theta}(n\sqrt{m}\sigma_0^{-1}\sigma_\xi^{-1} d^{-3/2}\eta^{-1})$ such that for all $k \leq T_1$, (1) $|\langle \mathbf{w}_{r,t}^k, \boldsymbol{\mu}_j\rangle| = \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|)$ (2) $|\langle \mathbf{w}_{r,t}^k, \boldsymbol{\xi}_i\rangle| = \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$ and (3) $\|\mathbf{w}_{r,t}^k\|^2 = \Theta(\sigma_0^2 d)$ for all $r \in [m], j = \pm 1, i \in [n]$. for all $j = \pm 1$, $r \in [m], i \in [n]$. Furthermore, we can show*

- $\langle \mathbf{w}_{r,t}^{T_1}, \boldsymbol{\mu}_j \rangle = \Theta(\langle \mathbf{w}_{r',t}^{T_1}, \boldsymbol{\mu}_{j'} \rangle)$,
- $\langle \mathbf{w}_{r,t}^{T_1}, \boldsymbol{\xi}_i \rangle = \Theta(\langle \mathbf{w}_{r',t}^{T_1}, \boldsymbol{\xi}_{i'} \rangle)$,
- $|\langle \mathbf{w}_{r,t}^{T_1}, \boldsymbol{\mu}_j \rangle| / |\langle \mathbf{w}_{r,t}^{T_1}, \boldsymbol{\xi}_i \rangle| = \Theta(n \cdot \mathrm{SNR}^2)$,

for all $j, j' = \pm 1, r, r' \in [m], i, i' \in [n]$.

Lemma 4.1 verifies that at the end of the first stage, due to the linear dynamics, all the key inner products and norms are still close to their initialization. In the meantime, all the neurons are concentrated in terms of signal and noise learning and the ratio is precisely determined by $n \cdot \mathrm{SNR}^2$. This is critically different compared to the case of classification where signal and noise inner product exhibits exponential growth as we show later and thus allows a clear scale difference at the end of the first stage.

Second stage. The second stage shows that there exists a stationary point such that the relative scale at the end of the first stage is preserved, namely concentration of neurons $\langle \mathbf{w}_{r,t}^k, \boldsymbol{\mu}_j\rangle = \Theta(\langle \mathbf{w}_{r',t}^k, \boldsymbol{\mu}_{j'}\rangle)$, $\langle \mathbf{w}_{r,t}^k, \boldsymbol{\xi}_i\rangle = \Theta(\langle \mathbf{w}_{r',t}^k, \boldsymbol{\xi}_{i'}\rangle)$ for all $j, j' = \pm 1, r, r' \in [m], i, i' \in [n]$ as well as the ratio of signal to noise learning respects the order of $n \cdot \mathrm{SNR}^2$. Towards this end, we first identify the required conditions for a stationary point $\mathbf{w}_{r,t}^*$ under the concentration of neurons, which leads to

$$\begin{aligned}&\sqrt{m}\Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle^5 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle^3 \|\mathbf{w}_{r,t}^*\|^2 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle \\ &\quad + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle^2 \|\mathbf{w}_{r,t}^*\|^2 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle^3 \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\mu}_j\|^2 + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle \|\boldsymbol{\mu}_j\|^2 \Big) \\ &= \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle^2 + \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\mu}_j\|^2 \Big), \\ &\sqrt{m}\Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle^5 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle^3 \|\mathbf{w}_{r,t}^*\|^2 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle \\ &\quad + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle^2 \|\mathbf{w}_{r,t}^*\|^2 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle + \frac{1}{n} \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle^3 \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\xi}_i\|^2 + \frac{1}{n} \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle \|\boldsymbol{\xi}_i\|^2 \Big), \\ &= \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle^2 + \frac{1}{n} \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\xi}_i\|^2 \Big) + \widetilde{O}(n\sqrt{m} d^{-1/2}).\end{aligned}$$

We then separately analyze the three SNR conditions: (1) $n \cdot \mathrm{SNR}^2 = \Theta(1)$, (2) $n \cdot \mathrm{SNR}^2 = \widetilde{\Omega}(1)$ and (3) $n^{-1} \cdot \mathrm{SNR}^{-2} = \widetilde{\Omega}(1)$. We show

- When $n \cdot \mathrm{SNR}^2 = \Theta(1)$, $|\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle| / |\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle| = \Theta(1)$.
- When $n \cdot \mathrm{SNR}^2 = \widetilde{\Omega}(1)$, $\Omega(1) = |\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j\rangle| / |\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i\rangle| \leq \Theta(n \cdot \mathrm{SNR}^2)$.

- When $n^{-1} \cdot \mathrm{SNR}^{-2} = \widetilde{\Omega}(1)$, $\Omega(1) = |\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle| / |\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle| \leq \Theta(n^{-1} \cdot \mathrm{SNR}^{-2})$.

This demonstrates that if $n \cdot \mathrm{SNR}^2$ is a constant, then the relative magnitude of signal learning to noise learning is also a constant. If $n \cdot \mathrm{SNR}^2$ is lower bounded by a log order, then the ratio cannot be larger than the order of $n \cdot \mathrm{SNR}^2$. On the other hand, if $n^{-1} \cdot \mathrm{SNR}^{-2}$ is lower bounded by a log order, then the ratio cannot be smaller than the order of $n \cdot \mathrm{SNR}^2$. This suggests there exists a stationary point that preserves the scale.

4.2 CLASSIFICATION

Let $\mathcal{S}_y := \{i \in [n] : y_i = y\}$ for $y = \pm 1$ and $\ell_i'^k = \ell'\big(y_i f(\mathbf{W}^k, \mathbf{x})\big)$. Then we can rewrite the gradient descent updates in terms of the signal and noise inner products:

$$\langle \mathbf{w}^{k+1}_{j,r}, \boldsymbol{\mu}_y \rangle = \langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_y \rangle - \frac{\eta |\mathcal{S}_y|}{nm} \ell_i'^k \langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_y \rangle j y \|\boldsymbol{\mu}\|^2 = (1 - \frac{\eta |\mathcal{S}_y| \|\boldsymbol{\mu}\|^2}{nm} \ell_i'^k j y) \langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_y \rangle, \quad (1)$$

$$\langle \mathbf{w}^{k+1}_{j,r}, \boldsymbol{\xi}_i \rangle = \langle \mathbf{w}^k_{j,r}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \ell_i'^k \langle \mathbf{w}^k_{j,r}, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2 j y_i - \frac{\eta}{nm} \sum_{i' \neq i} \ell_{i'}'^k \langle \mathbf{w}^k_{j,r}, \boldsymbol{\xi}_{i'} \rangle j y_{i'} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle, \quad (2)$$

for all $j, y = \pm 1, r \in [m], i \in [n]$. The iterative updates of signal inner product suggests that for any $j = \pm 1$, $\mathbf{w}_{j,r}$ specializes the learning of $\boldsymbol{\mu}_j$ because by the fact that $\ell_i'^k < 0$, $|\langle \mathbf{w}^{k+1}_{j,r}, \boldsymbol{\mu}_y \rangle| = (1 - \frac{\eta |\mathcal{S}_y| \|\boldsymbol{\mu}\|^2}{nm} \ell_i'^k j y) |\langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_y \rangle| > |\langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_y \rangle|$ only when $j = y$. For the noise inner product, the growth is dominated by the second term where $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| = \widetilde{O}(d^{-1/2}) \|\boldsymbol{\xi}_i\|^2$ is significantly smaller. Therefore, we can show $|\langle \mathbf{w}^{k+1}_{j,r}, \boldsymbol{\xi}_i \rangle|$ grows only for $j = y_i$ while for $j = -y_i$, the magnitude cannot increase relative to the scale of initialization. Next, we decompose the analysis into two stages.

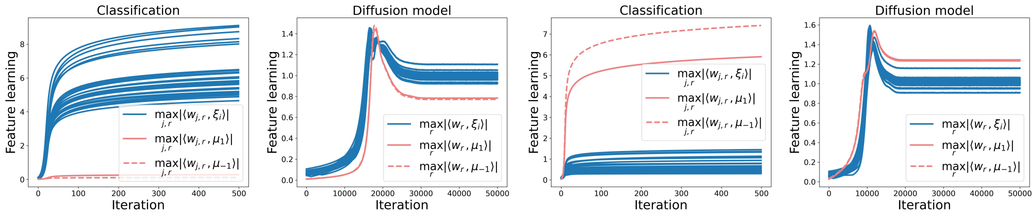
First stage. In the first stage before the maximum of signal and noise inner product reaches constant order, the loss derivatives can be lower bounded by an absolute constant, i.e., $|\ell_i'^k| \geq C_\ell$, for all $k \leq T_1$. As a result, both signal and noise inner product can grow exponentially and the relative growth rates are precisely characterized by the condition on $n \cdot \mathrm{SNR}^2$. A constant order of difference in the growth rate is sufficient to ensure at the end of first stage, there exists a scale separation in signal and noise learning, where either signal or noise inner product reaches a constant order.

Different to existing analysis that only shows maximum inner product reaches constant order (Cao et al., 2022), we also show the average inner product reach constant order at the same time. Such a stronger result is required for the analysis under the constant order of $n \cdot \mathrm{SNR}^2$, which reduces the required iteration number in the second stage by an order of m.

For the case of signal learning, we can readily obtain the same bound for the average inner product and maximum inner product based on (1). Nevertheless, this becomes challenging for noise learning due to the cross term in (2). Thus, we rely on an anti-concentration result that lower bounds the $|\langle \mathbf{w}^0_{j,r}, \boldsymbol{\xi}_i \rangle|$ at initialization, which is sufficient to ensure the sign invariance across the whole optimization process, i.e., $\mathrm{sign}(\langle \mathbf{w}^k_{j,r}, \boldsymbol{\xi}_i \rangle) = \mathrm{sign}(\langle \mathbf{w}^0_{j,r}, \boldsymbol{\xi}_i \rangle)$ for all k. The following lemma provides a formal characterization at the end of first stage.

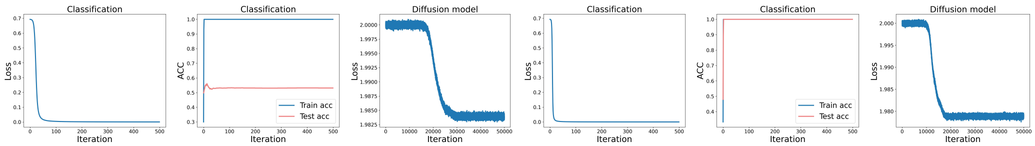
Lemma 4.2. *Under Condition 3.1: (1) When* $n \cdot \mathrm{SNR}^2 = \Omega(1)$*, there exists* $T_1 = \widetilde{\Theta}(\eta^{-1} m \|\boldsymbol{\mu}\|^{-2})$*, such that* $\frac{1}{m} \sum_{r=1}^m |\langle \mathbf{w}^{T_1}_{j,r}, \boldsymbol{\mu}_j \rangle| \geq 2$ *for all* $j = \pm 1$ *and* $\max_{j,r,i} |\langle \mathbf{w}^{T_1}_{j,r}, \boldsymbol{\xi}_i \rangle| = o(1)$*. (2) When* $n^{-1} \cdot \mathrm{SNR}^{-2} = \Omega(1)$*, there exists* $T_1 = \widetilde{\Theta}(\eta^{-1} n m \sigma_\xi^{-2} d^{-1})$ *such that* $\frac{1}{m} \sum_{r=1}^m |\langle \mathbf{w}^{T_1}_{y_i,r}, \boldsymbol{\xi}_i \rangle| \geq 4$ *for all* $i \in [n]$ *and* $\max j, r, y |\langle \mathbf{w}^{T_1}_{j,r}, \boldsymbol{\mu}_y \rangle| = o(1)$.

Second stage. Lemma 4.2 already shows a scale difference in the signal and noise learning. In the second stage, we follow the standard analysis (Cao et al., 2022; Kou et al., 2023) to show the loss converges while the scale difference is maintained. Because $n \cdot \mathrm{SNR}^2$ can be a constant, we require to carefully bound the loss derivatives in the second stage particularly for establishing the upper bound for $|\langle \mathbf{w}^k_{j,r}, \boldsymbol{\xi}_i \rangle|$ when $n \cdot \mathrm{SNR}^2 = \Omega(1)$. The naïve bound $\max_i |\ell_i'^k| \leq \max_i |\ell_i^k| \leq n L_S(\mathbf{W}^k)$ used in (Cao et al., 2022) no longer works as it introduces an additional factor of n. To provide a tighter bound, we show the ratio of loss derivatives in the case of $n \cdot \mathrm{SNR}^2 = \Omega(1)$, i.e., $|\ell_i'^k| / |\ell_{i'}'^k| \leq C_1$ for all $i, i' \in [n]$ with $y_i = y_{i'}$, $k \geq T_1$, where $C_1 > 0$ is a constant. This is possible because the network output is dominated by the signal, which is shared across samples with the same label. This allows to bound $\max_i |\ell_i'^k| = \Theta\big(|\mathcal{S}_{y_{i^*}}|^{-1} \sum_{i \in \mathcal{S}_{y_{i^*}}} |\ell_i'^k|\big) \leq \Theta(L_S(\mathbf{W}^k))$.



Low SNR ($n \cdot \mathrm{SNR}^2 = 0.75$) High SNR ($n \cdot \mathrm{SNR}^2 = 6.75$)

Figure 2: Experiments on the synthetic dataset with both low SNR ($n \cdot \mathrm{SNR}^2 = 0.75$) and high SNR ($n \cdot \mathrm{SNR}^2 = 6.75$). In the low SNR setting, we see noise learning quickly dominates signal learning for the classification task and in the high SNR setting, signal learning quickly dominates noise learning. Meanwhile diffusion model converges to a stationary point that with signal-to-noise learning ratio respects the order of $n \cdot \mathrm{SNR}^2$.



$n \cdot \mathrm{SNR}^2 = 0.75$ $n \cdot \mathrm{SNR}^2 = 6.75$

Figure 3: Experiments on the synthetic dataset with both low SNR ($n \cdot \mathrm{SNR}^2 = 0.75$) and high SNR ($n \cdot \mathrm{SNR}^2 = 6.75$). The loss for both diffusion model and classification and training/test accuracy for classification.

5 NUMERICAL EXPERIMENTS

We conduct both synthetic and real-world experiments to verify the difference between diffusion model and classification in terms of signal and noise learning.

5.1 SYNTHETIC EXPERIMENT

We follow the data distribution in Definition 2.1 to generate a synthetic dataset for both diffusion model and classification. Specifically, we set data dimension $d = 1000$ and let $\boldsymbol{\mu}_1 = [\mu, 0, \cdots, 0] \in \mathbb{R}^d$ and $\boldsymbol{\mu}_{-1} = [0, \mu, 0, \cdots, 0] \in \mathbb{R}^d$. We sample the noise patch $\boldsymbol{\xi}_i \sim \mathcal{N}(0, \mathbf{I}_d)$, $i \in [n]$ (i.e., $\sigma_\xi = 1$). We set sample size and network width to be $n = 30$ and $m = 20$ and initialize the weights to be Gaussian with a standard deviation $\sigma_0 = 0.001$. Such a setting is aligned with the Condition 3.1. We vary the choice of μ to create two problem settings: (1) low SNR with $\mu = 5$, which leads to $n \cdot \mathrm{SNR}^2 = 0.75$ and (2) high SNR with $\mu = 15$, which leads to $n \cdot \mathrm{SNR}^2 = 6.75$. We use the same two-layer networks introduced in Section 2. For classification, we set a learning rate of $\eta = 0.1$ and train for 500 iterations. We also measure the in-distribution test accuracy with 3000 test samples. For diffusion model, instead of using the expected loss, we train the DDPM loss by averaging the added diffusion noise, following the standard training of diffusion model. In particular, for each sample, we samples $n_\epsilon = 2000$ noise at each iteration and the loss is calculated by taking an average over the noise. For the noise coefficients, we consider a time $t = 0.2$ and set $\alpha_t = \exp(-t) = 0.82$ and $\beta_t = \sqrt{1 - \exp(-2t)} = 0.57$.

In Figure 2, we compare signal and noise learning dynamics (visualized through maximum signal and noise inner product) between classification and diffusion model. In Figure 3, we also include training loss convergence for both the tasks as well as training and test accuracy for classification.

We see both classification and diffusion model are able to converge in loss, although diffusion model only finds a stationary point. In the low SNR setting, classification is able to perfectly fit the training samples with a 100% classification accuracy. However because it primarily focuses on learning noise, the generalization is poor with a test accuracy of around 50%. For the high-SNR case, both training and test sets can be perfectly classified due to the signal learning.

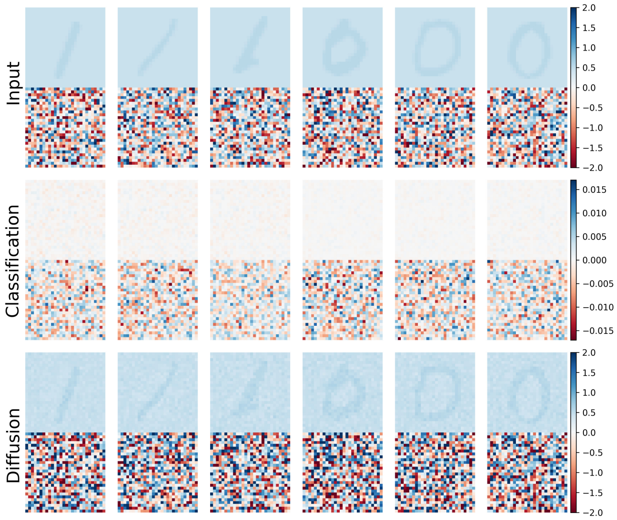


Figure 4: Experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.1$. (First row): Test Noisy-MNIST images; (Second row): Illustration of input gradient, i.e., $\nabla_{\mathbf{x}} F_{+1}(\mathbf{W}, \mathbf{x})$ when $y = 1$ and $\nabla_{\mathbf{x}} F_{-1}(\mathbf{W}, \mathbf{x})$ when $y = 0$. (Third row): denoised image from diffusion model. In this low-SNR case, we see classification tends to predominately learn noise while diffusion learns both signals and noise.

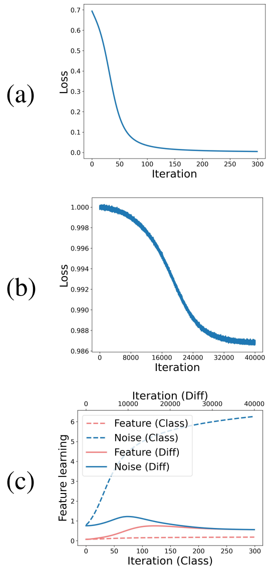


Figure 5: Experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.1$. (a) Train loss for classification. (b) Train loss for diffusion model. (c) Feature learning dynamics.

Regarding feature learning in classification, noise learning quickly dominates signal learning by exhibiting a significant larger growth in the first stage (up to around 20 iterations). This ensures noise learning to reach a constant order while signal learning is still very small. The second stage corresponds to loss convergence and the growth of both signal and noise learning is upper bounded by a log order. For diffusion model, in the first stage, where loss does not materially change, both signal and noise learning increases linearly which remains on the same order. In the second stage where loss significantly decreases, signal and noise learning grow at an exponential rate and in the third stage, due to the regularization term on the weight, noise and signal reach a stationary point that preserves the scale of $n \cdot \mathrm{SNR}^2$.

5.2 Real-world experiment

In addition, we also verify the feature learning comparisons on MNIST dataset (Lecun et al., 1998). In order to better control the SNR, we create a noisy version of MNIST dataset (and called Noisy-MNIST) where we view each original MNIST image as a clean signal patch and then we concatenate a standard Gaussian noise patch with the same size, i.e., 28×28. In addition, we scale the signal patch by a constant, which we denote as $\widetilde{\mathrm{SNR}}$. Because the noise scale is fixed, higher $\widetilde{\mathrm{SNR}}$ corresponds to higher SNR. Some sample images with $\widetilde{\mathrm{SNR}} = 0.1$ are shown in the first row of Figure 4. We select 50 samples each from digit 0 and 1 respectively (i.e., $n = 100$). We consider the same neural networks as in the synthetic example, where we set $m = 100$ and initialize the weights with $\sigma_0 = 0.01$. For diffusion model, we choose the same α_t, β_t as in the synthetic experiment. In the main paper, we present the results for $\widetilde{\mathrm{SNR}} = 0.1$, which corresponds to low SNR setting. Figure 5 shows both classification and diffusion model converges in losses. In addition, we also plot the signal and noise learning dynamics in Figure 5(c). Because each image is composed of unique signal $\boldsymbol{\mu}_i$ and noise patch $\boldsymbol{\xi}_i$ for $i \in [n]$, we measure the signal and noise learning by computing $\frac{1}{n}\sum_{i=1}^n \max_r |\langle \mathbf{w}_r, \boldsymbol{\mu}_i \rangle|$ and $\frac{1}{n}\sum_{i=1}^n \max_r |\langle \mathbf{w}_r, \boldsymbol{\xi}_i \rangle|$ respectively. We notice that due to the low SNR, when convergence, noise learning in classification dominates signal learning while diffusion model learns a more balanced ratio. This corroborates our theoretical findings.

To visualize the patterns learned by the neural networks, for classification, we use a similar idea of Grad-CAM (Selvaraju et al., 2020) by probing into the gradient of output with respect to the input. In particular, for samples of digit 0, we plot the gradient of negative function output, i.e., $\nabla_{\mathbf{x}} F_{-1}(\mathbf{W}, \mathbf{x})$ and for samples of digit 1, we plot $\nabla_{\mathbf{x}} F_{+1}(\mathbf{W}, \mathbf{x})$. In the second row of Figure 4, the gradients with respect to six test images suggest that classification learns significantly more noise compared to the signal patch. On the other hand, for diffusion model, we first add diffusion noise to the input images and use the network to predict the added noise. Then we plot the predicted input using the formula $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \beta_t \hat{\boldsymbol{\epsilon}}(\mathbf{x}_t))/\alpha_t$, where $\hat{\boldsymbol{\epsilon}}(\mathbf{x}_t)$ denotes the predicted diffusion noise. In the third row of Figure 4, we see diffusion models learn both the signal and noise. In Appendix B.1, we also experiment on a high-SNR setting with $\widetilde{\mathrm{SNR}} = 0.5$ where we see the reverse pattern that classification predominately learns noise rather than signal while diffusion model still balances the learning of both signal and noise.

6 Conclusions and Discussions

This work presents a novel theoretical framework for analyzing the feature learning dynamics in diffusion models, marking the first such contribution to the existing literature. Through rigorous analysis, we demonstrate that diffusion models inherently promote the learning of more balanced features, in contrast to traditional classification methods, which tend to prioritize certain features over others. This suggests models trained for classification may be more sensitive to the change in SNR compared to diffusion models. Consequently, this may explain the inherent adversarial robustness of diffusion model in downstream applications, such as classification (Li et al., 2023a; Chen et al., 2024c;b), because such perturbations are less likely to significantly alter the feature learning outcomes of diffusion models compared to classification models.

Although our study focuses on two-patch data setup, the framework can be adapted to accommodate more complex data settings. For example, our analysis can be extended to multi-feature data distributions, where certain features appear more frequently (Zou et al., 2023) or possess larger norms than others (Lu et al., 2024). Such extensions may potentially uncover deeper insights into the mechanisms of feature learning in more realistic scenarios. We hypothesize that, despite the infrequent occurrence or smaller norm of these features, diffusion models can effectively learn them due to the nature of the denoising objective. This insight has significant implications for downstream tasks, such as out-of-distribution classification, where only these rare or weak features may be present.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Uuf2q9TfXGA`.

Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B Grosse. Regularized linear autoencoders recover the principal components, eventually. *Advances in Neural Information Processing Systems*, 33:6971–6981, 2020.

Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022.

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.

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.

Haoxuan Chen, Yinuo Ren, Lexing Ying, and Grant M Rotskoff. Accelerating diffusion models with parallel sampling: Inference at sub-linear time complexity. *arXiv:2405.15986*, 2024a.

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Your diffusion model is secretly a certifiably robust classifier. *arXiv:2402.02316*, 2024b.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *International Conference on Machine Learning*, 2024c.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023a.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=NsMLjcFaO8O`.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023b.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv:2404.18893*, 2024d.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=MaYzugDmQV`.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36, 2024.

Hugo Cui and Lenka Zdeborová. High-dimensional asymptotics of denoising autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes S Fischer, Vincent Tao Hu, and Bjorn Ommer. Diffusion models and representation learning: A survey. *arXiv:2407.00783*, 2024.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv:2404.18869*, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Shivam Gupta, Linda Cai, and Sitan Chen. Faster diffusion-based sampling with randomized midpoints: Sequential and parallel. *arXiv:2406.00924*, 2024.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=h8GeqOxtd4`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

Wei Huang, Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki. Graph neural networks provably benefit from structural information: A feature learning perspective. *arXiv:2306.13926*, 2023.

Xunpeng Huang, Difan Zou, Hanze Dong, Yi Zhang, Yi-An Ma, and Tong Zhang. Reverse transition kernel: A flexible framework to accelerate diffusion inference. *arXiv preprint arXiv:2405.16387*, 2024.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rmg0qMKYRQ.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pp. 9965–10040. PMLR, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Kevin Kögler, Alexander Shevchenko, Hamed Hassani, and Marco Mondelli. Compression of structured data with autoencoders: Provable benefit of nonlinearities and depth. *arXiv:2402.05013*, 2024.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In *International Conference on Machine Learning*, pp. 17615–17659. PMLR, 2023.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *International Conference on Machine Learning*, pp. 3560–3569. PMLR, 2019.

Y Lecun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985. PMLR, 2023.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *International Conference on Computer Vision*, pp. 2206–2217, 2023a.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv:2306.09251*, 2023b.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27942–27954. PMLR, 21–27 Jul 2024a.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv:2402.07802*, 2024b.

Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27474–27498. PMLR, 2024.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2097–2127, 2023c.

Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rate. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wYmvN3sQpG.

Xuran Meng, Yuan Cao, and Difan Zou. Per-example gradient regularization improves learning signals from noisy data. *arXiv:2303.17940*, 2023.

Xuran Meng, Difan Zou, and Yuan Cao. Benign Overfitting in Two-Layer ReLU Convolutional Neural Networks for XOR Data. In *International Conference on Machine Learning*. PMLR, 2024.

Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification. *arXiv:2307.08702*, 2023.

Phan-Minh Nguyen. Analysis of feature learning in weight-tied autoencoders via the mean field lens. *arXiv:2102.08373*, 2021.

Thanh V Nguyen, Raymond KW Wong, and Chinmay Hegde. Benefits of jointly training autoencoders: An improved neural tangent kernel analysis. *IEEE Transactions on Information Theory*, 67(7):4669–4692, 2021.

Reza Oftadeh, Jiayi Shen, Zhangyang Wang, and Dylan Shell. Eliminating the invariance on the loss landscape of linear autoencoders. In *International Conference on Machine Learning*, pp. 7405–7413. PMLR, 2020.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, pp. 4195–4205, 2023.

Arnu Pretorius, Steve Kroon, and Herman Kamper. Learning dynamics of linear denoising autoencoders. In *International Conference on Machine Learning*, pp. 4141–4150. PMLR, 2018.

Maria Refinetti and Sebastian Goldt. The dynamics of representation learning in shallow, non-linear autoencoders. In *International Conference on Machine Learning*, pp. 18499–18519. PMLR, 2022.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *arXiv:2402.16991*, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2020.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the DDPM objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

Aleksandr Shevchenko, Kevin Kögler, Hamed Hassani, and Marco Mondelli. Fundamental limits of two-layer autoencoders, and achieving them with gradient methods. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31151–31209. PMLR, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252. PMLR, 2023.

Harald Steck. Autoencoders that don't overfit towards the identity. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19598–19608. Curran Associates, Inc., 2020.

Yuqing Wang, Ye He, and Molei Tao. Evaluating the design space of diffusion-based generative models. *arXiv:2406.12839*, 2024.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *International Conference on Computer Vision*, pp. 15802–15812, 2023.

Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *International Conference on Computer Vision*, pp. 18938–18949, 2023.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60134–60178. PMLR, 2024.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *International Conference on Computer Vision*, pp. 5729–5739, 2023.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. In *International Conference on Machine Learning*, pp. 43423–43479. PMLR, 2023.

Appendix Contents

A Related works

Theoretical analysis of diffusion model. Existing theoretical guarantees for diffusion models focus on distribution estimation and sampling. For distribution estimation, Oko et al. (2023) show that diffusion model can achieve a nearly minimax optimal estimation error where the true density is defined over a bounded Besov space. The minimax optimality of diffusion model is later proved to hold for a more general class of densities that are sub-Gaussian and satisfy certain degree of smoothness (Zhang et al., 2024). Further, Oko et al. (2023); Chen et al. (2023a) prove that when the density is supported on a low-dimensional subspace, diffusion model avoids curse of dimensionality with an estimation rate that only depends on the intrinsic dimension. Besides statistical guarantees, several studies approach the distribution learning problem from an algorithmic perspective. Shah et al. (2023) shows gradient descent can provably learn the distribution of well-separated spherical Gaussian mixtures. Some other works study the distribution estimation of diffusion model trained by gradient descent dynamics, under the choice of a random feature model (Li et al., 2023c) and neural tangent kernel regime (Han et al., 2024). In addition, Gatmiry et al. (2024); Chen et al. (2024d) introduce efficient algorithms based on diffusion models for estimating the density of more general Gaussian mixture model. Finally, Wang et al. (2024) analyze the convergence of denoising score matching objective under gradient descent.

Apart from distribution estimation aspect of diffusion model, many works study the convergence guarantees for diffusion model sampling. Several results (Lee et al., 2022; 2023; Chen et al., 2023b; Li et al., 2023b) have shown (score-based) diffusion model attains polynomial convergence rate under sufficiently accurate score estimation. Recent literature has also aimed to accelerate the convergence in sampling via strategies such as consistency training (Song et al., 2023; Li et al., 2024b), advanced design of the reverse transition kernel (Huang et al., 2024), higher-order approximation (Li et al., 2024a) and parallelization (Chen et al., 2024a; Gupta et al., 2024). In addition, Li & Chen (2024) theoretically verify the critical window of feature emergence during the sampling process assuming access to accurate score estimates.

Theoretical analysis on (denoising) autoencoders. Diffusion models can be viewed as multi-level denoising autoencoders (Xiang et al., 2023). There exists extensive research on theoretical guarantees for autoencoders *without denoising*. Most of the works focus on linear autoencoders (Kunin et al., 2019; Oftadeh et al., 2020; Steck, 2020; Bao et al., 2020) while only a few works analyzed non-linear autoencoders, either in the lazy training regime (Nguyen et al., 2021) or the mean-field regime Nguyen (2021). Training dynamics of non-linear autoencoders has also been studied under population gradient descent (Shevchenko et al., 2023; Kögler et al., 2024) and online gradient descent (Refinetti & Goldt, 2022). On the other hand, training dynamics of denoising autoencoder has been studied with a linear network (Pretorius et al., 2018) and in the high-dimensional asymptotic limit (Cui & Zdeborová, 2024). Thus, even for (denoising) autoencoders, feature learning dynamics is not well-understood.

Diffusion model for representation learning. Apart from the generative applications, diffusion models have been leveraged for representation learning. The intermediate representation of a pre-trained diffusion model is shown to possess significant discriminative power. Such an representation is useful for downstream tasks such as classification (Mukhopadhyay et al., 2023; Xiang et al., 2023; Li et al., 2023a; Clark & Jaini, 2024; Yang & Wang, 2023), semantic segmentation (Baranchuk et al., 2022; Zhao et al., 2023; Yang & Wang, 2023). Moreover many works have found intriguing properties of diffusion models used as classifier, including its ability to understand shape bias (Jaini et al., 2024) and improved adversarial robustness (Chen et al., 2024c). For more detailed exposition, we refer to the recent survey on this matter (Fuest et al., 2024).

B Additional experimental results

This section includes additional experiment results, supplementary to the results in the main text.

B.1 High SNR setting on Noisy-MNIST

Here we include experiment results when $\widetilde{\mathrm{SNR}} = 0.5$, which corresponds to the high SNR setting. The experiment settings are exactly the same as in the main experiment. Figure 7 shows both classification and diffusion model converge in terms of objective. In addition, we see the high

SNR encourages classification to learn primarily the signal while ignoring the noise. In contrast, diffusion model still learns both signal and noise to relatively the same order. Figure 6 suggests that classification learns more signal compared to noise while diffusion model still learns more balanced signal and noise. We also plot classification accuracy for both the low and high SNR cases. In the low-SNR case, because classification predominately learns noise, the generalization is poor with test accuracy around 50%. Conversely in the high-SNR case, where the model is able to learn signals, the classification demonstrates effective generalization with nearly 100% test accuracy.

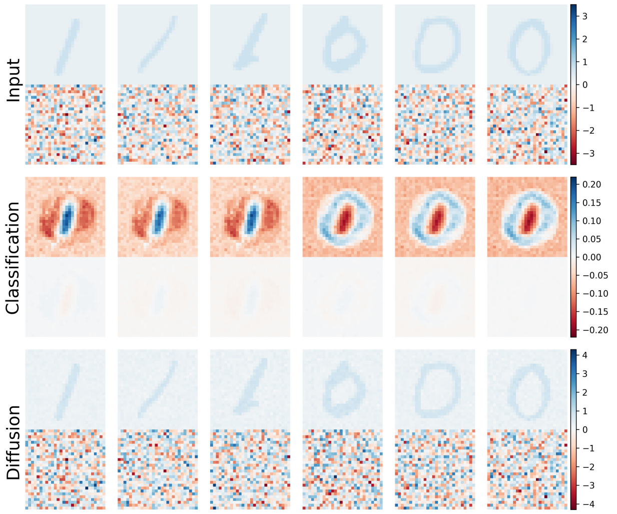


Figure 6: Experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.5$. (First row): Test Noisy-MNIST images; (Second row): Illustration of input gradient, i.e., $\nabla_{\mathbf{x}} F_{+1}(\mathbf{W}, \mathbf{x})$ when $y = 1$ and $\nabla_{\mathbf{x}} F_{-1}(\mathbf{W}, \mathbf{x})$ when $y = 0$. (Third row): denoised image from diffusion model. In this high-SNR case, we see classification tends to predominately learn signals while diffusion learns both signal and noise.

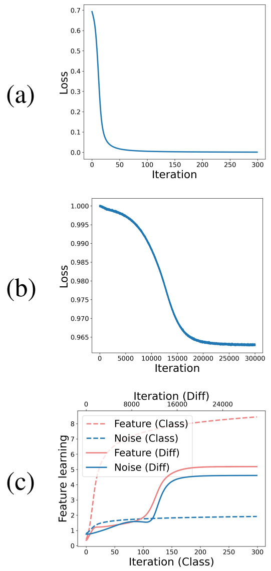
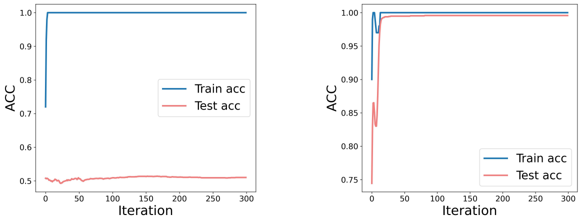


Figure 7: Experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.5$. (a) Train loss for classification. (b) Train loss for diffusion model. (c) Feature learning dynamics.



(a) ACC ($\widetilde{\mathrm{SNR}} = 0.1$)

(b) ACC ($\widetilde{\mathrm{SNR}} = 0.5$)

Figure 8: Classification accuracy on (a) low-SNR and (b) high-SNR noisy MNIST datasets. This demonstrates that when classification focuses on learning noise (as in the low-SNR case), the test accuracy hovers around 50%, thus suggesting failure to generalize. In contrast, when classification focuses on learning signals (as in the high-SNR case), classification generalizes effectively, achieving near-perfect accuracy.

B.2 Experiments with additional diffusion time step

Here we also test on additional diffusion time step for learning on noisy-MNIST dataset. In particular, we consider $t = 0.8$, which gives $\alpha_t = \exp(-t) = 0.45$ and $\beta_t = \sqrt{1 - \exp(-2t)} = 0.89$. We include the illustrations of denoised images as well as loss convergence and feature learning dynamics in Figure 9, 10, 11, 12. We see despite with a larger scale of added diffusion noise, diffusion model still learn both signals and noise unlike for the case of classification.

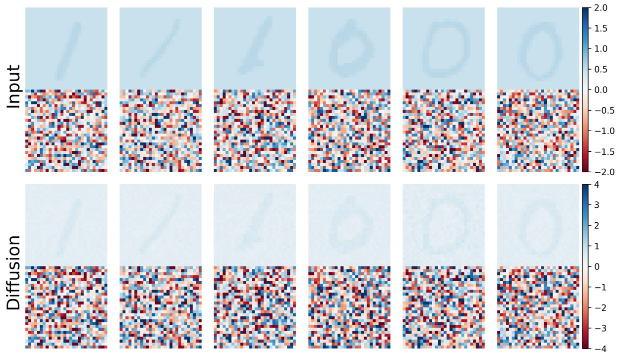


Figure 9: Additional experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.1$ and diffusion $t = 0.8$. (First row): Test Noisy-MNIST images; (Second row): denoised image from diffusion model. We see diffusion still learns both signals and noise even with large diffusion time step.

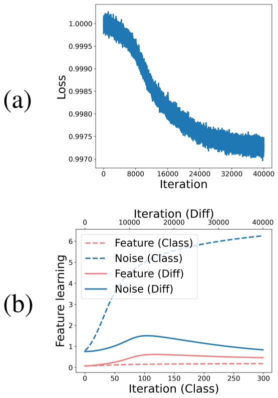


Figure 10: Additional experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.1$ and $t = 0.8$. (a) Train loss for diffusion model. (c) Feature learning dynamics.

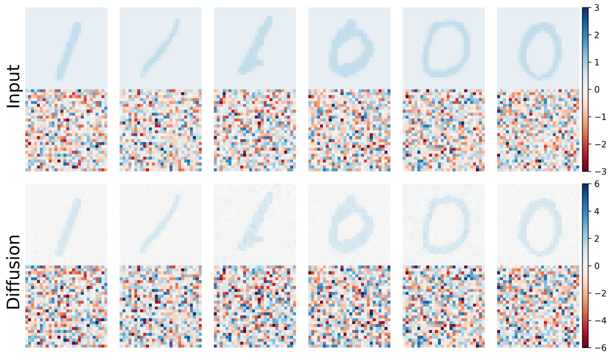


Figure 11: Additional experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.5$ and diffusion $t = 0.8$. (First row): Test Noisy-MNIST images; (Second row): denoised image from diffusion model. We see diffusion still learns both signals and noise even with large diffusion time step.

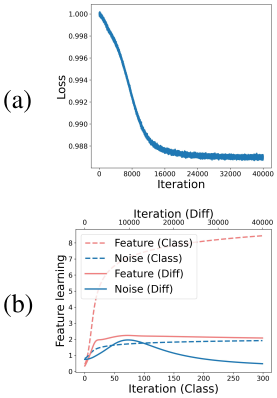


Figure 12: Additional experiments on Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.5$ and = $t = 0.8$. (a) Train loss for diffusion model. (c) Feature learning dynamics.

C Preliminary lemmas

Recall we define $\mathcal{S}_1 = \{i \in [n] : y_i = 1\}$ and $\mathcal{S}_{-1} = \{i \in [n] : y_i = -1\}$.

Lemma C.1. *Given arbitrary $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\frac{n}{2}\big(1 - \widetilde{O}(n^{-1/2})\big) \leq |\mathcal{S}_1|, |\mathcal{S}_{-1}| \leq \frac{n}{2}\big(1 + \widetilde{O}(n^{-1/2})\big)$$

Proof of Lemma C.1. The proof is the same as in (Cao et al., 2022; Kou et al., 2023) and we include here for completeness. Because $|\mathcal{S}_1| = \sum_{i=1}^n \mathbb{1}(y_i = 1)$ and $|\mathcal{S}_{-1}| = \sum_{i=1}^n \mathbb{1}(y_i = -1)$ and $\mathbb{P}(y_i = 1) = \mathbb{P}(y_i = -1) = 1/2$ for all $i \in [n]$, then $\mathbb{E}|\mathcal{S}_1| = \mathbb{E}|\mathcal{S}_{-1}| = n/2$. By Hoeffding's inequality, for arbitrary $a > 0$,

$$\mathbb{P}(||\mathcal{S}_{\pm 1}| - n/2| \geq a) \leq 2\exp(-2a^2 n^{-1}).$$

Setting $a = \sqrt{n \log(4/\delta)/2}$ and taking union bound, we have with probability at least $1 - \delta$,

$$\Big||\mathcal{S}_{\pm 1}| - \frac{n}{2}\Big| \leq \sqrt{\frac{n\log(4/\delta)}{2}}.$$

Hence the proof is complete. □

Lemma C.2. *Given arbitrary $\delta > 0$, with probability at least $1 - \delta$,*

$$\sigma_\xi^2 d(1 - \widetilde{O}(d^{-1/2})) \leq \|\boldsymbol{\xi}_i\|^2 \leq \sigma_\xi^2 d(1 + \widetilde{O}(d^{-1/2}))$$
$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \leq 2\sigma_\xi^2 \sqrt{d \log(4n^2/\delta)}$$

for all $i, i' \in [n]$.

Proof of Lemma C.2. The proof is the same as in (Cao et al., 2022; Kou et al., 2023) and we include here for completeness. By Bernstein's inequality, with probability at least $1 - \delta/(2n)$, we have

$$|\|\boldsymbol{\xi}_i\|^2 - \sigma_\xi^2 d| = O(\sigma_\xi^2 \sqrt{d \log(4n/\delta)}),$$

which shows the first result.

For the second claim, we can show by Bernstein's inequality, with probability at least $1 - \delta/(2n^2)$ that for any $i \neq i'$

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \leq 2\sigma_\xi^2 \sqrt{d \log(4n^2/\delta)}$$

Then we apply union bound to show the results hold for all $i, i' \in [n]$. □

D Classification

We track the inner product dynamics during the training of supervised classification to elucidate the signal learning and noise learning. We first write the gradient descent dynamics as follows.

$$\begin{aligned}
\mathbf{w}_{j,r}^{k+1} &= \mathbf{w}_{j,r}^k - \eta \nabla_{\mathbf{w}_{j,r}} L_S(\mathbf{W}^k) \\
&= \mathbf{w}_{j,r}^k - \frac{\eta}{nm}\sum_{i=1}^n \ell_i'^k \langle \mathbf{w}_{j,r}^k, \mathbf{x}_i^{(1)} \rangle j y_i \mathbf{x}_i^{(1)} - \frac{\eta}{nm}\sum_{i=1}^n \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i \rangle j y_i \boldsymbol{\xi}_i \\
&= \mathbf{w}_{j,r}^k - \frac{\eta}{nm}\sum_{i \in \mathcal{S}_1} \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_1 \rangle j \boldsymbol{\mu}_1 + \frac{\eta}{nm}\sum_{i \in \mathcal{S}_{-1}} \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_{-1} \rangle j \boldsymbol{\mu}_{-1} - \frac{\eta}{nm}\sum_{i=1}^n \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i \rangle j y_i \boldsymbol{\xi}_i
\end{aligned}$$

Here we restate the Condition 3.1 specific for the case of supervised classification.

Condition D.1. *Suppose that*

1. *Dimension d satisfies $d = \widetilde{\Omega}(\max\{n^2 m \sigma_\xi^{-1} \|\boldsymbol{\mu}\|, n^4 m\})$.*

2. *Training sample and network width satisfy* $m = \Omega(\log(n/\delta)), n = \Omega(\log(m/\delta))$.

3. *The initialization variation* σ_0 *satisfies* $\widetilde{O}(n^2 m \sigma_\xi^{-1} d^{-1}) \leq \sigma_0 \leq \widetilde{O}(\min\{\|\boldsymbol{\mu}\|^{-1}, \sigma_\xi^{-1} d^{-1/2}\})$.

4. *The learning rate satisfies* $\eta \leq \widetilde{O}(\min\{m\|\boldsymbol{\mu}\|^{-2}, nm\sigma_0\sigma_\xi^{-1}d^{-1/2}, nm\sigma_\xi^{-2}d^{-1}\})$

The lower bound on σ_0 is required for the noise memorization setting where we need to control the lower bound for the noise inner product at initialization. Thus to ensure the lower bound σ_0 is valid, we require further conditions on the dimension d apart from $d = \widetilde{\Omega}(n^2)$.

D.1 Useful lemmas

We first provide a lemma that bound the inner product at initialization.

Lemma D.1 (Cao et al. (2022)). *Suppose* $\delta > 0$ *and that* $d = \Omega(\log(mn/\delta)), m = \Omega(\log(1/\delta))$, *then with probability at least* $1 - \delta$,

$$|\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_{j'} \rangle| \leq \sqrt{2\log(8m/\delta)}\sigma_0\|\boldsymbol{\mu}\|$$
$$|\langle \mathbf{w}_{j',r}^0, \boldsymbol{\xi}_i \rangle| \leq 2\sqrt{\log(8mn/\delta)}\sigma_0\sigma_\xi\sqrt{d}$$

for all $j, j' \in \{\pm 1\}, r \in [m], i \in [n]$. *In addition,*

$$\max_{r\in[m]} |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_{j'} \rangle| \geq \sigma_0\|\boldsymbol{\mu}\|/2,$$
$$\max_{r\in[m]} |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\xi}_i \rangle| \geq \sigma_0\sigma_\xi\sqrt{d}/4$$

for all $j, j' \in \{\pm 1\}, i \in [n]$.

We decompose the weights into its signal components and noise components.

Lemma D.2. *The weight can be decomposed as*

$$\mathbf{w}_{j,r}^k = \mathbf{w}_{j,r}^0 + \zeta_1^k \boldsymbol{\mu}_1 + \zeta_{-1}^k \boldsymbol{\mu}_{-1} + \sum_{i=1}^n \rho_{j,r,i}^k \|\boldsymbol{\xi}_i\|^{-2}\boldsymbol{\xi}_i$$

where the noise coefficients $\rho_{j,r,i}^k$ *satisfy* $\rho_{j,r,i}^0 = 0$ *and*

$$\rho_{j,r,i}^{k+1} = \rho_{j,r,i}^k - \frac{\eta}{nm}\ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i \rangle j y_i \|\boldsymbol{\xi}_i\|^2$$

for all $j = \pm 1$, $r \in [m]$ *and* $i \in [n]$.

Proof of Lemma D.2. The proof follows from (Cao et al., 2022; Kou et al., 2023). First, we recall the gradient descent update as

$$\begin{aligned}
\mathbf{w}_{j,r}^{k+1} &= \mathbf{w}_{j,r}^k - \frac{\eta}{nm}\sum_{i\in\mathcal{S}_1} \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_1 \rangle j \boldsymbol{\mu}_1 + \frac{\eta}{nm}\sum_{i\in\mathcal{S}_{-1}} \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_{-1} \rangle j \boldsymbol{\mu}_{-1} - \frac{\eta}{nm}\sum_{i=1}^n \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i \rangle j y_i \boldsymbol{\xi}_i \\
&= \mathbf{w}_{j,r}^0 - \frac{\eta}{nm}\sum_{s=0}^k \sum_{i\in\mathcal{S}_1} \ell_i'^k \langle \mathbf{w}_{j,r}^s, \boldsymbol{\mu}_1 \rangle j \boldsymbol{\mu}_1 + \frac{\eta}{nm}\sum_{s=0}^k \sum_{i\in\mathcal{S}_{-1}} \ell_i'^k \langle \mathbf{w}_{j,r}^s, \boldsymbol{\mu}_{-1} \rangle j \boldsymbol{\mu}_{-1} \\
&\quad - \frac{\eta}{nm}\sum_{s=0}^k \sum_{i=1}^n \ell_i'^k \langle \mathbf{w}_{j,r}^s, \boldsymbol{\xi}_i \rangle j y_i \boldsymbol{\xi}_i.
\end{aligned}$$

By the data model, we have with probability 1, the vectors are linearly independent and thus the decomposition is unique with

$$\rho_{j,r,i}^k = -\frac{\eta}{nm}\sum_{s=0}^k \ell_i'^k \langle \mathbf{w}_{j,r}^s, \boldsymbol{\xi}_i \rangle j y_i \|\boldsymbol{\xi}_i\|^2$$

Then writing out the iterative update for $\rho_{j,r,i}^k$ completes the proof. □

Lemma D.3. *Let $x \sim \mathcal{N}(0, \sigma^2)$. Then $\mathbb{P}(|x| \leq c) \leq \operatorname{erf}\left(\frac{c}{\sqrt{2}\sigma}\right) \leq \sqrt{1 - \exp(-\frac{2c^2}{\pi\sigma^2})}$.*

Proof of Lemma D.3. The probability density function for x is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{x^2}{2\sigma^2}).$$

Then we know that

$$\mathbb{P}(|x| \leq c) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-c}^{c} \exp(-\frac{x^2}{2\sigma^2}) dx.$$

By the definition of erf function

$$\operatorname{erf}(c) = \frac{2}{\sqrt{\pi}} \int_{0}^{c} \exp(-x^2) dx,$$

and variable substitution yields

$$\operatorname{erf}(\frac{c}{\sqrt{2}\sigma}) = \frac{1}{\sqrt{2\pi}\sigma} \int_{0}^{c} \exp(-\frac{x^2}{2\sigma^2}) dx.$$

Therefore, we first conclude $\mathbb{P}(|x| \leq c) = 2\operatorname{erf}(\frac{c}{\sqrt{2}\sigma})$. Next, by the inequality $\operatorname{erf}(x) \leq \sqrt{1 - \exp(-4x^2/\pi)}$, we obtain the desired result. □

D.2 Scale of inner products

We first derive a global bound for the growth of inner products until convergence. To this end, we let $T^* = \eta^{-1}\mathrm{poly}(\|\boldsymbol{\mu}\|^{-1}, \sigma_\xi^{-2} d^{-1}, \sigma_0^{-1}, n, m, d)$ be the maximum number of iterations considered and let $\alpha = 2\log(T^*)$. We also denote $\beta \coloneqq 3\max_{j,r,i,y}\{|\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_y\rangle|, |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\xi}_i\rangle|\}$. Then from Lemma D.1 and from Condition D.1, we can bound

$$3\max\{\sigma_0\|\boldsymbol{\mu}\|/2, \sigma_0\sigma_\xi\sqrt{d}/4\} \leq \beta \leq 1/C \tag{3}$$

for some sufficiently large constant $C > 0$.

Proposition D.1. *Under Condition D.1, for all $0 \leq k \leq T^*$, we can bound*

$$|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle|, |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i\rangle|, |\rho_{y_i,r,i}^k| \leq \alpha, \tag{4}$$

$$|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_{-j}\rangle| \leq \beta, \tag{5}$$

$$|\langle \mathbf{w}_{-y_i,r}^k, \boldsymbol{\xi}_i\rangle|, |\rho_{-y_i,r,i}^k| \leq \beta + 12\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha \tag{6}$$

for all $i \in [n]$, $r \in [m]$ and $j = \pm 1$.

We will prove the bound by induction and we first derive several intermediate lemmas as follows.

Lemma D.4. *Suppose results in Proposition D.1 hold at iteration k, then we have $F_j(\mathbf{W}_j^k, \mathbf{x}_i) \leq 0.5$ for all $i \in [n]$, $j \neq y_i$.*

Proof of Lemma D.4. Recall that

$$\begin{aligned} F_j(\mathbf{W}_j^k, \mathbf{x}_i) &= \frac{1}{m}\sum_{r=1}^{m}\left(\langle \mathbf{w}_{j,r}^k, \mathbf{x}_i^{(1)}\rangle^2 + \langle \mathbf{w}_{j,r}^k, \mathbf{x}_i^{(2)}\rangle^2\right) \\ &= \frac{1}{m}\sum_{r=1}^{m}\left(\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_{y_i}\rangle^2 + \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle^2\right) \\ &\leq \beta^2 + \left(\beta + 12\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha\right)^2 \\ &\leq 0.5 \end{aligned}$$

where the second last inequality is by (5) and (6). The last inequality is by Condition D.1 such that $\beta \leq 1/C \leq 0.25$ and $d \geq 144n^2\alpha^2\log(4n^2/\delta)$. □

Lemma D.5. *Suppose results in Proposition D.1 hold at iteration k, then we have*

$$|\langle \mathbf{w}^k_{j,r} - \mathbf{w}^0_{j,r}, \boldsymbol{\xi}_i \rangle - \rho^k_{j,r,i}| \le 4\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha,$$

for all $j = \pm 1, r \in [m], i \in [n]$.

Proof. By Lemma D.2, we recall the decomposition as

$$\mathbf{w}^k_{j,r} = \mathbf{w}^0_{j,r} + \zeta^k_1 \boldsymbol{\mu}_1 + \zeta^k_{-1} \boldsymbol{\mu}_{-1} + \sum_{i=1}^n \rho^k_{j,r,i} \|\boldsymbol{\xi}_i\|^{-2} \boldsymbol{\xi}_i.$$

By the orthogonality, we can show

$$\langle \mathbf{w}^k_{j,r}, \boldsymbol{\xi}_i \rangle = \langle \mathbf{w}^0_{j,r}, \boldsymbol{\xi}_i \rangle + \rho^k_{j,r,i} + \sum_{i \neq i'} \rho^k_{j,r,i'} \|\boldsymbol{\xi}_{i'}\|^{-2} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle$$

By Lemma C.2 and suppose $d = \Omega(\log(n/\delta))$, then $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \|\boldsymbol{\xi}_i\|^{-2} \le 4\sqrt{\log(4n^2/\delta) d^{-1}}$. Thus we have

$$|\langle \mathbf{w}^k_{j,r} - \mathbf{w}^0_{j,r}, \boldsymbol{\xi}_i \rangle - \overline{\rho}^k_{j,r,i}| \le 4\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha,$$

where we use the upper bound on $|\overline{\rho}^k_{j,r,i}| \le \alpha$. □

Lemma D.6. *For any* $r \in [m], j, y = \pm 1$, *we have* $\mathrm{sign}(\langle \mathbf{w}^0_{j,r}, \boldsymbol{\mu}_y \rangle) = \mathrm{sign}(\langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_y \rangle)$ *for all* $0 \le k \le T^*$.

Proof of Lemma D.6. We prove the results by induction. First, it is clear at $k = 0$, the results are satisfied. Then suppose there exists an iteration $\widetilde{k}$ such that $\mathrm{sign}(\langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_y \rangle) = \mathrm{sign}(\langle \mathbf{w}^0_{j,r}, \boldsymbol{\mu}_y \rangle)$ holds for all $k \le \widetilde{k} - 1$, we show the sign invariance also holds at $\widetilde{k}$. Recall the gradient descent update as

$$\begin{aligned} \mathbf{w}^{k+1}_{j,r} &= \mathbf{w}^k_{j,r} - \frac{\eta}{nm} \sum_{i \in \mathcal{S}_1} \ell'^k_i \langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_1 \rangle j \boldsymbol{\mu}_1 + \frac{\eta}{nm} \sum_{i \in \mathcal{S}_{-1}} \ell'^k_i \langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_{-1} \rangle j \boldsymbol{\mu}_{-1} \\ &\quad - \frac{\eta}{nm} \sum_{i=1}^n \ell'^k_i \langle \mathbf{w}^k_{j,r}, \boldsymbol{\xi}_i \rangle j y_i \boldsymbol{\xi}_i. \end{aligned}$$

Then the update of the inner product is

$$\begin{aligned} \langle \mathbf{w}^{\widetilde{k}}_{j,r}, \boldsymbol{\mu}_y \rangle &= \langle \mathbf{w}^{\widetilde{k}-1}_{j,r}, \boldsymbol{\mu}_y \rangle - \frac{\eta}{nm} \sum_{i \in \mathcal{S}_y} \ell'^{\widetilde{k}-1}_i \langle \mathbf{w}^{\widetilde{k}-1}_{j,r}, \boldsymbol{\mu}_y \rangle j y \|\boldsymbol{\mu}\|^2 \\ &= \Big(1 - \frac{\eta}{nm} j y \sum_{i \in \mathcal{S}_y} \ell'^{\widetilde{k}-1}_i \|\boldsymbol{\mu}\|^2 \Big) \langle \mathbf{w}^{\widetilde{k}-1}_{j,r}, \boldsymbol{\mu}_y \rangle \end{aligned}$$

By the condition that $\eta \le C^{-1} m \|\boldsymbol{\mu}\|^{-2}$ for sufficiently large constant C, we have $|\frac{\eta}{nm} j y \sum_{i \in \mathcal{S}_y} \ell'^{\widetilde{k}-1}_i \|\boldsymbol{\mu}\|^2| < 1$. Thus we can guarantee the $\mathrm{sign}(\langle \mathbf{w}^{\widetilde{k}}_{j,r}, \boldsymbol{\mu}_y \rangle) = \mathrm{sign}(\langle \mathbf{w}^{\widetilde{k}-1}_{j,r}, \boldsymbol{\mu}_y \rangle) = \mathrm{sign}(\langle \mathbf{w}^0_{j,r}, \boldsymbol{\mu}_y \rangle)$. □

Proof of Proposition D.1. We prove the results by induction. For $\rho^k_{j,r,i}$, we prove a stronger result that $|\rho^k_{y_i,r,i}| \le 0.9\alpha \le \alpha$ and $|\rho^k_{-y_i,r,i}| \le 0.6\beta + 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha$. First it is clear at $t = 0$, the results are satisfied based on the definition of β and $\alpha \ge \beta$. Now suppose that there exists $\widetilde{T} \le T^*$ such that results hold for all $0 \le k \le \widetilde{T} - 1$. We wish to show the results also hold for $k = \widetilde{T}$.

First recall the gradient descent update as

$$\mathbf{w}^{k+1}_{j,r} = \mathbf{w}^k_{j,r} - \frac{\eta}{nm} \sum_{i \in \mathcal{S}_1} \ell'^k_i \langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_1 \rangle j \boldsymbol{\mu}_1 + \frac{\eta}{nm} \sum_{i \in \mathcal{S}_{-1}} \ell'^k_i \langle \mathbf{w}^k_{j,r}, \boldsymbol{\mu}_{-1} \rangle j \boldsymbol{\mu}_{-1}$$

$$-\frac{\eta}{nm}\sum_{i=1}^{n}\ell_i'^k\langle\mathbf{w}_{j,r}^k,\boldsymbol{\xi}_i\rangle j y_i\boldsymbol{\xi}_i.$$

Then based on the orthogonal data modelling assumption, we have for $y \neq j$, i.e., $y = -j$,

$$\begin{aligned}\langle\mathbf{w}_{j,r}^{k+1},\boldsymbol{\mu}_{-j}\rangle &= \langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{-j}\rangle + \frac{\eta}{nm}\sum_{i\in\mathcal{S}_{-j}}\ell_i'^k\langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{-j}\rangle\|\boldsymbol{\mu}\|^2 \\ &= \Big(1-\frac{\eta\|\boldsymbol{\mu}\|^2}{nm}\sum_{i\in\mathcal{S}_{-j}}|\ell_i'^k|\Big)\langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{-j}\rangle\end{aligned}$$

where the second equality is by $\ell_i'^k < 0$ for all i,k. From Lemma D.6, we have $\mathrm{sign}(\langle\mathbf{w}_{j,r}^{k+1},\boldsymbol{\mu}_{-j}\rangle) = \mathrm{sign}(\langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{-j}\rangle)$ and thus

$$|\langle\mathbf{w}_{j,r}^{\widetilde{T}},\boldsymbol{\mu}_{-j}\rangle| \le \left|\Big(1-\frac{\eta\|\boldsymbol{\mu}\|^2}{nm}\sum_{i\in\mathcal{S}_{-j}}|\ell_i'^{\widetilde{T}-1}|\Big)\right|\left|\langle\mathbf{w}_{j,r}^{\widetilde{T}-1},\boldsymbol{\mu}_{-j}\rangle\right| \le \left|\langle\mathbf{w}_{j,r}^{\widetilde{T}-1},\boldsymbol{\mu}_{-j}\rangle\right| \le \beta$$

On the other hand, for $y = j$, we have

$$\begin{aligned}\langle\mathbf{w}_{j,r}^{k+1},\boldsymbol{\mu}_{j}\rangle &= \langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{j}\rangle - \frac{\eta}{nm}\sum_{i\in\mathcal{S}_{j}}\ell_i'^k\langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{j}\rangle\|\boldsymbol{\mu}\|^2 \\ &= \langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{j}\rangle + \frac{\eta\|\boldsymbol{\mu}\|^2}{nm}\sum_{i\in\mathcal{S}_{j}}|\ell_i'^k|\langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_{j}\rangle\end{aligned}$$

Next, we notice that

$$\begin{aligned}|\ell_i'^k| &= \frac{1}{1+\exp\big(F_{y_i}(\mathbf{W}_{y_i}^k,\mathbf{x}_i) - F_{-y_i}(\mathbf{W}_{-y_i}^k,\mathbf{x}_i)\big)} \\ &\le \exp\big(-F_{y_i}(\mathbf{W}_{y_i}^k,\mathbf{x}_i) + F_{-y_i}(\mathbf{W}_{-y_i}^k,\mathbf{x}_i)\big) \\ &\le \exp\big(-F_{y_i}(\mathbf{W}_{y_i}^k,\mathbf{x}_i) + 0.5\big) \\ &= \exp\big(-\frac{1}{m}\sum_{r=1}^{m}\big(\langle\mathbf{w}_{y_i,r}^k,\boldsymbol{\mu}_{y_i}\rangle^2 + \langle\mathbf{w}_{y_i,r}^k,\boldsymbol{\xi}_i\rangle^2\big) + 0.5\big)\end{aligned} \tag{7}$$

where the last inequality is by Lemma D.4. Let $k_{j,r}$ be the last time $k \le T^*$ that $|\langle\mathbf{w}_{j,r}^k,\boldsymbol{\mu}_j\rangle| \le 0.5\alpha$. Then we have

$$\langle\mathbf{w}_{j,r}^{\widetilde{T}},\boldsymbol{\mu}_j\rangle = \langle\mathbf{w}_{j,r}^{k_{j,r}},\boldsymbol{\mu}_j\rangle + \underbrace{\frac{\eta\|\boldsymbol{\mu}\|^2}{nm}|\ell_i'^{k_{j,r}}|\langle\mathbf{w}_{j,r}^{k_{j,r}},\boldsymbol{\mu}_j\rangle}_{A_1} + \underbrace{\frac{\eta\|\boldsymbol{\mu}\|^2}{nm}\sum_{k_{j,r}<k\le\widetilde{T}-1}|\ell_i'^k|\langle\mathbf{w}_{j,r}^{k},\boldsymbol{\mu}_j\rangle}_{A_2}.$$

Without loss of generality, we suppose $\langle\mathbf{w}_{j,r}^0,\boldsymbol{\mu}_j\rangle \ge 0$, then by Lemma D.6, $\langle\mathbf{w}_{j,r}^k,\boldsymbol{\mu}_j\rangle \ge 0$ for all $k \ge 0$. Then we can bound

$$|A_1| \le \frac{\eta\|\boldsymbol{\mu}\|^2}{nm}0.5\alpha \le 0.25\alpha$$

where the last inequality is by the condition that $\eta \le nm\|\boldsymbol{\mu}\|^{-2}/2$. Furthermore,

$$\begin{aligned}|A_2| &\le \frac{\eta\|\boldsymbol{\mu}\|^2}{nm}\sum_{k_{j,r}<k\le\widetilde{T}-1}\exp(-F_{y_i}(\mathbf{W}_{y_i}^k,\mathbf{x}_i)+0.5)\langle\mathbf{w}_{j,r}^k,\boldsymbol{\mu}_j\rangle \\ &\le \frac{2\eta\|\boldsymbol{\mu}\|^2\alpha}{nm}T^*\exp(-\alpha^2/4) \\ &= \frac{2\eta\|\boldsymbol{\mu}\|^2\alpha}{nm}T^*\exp(-\log(T^*)) \\ &\le 0.25\alpha\end{aligned}$$

where the first inequality is by (7) and the second inequality is by upper bound on $\langle \mathbf{w}^{k}_{j,r}, \boldsymbol{\mu}_j \rangle \leq \alpha$ for all $k \leq \widetilde{T} - 1$. The equality is by the definition of $\alpha = 2\log(T^*)$ and the last inequality is by the condition $\eta \leq nm\|\boldsymbol{\mu}\|^{-2}/8$. Thus, we can show

$$\langle \mathbf{w}^{\widetilde{T}}_{j,r}, \boldsymbol{\mu}_j \rangle \leq 0.5\alpha + 0.25\alpha + 0.25\alpha = \alpha.$$

Next for the noise growth, from Lemma D.2, we have for $y_i \neq j$

$$\rho^{\widetilde{T}}_{-y_i,r,i} = \rho^{\widetilde{T}-1}_{-y_i,r,i} + \frac{\eta}{nm}\ell_i'^{k} \langle \mathbf{w}^{\widetilde{T}-1}_{-y_i,r}, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2. \tag{8}$$

When $|\rho^{\widetilde{T}-1}_{-y_i,r,i}| \leq 1.5\big(0.3\beta + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\big)$, we have

$$|\rho^{\widetilde{T}}_{-y_i,r,i}| \leq |\rho^{\widetilde{T}-1}_{-y_i,r,i}| + \frac{2\eta\sigma_\xi^2 d\alpha}{nm} \leq |\rho^{\widetilde{T}-1}_{-y_i,r,i}| + 0.15\beta \leq 0.6\beta + 8\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha$$

where the second inequality is by triangle inequality and $|\ell_i'^{k}| \leq 1$ and Lemma C.2. The third inequality is by the lower bound on β in (3) and the condition that $\eta \leq 0.05nm\sigma_0\sigma_\xi^{-1}d^{-1/2}\alpha$.

Further, because $|\langle \mathbf{w}^{0}_{-y_i,r}, \boldsymbol{\xi}_i \rangle| \leq 0.3\beta$, when $1.5\big(0.3\beta + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\big) \leq |\rho^{\widetilde{T}-1}_{-y_i,r,i}| \leq 0.6\beta + 8\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha$, we can show from Lemma D.5 that if $\rho^{\widetilde{T}-1}_{-y_i,r,i} > 0$, then

$$\frac{1}{3}\rho^{\widetilde{T}-1}_{-y_i,r,i} \leq \langle \mathbf{w}^{\widetilde{T}-1}_{-y_i,r}, \boldsymbol{\xi}_i \rangle \leq \frac{4}{3}\rho^{\widetilde{T}-1}_{-y_i,r,i}$$

Then (8) suggests

$$\rho^{\widetilde{T}}_{-y_i,r,i} \leq \big(1 - \frac{\eta\|\boldsymbol{\xi}_i\|^2}{3nm}|\ell_i'^{k}|\big)\rho^{\widetilde{T}-1}_{-y_i,r,i} \leq \rho^{\widetilde{T}-1}_{-y_i,r,i} \leq 0.6\beta + 8\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha$$

If $\rho^{\widetilde{T}-1}_{-y_i,r,i} < 0$, then

$$\frac{4}{3}\rho^{\widetilde{T}-1}_{-y_i,r,i} \leq \langle \mathbf{w}^{\widetilde{T}-1}_{-y_i,r}, \boldsymbol{\xi}_i \rangle \leq \frac{1}{3}\rho^{\widetilde{T}-1}_{-y_i,r,i}$$

Then (8) suggests

$$\rho^{\widetilde{T}}_{-y_i,r,i} \geq \big(1 - \frac{\eta\|\boldsymbol{\xi}_i\|^2}{3nm}|\ell_i'^{k}|\big)\rho^{\widetilde{T}-1}_{-y_i,r,i} \geq \rho^{\widetilde{T}-1}_{-y_i,r,i} \geq -0.6\beta - 8\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha$$

Thus this completes the proof that $|\rho^{\widetilde{T}}_{-y_i,r,i}| \leq 0.6\beta + 8\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha$.

Finally, by Lemma D.5 we have for all $k \geq 0$

$$|\langle \mathbf{w}^{k}_{-y_i,r}, \boldsymbol{\xi}_i \rangle| \leq |\langle \mathbf{w}^{0}_{-y_i,r}, \boldsymbol{\xi}_i \rangle| + |\rho^{k}_{-y_i,r,i}| + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \leq 0.9\beta + 12\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha$$

which proves the upper bound for $|\langle \mathbf{w}^{\widetilde{T}}_{-y_i,r}, \boldsymbol{\xi}_i \rangle|$ and $|\rho^{\widetilde{T}}_{-y_i,r,i}|$.

Next, from Lemma D.2, we have for $y_i = j$,

$$\rho^{k+1}_{y_i,r,i} = \rho^{k}_{y_i,r,i} - \frac{\eta}{nm}\ell_i'^{k} \langle \mathbf{w}^{k}_{y_i,r}, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2. \tag{9}$$

Let $\tilde{k}_{r,i}$ be the last time $k < T^*$ that $|\rho^{k}_{y_i,r,i}| \leq 0.6\alpha$. Then it can be verified that for $k \geq \tilde{k}_{r,i}$,

$$|\langle \mathbf{w}^{k}_{y_i,r}, \boldsymbol{\xi}_i \rangle| \geq |\rho^{k}_{y_i,r,i}| - |\langle \mathbf{w}^{0}_{y_i,r}, \boldsymbol{\xi}_i \rangle| - 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \geq 0.5\alpha$$

where the first inequality is by Lemma D.5 and the last inequality is by $|\langle \mathbf{w}^{0}_{y_i,r}, \boldsymbol{\xi}_i \rangle| + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \leq 1 \leq 0.1\alpha$.

We now expand (9) as

$$\rho_{y_i,r,i}^{\widetilde{T}} = \rho_{y_i,r,i}^{\tilde{k}_{r,i}} + \underbrace{\frac{\eta}{nm}|\ell_i'^{\tilde{k}_{r,i}}|\langle \mathbf{w}_{y_i,r}^{\tilde{k}_{r,i}}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2}_{A_3} + \underbrace{\frac{\eta}{nm}\sum_{\tilde{k}_{r,i}<k\le\widetilde{T}-1}|\ell_i'^{k}|\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2}_{A_4}$$

Then we can bound

$$\begin{aligned}|A_3| &\le \frac{2\eta\sigma_\xi^2 d}{nm}|\langle \mathbf{w}_{y_i,r}^{\tilde{k}_{r,i}}, \boldsymbol{\xi}_i\rangle| \le \frac{2\eta\sigma_\xi^2 d}{nm}\left(|\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i\rangle| + 0.6\alpha + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\right)\\ &\le \frac{2\eta\sigma_\xi^2 d}{nm}0.7\alpha\\ &\le 0.15\alpha\end{aligned}$$

where the first inequality is by $|\ell_i'^k| \le 1$ and Lemma C.2 with $d = \Omega(\log(n/\delta))$ and the second inequality is by Lemma D.5. The last inequality is by the condition $\eta \le C^{-1}nm\sigma_\xi^{-2}d^{-1}$ for sufficiently large constant C.

In addition, we bound

$$\begin{aligned}|A_4| &\le \frac{2\eta\sigma_\xi^2 d\alpha}{nm}\sum_{k_{j,r}<k\le\widetilde{T}-1}\exp(-F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i) + 0.5)\\ &\le \frac{4\eta\sigma_\xi^2 d\alpha}{nm}T^*\exp(-\alpha^2/4)\\ &\le \frac{4\eta\sigma_\xi^2 d\alpha}{nm}\\ &\le 0.15\alpha\end{aligned}$$

where the first inequality is by (7) and the second inequality is by $|\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i\rangle| \ge 0.6\alpha - 0.1\alpha = 0.5\alpha$. The last inequality is by the condition $\eta \le C^{-1}nm\sigma_\xi^{-2}d^{-1}$ for sufficiently large constant C.

Combining the bound on $|A_3|$ and $|A_4|$, we have

$$|\rho_{y_i,r,i}^{\widetilde{T}}| \le 0.6\alpha + 0.15\alpha + 0.15\alpha = 0.9\alpha.$$

Lastly, we bound

$$\begin{aligned}|\langle \mathbf{w}_{y_i,r}^{\widetilde{T}}, \boldsymbol{\xi}_i\rangle| \le |\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i\rangle| + |\rho_{y_i,r,i}^{\widetilde{T}}| + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha &\le 0.3\beta + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha + 0.9\alpha\\ &\le \alpha.\end{aligned}$$

This shows the upper bound as $|\langle \mathbf{w}_{y_i,r}^{\widetilde{T}}, \boldsymbol{\xi}_i\rangle|, |\rho_{y_i,r,i}^{\widetilde{T}}| \le \alpha$. □

We require the following lemma that lower bound the loss derivatives in the first stage before the inner products reach constant order.

Lemma D.7. *If* $\max_{r,i,y}\{\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_y\rangle, \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle\} = O(1)$*, there exists a constant* $C_\ell > 0$ *such that* $|\ell_i'^k| \ge C_\ell$ *for all* $i \in [n]$*.*

Proof of Lemma D.7. If $\max_{r,i,y}\{\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_y\rangle, \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle\} = O(1)$, we can bound for all $j = \pm 1$

$$F_j(\mathbf{W}_j^k, \mathbf{x}_i) = \frac{1}{m}\sum_{r=1}^m\left(\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_{y_i}\rangle^2 + \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle^2\right) \le O(1)$$

Therefore, we can bound $|\ell_i'^k| = (1 + \exp(F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i) - F_{-y_i}(\mathbf{W}_{-y_i}^k, \mathbf{x}_i)))^{-1} \ge \Omega(1)$. □

We also prove the following upper bound on the gradient norm.

Lemma D.8 (Proof of Lemma D.8)**.** *Under Condition D.1, for* $0 \leq k \leq T^*$*, we can bound*

$$\|\nabla L_S(\mathbf{W}^k)\|^2 = O(\max\{\|\boldsymbol{\mu}\|^2, \sigma_\xi^2 d\}) L_S(\mathbf{W}^k)$$

Proof of Lemma D.8. The proof adopts a similar argument as in (Cao et al., 2022, Lemma C.7) and we include here for completeness. We first bound

$$\begin{aligned}
\|\nabla f(\mathbf{W}^k, \mathbf{x}_i)\| &\leq \frac{2}{m} \sum_{j,r} \big\| \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_{y_i} \rangle \boldsymbol{\mu}_{y_i} + \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i \rangle \boldsymbol{\xi}_i \big\| \\
&\leq \frac{2}{m} \sum_r |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\mu}_{y_i} \rangle| \|\boldsymbol{\mu}\| + \frac{2}{m} \sum_r |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle| \|\boldsymbol{\xi}_i\| \\
&\quad + \frac{2}{m} \sum_r |\langle \mathbf{w}_{-y_i,r}^k, \boldsymbol{\mu}_{y_i} \rangle| \|\boldsymbol{\mu}\| + \frac{2}{m} \sum_r |\langle \mathbf{w}_{-y_i,r}^k, \boldsymbol{\xi}_i \rangle| \|\boldsymbol{\xi}_i\| \\
&\leq \frac{2}{m} \sum_{r=1}^m \Big(|\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\mu}_{y_i} \rangle| + |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle| \Big) \max\{\|\boldsymbol{\mu}\|, 2\sigma_\xi \sqrt{d}\} \\
&\quad + \frac{2}{m} \sum_{r=1}^m \Big(|\langle \mathbf{w}_{-y_i,r}^k, \boldsymbol{\mu}_{y_i} \rangle| + |\langle \mathbf{w}_{-y_i,r}^k, \boldsymbol{\xi}_i \rangle| \Big) \max\{\|\boldsymbol{\mu}\|, 2\sigma_\xi \sqrt{d}\} \\
&\leq 2 \Big(\sqrt{F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i)} + \sqrt{F_{-y_i}(\mathbf{W}_{-y_i}^k, \mathbf{x}_i)} \Big) \max\{\|\boldsymbol{\mu}\|, 2\sigma_\xi \sqrt{d}\} \\
&\leq 2 \Big(\sqrt{F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i)} + 1 \Big) \max\{\|\boldsymbol{\mu}\|, 2\sigma_\xi \sqrt{d}\}
\end{aligned}$$

where the third inequality is by Lemma C.2 and the fourth inequality is by Jensen's inequality and the last inequality is by Lemma D.4 that $F_{-y_i}(\mathbf{W}_{-y_i}^k, \mathbf{x}_i)$ for all $i \in [n]$. Then we have

$$\begin{aligned}
&- \ell'(y_i f(\mathbf{W}^k, \mathbf{x}_i)) \|\nabla f(\mathbf{W}^k, \mathbf{x}_i)\|^2 \\
&\leq -\ell'\big(F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i) - 0.5\big) \Big(2 \Big(\sqrt{F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i)} + 1 \Big) \max\{\|\boldsymbol{\mu}\|, 2\sigma_\xi \sqrt{d}\} \Big)^2 \\
&= -4\ell'\big(F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i) - 0.5\big) \big(\sqrt{F_{y_i}(\mathbf{W}_{y_i}^k, \mathbf{x}_i)} + 1 \big)^2 \max\{\|\boldsymbol{\mu}\|^2, 4\sigma_\xi^2 d\} \\
&\leq \max_{z > 0} \{-4\ell'(z - 0.5)(\sqrt{z} + 1)^2\} \max\{\|\boldsymbol{\mu}\|^2, 4\sigma_\xi^2 d\} \\
&= O(\max\{\|\boldsymbol{\mu}\|^2, \sigma_\xi^2 d\})
\end{aligned}$$

where the last equality is by $\max_{z>0}\{-4\ell'(z - 0.5)(\sqrt{z} + 1)^2\} < \infty$ because ℓ' has an exponentially decaying tail. Then we can bound

$$\begin{aligned}
\|\nabla L_S(\mathbf{W}^k)\|^2 &\leq \Big(\frac{1}{n} \sum_{i=1}^n \ell'(y_i f(\mathbf{W}^k, \mathbf{x}_i)) \|\nabla f(\mathbf{W}^k, \mathbf{x}_i)\| \Big)^2 \\
&\leq \Big(\frac{1}{n} \sum_{i=1}^n \sqrt{-O(\max\{\|\boldsymbol{\mu}\|^2, \sigma_\xi^2 d\}) \ell'(y_i f(\mathbf{W}^k, \mathbf{x}_i))} \Big)^2 \\
&\leq O(\max\{\|\boldsymbol{\mu}\|^2, \sigma_\xi^2 d\}) \frac{1}{n} \sum_{i=1}^n -\ell'(y_i f(\mathbf{W}^k, \mathbf{x}_i)) \\
&\leq O(\max\{\|\boldsymbol{\mu}\|^2, \sigma_\xi^2 d\}) L_S(\mathbf{W}^k)
\end{aligned}$$

where the third inequality is by Cauchy-Schwartz inequality and the last inequality is by $-\ell' \leq \ell$ for cross-entropy loss. □

D.3 SIGNAL LEARNING

We first analyze the setting, where $n \cdot \mathrm{SNR}^2 \geq C'$ for some constant $C' > 0$, which allows signal learning to dominate noise memorization, thus reaching benign overfitting.

For the purpose of signal learning, we derive an anti-concentration result that provides a lower bound for signal inner product at initialization.

Lemma D.9. *Suppose $\delta > 0$ and $m = \Omega(\log(1/\delta))$. Then with probability at least $1 - \delta$, we have for all $j, y = \pm 1$*

$$\sigma_0 \|\boldsymbol{\mu}\|/2 \leq \frac{1}{m} \sum_{r=1}^{n} |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_y \rangle| \leq \sigma_0 \|\boldsymbol{\mu}\|$$

Proof of Lemma D.16. First notice that for any $j = \pm 1$, $\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_y \rangle \sim \mathcal{N}(0, \sigma_0^2 \|\boldsymbol{\mu}\|^2)$ and thus we have $\mathbb{E}[|\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_y \rangle|] = \sigma_0 \|\boldsymbol{\mu}\| \sqrt{2/\pi}$. By sub-Gaussian tail bound, with probability at least $1 - \delta/8$, for any $j, y = \pm 1$

$$\left| \frac{1}{m} \sum_{r=1}^{m} |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_y \rangle| - \sigma_0 \|\boldsymbol{\mu}\| \sqrt{2/\pi} \right| \leq \sqrt{\frac{2\sigma_0^2 \|\boldsymbol{\mu}\|^2 \log(8/\delta)}{m}}$$

Choosing $m = \Omega(\log(1/\delta))$, we have

$$\sigma_0 \|\boldsymbol{\mu}\| \sqrt{2/\pi} 0.99 \leq \frac{1}{m} \sum_{r=1}^{n} |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_y \rangle| \leq \sigma_0 \|\boldsymbol{\mu}\| \sqrt{2/\pi} 1.01.$$

Then we have $\sigma_0 \|\boldsymbol{\mu}\|/2 \leq \frac{1}{m} \sum_{r=1}^{n} |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\xi}_i \rangle| \leq \sigma_0 \|\boldsymbol{\mu}\|$. Finally taking the union bound for all $j, y = \pm 1$ completes the proof. □

We have established several preliminary lemmas that hold with high probability, including Lemma C.1, Lemma C.2, Lemma D.1, Lemma D.9. We let $\mathcal{E}_{\text{prelim}}$ be the event such that all the results in these lemmas hold for a given δ. Then by applying union bound, we have $\mathbb{P}(\mathcal{E}_{\text{prelim}}) \geq 1 - 4\delta$. The subsequent analysis are conditioned on the event $\mathcal{E}_{\text{prelim}}$.

D.3.1 First stage

In the first stage where $\max_{r,i,y} \{ \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_y \rangle, \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i \rangle \} = O(1)$, we show in Lemma D.7 that we can lower bound the loss derivatives by a constant C_ℓ, i.e., $|\ell_i'^k| \geq C_\ell$, for all $i \in [n], k \leq T_1$.

Theorem D.1. *Under Condition D.1, suppose $n \cdot \mathrm{SNR}^2 \geq C'$ for some $C' \geq 0$. Then there exists a time $T_1 = \widetilde{\Theta}(\eta^{-1} m \|\boldsymbol{\mu}\|^{-2})$, such that (1) $\max_r |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j \rangle| \geq 2$, for all $j = \pm 1$, (2) $\frac{1}{m} \sum_{r=1}^{m} |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j \rangle| \geq 2$, for all $j = \pm 1$ (3) $\max_{r,i} |\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i \rangle| = \widetilde{O}(n^{-1/2})$.*

Proof of Theorem D.1. We first upper bound the growth of noise by analyzing inner product dynamics

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle &= \langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \sum_{i'=1}^{n} \ell_{i'}'^{k-1} \langle \mathbf{w}_{j,r}^{k-1}, \boldsymbol{\xi}_{i'} \rangle \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&= \langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \ell_i'^k \langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2 - \frac{\eta}{nm} \sum_{i' \neq i} \ell_{i'}'^{k-1} \langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_{i'} \rangle \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle
\end{aligned}$$

This suggests

$$|\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle| \leq |\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle| + \frac{\eta}{nm} |\ell_i'^k| |\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle| \|\boldsymbol{\xi}_i\|^2 + \frac{\eta}{nm} \sum_{i' \neq i} |\ell_{i'}'^k| |\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_{i'} \rangle| |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \tag{10}$$

Next, from Lemma D.7 and Lemma C.2, we have for any $i' \neq i \in [n]$ and $k \leq T_1$,

$$\frac{|\ell_{i'}'^k| \cdot |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle|}{|\ell_i'^k| \cdot \|\boldsymbol{\xi}_i\|^2} \leq \frac{2\sigma_\xi^2 \sqrt{d \log(4n^2/\delta)}}{C_\ell 0.99 \sigma_\xi^2 d} = 2.1 C_\ell^{-1} \sqrt{\frac{\log(4n^2/\delta)}{d}}$$

where we use the lower and upper bound on loss derivatives during the first stage, as well as Lemma C.2. Then taking the maximum of (10) over the neurons and samples, we let $B^k := \max_{r,i} |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle|$ and obtain

$$B^k \leq B^{k-1} + \frac{\eta}{nm} \Big(1 + 2.1 C_\ell^{-1} n \sqrt{\frac{\log(4n^2/\delta)}{d}} \Big) |\ell_i'^k| \|\boldsymbol{\xi}_i\|^2 B^{k-1}$$

$$
\begin{aligned}
&\le \Big(1 + \frac{1.01\eta\|\boldsymbol{\xi}_i\|^2}{nm}\Big) B^{k-1} \\
&\le \Big(1 + \frac{1.02\eta\sigma_\xi^2 d}{nm}\Big)^k B^0
\end{aligned}
$$

where the second inequality is by $d = \widetilde{\Omega}(n^2)$ sufficiently large and $|\ell_i'^k| \le 1$. The third inequality is by Lemma C.2.

We then consider the propagation of $\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_y \rangle$. From the gradient update we can show for $j = y$,

$$
\begin{aligned}
\langle \mathbf{w}_{j,r}^{k+1}, \boldsymbol{\mu}_j \rangle &= \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle - \frac{\eta}{nm} \sum_{i \in \mathcal{S}_j} \ell_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle \|\boldsymbol{\mu}\|^2 \\
&\ge \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle + \frac{\eta C_\ell |\mathcal{S}_1| \|\boldsymbol{\mu}\|^2}{nm} \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle \\
&\ge \big(1 + 0.49 \frac{\eta C_\ell \|\boldsymbol{\mu}\|^2}{m}\big) \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle
\end{aligned}
$$

where the first inequality is by loss derivative lower bound and the the second inequality is by Lemma C.1 and $n = \widetilde{\Omega}(1)$ sufficiently large. This implies that

$$
|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle| \ge \big(1 + 0.49 \frac{\eta C_\ell \|\boldsymbol{\mu}\|^2}{m}\big) |\langle \mathbf{w}_{j,r}^{k-1}, \boldsymbol{\mu}_j \rangle| \ge \big(1 + 0.49 \frac{\eta C_\ell \|\boldsymbol{\mu}\|^2}{m}\big)^k |\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_j \rangle|
$$

Applying Lemma D.1 and Lemma D.9, we have for all $j = \pm 1$,

$$
\begin{aligned}
\max_r |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle| &\ge \Big(1 + 0.49 \frac{\eta C_\ell \|\boldsymbol{\mu}\|^2}{m}\Big)^k \sigma_0 \|\boldsymbol{\mu}\|/2 \\
\frac{1}{m} \sum_{r=1}^m |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle| &\ge \Big(1 + 0.49 \frac{\eta C_\ell \|\boldsymbol{\mu}\|^2}{m}\Big)^k \sigma_0 \|\boldsymbol{\mu}\|/2
\end{aligned}
$$

Consider

$$
T_1 = \log(4m\sigma_0^{-1}\|\boldsymbol{\mu}\|^{-1}) / \log\big(1 + 0.49 \frac{\eta C_\ell \|\boldsymbol{\mu}\|^2}{m}\big) = \Theta(\eta^{-1} m \|\boldsymbol{\mu}\|^{-2} \log(4m\sigma_0^{-1}\|\boldsymbol{\mu}\|^{-1}))
$$

for η sufficiently small. Then we can verify that for $j = \pm 1$, we have

$$
\max_r |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j \rangle| \ge 2, \quad \frac{1}{m} \sum_{r=1}^m |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j \rangle| \ge 2,
$$

Now under the SNR condition, we can bound the growth of noise as

$$
\begin{aligned}
B^{T_1} &\le \Big(1 + 1.01 \frac{\eta\sigma_\xi^2 d}{nm}\Big)^{T_1} 2\sigma_0\sigma_\xi\sqrt{d}\sqrt{\log(8mn/\delta)} \\
&= \exp\left(\frac{\log(1 + 1.02\frac{\eta\sigma_\xi^2 d}{nm})}{\log(1 + 0.49\frac{\eta C_\ell \|\boldsymbol{\mu}\|^2}{m})} \log(4\sigma_0^{-1}\|\boldsymbol{\mu}\|^{-1}) \right) 2\sigma_0\sigma_\xi\sqrt{d}\sqrt{\log(8mn/\delta)} \\
&\le \exp\Big(\big(2.1/C_\ell n^{-1}\mathrm{SNR}^{-2} + \widetilde{O}(n\mathrm{SNR}^2\eta)\big)\log(4\sigma_0^{-1}\|\boldsymbol{\mu}\|^{-1})\Big) 2\sigma_0\sigma_\xi\sqrt{d}\sqrt{\log(8mn/\delta)} \\
&\le \exp\big(\big(2.1/C_\ell n^{-1}\mathrm{SNR}^{-2} + 0.01\big)\log(4\sigma_0^{-1}\|\boldsymbol{\mu}\|^{-1})\big) 2\sigma_0\sigma_\xi\sqrt{d}\sqrt{\log(8mn/\delta)} \\
&\le 8\mathrm{SNR}^{-1}\sqrt{\log(8mn/\delta)} \\
&= \widetilde{O}(n^{-1/2})
\end{aligned}
$$

where the first inequality is by Lemma D.1 and the second inequality is by Taylor expansion around $\eta = 0$. The third inequality is by choosing η sufficiently small and the fourth inequality is by the SNR condition that $n \cdot \mathrm{SNR}^2 \ge C' \ge 2.5 C_\ell^{-1}$. □

D.3.2 Second stage

First, at the end of first stage, we have

- $\max_r |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j\rangle| \geq 2$ for all $j = \pm 1$.
- $\frac{1}{m}\sum_{r=1}^m |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j\rangle| \geq 2$ for all $j = \pm 1$.
- $\max_{r,i} |\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i\rangle| = \widetilde{O}(n^{-1/2})$
- $\max_{r,i} |\langle \mathbf{w}_{-y_i,r}^{T_1}, \boldsymbol{\xi}_i\rangle| \leq \beta + 12\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha$.

Next we define

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^0 + 2\log(4/\epsilon)\mathrm{sign}(\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_j\rangle)\frac{\boldsymbol{\mu}_j + \boldsymbol{\mu}_{-j}}{\|\boldsymbol{\mu}\|^2}$$

We first show the monotonicity of signal inner product in the second stage.

Lemma D.10. *Under the same conditions as in Theorem D.1, we have for all $j = \pm 1, r \in [m]$, $T_1 \leq k \leq T$, $|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle| \geq |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j\rangle| \geq 2$.*

Proof of Lemma D.10. From the update of signal inner product, we have for all $j = \pm 1, r \in [m]$, $T_1 \leq k \leq T$

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{k+1}, \boldsymbol{\mu}_j\rangle &= \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle - \frac{\eta}{nm}\sum_{i\in\mathcal{S}_j}\ell_i'^k\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle\|\boldsymbol{\mu}\|^2 \\ &= \big(1 - \frac{\eta\|\boldsymbol{\mu}\|^2}{nm}\sum_{i\in\mathcal{S}_j}\ell_i'^k\big)\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle.\end{aligned}$$

Thus $|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle| \geq |\langle \mathbf{w}_{j,r}^{k-1}, \boldsymbol{\mu}_j\rangle| \geq |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j\rangle| \geq 2$ for all $j = \pm 1, r \in [m], T_1 \leq k \leq T$. □

We then bound the distance between $\mathbf{W}^{T_1}$ to $\mathbf{W}^*$.

Lemma D.11. *Under Condition D.1, we can bound $\|\mathbf{W}^{T_1} - \mathbf{W}^*\| = O(\sqrt{m}\log(1/\epsilon)\|\boldsymbol{\mu}\|^{-1})$.*

Proof of Lemma D.11. Let $\mathbf{P}_{\boldsymbol{\xi}}$ be the projection matrix to the direction of $\boldsymbol{\xi}$, i.e., $\mathbf{P}_{\boldsymbol{\xi}} = \frac{\boldsymbol{\xi}\boldsymbol{\xi}^\top}{\|\boldsymbol{\xi}\|^2}$. Then we can represent

$$\begin{aligned}\mathbf{w}_{j,r}^k - \mathbf{w}_{j,r}^0 = {}& \mathbf{P}_{\boldsymbol{\mu}_1}(\mathbf{w}_{j,r}^k - \mathbf{w}_{j,r}^0) + \mathbf{P}_{\boldsymbol{\mu}_{-1}}(\mathbf{w}_{j,r}^k - \mathbf{w}_{j,r}^0) + \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}(\mathbf{w}_{j,r}^k - \mathbf{w}_{j,r}^0) \\ & + \Big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mu}_1} - \mathbf{P}_{\boldsymbol{\mu}_{-1}} - \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}\Big)(\mathbf{w}_{j,r}^k - \mathbf{w}_{j,r}^0).\end{aligned}$$

By the scale difference at T_1 and the fact that gradient descent only updates in the direction of $\boldsymbol{\mu}_j$, $j = \pm 1$ and $\boldsymbol{\xi}_i$, we can bound

$$\begin{aligned}&\|\mathbf{W}^{T_1} - \mathbf{W}^0\|^2 \\ &\leq \sum_{j=\pm 1, r\in[m]}\Big(\frac{\langle \mathbf{w}_{j,r}^{T_1} - \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_1\rangle^2}{\|\boldsymbol{\mu}\|^2} + \frac{\langle \mathbf{w}_{j,r}^{T_1} - \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_{-1}\rangle^2}{\|\boldsymbol{\mu}\|^2} + \sum_{i=1}^n \frac{\langle \mathbf{w}_{j,r}^{T_1} - \mathbf{w}_{j,r}^0, \boldsymbol{\xi}_i\rangle^2}{\|\boldsymbol{\xi}_i\|^2}\Big) \\ &\quad + \sum_{j=\pm 1, r\in[m]}\bigg\|\Big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mu}_1} - \mathbf{P}_{\boldsymbol{\mu}_{-1}} - \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}\Big)(\mathbf{w}_{j,r}^{T_1} - \mathbf{w}_{j,r}^0)\bigg\|^2 \\ &\leq 2m\Big(\frac{2\max_r\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j\rangle^2}{\|\boldsymbol{\mu}\|^2} + \frac{2\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_{-j}\rangle^2 + 2\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_{-j}\rangle^2 + 2\langle \mathbf{w}_{j,r}^0, \boldsymbol{\mu}_j\rangle^2}{\|\boldsymbol{\mu}\|^2}\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{n} \frac{2\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\xi}_i\rangle^2 + 2\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\xi}_i\rangle^2}{\|\boldsymbol{\xi}_i\|^2}\Big) + \sum_{j=\pm 1, r\in[m]} \Big\| \Big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mu}_1} - \mathbf{P}_{\boldsymbol{\mu}_{-1}} - \sum_{i=1}^{n} \mathbf{P}_{\boldsymbol{\xi}_i}\Big)(\mathbf{w}_{j,r}^{T_1} - \mathbf{w}_{j,r}^{0}) \Big\|^2 \\
&\le O(m\|\boldsymbol{\mu}\|^{-2})
\end{aligned}
$$

where we have use the scale difference at T_1. Therefore,

$$
\begin{aligned}
\|\mathbf{W}^{T_1} - \mathbf{W}^*\| &\le \|\mathbf{W}^{T_1} - \mathbf{W}^0\| + \|\mathbf{W}^0 - \mathbf{W}^*\| \\
&\le O(\sqrt{m}\|\boldsymbol{\mu}\|^{-1}) + O(\sqrt{m}\log(1/\epsilon)\|\boldsymbol{\mu}\|^{-1}) \\
&\le O(\sqrt{m}\log(1/\epsilon)\|\boldsymbol{\mu}\|^{-1})
\end{aligned}
$$

where we use the definition of $\mathbf{W}^*$. □

Lemma D.12. *Under Condition D.1, we have for all $T_1 \le k \le T^*$*

$$
\|\mathbf{W}^k - \mathbf{W}^*\|^2 - \|\mathbf{W}^{k+1} - \mathbf{W}^*\|^2 \ge 2\eta L_S(\mathbf{W}^t) - \eta\epsilon
$$

Proof of Lemma D.12. The proof is similar as in Cao et al. (2022). We first show a lower bound on $y_i\langle \nabla f(\mathbf{W}^t, \mathbf{x}_i), \mathbf{W}^*\rangle$ for any $i \in [n]$ for all $T_1 \le k \le T^*$.

$$
\begin{aligned}
y_i\langle \nabla f(\mathbf{W}^k, \mathbf{x}_i), \mathbf{W}^*\rangle &= \frac{1}{m}\sum_{j,r} j y_i \langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_{y_i}\rangle\langle \boldsymbol{\mu}_{y_i}, \mathbf{w}_{j,r}^*\rangle + \frac{1}{m}\sum_{j,r} j y_i \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle\langle \boldsymbol{\xi}_i, \mathbf{w}_{j,r}^*\rangle \\
&= \frac{1}{m}\sum_{r=1}^{m}\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\mu}_{y_i}\rangle\langle \mathbf{w}_{y_i,r}^*, \boldsymbol{\mu}_{y_i}\rangle - \frac{1}{m}\sum_{r=1}^{m}\langle \mathbf{w}_{-y_i,r}^k, \boldsymbol{\mu}_{y_i}\rangle\langle \mathbf{w}_{-y_i,r}^*, \boldsymbol{\mu}_{y_i}\rangle \\
&\quad + \frac{1}{m}\sum_{j,r} j y_i \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle\langle \boldsymbol{\xi}_i, \mathbf{w}_{j,r}^0\rangle \\
&= \underbrace{\frac{1}{m}\sum_{r=1}^{m} |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\mu}_{y_i}\rangle| 2\log(4/\epsilon)}_{A_5} + \underbrace{\frac{1}{m}\sum_{r=1}^{m}\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\mu}_{y_i}\rangle\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\mu}_{y_i}\rangle}_{A_6} \\
&\quad \underbrace{- \frac{1}{m}\sum_{r=1}^{m}\langle \mathbf{w}_{-y_i,r}^k, \boldsymbol{\mu}_{y_i}\rangle\langle \mathbf{w}_{-y_i,r}^*, \boldsymbol{\mu}_{y_i}\rangle}_{A_7} + \underbrace{\frac{1}{m}\sum_{j,r} j y_i \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle\langle \boldsymbol{\xi}_i, \mathbf{w}_{j,r}^0\rangle}_{A_8}
\end{aligned}
$$

where the second equality is by definition of $\mathbf{W}^*$. The third equality is by Lemma D.6. We next bound

$$
\begin{aligned}
&|A_6| \le \sigma_0\|\boldsymbol{\mu}\|\sqrt{2\log(8m/\delta)}\alpha = \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|) \\
&|A_7| \le \frac{1}{m}\sum_{r=1}^{m} |\mathbf{w}_{-y_i,r}^k, \boldsymbol{\mu}_{y_i}|\big(|\langle \mathbf{w}_{-y_i,r}^0, \boldsymbol{\mu}_{y_i}\rangle| + 2\log(2/\epsilon)\big) = \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|) \\
&|A_8| \le \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})
\end{aligned}
$$

where we use the global bound on the inner product by $\widetilde{O}(1)$. Next, by Theorem D.1 and Lemma D.10, we can show $\frac{1}{m}\sum_{r=1}^{m} |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\mu}_{y_i}\rangle| \ge 2$ for all $i \in [n]$ and we can lower bound $A_5 \ge 4\log(4/\epsilon)$ and thus

$$
y_i\langle \nabla f(\mathbf{W}^k, \mathbf{x}_i), \mathbf{W}^*\rangle \ge 4\log(4/\epsilon) - 2\log(4/\epsilon) = 2\log(4/\epsilon) \tag{11}
$$

where we bound $|A_6| + |A_7| + |A_8| \le 2\log(4/\epsilon)$ under Condition D.1.

Further, we derive

$$
\begin{aligned}
&\|\mathbf{W}^k - \mathbf{W}^*\|^2 - \|\mathbf{W}^{k+1} - \mathbf{W}^*\|^2 \\
&= 2\eta\langle \nabla L_S(\mathbf{W}^k), \mathbf{W}^k - \mathbf{W}^*\rangle - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2 \\
&= \frac{2\eta}{n}\sum_{i=1}^{n} \ell_i'^k y_i\big(2f(\mathbf{W}^k, \mathbf{x}_i) - \langle \nabla f(\mathbf{W}^k, \mathbf{x}_i), \mathbf{W}^*\rangle\big) - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{2\eta}{n}\sum_{i=1}^{n} \ell_i'^k\big(2y_i f(\mathbf{W}^k,\mathbf{x}_i) - 2\log(2/\epsilon)\big) - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2 \\
&\geq \frac{4\eta}{n}\sum_{i=1}^{n} \big(\ell(y_i f(\mathbf{W}^k,\mathbf{x}_i)) - \epsilon/4\big) - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2 \\
&\geq 2\eta L_S(\mathbf{W}^k) - \eta\epsilon
\end{aligned}
$$

where the first inequality is by (11) and the second inequality is by convexity of cross-entropy function and the last inequality is by Lemma D.8. □

Before proving the second stage convergence, we require the following lemma in order to bound the ratio of loss derivatives among different samples.

Lemma D.13 (Kou et al. (2023))**.** *Let $g(z) = \ell'(z) = -(1+\exp(z))^{-1}$. Then for any $z_2 - c \geq z_1 \geq -1$ where $c \geq 0$, we have $g(z_1)/g(z_2) \leq \exp(c)$.*

Theorem D.2. *Under the same settings as in Theorem D.1, let $T = T_1 + \lfloor \frac{\|\mathbf{W}^{T_1}-\mathbf{W}^*\|^2}{\eta\epsilon} \rfloor = T_1 + O(\eta^{-1}\epsilon^{-1}m\|\boldsymbol{\mu}\|^{-2})$. Then we have*

- *there exists $T_1 \leq k \leq T$ such that $L_S(\mathbf{W}^k) \leq 0.1$.*
- $\max_{j,r,i}|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle| = o(1)$ *for all $T_1 \leq k \leq T$.*
- $\max_r |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle| \geq 2$ *for all $j = \pm 1, T_1 \leq k \leq T$.*

Proof of Theorem D.2. By Lemma D.12, for any $T_1 \leq k \leq T$, we have

$$\|\mathbf{W}^k - \mathbf{W}^*\|^2 - \|\mathbf{W}^{k+1} - \mathbf{W}^*\|^2 \geq 2\eta L_S(\mathbf{W}^k) - \eta\epsilon$$

for all $s \leq k$. Then summing over the inequality gives

$$\frac{1}{T - T_1 + 1}\sum_{k=T_1}^{T} L_S(\mathbf{W}^k) \leq \frac{\|\mathbf{W}^{T_1} - \mathbf{W}^*\|^2}{2\eta(T - T_1 + 1)} + \frac{\epsilon}{2} \leq \epsilon$$

where the last inequality is by the choice $T = T_1 + \lfloor \frac{\|\mathbf{W}^{T_1}-\mathbf{W}^*\|^2}{\eta\epsilon} \rfloor = T_1 + \Omega(\eta^{-1}\epsilon^{-1}m\log(1/\epsilon)\|\boldsymbol{\mu}\|^{-2})$. Then we can claim that there exists a $k \in [T_1, T]$ such that $L_S(\mathbf{W}^k) \leq \epsilon$. Setting $\epsilon = 0.1$ shows the desired convergence.

Next, we show the upper bound on $\max_{j,r,i}|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle|$ for all $k \in [T_1, T]$. Notice that by Proposition D.1, we already have $\max_{j,r}|\langle \mathbf{w}^k_{-y_i,r}, \boldsymbol{\xi}_i\rangle| \leq \vartheta$, where we let $\vartheta := 3\max\{\max_{r,i}|\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i\rangle|, \beta, 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\}$. Then we only focus on bounding $\max_{y_i,i}|\langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle|$. From the scale difference at T_1, we know that $\vartheta = \widetilde{O}(\max\{n^{-1/2}, \sigma_0\sigma_\xi\sqrt{d}, \sigma_0\|\boldsymbol{\mu}\|, nd^{-1/2}\}) = o(1)$. Next, we can bound

$$\sum_{k=T_1}^{T} L_S(\mathbf{W}^k) \leq \frac{\|\mathbf{W}^{T_1} - \mathbf{W}^*\|^2}{\eta} = O(\eta^{-1}m\log(1/\epsilon)\|\boldsymbol{\mu}\|^{-2}) \tag{12}$$

where we use Lemma D.11 for the last equality.

Then, we first prove $\max_{r,i}|\rho_{y_i,r,i}^k| \leq 2\vartheta$ for all $T_1 \leq k \leq T$. First it is easy to see that at T_1, we have

$$\max_{r,i}|\rho_{y_i,r,i}^{T_1}| \leq \max_{r,i}|\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i\rangle| + \max_{r,i}|\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i\rangle| + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \leq \vartheta \leq 2\vartheta$$

Then suppose there $\widetilde{T} \in [T_1, T]$ such that $\max_{r,i}|\rho_{y_i,r,i}^{T_1}| \leq 2\vartheta$ for all $k \in [T_1, \widetilde{T} - 1]$. Now we let $\phi^k := \max_{r,i}|\rho_{y_i,r,i}^k|$ and thus by the update of noise coefficient

$$\phi^{k+1} \leq \phi^k + \frac{\eta}{nm}|\ell_i'^k|\Big(\phi^k + \beta/3 + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\Big)\|\boldsymbol{\xi}_i\|^2$$

$$\le \phi^k + \frac{\eta}{nm}\max_i |\ell_i'^k|\Big(\phi^k + \beta/3 + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\Big)O(\sigma_\xi^2 d).$$

where we use Lemma D.5 in the first inequality. Then taking the summation from T_1 to $\widetilde{T}$ gives

$$\phi^{\widetilde{T}} \le \phi^{T_1} + \frac{\eta}{nm}\sum_{k=T_1}^{\widetilde{T}-1}\max_i |\ell_i'^k| O(\sigma_\xi^2 d)\vartheta \tag{13}$$

where the first inequality is by the induction condition. Next, the aim is bound $\sum_{k=T_1}^{\widetilde{T}-1}\max_i |\ell_i'^k|$. First, for any $i, i' \in [n]$ such that $y_i = y_{i'}$, we can bound for all $T_1 \le k \le \widetilde{T} - 1$

$$\begin{aligned}
& y_i f(\mathbf{W}^k, \mathbf{x}_i) - y_{i'} f(\mathbf{W}^k, \mathbf{x}_{i'}) \\
& = F_{y_i}(\mathbf{W}^k_{y_i}, \mathbf{x}_i) - F_{-y_i}(\mathbf{W}^k_{-y_i}, \mathbf{x}_i)) - F_{y_{i'}}(\mathbf{W}^k_{y_{i'}}, \mathbf{x}_{i'}) + F_{-y_{i'}}(\mathbf{W}^k_{-y_{i'}}, \mathbf{x}_{i'})) \\
& \le \frac{1}{m}\sum_{r=1}^m \big(\langle \mathbf{w}^k_{y_i,r}, \boldsymbol{\mu}_{y_i}\rangle^2 + \langle \mathbf{w}^k_{y_i,r}, \boldsymbol{\xi}_i\rangle^2\big) - \frac{1}{m}\sum_{r=1}^m \big(\langle \mathbf{w}^k_{y_i,r}, \boldsymbol{\mu}_{y_i}\rangle^2 + \langle \mathbf{w}^k_{y_i,r}, \boldsymbol{\xi}_{i'}\rangle^2\big) + 1/C_1 \\
& = \frac{1}{m}\sum_{r=1}^m \big(\langle \mathbf{w}^k_{y_i,r}, \boldsymbol{\xi}_i\rangle^2 - \langle \mathbf{w}^k_{y_i,r}, \boldsymbol{\xi}_{i'}\rangle^2\big) + 1/C_1 \\
& \le \max_{r,i} \langle \mathbf{w}^k_{y_i,r}, \boldsymbol{\xi}_i\rangle^2 + 1/C_1 \\
& \le \max_{r,i}\big(|\rho^k_{y_i,r,i}| + \max_{r,i}|\langle \mathbf{w}^0_{y_i,r}, \boldsymbol{\xi}_i\rangle| + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\big)^2 \\
& \le 6\vartheta^2 \le \vartheta
\end{aligned}$$

where in the first inequality we notice that $F_{-y_i}(\mathbf{W}^k_{-y_i}, \mathbf{x}_i)) \ge 0$, $y_i = y_{i'}$ and we recall that $F_{-y_i}(\mathbf{W}^k_j, \mathbf{x}_i) \le \beta^2 + \big(\beta + 12\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\big)^2 = 1/C_1$ for some sufficiently large constant $C_1 > 0$. The second last inequality is by induction condition and the last inequality is by choosing $\vartheta \le 1/6$. Then we can bound the ratio of loss derivatives (based on Lemma D.13) that

$$|\ell_{i'}'^k|/|\ell_i'^k| \le \exp\big(y_i f(\mathbf{W}^k, \mathbf{x}_i) - y_{i'} f(\mathbf{W}^k, \mathbf{x}_{i'})\big) \le \exp(\vartheta)$$

This suggests $1 - O(\vartheta) \le |\ell_{i'}'^k|/|\ell_i'^k| \le 1 + O(\vartheta)$ for all $i, i' \in [n]$, $T_1 \le k \le \widetilde{T} - 1$. Then let $i^* = \arg\max_i |\ell_i'^k|$, we have

$$\begin{aligned}
\sum_{T_1}^{T}\max_i |\ell_i'^k| = \sum_{T_1}^{T}\Theta\Big(\frac{1}{|\mathcal{S}_{y_{i^*}}|}\sum_{i\in\mathcal{S}_{y_{i^*}}}|\ell_i'^k|\Big) \le \sum_{T_1}^{T}\Theta\Big(\frac{1}{|\mathcal{S}_{y_{i^*}}|}\sum_{i\in\mathcal{S}_{y_{i^*}}}\ell_i^k\Big) \le \sum_{T_1}^{T}\Theta\Big(\frac{n}{|\mathcal{S}_{y_{i^*}}|}L_S(\mathbf{W}^k)\Big) \\
= \widetilde{O}(\eta^{-1} m \log(1/\epsilon)\|\boldsymbol{\mu}\|^{-2})
\end{aligned} \tag{14}$$

where the first inequality is by $|\ell'| \le \ell$ and the last equality is from (12) and $|\mathcal{S}_{y_{i^*}}| \ge 0.49n$ (based on Lemma C.1).

This allows to bound (13) as

$$\begin{aligned}
\phi^{\widetilde{T}} & \le \phi^{T_1} + \frac{\eta}{nm}\sum_{s=T_1}^{\widetilde{T}-1}\max_i |\ell_i'^k| O(\sigma_\xi^2 d)\vartheta \\
& \le \phi^{T_1} + O(n^{-1}\sigma_\xi^2 d \log(1/\epsilon)\|\boldsymbol{\mu}\|^{-2})\cdot\vartheta \\
& \le \vartheta + O(n^{-1}\mathrm{SNR}^{-2}\log(1/\epsilon))\cdot\vartheta \\
& \le 2\vartheta
\end{aligned}$$

and the second inequality is by (14) and the last inequality is by setting $\epsilon = 0.1$ and $n \cdot \mathrm{SNR}^2 \ge C'$ for sufficiently large constant C'. Thus, we have $\max_{r,i}|\langle \mathbf{w}^{\widetilde{T}}_{y_i,r}, \boldsymbol{\xi}_i\rangle| \le \max_{r,i}|\rho^{\widetilde{T}}_{y_i,r,i}| + \beta + 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \le 3\vartheta = o(1)$. The lower bound on signal inner product is directly from Lemma D.10. □

D.4 Noise memorization

We also analyze the setting where $n^{-1}\mathrm{SNR}^{-2} \geq C'$ for some constant $C' > 0$, which allows the noise memorization to dominate signal learning, thus reaching harmful overfitting.

We first require the following anti-concentration result for the noise inner product, which is required to ensure the sign invariance of the inner product along training.

Lemma D.14. *Suppose $\delta > 0$ and $\sigma_0 \geq \Omega(\log(n^2/\delta)n^2 m\alpha d^{-1}\sigma_\xi^{-1})$, we have for all $j = \pm 1, r \in [m], i \in [n]$, $|\langle \mathbf{w}_{j,r}^0, \boldsymbol{\xi}_i \rangle| \geq 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha$.*

Proof of Lemma D.14. For any $j = \pm 1, r \in [m], i \in [n]$, we have $\langle \mathbf{w}_{j,r}^0, \boldsymbol{\xi}_i \rangle \sim \mathcal{N}(0, \sigma_0^2 \|\boldsymbol{\xi}_i\|^2)$. Then applying Lemma D.3 by setting RHS to $\delta/(2mn)$ and $c = 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha$, we require

$$d^2 \geq 42 \log(4n^2/\delta) n^2 \alpha^2 \sigma_0^{-2} \sigma_\xi^{-2} / \log(\frac{4m^2n^2}{4m^2n^2 - \delta^2})$$

where we use Lemma C.2 that $\|\boldsymbol{\xi}_i\|^2 \geq 0.99\sigma_\xi^2 d$. Finally noticing that $1/\log(4m^2n^2/(4m^2n^2 - \delta^2)) \leq \Theta(m^2n^2)$ and taking the union bound completes the proof. □

An immediate consequence of Lemma D.14 is the following result that allows to derive the sign invariance for $\langle \mathbf{w}_{y_i,r,i}^k, \boldsymbol{\xi}_i \rangle$ for all iterations.

Lemma D.15. *Under Condition D.1, for any $i \in [n], r \in [m]$, we have* $\mathrm{sign}(\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle) = \mathrm{sign}(\rho_{y_i,r,i}^k) = \mathrm{sign}(\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle)$ *for all* $0 \leq k \leq T^*$.

Proof of Lemma D.15. First by Lemma D.14 and Lemma D.5, we can bound if $\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \geq 0$,

$$\rho_{y_i,r,i}^k + \frac{1}{2}\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \leq \langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle \leq \rho_{y_i,r,i}^k + \frac{3}{2}\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle$$

and if $\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \leq 0$,

$$\rho_{y_i,r,i}^k + \frac{3}{2}\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \leq \langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i \rangle \leq \rho_{y_i,r,i}^k + \frac{1}{2}\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle$$

Next we use induction to show the sign invariance. First it is clear when $k = 0$, the sign invariance is trivially satisfied. At $k = 1$, we have by the iterative update of the coefficients,

$$\rho_{y_i,r,i}^1 = \rho_{y_i,r,i}^0 + \frac{\eta}{nm}|\ell_i'^0| \langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2 = \frac{\eta}{nm}|\ell_i'^0| \langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2$$

and thus $\mathrm{sign}(\rho_{y_i,r,i}^1) = \mathrm{sign}(\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle)$. Further, by Lemma D.5, and without loss of generality that $\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \geq 0$, we have

$$\langle \mathbf{w}_{y_i,r}^1, \boldsymbol{\xi}_i \rangle \geq \rho_{y_i,r,i}^1 + \langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle - 4\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha \geq \rho_{y_i,r,i}^1 + \frac{1}{2}\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle \geq 0.$$

Similar argument also holds for $\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle < 0$. Then we show at $k = 1$, $\mathrm{sign}(\rho_{y_i,r,i}^1) = \mathrm{sign}(\langle \mathbf{w}_{y_i,r}^1, \boldsymbol{\xi}_i \rangle) = \mathrm{sign}(\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle)$. Suppose there exists a time $\widetilde{T}$ such that for all $k \leq \widetilde{T} - 1$, the sign invariance holds. Then for $k = \widetilde{T}$, suppose $\mathrm{sign}(\langle \mathbf{w}_{y_i,r}^{\widetilde{T}-1}, \boldsymbol{\xi}_i \rangle) = \mathrm{sign}(\rho_{y_i,r,i}^{\widetilde{T}-1}) = \mathrm{sign}(\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle) = +1$,

$$\begin{aligned}
\rho_{y_i,r,i}^{\widetilde{T}} &= \rho_{y_i,r,i}^{\widetilde{T}-1} + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}| \langle \mathbf{w}_{y_i,r}^{\widetilde{T}-1}, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2 \\
&\geq \rho_{y_i,r,i}^{\widetilde{T}-1} + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}| \big(\rho_{y_i,r,i}^{\widetilde{T}-1} + \langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle - 4\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha\big) \|\boldsymbol{\xi}_i\|^2 \\
&\geq \rho_{y_i,r,i}^{\widetilde{T}-1} + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}| \big(\rho_{y_i,r,i}^{\widetilde{T}-1} + \frac{1}{2}\langle \mathbf{w}_{y_i,r}^0, \boldsymbol{\xi}_i \rangle\big) \|\boldsymbol{\xi}_i\|^2 \\
&\geq 0
\end{aligned}$$

Further,

$$\langle \mathbf{w}_{y_i,r}^{\widetilde{T}}, \boldsymbol{\xi}_i \rangle \geq \rho_{y_i,r,i}^{\widetilde{T}} + \langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i \rangle - 4\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha \geq \rho_{y_i,r,i}^{\widetilde{T}} + \frac{1}{2}\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i \rangle \geq 0.$$

and thus completes the induction that $\mathrm{sign}(\langle \mathbf{w}_{y_i,r}^{\widetilde{T}}, \boldsymbol{\xi}_i \rangle) = \mathrm{sign}(\rho_{y_i,r,i}^{\widetilde{T}}) = \mathrm{sign}(\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i \rangle)$. Similar argument holds when $\mathrm{sign}(\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i \rangle) = -1$. □

We also derive the following concentration result for the average noise inner product at initialization.

Lemma D.16. *Suppose $\delta > 0$ and $m = \Omega(\log(n/\delta))$. Then with probability at least $1-\delta$, we have for all $j = \pm 1, i \in [n]$*

$$\sigma_0 \sigma_\xi \sqrt{d}/2 \leq \frac{1}{m}\sum_{r=1}^{n} |\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\xi}_i \rangle| \leq \sigma_0 \sigma_\xi \sqrt{d}$$

Proof of Lemma D.16. First notice that for any $i \in [n]$, $\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\xi}_i \rangle \sim \mathcal{N}(0, \sigma_0^2 \|\boldsymbol{\xi}_i\|^2)$ and thus we have $\mathbb{E}[|\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\xi}_i \rangle|] = \sigma_0 \|\boldsymbol{\xi}_i\| \sqrt{2/\pi}$. By sub-Gaussian tail bound, with probability at least $1-\delta/(2n)$, for any $i \in [n]$

$$\left| \frac{1}{m}\sum_{r=1}^{m} |\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\xi}_i \rangle| - \sigma_0 \|\boldsymbol{\xi}_i\| \sqrt{2/\pi} \right| \leq \sqrt{\frac{2\sigma_0^2 \|\boldsymbol{\xi}_i\|^2 \log(4n/\delta)}{m}}$$

Choosing $m = \Omega(\log(n/\delta))$, we have

$$\sigma_0 \|\boldsymbol{\xi}_i\| \sqrt{2/\pi} 0.99 \leq \frac{1}{m}\sum_{r=1}^{n} |\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\xi}_i \rangle| \leq \sigma_0 \|\boldsymbol{\xi}_i\| \sqrt{2/\pi} 1.01.$$

Because from Lemma C.2, we have $0.99\sigma_\xi \sqrt{d} \leq \|\boldsymbol{\xi}_i\| \leq 1.01\sigma_\xi\sqrt{d}$ by choosing $d = \widetilde{\Omega}(1)$ sufficiently large. Then we have $\sigma_0\sigma_\xi\sqrt{d}/2 \leq \frac{1}{m}\sum_{r=1}^{n} |\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\xi}_i \rangle| \leq \sigma_0\sigma_\xi\sqrt{d}$. Finally taking the union bound for all $j = \pm 1, i \in [n]$ completes the proof. □

We have established several preliminary lemmas that hold with high probability, including Lemma C.1, Lemma C.2, Lemma D.1, Lemma D.14, Lemma D.16. We let $\mathcal{E}_{\mathrm{prelim}}$ be the event such that all the results in these lemmas hold for a given δ. Then by applying union bound, we have $\mathbb{P}(\mathcal{E}_{\mathrm{prelim}}) \geq 1 - 5\delta$. The subsequent analysis are conditioned on the event $\mathcal{E}_{\mathrm{prelim}}$.

D.4.1 First stage

Theorem D.3. *Under Condition D.1, suppose $n^{-1} \cdot \mathrm{SNR}^{-2} \geq C'$ for some constant $C' > 0$. Then there exists a time $T_1 = \widetilde{\Theta}(\eta^{-1} nm \sigma_\xi^{-2} d^{-1})$, such that (1) $\max_r |\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i \rangle| \geq 2$ for all $i \in [n]$, (2) $\frac{1}{m}\sum_{r=1}^{m} |\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i \rangle| \geq 4$ for all $i \in [n]$ and (3) $\max_{j,r,y} |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_y \rangle| = \widetilde{O}(n^{-1/2})$.*

Proof of Theorem D.3. We first bound the growth of signal as follows. From the gradient descent update, we have

$$\begin{aligned} |\langle \mathbf{w}_{j,r}^{k}, \boldsymbol{\mu}_j \rangle| &= |\langle \mathbf{w}_{j,r}^{k-1}, \boldsymbol{\mu}_j \rangle| + \frac{\eta |\mathcal{S}_j|}{nm} |\langle \mathbf{w}_{j,r}^{k-1}, \boldsymbol{\mu}_j \rangle| \|\boldsymbol{\mu}\|^2 \\ &\leq \Big(1 + 0.51\frac{\eta\|\boldsymbol{\mu}\|^2}{m}\Big) |\langle \mathbf{w}_{j,r}^{k-1}, \boldsymbol{\mu}_j \rangle| \\ &\leq \Big(1 + 0.51\frac{\eta\|\boldsymbol{\mu}\|^2}{m}\Big)^k |\langle \mathbf{w}_{j,r}^{0}, \boldsymbol{\mu}_j \rangle| \end{aligned} \tag{15}$$

where the first inequality is by $|\ell_i'^k| \leq 1$ and the second inequality is by Lemma C.1 with $n = \widetilde{\Omega}(1)$ sufficiently large.

On the other hand, for the growth of noise, we have from the inner product update, for any $i \in [n]$

$$\begin{aligned}\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i \rangle &= \langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \sum_{i'=1}^{n} \ell_{i'}'^{k-1} \langle \mathbf{w}_{j,r}^{k-1}, \boldsymbol{\xi}_{i'} \rangle \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\ &= \Big(1 - \frac{\eta}{nm} \ell_i'^{k} \|\boldsymbol{\xi}_i\|^2 \Big) \langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \sum_{i' \neq i} \ell_{i'}'^{k-1} \langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_{i'} \rangle \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle\end{aligned}$$

Then this suggests

$$|\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i \rangle| \geq \Big(1 - \frac{\eta}{nm} \ell_i'^{k} \|\boldsymbol{\xi}_i\|^2 \Big) |\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i \rangle| - \frac{\eta}{nm} \sum_{i' \neq i} |\ell_{i'}'^{k-1}| \cdot |\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_{i'} \rangle| \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \quad (16)$$

We first prove for any $i \in [n]$, $\max_r |\langle \mathbf{w}_{y_i,r}^{k+1}, \boldsymbol{\xi}_i \rangle| \geq \max_r |\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i \rangle| \geq \max_r |\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i \rangle|$ for all $k \leq T_1$. We prove such a result by induction. It is clear that at $k = 0$, the result is satisfied. Now suppose there exists an iteration $\tilde{k}$ such that

$$\max_r |\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i \rangle| \geq \max_r |\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i \rangle| \geq \sigma_0 \sigma_\xi \sqrt{d}/4$$

for all $k \leq \tilde{k} - 1$, where the last inequality is by Lemma D.1. Then we can bound based on Lemma D.7 and Lemma C.2, we have for any $i' \neq i \in [n]$ and

$$\begin{aligned}\frac{n |\ell_{i'}'^{\tilde{k}-1}| \cdot |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \cdot |\langle \mathbf{w}_{y_i,r}^{\tilde{k}-1}, \boldsymbol{\xi}_{i'} \rangle|}{|\ell_i'^{\tilde{k}-1}| \cdot \|\boldsymbol{\xi}_i\|^2 \max_r |\langle \mathbf{w}_{y_i,r}^{\tilde{k}-1}, \boldsymbol{\xi}_i \rangle|} &\leq \frac{2\sigma_\xi^2 \sqrt{d \log(4n^2/\delta)}}{C_\ell 0.99 \sigma_\xi^2 d} n \alpha \sigma_0^{-1} \sigma_\xi^{-1} d^{-1/2} \\ &= 8.4 C_\ell^{-1} n \alpha \frac{\sqrt{\log(4n^2/\delta)}}{d \sigma_0 \sigma_\xi} \\ &\leq 0.01 \end{aligned} \quad (17)$$

where we use the lower and upper bound on loss derivatives during the first stage, as well as Lemma C.2 and Lemma D.1. The last inequality is by $\sigma_0 \geq 840 n C_\ell^{-1} d^{-1} \sigma_\xi^{-1} \alpha \sqrt{\log(4n^2/\delta)}$. Then we have

$$\begin{aligned}\max_r |\langle \mathbf{w}_{y_i,r}^{\tilde{k}}, \boldsymbol{\xi}_i \rangle| &\geq \Big(1 - \frac{\eta}{nm} \ell_i'^{\tilde{k}-1} \|\boldsymbol{\xi}_i\|^2 \Big) \max_r |\langle \mathbf{w}_{y_i,r}^{\tilde{k}-1}, \boldsymbol{\xi}_i \rangle| - \frac{\eta}{nm} \sum_{i' \neq i} |\ell_{i'}'^{\tilde{k}-1}| \cdot |\langle \mathbf{w}_{y_i,r}^{\tilde{k}-1}, \boldsymbol{\xi}_{i'} \rangle| \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \\ &\geq \Big(1 + \frac{\eta}{nm} 0.99 |\ell_i'^{\tilde{k}-1}| \|\boldsymbol{\xi}_i\|^2 \Big) \max_r \Big| \langle \mathbf{w}_{y_i,r}^{\tilde{k}-1}, \boldsymbol{\xi}_i \rangle \Big| \\ &\geq \max_r \Big| \langle \mathbf{w}_{y_i,r}^{\tilde{k}-1}, \boldsymbol{\xi}_i \rangle \Big| \\ &\geq \max_r |\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i \rangle|\end{aligned}$$

Let $B_i^k := \max_r |\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i \rangle|$ and we obtain for any $k \leq T_1$,

$$\begin{aligned}B_i^k \geq \Big(1 + \frac{\eta}{nm} 0.99 |\ell_i'^{\tilde{k}-1}| \|\boldsymbol{\xi}_i\|^2 \Big) B_i^{k-1} &\geq \Big(1 + \frac{\eta \sigma_\xi^2 d}{nm} 0.98 C_\ell \Big) B_i^{k-1} \\ &\geq \Big(1 + \frac{\eta \sigma_\xi^2 d}{nm} 0.98 C_\ell \Big)^k B_i^0 \\ &\geq \Big(1 + \frac{\eta \sigma_\xi^2 d}{nm} 0.98 C_\ell \Big)^k \sigma_0 \sigma_\xi \sqrt{d}/4\end{aligned}$$

where we use (17), which holds for iteration k and Lemma D.1. Consider

$$T_1 = \log(8 \sigma_0^{-1} \sigma_\xi^{-1} d^{-1/2}) / \log\big(1 + \frac{\eta \sigma_\xi^2 d}{nm} 0.98 C_\ell\big) = \Theta(\eta^{-1} n m \sigma_\xi^{-2} d^{-1} \log(8 \sigma_0^{-1} \sigma_\xi^{-1} d^{-1/2}))$$

for η sufficiently small. Then it can be shown that

$$B_i^{T_1} = \max_r |\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i \rangle| \geq 2$$

In addition, we show the average also grows to a constant order with a similar argument. In particular, from (16), we have

$$\frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i\rangle| \geq \Big(1 - \frac{\eta}{nm}\ell_i'^{k}\|\boldsymbol{\xi}_i\|^2\Big)\frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i\rangle|$$
$$- \frac{\eta}{nm}\sum_{i'\neq i}|\ell_{i'}'^{k-1}| \cdot \frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_{i'}\rangle| \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle|$$

Using a similar induction argument, we can show

$$\frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{k}, \boldsymbol{\xi}_i\rangle| \geq \frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{k-1}, \boldsymbol{\xi}_i\rangle| \geq \frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{0}, \boldsymbol{\xi}_i\rangle| \geq \sigma_0\sigma_\xi\sqrt{d}/2$$

for all $k \leq T_1$, where the last inequality follows from Lemma D.16. Then we can show at T_1,

$$\frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i\rangle| \geq \left(1 + \frac{\eta\sigma_\xi^2 d}{nm}0.98C_\ell\right)^{T_1}\sigma_0\sigma_\xi\sqrt{d}/2 \geq 4.$$

In the meantime, (15) allows to bound the growth of signal learning as for any $j = \pm 1$,

$$\begin{aligned}
&\max_r |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j\rangle| \\
&\leq \Big(1 + 0.51\frac{\eta\|\boldsymbol{\mu}\|^2}{m}\Big)^{T_1}\sqrt{2\log(8m/\delta)}\sigma_0\|\boldsymbol{\mu}\| \\
&= \exp\Big(\frac{\log(1+0.51\frac{\eta\|\boldsymbol{\mu}\|^2}{m})}{\log(1+0.98\frac{\eta\sigma_\xi^2 dC_\ell}{nm})}\log\big(8\sigma_0^{-1}\sigma_\xi^{-1}d^{-1/2}\big)\Big)\sqrt{2\log(8m/\delta)}\sigma_0\|\boldsymbol{\mu}\| \\
&\leq \exp\Big(\big(0.53C_\ell^{-1}n\mathrm{SNR}^2 + \widetilde{O}(n^{-1}\mathrm{SNR}^{-2}\eta)\big)\log\big(8\sigma_0^{-1}\sigma_\xi^{-1}d^{-1/2}\big)\Big)\sqrt{2\log(8m/\delta)}\sigma_0\|\boldsymbol{\mu}\| \\
&\leq 8\sqrt{2\log(8m/\delta)}\mathrm{SNR} \\
&= \widetilde{O}(n^{-1/2})
\end{aligned}$$

where the first inequality is by Lemma D.1 and the second inequality is by Taylor expansion around $\eta = 0$. The third inequality is by choosing η sufficiently small and based on the condition that $n^{-1}\mathrm{SNR}^{-2} \geq 0.55C_\ell^{-1}$. The last equality is by the SNR condition. □

D.4.2 SECOND STAGE

We choose $\mathbf{W}^*$ to be

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^0 + 2\log(4/\epsilon)\sum_{i=1}^{n}\mathbb{1}(y_i = j)\mathrm{sign}(\langle \mathbf{w}_{j,r}^0, \boldsymbol{\xi}_i\rangle)\frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|^2}$$

First we show the invariance of sign of noise inner product after the first stage.

Lemma D.17. *Under the same settings as in Theorem D.3, we have* $\max_r |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i\rangle| \geq 1$ *and* $\frac{1}{m}\sum_{r=1}^m |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i\rangle| \geq 2$ *for all* $T_1 \leq k \leq T^*$ *and any* $i \in [n]$.

Proof of Lemma D.17. In addition to the two results, we also prove $\max_r |\rho_{y_i,r,i}^k| \geq 1.5$ and $\frac{1}{m}\sum_{r=1}^m |\rho_{y_i,r,i}^k| \geq 3$. We prove these results by induction. First, it is clear that at $k = T_1$, the bound regarding inner products are trivially satisfied by Theorem D.3. Then by Lemma D.5, we have

$$\max_r |\rho_{y_i,r,i}^{T_1}| \geq \max_r |\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i\rangle| - \beta - 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \geq 2 - 0.5 = 1.5$$
$$\frac{1}{m}\sum_{r=1}^{m}|\rho_{y_i,r,i}^{T_1}| \geq \frac{1}{m}\sum_{r=1}^{m}|\langle \mathbf{w}_{y_i,r}^{T_1}, \boldsymbol{\xi}_i\rangle| - \beta - 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \geq 4 - 1 = 3$$

where the last inequalities are by Condition D.1 for sufficiently large constant C.

Now suppose there exists a time $T_1 \leq \widetilde{T} \leq T^*$ such that the results hold for all $k \leq \widetilde{T} - 1$. Then at $k = \widetilde{T}$, recall the coefficient update as

$$\rho^{\widetilde{T}}_{y_i,r,i} = \rho^{\widetilde{T}-1}_{y_i,r,i} + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}|\langle \mathbf{w}^{\widetilde{T}-1}_{y_i,r}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2 \tag{18}$$

If $\langle \mathbf{w}^0_{y_i,r}, \boldsymbol{\xi}_i\rangle > 0$, by Lemma D.15 we have $\langle \mathbf{w}^{\widetilde{T}-1}_{y_i,r}, \boldsymbol{\xi}_i\rangle, \rho^{\widetilde{T}-1}_{y_i,r,i} > 0$. Then

$$\begin{aligned}\rho^{\widetilde{T}}_{y_i,r,i} &= \rho^{\widetilde{T}-1}_{y_i,r,i} + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}|\langle \mathbf{w}^{\widetilde{T}-1}_{y_i,r}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2 \\ &\geq \rho^{\widetilde{T}-1}_{y_i,r,i} + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}|\big(\rho^{\widetilde{T}-1}_{y_i,r,i} + \langle \mathbf{w}^0_{y_i,r}, \boldsymbol{\xi}_i\rangle - 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\big)\|\boldsymbol{\xi}_i\|^2 \\ &\geq \rho^{\widetilde{T}-1}_{y_i,r,i} + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}|\big(\rho^{\widetilde{T}-1}_{y_i,r,i} + \frac{1}{2}\langle \mathbf{w}^0_{y_i,r}, \boldsymbol{\xi}_i\rangle\big)\|\boldsymbol{\xi}_i\|^2.\end{aligned}$$

Then taking maximum over r,

$$\max_r |\rho^{\widetilde{T}}_{y_i,r,i}| \geq \max_r |\rho^{\widetilde{T}-1}_{y_i,r,i}| + \frac{\eta\|\boldsymbol{\xi}_i\|^2}{2nm}|\ell_i'^{\widetilde{T}-1}|\max_r |\rho^{\widetilde{T}-1}_{y_i,r,i}| \geq \max_r |\rho^{\widetilde{T}-1}_{y_i,r,i}| \geq 1.5$$

where the first inequality follows from $\langle \mathbf{w}^0_{y_i,r}, \boldsymbol{\xi}_i\rangle/2 \leq 0.5 \leq \max_r |\rho^{\widetilde{T}-1}_{y_i,r,i}|/2$ based on Condition D.1. Similarly, when $\langle \mathbf{w}^0_{y_i,r}, \boldsymbol{\xi}_i\rangle < 0$, we can obtain the same result. Then, we have

$$\max_r |\langle \mathbf{w}^{\widetilde{T}}_{y_i,r}, \boldsymbol{\xi}_i\rangle| \geq \max_r |\rho^{\widetilde{T}}_{y_i,r,i}| - \beta - 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \geq 1.5 - 0.5 = 1.$$

Furthermore, we prove the results for the average quantities in a similar manner. First, from the coefficient update, and by Lemma D.15, $\mathrm{sign}(\rho^{\widetilde{T}-1}_{y_i,r,i}) = \mathrm{sign}(\langle \mathbf{w}^{\widetilde{T}-1}_{y_i,r}, \boldsymbol{\xi}_i\rangle)$ and thus taking the average of absolute value on both sides of (18), we get

$$\begin{aligned}\frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}}_{y_i,r,i}| &= \frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}-1}_{y_i,r,i}| + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}|\frac{1}{m}\sum_{r=1}^m |\langle \mathbf{w}^{\widetilde{T}-1}_{y_i,r}, \boldsymbol{\xi}_i\rangle|\|\boldsymbol{\xi}_i\|^2 \\ &\geq \frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}-1}_{y_i,r,i}| + \frac{\eta}{nm}|\ell_i'^{\widetilde{T}-1}|(\frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}-1}_{y_i,r,i}| - \beta - 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\big)\|\boldsymbol{\xi}_i\|^2 \\ &\geq \frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}-1}_{y_i,r,i}| + \frac{\eta}{2nm}|\ell_i'^{\widetilde{T}-1}|\frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}-1}_{y_i,r,i}|\|\boldsymbol{\xi}_i\|^2 \\ &\geq \frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}-1}_{y_i,r,i}| \geq 3\end{aligned}$$

where we use $|a+b| = |a| + |b|$ when $\mathrm{sign}(a) = \mathrm{sign}(b)$. Then, we have

$$\frac{1}{m}\sum_{r=1}^m |\langle \mathbf{w}^{\widetilde{T}}_{y_i,r}, \boldsymbol{\xi}_i\rangle| \geq \frac{1}{m}\sum_{r=1}^m |\rho^{\widetilde{T}}_{y_i,r,i}| - \beta - 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \geq 3 - 1 = 2.$$

where the inequality is by Condition D.1. □

Lemma D.18. *Under Condition D.1, we have* $\|\mathbf{W}^{T_1} - \mathbf{W}^*\| = O(\sqrt{nm}\log(1/\epsilon)\sigma_\xi^{-1}d^{-1/2})$.

Proof of Lemma D.18. The proof follows similarly as in Lemma D.11. Let $\mathbf{P}_{\boldsymbol{\xi}}$ be the projection matrix to the direction of $\boldsymbol{\xi}$, i.e., $\mathbf{P}_{\boldsymbol{\xi}} = \frac{\boldsymbol{\xi}\boldsymbol{\xi}^\top}{\|\boldsymbol{\xi}\|^2}$. Then we can represent

$$\mathbf{w}^k_{j,r} - \mathbf{w}^0_{j,r} = \mathbf{P}_{\boldsymbol{\mu}_1}(\mathbf{w}^k_{j,r} - \mathbf{w}^0_{j,r}) + \mathbf{P}_{\boldsymbol{\mu}_{-1}}(\mathbf{w}^k_{j,r} - \mathbf{w}^0_{j,r}) + \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}(\mathbf{w}^k_{j,r} - \mathbf{w}^0_{j,r})$$

$$+\Big(\mathbf{I}-\mathbf{P}_{\boldsymbol{\mu}_1}-\mathbf{P}_{\boldsymbol{\mu}_{-1}}-\sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}\Big)(\mathbf{w}_{j,r}^k-\mathbf{w}_{j,r}^0).$$

By the scale difference at T_1 and the fact that gradient descent only updates in the direction of $\boldsymbol{\mu}_j$, $j=\pm 1$ and $\boldsymbol{\xi}_i$, we can bound

$$\begin{aligned}
&\|\mathbf{W}^{T_1}-\mathbf{W}^0\|^2\\
&\le \sum_{j=\pm1, r\in[m]}\Big(\frac{\langle \mathbf{w}_{j,r}^{T_1}-\mathbf{w}_{j,r}^0,\boldsymbol{\mu}_1\rangle^2}{\|\boldsymbol{\mu}\|^2}+\frac{\langle \mathbf{w}_{j,r}^{T_1}-\mathbf{w}_{j,r}^0,\boldsymbol{\mu}_{-1}\rangle^2}{\|\boldsymbol{\mu}\|^2}+\sum_{i=1}^n\frac{\langle \mathbf{w}_{j,r}^{T_1}-\mathbf{w}_{j,r}^0,\boldsymbol{\xi}_i\rangle^2}{\|\boldsymbol{\xi}_i\|^2}\Big)\\
&\quad+\sum_{j=\pm1, r\in[m]}\Big\|\Big(\mathbf{I}-\mathbf{P}_{\boldsymbol{\mu}_1}-\mathbf{P}_{\boldsymbol{\mu}_{-1}}-\sum_{i=1}^n\mathbf{P}_{\boldsymbol{\xi}_i}\Big)(\mathbf{w}_{j,r}^{T_1}-\mathbf{w}_{j,r}^0)\Big\|^2\\
&\le 2m\Big(\frac{2\langle \mathbf{w}_{j,r}^{T_1},\boldsymbol{\mu}_j\rangle^2+2\langle \mathbf{w}_{j,r}^{T_1},\boldsymbol{\mu}_{-j}\rangle^2+2\langle \mathbf{w}_{j,r}^{0},\boldsymbol{\mu}_{-j}\rangle^2+2\langle \mathbf{w}_{j,r}^{0},\boldsymbol{\mu}_{j}\rangle^2}{\|\boldsymbol{\mu}\|^2}\\
&\quad+n\max_{j,r}\frac{2\langle \mathbf{w}_{j,r}^{T_1},\boldsymbol{\xi}_i\rangle^2+2\langle \mathbf{w}_{j,r}^{0},\boldsymbol{\xi}_i\rangle^2}{\|\boldsymbol{\xi}_i\|^2}\Big)+\sum_{j=\pm1, r\in[m]}\Big\|\Big(\mathbf{I}-\mathbf{P}_{\boldsymbol{\mu}_1}-\mathbf{P}_{\boldsymbol{\mu}_{-1}}-\sum_{i=1}^n\mathbf{P}_{\boldsymbol{\xi}_i}\Big)(\mathbf{w}_{j,r}^{T_1}-\mathbf{w}_{j,r}^0)\Big\|^2\\
&\le O(mn\sigma_\xi^{-2}d^{-1})
\end{aligned}$$

where we use the scale difference at T_1. Therefore,

$$\begin{aligned}
\|\mathbf{W}^{T_1}-\mathbf{W}^*\| &\le \|\mathbf{W}^{T_1}-\mathbf{W}^0\|+\|\mathbf{W}^0-\mathbf{W}^*\|\\
&\le O(\sqrt{mn}\sigma_\xi^{-1}d^{-1/2})+O(\sqrt{nm}\log(1/\epsilon)\sigma_\xi^{-1}d^{-1/2})\\
&\le O(\sqrt{nm}\log(1/\epsilon)\sigma_\xi^{-1}d^{-1/2})
\end{aligned}$$

where we use the definition of $\mathbf{W}^*$. □

Lemma D.19. *Under Condition D.1, we have for all $T_1\le k\le T^*$*

$$\|\mathbf{W}^k-\mathbf{W}^*\|^2-\|\mathbf{W}^{k+1}-\mathbf{W}^*\|^2\ge 2\eta L_S(\mathbf{W}^t)-\eta\epsilon$$

Proof of Lemma D.19. The proof follows from similar arguments as for Lemma D.12. We first obtain a lower bound on $y_i\langle\nabla f(\mathbf{W}^t,\mathbf{x}_i),\mathbf{W}^*\rangle$ for any $i\in[n]$ for all $T_1\le k\le T^*$.

$$\begin{aligned}
y_i\langle\nabla f(\mathbf{W}^k,\mathbf{x}_i),\mathbf{W}^*\rangle &= \frac{1}{m}\sum_{j,r} jy_i\langle \mathbf{w}_{j,r}^k,\boldsymbol{\mu}_{y_i}\rangle\langle\boldsymbol{\mu}_{y_i},\mathbf{w}_{j,r}^*\rangle+\frac{1}{m}\sum_{j,r} jy_i\langle \mathbf{w}_{j,r}^k,\boldsymbol{\xi}_i\rangle\langle\boldsymbol{\xi}_i,\mathbf{w}_{j,r}^*\rangle\\
&= \frac{1}{m}\sum_{j,r} jy_i\langle \mathbf{w}_{j,r}^k,\boldsymbol{\mu}_{y_i}\rangle\langle\boldsymbol{\mu}_{y_i},\mathbf{w}_{j,r}^0\rangle+\frac{1}{m}\sum_{j,r} jy_i\langle \mathbf{w}_{j,r}^k,\boldsymbol{\xi}_i\rangle\langle\boldsymbol{\xi}_i,\mathbf{w}_{j,r}^0\rangle\\
&\quad+\frac{1}{m}\sum_{j=\pm1}\sum_{r=1}^m\sum_{i'=1}^n jy_i\langle \mathbf{w}_{j,r}^k,\boldsymbol{\xi}_i\rangle\mathbb{1}(j=y_{i'})\frac{\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|^2}2\log(4/\epsilon)\\
&= \underbrace{\frac{1}{m}\sum_{r=1}^m|\langle \mathbf{w}_{y_i,r}^k,\boldsymbol{\xi}_i\rangle|2\log(4/\epsilon)}_{A_9}+\underbrace{\frac{1}{m}\sum_{j,r}\sum_{i'\neq i}\langle \mathbf{w}_{y_i,r}^k,\boldsymbol{\xi}_i\rangle 2\log(4/\epsilon)\frac{\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|^2}}_{A_{10}}\\
&\quad+\underbrace{\frac{1}{m}\sum_{j,r} jy_i\langle \mathbf{w}_{j,r}^k,\boldsymbol{\mu}_{y_i}\rangle\langle\boldsymbol{\mu}_{y_i},\mathbf{w}_{j,r}^0\rangle}_{A_{11}}+\underbrace{\frac{1}{m}\sum_{j,r} jy_i\langle \mathbf{w}_{j,r}^k,\boldsymbol{\xi}_i\rangle\langle\boldsymbol{\xi}_i,\mathbf{w}_{j,r}^0\rangle}_{A_{12}}
\end{aligned}$$

where the second equality is by definition of $\mathbf{W}^*$. The third equality is by Lemma D.17 and Lemma D.15 on the sign invariance. We next bound based on the scale difference and Lemma C.2,

$$|A_{10}|=\widetilde{O}(nd^{-1/2}),\quad |A_{11}|=\widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|),\quad |A_{12}|\le\widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$$

where we use the global bound on the inner product by $\widetilde{O}(1)$. Next, by Theorem D.3 and Lemma D.17, we can show $\frac{1}{m}\sum_{r=1}^m |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\mu}_{y_i}\rangle| \geq 2$ for all $i \in [n], k \geq T_1$ and we can bound

$$A_9 \geq 4\log(4/\epsilon)$$

Combining the bound for $A_9, A_{10}, A_{11}, A_{12}$, we have

$$y_i\langle \nabla f(\mathbf{W}^k, \mathbf{x}_i), \mathbf{W}^*\rangle \geq 2\log(4/\epsilon) \tag{19}$$

where we bound $|A_{10}| + |A_{11}| + |A_{12}| \leq 2\log(4/\epsilon)$ under Condition D.1.

Further, we derive

$$\begin{aligned}
&\|\mathbf{W}^k - \mathbf{W}^*\|^2 - \|\mathbf{W}^{k+1} - \mathbf{W}^*\|^2 \\
&= 2\eta\langle \nabla L_S(\mathbf{W}^k), \mathbf{W}^k - \mathbf{W}^*\rangle - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2 \\
&= \frac{2\eta}{n}\sum_{i=1}^n \ell_i'^k y_i\big(2f(\mathbf{W}^k, \mathbf{x}_i) - \langle \nabla f(\mathbf{W}^k, \mathbf{x}_i), \mathbf{W}^*\rangle\big) - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2 \\
&\geq \frac{2\eta}{n}\sum_{i=1}^n \ell_i'^k\big(2y_i f(\mathbf{W}^k, \mathbf{x}_i) - 2\log(2/\epsilon)\big) - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2 \\
&\geq \frac{4\eta}{n}\sum_{i=1}^n \big(\ell(y_i f(\mathbf{W}^k, \mathbf{x}_i)) - \epsilon/4\big) - \eta^2\|\nabla L_S(\mathbf{W}^k)\|^2 \\
&\geq 2\eta L_S(\mathbf{W}^k) - \eta\epsilon
\end{aligned}$$

where the first inequality is by (19) and the second inequality is by convexity of cross-entropy function and the last inequality is by Lemma D.8. □

Theorem D.4. *Under the same settings as in Theorem D.3, let* $T = T_1 + \lfloor \frac{\|\mathbf{W}^{T_1} - \mathbf{W}^*\|^2}{\eta\epsilon} \rfloor = T_1 + O(\eta^{-1}\epsilon^{-1}mn\sigma_\xi^{-2}d^{-1})$. *Then we have*

- *there exists* $T_1 \leq k \leq T$ *such that* $L_S(\mathbf{W}^k) \leq 0.1$.
- $\max_{j,r,y} |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_y\rangle| = o(1)$ *for all* $T_1 \leq k \leq T$.
- $\max_r |\langle \mathbf{w}_{y_i,r}^k, \boldsymbol{\xi}_i\rangle| \geq 1$ *for all* $i \in [n], T_1 \leq k \leq T$.

Proof of Theorem D.4. The proof is similar as in Theorem D.2. By Lemma D.19, for any $T_1 \leq k \leq T$, we have

$$\|\mathbf{W}^k - \mathbf{W}^*\|^2 - \|\mathbf{W}^{k+1} - \mathbf{W}^*\|^2 \geq 2\eta L_S(\mathbf{W}^k) - \eta\epsilon$$

for all $s \leq k$. Then summing over the inequality gives

$$\frac{1}{T - T_1 + 1}\sum_{k=T_1}^T L_S(\mathbf{W}^k) \leq \frac{\|\mathbf{W}^{T_1} - \mathbf{W}^*\|^2}{2\eta(T - T_1 + 1)} + \frac{\epsilon}{2} \leq \epsilon$$

where the last inequality is by the choice $T = T_1 + \lfloor \frac{\|\mathbf{W}^{T_1} - \mathbf{W}^*\|^2}{\eta\epsilon} \rfloor = T_1 + \Omega(\eta^{-1}\epsilon^{-1}nm^3\log(1/\epsilon)\sigma_\xi^{-2}d^{-1})$. Then we can claim that there exists a $k \in [T_1, T]$ such that $L_S(\mathbf{W}^k) \leq \epsilon$. Setting $\epsilon = 0.1$ shows the desired convergence.

Next, we show the upper bound on $\max_{j,y,r} |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_y\rangle|$ for all $k \in [T_1, T]$. Notice that by Proposition D.1, we already have $\max_{j,r} |\langle \mathbf{w}_{-j,r}^k, \boldsymbol{\mu}_j\rangle| \leq \vartheta$, where we let

$$\vartheta := 3\max\{\max_{j,r} |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j\rangle|, \beta, 4\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha\} = \widetilde{O}(\max\{n^{-1/2}, \sigma_0\sigma_\xi\sqrt{d}, \sigma_0\|\boldsymbol{\mu}\|, nd^{-1/2}\})$$

Subsequently, we use induction to prove $\max_{j,r} |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j\rangle| \leq 2\vartheta$. First we notice that

$$\sum_{k=T_1}^T L_S(\mathbf{W}^k) \leq \frac{\|\mathbf{W}^{T_1} - \mathbf{W}^*\|^2}{\eta} = O(\eta^{-1}nm\sigma_\xi^{-2}d^{-1}) \tag{20}$$

where the equality is by Lemma D.11 where we choose $\epsilon = 0.1$.

At $k = T_1$, we have $\max_{j,r} |\langle \mathbf{w}_{j,r}^{T_1}, \boldsymbol{\mu}_j \rangle| \leq \vartheta \leq 2\vartheta$. Suppose there $\widetilde{T} \in [T_1, T]$ such that $\max_{r,i} |\rho_{y_i,r,i}^{T_1}| \leq 2\vartheta$ for all $k \in [T_1, \widetilde{T} - 1]$. Now we let $\Psi^k := \max_{j,r} |\langle \mathbf{w}_{j,r}^k, \boldsymbol{\mu}_j \rangle|$ and thus by the update of inner product

$$
\begin{aligned}
\Psi^{k+1} &\leq \Psi^k + \frac{\eta}{nm} \sum_{i \in \mathcal{S}_j} |\ell_i'^k| \Psi^k \|\boldsymbol{\mu}\|^2 \\
&\leq \Psi^k + \frac{\eta}{nm} \sum_{i \in [n]} \ell_i^k \Psi^k \|\boldsymbol{\mu}\|^2 \\
&= \Psi^k + \frac{2\eta \|\boldsymbol{\mu}\|^2}{m} L_S(\mathbf{W}^k) \Psi^k.
\end{aligned}
$$

where we use $|\ell'| \leq \ell$ in the second inequality. Taking the summation from T_1 to $\widetilde{T}$ gives

$$
\begin{aligned}
\Psi^{\widetilde{T}} &\leq \Psi^{T_1} + \frac{2\eta \|\boldsymbol{\mu}\|^2}{m} \sum_{k=T_1}^{\widetilde{T}-1} L_S(\mathbf{W}^k) \cdot m^2 \vartheta \\
&\leq \Psi^{T_1} + O(n\mathrm{SNR}^2) \cdot 2\vartheta \\
&\leq 2\vartheta
\end{aligned}
$$

where the second inequality is by (20) and the last inequality is by $n^{-1} \cdot \mathrm{SNR}^{-2} \geq C'$ for sufficiently large constant $C' > 0$. The lower bound for noise inner product is directly from Lemma D.17. □

E Diffusion model

For the analysis of diffusion model, we restate 3.1 specifically for the case of diffusion model.

Condition E.1. *Suppose there exists a sufficiently large constant $C > 0$ such that the following hold:*

1. *The dimension d satisfies $d = \widetilde{\Omega}(\max\{\|\boldsymbol{\mu}\|^2, n^2\})$.*
2. *The training sample and network width satisfy $m, n = \widetilde{\Omega}(1)$.*
3. *The initialization variation σ_0 satisfies $\sigma_0 \leq \widetilde{O}(\min\{\|\boldsymbol{\mu}\|^{-1}, \sigma_\xi^{-1} d^{-1/2}, m^{-3} d^{-1/2}\})$.*
4. *The noise coefficients α_t, β_t satisfy $\alpha_t, \beta_t = \Theta(1)$.*

E.1 Useful lemmas

Lemma E.1. *Suppose $\delta > 0$. Then with probability at least $1 - \delta$, for any t,*

$$
\begin{aligned}
&\sigma_0^2 d(1 - \widetilde{O}(d^{-1/2})) \leq \|\mathbf{w}_{r,t}^0\|^2 \leq \sigma_0^2 d(1 + \widetilde{O}(d^{-1/2})) \\
&|\langle \mathbf{w}_{r,t}^0, \boldsymbol{\mu}_j \rangle| \leq \sqrt{2\log(16m/\delta)} \sigma_0 \|\boldsymbol{\mu}_j\|, \\
&|\langle \mathbf{w}_{r,t}^0, \boldsymbol{\xi}_i \rangle| \leq 2\sqrt{\log(16mn/\delta)} \sigma_0 \sigma_\xi \sqrt{d} \\
&|\langle \mathbf{w}_{r,t}^0, \mathbf{w}_{r',t}^0 \rangle| \leq 2\sqrt{\log(16m^2/\delta)} \sigma_0^2 \sqrt{d}, \quad r \neq r'
\end{aligned}
$$

for all $r, r' \in [m]$ and $i \in [n]$. and $j = 1, 2$

Proof of Lemma E.1. The proof is the same as in (Kou et al., 2023) and we include here for completeness. Because at initialization $\mathbf{w}_{r,t}^0 \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$, by Bernstein's inequality, with probability at least $1 - \delta/(8m)$, we have

$$
|\|\mathbf{w}_{r,t}^0\|_2^2 - \sigma_0^2 d| = O(\sigma_0^2 \sqrt{d \log(16m/\delta)})
$$

Then taking the union bound yields for all $r \in [m]$, we have with probability at least $1 - \delta/4$ that

$$
\sigma_0^2 d(1 - \widetilde{O}(d^{-1/2})) \leq \|\mathbf{w}_{r,t}^0\|_2^2 \leq \sigma_0^2 d(1 + \widetilde{O}(d^{-1/2})).
$$

Further, because $\langle \mathbf{w}^0_{r,t}, \boldsymbol{\mu}_j \rangle \sim \mathcal{N}(0, \sigma_0^2 \|\boldsymbol{\mu}_j\|_2^2)$ for $j = 1, 2$, then by Gaussian tail bound and union bound, we have with probability at least $1 - \delta/4$, for all $j = 1, 2$, $r \in [m]$,

$$|\langle \mathbf{w}^0_{r,t}, \boldsymbol{\mu}_j \rangle| \le \sqrt{2\log(16m/\delta)}\sigma_0 \|\boldsymbol{\mu}\|_2$$

Finally, following similar argument and noticing that $\|\boldsymbol{\xi}_i\|_2^2 = \Theta(\sigma_\xi^2 d)$ and $\|\mathbf{w}^0_{r,t}\|_2^2 = \Theta(\sigma_0^2 d)$, we have with probability at least $1 - \delta/4$ that for all $i \in [n]$, $|\langle \mathbf{w}^0_{r,t}, \boldsymbol{\xi}_i \rangle| \le 2\sqrt{\log(16mn/\delta)}\sigma_0\sigma_\xi\sqrt{d}$ and $|\langle \mathbf{w}^0_{r,t}, \mathbf{w}^0_{r',t} \rangle| \le 2\sqrt{\log(16m^2/\delta)}\sigma_0^2\sqrt{d}$. □

E.2 Derivation of loss function and gradient

We first simplify the objective through taking the expectation over the added diffusion noise.

Lemma E.2. *The DDPM loss can be simplified under expectation as*

$$L(\mathbf{W}_t) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j\in[2]}\Big(d + L^{(j)}_{1,i}(\mathbf{W}_t) + L^{(j)}_{2,i}(\mathbf{W}_t)\Big),$$

where

$$L^{(j)}_{1,i}(\mathbf{W}_t) = \frac{1}{m}\sum_{r=1}^{m}\|\mathbf{w}_{r,t}\|^2\Big(\alpha_t^4\langle \mathbf{w}_{r,t}, \mathbf{x}^{(j)}_{0,i}\rangle^4 + 6\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}^{(j)}_{0,i}\rangle^2\|\mathbf{w}_{r,t}\|^2 + 3\beta_t^4\|\mathbf{w}_{r,t}\|^4 - 4\sqrt{m}\alpha_t\beta_t\langle \mathbf{w}_{r,t}, \mathbf{x}^{(j)}_{0,i}\rangle\Big)$$

$$L^{(j)}_{2,i}(\mathbf{W}_t) = \frac{2}{m}\sum_{r=1}^{m}\sum_{r'\neq r}\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle\Big(\big(\alpha_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}^{(j)}_{0,i}\rangle^2 + \beta_t^2\|\mathbf{w}_{r,t}\|^2\big)\big(\alpha_t^2\langle \mathbf{w}_{r',t}, \mathbf{x}^{(j)}_{0,i}\rangle^2 + \beta_t^2\|\mathbf{w}_{r',t}\|^2\big) + 2\beta_t^4\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle^2 + 4\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r',t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle\Big)$$

corresponding to the learning of r-th neuron and alignment of r-th neuron with other neurons respectively.

Proof of Lemma E.2. Without loss of generality, we consider for a single sample $\mathbf{x}_{t,i}$. We first write the objective as

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{f}_p(\mathbf{W}_t, \mathbf{x}^{(p)}_{t,i}) - \boldsymbol{\epsilon}^{(p)}_{t,i}\|^2 \\
&= \underbrace{\mathbb{E}\|\boldsymbol{\epsilon}^{(p)}_{t,i}\|^2}_{I_1} + \underbrace{\mathbb{E}\left\|\frac{1}{\sqrt{m}}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{t,i}\rangle)\mathbf{w}_{r,t}\right\|^2}_{I_2} - 2\underbrace{\mathbb{E}\left[\frac{1}{\sqrt{m}}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{t,i}\rangle)\langle \mathbf{w}_{r,t}, \boldsymbol{\epsilon}_{t,i}\rangle\right]}_{I_3}
\end{aligned}$$

where we omit the subscript for the expectation for clarity.

First, we can see $I_1 = d$. Then

$$\begin{aligned}
I_3 &= \frac{1}{\sqrt{m}}\sum_{r=1}^{m}\mathbb{E}\Big[((\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{t,i}\rangle))^2\langle \mathbf{w}_{r,t}, \boldsymbol{\epsilon}_{t,i}\rangle\Big] \\
&= \frac{1}{\sqrt{m}}\sum_{r=1}^{m}\sum_{i'=1}^{d}\mathbb{E}\Big[((\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{t,i}\rangle))^2\mathbf{w}_{r,t}[i']\boldsymbol{\epsilon}_{t,i}[i']\Big] \\
&= \frac{2\beta_t}{\sqrt{m}}\sum_{r=1}^{m}\sum_{i'=1}^{d}\mathbb{E}\Big[((\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{t,i}\rangle))\mathbf{w}_{r,t}[i']^2\Big] \\
&= \frac{2\beta_t}{\sqrt{m}}\sum_{r=1}^{m}\|\mathbf{w}_{r,t}\|^2\mathbb{E}\Big[\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{t,i}\rangle\Big] \\
&= \frac{2\alpha_t\beta_t}{\sqrt{m}}\sum_{r=1}^{m}\|\mathbf{w}_{r,t}\|^2\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle
\end{aligned}$$

where the third equality uses Stein's Lemma.

Next, we consider I_2 by writing

$$I_2 = \frac{1}{m}\sum_{r=1}^{m}\mathbb{E}\big[(\langle \mathbf{w}_{r,t}, \mathbf{x}_{t,i}^{(p)}\rangle)^4\big]\,\|\mathbf{w}_{r,t}\|^2 + \frac{2}{m}\sum_{r=1}^{m}\sum_{r'\neq r}\mathbb{E}\big[(\langle \mathbf{w}_{r,t}, \mathbf{x}_{t,i}^{(p)}\rangle)^2(\langle \mathbf{w}_{r',t}, \mathbf{x}_{t,i}^{(p)}\rangle)^2\big]\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle.$$

Next, we compute the two terms $\mathbb{E}\big[(\langle \mathbf{w}_{r,t}, \mathbf{x}_{t,i}^{(p)}\rangle)^4\big]$ and $\mathbb{E}\big[(\langle \mathbf{w}_{r,t}, \mathbf{x}_{t,i}^{(p)}\rangle)^2(\langle \mathbf{w}_{r',t}, \mathbf{x}_{t,i}^{(p)}\rangle)^2\big]$ respectively. For notation simplicity, we let $a_r := \alpha_t\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}^{(p)}\rangle$, $b_r := \beta_t\|\mathbf{w}_{r,t}\|$ and $z_r := \beta_t\langle \mathbf{w}_{r,t}, \boldsymbol{\epsilon}_{t,i}\rangle$. We first compute $\mathbb{E}[z_r] = 0$ and $\mathbb{E}[z_r^2] = \beta_t^2\|\mathbf{w}_{r,t}\|^2$, $\mathbb{E}[z_r^4] = 3\beta_t^4\|\mathbf{w}_{r,t}\|^4$. For the first term,

$$\begin{aligned}
\mathbb{E}\big[(\langle \mathbf{w}_{r,t}, \mathbf{x}_{t,i}^{(p)}\rangle)^4\big] &= \mathbb{E}\big[(a_r + z_r)^4\big] \\
&= \mathbb{E}[a_r^4 + 4a_r^3 z_r + 6a_r^2 z_r^2 + 4a_r z_r^3 + z_r^4] \\
&= a_r^4 + 6a_r^2\mathbb{E}[z_r^2] + \mathbb{E}[z_r^4] \\
&= a_r^4 + 6a_r^2 b_r^2 + 3b_r^4 \\
&= \alpha_t^4\langle \mathbf{w}_{r,t}, \mathbf{x}_{t,i}^{(p)}\rangle^4 + 6\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}^{(p)}\rangle^2\|\mathbf{w}_{r,t}\|^2 + 3\beta_t^4\|\mathbf{w}_{r,t}\|^4
\end{aligned}$$

Next, for $\mathbb{E}_{\boldsymbol{\epsilon}_{t,i}\sim\mathcal{N}(0,\mathbf{I})}[\langle \mathbf{w}_{r,t}, \alpha_t\mathbf{x}_{0,i} + \beta_t\boldsymbol{\epsilon}_{t,i}\rangle^2\langle \mathbf{w}_{r',t}, \alpha_t\mathbf{x}_{0,i} + \beta_t\boldsymbol{\epsilon}_{t,i}\rangle^2]$, we note that

$$\begin{aligned}
&\mathbb{E}[z_r z_{r'}] = \beta_t^2\mathbb{E}[\boldsymbol{\epsilon}_{t,i}^\top \mathbf{w}_{r,t}\mathbf{w}_{r',t}^\top\boldsymbol{\epsilon}_{t,i}] = \beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle, \\
&\mathbb{E}[z_r z_{r'}^2] = 0 \\
&\mathbb{E}[z_r^2 z_{r'}^2] = \mathbb{E}[z_r^2]\mathbb{E}[z_{r'}^2] + 2\mathbb{E}[z_r z_{r'}]^2 = \beta_t^4\|\mathbf{w}_{r,t}\|^2\|\mathbf{w}_{r',t}\|^2 + 2\beta_t^4\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle^2
\end{aligned}$$

where the second and third results follow from Isserlis Theorem. Then we can simplify

$$\begin{aligned}
&\mathbb{E}[\langle \mathbf{w}_{r,t}, \alpha_t\mathbf{x}_{0,i} + \beta_t\boldsymbol{\epsilon}_{t,i}\rangle^2\langle \mathbf{w}_{r',t}, \alpha_t\mathbf{x}_{0,i} + \beta_t\boldsymbol{\epsilon}_{t,i}\rangle^2] \\
&= \mathbb{E}[(a_r + z_r)^2(a_{r'} + z_{r'})^2] \\
&= a_r^2 a_{r'}^2 + a_r^2\mathbb{E}[z_{r'}^2] + 4a_r a_{r'}\mathbb{E}[z_r z_{r'}] + a_{r'}^2\mathbb{E}[z_r^2] + \mathbb{E}[z_r^2 z_{r'}^2] \\
&= \alpha_t^4\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}\rangle^2\langle \mathbf{w}_{r',t}, \mathbf{x}_{0,i}\rangle^2 + \alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}\rangle^2\|\mathbf{w}_{r',t}\|^2 + 4\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r',t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle \\
&\quad + \alpha_t^2\beta_t^2\langle \mathbf{w}_{r',t}, \mathbf{x}_{0,i}\rangle^2\|\mathbf{w}_{r,t}\|^2 + \beta_t^4\|\mathbf{w}_{r,t}\|^2\|\mathbf{w}_{r',t}\|^2 + 2\beta_t^4\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle^2
\end{aligned}$$

Combining I_1, I_2, I_3 gives

$$\begin{aligned}
&\mathbb{E}\|s_t(\mathbf{x}_{t,i}^{(p)}) - \boldsymbol{\epsilon}_{t,i}\|^2 \\
&= d + \underbrace{\frac{1}{m}\sum_{r=1}^{m}\|\mathbf{w}_{r,t}\|^2\Big(\alpha_t^4\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}^{(p)}\rangle^4 + 6\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}^{(p)}\rangle^2\|\mathbf{w}_{r,t}\|^2 + 3\beta_t^4\|\mathbf{w}_{r,t}\|^4 - 4\sqrt{m}\alpha_t\beta_t\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}^{(p)}\rangle\Big)}_{L_{1,i}^{(p)}(\mathbf{w}_{r,t})} \\
&\quad \underbrace{\begin{aligned}&+ \frac{2}{m}\sum_{r=1}^{m}\sum_{r'\neq r}\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle\Big(\big(\alpha_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}^{(p)}\rangle^2 + \beta_t^2\|\mathbf{w}_{r,t}\|^2\big)\big(\alpha_t^2\langle \mathbf{w}_{r',t}, \mathbf{x}_{0,i}^{(p)}\rangle^2 + \beta_t^2\|\mathbf{w}_{r',t}\|^2\big) \\ &+ 2\beta_t^4\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle^2 + 4\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r',t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle\Big)\end{aligned}}_{L_{2,i}^{(p)}(\mathbf{w}_{r,t})}
\end{aligned}$$

where we respectively denote the two composing loss terms as $L_{1,i}^{(p)}$ (corresponding to the learning of r-th neuron) and $L_{2,i}^{(p)}$ (alignment with other neurons). □

We next compute the gradient of the DDPM loss in expectation.

Lemma E.3. *The gradient of expected DDPM loss in Lemma E.2 can be computed as*

$$\nabla L(\mathbf{W}_t) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{p\in[2]}\big(\nabla L_{1,i}^{(p)}(\mathbf{W}_t) + \nabla L_{2,i}^{(p)}(\mathbf{W}_t)\big)$$

where

$$
\begin{aligned}
&\nabla L^{(p)}_{1,i}(\mathbf{w}_{r,t}) \\
&= \frac{2}{m}\Big(\alpha_t^4\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle^4 + 12\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle^2\|\mathbf{w}_{r,t}\|^2 + 9\beta_t^4\|\mathbf{w}_{r,t}\|^4 - 4\sqrt{m}\alpha_t\beta_t\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle\Big)\mathbf{w}_{r,t} \\
&\quad + \frac{2}{m}\Big(2\alpha_t^4\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle^3\|\mathbf{w}_{r,t}\|^2 + 6\alpha_t^2\beta_t^2\|\mathbf{w}_{r,t}\|^4\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle - 2\sqrt{m}\alpha_t\beta_t\|\mathbf{w}_{r,t}\|^2\Big)\mathbf{x}^{(p)}_{0,i} \\
&\nabla L^{(p)}_{2,i}(\mathbf{w}_{r,t}) \\
&= \frac{2}{m}\sum_{r'\neq r}\Big(\big(\alpha_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle^2 + \beta_t^2\|\mathbf{w}_{r,t}\|^2\big)\big(\alpha_t^2\langle \mathbf{w}_{r',t}, \mathbf{x}^{(p)}_{0,i}\rangle^2 + \beta_t^2\|\mathbf{w}_{r',t}\|^2\big) + 2\beta_t^4\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle^2 \\
&\qquad\qquad + 4\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r',t}, \mathbf{x}_{0,i}\rangle\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle\Big)\mathbf{w}_{r',t} \\
&\quad + \frac{2}{m}\sum_{r'\neq r}\big(\alpha_t^2\langle \mathbf{w}_{r',t}, \mathbf{x}^{(p)}_{0,i}\rangle^2 + \beta_t^2\|\mathbf{w}_{r',t}\|^2\big)\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle\big(2\alpha_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}^{(p)}_{0,i}\rangle\mathbf{x}^{(p)}_{0,i} + 2\beta_t^2\mathbf{w}_{r,t}\big) \\
&\quad + \frac{2}{m}\sum_{r'\neq r}\langle \mathbf{w}_{r,t}, \mathbf{w}_{r',t}\rangle^2\big(4\beta_t^2\mathbf{w}_{r',t} + 8\alpha_t^2\beta_t^2\langle \mathbf{w}_{r,t}, \mathbf{x}_{0,i}\rangle\mathbf{x}_{0,i}\big)
\end{aligned}
$$

Proof of Lemma E.3. The proof is straightforward and thus omitted for clarity. □

E.3 First stage

Lemma E.4. *Under Condition E.1, suppose $n \cdot \mathrm{SNR}^2 = \widetilde{O}(1), n^{-1} \cdot \mathrm{SNR}^{-2} = \widetilde{O}(1)$. There exists an iteration $T_1^- = \max\{T_\mu, T_\xi\}$, where $T_\mu = \widetilde{\Theta}(\sqrt{m}\sigma_0^{-1}d^{-1}\|\boldsymbol{\mu}\|^{-1}\eta^{-1})$ and $T_\xi = \widetilde{\Theta}(n\sqrt{m}\sigma_0^{-1}\sigma_\xi^{-1}d^{-3/2}\eta^{-1})$ such that for all $0 \leq k \leq T_1^-$, (1) $|\langle \mathbf{w}^k_{r,t}, \boldsymbol{\mu}_j\rangle| = \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|)$ (2) $|\langle \mathbf{w}^k_{r,t}, \boldsymbol{\xi}_i\rangle| = \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$ and (3) $\|\mathbf{w}^k_{r,t}\|^2 = \Theta(\sigma_0^2 d)$ for all $r \in [m], j = \pm 1, i \in [n]$. and (4) the signal and noise learning dynamics can be simplified to*

$$
\begin{aligned}
\langle \mathbf{w}^{k+1}_{r,t}, \boldsymbol{\mu}_j\rangle &= \langle \mathbf{w}^{k}_{r,t}, \boldsymbol{\mu}_j\rangle + \frac{4\eta\alpha_t\beta_t|\mathcal{S}_j|}{n\sqrt{m}}\|\mathbf{w}^k_{r,t}\|^2\|\boldsymbol{\mu}_j\|^2 + \widetilde{O}(\eta\sigma_0^5 d^2\|\boldsymbol{\mu}_j\|^3), \\
\langle \mathbf{w}^{k+1}_{r,t}, \boldsymbol{\xi}_i\rangle &= \langle \mathbf{w}^{k}_{r,t}, \boldsymbol{\xi}_i\rangle + \frac{4\eta\alpha_t\beta_t}{n\sqrt{m}}\|\mathbf{w}^k_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2 + \widetilde{O}(\eta\sigma_0^5\sigma_\xi^3 d^{7/2}n^{-1})
\end{aligned}
$$

for all $j = \pm 1$, $r \in [m], i \in [n]$. Furthermore, we can show

- $\langle \mathbf{w}^{T_1^-}_{r,t}, \boldsymbol{\mu}_j\rangle = \Theta(\langle \mathbf{w}^{T_1^-}_{r',t}, \boldsymbol{\mu}_{j'}\rangle)$,
- $\langle \mathbf{w}^{T_1^-}_{r,t}, \boldsymbol{\xi}_i\rangle = \Theta(\langle \mathbf{w}^{T_1^-}_{r',t}, \boldsymbol{\xi}_{i'}\rangle)$,
- $\langle \mathbf{w}^{T_1^-}_{r,t}, \mathbf{w}^{T_1^-}_{r',t}\rangle = \Theta(\|\mathbf{w}^{T_1^-}_{r,t}\|^2)$
- $|\langle \mathbf{w}^{T_1^-}_{r,t}, \boldsymbol{\mu}_j\rangle|/|\langle \mathbf{w}^{T_1^-}_{r,t}, \boldsymbol{\xi}_i\rangle| = \Theta(n \cdot \mathrm{SNR}^2)$

for all $j, j' = \pm 1, r, r' \in [m], i, i' \in [n]$.

Proof of Lemma E.4. We prove the results by induction. To this end, we first compute the scale of the gradients projected to the space of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1}$ and $\boldsymbol{\xi}_i$, for $i \in [n]$ under the scale of (1)-(3). For notation clarity, we omit the index k.

Signal. First for $\boldsymbol{\mu}_j$, and for any $i \in [n]$, we compute

$$
\begin{aligned}
\frac{1}{2n}\sum_{i=1}^{n}\langle \nabla L^{(1)}_{1,i}(\mathbf{w}_{r,t}), \boldsymbol{\mu}_j\rangle &= \widetilde{O}(\sigma_0^4\|\boldsymbol{\mu}_j\|^4 + \sigma_0^4\|\boldsymbol{\mu}_j\|^2 d + \sigma_0^4 d^2 + \sigma_0\|\boldsymbol{\mu}_j\|)\widetilde{O}(\sigma_0\|\boldsymbol{\mu}_j\|) \\
&\quad + \widetilde{O}(\sigma_0^5\|\boldsymbol{\mu}_j\|^3 d + \sigma_0^5 d^2\|\boldsymbol{\mu}_j\| + \sigma_0^2 d)\|\boldsymbol{\mu}_j\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= \widetilde{O}(\sigma_0^2 \|\boldsymbol{\mu}_j\|^2) + \widetilde{O}(\sigma_0^2 d \|\boldsymbol{\mu}_j\|^2) \\
&= \widetilde{O}(\sigma_0^2 d \|\boldsymbol{\mu}_j\|^2)
\end{aligned}
$$

where the dominating term is $-4\sqrt{m}\alpha_t\beta_t\|\mathbf{w}_{r,t}\|^2\|\boldsymbol{\mu}_j\|^2$. It is also worth highlighting that the second dominating term is $6\alpha_t^2\beta_t^2\|\mathbf{w}_{r,t}\|^4\langle\mathbf{w}_{r,t}, \boldsymbol{\mu}_j\rangle\|\boldsymbol{\mu}_j\|^2$, which on the order of $\widetilde{O}(\sigma_0^5 d^2 \|\boldsymbol{\mu}_j\|^3)$.

Further, we have due to the orthogonality between signal and noise vectors,

$$
\begin{aligned}
\frac{1}{2n}\sum_{i=1}^n \langle \nabla L_{1,i}^{(2)}(\mathbf{w}_{r,t}), \boldsymbol{\mu}_j\rangle &= \widetilde{O}(\sigma_0^4\sigma_\xi^4 d^2 + \sigma_0^4\sigma_\xi^2 d^2 + \sigma_0^4 d^2 + \sigma_0\sigma_\xi\sqrt{d})\widetilde{O}(\sigma_0\|\boldsymbol{\mu}_j\|) \\
&= \widetilde{O}(\sigma_0^2\sigma_\xi\|\boldsymbol{\mu}_j\|\sqrt{d})
\end{aligned}
$$

where the dominating term is $-4\sqrt{m}\alpha_t\beta_t\langle\mathbf{w}_{r,t}, \boldsymbol{\xi}_i\rangle\langle\mathbf{w}_{r,t}, \boldsymbol{\mu}_j\rangle$.

In addition, we have

$$
\begin{aligned}
\frac{1}{2n}\sum_{i=1}^n \langle \nabla L_{2,i}^{(1)}(\mathbf{w}_{r,t}), \boldsymbol{\mu}_j\rangle &= \widetilde{O}\Big(\big(\sigma_0^2\|\boldsymbol{\mu}_j\|^2 + \sigma_0^2 d\big)^2 + \sigma_0^4 d + \sigma_0^4\|\boldsymbol{\mu}\|^2\sqrt{d}\Big)\widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|) \\
&\quad + \widetilde{O}\Big((\sigma_0^2\|\boldsymbol{\mu}_j\|^2 + \sigma_0^2 d)\sigma_0^2\sqrt{d}(\sigma_0\|\boldsymbol{\mu}_j\|^3 + \sigma_0\|\boldsymbol{\mu}_j\|)\Big) \\
&\quad + \widetilde{O}\Big(\sigma_0^4 d(\sigma_0\|\boldsymbol{\mu}_j\| + \sigma_0\|\boldsymbol{\mu}_j\|^3)\Big) \\
&= \widetilde{O}\big(\sigma_0^5 d^2\|\boldsymbol{\mu}_j\|\big) + \widetilde{O}(\sigma_0^5 d^{3/2}\|\boldsymbol{\mu}_j\|^3) + \widetilde{O}(\sigma_0^5 d\|\boldsymbol{\mu}_j\|^3)
\end{aligned}
$$

Further,

$$
\begin{aligned}
\frac{1}{2n}\sum_{i=1}^n \langle \nabla L_{2,i}^{(2)}(\mathbf{w}_{r,t}), \boldsymbol{\mu}_j\rangle &= \widetilde{O}\Big(\big(\sigma_0^2\sigma_\xi^2 d + \sigma_0^2 d\big)^2 + \sigma_0^4 d + \sigma_0^4\sigma_\xi^2 d^{3/2}\Big)\widetilde{O}(\sigma_0\|\boldsymbol{\mu}_j\|) \\
&\quad + \widetilde{O}\Big((\sigma_0^2\sigma_\xi^2 d + \sigma_0^2 d)\sigma_0^2\sqrt{d}(\sigma_0\|\boldsymbol{\mu}_j\|)\Big) \\
&\quad + \widetilde{O}\Big(\sigma_0^4 d(\sigma_0\|\boldsymbol{\mu}_j\|)\Big) \\
&= \widetilde{O}(\sigma_0^5 d^2\|\boldsymbol{\mu}_j\|)
\end{aligned}
$$

Then according to the definition of $|\mathcal{S}_{\pm 1}|$ and $\boldsymbol{\mu}_{\pm 1}$, we can simplify the dynamics of $\boldsymbol{\mu}_j$ learning at initialization as

$$
\langle \mathbf{w}_{r,t}^{k+1}, \boldsymbol{\mu}_j\rangle = \langle \mathbf{w}_{r,t}^{k}, \boldsymbol{\mu}_j\rangle + \frac{4\eta\alpha_t\beta_t|\mathcal{S}_j|}{n\sqrt{m}}\|\mathbf{w}_{r,t}^k\|^2\|\boldsymbol{\mu}_j\|^2 + \widetilde{O}(\sigma_0^5 d^2\|\boldsymbol{\mu}_j\|^3)
$$

where the second dominating term is $6\alpha_t^2\beta_t^2\|\mathbf{w}_{r,t}^k\|^4\langle\mathbf{w}_{r,t}^k, \boldsymbol{\mu}_j\rangle\|\boldsymbol{\mu}_j\|^2$, which on the order of $\widetilde{O}(\sigma_0^5 d^2\|\boldsymbol{\mu}_j\|^3)$.

Noise. Similarly, we can also show for the noise learning

$$
\begin{aligned}
\frac{1}{2n}\sum_{i'=1}^n \langle \nabla L_{1,i'}^{(1)}(\mathbf{w}_{r,t}), \boldsymbol{\xi}_i\rangle &= \widetilde{O}(\sigma_0^4\|\boldsymbol{\mu}_j\|^4 + \sigma_0^4\|\boldsymbol{\mu}_j\|^2 d + \sigma_0^4 d^2 + \sigma_0\|\boldsymbol{\mu}_j\|)\widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}) \\
&= \widetilde{O}(\sigma_0^2\sigma_\xi\sqrt{d}\|\boldsymbol{\mu}_j\|)
\end{aligned}
$$

where the dominating term is $-4\sqrt{m}\alpha_t\beta_t\langle\mathbf{w}_{r,t}, \boldsymbol{\mu}_j\rangle\langle\mathbf{w}_{r,t}, \boldsymbol{\xi}_i\rangle$.

$$
\begin{aligned}
\frac{1}{2n}\sum_{i'=1}^n \langle \nabla L_{1,i'}^{(2)}(\mathbf{w}_{r,t}), \boldsymbol{\xi}_i\rangle &= \widetilde{O}(\sigma_0^4\sigma_\xi^4 d^2 + \sigma_0^4\sigma_\xi^2 d^2 + \sigma_0^4 d^2 + \sigma_0\sigma_\xi\sqrt{d})\widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}) \\
&\quad + \widetilde{O}(\sigma_0^5\sigma_\xi^3 d^{5/2} + \sigma_0^5\sigma_\xi d^{5/2} + \sigma_0^2 d)O(\sigma_\xi^2 d(n^{-1} + d^{-1/2})) \\
&= \widetilde{O}(\sigma_0^2\sigma_\xi^2 d) + \widetilde{O}(\sigma_0^2\sigma_\xi^2 d^2)
\end{aligned}
$$

$$= \widetilde{O}(n^{-1}\sigma_0^2\sigma_\xi^2 d^2)$$

where the dominating term is $-4\sqrt{m}\alpha_t\beta_t\|\mathbf{w}_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2/n$. The next dominating term is $6\alpha_t^2\beta_t^2\|\mathbf{w}_{r,t}\|^4\langle \mathbf{w}_{r,t}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2/n$, which is on the order of $\sigma_0^5\sigma_\xi^3 d^{7/2}/n$.

Further, we can show

$$\begin{aligned}
\frac{1}{2n}\sum_{i'=1}^{n}\langle \nabla L_{2,i}^{(1)}(\mathbf{w}_{r,t}), \boldsymbol{\xi}_i\rangle &= \widetilde{O}\Big(\big(\sigma_0^2\|\boldsymbol{\mu}_j\|^2 + \sigma_0^2 d\big)^2 + \sigma_0^4 d + \sigma_0^4\|\boldsymbol{\mu}\|^2\sqrt{d}\Big)\widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}) \\
&\quad + \widetilde{O}\Big((\sigma_0^2\|\boldsymbol{\mu}_j\|^2 + \sigma_0^2 d)\sigma_0^2\sqrt{d}(\sigma_0\sigma_\xi\sqrt{d})\Big) \\
&\quad + \widetilde{O}\Big(\sigma_0^4 d(\sigma_0\sigma_\xi\sqrt{d})\Big) \\
&= \widetilde{O}(\sigma_0^5\sigma_\xi d^2)
\end{aligned}$$

Lastly,

$$\begin{aligned}
\frac{1}{2n}\sum_{i'=1}^{n}\langle \nabla L_{2,i}^{(2)}(\mathbf{w}_{r,t}), \boldsymbol{\xi}_i\rangle &= \widetilde{O}\Big(\big(\sigma_0^2\sigma_\xi^2 d + \sigma_0^2 d\big)^2 + \sigma_0^4 d + \sigma_0^4\sigma_\xi^2 d^{3/2}\Big)\widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}) \\
&\quad + \widetilde{O}\Big((\sigma_0^2\sigma_\xi^2 d + \sigma_0^2 d)\sigma_0^2\sqrt{d}(\sigma_0\sigma_\xi^3 d^{3/2}(n^{-1/2} + d^{-1/2}) + \sigma_0\sigma_\xi\sqrt{d})\Big) \\
&\quad + \widetilde{O}\Big(\sigma_0^4 d(\sigma_0\sigma_\xi\sqrt{d} + \sigma_0\sigma_\xi\sqrt{d}\sigma_\xi^2 d(n^{-1/2} + d^{-1/2}))\Big) \\
&= \widetilde{O}(\sigma_0^5\sigma_\xi d^{5/2}) + \widetilde{O}(\sigma_0^5\sigma_\xi^3 d^3 n^{-1/2}) + \widetilde{O}(\sigma_0^5\sigma_\xi^3 d^{5/2} n^{-1/2}).
\end{aligned}$$

This suggests the the dynamics of noise learning at initialization is

$$\langle \mathbf{w}_{r,t}^{k+1}, \boldsymbol{\xi}_i\rangle = \langle \mathbf{w}_{r,t}^{k}, \boldsymbol{\xi}_i\rangle + \frac{4\eta\alpha_t\beta_t}{n\sqrt{m}}\|\mathbf{w}_{r,t}^k\|^2\|\boldsymbol{\xi}_i\|^2 + \widetilde{O}(\sigma_0^5\sigma_\xi^3 d^{7/2} n^{-1})$$

where the second dominating term is $6\alpha_t^2\beta_t^2\|\mathbf{w}_{r,t}\|^4\langle \mathbf{w}_{r,t}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2/n$, which is on the order of $\sigma_0^5\sigma_\xi^3 d^{7/2}/n$.

Next, let $T_\mu = \Theta(\frac{\sqrt{m\log(16m/\delta)}}{\sigma_0 d\|\boldsymbol{\mu}\|\eta\alpha_t\beta_t})$ and $T_\xi = \Theta(\frac{n\sqrt{m\log(16mn/\delta)}}{\sigma_0\sigma_\xi d^{3/2}\eta\alpha_t\beta_t})$ and we prove the results (1)-(4) hold for all $0 \le k \le T_1^-$ via induction. We first show for all $0 \le k \le T_1^-$ that

$$\begin{aligned}
\langle \mathbf{w}_{r,t}^{k+1}, \boldsymbol{\mu}_j\rangle &= \langle \mathbf{w}_{r,t}^{k}, \boldsymbol{\mu}_j\rangle + \frac{4\eta\alpha_t\beta_t|\mathcal{S}_j|}{n\sqrt{m}}\|\mathbf{w}_{r,t}^k\|^2\|\boldsymbol{\mu}_j\|^2 + \widetilde{O}(\eta\sigma_0^5 d^2\|\boldsymbol{\mu}_j\|^3) \\
\langle \mathbf{w}_{r,t}^{k+1}, \boldsymbol{\xi}_i\rangle &= \langle \mathbf{w}_{r,t}^{k}, \boldsymbol{\xi}_i\rangle + \frac{4\eta\alpha_t\beta_t}{n\sqrt{m}}\|\mathbf{w}_{r,t}^k\|^2\|\boldsymbol{\xi}_i\|^2 + \widetilde{O}(\eta\sigma_0^5\sigma_\xi^3 d^{5/2} n^{-1})
\end{aligned}$$

First it is clear that at $k = 0$, we have from Lemma E.1 that $\|\mathbf{w}_{r,t}^0\|^2 = \Theta(\sigma_0^2 d)$ and

$$\begin{aligned}
|\langle \mathbf{w}_{r,t}^0, \boldsymbol{\mu}_j\rangle| &\le \sqrt{2\log(16m/\delta)}\sigma_0\|\boldsymbol{\mu}\| = \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|) \\
|\langle \mathbf{w}_{r,t}^0, \boldsymbol{\xi}_i\rangle| &\le 2\sqrt{\log(16mn/\delta)}\sigma_0\sigma_\xi\sqrt{d} = \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})
\end{aligned}$$

Suppose there exists an iteration $\widetilde{T}_\mu \le \min\{T_\mu, T_\xi\}$ such that $\|\mathbf{w}_{r,t}^0\|^2 = \Theta(\sigma_0^2 d)$ and $|\langle \mathbf{w}_{r,t}^0, \boldsymbol{\mu}_j\rangle| = \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|)$ for all $0 \le k \le \widetilde{T}_\mu - 1$. Then we have from the previous analysis, we can approximate with a linear dynamics by omitting the higher order terms $\widetilde{O}(\sigma_0^5 d^2\|\boldsymbol{\mu}\|^3)$ as

$$\begin{aligned}
\langle \mathbf{w}_{r,t}^{\widetilde{T}}, \boldsymbol{\mu}_j\rangle = \langle \mathbf{w}_{r,t}^{\widetilde{T}-1}, \boldsymbol{\mu}_j\rangle + \frac{\eta\alpha_t\beta_t}{\sqrt{m}}\Theta(\sigma_0^2 d)\|\boldsymbol{\mu}\|^2 &= \langle \mathbf{w}_{r,t}^0, \boldsymbol{\mu}_j\rangle + \frac{\eta\alpha_t\beta_t}{\sqrt{m}}\Theta(\sigma_0^2 d)\|\boldsymbol{\mu}\|^2\widetilde{T}_\mu \\
&\le \langle \mathbf{w}_{r,t}^0, \boldsymbol{\mu}_j\rangle + \frac{\eta\alpha_t\beta_t}{\sqrt{m}}\Theta(\sigma_0^2 d)\|\boldsymbol{\mu}\|^2 T_\mu \\
&= \langle \mathbf{w}_{r,t}^0, \boldsymbol{\mu}_j\rangle + \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|) \\
&= \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|)
\end{aligned}$$

where we use the Lemma C.1 that $|\mathcal{S}_j| = \Theta(n/2)$.

For the same argument, suppose there exists an iteration $\widetilde{T}_\xi \leq \min\{T_\mu, T_\xi\}$ such that $\|\mathbf{w}_{r,t}^0\|^2 = \Theta(\sigma_0^2 d)$ and $|\langle \mathbf{w}_{r,t}^0, \boldsymbol{\xi}_i \rangle| = \widetilde{O}(\sigma_0 \sigma_\xi \sqrt{d})$ hold for all $0 \leq k \leq \widetilde{T}_\xi - 1$. Then we we can approximate with a linear dynamics by omitting the higher order terms $\widetilde{O}(\sigma_0^5 \sigma_\xi^3 d^{7/2})$

$$\begin{aligned}
\langle \mathbf{w}_{r,t}^{\widetilde{T}}, \boldsymbol{\xi}_i \rangle = \langle \mathbf{w}_{r,t}^{\widetilde{T}-1}, \boldsymbol{\xi}_i \rangle + \frac{\eta \alpha_t \beta_t}{n\sqrt{m}} \Theta(\sigma_0^2 \sigma_\xi^2 d^2) &\leq \langle \mathbf{w}_{r,t}^0, \boldsymbol{\xi}_i \rangle + \frac{\eta \alpha_t \beta_t}{n\sqrt{m}} \Theta(\sigma_0^2 \sigma_\xi^2 d^2) T_\xi \\
&= \widetilde{O}(\sigma_0 \sigma_\xi \sqrt{d})
\end{aligned}$$

where we use Lemma E.1 that $\|\boldsymbol{\xi}_i\|^2 = \Theta(\sigma_\xi^2 d)$ for all $i \in [n]$. Next, denote $\mathbf{P}_{\boldsymbol{\xi}} = \frac{\boldsymbol{\xi}\boldsymbol{\xi}^\top}{\|\boldsymbol{\xi}\|^2}$ be the projection matrix onto the direction of $\boldsymbol{\xi}$ and we express $\mathbf{w}_{r,t}^{\widetilde{T}} = \mathbf{P}_{\boldsymbol{\mu}_1} \mathbf{w}_{r,t}^{\widetilde{T}} + \mathbf{P}_{\boldsymbol{\mu}_{-1}} \mathbf{w}_{r,t}^{\widetilde{T}} + \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i} \mathbf{w}_{r,t}^{\widetilde{T}} + \big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mu}_1} - \mathbf{P}_{\boldsymbol{\mu}_{-1}} - \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}\big) \mathbf{w}_{r,t}^{\widetilde{T}}$ and due to the orthogonality of the decomposition, we have

$$\begin{aligned}
\|\mathbf{w}_{r,t}^{\widetilde{T}}\|^2 &= \frac{\langle \mathbf{w}_{r,t}^{\widetilde{T}}, \boldsymbol{\mu}_1 \rangle^2}{\|\boldsymbol{\mu}\|^2} + \frac{\langle \mathbf{w}_{r,t}^{\widetilde{T}}, \boldsymbol{\mu}_{-1} \rangle^2}{\|\boldsymbol{\mu}\|^2} + \left\| \sum_{i=1}^n \frac{\langle \mathbf{w}_{r,t}^{\widetilde{T}}, \boldsymbol{\xi}_i \rangle}{\|\boldsymbol{\xi}\|^2} \right\|^2 + \\
&\quad + \left\| \Big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mu}_1} - \mathbf{P}_{\boldsymbol{\mu}_{-1}} - \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}\Big) \mathbf{w}_{r,t}^{\widetilde{T}} \right\|^2 \\
&= \widetilde{O}(\sigma_0^2) + \widetilde{O}(n\sigma_0^2) + \Theta(\sigma_0^2 d) \\
&= \Theta(\sigma_0^2 d)
\end{aligned}$$

where we use the induction results that $|\langle \mathbf{w}_{r,t}^{\widetilde{T}}, \boldsymbol{\mu}_j \rangle| = \widetilde{O}(\sigma_0 \|\boldsymbol{\mu}\|)$ and $|\langle \mathbf{w}_{r,t}^{\widetilde{T}}, \boldsymbol{\xi}_i \rangle| = \widetilde{O}(\sigma_0 \sigma_\xi \sqrt{d})$, and the $\big\| \big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mu}_1} - \mathbf{P}_{\boldsymbol{\mu}_{-1}} - \sum_{i=1}^n \mathbf{P}_{\boldsymbol{\xi}_i}\big) \mathbf{w}_{r,t}^{\widetilde{T}} \big\|^2$ is dominated by $\|\mathbf{w}_{r,t}^0\|^2 = \Theta(\sigma_0^2 d)$.

This completes the induction that for all $k \leq \min\{T_\mu, T_\xi\}$ the results (1) are satisfied. Next, we examine the iteration $\min\{T_\mu, T_\xi\} \leq k \leq \max\{T_\mu, T_\xi\} = T_1^-$. The magnitude comparison between T_μ and T_ξ depends on the condition on $n \cdot \mathrm{SNR}^2$. Because $n \cdot \mathrm{SNR}^2, n^{-1} \cdot \mathrm{SNR}^{-2} = \widetilde{O}(1)$, then $T_\mu / T_\xi = \widetilde{\Theta}(n^{-1/2} \sqrt{n^{-1} \mathrm{SNR}^{-2}}) = \widetilde{O}(1)$ and $T_\xi / T_\mu = \widetilde{\Theta}(n^{1/2} \sqrt{n \mathrm{SNR}^2}) = \widetilde{O}(1)$, where we use the condition on $n = \widetilde{O}(1)$. Hence using a similar induction argument for the iteration $\min\{T_\mu, T_\xi\} \leq k \leq \max\{T_\mu, T_\xi\}$ completes the proof that for all $0 \leq k \leq T_1^-$, $\|\mathbf{w}_{r,t}^k\|^2 = \Theta(\sigma_0^2 d)$, $|\langle \mathbf{w}_{r,t}^k, \boldsymbol{\mu}_j \rangle| = \widetilde{O}(\sigma_0 \|\boldsymbol{\mu}\|)$ and $|\langle \mathbf{w}_{r,t}^k, \boldsymbol{\xi}_i \rangle| = \widetilde{O}(\sigma_0 \sigma_\xi \sqrt{d})$.

Furthermore, at $k = T_1^-$, we have for all $r \in [m]$, $j = \pm 1$ and $i \in [n]$, the growth term dominates the initialization term and thus

$$\begin{aligned}
\langle \mathbf{w}_{r,t}^{T_1^-}, \boldsymbol{\mu}_j \rangle &= \Theta(\eta \alpha_t \beta_t m^{-1/2} \sigma_0^2 d \|\boldsymbol{\mu}\|^2 T_1^-) \\
\langle \mathbf{w}_{r,t}^{T_1^-}, \boldsymbol{\xi}_i \rangle &= \Theta(\eta \alpha_t \beta_t n^{-1} m^{-1/2} \sigma_0^2 d \sigma_\xi^2 d T_1^-)
\end{aligned}$$

Thus, we verify the concentration of inner products at the end of first stage as well as the ratio $\langle \mathbf{w}_{r,t}^{T_1^-}, \boldsymbol{\mu}_j \rangle / \langle \mathbf{w}_{r,t}^{T_1^-}, \boldsymbol{\xi}_i \rangle = \Theta(n \mathrm{SNR}^2)$ as well as $\langle \mathbf{w}_{r,t}^{T_1^-}, \mathbf{w}_{r',t}^{T_1^-} \rangle = \Theta(\|\mathbf{w}_{r,t}^{T_1^-}\|^2)$. □

E.4 SECOND STAGE

In the second stage, we show there exists some $\mathbf{W}_t^*$ such that $\nabla L(\mathbf{W}_t^*)$ satisfies $\langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*), \boldsymbol{\mu}_j \rangle = \langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*), \boldsymbol{\xi}_i \rangle = 0$ for all $j = \pm 1$, $r \in [m]$, $i \in [n]$. In other words $\mathbf{W}_t^*$ is a stationary point whose gradients along the signal and noise directions are zero.

Theorem E.1. *Under Condition E.1, there exists a stationary point $\mathbf{W}_t^*$, i.e., $\nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*) = 0$ that satisfies (1) $\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle = \Theta(\langle \mathbf{w}_{r',t}^*, \boldsymbol{\mu}_{j'} \rangle)$, (2) $\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle = \Theta(\langle \mathbf{w}_{r',t}^*, \boldsymbol{\xi}_{i'} \rangle)$, (3) $\langle \mathbf{w}_{r,t}^*, \mathbf{w}_{r',t}^* \rangle = \Theta(\|\mathbf{w}_{r,t}^*\|^2)$ and*

$$|\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle| / |\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle| = \Theta(n \cdot \mathrm{SNR}^2)$$

for all $j = \pm 1, r \in [m]$, $i \in [m]$.

Proof of Theorem E.1. The proof starts by assuming the concentration of neurons. This allows to simplify the expression for $\langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*), \boldsymbol{\mu}_j \rangle$, $\langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*), \boldsymbol{\xi}_i \rangle$ as follows.

Signal. Recall $\|\boldsymbol{\mu}_1\| = \|\boldsymbol{\mu}_{-1}\| = \|\boldsymbol{\mu}\|$. For $j = \pm 1$,

$$\langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*), \boldsymbol{\mu}_j \rangle = \frac{1}{2n} \sum_{i=1}^n \langle \nabla L_{1,i}^{(1)}(\mathbf{w}_{r,t}^*) + \nabla L_{1,i}^{(2)}(\mathbf{w}_{r,t}^*) + \nabla L_{2,i}^{(1)}(\mathbf{w}_{r,t}^*) + \nabla L_{2,i}^{(2)}(\mathbf{w}_{r,t}^*), \boldsymbol{\mu}_j \rangle$$

Then we can simplify

$$\begin{aligned}
&\frac{1}{2n} \sum_{i=1}^n \langle \nabla L_{1,i}^{(1)}(\mathbf{w}_{r,t}^*), \boldsymbol{\mu}_j \rangle \\
&= \frac{1}{m} \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^5 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^3 \|\mathbf{w}_{r,t}^*\|^2 + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle - \sqrt{m} \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^2 \\
&\quad + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^3 \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\mu}_j\|^2 + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \|\boldsymbol{\mu}_j\|^2 - \sqrt{m} \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\mu}_j\|^2 \Big)
\end{aligned}$$

where we have used Lemma C.1. And

$$\begin{aligned}
\frac{1}{2n} \sum_{i=1}^n \langle \nabla L_{1,i}^{(2)}(\mathbf{w}_{r,t}^*), \boldsymbol{\mu}_j \rangle &= \frac{1}{m} \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle^2 \|\mathbf{w}_{r,t}^*\|^2 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \\
&\quad + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle - \sqrt{m} \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \Big)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{2n} \sum_{i=1}^n \langle \nabla L_{2,i}^{(1)}(\mathbf{w}_{r,t}^*), \boldsymbol{\mu}_j \rangle &= \frac{m-1}{m} \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^5 + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^3 \|\mathbf{w}_{r,t}^*\|^2 \\
&\quad + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^3 \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\mu}_j\|^2 + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \|\boldsymbol{\mu}_j\|^2 \Big)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{2n} \sum_{i=1}^n \langle \nabla L_{2,i}^{(2)}(\mathbf{w}_{r,t}^*), \boldsymbol{\mu}_j \rangle &= \frac{m-1}{m} \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \\
&\quad + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle^2 \|\mathbf{w}_{r,t}^*\|^2 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \Big).
\end{aligned}$$

Setting $\langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*), \boldsymbol{\mu}_j \rangle = 0$ yields

$$\begin{aligned}
&\sqrt{m} \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^5 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^3 \|\mathbf{w}_{r,t}^*\|^2 + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^3 \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\mu}_j\|^2 \\
&\quad + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \|\boldsymbol{\mu}_j\|^2 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle^2 \|\mathbf{w}_{r,t}^*\|^2 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \Big) \\
&= \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^2 + \|\mathbf{w}_{r,t}^*\|^2 \|\boldsymbol{\mu}_j\|^2 + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \Big)
\end{aligned}$$

Noise. Similarly, using the same argument, we can show for noise direction,

$$\langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}_t^*), \boldsymbol{\xi}_i \rangle = \frac{1}{2n} \sum_{i'=1}^n \langle \nabla L_{1,i'}^{(1)}(\mathbf{w}_{r,t}^*) + \nabla L_{1,i'}^{(2)}(\mathbf{w}_{r,t}^*) + \nabla L_{2,i'}^{(1)}(\mathbf{w}_{r,t}^*) + \nabla L_{2,i'}^{(2)}(\mathbf{w}_{r,t}^*), \boldsymbol{\xi}_i \rangle.$$

Then we can simplify

$$\begin{aligned}
\frac{1}{2n} \sum_{i'=1}^n \langle \nabla L_{1,i'}^{(1)}(\mathbf{w}_{r,t}^*), \boldsymbol{\xi}_i \rangle &= \frac{1}{m} \Theta\Big(\langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle + \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle^2 \|\mathbf{w}_{r,t}^*\|^2 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle \\
&\quad + \|\mathbf{w}_{r,t}^*\|^4 \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle - \sqrt{m} \langle \mathbf{w}_{r,t}^*, \boldsymbol{\mu}_j \rangle \langle \mathbf{w}_{r,t}^*, \boldsymbol{\xi}_i \rangle \Big)
\end{aligned}$$

$$\frac{1}{2n} \sum_{i'=1}^n \langle \nabla L_{1,i'}^{(2)}(\mathbf{w}_{r,t}^*), \boldsymbol{\xi}_i \rangle$$

$$
\begin{aligned}
&= \frac{1}{m}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2 + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i - \sqrt{m}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\Big) \\
&\quad + \frac{1}{nm}\sum_{i'=1}^{n}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2 + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle - \sqrt{m}\|\mathbf{w}^*_{r,t}\|^2\Big)\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle \\
&= \frac{1}{m}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2 + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle - \sqrt{m}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\Big) \\
&\quad + \frac{1}{nm}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2 + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2 - \sqrt{m}\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2\Big) + \widetilde{O}(nd^{-1/2})
\end{aligned}
$$

where the second equality is due to $\sum_{i'=1}^{n}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle = (1 + \widetilde{O}(nd^{-1/2}))\|\boldsymbol{\xi}_i\|^2$ by Lemma C.2. Furthermore, we have

$$
\begin{aligned}
\frac{1}{2n}\sum_{i'=1}^{n}\langle \nabla L^{(1)}_{2,i'}(\mathbf{w}^*_{r,t}), \boldsymbol{\xi}_i\rangle = \frac{m-1}{m}\Theta\Big(&\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle \\
&+ \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^2\|\mathbf{w}^*_{r,t}\|^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle\Big)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{2n}\sum_{i'=1}^{n}\langle \nabla L^{(2)}_{2,i'}(\mathbf{w}^*_{r,t}), \boldsymbol{\xi}_i\rangle &= \frac{m-1}{m}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5 + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2\Big) \\
&\quad + \frac{m-1}{nm}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2 + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2\Big) + \widetilde{O}(nd^{-1/2}).
\end{aligned}
$$

Setting $\langle \nabla_{\mathbf{w}_{r,t}} L(\mathbf{W}^*_t), \boldsymbol{\xi}_i\rangle = 0$ yields

$$
\begin{aligned}
&\sqrt{m}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^2\|\mathbf{w}^*_{r,t}\|^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle \\
&\quad + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \frac{1}{n}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2 + \frac{1}{n}\|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2\Big) + \widetilde{O}(n\sqrt{m}d^{-1/2}) \\
&= \Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2 + \frac{1}{n}\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2\Big)
\end{aligned}
$$

Combining the above results for both signal and noise, we require to solve the following equations to compute the stationary point.

$$
\begin{aligned}
&\sqrt{m}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^5 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^3\|\mathbf{w}^*_{r,t}\|^2 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle \\
&\quad + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\|\mathbf{w}^*_{r,t}\|^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^3\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\mu}_j\|^2 + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle\|\boldsymbol{\mu}_j\|^2\Big) \\
&= \Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^2 + \|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\mu}_j\|^2\Big) \qquad (21) \\
&\sqrt{m}\Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2 + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle \\
&\quad + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^2\|\mathbf{w}^*_{r,t}\|^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \frac{1}{n}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^3\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2 + \frac{1}{n}\|\mathbf{w}^*_{r,t}\|^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|^2\Big) \\
&= \Theta\Big(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2 + \frac{1}{n}\|\mathbf{w}^*_{r,t}\|^2\|\boldsymbol{\xi}_i\|^2\Big) + \widetilde{O}(n\sqrt{m}d^{-1/2}) \qquad (22)
\end{aligned}
$$

Because we require $d = \widetilde{\Omega}(n^2 m)$, we can first ignore the term $\widetilde{O}(n\sqrt{m}d^{-1/2})$ when computing the stationary point.

In order to solve such equations, we let $\tau := \frac{\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle}{\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle}$ for any i, j and further consider the decomposition of $\mathbf{w}^*_{r,t} = \gamma_1\boldsymbol{\mu}_1\|\boldsymbol{\mu}_1\|^{-2} + \gamma_{-1}\boldsymbol{\mu}_{-1}\|\boldsymbol{\mu}_{-1}\|^{-2} + \sum_{i=1}^{n}\rho_{r,i}\boldsymbol{\xi}_i\|\boldsymbol{\xi}_i\|^{-2}$ based on the gradient descent updates of $\mathbf{w}_{r,t}$ starting from small initialization. Then we can see $\gamma_j = \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle$ for $j = \pm 1$ and $\rho_{r,i} = \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle + \widetilde{O}(d^{-1/2})$, where we use Lemma C.2 that $n|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle|\|\boldsymbol{\xi}_i\|^{-2} = \widetilde{O}(nd^{-1/2})$ for any $i \neq i'$. Then we can show

$$
\|\mathbf{w}^*_{r,t}\|^2 = 2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j\rangle^2\|\boldsymbol{\mu}_j\|^{-2} + \|\sum_{i=1}^{n}\rho_{r,i}\boldsymbol{\xi}_i\|\boldsymbol{\xi}_i\|^{-2}\|^2
$$

$$= 2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle^2 \|\boldsymbol{\mu}_j\|^{-2} + n \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2 \|\boldsymbol{\xi}_i\|^{-2} + \widetilde{O}(nd^{-1/2})$$
$$= 2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle^2 \|\boldsymbol{\mu}_j\|^{-2} + n \cdot \mathrm{SNR}^2 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2 \|\boldsymbol{\mu}_j\|^{-2} + \widetilde{O}(nd^{-1/2}) \tag{23}$$

where the first equality is by orthogonality of $\boldsymbol{\mu}_j$ to $\boldsymbol{\xi}_i$ and the second equality is by $\rho_{r,i} = \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle + \widetilde{O}(d^{-1/2})$ and $n|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \|\boldsymbol{\xi}_i\|^{-2} = \widetilde{O}(nd^{-1/2})$.

Next, we separately consider three SNR conditions, namely (1) $n \cdot \mathrm{SNR}^2 = \Theta(1)$; (2) $n \cdot \mathrm{SNR}^2 \geq \widetilde{\Omega}(1)$; and (3) $n^{-1} \cdot \mathrm{SNR}^{-2} \geq \widetilde{\Omega}(1)$.

1. When $n \cdot \mathrm{SNR}^2 = \Theta(1)$: we first can derive
$$\|\mathbf{w}^*_{r,t}\|^2 = \max\{\Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu} \rangle^2), \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2)\} \|\boldsymbol{\mu}\|^{-2}$$
where we ignore the $\widetilde{O}(nd^{-1/2})$ Next, we can simplify (21) and (22) depending on the scale of τ.

    - When $\tau = \widetilde{\Omega}(1)$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle^2) \|\boldsymbol{\mu}_j\|^{-2}$ and the equations reduce to
    $$\begin{cases} \Theta(\sqrt{m}\tau^5 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\tau^2 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \\ \Theta(\sqrt{m}\tau^4 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\tau^2 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \end{cases}$$
    It is clear to see for $\tau = \widetilde{\Omega}(1)$, the equations cannot be jointly satisfied.
    - When $\tau^{-1} = \widetilde{\Omega}(1)$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \|\boldsymbol{\mu}_j\|^{-2}$ and the equations reduce to
    $$\begin{cases} \Theta(\sqrt{m}\tau \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \\ \Theta(\sqrt{m} \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \end{cases}$$
    which cannot be satisfied simultaneously for $\tau^{-1} = \widetilde{\Omega}(1)$.
    - When $\tau = \Theta(1)$, $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle^2) \|\boldsymbol{\mu}_j\|^{-2} = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \|\boldsymbol{\mu}_j\|^{-2}$ and thus we can simplify the equations to
    $$\begin{cases} \Theta(\sqrt{m} \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \\ \Theta(\sqrt{m} \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \end{cases}$$
    which has a solution with $\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle = \Theta(m^{-6}) = \langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle$.

    Finally, we notice the term ignored has an order of $\widetilde{O}(n\sqrt{m}d^{-1/2})$.

2. When $n \cdot \mathrm{SNR}^2 = \widetilde{\Omega}(1)$: we first derive
$$\|\mathbf{w}^*_{r,t}\|^2 = \max\{\Theta(\tau^2), \Theta(n\mathrm{SNR}^2)\} \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2 \|\boldsymbol{\mu}\|^{-2}.$$
We consider the scale of τ as follows.

    - When $\tau \geq n\mathrm{SNR}^2$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(\tau^2) \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2 \|\boldsymbol{\mu}_j\|^{-2}$ we can simplify the equations to
    $$\begin{cases} \Theta(\sqrt{m}\tau^5 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\tau^2 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \\ \Theta(\sqrt{m}\tau^4 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\tau^2 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2 n^{-1} \mathrm{SNR}^{-2}) \end{cases}$$
    In order to satisfy both equations, we require $\tau = \Theta(n\mathrm{SNR}^2)$. On the other hand, if $\sqrt{n\mathrm{SNR}^2} \leq \tau \leq n\mathrm{SNR}^2$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(n\mathrm{SNR}^2) \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2 \|\boldsymbol{\mu}\|^{-2}$ and we can simplify the equations to
    $$\begin{cases} \Theta(\sqrt{m}\tau^5 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\tau^2 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \\ \Theta(\sqrt{m}\tau^4 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\tau \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \end{cases}$$
    which can be satisfied for any $\sqrt{n\mathrm{SNR}^2} \leq \tau \leq n\mathrm{SNR}^2$.

    Finally, if $\Omega(1) \leq \tau \leq \sqrt{n\mathrm{SNR}^2}$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(n\mathrm{SNR}^2) \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2 \|\boldsymbol{\mu}\|^{-2}$ and we can simplify (21) as
    $$\begin{cases} \Theta(\sqrt{m}(\tau^3 n\mathrm{SNR}^2 + \tau n^2 \mathrm{SNR}^4 \|\boldsymbol{\mu}\|^{-2}) \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(n\mathrm{SNR}^2 \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \\ \Theta(\sqrt{m}(\tau^4 + n^2 \mathrm{SNR}^4 \|\boldsymbol{\mu}\|^{-4} + \tau^2 n\mathrm{SNR}^2 \|\boldsymbol{\mu}\|^{-2}) \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^5) = \Theta(\tau \langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle^2) \end{cases} \tag{24}$$

    To analyze the solution to the above equations, we consider two cases:

  - When $\|\boldsymbol{\mu}\|^2 \leq n\mathrm{SNR}^2\tau^{-2}$, the first equation of (24) can be simplified as $\Theta(\sqrt{m}\tau n^2\mathrm{SNR}^4\|\boldsymbol{\mu}\|^{-2}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(n\mathrm{SNR}^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)$ and the second equation of (24) can be reduced to $\Theta(\sqrt{m}n^2\mathrm{SNR}^4\|\boldsymbol{\mu}\|^{-4}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\tau\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)$ and in order for both equations hold, we must have $\tau = \Theta(\sqrt{n\mathrm{SNR}^2\|\boldsymbol{\mu}\|^{-2}})$.
  - When $\|\boldsymbol{\mu}\|^2 \geq n\mathrm{SNR}^2\tau^{-2}$, the first equation of (24) can be simplified as $\Theta(\sqrt{m}\tau^3 n\mathrm{SNR}^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(n\mathrm{SNR}^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)$ and the second equation of (24) can be reduced to $\Theta(\sqrt{m}\tau^4\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\tau\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)$, which holds for any τ.

  Hence in summary we have $\tau \geq \sqrt{n\mathrm{SNR}^2\|\boldsymbol{\mu}\|^{-2}} \geq \Omega(1)$ can satisfy both equations.

- When $\tau^{-1} = \widetilde{\Omega}(1)$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(n\mathrm{SNR}^2)\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\|\boldsymbol{\mu}\|^{-2}$ and thus

$$\begin{cases}\Theta(\sqrt{m}(\tau + \tau^3 n\mathrm{SNR}^2 + \tau n^2\mathrm{SNR}^4\|\boldsymbol{\mu}\|^{-2})\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(n\mathrm{SNR}^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\\ \Theta(\sqrt{m}(1 + n\mathrm{SNR}^2\|\boldsymbol{\mu}\|^{-2} + n^2\mathrm{SNR}^4\|\boldsymbol{\mu}\|^{-4})\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\end{cases}$$

  which cannot be satisfied for $\tau^{-1} = \widetilde{\Omega}(1)$ because the coefficient of the first equality is order smaller than the coefficient of the second equality, i.e., $\tau n^{-1}\mathrm{SNR}^{-2} + \tau^3 + \tau n\mathrm{SNR}^2\|\boldsymbol{\mu}\|^{-2} \ll 1 + n\mathrm{SNR}^2\|\boldsymbol{\mu}\|^{-2} + n^2\mathrm{SNR}^4\|\boldsymbol{\mu}\|^{-4}$.

3. When $n^{-1}\cdot\mathrm{SNR}^{-2} = \widetilde{\Omega}(1)$: we first derive

$$\|\mathbf{w}^*_{r,t}\|^2 = \max\{\Theta(\tau^2), \Theta(n\mathrm{SNR}^2)\}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\|\boldsymbol{\mu}_j\|^{-2}.$$

   We consider the scale of τ as follows.

   - When $\tau = \widetilde{\Omega}(1)$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(\tau^2)\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\|\boldsymbol{\mu}_j\|^{-2}$ and thus we can simplify the equations to

$$\begin{cases}\Theta(\sqrt{m}\tau^5\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\tau^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\\ \Theta(\sqrt{m}(\tau^4 + n^{-1}\mathrm{SNR}^{-2}\tau^2 + \tau^4 n^{-1}\mathrm{SNR}^{-2}\|\boldsymbol{\mu}\|^{-2})\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(n^{-1}\mathrm{SNR}^{-2}\tau^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\end{cases}$$

     which cannot be satisfied when $\tau = \widetilde{\Omega}(1)$ because the first equation has a coefficient τ^3 while the second equation has a coefficient of $\tau^2 n\mathrm{SNR}^2 + 1 + \tau^2$, which is much smaller given $n^{-1}\mathrm{SNR}^{-2} = \widetilde{\Omega}(1)$.
   - When $\tau^{-1} \geq n^{-1}\mathrm{SNR}^{-2}$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(n\mathrm{SNR}^2)\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\|\boldsymbol{\mu}_j\|^{-2}$ and we can simplify

$$\begin{cases}\Theta(\sqrt{m}\tau\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(n\mathrm{SNR}^2\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\\ \Theta(\sqrt{m}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\end{cases}$$

     which can only be satisfied if $\tau = \Theta(n\mathrm{SNR}^2)$.
   - When $\sqrt{n^{-1}\mathrm{SNR}^{-2}} \leq \tau^{-1} \leq n^{-1}\mathrm{SNR}^{-2}$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(n\mathrm{SNR}^2)\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\|\boldsymbol{\mu}_j\|^{-2}$ and we can simplify

$$\begin{cases}\Theta(\sqrt{m}\tau\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\tau\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\\ \Theta(\sqrt{m}\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\end{cases}$$

     which can be satisfied for any $\sqrt{n^{-1}\mathrm{SNR}^{-2}} \leq \tau^{-1} \leq n^{-1}\mathrm{SNR}^{-2}$.
   - When $\Omega(1) \leq \tau^{-1} \leq \sqrt{n^{-1}\mathrm{SNR}^{-2}}$, we have $\|\mathbf{w}^*_{r,t}\|^2 = \Theta(\tau^2)\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2\|\boldsymbol{\mu}_j\|^{-2}$ and we can simplify

$$\begin{cases}\Theta(\sqrt{m}\tau\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta(\tau\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\\ \Theta(\sqrt{m}(1 + \tau^2 n^{-1}\mathrm{SNR}^{-2})\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^5) = \Theta((1 + \tau^2 n^{-1}\mathrm{SNR}^{-2})\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i\rangle^2)\end{cases}$$

     which can be satisfied for any $\Omega(1) \leq \tau^{-1} \leq \sqrt{n^{-1}\mathrm{SNR}^{-2}}$.

In summary, the above scale analysis reveals that

1. When $n \cdot \mathrm{SNR}^2 = \Theta(1)$, $|\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle| / |\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle| = \Theta(1)$ and $\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle = \Theta(m^{-6})$.
2. When $n \cdot \mathrm{SNR}^2 = \widetilde{\Omega}(1)$, $\Omega(1) = |\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle| / |\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle| \le \Theta(n \cdot \mathrm{SNR}^2)$.
3. When $n^{-1} \cdot \mathrm{SNR}^{-2} = \widetilde{\Omega}(1)$, $\Omega(1) = |\langle \mathbf{w}^*_{r,t}, \boldsymbol{\xi}_i \rangle| / |\langle \mathbf{w}^*_{r,t}, \boldsymbol{\mu}_j \rangle| \le \Theta(n^{-1} \cdot \mathrm{SNR}^{-2})$.

□

F ON THE FEATURE LEARNING OF VAE

To analyze variational auto-encoder (VAE) (Kingma & Welling, 2014), we consider the following problem setup. We follow the common practice by setting the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\phi}(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}(\mathbf{x})^2))$. Here $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ is parameterized by an encoder network. Then we can show the objective of VAE is given by a reconstruction loss plus a regularization term:

$$\begin{aligned} L &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \|\mathbf{x} - \boldsymbol{\psi}(\mathbf{z})\|^2 + \frac{1}{2}\|\boldsymbol{\phi}(\mathbf{x})\|^2 - \frac{1}{2}\log\big(\boldsymbol{\sigma}(\mathbf{x})^2\big)^\top \mathbf{1} + \frac{1}{2}\boldsymbol{\sigma}(\mathbf{x})^{2\top}\mathbf{1} + \mathrm{const} \\ &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \|\mathbf{x} - \boldsymbol{\psi}(\mathbf{z})\|^2 + L_{\mathrm{reg}} \end{aligned}$$

where $\boldsymbol{\psi} : \mathbb{R}^m \to \mathbb{R}^d$ is a decoder network and we use $L_{\mathrm{reg}} = \frac{1}{2}\|\boldsymbol{\phi}(\mathbf{x})\|^2 - \frac{1}{2}\log\big(\boldsymbol{\sigma}(\mathbf{x})^2\big)^\top \mathbf{1} + \frac{1}{2}\boldsymbol{\sigma}(\mathbf{x})^{2\top}\mathbf{1} + \mathrm{const}$ to denote the KL regularization term. Consider a dataset of $\{\mathbf{x}_i\}_{i=1}^n$ following the same distribution as in Definition 2.1, and leveraging the reparameterized trick, we can rewrite the loss as

$$\begin{aligned} L &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{N}(0,\mathbf{I})} \|\mathbf{x}_i - \boldsymbol{\psi}\big(\boldsymbol{\phi}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) \odot \boldsymbol{\xi}\big)\|^2 + L_{\mathrm{reg}} \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{p=1}^2 \mathbb{E}_{\boldsymbol{\xi}} \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top\big((\mathbf{W}\mathbf{x}_i^{(p)})^2 + \boldsymbol{\sigma}(\mathbf{x}) \odot \boldsymbol{\xi}\big)\|^2 + L_{\mathrm{reg}} \end{aligned}$$

where we set the network models following our setting for diffusion models, i.e., $\boldsymbol{\phi}(\mathbf{x}) = (\mathbf{W}\mathbf{x})^2$ and $\boldsymbol{\psi}(\mathbf{z}) = \mathbf{W}^\top\mathbf{z}$ for $\mathbf{W} \in \mathbb{R}^{m \times d}$, with shared weights for encoder and decoder and quadratic activation function.

We can readily observe that the loss of VAE is similar to the DDPM loss for diffusion models as a form of denoising except that the noise is added in the latent space.

Then following a similar trjactory based analysis developed for diffusion models, we expect similar VAEs also learn balanced features. To see this, we consider the following two settings:

(1) When $\boldsymbol{\sigma}(\mathbf{x}) = \mathbf{I}$, then the loss can be simplified by taking the expectation over $\boldsymbol{\xi}$. For a single patch of a sample, we can simplify

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}} \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2 - \mathbf{W}^\top\boldsymbol{\xi}\|^2 \\ &= \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2\|^2 + \mathbb{E}_{\boldsymbol{\xi}}\|\mathbf{W}^\top\boldsymbol{\xi}\|^2 - 2\mathbb{E}_{\boldsymbol{\xi}}\langle \mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2, \mathbf{W}^\top\boldsymbol{\xi}\rangle \\ &= \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2\|^2 + \|\mathbf{W}\|^2, \end{aligned}$$

which reduces to an auto-encoder with L_2 regularization, i.e.,

$$L = \frac{1}{n}\sum_{i=1}^n \sum_{p=1}^2 \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2\|^2 + L_{\mathrm{reg}} + \|\mathbf{W}\|^2.$$

(2) When $\boldsymbol{\sigma}(\mathbf{x})$ is general, then we can simplify the loss as

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}} \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2 - \mathbf{W}^\top(\boldsymbol{\sigma}(\mathbf{x}_i^{(p)}) \odot \boldsymbol{\xi})\|^2 \\ &= \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2\|^2 + \mathbb{E}_{\boldsymbol{\xi}}\|\mathbf{W}^\top(\boldsymbol{\sigma}(\mathbf{x}_i^{(p)}) \odot \boldsymbol{\xi})\|^2 \end{aligned}$$

$$= \|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2\|^2 + \langle \mathbf{W}, \mathrm{diag}\big(\boldsymbol{\sigma}(\mathbf{x}_i^{(p)})^2\big)\mathbf{W}\rangle$$

where we notice the cross term vanishes. This leads to the loss

$$L = \frac{1}{n}\sum_{i=1}^{n}\sum_{p=1}^{2}\|\mathbf{x}_i^{(p)} - \mathbf{W}^\top(\mathbf{W}\mathbf{x}_i^{(p)})^2\|^2 + L_{\mathrm{reg}} + \langle \mathbf{W}, \mathrm{diag}\big(\boldsymbol{\sigma}(\mathbf{x}_i^{(p)})^2\big)\mathbf{W}\rangle.$$

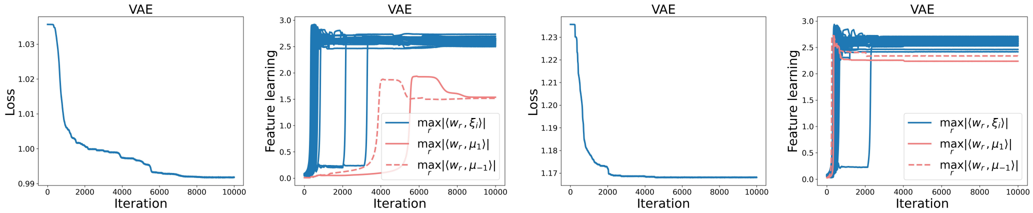
It can be seen the VAE loss indeed comprises of a reconstruction term plus some regularization terms. The reconstruction loss forces the model to learn both signal and noise. To see this, we can compute the dominant term in the gradient directions of the reconstruction loss and simplify the updates in the early-stage the same as for diffusion models:

$$\langle \mathbf{w}_r^{k+1}, \boldsymbol{\mu}_j\rangle = \langle \mathbf{w}_r^{k}, \boldsymbol{\mu}_j\rangle + \Theta(\eta\langle \mathbf{w}_r^{k}, \boldsymbol{\mu}_j\rangle^3)$$
$$\langle \mathbf{w}_r^{k+1}, \boldsymbol{\xi}_i\rangle = \langle \mathbf{w}_r^{k}, \boldsymbol{\xi}_i\rangle + \Theta(\eta\langle \mathbf{w}_r^{k}, \boldsymbol{\xi}_i\rangle^3)$$

Then we can follow the analysis as in diffusion model to characterize the feature learning dynamics of VAE.

F.1 Experiments

We proceed with experiments to investigate the feature learning dynamics of the VAE model. Using the same data generation setup as outlined in Section 5.1, we align with the diffusion model approach by averaging over 2000 sampled noise vectors, ξ, for each data point. The VAE loss is then optimized with a regularization term, λL_{reg}, setting $\lambda = 0.001$. To parameterize the variance, we define it as $\boldsymbol{\sigma}(\mathbf{x}) = \mathrm{diag}(\mathbf{W}_v\mathbf{x})$, where $\mathbf{W}_v$ is a trainable matrix distinct from $\mathbf{W}$. Feature learning is evaluated across two signal-to-noise ratio (SNR) regimes, specifically, $n \cdot \mathrm{SNR}^2 = 0.75$ and $n \cdot \mathrm{SNR}^2 = 6.75$, as described in Section 5.1. The results of these experiments are presented in Figure 13. We observe that similar to diffusion model, VAE learns features following the scale of SNR.



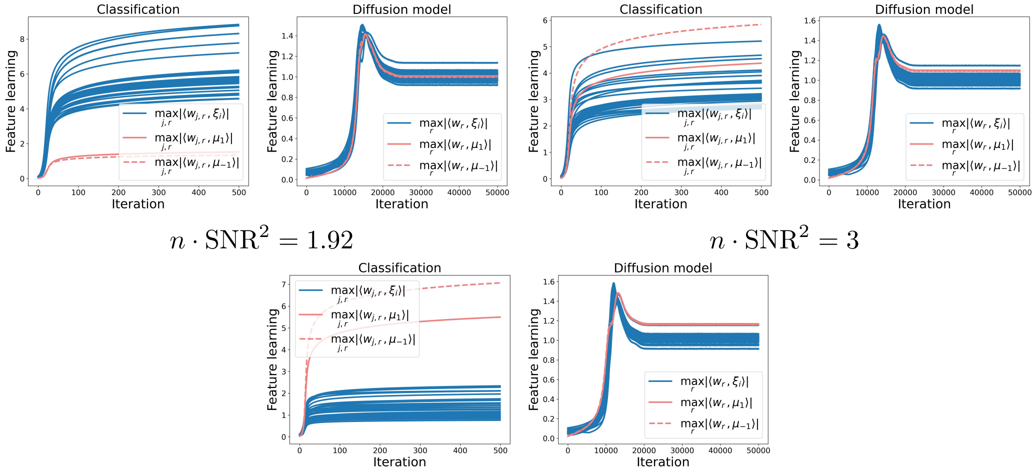
Low SNR ($n \cdot \mathrm{SNR}^2 = 0.75$) High SNR ($n \cdot \mathrm{SNR}^2 = 6.75$)

Figure 13: Loss and feature learning of VAE on the synthetic dataset with both low SNR ($n \cdot \mathrm{SNR}^2 = 0.75$) and high SNR ($n \cdot \mathrm{SNR}^2 = 6.75$). We observe the scale of feature learning matches the scale of $n \cdot \mathrm{SNR}^2$.

G Feature learning comparison under varying SNRs

In this section, we compare the feature learning dynamics of classification and diffusion models on additional settings of SNR. Apart from the $n \cdot \mathrm{SNR}^2 = 0.75$ and $n \cdot \mathrm{SNR}^2 = 6.75$ as shown in the main text, we additionally test on (1) $n \cdot \mathrm{SNR}^2 = 1.92$, (2) $n \cdot \mathrm{SNR}^2 = 3$ (3) $n \cdot \mathrm{SNR}^2 = 4.32$. The feature learning dynamics under the corresponding SNR settings are shown in Figure 14.

From the figures, we can see that classification indeed is more sensitive to the SNR scale, where it easily overfit to either signal or noise (except for the case where $n \cdot \mathrm{SNR}^2 = 3$ where classification learns signal and noise to approximately the same scale). On the other hand, we can verify that at stationarity, diffusion model learns in a more balanced scale for signal and noise.



$n \cdot \mathrm{SNR}^2 = 1.92$

$n \cdot \mathrm{SNR}^2 = 3$

$n \cdot \mathrm{SNR}^2 = 4.32$

Figure 14: Experiments on the synthetic dataset with varying SNRs.

H Feature learning comparison with three-layer neural networks

This section performs additional experiments by changing the neural network for both classification and diffusion model from two layers to three layers. In addition, we switch from quadratic activation to more practically used ReLU activation.

Specifically, for **diffusion model**, we consider

$$\boldsymbol{f}(\mathbf{W}_{t,1}, \mathbf{W}_{t,2}, \mathbf{x}_t) = \Big[\boldsymbol{f}_1(\mathbf{W}_t, \mathbf{x}_t^{(1)})^\top, \boldsymbol{f}_2(\mathbf{W}_t, \mathbf{x}_t^{(2)})^\top\Big]^\top \in \mathbb{R}^{2d}$$

$$\text{where } \boldsymbol{f}_p(\mathbf{W}_{t,1}, \mathbf{W}_{t,2}, \mathbf{x}_t^{(p)}) = \mathbf{W}_{t,1}^\top \sigma_R(\mathbf{W}_{t,2}\sigma_R(\mathbf{W}_{t,1}\mathbf{x}_t^{(1)})), \quad p = 1, 2$$

where $\sigma_R(\cdot)$ denotes the ReLU activation and $\mathbf{W}_{t,1} \in \mathbb{R}^{m\times d}$, $\mathbf{W}_{t,2} \in \mathbb{R}^{m\times m}$.

For **classification**, we consider

$$f(\mathbf{W}_0, \mathbf{W}, \mathbf{x}) = F_1(\mathbf{W}_0, \mathbf{W}_1, \mathbf{z}) - F_{-1}(\mathbf{W}_0, \mathbf{W}_{-1}, \mathbf{z})$$

$$\text{where } F_j(\mathbf{W}_0, \mathbf{W}_j, \mathbf{z}) = \frac{1}{m}\sum_{r=1}^m \sigma_R(\langle \mathbf{w}_{j,r}, \mathbf{z}^{(1)}\rangle) + \frac{1}{m}\sum_{r=1}^m \sigma_R(\langle \mathbf{w}_{j,r}, \mathbf{z}^{(2)}\rangle)$$

$$\mathbf{z}^{(p)} = \sigma_R(\mathbf{W}_0\mathbf{x}^{(p)}), \quad p = 1, 2$$

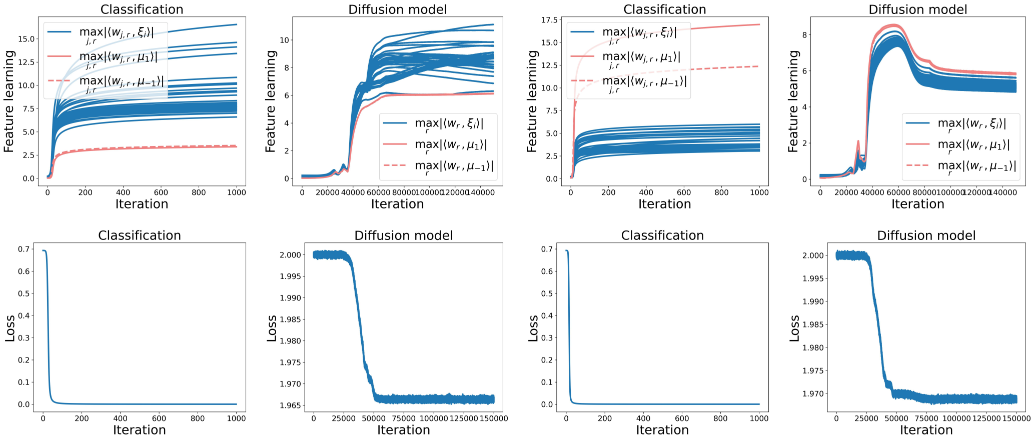
where $\mathbf{W}_0 \in \mathbb{R}^{d\times m}$ is the first layer weight and we use ReLU activation $\sigma_R(\cdot)$.

Here we measure the signal and noise learning by tracking the signal and noise inner products with the *first-layer* weight, which directly extracts features from the data.

We use the same synthetic data setups as in Section 5.1 under the two SNR cases. The results are shown in Figure 15. We observe that although we include another layer and change the activation from quadratic to ReLU, we still observe a similar pattern as for the two-layer network setup. In particular, we verify that classification similarly bias the learning towards the one feature depending on SNR and the resulting gap is significantly larger compared to diffusion model, where all features are learned to a relatively the same scale.

I On the feature learning with 10-class MNIST

In the main paper, we only conduct experiments on Noisy-MNIST restricted to two classes. In this section, we experiment over the 10-class MNIST dataset, which contains more features and is more challenging for both diffusion model and classification.



$n \cdot \mathrm{SNR}^2 = 0.75$ $n \cdot \mathrm{SNR}^2 = 6.75$

Figure 15: Comparison of feature learning dynamics with three-layer neural network. We observe a similar pattern as for the two-layer networks, i.e., classification is prone to learning one feature over another while diffusion model tends to learn much balanced features.

We adopt the same data processing pipelines as in Section 5.2 except that for each class, we select 10 images. We set the scaled SNR $\widetilde{\mathrm{SNR}} = 0.1$, consistent with the main paper. While the diffusion model remains unchanged, the classification model requires modification. Specifically, the second layer's weight matrix has dimensions $m \times 10$, with entries fixed uniformly to values in $\{-1, 1\}$. Furthermore, we employ cross-entropy loss for training the classification model.

We plot the visualization of feature learning in Figure 16. We observe that, even with additional features and labels, the similar learning patterns are observed, i.e., diffusion model learns both signals and noise in order to reconstruct the input distribution while classification model learns primarily noise for loss minimization. From Figure 17(c), we notice that diffusion model learns features to relatively the same scale while for classification, the growth of feature learning is dominated by noise learning.

J ON THE FEATURE LEARNING OF CLASSIFICATION WITH ADDED GAUSSIAN NOISE

In this section, we examine the feature learning of classification models when injecting Gaussian noise into the inputs. Let $L_S(\mathbf{W})$ be the empirical logistic loss without input noise and let $\widetilde{L}_S(\mathbf{W})$ be the empirical logistic loss with input noise, i.e.,

$$L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell(y_i f(\mathbf{W}, \mathbf{x}_i)), \quad \widetilde{L}_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell(y_i f(\mathbf{W}, \mathbf{x}_i + \boldsymbol{\epsilon}_i))$$

where we highlight that the added noise $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \mathbf{I})$ is randomly sampled every iteration. We assume that the added noise has unit variance without loss of generality.

Next we compute the gradient as

$$\nabla_{\mathbf{w}_{j,r}} \widetilde{L}_S(\mathbf{W}^k)$$
$$= \frac{1}{nm}\sum_{i=1}^n \tilde{\ell}_i'^k \langle \mathbf{w}_{j,r}^k, \mathbf{x}_i^{(1)} + \boldsymbol{\epsilon}_{i,1}^k \rangle j y_i (\mathbf{x}_i^{(1)} + \boldsymbol{\epsilon}_{i,1}^k) + \frac{1}{nm}\sum_{i=1}^n \tilde{\ell}_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{i,2}^k \rangle j y_i (\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{i,2}^k)$$

Then taking the expectation over the added noise and assuming that $\tilde{\ell}_i'^k$ is bounded (as in the first stage), we obtain

$$\mathbb{E}_{\boldsymbol{\epsilon}_{i,1}^k, \boldsymbol{\epsilon}_{i,2}^k}[\nabla_{\mathbf{w}_{j,r}} \widetilde{L}_S(\mathbf{W}^k)]$$

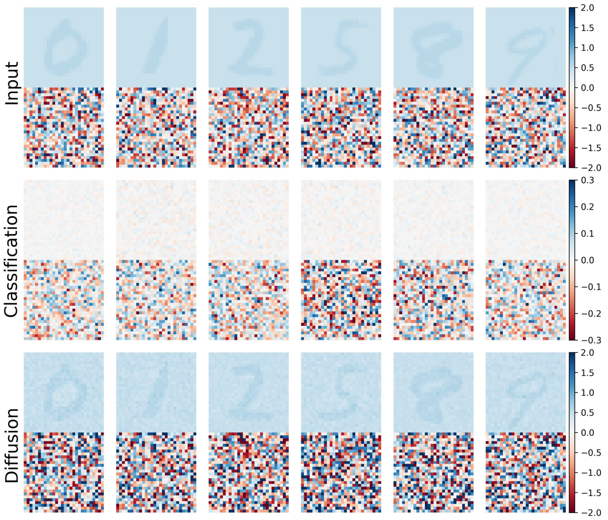


Figure 16: Experiments on 10-class Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.1$. (First row): Test Noisy-MNIST images; (Second row): Illustration of gradient of output (for the true class) with respect to the input. (Third row): denoised image from diffusion model. In this low-SNR case, we see classification tends to predominately learn noise while diffusion learns both signals and noise.

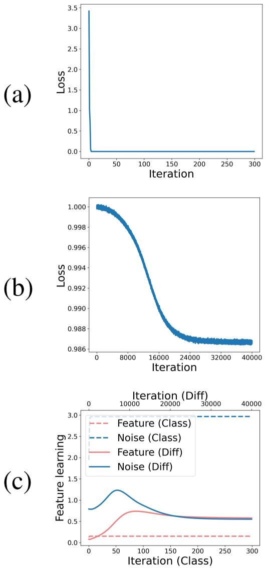


Figure 17: Experiments on 10-class Noisy-MNIST with $\widetilde{\mathrm{SNR}} = 0.1$. (a) Train loss for classification. (b) Train loss for diffusion model. (c) Feature learning dynamics.

$$= \frac{1}{nm}\sum_{i=1}^{n} \tilde{\ell}_i'^k \langle \mathbf{w}_{j,r}^k, \mathbf{x}_i^{(1)}\rangle j y_i \mathbf{x}_i^{(1)} + \frac{1}{nm}\sum_{i=1}^{n} \tilde{\ell}_i'^k \langle \mathbf{w}_{j,r}^k, \boldsymbol{\xi}_i\rangle j y_i \boldsymbol{\xi}_i + \frac{1}{nm}\sum_{i=1}^{n} \tilde{\ell}_i'^k \mathbf{w}_{j,r}^k$$

where we use the fact that $\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \mathbf{I}$.

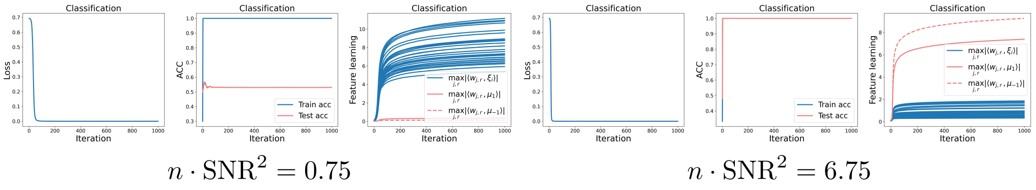
Then we can show that

$$\mathbb{E}_{\boldsymbol{\epsilon}_{i,1}^k, \boldsymbol{\epsilon}_{i,2}^k}[\nabla_{\mathbf{w}_{j,r}} \widetilde{L}_S(\mathbf{W}^k)] \approx \nabla_{\mathbf{w}_{j,r}} L_S(\mathbf{W}^k) + \frac{1}{nm}(\sum_{i=1}^{n} \ell_i'^k)\mathbf{w}_{j,r}^k$$

which can be viewed as the original objective function with an L_2 regularization.

In particular, we follow the same problem setups as in Section 5.1 for generating the data. Each iteration, we use input with randomly sampled Gaussian noise for classification models, i.e., $\mathbf{x} \to \alpha_t \mathbf{x} + \beta_t \boldsymbol{\epsilon}$, where $\alpha_t = \exp(-t) = 0.82$ and $\beta_t = \sqrt{1 - \exp(-2t)} = 0.57$, (the same as for diffusion models).

We plot the feature learning dynamics of classification in Figure 18. From the results, we see despite the presence of input Gaussian noise, classification still bias learning towards one feature over the others.



$n \cdot \mathrm{SNR}^2 = 0.75$ $\quad$ $n \cdot \mathrm{SNR}^2 = 6.75$

Figure 18: Experiments of classification on the synthetic dataset with both low SNR ($n{\cdot}\mathrm{SNR}^2 = 0.75$) and high SNR ($n \cdot \mathrm{SNR}^2 = 6.75$). The inputs to the classification include Gaussian noise. We see the classification still primarily learn one feature even with the presencen of input noise.