Provable Guarantees for Neural Networks via Gradient Feature Learning

Zhenmei Shi*, Junyi Wei*, Yingyu Liang
University of Wisconsin, Madison
zhmeishi@cs.wisc.edu, jwei53@wisc.edu, yliang@cs.wisc.edu

Abstract

Neural networks have achieved remarkable empirical performance, while the current theoretical analysis is not adequate for understanding their success, e.g., the Neural Tangent Kernel approach fails to capture their key feature learning ability, while recent analyses on feature learning are typically problem-specific. This work proposes a unified analysis framework for two-layer networks trained by gradient descent. The framework is centered around the principle of feature learning from gradients, and its effectiveness is demonstrated by applications in several prototypical problems such as mixtures of Gaussians and parity functions. The framework also sheds light on interesting network learning phenomena such as feature learning beyond kernels and the lottery ticket hypothesis.

1 Introduction

Neural network (NN) learning has achieved remarkable empirical success and has been a main driving force for the recent progress in machine learning and artificial intelligence. On the other hand, theoretical understandings significantly lag behind. Traditional analysis approaches are not adequate due to the overparameterization of practical networks and the non-convex optimization in the training via gradient descent. One line of work (e.g. [9, 31, 38, 60, 71, 123] and many others) shows under proper conditions, heavily overparameterized networks are approximately linear models over data-independent features, i.e., a linear function on the Neural Tangent Kernel (NTK). While making weak assumptions about the data and thus applicable to various settings, this approach requires the network learning to be approximately using fixed data-independent features (i.e., the kernel regime, or fixed feature methods). It thus fails to capture the feature learning ability of networks (i.e., to learn a feature mapping for the inputs which allow accurate prediction), which is widely believed to be the key factor to their empirical success in many applications (e.g., [54, 77, 117, 119]). To study feature learning in networks, a recent line of work (e.g. [5, 6, 14, 33, 52, 72, 76, 116] and others) shows examples where networks provably enjoy advantages over fixed feature methods (including NTK), under different settings and assumptions. While providing more insights, these studies typically focus on specific problems, and their analyses exploit the specific properties of the problems and appear to be unrelated to each other. Is there a common principle for feature learning in networks via gradient descent? Is there a unified analysis framework that can clarify the principle and also lead to provable error guarantees for prototypical problem settings?

In this work, we take a step toward this goal by proposing a gradient feature learning framework for analyzing two-layer network learning by gradient descent. (1) The framework makes essentially no assumption about the data distribution and can be applied to various problems. Furthermore, it is centered around features from gradients, clearly illustrating how gradient descent leads to feature learning in networks and subsequently accurate predictions. (2) It leads to error guarantees competitive with the optimal in a family of networks that use the features induced by gradients on the

*Equal contribution.

data distribution. Then for a specific problem with structured data distributions, if the optimal in the induced family is small, the framework gives a small error guarantee.

We then apply the framework to several prototypical problems: mixtures of Gaussians, parity functions, linear data, and multiple-index models. These have been used for studying network learning (in particular, for the feature learning ability), but with different and seemingly unrelated analyses. In contrast, straightforward applications of our framework give small error guarantees, where the main effort is to compute the optimal in the induced family. Furthermore, in some cases, such as parities, we can handle more general data distributions than in the existing work.

Finally, we also demonstrate that the framework sheds light on several interesting network learning phenomena or implications such as feature learning beyond the kernel regime, lottery ticket hypothesis (LTH), simplicity bias, learning over different data distributions, and new perspectives about roadmaps forward. Due to space limitations, we present implications about features beyond the kernel regime and LTH in the main body but defer the other implications in Appendix C with a brief here. (1) For simplicity bias, it is generally believed that the optimization has some implicit regularization effect that restricts learning dynamics to a low capacity subset of the whole hypothesis class, so can lead to good generalization [53, 90]. Our framework provides an explanation that the learning first learns simpler functions and then more sophisticated ones. (2) For learning over different data distributions, we provide data-dependent non-vacuous guarantees, as our framework can be viewed as using the optimal gradient-induced NN to measure or quantify the “complexity” of the problem. For easier problems, this quantity is smaller, and our framework can give a better error bound to derive guarantees. (3) For new perspectives about roadmaps forward, our framework suggests the strong representation power of NN is actually the key to successful learning, while traditional ones suggest strong representation power leads to vacuous generalization bounds [19, 33]. Thus, we suggest a different analysis road. Traditional analysis typically first reasons about the optimal based on the whole function class then analyzes how NN learns proper features and reaches the optimal. In contrast, our framework defines feature family first, and then reasons about the optimal based on it.

2 Related Work

Neural Networks Learning Analysis. Recently there has been an increasing interest in the analysis of network learning. One line of work connects the sufficiently over-parameterized neural network to linear methods around its initialization like NTK (e.g. [9, 11, 20, 21, 31, 38, 49, 60, 62, 69, 71, 78, 82, 91, 93, 95, 114, 121, 122] and more), so that the neural network training is a convex problem. The key idea is that it suffices to consider the first-order Tyler expansion of the neural network around the origin when the initialization is large enough. However, NTK lies in the lazy training (kernel) regime that excludes feature learning [29, 50, 68, 113]. Many studies (e.g. [2, 5, 6, 8, 12, 14, 22, 26, 33, 37, 51, 52, 57, 58, 70, 72, 73, 76, 99, 112, 115, 116] and more) show that neural networks take advantage over NTK empirically and theoretically. Another line of work is the mean-field (MF) analysis of neural networks (e.g. [27, 28, 36, 79, 80, 100, 106] and more). The insight is to see the training dynamics of a sufficiently large-width neural network as a PDE. It uses a smaller initialization than the NTK so that the parameters may move away from the initialization. However, the MF does not provide explicit convergence rates and requires an unrealistically large width of the neural network. One more line of work is neural networks max-margin analysis (e.g. [30, 47, 48, 56, 61, 63, 74, 75, 83, 85, 86, 107, 109] and more). They need a strong assumption that the convergence starts from weights having perfect training accuracy, while feature learning happens in the early stage of training. To explain the success of neural networks beyond the limitation mentioned above, some work introduces the low intrinsic dimension of data distributions [17, 18, 23, 24, 25, 44, 67, 104, 108, 124]. Another recent line of work is that a trained network can exactly recover the ground truth or optimal solution or teacher network [3, 4, 10, 39, 84, 87, 94, 96, 120], but they have strong assumptions on data distribution or model structure, e.g., Gaussian marginals. [1, 40, 55, 110, 111] show that training dynamics of neural networks have multiple phases, e.g., feature learning at the beginning, and then dynamics in convex optimization which requires proxy convexity [43] or PL condition [65] or special data structure.

Feature Learning Based on Gradient Analysis. A recent line of work is studying how features emerge from the gradient. [7, 46] consider linear separable data and show that the first few gradient steps can learn good features, and the later steps learn a good network on neurons with these features. [33, 45, 105] have similar conclusions on non-linear data (e.g., parity functions), while in their problems one feature is sufficient for accurate prediction (i.e., single-index data model).
The first idea of our framework is to provide guarantees compared to the optimal in a family of networks. Here let us consider networks with specific weights for the first layer: $W$ after one gradient update, i.e., no updates to $W$.

Then the training process is summarized in Algorithm 1.

**Problem Setup.** We denote $[n] := \{1, 2, \ldots, n\}$ and $\tilde{O}(\cdot), \tilde{\Theta}(\cdot), \tilde{\Omega}(\cdot)$ to omit the log term inside.

**Definition 3.3.** For some fixed $W \in \mathbb{R}^{d \times m}$, let $\mathcal{F}_{d,m} := \{f(a,W,b) \mid f(a,W,b)(x) := a^\top [\sigma(W^\top x - b)] = \sum_{i \in [m]} a_i [\sigma((w_i, x) - b_i)], \}$. (1)

where $\sigma(z) = \max(z, 0)$ is the ReLU activation function, $a \in \mathbb{R}^m$ is the second layer weight, $W \in \mathbb{R}^{d \times m}$ is the first layer weight, $w_i$ is the $i$-th column of $W$ (i.e., the weight for the $i$-th neuron), and $b \in \mathbb{R}^m$ is the bias for the neurons. For technical simplicity, we only train $a$, $W$ but not $b$. Let superscript $(t)$ denote the time step, e.g., $f(a^{(t)},W^{(t)},b)$ denote the network at time step $t$. Denote $\Xi := \{(x,y)\}$, $\Xi^{(t)} := \{(x,y)^{(t)}\}$. The goal of neural network learning is to minimize the expected risk, i.e., $\mathcal{L}_D(f) := \mathbb{E}_{(x,y) \sim D} \mathcal{L}_f(x,y)$, where $\mathcal{L}_f(x,y) = \mathcal{L}(f)$ is the loss on an example $(x,y)$ for some loss function $\mathcal{L}(\cdot)$, e.g., the hinge loss $\ell(z) = \max[0, 1 - z]$ and the logistic loss $\ell(z) = \log[1 + \exp(-z)]$. We also consider $\ell_2$ regularization. The regularized loss with regularization coefficient $\lambda$ is $\mathcal{L}_D^{\lambda}(f) := \mathcal{L}_D(f) + \frac{\lambda}{2} (\|W\|^2_2 + \|a\|^2_2)$. Given a training set with $n$ i.i.d. samples $\mathcal{Z} = \{(x^{(i)},y^{(i)})\}_{i \in [n]}$ from $D$, the empirical risk and its regularized version are:

$$\mathcal{L}_Z(f) := \frac{1}{n} \sum_{i \in [n]} \mathcal{L}_D(x^{(i)},y^{(i)}), \quad \mathcal{L}_Z^{\lambda}(f) := \mathcal{L}_Z(f) + \frac{\lambda}{2} (\|W\|^2 + \|a\|^2_2). \quad (2)$$

Then the training process is summarized in Algorithm 1.

**Algorithm 1** Network Training via Gradient Descent

```
Initialize $(a^{(0)}, W^{(0)}, b)$
for $t = 1$ to $T$
    Sample $\mathcal{Z}^{(t-1)} \sim D^n$
    $a^{(t)} = a^{(t-1)} - \eta^{(t)} \nabla_{a} \mathcal{L}_{Z^{(t-1)}}^{\lambda^{(t)}}(f_{\mathcal{Z}^{(t-1)}}), \quad W^{(t)} = W^{(t-1)} - \eta^{(t)} \nabla_{W} \mathcal{L}_{Z^{(t-1)}}^{\lambda^{(t)}}(f_{\mathcal{Z}^{(t-1)}})$
end for
```

In the whole paper, we need some natural assumptions about the data and the loss.

**Assumption 3.1.** We assume $\mathbb{E}[\|x\|_2^2] \leq B_x$, $\mathbb{E}[\|x\|_2^3] \leq B_x$, $\|x\|_2 \leq B_x$ and for any label $y$, we have $|y| \leq 1$. We assume the loss function $\mathcal{L}(\cdot)$ is a 1-Lipschitz convex decreasing function, normalized $\ell(0) = 1, \ell'(0) = 0$, and $\ell(\infty) = 0$.

**Remark 3.2.** The above are natural assumptions. Most input distributions have the bounded norms required, and the typical binary classification $Y = \{\pm 1\}$ satisfies the requirement. Also, the most popular loss functions satisfy the assumption, e.g., the hinge loss and logistic loss.

### 3.1 Warm Up: A Simple Setting with Frozen First Layer

To illustrate some high-level intuition, we first consider a simple setting where the first layer is frozen after one gradient update, i.e., no updates to $W$ for $t \geq 2$ in Algorithm 1.

The first idea of our framework is to provide guarantees compared to the optimal in a family of networks. Here let us consider networks with specific weights for the first layer:

**Definition 3.3.** For some fixed $W \in \mathbb{R}^{d \times m}$, $b \in \mathbb{R}^d$, and a parameter $B_{a_2}$, consider the following family of networks $\mathcal{F}_{W,b,B_{a_2}}$ and the optimal approximation network loss in this family:

$$\mathcal{F}_{W,b,B_{a_2}} := \{f(a,W,b) \in \mathcal{F}_{d,m} \mid \|a\|_2 \leq B_{a_2}\}, \quad \text{OPT}_{W,b,B_{a_2}} := \min_{f \in \mathcal{F}_{W,b,B_{a_2}}} \mathcal{L}_D(f). \quad (3)$$
The second idea is to compare to networks using features from gradient descent. As an illustrative example, we now provide guarantees compared to networks with first layer weights $W^{(1)}$ (i.e., the weights after the first gradient step):

\begin{theorem}[Simple Setting] \label{thm:simple}
Assume $\tilde{L}_z(f, W^{(1)}, b)$ is $L$-smooth to $a$. Let $\eta^{(t)} = \frac{1}{L}, \lambda^{(t)} = 0, \text{ for all } t \in \{2, 3, \ldots, T\}$. Training by Algorithm 1 with no updates for the first layer after the first gradient step, w.h.p., there exists $t \in [T]$ such that

$$\mathcal{L}_D(f_{(a^{(t)}, W^{(1)}, b)}) \leq \text{OPT}_{W^{(1)}, b, B, z} + O\left(\frac{L(f(a^{(t)}, W^{(1)}, b))}{L(f_{a^{(t)}, W^{(1)}, b})} + B^2 + \|b\|^2\right).$$

\end{theorem}

Intuitively, the theorem shows that if the weight $W^{(1)}$ after a one-step gradient gives a good set of neurons in the sense that there exists a classifier on top of these neurons with low loss, then the network will learn to approximate this good classifier and achieve low loss. The proof is based on standard convex optimization and the Rademacher complexity (details in Appendix D.1).

Such an approach, while simple, has been used to obtain interesting results on network learning in existing work, which shows that $W^{(1)}$ can indeed give good neurons due to the structure of the special problems considered (e.g., parities on uniform inputs [15], or polynomials on a subspace [32]). However, it is unclear whether such intuition can still yield useful guarantees for other problems. So, for our purpose of building a general framework covering more prototypical problems, the challenge is what features from gradient descent should be considered so that the family of networks for comparison can achieve a low loss on other problems. The other challenge is that we would like to consider the typical case where the first layer weights are not frozen. In the following, we will introduce the core concept of Gradient Features to address the first challenge, and stipulate proper geometric properties of Gradient Features for the second challenge.

### 3.2 Core Concepts in the Gradient Feature Learning Framework

Now, we will introduce the core concept in our framework, Gradient Features, and use it to build the family of networks to derive guarantees. As mentioned, we consider the setting where the first layer is not frozen. After the network learns good features, to ensure the updates in later gradient steps of the first layer are still benign for feature learning, we need some geometric conditions about the gradient features, which are measured by parameters in the definition of Gradient Features. The conditions are general enough, so that, as shown in Section 4, many prototypical problems satisfy them and the induced family of networks enjoys low loss, leading to useful guarantees. We begin by considering what features can be learned via gradients. Note that the gradient w.r.t. $w_i$ is

$$\frac{\partial \mathcal{L}_D(f)}{\partial w_i} = a_i \mathbb{E}_{(x, y)} \left[ \ell'(y f(x)) y [\sigma'(\langle w_i, x \rangle - b_i)] x \right] = a_i \mathbb{E}_{(x, y)} \left[ \ell'(y f(x)) y x I[\langle w_i, x \rangle > b_i] \right].$$

Inspired by this, we define the following notion:

\begin{definition}[Simplified Gradient Vector] \label{def:simplified}
For any $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, a Simplified Gradient Vector is

$$G(w, b) := \mathbb{E}_{(x, y) \sim D}[y x I[\langle w, x \rangle > b]]. \quad (4)$$

\end{definition}

\begin{remark}
Note that the definition of $G(w, b)$ ignores the term $\ell'(y f(x))$ in the gradient, where $f$ is the model function. In the early stage of training (or the first gradient step), $\ell'(\cdot)$ is approximately a constant, i.e., $\ell'(y f(x)) \approx \ell'(0)$ due to the symmetric initialization (see Equation (8)).

\end{remark}

\begin{definition}[Gradient Feature] \label{def:gradient}
For a unit vector $D \in \mathbb{R}^d$ with $\|D\|_2 = 1$, and $a \in (0, 1)$, a direction neighborhood (cone) $\mathcal{C}_{D, a}$ is defined as:

$$\mathcal{C}_{D, a} := \{w \mid \|\langle w, D \rangle\|_2 / \|w\|_2 > (1 - a)\}. \quad (5)$$

\end{definition}

Figure 1: An illustration of Gradient Feature, i.e., Definition 3.7 with random initialization (Gaussian), under Mixture of three Gaussian clusters in 3-dimension data space with blue/green/orange color. The Gradient Feature stays in three cones, where each center of the cone aligns with the corresponding Gaussian cluster center.
Let \( w \in \mathbb{R}^d, b \in \mathbb{R} \) be random variables drawn from some distribution \( \mathcal{W}, \mathcal{B} \). A Gradient Feature set with parameters \( p, \gamma, B_G \) is defined as:
\[
S_{p, \gamma, B_G}(\mathcal{W}, \mathcal{B}) := \{(D, s) \mid \Pr_{w, b}[G(w, b) \in \mathcal{C}_{D, \gamma}, \|G(w, b)\|_2 \geq B_G, s = b/\|b\| \geq p]\}.
\]

**Remark 3.8.** When clear from context, write it as \( S_{p, \gamma, B_G} \). Gradient features (see Figure 1 for illustration) are simply normalized vectors \( D \) that are given (approximately) by the simplified gradient vectors. (Similarly, the normalized scalar \( s \) is given by the bias \( b \).) To be a useful gradient feature, we require the direction to be “hit” by sufficiently large simplified gradient vectors with sufficient large probability, so as to be distinguished from noise and remain useful throughout the gradient steps. Later we will use the gradient features when \( \mathcal{W}, \mathcal{B} \) are the initialization distributions.

To make use of the gradient features, we consider the following family of networks using these features and with bounded norms, and will provide guarantees compared to the best in this family:

**Definition 3.9 (Gradient Feature Induced Networks).** The Gradient Feature Induced Networks are:
\[
\mathcal{F}_{d,m,B_F,S} := \{ f_{(a,b)}(w) \in \mathcal{F}_{d,m} \mid \forall i \in [m], |a_i| \leq B_{a_i}, |a_i|_2 \leq B_{a_2}, (w_i, b_i/|b_i|) \in S, |b_i| \leq B_b \},
\]
where \( S \) is some Gradient Feature set and \( B_F := (B_{a_1}, B_{a_2}, B_{b_1}) \) are some parameters.

**Remark 3.10.** In above definition, the weight and bias of a neuron are simply the scalings of some item in the feature set \( S \) (for simplicity the scaling of \( w_i \) is absorbed into the scaling of \( a_i \) and \( b_i \)).

**Definition 3.11 (Optimal Approximation via Gradient Features).** The optimal approximation network and loss using Gradient Feature Induced Networks \( \mathcal{F}_{d,r,B_F,S} \) are defined as:
\[
f^* \overset{\text{def}}{=} \arg \min_{f \in \mathcal{F}_{d,r,B_F,S}} \mathcal{L}_{D}(f), \quad \text{OPT}_{d,r,B_F,S} := \min_{f \in \mathcal{F}_{d,r,B_F,S}} \mathcal{L}_{D}(f).
\]

### 3.3 Provable Guarantee via Gradient Feature Learning

To obtain the guarantees, we first specify the symmetric initialization. It is canonical for the analysis and is typical in existing analysis (e.g., [7, 32, 33, 105]), though some other initialization can also work. Formally, we train a two-layer network with \( 4m \) neurons, \( f_{(a,b)}(w) \in \mathcal{F}_{d,4m} \). We initialize \( a_i^{(0)}, w_i^{(0)} \) from Gaussians and \( b_i \) from a constant for \( i \in \{1, \ldots, m\} \), and initialize the parameters for \( i \in \{m+1, \ldots, 4m\} \) accordingly to get a zero output initial network. Specifically:

- for \( i \in \{1, \ldots, m\} \) : \( a_i^{(0)} \sim \mathcal{N}(0, \sigma^2_a), w_i^{(0)} \sim \mathcal{N}(0, \sigma^2_w I), b_i = \bar{b} \),
- for \( i \in \{m+1, \ldots, 2m\} \) : \( a_i^{(0)} = -a_{i-m}^{(0)}, w_i^{(0)} = -w_{i-m}^{(0)}, b_i = -b_{i-m} \),
- for \( i \in \{2m+1, \ldots, 4m\} \) : \( a_i^{(0)} = -a_{i-2m}^{(0)}, w_i^{(0)} = w_{i-2m}^{(0)}, b_i = b_{i-2m} \),

where \( \sigma_a^2, \sigma_w^2, \bar{b} > 0 \) are hyper-parameters. After initialization, \( a, W \) are updated as in Algorithm 1. We are now ready to present our main result in the framework.

**Theorem 3.12 (Main Result).** Assume Assumption 3.1. For any \( \epsilon, \delta \in (0, 1) \), if \( m \leq e^d \) and
\[
m = \Omega \left( \frac{1}{p^4} \left( r B_{a_1} B_{x_1} \sqrt{B_b/B_G} \right)^4 + \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{T}{\delta} \right)^2 \right) \right),
\]
\[
T = \Omega \left( \frac{1}{\epsilon^2} \left( \sqrt{r} B_{x_2} B_{x_1} B_{x_2} \frac{m^2}{(mp)^{\frac{1}{2}}} + \bar{m} \right) \left( \frac{\sqrt{\log m}}{\sqrt{B_b B_G}} + \frac{1}{B_{x_1}(mp)^{\frac{1}{2}}} \right) \right),
\]
\[
\frac{n}{\log n} = \tilde{\Omega} \left( \frac{m^3 p B_{a_2} B_{x_2} B_{x_1}}{e^2 r^2 B_{a_1} B_G} + \frac{(mp)^{\frac{1}{2}} B_{x_2} B_{x_2}}{B_b B_G} + \frac{B_{x_2}^2}{B_{x_2}^2 + \frac{1}{p} + \frac{1}{B_{x_1}^2} + \frac{1}{B_{x_2}^2} \left( \mathcal{O}(0) \right)^2 + T m/\delta} \right),
\]
then with initialization (8) and proper hyper-parameter values, we have with probability \( \geq 1 - \delta \) over the initialization and training samples, there exists \( t \in [T] \) in Algorithm 1 with:
\[
\Pr[\text{sign}(f_{3\omega}(x)) \neq y] \leq \mathcal{L}_{D}(f_{3\omega}) \leq \text{OPT}_{d,r,B_F,S,B_{p,\gamma,B_G}} + B_{a_1} B_{x_1} \sqrt{2\gamma + O \left( \frac{B_{x_2} \log n}{B_G \mathcal{O}(0)n^{\frac{1}{2}}} \right) + \epsilon}.
\]
Intuitively, the theorem shows when a data distribution admits a small approximation error by some “ground-truth” network with $r$ neurons using gradient features from $S_{p, \gamma, B_G}$ (i.e., a small optimal approximate loss $\text{OPT}_{d, r, B_F, S_{p, \gamma, B_G}}$), the gradient descent training can successfully learn good neural networks with sufficiently many $m$ neurons.

Now we discuss the requirements and the error guarantee. Viewing boundedness parameters $B_{a_1}, B_{z_1}$ etc. as constants, then the number $m$ of neurons learned is roughly $\Theta \left( \frac{m^2}{r^2} \right)$, a polynomial overparameterization compared to the “ground-truth” network. The proof shows that such an overparameterization is needed such that some neurons can capture the gradient features given by gradient descent. This is consistent with existing analysis about overparameterization network learning, and also consistent with existing empirical observations.

The error bound consists of three terms. The last term $\epsilon$ can be made arbitrarily small, while the other two depend on the concrete data distribution. Specifically, with larger $r$ and $\gamma$, the second term increases. While the first term (the optimal approximation loss) decreases, since a larger $r$ means a larger “ground-truth” network family, and a larger $\gamma$ means a larger Gradient Feature set $S_{p, \gamma, B_G}$. So, there is a trade-off between these two terms. When we later apply the framework to concrete problems (e.g., mixtures of Gaussians, parity functions), we will show that depending on the specific data distribution, we can choose the proper values for $r, \gamma$ to make the error small. This then leads to error guarantees for the concrete problems and demonstrates the unifying power of the framework. Please refer to Appendix D.3 for more discussion about our problem setup and our core concept, e.g., parameter choice, early stopping, the role of $s$, activation functions, and so on.

**Proof Sketch.** The intuition in the proof of Theorem 3.12 is closely related to the notion of Gradient Features. First, the gradient descent will produce gradients that approximate the features in $S_{p, \gamma, B_G}$.

Then, the gradient descent update gives a good set of neurons, such that there exists an accurate convex learning using the good set of neurons. Technically, we adopt the online convex optimization classifier. Intuitively, with small step sizes $\eta$, the values, with probability at least $1 - \epsilon$, the weights of the first layer $w_i$ is initialized by Gaussians, by $\nabla_i$ (0), $a_i = \Theta\left( \frac{m^2}{r^2} \right)$, a polynomial overparameterization compared to the “ground-truth” network. The proof shows that such an overparameterization is needed such that some neurons can capture the gradient features given by gradient descent. This is consistent with existing analysis about overparameterization network learning, and also consistent with existing empirical observations.

The intuition in the proof of Theorem 3.12 is closely related to the notion of Gradient Features. First, the gradient descent will produce gradients that approximate the features in $S_{p, \gamma, B_G}$. Then, the gradient descent update gives a good set of neurons, such that there exists an accurate classifier using these neurons with loss comparable to the optimal approximation loss. Finally, the training will learn to approximate the accurate classifier, resulting in the desired error guarantee. The complete proof is in Appendix D (the population version in Appendix D.2 and the empirical version in Appendix D.4), including the proper values for hyper-parameters such as $\eta^{(i)}$ in Theorem D.17. Below, we briefly sketch the key ideas and omit the technical details.

We first show that a large subset of neurons has gradients at the first step as good features. (The claim can be extended to multiple steps; for simplicity, we follow existing work (e.g., [33, 105]) and present only the first step.) Let $\nabla_i$ denote the gradient of the $i$-th neuron $\nabla_i f, L_D(f^{(0)})$. Denote the subset of neurons with nice gradients approximating feature $(D, s)$ as:

$$G_{(D, s), \text{Nice}} := \left\{ i \in [2m] : s = b_i, ||\nabla_i f, (D, s)|| \geq \frac{\|G_i^{(0)}\|}{B_G} \right\}. \quad (9)$$

**Lemma 3.13** (Feature Emergence). For any $r$ size subset $(\{D_1, s_1\}, \ldots, (D_r, s_r)) \subseteq S_{p, \gamma, B_G}$, with probability at least $1 - re^{-\Theta(mp)}$, for all $j \in [r]$, we have $|G_{(D_j, s_j), \text{Nice}}| \geq \frac{mp}{4}$. This is because $\nabla_i = \ell'(0) a_i^{(0)} \mathbb{E}_{(x, y)} \left[ y \sigma' \left( \left( w_i^{(0)}, x \right) - b_j \right) x \right] = \ell'(0) a_i^{(0)} G_i^{(0)}, b_i$). Now consider $s_j = +1$ (the case $-1$ is similar). Since $w_i$ is initialized by Gaussians, by $\nabla_i$’s connection to Gradient Features, we can see that for all $i \in [m]$, $Pr \left[ i \in G_{(D_j, +1), \text{Nice}} \right] \geq \frac{\epsilon}{2}$. The lemma follows from concentration via a large enough $m$, i.e., sufficient overparameterization. The gradients allow obtaining a set of neurons approximating the “ground-truth” network with comparable loss:

**Lemma 3.14** (Existence of Good Networks). For any $\delta \in (0, 1)$, with proper hyper-parameter values, with probability at least $1 - \delta$, there is $\tilde{a}$ such that $\|\tilde{a}\|_0 = O\left( \sqrt{\frac{1}{mp}} \right)$ and $f(\tilde{a}, w_i^{(1)}, b) = \sum_{i=1}^{4m} \tilde{a}_i \sigma \left( \left( w_i^{(1)}, x \right) - b_i \right)$ satisfies:

$$L_D(f(\tilde{a}, w_i^{(1)}, b)) \leq \text{OPT}_{d, r, B_F, S_{p, \gamma, B_G}} + \sqrt{2rB_{a_1}B_{z_1}} \left( \sqrt{\gamma} + \frac{2B_b}{\sqrt{mpB_G}} \right).$$

Given the good set of neurons, we finally show that the remaining gradient steps can learn an accurate classifier. Intuitively, with small step sizes $\eta^{(i)}$, the weights of the first layer $w_i$ do not change too much (stay in a neighborhood) while the second layer weights grow, and thus the learning is similar to convex learning using the good set of neurons. Technically, we adopt the online convex optimization analysis (Theorem D.5) in [33] to get the final loss guarantee in Theorem 3.12.
4 Applications in Special Cases

In this section we will apply the gradient feature learning framework to some specific problems, corresponding to concrete data distributions $D$. We primarily focus on prototypical problems for analyzing feature learning in networks. We will present here the results for mixtures of Gaussians and parity functions, and include the complete proofs and some other results in Appendix E.

4.1 Mixtures of Gaussians

Mixtures of Gaussians are among the most fundamental and widely used statistical models. Recently, it has been used to study neural network learning, in particular, the effect of gradient descent for feature learning of two-layer neural networks and the advantage over fixed feature methods [46, 99].

Data Distributions. We follow notations from [99]. The data are from a mixture of $r$ high-dimensional Gaussians, and each Gaussian is assigned to one of two possible labels in $\mathcal{Y} = \{\pm 1\}$. Let $\mathcal{S}(y) \subseteq [r]$ denote the set of indices of Gaussians associated with the label $y$. The data distribution is then: $q(x, y) = q(y)q(x|y), q(x|y) = \sum_{j\in \mathcal{S}(y)} p_j N_j(x)$, where $N_j(x)$ is a multivariate normal distribution with mean $\mu_j$, covariance $\Sigma_j$, and $p_j$ are chosen such that $q(x, y)$ is correctly normalized.

We will make some assumptions about the Gaussians, for which we first introduce some notations.

$$D_j := \frac{\mu_j}{\|\mu_j\|_2}, \quad \tilde{\mu}_j := \mu_j/\sqrt{d}, \quad B_{\mu_1} := \min_{j\in [r]} \|\tilde{\mu}_j\|_2, \quad B_{\mu_2} := \max_{j\in [r]} \|\tilde{\mu}_j\|_2, \quad p_B := \min_{j\in [r]} p_j.$$

Assumption 4.1. Let $8 \leq \tau \leq d$ be a parameter that will control our final error guarantee. Assume

- Equi-probable labels: $q(-1) = q(+1) = 1/2$.
- For all $j \in [r]$, $\Sigma_j = \sigma_j I_{d \times d}$. Let $\sigma_B := \max_{j\in [r]} \sigma_j$ and $\sigma_{B+} := \max \{\sigma_B, B_{\mu_2}\}$.
- $r \leq 2d$, $p_B \geq \frac{1}{2d}$. $\Omega \left(1/d + \sqrt{\tau \sigma_B^2 \log d/d}\right) \leq B_{\mu_1} \leq B_{\mu_2} \leq d$.
- The Gaussians are well-separated: for all $i \neq j \in [r]$, we have $-1 \leq \langle D_i, D_j \rangle \leq \theta$, where $0 \leq \theta \leq \min \left\{\frac{1}{2\tau}, \frac{\sigma_{B+}}{B_{\mu_2}^2} \sqrt{\tau \log d/d}\right\}$.

Remark 4.2. The first two assumptions are for simplicity; they can be relaxed. We can generalize our analysis to the mixture of Gaussians with unbalanced label probabilities and general covariances. The third assumption is to make sure that each Gaussian has a good amount of probability mass to be learned. The remaining assumptions are to make sure that the Gaussians are well-separated and can be distinguished by the learning algorithm.

We are now ready to apply the framework to these data distributions, for which we only need to compute the Gradient Feature set and the corresponding optimal approximation loss.

Lemma 4.3 (Mixtures of Gaussians: Gradient Features). $(D_j, +1) \in S_{p, \gamma, B_G}$ for all $j \in [r]$, where

$$p = \frac{B_{\mu_1}}{\sqrt{\tau \log d} \sigma_{B+}} \cdot d^{\Theta(\sigma_{B+}^2/B_{\mu_1}^2)} \cdot \gamma = \frac{1}{d^{0.9\tau - 1.5}}, \quad B_G = p_B B_{\mu_1} \sqrt{d} - O\left(\frac{\sigma_{B+}}{d^{0.9\tau}}\right).$$

Let $f^*(x) = \sum_{j=1}^r \frac{B_{\mu_1}}{\sqrt{\tau \log d} \sigma_{B+}} \left[\sigma \left(\langle D_j, x \rangle - 2\sqrt{\tau \log d} \sigma_{B+}\right)\right]$ whose hinge loss is at most $\frac{3}{\tau \tau^*} + \frac{d^{\Theta(\sigma_{B+}^2/B_{\mu_1}^2)} \cdot r}{d^{0.9\tau - 4} \sqrt{\tau \log d}}$.

Given the values on gradient feature parameters $p, \gamma, B_G$ and the optimal approximation loss $\text{OPT}_{d, r, B_F, S_{p, \gamma, B_G}}$, the framework immediately leads to the following guarantee:

Theorem 4.4 (Mixtures of Gaussians: Main Result). Assume Assumption 4.1. For any $\epsilon, \delta \in (0, 1)$, when Algorithm 1 uses hinge loss with

$$m = \text{poly} \left(\frac{1}{\delta}, \frac{1}{\epsilon}, d^{\Theta(\sigma_{B+}^2/B_{\mu_1}^2)}, r, \frac{1}{p_B}\right) \leq e^d, \quad T = \text{poly} \left(m\right), \quad n = \text{poly} \left(m\right)$$

and proper hyper-parameters, then with probability at least $1 - \delta$, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f^*_{E^{(t)}(x)}) \neq y] \leq \frac{\sqrt{2r}}{d^{0.4\tau - 0.8}} + \epsilon.$$

7
The theorem shows that gradient descent can learn to a small error via learning the gradient features, given proper hyper-parameters. In particular, we need sufficient overparameterization (a sufficiently large number $m$ of neurons). When $\sigma_{B+2}/B^2_{m+1}$ is a constant which is the prototypical interesting case, and we choose a constant $\tau$, then $m$ is polynomial in the key parameters $\frac{1}{\tau^2}, \frac{1}{\tau}, d, r, \frac{1}{\rho_B}$, and the error bound is inverse polynomial in $d$. The complete proof is given in Appendix E.2.

[46] studies (almost) linear separable cases while our setting includes non-linear separable cases, e.g., XOR. [99] mainly studies neural network classification on 4 Gaussian clusters with XOR structured labels, while our setting is much more general, e.g., our cluster number can extend up to $2d$.

4.1.1 Mixtures of Gaussians: Beyond the Kernel Regime

As discussed in the introduction, it is important for the analysis to go beyond fixed feature methods such as NTK (i.e., the kernel regime), so as to capture the feature learning ability which is believed to be the key factor for the empirical success. We first review the fixed feature methods. Following [33, 76, 105], our framework only needs roughly $\Omega((\log d)^2)$ neurons and $\Omega((\log d)^2)$ samples to achieve arbitrary small constant error (see Theorem E.18 when $\sigma_B = 1$), while fixed feature methods need $\Omega(d^2)$ features and $\Omega(d^2)$ samples to achieve nontrivial errors (as proved in [99]). Moreover, [99] uses ODE to simulate the optimization process for the 2-layer networks learning XOR-shaped Gaussian mixture with $\Omega(1)$ neurons and gives convincing evidence that $\Omega(d)$ samples is enough to learn it, yet they do not give a rigorous convergence guarantee for this problem. We successfully derive a convergence guarantee and we require a much smaller sample size $\Omega((\log d)^2)$. For the proof (detailed in Appendix E.3), we only need to calculate the $p, \gamma, B_G$ of the data distribution carefully and then inject these numbers into Theorem 3.12.

4.2 Parity Functions

Parity functions are a canonical family of learning problems in computational learning theory, usually for theoretical computational barriers [103]. The typical sparse parities over $d$-dim binary inputs $\phi \in \{-1, 1\}^d$ are $\prod_{j \in A} \phi_j$, where $A \subseteq [d]$ is a subset of dimensions. Recent studies have shown that when the distribution of inputs $\phi$ has structures rather than uniform, neural networks can perform feature learning and finally learn parity functions with a small error, while methods without feature learning, e.g. NTK, cannot achieve as good results [33, 76, 105]. Thus, this has been a prototypical setting for studying feature learning phenomena in networks. Here we consider a generalization of this problem and show that our framework can show successful learning via gradient descent.

Data Distributions. Suppose $M \in \mathbb{R}^{d \times D}$ is an unknown dictionary with $D$ columns that can be regarded as patterns. For simplicity, assume $d = D$ and $M$ is orthonormal. Let $\phi \in \mathbb{R}^d$ be a hidden representation vector. Let $A \subseteq [D]$ be a subset of size $rk$ corresponding to the class relevant patterns and $r$ is an odd number. Then the input is generated by $M\phi$, and some function on $\phi_A$ generates the label. WLOG, let $A = \{1, \ldots, rk\}$, $A^+ = \{rk + 1, \ldots, d\}$. Also, we split $A$ such that for all $j \in [r], A_j = \{(j-1)k+1, \ldots, jk\}$. Then the input $x$ and the class label $y$ are given by:

$$x = M\phi, \quad y = g^*(\phi_A) = \text{sign} \left( \sum_{j \in [r]} \text{XOR}(\phi_{A_j}) \right), \quad (10)$$

where $g^*$ is the ground-truth labeling function mapping from $\mathbb{R}^{rk}$ to $Y = \{-1, 1\}, \phi_A$ is the sub-vector of $\phi$ with indices in $A$, and $\text{XOR}(\phi_{A_j}) = \prod_{\ell \in A_j} \phi_\ell$ is the parity function. We still need to specify the distribution $X$ of $\phi$, which determines the structure of the input distribution:

$$X := (1 - 2rp_A)X_U + \sum_{j \in [r]} p_A(X_{j,+} + X_{j,-}). \quad (11)$$
For all corresponding $\phi_{A^+}$ in $\mathcal{X}$, we have $\forall l \in A^+$, independently: 
$$
\phi_l = \begin{cases} 
+1, & \text{w.p. } p_o \\
-1, & \text{w.p. } p_o \\
0, & \text{w.p. } 1 - 2p_o 
\end{cases},
$$

where $p_o$ controls the signal noise ratio: if $p_o$ is large, then there are many nonzero entries in $A^+$ which are noise interfering with the learning of the ground-truth labeling function on $A$. For corresponding $\phi_A$, any $j \in [r]$, we have

- In $\mathcal{X}_{j,+}^r$, $\phi_{A_j} = [+1, +1, \ldots, +1]^T$ and $\phi_{A \setminus A_j}$ only have zero elements.
- In $\mathcal{X}_{j,-}^r$, $\phi_{A_j} = [-1, -1, \ldots, -1]^T$ and $\phi_{A \setminus A_j}$ only have zero elements.
- In $\mathcal{X}_u$, we have $\phi_A$ drawn from $\{+1, -1\}^r$ uniformly.

In short, we have $r$ parity functions each corresponding to a block of $k$ dimensions; $\mathcal{X}_{j,+}$ and $\mathcal{X}_{j,-}$ stands for the component providing a strong signal for the $j$-th parity; $\mathcal{X}_u$ corresponds to uniform distribution unrelated to any parity and providing weak learning signal; $A^+$ is the noise part. The label depends on the sum of the $r$ parity functions.

**Assumption 4.5.** Let $8 \leq \tau \leq d$ be a parameter that will control our final error guarantee. Assume $k$ is an odd number and: $k \geq \Omega(\tau \log d)$, $d \geq rk + \Omega(\tau r \log d)$, $p_o = O\left(\frac{r}{d^2}\right)$, $p_A \geq \frac{1}{d}$.

**Remark 4.6.** We set up the problem to be more general than the parity function learning in existing work. If $r = 1$, the labeling function reduces to the traditional $k$-sparse parties of $d$ bits. The assumptions require $k$, $d$, and $p_A$ to be sufficiently large so as to provide enough large signals for learning. Note that when $k = \frac{d}{r}$, $r = 1$, $p_o = \frac{1}{k}$, our analysis also holds, which shows our framework is beyond the kernel regime (discuss in detail in Section 4.2.1).

To apply our framework, again we only need to compute the Gradient Feature set and the corresponding optimal loss. We first define the Gradient Features: For all $j \in [r]$, let $D_j = \frac{\sum_{i \in A_j} M_i}{\|\sum_{i \in A_j} M_i\|_2}$.

**Lemma 4.7 (Parity Functions: Gradient Features).** We have $(D_j, +1), (D_j, -1) \in S_{p, \gamma, B_G}$ for all $j \in [r]$, where

$$
p = \Theta\left(\frac{1}{\sqrt{\tau^2 \log d \cdot d^{3\theta(\tau)}}}\right), \quad \gamma = \frac{1}{d^{\tau - 2}}, \quad B_G = \sqrt{k}p_A - O\left(\frac{\sqrt{k}}{d^\tau}\right). \tag{12}
$$

With gradient features from $S_{p, \gamma, B_G}$, let $f^*(x) = \sum_{j=1}^r \sum_{i=0}^k (-1)^{i+1} \sqrt{\tau} \left[ \sigma \left( (D_j, x) - \frac{2i - k - 1}{\sqrt{k}} \right) - 2\sigma \left( (D_j, x) - \frac{2i - k + 1}{\sqrt{k}} \right) \right]$ whose hinge loss is $0$.

Above, we show that $D_j$ is the “indicator function” for the subset $A_j$ so that we can build the optimal neural network based on such directions. Given the values on gradient feature parameters and the optimal approximation loss, the framework immediately leads to the following guarantee:

**Theorem 4.8 (Parity Functions: Main Result).** Assume Assumption 4.5. For any $\epsilon, \delta \in (0, 1)$, when Algorithm 1 uses hinge loss with

$$
m = \text{poly}\left(\frac{1}{\delta}, \frac{1}{\epsilon}, d^{3\theta(\tau)}, k, \frac{1}{p_A}\right) \leq e^d, \quad T = \text{poly}(m), \quad n = \text{poly}(m)
$$

and proper hyper-parameters, then with probability at least $1 - \delta$, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f_{\Xi(t)}(x)) \neq y] \leq \frac{3\tau \sqrt{K}}{d^{(\tau - 3)/2}} + \epsilon.$$

The theorem shows that gradient descent can learn to a small error in this problem. We also need sufficient overparameterization: When $r$ is a constant (e.g., $r = 1$ in existing work), and we choose a constant $\tau$, $m$ is polynomial in $\frac{1}{\delta}, \frac{1}{\epsilon}, d, k, \frac{1}{p_A}$, and the error bound is inverse polynomial in $d$. The proof is in Appendix E.4. Our setting is more general than that in [33, 76] which corresponds to $M = I, r = 1, p_A = \frac{1}{\tau}, p_o = \frac{1}{2}$. [105] study single index learning, where one feature direction is enough for a two-layer network to recover the label, while our setting considers $r$ directions $D_1, \ldots, D_r$, so the network needs to learn multiple directions to get a small error.
4.2.1 Parity Functions: Beyond the Kernel Regime

Again, we show that our framework indeed goes beyond fixed features under parity functions. Our problem setting in Section 4.2 is general enough to include the problem setting in [33]. Their lower bound for fixed feature methods directly applies to our case and leads to the following:

**Proposition 4.9.** There exists a data distribution in the parity learning setting in Section 4.2 with $M = I, r = 1, p_A = \frac{1}{7}, k = \frac{4}{16}, p_o = \frac{1}{2}$, such that all $h \in \mathcal{H}_B$ have hinge-loss at least $\frac{1}{2} - \frac{\sqrt{NB}}{2^k \sqrt{2}}$.

This means to get an inverse-polynomially small loss, fixed feature models need to have an exponentially large size, i.e., either the number of features $N$ or the norm $B$ needs to be exponential in $k$. In contrast, Theorem 4.8 shows our framework guarantees a small loss with a polynomially large model, runtime, and sample complexity. Clearly, our framework is beyond the fixed feature methods.

**Parities on Uniform Inputs.** When $r = 1, p_A = 0$, our problem setting will degenerate to the classic sparse parity function on a uniform input distribution. This has also been used for analyzing network learning [16]. For this case, our framework can get a $k2^O(k) \log(k)$ network width bound and a $O(d^k)$ sample complexity bound, matching those in [16]. This then again confirms the advantage of network learning over kernel methods that requires $d^\Omega(k)$ dimensions as shown in [16]. See the full statement in Theorem E.31, details in Appendix E.5, and alternative analysis in Appendix E.6.

5 Further Implications and Conclusion

Our general framework sheds light on several interesting phenomena in NN learning observed in practice. Feature learning beyond the kernel regime has been discussed in Section 4.1.1 and Section 4.2.1. Here we discuss the LTH and defer more implications such as simplicity bias, learning over different data distributions, and new perspectives about roadmaps forward in Appendix C.

**Lottery Ticket Hypothesis (LTH).** Another interesting phenomenon is the LTH [41]: randomly-initialized networks contain subnetworks that when trained in isolation reach test accuracy comparable to the original network in a similar number of iterations. Later studies (e.g., [42]) show that LTH is more stable when subnetworks are found in the network after a few gradient steps.

Our framework provides an explanation for two-layer networks: the lottery ticket subnetwork contains exactly those neurons whose gradient feature approximates the weights of the “ground-truth” network $f^*$; they may not exist at initialization but can be found after the first gradient step. More precisely, Lemma 3.14 shows that after the first gradient step, there is a sparse second-layer weight $\tilde{a}$ with $\|a\|_0 = O(r \sqrt{mp})$, such that using this weight on the hidden neurons gives a network with a small loss. Let $U$ be the support of $\tilde{a}$. Equivalently, there is a small-loss subnetwork $f_U$ with only neurons in $U$ and with second-layer weight $\tilde{a}_U$ on these neurons. Following the same proof of Theorem 3.12:

**Proposition 5.1.** In the same setting of Theorem 3.12 but only considering the subnetwork supported on $U$ after the first gradient step, with the same requirements on $m$ and $T$, with proper hyperparameter values, we have the same guarantee: with probability $\geq 1 - \delta$, there is $t \in [T]$ with

$$\Pr[\text{sign}(f_{U}^T(x)) \neq y] \leq \text{OPT}_{d,r,B_F,S,p,\gamma,B_G} + rB_{a1}B_{x1} \left(2\gamma + O\left(\frac{\sqrt{B_{F}\log n}}{B_G \sqrt{n}}\right)\right) + \epsilon.$$  

This essentially formally proves LTH for two-layer networks, showing (a) the existence of the winning lottery subnetwork and (b) that gradient descent on the subnetwork can learn to similar loss in similar runtime as on the whole network. In particular, (b) is novel and not analyzed in existing work.

We provide our work’s broader impacts and limitations (e.g., statement of recovering existing results and some failure cases beyond our framework) in Appendix A and Appendix B respectively.

**Conclusion.** We propose a general framework for analyzing two-layer neural network learning by gradient descent and show that it can lead to provable guarantees for several prototypical problem settings for analyzing network learning. In particular, our framework goes beyond fixed feature methods, e.g., NTK. It sheds light on several interesting phenomena in NN learning, e.g., the lottery ticket hypothesis and simplicity bias. Future directions include: (1) How to extend the framework to deeper networks? (2) While the current framework focuses on the gradient features in the early gradient steps, whether feature learning also happens in later steps and if so how to formalize that?
Acknowledgements

The work is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants 2008559-IIS, 2023239-DMS, and CCF-2046710.

References


## Contents

1 Introduction 1

2 Related Work 2

3 Gradient Feature Learning Framework 3
   3.1 Warm Up: A Simple Setting with Frozen First Layer 3
   3.2 Core Concepts in the Gradient Feature Learning Framework 4
   3.3 Provable Guarantee via Gradient Feature Learning 5

4 Applications in Special Cases 7
   4.1 Mixtures of Gaussians 7
      4.1.1 Mixtures of Gaussians: Beyond the Kernel Regime 8
   4.2 Parity Functions 8
      4.2.1 Parity Functions: Beyond the Kernel Regime 10

5 Further Implications and Conclusion 10

A Broader Impacts 19

B Limitations 19

C More Further Implications 20

D Gradient Feature Learning Framework 21
   D.1 Simplified Gradient Feature Learning Framework 21
   D.2 Gradient Feature Learning Framework under Expected Risk 22
      D.2.1 Feature Learning 23
      D.2.2 Good Network Exists 24
      D.2.3 Learning an Accurate Classifier 26
   D.3 More Discussion about Setting 30
   D.4 Gradient Feature Learning Framework under Empirical Risk with Sample Complexity 31

E Applications in Special Cases 38
   E.1 Linear Data 38
   E.2 Mixtures of Gaussians 40
      E.2.1 Problem Setup 40
      E.2.2 Mixtures of Gaussians: Feature Learning 41
      E.2.3 Mixtures of Gaussians: Final Guarantee 46
   E.3 Mixtures of Gaussians - XOR 48
      E.3.1 Mixtures of Gaussians - XOR: Feature Learning 48
Appendix A discusses the potential societal impact of our work. Appendix B describes the limitations of our work. In Appendix C, we present our framework implications about simplicity bias. The complete proof of our main results is given in Appendix D. We present the case study of linear data in Appendix E.1, mixtures of Gaussians in Appendix E.2 and Appendix E.3, parity functions in Appendix E.4, Appendix E.5 and Appendix E.6, and multiple-index models in Appendix E.7. We put the auxiliary lemmas in Appendix F.

A Broader Impacts

Our paper is purely theoretical in nature, and thus we do not anticipate an immediate negative ethical impact. We provide a unified theoretical framework that can be applied to different theoretical problems. We propose the two key ideas of gradient feature and gradient feature-induced neural networks not only to show their ability to unify several current works but also to open a new direction of thinking with respect to the learning process. These notations have the potential to be extended to multi-layer gradient features and multi-step learning, and this work is only our first step.

On the other hand, this work may lead to a better understanding and inspire the development of improved network learning methods, which may have a positive impact on the theoretical machine-learning community. It may also be beneficial to engineering-inclined machine-learning researchers.

B Limitations

Recover Existing Results. The framework may or may not recover the width or sample complexity bounds in existing work.

1. The framework can give matching bounds as the existing work in some cases, like parities over uniform inputs (Appendix E.5).
2. In some other cases, it gives polynomial error bounds not the same as those in the existing work (e.g., for parities over structured inputs). This is because our work is analyzing general cases, and thus may not give better than or the same bounds as those in special cases, since special cases have more properties that can be exploited to get potentially better bounds.

We would like to emphasize that our contribution is providing an analysis framework that can (1) formalize the unifying principles of learning features from gradients in network training, and (2) give
polynomial error bounds for prototypical problems. Our focus is not to recover the guarantees in existing work.

**Failure Cases.** There are some failure cases that gradient feature learning framework cannot cover:

1. In [101], they constructed a function that is easy to approximate using a 3-layer network but not approximable by any 2-layer network. Since the function is not approximable by any 2-layer network, it cannot be approximated by the gradient-induced networks as well, so OPT will be large. As a result, the final error will be large.

2. In uniform parity data distribution, considering an odd number of features rather than even, i.e., \( k \) is an odd number in Assumption E.28, we can show that our gradient feature set is empty even when \( p \) in Equation (6) is exponentially small, thus the OPT is a positive constant since the gradient induced network can only be constants. Meanwhile, the neural network won’t be able to learn this data distribution because its gradient is always 0 through the training, and the final error equals OPT.

The first case corresponds to the approximation hardness of 2-layer networks, while the second case gives a learning hardness example. The above two cases show that if there is an approximation or learning hardness, our gradient feature learning framework may be vacuous because the optimal model in the gradient feature class has a large risk, then the ground-truth mapping from inputs to labels is not learnable by gradient descent. These analyses are consistent with previous works [15, 101].

**C More Further Implications**

Our general framework also sheds some light on several interesting phenomena in neural network (NN) learning observed in practice. Feature learning beyond the kernel regime has been discussed in Section 4.1.1 and Section 4.2.1. The lottery ticket hypothesis (LTH) has been discussed in Section 5. Below we discuss other implications.

**Implicit Regularization/Simplicity Bias.** It is now well known that practical NN are overparameterized and traditional uniform convergence bounds cannot adequately explain their generalization performance [59, 88, 118]. It is generally believed that the optimization has some implicit regularization effect that restricts learning dynamics to a subset of the whole hypothesis class, which is not of high capacity so can lead to good generalization [53, 90]. Furthermore, learning dynamics tend to first learn simple functions and then learn more and more sophisticated ones (referred to as simplicity bias) [89, 102]. However, it remains elusive to formalize such simplicity bias.

Our framework provides a candidate explanation: the learning dynamics first learn to approximate the best network in a smaller family of gradient feature induced networks \( \mathcal{F}_{d,r_1,B,F,S} \) and then learn to approximate the best in a larger family. Consider the number of neurons \( r \) for illustration. Let \( r_1 \ll r_2 \), and let \( T_1 \) and \( T_2 \) be their corresponding runtime bounds for \( T \) in the main Theorem 3.12. Clearly, \( T_1 \ll T_2 \). Then, at time \( T_1 \), the theorem guarantees the learning dynamics learn to approximate the best in the family \( \mathcal{F}_{d,r_1,B,F,S} \) with \( r_1 \) neurons, but not for the larger family \( \mathcal{F}_{d,r_2,B,F,S} \). Later, at time \( T_2 \), the learning dynamics learn to approximate the best in the larger family \( \mathcal{F}_{d,r_2,B,F,S} \). That is, the learning first learns simpler functions and then more sophisticated ones where the simplicity bias is measured by the size of the family of gradient feature-induced networks. The implicit regularization is then restricting to networks approximating smaller families of gradient feature-induced networks. Furthermore, we can also conclude that for an SGD-optimized NN, its actual representation power is from the subset of NN based on gradient features, instead of the whole set of NN. This view helps explain the simplicity bias/implicit regularization phenomenon of NN learning in practice.

**Learning over Different Data Distributions.** Our framework articulates the following key principles (pointed out for specific problems in existing work but not articulated more generally):

- Role of gradient: the gradient leads to the emergence of good features, which is useful for the learning of upper layers in later stages.
- From features to solutions: learned features in early steps will not be distorted, if not improved, in later stages. The training dynamic for upper layers will eventually learn a good combination of hidden neurons based on gradient features, giving a good solution.
Then, more interesting insights are obtained from the generality of the framework. To build a general framework, the meaningful error guarantees should be data-dependent, since NN learning on general data distributions is hard and data-independent guarantees will be vacuous [34, 35]. Comparing the optimal in a family of “ground-truth” functions (inspired by agnostic learning in learning theory) is a useful method to obtain the data-dependent bound. We further construct the “ground-truth” functions using properties of the training dynamics, i.e., gradient features. This greatly facilitates the analysis of the training dynamics and is the key to obtaining the final guarantees. On the other hand, the framework can also be viewed as using the optimal by gradient-induced NN to measure or quantify the “complexity” of the problem. For easier problems, this quantity is smaller, and our framework can give a better error bound. So this provides a united way to derive guarantees for specific problems.

**New Perspectives about Roadmaps Forward.** We argue a new perspective about the connection between the strong representation power and the successful learning of NN. Traditionally, the strong representation power of NN is the key reason for hardness results of NN learning: NN has strong representation power and can encode hard learning questions, so they are hard to learn. See the proof in SQ bound from [33] or NP-hardness from [19]. The strong representation power also causes trouble for the statistical aspect: it leads to vacuous generalization bounds when traditional uniform convergence tools are used.

Our framework suggests a perspective in sharp contrast: the strong representation power of NN with gradient features is actually the key to successful learning. More concretely, the optimal error of the gradient feature-induced NN being small (i.e., strong representation power for a given data distribution) can lead to a small guarantee, which is the key to successful learning. The above new perspective suggests a different analysis road than traditional ones. Traditional analysis typically first reasons about the optimal based on the whole function class, i.e., the ground truth, then analyze how NN learns proper features and reaches the optimal. In contrast, our framework defines feature family first, and then reasons about the optimal based on it.

Our framework provides the foundation for future work on analyzing gradient-based NN learning, which may inspire future directions including but not limited to (1) defining a new feature family for 2-layer NN rather than gradient feature, (2) considering deep NN and introducing new gradient features (e.g., gradient feature notion for upper layers), (3) defining different gradient feature family at different training stages (e.g., gradient feature notion for later stages). In particular, the challenges in the later-stage analysis are: (a) the weights in the later stage will not be as normal as the initialization, and we need new tools to analyze their properties; (b) to show that the later-stage features eventually lead to a good solution, we may need new analysis tools for the non-convex optimization due to the changes in the first layer weights.

**D Gradient Feature Learning Framework**

We first prove a Simplified Gradient Feature Learning Framework in Appendix D.1, which only considers one-step gradient feature learning. Then, we prove our Gradient Feature Learning Framework, e.g., no freezing of the first layer. In Appendix D.2, we consider population loss to simplify the proof. Then, we provide more discussion about our problem setup and our core concept in Appendix D.3. Finally, we prove our Gradient Feature Learning Framework under empirical loss considering sample complexity in Appendix D.4.

**D.1 Simplified Gradient Feature Learning Framework**

Algorithm 2 Training by Algorithm 1 with no updates for the first layer after the first gradient step

\[
\text{Initialize } f(a^{(0)}, W^{(0)}, b) \in \mathcal{F}_{d, m}; \text{ Sample } \tilde{Z} \sim \mathcal{D}^{n} \\
\text{Get } (a^{(1)}, W^{(1)}, b) \text{ by one gradient step update and fix } W^{(1)}, b \\
\text{for } t = 2 \text{ to } T \text{ do } \\
a^{(t)} = a^{(t-1)} - \eta^{(t)} \nabla a \tilde{L}(f_{\tilde{Z}}(t)) \\
\text{end for}
\]

**Theorem 3.4 (Simple Setting).** Assume \( \tilde{L}(f_{(a,W^{(t)},b)}) \) is \( L \)-smooth to \( a \). Let \( \eta^{(t)} = \frac{1}{L} \), \( \lambda^{(t)} = 0 \), for all \( t \in \{2,3,\ldots,T\} \). Training by Algorithm 1 with no updates for the first layer after the
We consider the following training process under population loss to simplify the proof. We prove that under Algorithm 3, we have the following results. Then our theorem gets proved by Lemma F.9 and generalization bounds based on Rademacher complexity.

Proof of Theorem 3.4. Recall that

\[\mathcal{F}_{W,b,B_{a2}} := \{f(a,W,b) \in \mathcal{F}_{d,m} \mid \|a\|_2 \leq B_{a2}\}, \quad \text{OPT}_{W,b,B_{a2}} := \min_{f \in \mathcal{F}_{W,b,B_{a2}}} \mathcal{L}_D(f)\]  

(13)

We denote \(f^* = \arg\min_{f \in \mathcal{F}_{W,b,B_{a2}}} \mathcal{L}_D(f)\) and \(\tilde{f}^* = \arg\min_{f \in \mathcal{F}_{W,b,B_{a2}}} \tilde{\mathcal{L}}_Z(f)\). We use \(a^*\) and \(\tilde{a}^*\) to denote their second layer weights respectively. Then, we have

\[
\mathcal{L}_D(f(a^{(t)},W^{(1)},b)) = \mathcal{L}_D(f(a^{(t)},W^{(1)},b)) - \tilde{\mathcal{L}}_Z(f(a^{(t)},W^{(1)},b))
\]  

(14)

\[
+ \tilde{\mathcal{L}}_Z(f(a^{(t)},W^{(1)},b)) - \tilde{\mathcal{L}}_Z(f(\tilde{a}^*,W^{(1)},b))
\]  

(15)

\[
+ \tilde{\mathcal{L}}_Z(f(\tilde{a}^*,W^{(1)},b)) - \mathcal{L}_D(f(a^{(t)},W^{(1)},b))
\]  

(16)

\[
+ \mathcal{L}_D(f(a^{(t)},W^{(1)},b))
\]  

(17)

\[
\leq \left| \mathcal{L}_D(f(a^{(t)},W^{(1)},b)) - \tilde{\mathcal{L}}_Z(f(a^{(t)},W^{(1)},b)) \right|
\]  

(18)

\[
+ \left| \tilde{\mathcal{L}}_Z(f(a^{(t)},W^{(1)},b)) - \tilde{\mathcal{L}}_Z(f(\tilde{a}^*,W^{(1)},b)) \right|
\]  

(19)

\[
+ 0
\]  

(20)

\[
+ \left| \tilde{\mathcal{L}}_Z(f(\tilde{a}^*,W^{(1)},b)) - \mathcal{L}_D(f(a^{(t)},W^{(1)},b)) \right|
\]  

(21)

\[
+ \mathcal{L}_D(f(a^{(t)},W^{(1)},b))
\]  

(22)

\[
+ \text{OPT}_{W^{(1)},b,B_{a2}}
\]  

(23)

Fixing \(W^{(1)}, b\) and optimizing \(a\) only is a convex optimization problem. Note that \(\eta \leq \frac{1}{T}\), where \(\tilde{\mathcal{L}}_Z\) is \(L\)-smooth to \(a\). Thus with gradient descent, we have

\[
\frac{1}{T} \sum_{t=1}^{T} \tilde{\mathcal{L}}_Z(f(a^{(t)},W^{(1)},b)) - \tilde{\mathcal{L}}_Z(f(\tilde{a}^*,W^{(1)},b)) \leq \frac{\|a^{(1)} - a^*\|^2_2}{2T \eta}.
\]  

(24)

Then our theorem gets proved by Lemma F.9 and generalization bounds based on Rademacher complexity.

\[\square\]

D.2 Gradient Feature Learning Framework under Expected Risk

We consider the following training process under population loss to simplify the proof. We prove our Gradient Feature Learning Framework under empirical loss considering sample complexity in Appendix D.4.

**Algorithm 3** Network Training via Gradient Descent

```plaintext
 initialized (a^{(0)}, W^{(0)}, b) as in Equation (8)
 for t = 1 to T do
   a^{(t)} = a^{(t-1)} - \eta^{(t)} \nabla_a \mathcal{L}_D(f_{\Xi^{(t-1)}})
   W^{(t)} = W^{(t-1)} - \eta^{(t)} \nabla_w \mathcal{L}_D(f_{\Xi^{(t-1)}})
 end for
```

Given an input distribution, we can get a Gradient Feature set \(S_{p,\gamma,B_G}\) and \(f^*(x) = \sum_{j=1}^{d} a^*_j \sigma(\langle w^*_j, x \rangle - b^*_j)\), where \(f^* \in \mathcal{F}_{d,r,B_F,S_{p,\gamma,B_G}}\) is a Gradient Feature Induced networks defined in Definition 3.11. Considering training by Algorithm 3, we have the following results.
Algorithm 3

For any $\epsilon, \delta$ we complete the proof by union bound.

Theorem D.1 (Gradient Feature Learning Framework under Expected Risk). Assume Assumption 3.1. For any $\epsilon, \delta \in (0,1)$, if $m \leq \epsilon^d$ and

$$m = \Omega \left( \frac{1}{p} \left( \frac{r B_{a1} B_{x1}}{\epsilon} \sqrt{B_{b}} \right) + \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{t}{\delta} \right) \right)^2 \right), \quad \text{(25)}$$

$$T = \Omega \left( \frac{1}{\epsilon} \left( \frac{\sqrt{r B_{a2} B_{b} B_{x1}}}{(mp)^{\frac{1}{4}}} + mb \left( \frac{\sqrt{\log m}}{\sqrt{B_{b}} B_{G}} + \frac{1}{B_{x1}} \right) \right) \right), \quad \text{(26)}$$

then with proper hyper-parameter values, we have with probability $\geq 1 - \delta$, there exists $t \in [T]$ in Algorithm 3 with

$$\Pr[\text{sign}(f_{\Xi(t)}(x)) \neq y] \leq \mathcal{L}_D(f_{\Xi(t)}) \leq \text{OPT}_{d,r,B_F,S_{p,\gamma},B_G} + r B_{a1} B_{x1} \sqrt{2\gamma} + \epsilon. \quad \text{(27)}$$

See the full statement and proof in Theorem D.9. Below, we show some lemmas used in the analysis of population loss.

D.2.1 Feature Learning

We first show that a large subset of neurons has gradients at the first step as good features.

Definition D.2 (Nice Gradients Set. Equivalent to Equation (9)). We define

$$G_{(D,+1),\text{Nice}} := \left\{ i \in [m] : \left< w_i^{(1)}, D \right> > (1 - \gamma) \| w_i^{(1)} \|_2, \| w_i^{(1)} \|_2 \geq \| \eta^{(1)} \ell'(0) a_i^{(0)} \| B_G \right\}$$

$$G_{(D,-1),\text{Nice}} := \left\{ i \in [2m] \setminus [m] : \left< w_i^{(1)}, D \right> > (1 - \gamma) \| w_i^{(1)} \|_2, \| w_i^{(1)} \|_2 \geq \| \eta^{(1)} \ell'(0) a_i^{(0)} \| B_G \right\}$$

where $\gamma, B_G$ is the same in the Definition 3.7.

Lemma D.3 (Feature Emergence. Full Statement of Lemma 3.13). Let $\lambda^{(1)} = \frac{1}{\eta^{(1)}}$. For any $r$ size subset $\{(D_1, s_1), \ldots, (D_r, s_r)\} \subseteq S_{p,\gamma,B_G}$, with probability at least $1 - 2re^{-cmp}$ where $c > 0$ is a universal constant, we have that for all $j \in [r]$, $|G_{(D_j,s_j),\text{Nice}}| \geq \frac{mp}{r^2}$.

Proof of Lemma D.3. By symmetric initialization and Lemma F.1, we have for all $i \in [2m]

$$w_i^{(1)} = -\eta^{(1)} \ell'(0) a_i^{(0)} \mathbb{E}(x, b) \left[ y \sigma' \left( \left< w_i^{(0)}, x \right> - b \right) x \right] \quad \text{(28)}$$

$$= -\eta^{(1)} \ell'(0) a_i^{(0)} G(w_i^{(0)}, b_i). \quad \text{(29)}$$

For all $j \in [r]$, as $(D_j, s_j) \in S_{p,\gamma,B_G}$, by Lemma F.3,

1. if $s_j = +1$, for all $i \in [m]$, we have

$$\Pr \left[ i \in G_{(D_j, s_j), \text{Nice}} \right] \geq \frac{p}{2}, \quad \text{(30)}$$

2. if $s_j = -1$, for all $i \in [2m] \setminus [m]$, similarly we have

$$\Pr \left[ i \in G_{(D_j, s_j), \text{Nice}} \right] \geq \frac{p}{2}. \quad \text{(34)}$$

By concentration inequality, (Chernoff’s inequality under small deviations), we have

$$\Pr \left[ |G_{(D_j, s_j), \text{Nice}}| < \frac{mp}{4} \right] \leq 2e^{-c mp}. \quad \text{(35)}$$

We complete the proof by union bound. \(\square\)
D.2.2 Good Network Exists

Then, the gradients allow for obtaining a set of neurons approximating the “ground-truth” network with comparable loss.

**Lemma D.4 (Existence of Good Networks. Full Statement of Lemma 3.14).** Let \( \lambda^{(1)} = \frac{1}{m^2} \). For any \( B_\epsilon \in (0, B_0) \), let \( \sigma_\epsilon = \Theta \left( \frac{\tilde{b}}{-\ell'(0)\eta^{(1)} B_\epsilon} \right) \) and \( \delta = 2re^{-\sqrt{mp}} \). Then, with probability at least \( 1 - \delta \) over the initialization, there exists \( \tilde{a}_i \)'s such that \( f(\tilde{a}, \tilde{W}^{(1)}; \tilde{b}) = \sum_{i=1}^{4m} \tilde{a}_i \sigma \left( \langle w_i^{(1)}, x \rangle - b_i \right) \) satisfies

\[
\mathcal{L}_D(f(\tilde{a}, \tilde{W}^{(1)}; \tilde{b})) \leq rB_{\epsilon 1} \left( \frac{B_2^2 B_0}{\sqrt{mp} B_\epsilon} + B_{\epsilon 1} \sqrt{2\gamma} + B_\epsilon \right) + \text{OPT}_{d, r, B_F, S_{\epsilon, \gamma}}.
\]

and \( \|\tilde{a}\|_0 = O \left( r(mp)^{\frac{1}{2}} \right), \|\tilde{a}\|_2 = O \left( \frac{B_{\epsilon 2} B_0}{b(mp)^{\frac{1}{2}}} \right) \|\tilde{a}\|_\infty = O \left( \frac{B_{\epsilon 1} B_0}{b(mp)^{\frac{1}{2}}} \right) \).

**Proof of Lemma D.4.** Recall \( f^*(x) = \sum_{j=1}^{r} a_j^* \sigma(\langle w_j^*, x \rangle - b_j^*) \), where \( f^* \in \mathcal{F}_{d, r, B_F, S_{\epsilon, \gamma}} \) is defined in Definition 3.11 and let \( s_j^* = \frac{b_j^*}{|b_j^*|} \). By Lemma D.3, with probability at least \( 1 - \delta_1, \delta_1 = 2re^{-\sqrt{mp}} \), for all \( j \in [r] \), we have \( |G(w_j^*, s_j^*)|_{\text{Nice}} \geq \frac{mp}{4} \). Then for all \( i \in G(w_j^*, s_j^*) \), we have

\[
-\ell'(0)\eta^{(1)} G(w_i^{(0)}, b_i) b_j^* \ 	ext{only depend on } w_i^{(0)} \text{ and } b_i, \text{ which is independent of } a_i^{(0)}. \]

Given Definition 3.7, we have

\[
-\ell'(0)\eta^{(1)} ||G(w_i^{(0)}, b_i)||_2 b_j^* = \left[ \ell'(0)\eta^{(1)} B_{\epsilon 1} \frac{B_0}{b} - \ell'(0)\eta^{(1)} B_{\epsilon 1} \frac{B_0}{b} \right].
\]

We split \( [r] \) into \( \Gamma = \{ j \in [r] : |b_j^*| < B_\epsilon \}, \Gamma_- = \{ j \in [r] : b_j^* \leq -B_\epsilon \} \) and \( \Gamma_+ = \{ j \in [r] : b_j^* \geq B_\epsilon \} \). Let \( \epsilon_a = \frac{B_{\epsilon 1} B_0}{\sqrt{mp} B_\epsilon} \). Then we know that for all \( j \in \Gamma_+ \cup \Gamma_- \), for all \( i \in G(w_j^*, s_j^*) \), Nice, we have

\[
\Pr_{a_i^{(0)} \sim \mathcal{N}(0, \sigma^2_i)} \left[ \left| -a_i^{(0)} \ell'(0)\eta^{(1)} ||G(w_i^{(0)}, b_i)||_2 \frac{|b_j^*|}{b} - 1 \right| \leq \epsilon_a \right] = \Pr_{g \sim \mathcal{N}(0, 1)} \left[ 1 - \epsilon_a \leq g \Theta \left( \frac{||G(w_i^{(0)}, b_i)||_2 |b_j^*|}{B_G B_\epsilon} \right) \leq 1 + \epsilon_a \right]
\]

\[
= \Pr_{g \sim \mathcal{N}(0, 1)} \left[ 1 - \epsilon_a \leq g \Theta \left( \frac{||G(w_i^{(0)}, b_i)||_2 |b_j^*|}{B_G B_\epsilon} \right) \leq 1 + \epsilon_a \right]
\]

\[
= \Theta \left( \frac{\epsilon_a B_G B_\epsilon}{||G(w_i^{(0)}, b_i)||_2 |b_j^*|} \right)
\]

\[
\Omega \left( \frac{\epsilon_a B_G B_\epsilon}{B_{\epsilon 1} B_0} \right)
\]

\[
\Omega \left( \frac{\epsilon_a B_G B_\epsilon}{B_{\epsilon 1} B_0} \right)
\]

Thus, with probability \( \Omega \left( \frac{1}{\sqrt{mp}} \right) \) over \( a_i^{(0)} \), we have

\[
\left| -a_i^{(0)} \ell'(0)\eta^{(1)} ||G(w_i^{(0)}, b_i)||_2 \frac{|b_j^*|}{b} - 1 \right| \leq \epsilon_a, \ |a_i^{(0)}| = O \left( \frac{\tilde{b}}{-\ell'(0)\eta^{(1)} B_G B_\epsilon} \right).
\]

Similarly, for \( j \in \Gamma, \) for all \( i \in G(w_j^*, s_j^*), \text{Nice}, \) with probability \( \Omega \left( \frac{1}{\sqrt{mp}} \right) \) over \( a_i^{(0)} \), we have

\[
\left| -a_i^{(0)} \ell'(0)\eta^{(1)} ||G(w_i^{(0)}, b_i)||_2 \frac{B_j}{b} - 1 \right| \leq \epsilon_a, \ |a_i^{(0)}| = O \left( \frac{\tilde{b}}{-\ell'(0)\eta^{(1)} B_G B_\epsilon} \right).
\]
For all $j \in [r]$, let $\Lambda_j \subseteq G(w_j^*, s_j^*, N_{i,j})$ be the set of $i$'s such that condition Equation (44) or Equation (45) are satisfied. By Chernoff bound and union bound, with probability at least $1 - \delta_2$, $\delta_2 = re^{-\sqrt{mp}}$, for all $j \in [r]$ we have $|\Lambda_j| \geq \Omega(\sqrt{mp})$.

We have for $\forall j \in \Gamma_+ \cup \Gamma_-, \forall i \in \Lambda_j,$

$$\frac{|b_j^*|}{b} \left\langle w_i^{(1)}, x \right\rangle - \left\langle w_j^*, x \right\rangle \right\rangle \leq \left\| -a_i^j \ell'(0)x \right\| \|G(w_i^{(0)}, b_i)\|_2 \left\| \frac{b_j^*}{b} w_i^{(1)} - \frac{w_j^*}{w_i^{(1)}} \right\|_2 + \frac{w_i^{(1)}}{\| w_i^{(1)} \|_2} - w_j^* \right\|_2 \right\| x \|_2 \right\| \leq (\epsilon_a + \sqrt{2\gamma}) ||x||_2. \right\|$$

Similarly, for $\forall j \in \Gamma, \forall i \in \Lambda_j$,

$$||B_e \left\langle w_i^{(1)}, x \right\rangle - \left\langle w_j^*, x \right\rangle || \leq (\epsilon_a + \sqrt{2\gamma}) ||x||_2. \right\|$$

If $i \in \Lambda_j$, $j \in \Gamma_+ \cup \Gamma_-$, set $\bar{a}_i = a_i^j |\Lambda_j|b$, if $i \in \Lambda_j$, $j \in \Gamma$, set $\bar{a}_i = a_i^j |\Lambda_j|b$, otherwise set $\bar{a}_i = 0$.

we have $||\bar{a}||_0 = O (r(mp)\frac{1}{2})$, $||\bar{a}||_2 = O \left( \frac{B_e b_i}{\max(mp)^2} \right)$, $||\bar{a}||_\infty = O \left( \frac{B_e b_i}{\max(mp)^2} \right)$.

Finally, we have

$$\mathcal{L}_D(f_\bar{a},w^{(1)},b) = \mathcal{L}_D(f_\bar{a},w^{(1)},b) - \mathcal{L}_D(f^*) + \mathcal{L}_D(f^*) \right\} \leq \mathcal{E}_C_{(x,y)} \left\{ \left| f_\bar{a}(x) - f^*(x) \right| \right\} + \mathcal{L}_D(f^*) \right\} \leq \mathcal{E}_C_{(x,y)} \left\{ \sum_{j=1}^{m} \bar{a}_i \sigma \left( \left\langle w_i^{(1)}, x \right\rangle - \bar{b} \right) + \sum_{i=m+1}^{m} \bar{a}_i \sigma \left( \left\langle w_i^{(1)}, x \right\rangle + \bar{b} \right) - \sum_{j=1}^{r} a_i^j \sigma \left( \left\langle w_j^*, x \right\rangle - b_j^* \right) \right\} \right\}$$

$$+ \mathcal{L}_D(f^*) \right\} \leq \mathcal{E}_C_{(x,y)} \left\{ \sum_{j=1}^{m} \bar{a}_i \sigma \left( \left\langle w_i^{(1)}, x \right\rangle - \bar{b} \right) - \sigma \left( \left\langle w_j^*, x \right\rangle - b_j^* \right) \right\} \right\}$$

$$+ \mathcal{E}_C_{(x,y)} \left\{ \sum_{j=1}^{m} \bar{a}_i \sigma \left( \left\langle w_i^{(1)}, x \right\rangle + \bar{b} \right) - \sigma \left( \left\langle w_j^*, x \right\rangle - b_j^* \right) \right\} \right\} + \mathcal{L}_D(f^*) \right\} \leq \mathcal{E}_C_{(x,y)} \left\{ \sum_{j=1}^{m} \bar{a}_i \sigma \left( \left\langle w_i^{(1)}, x \right\rangle - \bar{b} \right) \right\}$$

$$+ \mathcal{E}_C_{(x,y)} \left\{ \sum_{j=1}^{m} \bar{a}_i \sigma \left( \left\langle w_i^{(1)}, x \right\rangle + \bar{b} \right) \right\} \right\} + \mathcal{L}_D(f^*) \right\} \leq \mathcal{E}_C_{(x,y)} \left\{ r||\bar{a}||_\infty (\epsilon_a + \sqrt{2\gamma}) ||x||_2 + ||\Gamma|| ||\bar{a}||_\infty B_e + \mathcal{L}_D(f^*) \right\}$$

$$\leq rB_e b_i (\epsilon_a + \sqrt{2\gamma}) + ||\Gamma|| B_e b_i + \text{OPT}_{d,r,B,F,S,p,\gamma,b_o}.$$

We finish the proof by union bound and $\delta \geq \delta_1 + \delta_2$. \qed
D.2.3 Learning an Accurate Classifier

We will use the following theorem from existing work to prove that gradient descent learns a good classifier (Theorem D.9). Theorem D.1 is simply a direct corollary of Theorem D.9.

Theorem D.5 (Theorem 13 in [33]). Fix some \( \eta \), and let \( f_1, \ldots, f_T \) be some sequence of convex functions. Fix some \( \theta_1 \), and assume we update \( \theta_{t+1} = \theta_t - \eta \nabla f_t(\theta_t) \). Then for every \( \theta^* \) the following holds:

\[
\frac{1}{T} \sum_{t=1}^{T} f_t(\theta_t) \leq \frac{1}{T} \sum_{t=1}^{T} f_t(\theta^*) + \frac{1}{2\eta T} \| \theta^* \|^2 + \| \theta_1 \|^2 + \frac{1}{2 T} \sum_{t=1}^{T} \| \nabla f_t(\theta_t) \|_2 + \eta \frac{1}{T} \sum_{t=1}^{T} \| \nabla f_t(\theta_t) \|_2^2.
\]

To apply the theorem we first present a few lemmas bounding the change in the network during steps.

Lemma D.6 (Bound of \( \Xi^{(t)} \), \( \Xi^{(1)} \)). Assume the same conditions as in Lemma D.4, and let \( d \geq \log m \), with probability at least \( 1 - \frac{1}{m^2} \) over the initialization, \( \| a^{(0)} \|_\infty = O \left( \frac{b}{\sqrt{\log m} \eta B_\sigma B_\epsilon} \right) \), and for all \( i \in [4m] \), we have \( \| w_i^{(0)} \|_2 = O \left( \sigma_w \sqrt{d} \right) \). Finally, \( \| a_i^{(1)} \|_\infty = O \left( -\eta^{(1)} \ell'(0) (B_{x1} \sigma_w \sqrt{d} + \tilde{b}) \right) \), and for all \( i \in [4m] \), \( \| w_i^{(1)} \|_2 = O \left( \frac{b}{\sqrt{\log m} B_{x1} B_\sigma B_\epsilon} \right) \).

Proof of Lemma D.6. By Lemma F.4, we have \( \| a^{(0)} \|_\infty = O \left( \frac{b}{\sqrt{\log m} \eta B_\sigma B_\epsilon} \right) \) with probability at least \( 1 - \frac{1}{m^2} \) by property of maximum i.i.d Gaussians. For any \( i \in [4m] \), by Lemma F.5 and \( d \geq \log m \), we have

\[
\Pr \left( \frac{1}{\sigma_w^2} \left\| w_i^{(0)} \right\|_2^2 \geq d + 2 \sqrt{4d \log(m) + 8 \log(m)} \right) \leq O \left( \frac{1}{m^2} \right).
\]

Thus, by union bound, with probability at least \( 1 - \frac{1}{2m^2} \), for all \( i \in [4m] \), we have \( \| w_i^{(0)} \|_2 = O \left( \sigma_w \sqrt{d} \right) \).

For all \( i \in [4m] \), we have

\[
| a_i^{(1)} | = -\eta^{(1)} \ell'(0) \left[ E_{(x,y)} \left[ y \sigma \left( \left\langle w_i^{(0)}, x \right\rangle - b_i \right) \right] \right],
\]

\[
\leq -\eta^{(1)} \ell'(0) \left( \| w_i^{(0)} \|_2 E_{(x,y)} \| x \|_2 + \tilde{b} \right),
\]

\[
\leq O \left( -\eta^{(1)} \ell'(0) (B_{x1} \sigma_w \sqrt{d} + \tilde{b}) \right).
\]

\[
\left\| w_i^{(1)} \right\|_2 = -\eta^{(1)} \ell'(0) \left\| a_i^{(0)} E_{(x,y)} \left[ y \sigma' \left( \left\langle w_i^{(0)}, x \right\rangle - b_i \right) \cdot x \right] \right\|_2
\leq O \left( \frac{\tilde{b}}{\sqrt{\log m} B_{x1}} \right).
\]

Lemma D.7 (Bound of \( \Xi^{(t)} \)). Assume the same conditions as in Lemma D.6, and let \( \eta = \eta^{(t)} \) for all \( t \in \{2, 3, \ldots, T\} \), \( 0 < T \eta B_{x1} \leq o(1) \), and \( 0 = \lambda = \lambda^{(t)} \) for all \( t \in \{2, 3, \ldots, T\} \), for all \( i \in [4m] \), we have

\[
| a_i^{(t)} | \leq O \left( | a_i^{(1)} | + \left\| w_i^{(1)} \right\|_2 + \frac{\tilde{b}}{B_{x1}} + \eta^{(t)} \right)
\]

\[
\left\| w_i^{(t)} - w_i^{(1)} \right\|_2 \leq O \left( t \eta B_{x1} | a_i^{(1)} | + t \eta^2 B_{x1}^2 \left\| w_i^{(1)} \right\|_2 + t \eta^2 \right.
\]

26
Proof of Lemma D.7. For all $i \in [4n]$, by Lemma D.6,

$$|a_i^{(t)}| = \left| (1 - \eta \lambda) a_i^{(t-1)} - \eta \mathbb{E}_{(x,y)} \left[ \ell' \left( y \mathcal{F}(z; (x, y), \xi) \right) y \left( \langle w_i^{(t-1)}, x \rangle - b_i \right) \right] \right|$$ (70)

$$\leq \left| (1 - \eta \lambda) a_i^{(t-1)} \right| + \eta \mathbb{E}_{(x,y)} \left[ \left| \sigma \left( \langle w_i^{(t-1)}, x \rangle - b_i \right) \right| \right]$$ (71)

$$\leq |a_i^{(t-1)}| + \eta (B_{x1} \|w_i^{(t-1)}\|_2 + \tilde{b})$$ (72)

$$\leq |a_i^{(t-1)}| + \eta B_{x1} \|w_i^{(t-1)} - w_i^{(1)}\|_2 + \eta B_{x1} \|w_i^{(1)}\|_2 \eta \tilde{b}$$ (73)

$$= |a_i^{(t-1)}| + \eta B_{x1} \|w_i^{(t-1)} - w_i^{(1)}\|_2 + \eta Z_i$$ (74)

where we denote $Z_i = B_{x1} \|w_i^{(1)}\|_2 + \tilde{b}$.

Then we give a bound of the first layer’s weight change,

$$\|w_i^{(t)} - w_i^{(1)}\|_2$$ (75)

$$= \left\| (1 - \eta \lambda) w_i^{(t-1)} - \eta a_i^{(t-1)} \mathbb{E}_{(x,y)} \left[ \ell' \left( y \mathcal{F}(z; (x, y), \xi) \right) y \sigma' \left( \langle w_i^{(t-1)}, x \rangle - b_i \right) \right] - w_i^{(1)} \right\|_2$$ (76)

$$\leq \|w_i^{(t-1)} - w_i^{(1)}\|_2 + \eta B_{x1} |a_i^{(t-1)}|.$$ (77)

Combine two bounds, we can get

$$|a_i^{(t)}| \leq |a_i^{(t-1)}| + \eta Z_i + (\eta B_{x1})^2 \sum_{l=1}^{t-2} |a_i^{(l)}|$$ (78)

$$\Leftrightarrow \sum_{l=1}^{t} |a_i^{(l)}| \leq 2 \left( \sum_{l=1}^{t-1} |a_i^{(l)}| \right) - (1 - (\eta B_{x1})^2) \left( \sum_{l=1}^{t-2} |a_i^{(l)}| \right) + \eta Z_i.$$ (79)

Let $h(1) = |a_i^{(1)}|$, $h(2) = 2 |a_i^{(1)}| + \eta Z_i$ and $h(t + 2) = 2 h(t + 1) - (1 - (\eta B_{x1})^2) h(t) + \eta Z_i$ for $n \in \mathbb{N}_+$, by Lemma F.8, we have

$$h(t) = - 1 - \frac{Z_i}{\eta B_{x1}} + c_1 (1 - \eta B_{x1})^{(t-1)} + c_2 (1 + \eta B_{x1})^{(t-1)}$$ (80)

$$c_1 = \frac{1}{2} \left( |a_i^{(1)}| + \frac{Z_i}{\eta B_{x1}} - |a_i^{(1)}| + \frac{\eta Z_i}{\eta B_{x1}} \right)$$ (81)

$$c_2 = \frac{1}{2} \left( |a_i^{(1)}| + \frac{Z_i}{\eta B_{x1}} + |a_i^{(1)}| + \frac{\eta Z_i}{\eta B_{x1}} \right).$$ (82)

Thus, by $|c_1| \leq c_2$, and $0 < T \eta B_{x1} \leq o(1)$, we have

$$|a_i^{(t)}| \leq h(t) - h(t-1)$$ (83)

$$= - \eta B_{x1} c_1 (1 - \eta B_{x1})^{(t-2)} + \eta B_{x1} c_2 (1 + \eta B_{x1})^{(t-2)}$$ (84)

$$\leq 2 \eta B_{x1} c_2 (1 + \eta B_{x1})^{t}$$ (85)

$$\leq O(2 \eta B_{x1} c_2).$$ (86)

Similarly, by binomial approximation, we also have

$$\|w_i^{(t)} - w_i^{(1)}\|_2 \leq \eta B_{x1} h(t - 1)$$ (87)

$$= \eta B_{x1} \left( - \frac{Z_i}{\eta B_{x1}} + c_1 (1 - \eta B_{x1})^{(t-2)} + c_2 (1 + \eta B_{x1})^{(t-2)} \right)$$ (88)

$$\leq \eta B_{x1} O \left( - \frac{Z_i}{\eta B_{x1}} + c_1 (1 - (t - 2) \eta B_{x1}) + c_2 (1 + (t - 2) \eta B_{x1}) \right)$$ (89)

$$\leq \eta B_{x1} O \left( - \frac{Z_i}{\eta B_{x1}} + c_1 + c_2 + (c_2 - c_1) t \eta B_{x1} \right)$$ (90)

$$\leq \eta B_{x1} O \left( |a_i^{(1)}| + |a_i^{(1)}| + \frac{\eta Z_i}{\eta B_{x1}} \right)$$ (91)

$$\leq O \left( (\eta |a_i^{(1)}| + \eta^2 Z_i) t B_{x1} \right).$$ (92)
We finish the proof by plugging $Z_i, c_i$ into the bound.

**Lemma D.8 (Bound of Loss Gap and Gradient).** Assume the same conditions as in Lemma D.7, for all $t \in [T]$, we have

$$|\mathcal{L}_D(f_{\hat{a}, \mathbf{w}}) - \mathcal{L}_D(f_{\mathbf{w}})| \leq B_{z1} \|\hat{a}\|_2 \sqrt{\|\mathbf{w}\|_0 \max_{i \in [4m]} \|w_i^{(t)} - w_i^{(1)}\|_2}$$

(93)

and for all $t \in [T]$, for all $i \in [4m]$, we have

$$\left| \frac{\partial \mathcal{L}_D(f_{\hat{a}, \mathbf{w}})}{\partial a_i^{(t)}} \right| \leq B_{z1} \|w_i^{(t)} - w_i^{(1)}\|_2 + \|w_i^{(1)}\|_2 + \tilde{b}.$$  

(94)

**Proof of Lemma D.8.** It follows from that

$$|\mathcal{L}_D(f_{\hat{a}, \mathbf{w}}) - \mathcal{L}_D(f_{\mathbf{w}})| $$

(95)

\[
\leq E_{(x,y)}|f_{\hat{a}, \mathbf{w}}(x) - f_{\hat{a}, \mathbf{w}}(x)|
\]

(96)

$$\leq E_{(x,y)} \left[ \|\hat{a}\|_2 \sqrt{\|\mathbf{w}\|_0 \max_{i \in [4m]} |\sigma \left( \langle w_i^{(t)}, x \rangle - \langle w_i^{(1)}, x \rangle - b_i \right) |} \right]$$

(97)

$$\leq B_{z1} \|\hat{a}\|_2 \sqrt{\|\mathbf{w}\|_0 \max_{i \in [4m]} \|w_i^{(t)} - w_i^{(1)}\|_2}.$$  

(98)

Also, we have

$$\left| \frac{\partial \mathcal{L}_D(f_{\hat{a}, \mathbf{w}})}{\partial a_i^{(t)}} \right| = \left| E_{(x,y)} \left[ \ell'(y, f_{\hat{a}, \mathbf{w}}(x)) y \left( \sigma \left( \langle w_i^{(t)}, x \rangle - b_i \right) \right) \right] \right|$$

(99)

$$\leq B_{z1} \|w_i^{(t)}\|_2 + \tilde{b}$$

(100)

$$\leq B_{z1} \|w_i^{(t)} - w_i^{(1)}\|_2 + \|w_i^{(1)}\|_2 + \tilde{b}. $$

(101)

We are now ready to prove the main theorem.

**Theorem D.9 (Online Convex Optimization. Full Statement of Theorem D.1).** Consider training by Algorithm 3, and any $\delta \in (0, 1)$. Assume $d \geq \log m$. Set

$$\sigma_w > 0, \quad \tilde{b} > 0, \quad \eta^{(t)} = \eta, \quad \lambda^{(t)} = 0 \text{ for all } t \in \{2, 3, \ldots, T\},$$

$$\eta^{(t)} = \Theta \left( \frac{\min(O(\eta), O(\tilde{b}))}{-\ell'(0) (B_1 \sigma_w \sqrt{d + b})} \right), \quad \lambda^{(t)} = \frac{1}{\eta^{(1)}}, \quad \sigma_a = \Theta \left( \frac{\tilde{b}(mp)^{1/2}}{-\ell'(0) \eta^{(1)} B_1 \sqrt{B_1} \sqrt{B_2}} \right).$$

Let $0 < T \eta B_{x1} \leq o(1)$, $m = \Omega \left( \frac{1}{d^{1/4}} + \frac{1}{p} \left( \log \left( \frac{r}{\delta} \right) \right)^2 \right)$. With probability at least $1 - \delta$ over the initialization, there exists $t \in [T]$ such that

$$\mathcal{L}_D(f_{\hat{a}, \mathbf{w}}) \leq \text{OPT}_{d,r,B_F,p,G} + r B_{a1} \left( \frac{2B_{x1}}{(mp)^{1/2}} \sqrt{\frac{B_2}{B_1}} + B_{x1} \right)^{1/2}$$

(102)

$$+ \eta \left( \sqrt{r} B_{a2} B_{a1} T \eta B_{z1}^2 + \tilde{m} \right) \left( \frac{\sqrt{\log m B_{x1} (mp)^{1/2}}}{\sqrt{B_0 B_1}} + 1 \right) + O \left( \frac{B_{a2}^2 B_{b1}^2}{\eta T B_2 (mp)^{1/2}} \right).$$

Furthermore, for any $\epsilon \in (0, 1)$, set

$$\tilde{b} = \Theta \left( \frac{B_{b1}^2 B_{b2}^2}{\sqrt{r} B_{a1}} \right), \quad m = \Omega \left( \frac{1}{p \epsilon} \left( r B_{a1} B_{x1} \sqrt{\frac{B_2}{B_1}} \right)^4 + \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{r}{\delta} \right) \right)^2 \right),$$

(103)

$$\eta = \Theta \left( \frac{\epsilon}{\sqrt{r} B_{a2} B_{a1} (mp)^{1/2} + \tilde{m} \left( \frac{\sqrt{\log m B_{x1} (mp)^{1/2}}}{\sqrt{B_0 B_1}} + 1 \right)} \right), \quad T = \Theta \left( \frac{1}{\eta B_{x1} (mp)^{1/2}} \right),$$

(104)

we have there exists $t \in [T]$ with

$$\Pr[\text{sign}(f_{\hat{a}, \mathbf{w}}) \neq y] \leq \mathcal{L}_D(f_{\hat{a}, \mathbf{w}}) \leq \text{OPT}_{d,r,B_F,p,G} + r B_{a1} B_{x1} \sqrt{2\gamma} + \epsilon.$$  

(105)
Proof of Theorem D.9. By \( m = \Omega \left( \frac{d}{\sqrt{r}} + \frac{1}{\delta} \left( \log \left( \frac{T}{\delta} \right) \right)^2 \right) \) we have \( 2r e^{-\sqrt{mp}} + \frac{1}{m^2} \leq \delta \). For any \( B \in (0, B_0) \), when \( \sigma_a = \Theta \left( \frac{\delta}{-r'(0)(B_1 \sigma_w \sqrt{d} + \tilde{b})} \right) \), by Theorem D.5, Lemma D.4, Lemma D.8, with probability at least \( 1 - \delta \) over the initialization, we have

\[
\frac{1}{T} \sum_{t=1}^T \mathcal{L}_D \left( f_{\Xi(t)} \right) \leq \frac{1}{T} \sum_{t=1}^T \left( |\mathcal{L}_D (f_{\tilde{a}, w^{(1)}(t), b})| - \mathcal{L}_D (f_{\tilde{a}, w^{(1)}(t), b}) \right) + \mathcal{L}_D (f_{\tilde{a}, w^{(1)}(t), b}) \right) + \mathcal{L}_D (f_{\tilde{a}, w^{(1)}(t), b})
\]

\[
+ \frac{\|\tilde{a}\|_2^2}{2\eta T} + (2 \|a\|_2 \sqrt{m} + 4\eta m) \max_{i \in [4m]} \left| \partial \mathcal{L}_D (f_{\Xi(t)}) / \partial a_i \right| \leq \text{OPT}_{d,r,B,F,S,\gamma,B_G} + rB_0 \left( \frac{B_1^2 B_0}{\sqrt{mp} B_G} + B_1 \sqrt{2\gamma} + \tilde{b} \right)
\]

\[
+ B_1 \left( \|a\|_2 \sqrt{\|\tilde{a}\|_0} \max_{i \in [4m]} \|w_i^{(T)} - w_i^{(1)}\|_2 \right) + \frac{\|\tilde{a}\|_2^2}{2\eta T} + 4mB_1 (\|a\|_\infty + \eta) \left( \max_{i \in [4m]} \|w_i^{(T)} - w_i^{(1)}\|_2 + \max_{i \in [4m]} \|w_i^{(1)}\|_2 + \frac{\tilde{b}}{B_1} \right)
\]

By Lemma D.4, Lemma D.6, Lemma D.7, when \( \eta^{(1)} = \Theta \left( \min \{ O(\eta), O(\eta B_0) \} \right) \), we have

\[
\|\tilde{a}\|_0 = O \left( r(mp)^{\frac{1}{2}} \right), \quad \|\tilde{a}\|_2 = O \left( \frac{B_1^2 B_0}{(b(mp))^{\frac{1}{2}}} \right)
\]

\[
\|a\|_\infty = O \left( -\eta^{(1)} r'(0) (B_1 \sigma_w \sqrt{d} + \tilde{b}) \right)
\]

\[
= \min \{ O(\eta), O(p) \}
\]

\[
\max_{i \in [4m]} \|w_i^{(1)}\|_2 = O \left( \frac{\tilde{b} \log m B_1}{B_G B_0} \right)
\]

\[
\max_{i \in [4m]} \|w_i^{(T)} - w_i^{(1)}\|_2 = O \left( T_\eta B_1 \|a\|_\infty + T_\eta^2 B_1^2 \max_{i \in [4m]} \|w_i^{(1)}\|_2 + T_\eta^2 B_1 \tilde{b} \right)
\]

\[
= O \left( T_\eta^2 B_1 \left( \max_{i \in [4m]} \|w_i^{(1)}\|_2 + \frac{\tilde{b}}{B_1} \right) \right).
\]

Set \( B = \frac{B_1}{(mp)^{\frac{1}{2}}} \sqrt{\frac{B_0}{B_G}} \), we have \( \sigma_a = \Theta \left( \frac{\tilde{b}(mp)^{\frac{1}{2}}}{-r'(0) (B_1 \sigma_w \sqrt{d} + \tilde{b})} \right) \) which satisfy the requirements. Then,

\[
\frac{1}{T} \sum_{t=1}^T \mathcal{L}_D \left( f_{\Xi(t)} \right) \leq \text{OPT}_{d,r,B,F,S,\gamma,B_G} + rB_0 \left( \frac{2B_1}{(mp)^{\frac{1}{2}}} \sqrt{\frac{B_0}{B_G}} + B_1 \sqrt{2\gamma} \right)
\]

\[
+ \left( \sqrt{B_1^2 B_0 \eta^2 B_1^2} + m_\eta B_1 \right) O \left( \frac{\tilde{b} \log m B_1}{B_G B_0} + \frac{\tilde{b}}{B_1} \right) + O \left( \frac{B_0^2 B_1^2}{\eta T b^2 (mp)^{\frac{1}{2}}} \right)
\]

\[
\leq \text{OPT}_{d,r,B,F,S,\gamma,B_G} + rB_0 \left( \frac{2B_1}{(mp)^{\frac{1}{2}}} \sqrt{\frac{B_0}{B_G}} + B_1 \sqrt{2\gamma} \right)
\]

\[
+ \eta \left( \sqrt{B_1^2 B_0 \eta B_1^2} + m_\eta \right) O \left( \frac{\sqrt{\log m B_1^2}}{\sqrt{B_0 B_G}} + 1 \right) + O \left( \frac{B_1^2 B_0^2}{\eta T b^2 (mp)^{\frac{1}{2}}} \right).
\]
We finish the proof as the 0-1 classification error is bounded by the loss function, e.g.,
\[ m = \Omega \left( \frac{1}{\epsilon} \left( rB_{a1}B_{x1} \sqrt{B_b/B_G} \right)^4 + \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{r}{\delta} \right) \right)^2 \right), \] (122)
we have
\[ 1/T \sum_{t=1}^T \mathcal{L}_D (f_{x(t)}^\ast) \leq \text{OPT}_{d,r,B_{F},S_{p,\gamma,B_{G}}} + rB_{a1} \left( \frac{2B_{x1}}{(mp)^{\frac{1}{2}}} \sqrt{\frac{B_{b}}{B_{G}}} + B_{x1} \sqrt{2\gamma} \right) + \epsilon \] (124)
\[ + O \left( \frac{B_{x1}B_{a2}B_{b}^2}{b^2(mp)^{\frac{1}{2}}} \right) \] (125)
\[ \leq \text{OPT}_{d,r,B_{F},S_{p,\gamma,B_{G}}} + rB_{a1}B_{x1} \sqrt{2\gamma} + \epsilon. \] (126)
We finish the proof as the 0-1 classification error is bounded by the loss function, e.g.,
\[ \mathbb{I}[\sign(f(x)) \neq y] \leq \ell'(y(x))/\ell'(|y|), \] where \( \ell(0) = 1 \).

D.3 More Discussion about Setting

Range of \( \sigma_w \). In practice, the value of \( \sigma_w \) cannot be arbitrary, because its choice will have an effect on the Gradient Feature set \( S_{p,\gamma,B_{G}} \). On the other hand, \( d \geq \log m \) is a natural assumption, otherwise, the two-layer neural networks may fall in the NTK regime.

Parameter Choice. We use \( \lambda = 1/\eta \) in the first step so that the neural network will totally forget its initialization, leading to the feature emergence here. This is a common setting for analysis convenience in previous work, e.g., [32, 33, 105]. We can extend this to other choices (e.g., small initialization and large step size for the first few steps), as long as after the gradient update, the gradient dominates the neuron weights. We use \( \lambda = 0 \) afterward as the regularization effect is weak in our analysis. We can extend our analysis to \( \lambda \) being a small value.

Early Stopping. Our analysis divides network learning into two stages: the feature learning stage, and then classifier learning over the good features. The feature learning stage is simplified to one gradient step for the convenience of analysis, while in practice feature learning can happen in multiple steps. The current framework focuses on the gradient features in the early gradient steps, while feature learning can also happen in later steps, in particular for more complicated data. It is an interesting direction to extend the analysis to a longer training horizon.

Role of \( s \). The \( s \) encodes the sign of the bias term, which is important. Recall that we do not update the bias term for simplicity. Let’s consider a simple toy example. Assume we have \[ f_1(x) = a_1\sigma(w_1^T x + 1), f_2(x) = a_2\sigma(w_2^T x - 1) \] and \[ f_3(x) = a_3\sigma(w_3^T x + 2), \] where \( \sigma \) is ReLU activation function which is a homogeneous function.

1. The sign of the bias term is important. We can see that we always have \( a_1\sigma(w_1^T x + 1) \neq a_2\sigma(w_2^T x - 1) \) for any \( a_1, w_1, a_2, w_2 \). This means that \( f_1(x) \) and \( f_2(x) \) are intrinsically different and have different active patterns. Thus, we need to handle the sign of the bias term carefully.

2. The scaling of the bias is absorbed. On the other hand, we can see that \( a_1\sigma(w_1^T x + 1) = a_3\sigma(w_3^T x + 2) \) when \( a_1 = 2a_3, 2w_1 = w_3 \). It means that the scale of the bias term is less important, which can be absorbed into other terms.

Thus, we only need to handle bias with different signs carefully.

Gradient Feature Distribution. We may define a gradient feature distribution rather than a gradient feature set. However, we find that the technical tools used in this continuous setting are pretty different from the discrete version.
**Activation Functions.** We can change the ReLU activation function to a sublinear activation function, e.g., leaky ReLU, sigmoid, to get a similar conclusion. First, we need to introduce a corresponding gradient feature set, and then we can make it by following the same analysis pipeline. For simplicity, we present ReLU only.

**D.4 Gradient Feature Learning Framework under Empirical Risk with Sample Complexity**

In this section, we consider training with empirical risk. Intuitively, the proof is straightforward from the proof for population loss. We can simply replace the population loss with the empirical loss, which will introduce an error term in the gradient analysis. We use concentration inequality to control the error term and show that the error term depends inverse-polynomially on the sample size $n$.

**Definition D.10 (Empirical Simplified Gradient Vector).** Recall $Z = \{ (x^{(i)}, y^{(i)}) \}_{i \in [n]}$, for any $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, an Empirical Simplified Gradient Vector is defined as

$$
\tilde{G}(w, b) := \frac{1}{n} \sum_{i \in [n]} [y^{(i)} x^{(i)} \mathbb{I}[w^T x^{(i)} > b]].
$$

**Definition D.11 (Empirical Gradient Feature).** Recall $Z = \{ (x^{(i)}, y^{(i)}) \}_{i \in [n]}$, let $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ be random variables drawn from some distribution $W, D$. An Empirical Gradient Feature set with parameters $p, \gamma, B_G$ is defined as:

$$
\tilde{S}_{p, \gamma, B_G}(W, D) := \left\{ (D, s) \mid \Pr_{W, b} \left[ \tilde{G}(w, b) \in C_{D, \gamma} \text{ and } \|\tilde{G}(w, b)\|_2 \geq B_G \right] \geq s = \frac{b}{|b|} \right\}.
$$

When clear from context, write it as $\tilde{S}_{p, \gamma, B_G}$.

Considering training by Algorithm 1, we have the following results.

**Theorem 3.12 (Main Result).** Assume Assumption 3.1. For any $\epsilon, \delta \in (0, 1)$, if $m \leq e^d$ and

$$
m = \Omega \left( \frac{1}{e^d} \left( r B_{a_1} B_{x_1} \sqrt{B_b \over B_G} \right)^4 + \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{r}{\delta} \right) \right)^2 \right),
$$

$$
T = \Omega \left( \frac{1}{\epsilon} \left( \sqrt{r} B_{a_2} B_{a_2} B_{x_1} \frac{1}{(mp)^2} + m b \log m \over \sqrt{B_b \over B_G} \over \sqrt{B_b \over B_G} + 1 \over B_{x_1} (mp)^2 \right) \right),
$$

$$
n \log n = \tilde{\Omega} \left( \frac{m^3 p B_{a_2}^4 B_b}{e^{2r} B_{a_2}^2 B_{a_2}^2 B_{x_1} B_{x_1} \over B_{x_1} B_{x_1} + 1 \over p} + \left( \frac{1}{B_G^2} \frac{1}{B_{x_1}^2} \frac{B_{x_2}}{|l^p(0)|^2} \frac{T m}{\delta} \right) \right),
$$

then with initialization (8) and proper hyper-parameter values, we have with probability $\geq 1 - \delta$ over the initialization and training samples, there exists $t \in [T]$ in Algorithm 1 with:

$$\Pr[\text{sign}(f^*(x)) \neq y] \leq \mathcal{L}_D(f^*(x)) \leq \text{OPT}_{d, r, B_F, S_{p, \gamma, B_G}} + r B_{a_1} B_{x_1} \sqrt{2 \gamma + O \left( \frac{\sqrt{B_{x_2} \log n}}{B_G |l^p(0)| n^2} \right) + \epsilon}.$$

See the full statement and proof in Theorem D.17. Below, we show some lemmas used in the analysis under empirical loss.

**Lemma D.12 (Empirical Gradient Concentration Bound).** When $n \log n > B_{y_1}^2$, with probability at least $1 - \Omega \left( 1 \over n \right)$ over training samples, for all $i \in [4m]$, we have

$$
\left\| \frac{\partial \tilde{L}_Z(f^*)}{\partial w_i} - \frac{\partial \mathcal{L}_D(f^*)}{\partial w_i} \right\|_2 \leq O \left( \frac{|a_i| \sqrt{B_{x_2}} \log n}{n^2} \right),
$$

$$
\left\| \frac{\partial \tilde{L}_Z(f^*)}{\partial a_i} - \frac{\partial \mathcal{L}_D(f^*)}{\partial a_i} \right\| \leq O \left( \frac{\|w_i\|_2 \sqrt{B_{x_2}} \log n}{n^2} \right),
$$

$$
\left\| \tilde{L}_Z(f^*) - \mathcal{L}_D(f^*) \right\| \leq O \left( \frac{\|a_i\|_\infty \max_{i \in [4m]} \|w_i\|_2 B_x + b + 1}{n^2} \sqrt{\log n} \right). \tag{130}
$$

31
Proof of Lemma D.12. First, we define,
\[
z^{(l)} = t'(y^{(l)} f_{\Xi}(x^{(l)})) y^{(l)} \left[ \sigma'(\left\langle w_i, x^{(l)} \right\rangle) - b_i \right] x^{(l)} \tag{131}
\]
\[
- \mathbb{E}_{(x,y)} [t'(y f_{\Xi}(x)) y [\sigma'(\left\langle w_i, x \right\rangle) - b_i]] x \tag{132}
\]
As \(|t'(z)| \leq 1, |y| \leq 1, |\sigma'(z)| \leq 1|, we have \(z^{(l)}\) is zero-mean random vector with \(\|z^{(l)}\|_2 \leq 2B_x\) as well as \(\mathbb{E} \left[ \|z^{(l)}\|^2 \right] \leq B_x^2\). Then by Vector Bernstein Inequality, Lemma 18 in [66], for \(0 < z < \frac{B_x}{B_x}\) we have
\[
\Pr \left( \left\| \frac{\partial \tilde{L}_Z(f_{\Xi})}{\partial a_i} - \frac{\partial \tilde{L}_D(f_{\Xi})}{\partial a_i} \right\|_2 \geq z \right) = \Pr \left( \left\| \frac{1}{n} \sum_{i=1}^n z^{(l)} \right\|_2 \geq z \right) \leq \exp \left(-n \cdot \frac{z^2}{8B_x^2} + \frac{1}{4} \right) \tag{133}
\]
Thus, let \(z = n^{-\frac{1}{2}} \sqrt{B_x^2 \log n}\), with probability at least \(1 - O\left(\frac{1}{n}\right)\), we have
\[
\left\| \frac{\partial \tilde{L}_Z(f_{\Xi})}{\partial a_i} - \frac{\partial \tilde{L}_D(f_{\Xi})}{\partial a_i} \right\|_2 \leq O \left(\frac{|a_i| \sqrt{B_x^2 \log n}}{n^2} \right). \tag{135}
\]
On the other hand, by Bernstein Inequality, for \(z > 0\) we have
\[
\Pr \left( \left\| \frac{\partial \tilde{L}_Z(f_{\Xi})}{\partial a_i} - \frac{\partial \tilde{L}_D(f_{\Xi})}{\partial a_i} \right\|_2 > z \|w_i\|_2 \right) \leq 2 \exp \left(-\frac{1}{2} n \frac{z^2}{B_x^2 + \frac{1}{3} B_x^2} \right) \tag{139}
\]
Thus, when \(\frac{n}{\log n} > \frac{n^2}{B_x^2}\), let \(z = n^{-\frac{1}{2}} \sqrt{B_x^2 \log n}\), with probability at least \(1 - O\left(\frac{1}{n}\right)\), we have
\[
\left\| \frac{\partial \tilde{L}_Z(f_{\Xi})}{\partial a_i} - \frac{\partial \tilde{L}_D(f_{\Xi})}{\partial a_i} \right\|_2 \leq O \left(\frac{|w_i| \sqrt{B_x^2 \log n}}{n^2} \right). \tag{140}
\]
Finally, we have
\[
\left| \tilde{L}_Z(f_{\Xi}) - \mathcal{L}_D(f_{\Xi}) \right| \tag{141}
\]
\[
= \frac{1}{n} \sum_{l=1}^n \left( t(y^{(l)} a^T \left[ \sigma(W^T x^{(l)} - b) \right] ) - \mathbb{E}_{(x,y)} \left[ t(y a^T \left[ \sigma(W^T x - b) \right] ) \right] \right) \tag{142}
\]
By Assumption 3.1, we have \(\ell(y^{(l)} a^T \left[ \sigma(W^T x^{(l)} - b) \right] ) - \mathbb{E}_{(x,y)} \left[ \ell(y a^T \left[ \sigma(W^T x - b) \right] ) \right]\) is a zero-mean random variable, with bound \(2\|a\|_0\|a\|_\infty (\max_{i\in[4m]} \|w_i\|_2 B_x + \hat{b}) + 2\). By Hoeffding’s inequality, for all \(z > 0\), we have
\[
\Pr \left( \left| \tilde{L}_Z(f_{\Xi}) - \mathcal{L}_D(f_{\Xi}) \right| \geq z \right) \leq 2 \exp \left(-\frac{z^2 n}{(\|a\|_0\|a\|_\infty (\max_{i\in[4m]} \|w_i\|_2 B_x + \hat{b}) + 1)^2} \right) \tag{131}
\]
Thus, with probability at least \(1 - O\left(\frac{1}{n}\right)\), we have
\[
\left| \tilde{L}_Z(f_{\Xi}) - \mathcal{L}_D(f_{\Xi}) \right| \leq O \left(\frac{\|a\|_0\|a\|_\infty (\max_{i\in[4m]} \|w_i\|_2 B_x + \hat{b}) + 1) \sqrt{\log n}}{n^2} \right) \tag{143}
\]
The gradients allow for obtaining a set of neurons approximating the “ground-truth” network with comparable loss.

**Lemma D.13** (Existence of Good Networks under Empirical Risk). Suppose \( \frac{n}{\log n} > \Omega \left( \frac{B^2}{n^2} + \frac{1}{v} + \frac{B e}{\sqrt{n} \log (n^2)} \right) \). Let \( \lambda^{(1)} = \frac{1}{n} \). For any \( B_e \in (0, B_b) \), let \( \sigma_a = \Theta \left( \frac{\tilde{b}}{-\alpha \log n + \tilde{d}} \right) \) and \( \delta = 2e^{-\sqrt{\frac{\eta}{2}}} \). Then, with probability at least \( 1 - \delta \) over the initialization and training samples, there exists \( \tilde{a}_i \)'s such that \( f_i(\tilde{a}_i, W_i(1), b_i) = \sum_{i=1}^{4m} \tilde{a}_i \sigma \left( \left( w_i^{(1)} \right) - b_i \right) \) satisfies

\[
L_D(f_i(\tilde{a}_i, W_i(1), b_i)) \leq \rho \frac{B_e^2 + B_e}{\sqrt{m^2 B_G B_e}} + B_e^{\frac{1}{2}}(2\gamma + O \left( \sqrt{B_e^2 \log n} \right) + B_e) + \text{OPT}_d(r, B, S, \eta), \tag{145}
\]

and \( \| \tilde{a} \|_2 = O \left( \frac{B_e}{b(m_p)^2} \right), \| \tilde{a} \|_\infty = O \left( \frac{B_e}{b(m_p)^2} \right) \).

**Proof of Lemma D.13.** Denote \( \rho = O \left( \frac{1}{m} \right) \) and \( \beta = O \left( \frac{\sqrt{B_e^2 \log n}}{n^2} \right) \). Note that by symmetric initialization, we have \( \ell'(y_G(x)) = |\ell'(0)| \) for any \( x \in \mathcal{X} \), so that, by Lemma D.12, we have \( \| G(w_i^{(0)}, b_i) - G(w_i^{(0)}, b_i) \|_2 \leq \frac{\beta}{|\ell'(0)|} \) with probability at least \( 1 - \rho \). Thus, by union bound, we can see that \( S_{p, \eta, B_G} \subseteq \tilde{S}_{p, \eta, B_G} \subseteq \text{OPT}_d(r, B, S, \eta) \). Consequently, we have \( \text{OPT}_d(r, B, S, \eta, B_G) \subseteq \text{OPT}_d(r, B, S, \eta, B_G) \). Exactly follow the proof in Lemma D.4 by replacing \( S_{p, \eta, B_G} \) to \( \tilde{S}_{p, \eta, B_G} \). Then, we finish the proof by \( \rho \leq \frac{\beta}{|\ell'(0)|} \leq \left( 1 - \sqrt{2} \right) B_G \).

We will use Theorem D.5 to prove that gradient descent learns a good classifier (Theorem D.17). Theorem D.12 is simply a direct corollary of Theorem D.17. To apply the theorem we first present a few lemmas bounding the change in the network during steps.

**Lemma D.14** (Bound of \( \Xi^{(0)}, \Xi^{(1)} \) under Empirical Risk). Assume the same conditions as in Lemma D.13, and \( d \geq \log m \), with probability at least \( 1 - \frac{1}{m^2} - O \left( \frac{m}{n} \right) \) over the initialization and training samples, \( \| \tilde{a}^{(0)} \|_\infty = O \left( \frac{\sigma_G \sqrt{d}}{n^2} \right) \), and for all \( i \in [4m] \), we have \( \| w_i^{(0)} \|_2 = O \left( \sigma_G \sqrt{d} \right) \). Finally, \( \| a_i^{(1)} \|_\infty = O \left( \eta \left( \frac{1}{\sqrt{d}} \left( \| B_x \|_2 \sigma_G \sqrt{d} + \tilde{b} \right) + \eta \left( \frac{1}{\sqrt{d}} \left( \| B_x \|_2 \sigma_G \sqrt{d} \right) \right) \right) \), and for all \( i \in [4m] \), \( \| w_i^{(1)} \|_2 = O \left( \frac{\sigma_G \sqrt{d}}{n^2} \left( \| B_x \|_2 + \sqrt{B_x \log n} \right) \right) \).

**Proof of Lemma D.14.** The proof exactly follows the proof of Lemma D.6 with Lemma D.12. \( \square \)

**Lemma D.15** (Bound of \( \Xi^{(1)} \) under Empirical Risk). Assume the same conditions as in Lemma D.14, and let \( \eta = \eta^{(1)} \) for all \( t \in \{ 2, 3, \ldots, T \} \), \( 0 < T \eta B_x \leq o(1) \), and \( 0 = \lambda = \lambda^{(1)} \) for all \( t \in \{ 2, 3, \ldots, T \} \). With probability at least \( 1 - O \left( \frac{m}{n} \right) \) over training samples, for all \( i \in [4m] \), for all \( t \in \{ 2, 3, \ldots, T \} \), we have

\[
| a_i^{(1)} | \leq O \left( | a_i^{(1)} | + \| w_i^{(1)} \|_2 + \frac{\tilde{b}}{B_x + \sqrt{B_x \log n} \frac{n}{\sqrt{2}}} + \eta \tilde{b} \right) \tag{146}
\]

\[
\| w_i^{(t)} - w_i^{(1)} \|_2 \leq O \left( t \eta \left( B_x + \frac{\sqrt{B_x \log n}}{n^2} \right) \right) \tag{147}
\]
Proof of Lemma D.15. The proof exactly follows the proof of Lemma D.7 with Lemma D.12. Note that, we have
\begin{equation}
|a_i^{(t)}| \leq |a_i^{(t-1)}| + \eta(B_{x1}\|w_i^{(t-1)}\|_2 + \tilde{b}) + \frac{\eta\|w_i^{(t-1)}\|_2\sqrt{B_{x2}\log n}}{n^{\frac{1}{2}}}
\end{equation}
(148)
\begin{equation}
\leq |a_i^{(t-1)}| + \eta\left(B_{x1} + \frac{\sqrt{B_{x2}\log n}}{n^{\frac{1}{2}}}\right)\|w_i^{(t-1)} - w_i^{(1)}\|_2 + \eta Z_i,
\end{equation}
(149)
where we denote $Z_i = \left(B_{x1} + \frac{\sqrt{B_{x2}\log n}}{n^{\frac{1}{2}}}\right)\|w_i^{(1)}\|_2 + \tilde{b}$. Similarly, we have
\begin{equation}
\|w_i^{(t)} - w_i^{(1)}\|_2 \leq \|w_i^{(t-1)} - w_i^{(1)}\|_2 + \eta \left(B_{x1} + \frac{\sqrt{B_{x2}\log n}}{n^{\frac{1}{2}}}\right)|a_i^{(t-1)}|.
\end{equation}
(150)
We finish the proof by following the same arguments in the proof of Lemma D.7 and union bound. □

Lemma D.16 (Bound of Loss Gap and Gradient under Empirical Risk). Assume the same conditions as in Lemma D.15. With probability at least $1 - O\left(\frac{L}{n}\right)$, for all $t \in [T]$, we have
\begin{equation}
\left|\hat{L}_{\Sigma^{(t)}}(f_{\hat{a},w^{(t)}},b) - \mathcal{L}_D(f_{\hat{a},w^{(t)}},b)\right| \leq O\left(\frac{\|\hat{a}\|_0\|\hat{a}\|_\infty\max_{i\in[4m]}\|w_i^{(t)}\|_2B_x + \tilde{b}}{n^{\frac{1}{2}}} + \frac{1}{\sqrt{\log n}}\right) + B_{x1}\|\hat{a}\|_2\sqrt{\|\hat{a}\|_0}\max_{i\in[4m]}\|w_i^{(t)} - w_i^{(1)}\|_2.
\end{equation}
(151)
We finish the proof by following the same arguments as in the proof of Lemma D.7 and union bound.

Proof of Lemma D.16. By Lemma D.8 and Lemma D.12, with probability at least $1 - O\left(\frac{L}{n}\right)$, for all $t \in [T]$, we have
\begin{equation}
\left|\hat{L}_{\Sigma^{(t)}}(f_{\hat{a},w^{(t)}},b) - \mathcal{L}_D(f_{\hat{a},w^{(t)}},b)\right| \leq \left|\hat{L}_{\Sigma^{(t)}}(f_{\hat{a},w^{(t)}},b) - \mathcal{L}_D(f_{\hat{a},w^{(t)}},b)\right| + \mathcal{L}_D(f_{\hat{a},w^{(t)}},b) - \mathcal{L}_D(f_{\hat{a},w^{(t)}},b)
\end{equation}
\begin{equation}
\leq O\left(\frac{\|\hat{a}\|_0\|\hat{a}\|_\infty\max_{i\in[4m]}\|w_i^{(t)}\|_2B_x + \tilde{b}}{n^{\frac{1}{2}}} + \frac{1}{\sqrt{\log n}}\right) + B_{x1}\|\hat{a}\|_2\sqrt{\|\hat{a}\|_0}\max_{i\in[4m]}\|w_i^{(t)} - w_i^{(1)}\|_2.
\end{equation}
(152)
(153)
(154)
We finish the proof by following the same arguments as in the proof of Lemma D.7 and union bound.

By Lemma D.8 and Lemma D.12, with probability at least $1 - O\left(\frac{L}{n}\right)$, for all $t \in [T]$, we have
\begin{equation}
\left|\|\hat{w}_i^{(t)} - w^{(1)}\|_2 + \frac{B_{x1}\|\hat{a}\|_2\sqrt{\|\hat{a}\|_0}\max_{i\in[4m]}\|w_i^{(t)} - w_i^{(1)}\|_2.}
\end{equation}
\begin{equation}
\end{equation}
(155)
(156)
(157)
(158)
\begin{equation}
\end{equation}
(159)
We finish the proof by following the same arguments as in the proof of Lemma D.7 and union bound.

Theorem D.17 (Online Convex Optimization under Empirical Risk. Full Statement of Theorem 3.12). Consider training by Algorithm 1, and any $\delta \in (0, 1)$. Assume $d \geq \log m$. Set
\begin{equation}
\sigma_w > 0, \quad \tilde{b} > 0, \quad \eta^{(1)} = \eta, \quad \lambda^{(1)} = 0 \text{ for all } t \in \{2, 3, \ldots, T\},
\end{equation}
\begin{equation}
\eta^{(1)} = \Theta\left(\frac{\min\{O(\eta), O(\eta\tilde{b})\}}{-\ell'(0)(B_{x1}\sigma_w\sqrt{d} + \tilde{b})}, \lambda^{(1)} = \frac{1}{\eta^{(1)}}, \quad \sigma_a = \Theta\left(\frac{\tilde{b}(mp)\frac{1}{2}}{-\ell'(0)\eta^{(1)}B_{x1}\sqrt{B_G^2B_0}}\right). \right.
\end{equation}

34
Let 0 < \( T\eta B_{x_1} \leq o(1) \), \( m = \Omega \left( \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{n}{m} \right) \right)^{2} \right) \) and \( \frac{n}{\log n} > \Omega \left( \frac{B_{x_2}^{2}}{B_{x_1}} + \frac{1}{p} + \left( \frac{1}{B_{G}^{2}} + \frac{1}{B_{x_1}^{2}} \right) \frac{T m}{\delta} \right) \). With probability at least 1 - \( \delta \) over the initialization and training samples, there exists \( t \in [T] \) such that

\[
\begin{align*}
\mathcal{L}_D(f_{\Xi(n)}) & \leq \text{OPT}_{d,r,B_F,S_{p,-\gamma,B_G}} + r B_{a_1} \left( 2\sqrt{2B_{x_1}} \sqrt{\frac{B_{b}}{B_{G}}} + B_{x_1} \right) \sqrt{2\gamma + O \left( \frac{\sqrt{B_{x_2} \log n}}{B_G|\ell'(0)|n^{1/2}} \right)} \\
& + \eta \left( \sqrt{T B_{a_2}B_{b}T \eta B_{x_1}^{2} + m} \hat{b} \right) O \left( \frac{\sqrt{\log m B_{x_1}(mp)^{1/4}}}{\sqrt{B_{b}B_{G}}} + \frac{1}{\sqrt{\log m B_{x_1}(mp)}} + 1 \right) + O \left( \frac{B_{a_2}^{2} B_{b}^{2}}{\eta T b^{2}(mp)^{1/4}} \right) \\
& + \frac{\sqrt{\log n}}{n^{1/2}} O \left( \frac{r B_{a_1} B_{b}}{B_{x_1}} + \eta \left( \frac{\sqrt{\log m B_{x_1}(mp)^{1/4}}}{\sqrt{B_{b}B_{G}}} + \frac{1}{\sqrt{\log m B_{x_1}(mp)}} + 1 \right) \right) \\
& + \frac{\sqrt{\log n}}{n^{1/2}} O \left( mn \left( \frac{\sqrt{\log m B_{x_1}(mp)^{1/4}}}{\sqrt{B_{b}B_{G}}} + \frac{1}{\sqrt{\log m B_{x_1}(mp)}} + 1 \right) \right) \tag{165}
\end{align*}
\]

Furthermore, for any \( \epsilon \in (0, 1) \), set

\[
\hat{b} = \Theta \left( \frac{B_{a_2}^{3/4} B_{b}^{3/4}}{\sqrt{r B_{a_1}}} \right), \quad m = \Omega \left( \frac{1}{p^{4/3}} \left( r B_{a_1} B_{x_1} \sqrt{B_{b}B_{G}} \right)^{4} + \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{n}{m} \right) \right)^{2} \right),
\]

\[
\eta = \Theta \left( \frac{\sqrt{T B_{a_2}B_{b}T \eta B_{x_1}^{2} + m} \hat{b}}{(mp)^{1/4}} \right) + \frac{\sqrt{\log m B_{x_1}(mp)^{1/4}}}{\sqrt{B_{b}B_{G}}} + 1 + 1 \right) \right) \), \quad T = \Theta \left( \frac{1}{\eta B_{x_1}(mp)^{1/4}} \right),
\]

\[
\frac{n}{\log n} = \Omega \left( \frac{m^{3} p B_{a_2}^{4} B_{b}^{4}(\log m)^{2}}{e^{2} r^{2} B_{a_1} B_{G}} + \frac{(mp)^{1/4} B_{x_2} \log m}{B_{b} B_{G}} + \frac{B_{x_2}^{2}}{B_{x_2}} + \frac{1}{p} + \frac{1}{B_{G}^{2}} + \frac{1}{B_{x_1}^{2}} \right) \frac{B_{x_2}}{|\ell'(0)|^{2} + \frac{T m}{\delta}} \),
\]

we have there exists \( t \in [T] \) with

\[
\Pr[\text{sign}(f_{\Xi(n)})(x) \neq y] \leq \mathcal{L}_D(f_{\Xi(n)}) \tag{166}
\]

\[
\leq \text{OPT}_{d,r,B_F,S_{p,-\gamma,B_G}} + r B_{a_1} B_{x_1} \sqrt{2\gamma + O \left( \frac{\sqrt{B_{x_2} \log n}}{B_G|\ell'(0)|n^{1/2}} \right)} + \epsilon. \tag{167}
\]

**Proof of Theorem D.17.** We follow the proof in Theorem D.9. By \( m = \Omega \left( \frac{1}{\sqrt{\delta}} + \frac{1}{p} \left( \log \left( \frac{n}{m} \right) \right)^{2} \right) \)
and \( \frac{n}{\log n} > \Omega \left( \frac{B_{x_2}^{2}}{B_{x_1}} + \frac{1}{p} + \left( \frac{1}{B_{G}^{2}} + \frac{1}{B_{x_1}^{2}} \right) \frac{T m}{\delta} \right) \), we have \( 2re^{-\sqrt{mp}} + \frac{1}{m^{2}} + O \left( \frac{T m}{n^{1/2}} \right) \leq \delta \). For any \( B_{e} \in (0, B_{b}) \), when \( \sigma_{a} = \Theta \left( \frac{\hat{b}}{(\ell'(0))^{1/4} B_{b} B_{G}} \right) \), by Theorem D.5, Lemma D.12, Lemma D.13,
Lemma D.16, with probability at least $1 - \delta$ over the initialization and training samples, we have

\begin{align}
\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_D(f_{\Xi(t)}) & \leq \frac{1}{T} \sum_{t=1}^{T} |\mathcal{L}_D(f_{\Xi(t)}) - \mathcal{L}_f(f_{\Xi(t)})| + \frac{1}{T} \sum_{t=1}^{T} |\mathcal{L}_f(f_{\Xi(t)})| \\
& \leq \frac{1}{T} \sum_{t=1}^{T} |\mathcal{L}_D(f_{\Xi(t)}) - \mathcal{L}_f(f_{\Xi(t)})| + \frac{1}{T} \sum_{t=1}^{T} |\mathcal{L}_f(f_{\Xi(t)}) - \mathcal{L}_D(f(\bar{a},w^{(1)}))| + \mathcal{L}_D(f(\bar{a},w^{(1)})) \\
& \quad + \mathcal{L}_D(f(\bar{a},w^{(1)})) + \frac{\|\bar{a}\|^2}{2\eta T} + \left(2\|a^{(1)}\|_2 \sqrt{m} + 4\eta m \max_{t \in [T], i \in [4m]} \frac{\partial \mathcal{L}_f(f_{\Xi(t)})}{\partial a^{(1)}} \right) \\
& \leq B_{x1} \|\bar{a}\|_2 \sqrt{\|\bar{a}\|_2 \max_{i \in [4m]} \|w_i^{(T)} - w_i^{(1)}\|_2} \\
& \quad + O \left(\frac{\|\bar{a}\|_0 \|\bar{a}\|_\infty + m \|a^{(T)}\|_\infty (\max_{i \in [4m]} \|w_i^{(T)}\|_2 B_x + \tilde{b}) + 2 \sqrt{\log n}}{n^{\frac{1}{2}}}\right) \\
& \quad + \text{OPT}_{d,r,B_x,S_p,S_w,B_G} + r B_{a1} \left(\frac{2B_x^2 B_b}{\sqrt{m} B_G B_x} + B_{x1} \sqrt{2\gamma + O \left(\frac{\sqrt{B_{x2} \log n}}{B_G |\hat{r}'(0)| n^{\frac{1}{2}}}\right)} + B_r \right) \\
& \quad + \frac{\|\hat{a}\|^2}{2\eta T} + 4m B_{x1} (\|a^{(1)}\|_\infty + \eta) \\
& \quad \cdot \left(\max_{i \in [4m]} \|w_i^{(T)} - w_i^{(1)}\|_2 + \max_{i \in [4m]} \|w_i^{(1)}\|_2 + \tilde{b} \frac{B_{x1}}{B_{x1}} + O \left(\frac{\max_{i \in [4m]} \|w_i^{(T)}\|_2 \sqrt{B_{x2} \log n}}{B_{x1} n^{\frac{1}{2}}}\right)\right).
\end{align}

Set $B_x = \frac{B_{x1}}{(mp)^{\frac{1}{4}}} \sqrt{\frac{2B_b}{B_G}}$, we have $\sigma_a = \Theta \left(\frac{\tilde{b}(mp)^{\frac{1}{2}}}{-\hat{r}'(0) B_{x1} \sqrt{B_G B_b}}\right)$ which satisfy the requirements. By Lemma D.13, Lemma D.14, Lemma D.15, $\frac{n}{\log n} > \Omega \left(\frac{B_{x2}}{B_{x1} |\hat{r}'(0)|^2}\right)$, when $\eta^{(1)} = \Theta \left(\min \{O(\eta), O(\eta \tilde{b})\}\right)$, we have

\begin{align}
\|\bar{a}\|_0 = & O \left(r \sqrt{mp}\right), \quad \|\bar{a}\|_2 = O \left(\frac{B_{a2} B_b}{\tilde{b}(mp)^{\frac{1}{2}}}\right), \quad \|\bar{a}\|_\infty = O \left(\frac{B_{a1} B_b}{\tilde{b}(mp)^{\frac{1}{2}}}\right) \\
\|a^{(1)}\|_\infty = & O \left(\eta |\hat{r}'(0)| (B_{x1} \sigma_w \sqrt{\tilde{d} + \tilde{b}} + \eta \frac{\sigma_w \sqrt{B_{x2} \log n}}{n^{\frac{1}{2}}} \right) \\
& \quad \quad \quad \quad = \min \{O(\eta), O(\eta \tilde{b})\} \\
\|a^{(T)}\|_\infty \leq & O \left(\|a^{(1)}\|_\infty + \max_{i \in [4m]} \|w_i^{(1)}\|_2 + \frac{\tilde{b}}{B_{x1} + \sqrt{\frac{B_{x2} \log n}{n^{\frac{1}{2}}}}} + \eta \tilde{b}\right) \\
& \quad \quad \quad \quad \leq O \left(\max_{i \in [4m]} \|w_i^{(1)}\|_2 + \frac{\tilde{b}}{B_{x1}}\right).
\end{align}
\[
\max_{i \in [4m]} \|w_i^{(T)} - w_i^{(1)}\|_2 = O \left( T\eta \left( B_{z_1} + \sqrt{B_{z_2} \log n} \right) \|a_i^{(1)}\| \right)
\]

\[
+ T\eta^2 \left( B_{z_1} + \sqrt{B_{z_2} \log n} \right)^2 \|w_i^{(1)}\|_2
\]

\[
+ T\eta^2 \left( B_{z_1} + \sqrt{B_{z_2} \log n} \right) \hat{b}
\]

\[
= O \left( T\eta^2 B_{z_1}^2 \left( \max_{i \in [4m]} \|w_i^{(1)}\|_2 + \frac{\hat{b}}{B_{z_1}} \right) \right).
\]

Then, following the proof in Theorem D.9, we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_D (J_{Z_t(t)})
\]

\[
\leq \text{OPT}_{d,r,B,F,r_p,\gamma,B_{G}} + r B_{a_1} \left( 2\sqrt{2} B_{z_1} \sqrt{\frac{B_b}{B_G}} + B_{z_1} \sqrt{2\gamma + O \left( \sqrt{B_{z_2} \log n} \right) B_{z_1} \left( \sqrt{\log m B_{z_1} (mp) \frac{1}{2}} + 1 \right) + O \left( \frac{B_{z_2}^2 B_{z_1}^2}{\eta T b^2 (mp) \frac{1}{2}} \right)} \right)
\]

\[
+ \eta \left( \sqrt{T} B_{z_2} B_b T\eta B_{z_1} + m \hat{b} \right) O \left( \frac{\sqrt{\log m B_{z_1} (mp) \frac{1}{2}}}{\sqrt{B_b B_G}} + 1 \right) + O \left( \frac{B_{z_2}^2 B_{z_1}^2}{\eta T b^2 (mp) \frac{1}{2}} \right)
\]

\[
+ O \left( \frac{m \eta \max_{i \in [4m]} \|w_i^{(T)}\|_2 \sqrt{2} \sqrt{B_{z_2} \log n}}{n^\frac{1}{2}} \right)
\]

\[
\leq \text{OPT}_{d,r,B,F,r_p,\gamma,B_{G}} + r B_{a_1} \left( 2\sqrt{2} B_{z_1} \sqrt{\frac{B_b}{B_G}} + B_{z_1} \sqrt{2\gamma + O \left( \sqrt{B_{z_2} \log n} \right) B_{z_1} \left( \sqrt{\log m B_{z_1} (mp) \frac{1}{2}} + 1 \right) + O \left( \frac{B_{z_2}^2 B_{z_1}^2}{\eta T b^2 (mp) \frac{1}{2}} \right)} \right)
\]

\[
+ \eta \left( \sqrt{T} B_{z_2} B_b T\eta B_{z_1} + m \hat{b} \right) O \left( \frac{\sqrt{\log m B_{z_1} (mp) \frac{1}{2}}}{\sqrt{B_b B_G}} + 1 \right) + O \left( \frac{B_{z_2}^2 B_{z_1}^2}{\eta T b^2 (mp) \frac{1}{2}} \right)
\]

\[
+ \sqrt{\log n} \left( \frac{r B_{a_1} B_b}{b} + m \left( \frac{b \sqrt{\log m (mp) \frac{1}{2}}}{\sqrt{B_b B_G}} + \frac{\hat{b}}{B_{z_1}} \right) \right)
\]

\[
\cdot \left( \frac{b \sqrt{\log m (mp) \frac{1}{2}}}{\sqrt{B_b B_G}} + T\eta B_{z_1} \hat{b} \right) B_x + \hat{b}
\]

\[
+ \sqrt{\log n} \left( m \eta \left( \frac{b \sqrt{\log m (mp) \frac{1}{2}}}{\sqrt{B_b B_G}} + T\eta B_{z_1} \hat{b} \right) \sqrt{B_{z_2}} \right).
\]
Furthermore, for any \( \epsilon \in (0, 1) \), set
\[
\hat{b} = \Theta \left( \frac{B_G^2 B_{a2} B_b^3}{\sqrt{r B_{a1}}} \right), \quad m = \Omega \left( \frac{1}{p \epsilon^4} \left( r B_{a1} B_x \sqrt{B_b \sqrt{B_G}} \right)^4 + \frac{1}{\sqrt{\epsilon}} + \frac{1}{p} \left( \log \left( \frac{r}{\lambda} \right) \right)^2 \right),
\]
\[
\eta = \Theta \left( \frac{\epsilon}{(\eta B_{a1} (mp)^{\frac{1}{4}})} \right), \quad T = \Theta \left( \frac{1}{\eta B_{x1} (mp)^{\frac{1}{4}}} \right),
\]
\[
\frac{n}{\log n} = \Omega \left( \frac{1}{\epsilon^2 \eta^2 B_{a1} B_G} \log (m) \right) + \frac{1}{(mp)^{\frac{1}{4}} B_{a1} B_G} + \frac{1}{B_{x2}^2} + \frac{1}{p} + \frac{1}{B_G} \right) \left( B_{x2} \| \ell'(0) \|_2^2 + \frac{T m}{\delta} \right),
\]
and note that \( B_G \leq B_{x1} \leq B_x \) and \( \sqrt{B_{x2}} \leq B_x \) naturally, we have
\[
\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_D (f_{\hat{z}(\cdot)}) \leq \text{OPT}_{\delta, r, B_P, S_{p, \gamma, B_G}} + r B_{a1} \left( \frac{2 \sqrt{B_{x1} (mp)^{\frac{1}{4}}} \sqrt{B_G} + B_{x1} \sqrt{2 \gamma + O \left( \frac{\sqrt{B_{x2} \log n}}{B_G} | \ell'(0) | n^2 \right)} \right)
\]
\[
+ \frac{\epsilon}{2} + O \left( \frac{B_{x2} \sqrt{B_{a1} B_G}^2}{B_{a2} \sqrt{B_{a2} B_G}^2} \right) + \frac{\sqrt{\log n}}{n^2} O \left( m B_{a1} B_{x2} \sqrt{B_{a1} B_G} \sqrt{\ell'(0)} \| n^2 \right)
\]
\[
+ \frac{\sqrt{\log n}}{n^2} O \left( \frac{\epsilon \sqrt{B_{x2} \log (mp)^{\frac{1}{4}}}}{\sqrt{B_{a1} B_G}} \right)
\]
\[
\leq \text{OPT}_{\delta, r, B_P, S_{p, \gamma, B_G}} + r B_{a1} B_{x1} \left( 2 \gamma + O \left( \frac{\sqrt{B_{x2} \log n}}{B_G} | \ell'(0) | n^2 \right) \right) + \epsilon.
\]

We finish the proof as the 0-1 classification error is bounded by the loss function, e.g.,
\[
\mathbb{I}[\text{sign}(f(x)) \neq y] \leq \frac{\ell(y(f(x)))}{\ell(0)}, \text{ where } \ell(0) = 1.
\]

\[\square\]

### E Applications in Special Cases

We present the case study of linear data in Appendix E.1, mixtures of Gaussians in Appendix E.2 and Appendix E.3, parity functions in Appendix E.4, Appendix E.5 and Appendix E.6, and multiple-index models in Appendix E.7.

In special case applications, we consider binary classification with hinge loss, e.g., \( \ell(z) = \max\{1 - z, 0\} \). Let \( \mathcal{X} = \mathbb{R}^d \) be the input space, and \( \mathcal{Y} = \{\pm 1\} \) be the label space.

**Remark E.1 (Hinge Loss and Logistic Loss).** Both hinge loss and logistic loss can be used in special cases and general cases. For convenience, we use hinge loss in special cases, where we can directly get the ground-truth NN close form of the optimal solution which has zero loss. For logistic loss, there is no zero-loss solution. We can still show that the OPT value has an exponentially small upper bound at the cost of more computation.

#### E.1 Linear Data

**Data Distributions.** Suppose two labels are equiprobable, i.e., \( E[y = -1] = E[y = +1] = \frac{1}{2} \).

The input data are linearly separable and there is a ground truth direction \( w^* \), where \( \|w^*\|_2 = 1 \), such that \( y \langle w^*, x \rangle > 0 \). We also assume \( E[y P_{w^*} x] = 0 \), where \( P_{w^*} \) is the projection operator on the complementary space of the ground truth, i.e., the components of input data being orthogonal with the ground truth are independent of the label \( y \). We define the input data signal level as \( \rho := E[y \langle w^*, x \rangle] > 0 \) and the margin as \( \beta := \min_{x,y} y \langle w^*, x \rangle > 0 \).

We call this data distribution \( D_{\text{linear}} \).
Lemma E.2 (Linear Data: Gradient Feature Set). Let \( \tilde{b} = d^r B_{x1} \sigma_w \), where \( \tau \) is any number large enough to satisfy \( d^{r/2-\frac{1}{4}} > \Omega \left( \frac{\sqrt{B_{x2}}}{\rho} \right) \). For \( \mathcal{D}_{\text{linear}} \) setting, we have \( (w^*, -1) \in \mathcal{S}_{p, \gamma, B_G} \) where

\[
p = \frac{1}{2}, \quad \gamma = \Theta \left( \frac{\sqrt{B_{x2}}}{\rho d^{r/2-\frac{1}{4}}} \right), \quad B_G = \rho - \Theta \left( \frac{\sqrt{B_{x2}}}{d^{r/2-\frac{1}{4}}} \right).
\]  

(203)

Proof of Lemma E.2. By data distribution, we have

\[
E_{(x,y)}[gy] = \rho w^*.
\]  

(204)

Define \( S_{\text{Sure}} : \{ i \in [m] : \| w_i^{(0)} \|_2 \leq 2\sqrt{d} \sigma_w \} \). For all \( i \in [m] \), we have

\[
Pr[i \in S_{\text{Sure}}] = Pr[\| w_i^{(0)} \|_2 \leq 2\sqrt{d} \sigma_w] \geq \frac{1}{2}.
\]  

(205)

For all \( i \in S_{\text{Sure}} \), by Markov’s inequality and considering neuron \( i + m \), we have

\[
Pr \left[ \left\langle w_{i+m}^{(0)}, x \right\rangle - b_{i+m} < 0 \right] = Pr \left[ \left\langle w_i^{(0)}, x \right\rangle + b_i < 0 \right] \leq Pr \left[ \| w_i^{(0)} \|_2 \| x \|_2 \geq b_i \right] \leq Pr \left[ \| x \|_2 \geq \frac{d^{r/2} B_{x1}}{2} \right] \leq \Theta \left( \frac{1}{d^{r/2-\frac{1}{4}}} \right).
\]  

(209)

For all \( i \in S_{\text{Sure}} \), by Hölder’s inequality, we have

\[
\left\| E_{(x,y)} \left[ g \left( 1 - \sigma' \left[ \left\langle w_i^{(0)}, x \right\rangle + b_i \right] \right) \right] x \right\|_2 \leq \sqrt{E[\| x \|_2^2] E \left[ (1 - \sigma' \left[ \left\langle w_i^{(0)}, x \right\rangle + b_i \right] )^2 \right]} \leq \Theta \left( \frac{\sqrt{B_{x2}}}{d^{r/2-\frac{1}{4}}} \right).
\]  

(213)

We have

\[
1 - \frac{\| G(w_{i+m}^{(0)}, b_{i+m}), w^* \|_2}{\| G(w_i^{(0)}, b_i) \|_2} = 1 - \frac{\| G(w_i^{(0)}, -b_i), w^* \|_2}{\| G(w_i^{(0)}, -b_i) \|_2} \leq 1 - \frac{\rho - \Theta \left( \frac{\sqrt{B_{x2}}}{d^{r/2-\frac{1}{4}}} \right)}{\rho + \Theta \left( \frac{\sqrt{B_{x2}}}{d^{r/2-\frac{1}{4}}} \right)} = \Theta \left( \frac{\sqrt{B_{x2}}}{\rho d^{r/2-\frac{1}{4}}} \right) = \gamma.
\]  

(216)

We finish the proof by \( \frac{b_{i+m}}{b_i} = -1 \). \( \square \)

Lemma E.3 (Linear Data: Existence of Good Networks). Assume the same conditions as in Lemma E.2. Define

\[
f^*(x) = \frac{1}{\beta} \sigma((w^*, x)) - \frac{1}{\beta} \sigma((-w^*, x)).
\]  

(217)

For \( \mathcal{D}_{\text{linear}} \) setting, we have \( f^* \in \mathcal{F}_{d, r, B_F, S_{p, \gamma, B_G}} \), where \( r = 2, B_F = (B_{a1}, B_{a2}, B_b) = \left( \frac{1}{\beta}, \frac{1}{\beta}, \frac{1}{B_{x1}} \right), p = \frac{1}{2}, \gamma = \Theta \left( \frac{\sqrt{B_{x2}}}{\rho d^{r/2-\frac{1}{4}}} \right), B_G = \rho - \Theta \left( \frac{\sqrt{B_{x2}}}{d^{r/2-\frac{1}{4}}} \right) \). We also have \( \text{OPT}_{d, r, B_F, S_{p, \gamma, B_G}} = 0 \).
Proof of Lemma E.3. By Lemma E.2 and Lemma F.3, we have $f^* \in \mathcal{F}_{d,r,B_F,S_p,\gamma,B_G}$. We also have
\[ \text{OPT}_{d,r,B_F,S_p,\gamma,B_G} \leq \mathcal{L}_{\text{linear}}(f^*) \leq \mathcal{L}_{\text{linear}}(f^*) = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{linear}}} \mathcal{L}(x,y)(f^*) = 0. \]

\[ \square \]

Theorem E.4 (Linear Data: Main Result). For $\mathcal{D}_{\text{linear}}$ setting, for any $\delta \in (0,1)$ and for any $\epsilon \in (0,1)$ when
\[ m = \text{poly} \left( \frac{1}{\delta \epsilon \beta \rho} \right) \leq \epsilon^d, \quad T = \text{poly} \left( (m, B_{x_1}) \right), \quad n = \text{poly} \left( (m, B_x, \frac{1}{\delta \epsilon \beta \rho}) \right), \]
trained by Algorithm 1 with hinge loss, with probability at least $1 - \delta$ over the initialization, with proper hyper-parameters, there exists $t \in [T]$ such that
\[ \Pr[\text{sign}(f_{\Xi(t)}(x)) \neq y] \leq \epsilon. \]

\[ \square \]

Proof of Theorem E.4. Let $\hat{b} = d^r B_{x_1} \sigma_w$, where $\tau$ is a number large enough to satisfy $d^{r/2} \geq \Omega \left( \frac{\sqrt{m}}{\rho} \right)$ and $O \left( \frac{B_{x_2} \sqrt{\tau}}{\beta \sqrt{m}} \right) \leq \frac{\epsilon}{3}$. By Lemma E.3, we have $f^* \in \mathcal{F}_{d,r,B_F,S_p,\gamma,B_G}$, where
\[ r = 2, B_F = (B_{x_1}, B_{x_2}, B_y) = \left( \frac{1}{\beta \sqrt{r}}, \frac{1}{B_{x_1}} \right), p = \frac{1}{2}, \gamma = \Theta \left( \frac{\sqrt{\tau} x_2}{\rho \sqrt{m}} \right), B_G = \rho - \Theta \left( \frac{\sqrt{\tau} x_2}{d^{r/2} \beta} \right). \]
We also have $\text{OPT}_{d,r,B_F,S_p,\gamma,B_G} = 0$.

Adjust $\sigma_w$ such that $\hat{b} = d^r B_{x_1} \sigma_w = \Theta \left( \frac{B_{x_2} \sqrt{\tau}}{\sqrt{r} B_{x_1}} \right)$. Injecting above parameters into Theorem 3.12, we have with probability at least $1 - \delta$ over the initialization, with proper hyper-parameters, there exists $t \in [T]$ such that
\[ \Pr[\text{sign}(f_{\Xi(t)}(x)) \neq y] \leq O \left( \frac{B_{x_2} \sqrt{\tau}}{\beta \sqrt{m}} \right) + O \left( \frac{B_{x_2} \sqrt{\tau} \log n}{\beta \sqrt{m} \frac{\epsilon}{3}} \right) + \epsilon/2 \leq \epsilon. \]

\[ \square \]

E.2 Mixture of Gaussians

We recap the problem setup in Section 4.1 for readers’ convenience.

E.2.1 Problem Setup

Data Distributions. We follow the notations from [99]. The data are from a mixture of $r$ high-dimensional Gaussians, and each Gaussian is assigned to one of two possible labels in $\mathcal{Y} = \{ \pm 1 \}$. Let $S(y) \subseteq [r]$ denote the set of indices of the Gaussians associated with the label $y$. The data distribution is then:
\[ q(x, y) = q(y) q(x | y), \quad q(x | y) = \sum_{j \in S(y)} p_j \mathcal{N}_j(x), \]
where $\mathcal{N}_j(x)$ is a multivariate normal distribution with mean $\mu_j$ and covariance $\Sigma_j$, and $p_j$ are chosen such that $q(x, y)$ is correctly normalized.

We call this data distribution $\mathcal{D}_{\text{mixture}}$.

We will make some assumptions about the Gaussians, for which we first introduce some notations. For all $j \in [r]$, let $y_{(j)} \in \{ +1, -1 \}$ be the label for $\mathcal{N}_j(x)$.
\[ D_j := \frac{\mu_j}{\| \mu_j \|_2}, \quad \bar{\mu}_j := \mu_j / \sqrt{d}, \quad B_{\mu_1} := \min_{j \in [r]} \| \bar{\mu}_j \|_2, \quad B_{\mu_2} := \max_{j \in [r]} \| \bar{\mu}_j \|_2, \quad p_B := \min_{j \in [r]} p_j. \]

Assumption E.5 (Mixture of Gaussians. Recap of Assumption 4.1). Let $8 \leq \tau \leq d$ be a parameter that will control our final error guarantee. Assume
• Equiprobable labels: \( q(-1) = q(+1) = 1/2 \).
• For all \( j \in [r] \), \( \Sigma_j = \sigma_{j \cdot d_{x \cdot d}} \). Let \( \sigma_B := \max_{j \in [r]} \sigma_j \) and \( \sigma_{B^+} := \max \{ \sigma_B, B_{\mu_2} \} \).

\[ r \le 2d, \quad p_B \ge \frac{1}{2^d}, \quad \Omega \left( \frac{1}{2} + \sqrt{\frac{\sigma_{B^+}^2 \log d}{d}} \right) \le B_{\mu_1} \le B_{\mu_2} \le d. \]

• The Gaussians are well-separated: for all \( i \neq j \in [r] \), we have \(-1 \le \langle D_i, D_j \rangle \le 0 \), where

\[ 0 \le \theta \le \min \left\{ \frac{1}{2\tau}, \frac{\sigma_{B^+}}{B_{\mu_2}} \sqrt{\frac{\tau \log d}{d}} \right\}. \]

Below, we define a sufficient condition that randomly initialized weights will fall in nice gradients set after the first gradient step update.

**Definition E.6** (Mixture of Gaussians: Subset of Nice Gradients Set). Recall \( w_i^{(0)} \) is the weight for the \( i \)-th neuron at initialization. For all \( j \in [r] \), let \( S_{D_j, \text{Sure}} \subseteq [m] \) be those neurons that satisfy

\[
\begin{align*}
\langle w_i^{(0)}, \mu_j \rangle & \ge C_{\text{Sure}, 1} b_i, \\
\langle w_i^{(0)}, \mu_{j'} \rangle & \le C_{\text{Sure}, 2} b_i, \text{ for all } j' \neq j, j' \in [r], \\
\| w_i^{(0)} \|_2 & \le \Theta(\sqrt{\sigma_w}).
\end{align*}
\]

### E.2.2 Mixture of Gaussians: Feature Learning

We show the important Lemma E.7 first and defer other Lemmas after it.

**Lemma E.7** (Mixture of Gaussians: Gradient Feature Set. Part statement of Lemma 4.3). Let \( C_{\text{Sure}, 1} = \frac{3}{2} \), \( C_{\text{Sure}, 2} = \frac{1}{2} \), \( \tilde{b} = C_b \sqrt{d \log d \sigma_w \sigma_{B^+}} \), where \( C_b \) is a large enough universal constant. For \( D_{\text{mixture}} \) setting, we have \( (D_{j}, +1) \in S_{p, \gamma, B_{G}} \) for all \( j \in [r] \), where

\[
\begin{align*}
p & = \Theta \left( \frac{B_{\mu_1}}{\sqrt{\gamma \log d \sigma_{B^+}} \cdot d^{\frac{9}{2} C_{\text{Sure}, 2}^2 \sigma_{B^+}^2 / (2 B_{\mu_1}^2)}} \right), \\
\gamma & = \frac{1}{d^{0.9 \tau - 1.5}}, \\
B_{G} & = p B_{\mu_1} \sqrt{d} - O \left( \frac{d \sigma_{B^+}}{d^{0.9 \tau}} \right).
\end{align*}
\]

**Proof of Lemma E.7.** For all \( j \in [r] \), by Lemma E.10, for all \( i \in S_{D_j, \text{Sure}} \),

\[
\begin{align*}
1 & - \frac{\left| \langle G(w_i^{(0)}, b_i), D_j \rangle \right|}{\| G(w_i^{(0)}, b_i) \|_2} \\
\le & 1 - \frac{\left| \langle G(w_i^{(0)}, b_i), D_j \rangle \right|}{\sqrt{\left| \langle G(w_i^{(0)}, b_i), D_j \rangle \right|^2 + \max_{D_j^+, \|D_j^+\|_2 = 1} \left| \langle G(w_i^{(0)}, b_i), D_j^+ \rangle \right|^2}} \\
\le & 1 - \frac{\left| \langle G(w_i^{(0)}, b_i), D_j \rangle \right|}{1 + \frac{B_{\mu_1} O \left( \frac{1}{\sqrt{\gamma \log d}} \right) + \sigma_{B^+} O \left( \frac{1}{d \sigma_{B^+}} \right)}{p_j B_{\mu_1} \sqrt{d} \left( 1 - O \left( \frac{1}{d \sigma_{B^+}} \right) \right) - B_{\mu_2} O \left( \frac{1}{d \sigma_{B^+}} \right)}} \\
\le & \frac{\sigma_{B^+} O \left( \frac{1}{d \sigma_{B^+}} \right)}{p_j B_{\mu_1} \sqrt{d} - \sigma_{B^+} O \left( \frac{1}{d \sigma_{B^+}} \right)} \\
\le & \frac{1}{d^{0.9 \tau - 1.5}} = \gamma,
\end{align*}
\]

where the last inequality follows \( B_{\mu_1} \ge \Omega \left( \sigma_{B^+} \sqrt{\frac{\tau \log d}{d}} \right) \).
Thus, we have $G(w^{(0)}_i, b_i) \in C_{D_j,\gamma}$ and $\left| \langle G(w^{(0)}_i, b_i), D_j \rangle \right| \leq \|G(w^{(0)}_i, b_i)\|_2 \leq B_{21}$, $\frac{b_i}{|b_i|} = +1$. Thus, by Lemma E.8, we have

$$\Pr_{w,b} \left[ G(w, b) \in C_{D_j,\gamma} \text{ and } \|G(w, b)\|_2 \geq B_G \text{ and } \frac{b}{|b|} = +1 \right] \geq \Pr \left[ i \in S_{D_j,\text{Sure}} \right] \geq p. \tag{233}$$

Thus, $(D_j, +1) \in S_{p,\gamma,B_G}$. We finish the proof. \hfill \Box

Below are Lemmas used in the proof of Lemma E.7. In Lemma E.8, we calculate $p$ used in $S_{p,\gamma,B_G}$.

**Lemma E.8** (Mixture of Gaussians: Geometry at Initialization. Lemma B.2 in [7]). Assume the same conditions as in Lemma E.7, recall for all $i \in [m]$, $w^{(0)}_i \sim \mathcal{N}(0, \sigma_o^2 I_{d \times d})$, over the random initialization, we have for all $i \in [m], j \in [r],$

$$\Pr \left[ i \in S_{D_j,\text{Sure}} \right] \geq \Theta \left( \frac{B_{\mu_1}}{\sqrt{r \log d \sigma_{B^+} \cdot d (9G^2 \tau \sigma_{B^+}^2 / (2B_{\mu_1}^2))} \right). \tag{236}$$

**Proof of Lemma E.8.** Recall for all $l \in [r], \hat{\mu}_l = \mu_l / \sqrt{d}$.

WLOG, let $j = r$. For all $l \in [r - 1]$. We define $Z_1 = \left\{ l \in [r - 1] : \langle D_l, D_r \rangle \geq -\theta \right\}$ and $Z_2 = \left\{ l \in [r - 1] : -1 < \langle D_l, D_r \rangle < -\theta \right\}$. WLOG, let $Z_1 = [r_1], Z_2 = \{r_1 + 1, \ldots, r_2\}$, where $0 \leq r_1 \leq r_2 \leq r - 1$. We define the following events

$$\zeta_l = \left\{ \left\langle w^{(0)}_l, \mu_l \right\rangle \leq C_{\text{Sure},2} b_l \right\}, \hat{\zeta}_l = \left\{ \left| \left\langle w^{(0)}_l, \hat{\mu}_l \right\rangle \right| \leq C_{\text{Sure},2} b_l \right\}. \tag{237}$$

We define space $A = \text{span}(\mu_1, \ldots, \mu_r)$ and $\hat{\mu}_r = P_{A^+} \mu_r$, where $P_{A^+}$ is the projection operator on the complementary space of $A$. For $l \in Z_2$, we also define $\hat{\mu}_l = \mu_l - \frac{\langle \mu_l, \mu_r \rangle \mu_r}{\|\mu_r\|_2^2}$, and the event

$$\hat{\zeta}_l = \left\{ \left\langle w^{(0)}_l, \hat{\mu}_l \right\rangle \leq C_{\text{Sure},2} b_l \right\}, \hat{\zeta}_l = \left\{ \left| \left\langle w^{(0)}_l, \hat{\mu}_l \right\rangle \right| \leq C_{\text{Sure},2} b_l \right\}. \tag{238}$$

For $l \in Z_2$, we have $\mu_l = \hat{\mu}_l - \rho \mu_r$, where $\rho \geq 0$. So $\langle w, \mu_l \rangle = \langle w, \hat{\mu}_l \rangle - \rho \langle w, \mu_r \rangle \leq \langle w, \hat{\mu}_l \rangle$ when $\langle w, \mu_r \rangle \geq 0$. As a result, we have

$$\hat{\zeta}_l \cap \left\{ \left\langle w^{(0)}_l, \mu_r \right\rangle \geq C_{\text{Sure},1} b_l \right\} \leq \zeta_l \cap \left\{ \left\langle w^{(0)}_l, \mu_r \right\rangle \geq C_{\text{Sure},1} b_l \right\}. \tag{239}$$

By Assumption 4.1, we have

$$\frac{1}{2} \leq 1 - r\theta \leq 1 - r_1 \theta \leq \frac{\|\mu_r\|_2^2}{\|\mu_r\|_2} \leq 1. \tag{240}$$

We also have,

$$\Pr \left[ \left\langle w^{(0)}_l, \mu_r \right\rangle \geq C_{\text{Sure},1} b_l, \zeta_1, \ldots, \zeta_{r-1} \right] \geq \Pr \left[ \left\langle w^{(0)}_l, \mu_r \right\rangle \geq C_{\text{Sure},1} b_l, \zeta_1, \ldots, \zeta_{r_2} \right] \geq \Pr \left[ \left\langle w^{(0)}_l, \mu_r \right\rangle \geq C_{\text{Sure},1} b_l, \zeta_1, \ldots, \zeta_{r_1, r_1+1}, \ldots, \zeta_{r_2} \right]$$

$$\geq \Pr \left[ \left\langle w^{(0)}_l, \mu_r \right\rangle \geq C_{\text{Sure},1} b_l, \zeta_1, \ldots, \zeta_{r_1}, \zeta_{r_1+1}, \ldots, \zeta_{r_2} \right] \geq \Pr \left[ \left\langle w^{(0)}_l, \mu_r \right\rangle \geq C_{\text{Sure},1} b_l, \zeta_1, \ldots, \zeta_{r_1}, \zeta_{r_1+1}, \ldots, \zeta_{r_2} \right] \left| \begin{array}{c} \rho \in [r_2] \rho_l \end{array} \right| \Pr \left[ \left\{ \zeta_1, \ldots, \zeta_{r_1}, \zeta_{r_1+1}, \ldots, \zeta_{r_2} \right\} \right]. \tag{244}$$
For the first condition in Definition E.6, we have,

\[
p_r = \Pr \left[ \left\langle w_i^{(0)}, \mu_r \right\rangle \geq C_{\text{Sure}, 1} b_i \left\hat{\zeta}_1, \ldots, \hat{\zeta}_r, \hat{\zeta}_{r+1}, \ldots, \hat{\zeta}_2 \right\rangle \right]
\]

\[
= \Pr \left[ \left\langle w_i^{(0)}, \hat{\mu}_r + \mu_r - \hat{\mu}_r \right\rangle \geq C_{\text{Sure}, 1} b_i \left\hat{\zeta}_1, \ldots, \hat{\zeta}_r \right\rangle \right]
\]

\[
\geq \Pr \left[ \left\langle w_i^{(0)}, \hat{\mu}_r + \mu_r - \hat{\mu}_r \right\rangle \geq C_{\text{Sure}, 1} b_i \left\langle w_i^{(0)}, \mu_r - \hat{\mu}_r \right\rangle \geq 0 \right] \geq \hat{\zeta}_1, \ldots, \hat{\zeta}_r \right\rangle
\]

\[
= \Pr \left[ \left\langle w_i^{(0)}, \hat{\mu}_r - \mu_r \right\rangle \geq 0 \right] \geq \hat{\zeta}_1, \ldots, \hat{\zeta}_r \right\rangle
\]

\[
= \frac{1}{2} \Pr \left[ \left\langle w_i^{(0)}, \hat{\mu}_r \right\rangle \geq C_{\text{Sure}, 1} b_i \left\langle w_i^{(0)}, \mu_r - \hat{\mu}_r \right\rangle \geq 0 \right] \geq \hat{\zeta}_1, \ldots, \hat{\zeta}_r \right\rangle
\]

\[
= \frac{1}{2} \Pr \left[ \left\langle w_i^{(0)}, \hat{\mu}_r \right\rangle \geq C_{\text{Sure}, 1} b_i \right] \geq \hat{\zeta}_1, \ldots, \hat{\zeta}_r \right\rangle
\]

\[
\geq \Theta \left( \frac{\|\hat{\mu}_r\|_2}{\sqrt{\tau \log d} \sigma_{B+} \cdot d(9C_i^2 \tau \sigma_B^2 + (2\|\hat{\mu}_r\|_2))} \right).
\]

where the last equality following that \( \hat{\mu}_r \) is orthogonal with \( \mu_1, \ldots, \mu_{r_1} \) and the property of the standard Gaussian vector, and the last inequality follows Lemma F.6.

For the second condition in Definition E.6, by Lemma F.6, we have,

\[
p_1 = \Pr \left[ \hat{\zeta}_1 \right] = 1 - \Theta \left( \frac{\|\hat{\mu}_1\|_2}{\sqrt{\tau \log d} \sigma_{B+} \cdot d(9C_i^2 \tau \sigma_B^2 + (8\|\hat{\mu}_1\|_2^2))} \right)
\]

\[
p_2 = \Pr \left[ \hat{\zeta}_2 | \hat{\zeta}_1 \right] = \Pr \left[ \hat{\zeta}_2 \right] \geq 1 - \Theta \left( \frac{\|\hat{\mu}_2\|_2}{\sqrt{\tau \log d} \sigma_{B+} \cdot d(9C_i^2 \tau \sigma_B^2 + (8\|\hat{\mu}_2\|_2^2))} \right)
\]

\[
\vdots
\]

\[
p_{r-1} = \Pr \left[ \hat{\zeta}_{r-1} | \hat{\zeta}_1, \ldots, \hat{\zeta}_{r-1}, \hat{\zeta}_{r+1}, \ldots, \hat{\zeta}_{2r} \right] \geq \Pr \left[ \hat{\zeta}_{r-1} \right] \geq \Pr \left[ \hat{\zeta}_{r-1} \right]
\]

\[
\geq 1 - \Theta \left( \frac{\|\hat{\mu}_{r-1}\|_2}{\sqrt{\tau \log d} \sigma_{B+} \cdot d(9C_i^2 \tau \sigma_B^2 + (8\|\hat{\mu}_{r-1}\|_2^2))} \right).
\]

On the other hand, if \( X \) is a \( \chi^2(k) \) random variable. Then we have

\[
\Pr(X \geq k + 2\sqrt{kx} + 2x) \leq e^{-x}.
\]

Therefore, by assumption \( B_{\mu_1} \geq \Omega \left( \sigma_{B+} \sqrt{\frac{\tau \log d}{d}} \right) \), we have

\[
\Pr \left( \frac{1}{\sigma_{\mu_1}^2} \left\| w_i^{(0)} \right\|_2^2 \geq d + 2 \sqrt{(9C_i^2 \tau \sigma_B^2 + (2B_{\mu_1}^2)) \log d} \right)
\]

\[
\leq O \left( \frac{1}{d(9C_i^2 \tau \sigma_B^2 + (2B_{\mu_1}^2))} \right).
\]

43
Recall $B_{μ1} = \min_{j∈[r]} ∥μ_j∥_2$, $B_{μ2} = \max_{j∈[r]} ∥μ_j∥_2$. Thus, by union bound, we have

$$\Pr \left[ i \in S_{D_j, \text{Sure}} \right] \geq \prod_{i \in [r]} \Pr \left[ 1 - \frac{1}{d^2 \cdot \left( \Theta(\sigma u^2) \right)} \right]$$

(263)

$$\geq \left( \frac{B_{μ1}}{\sqrt{\log d} \cdot \sigma_{B+} \cdot \left( \Theta(\sigma u^2) \right)} \right) \cdot \left( 1 - \frac{r B_{μ2}}{\sqrt{\log d} \cdot \sigma_{B+} \cdot \left( \Theta(\sigma u^2) \right)} \right)$$

(264)

(265)

$$\geq \left( \frac{B_{μ1}}{\sqrt{\log d} \cdot \sigma_{B+} \cdot \left( \Theta(\sigma u^2) \right)} \right) .$$

(266)

(267)

In Lemma E.9, we compute the activation pattern for the neurons in $S_{D_j, \text{Sure}}$.

**Lemma E.9 (Mixture of Gaussians: Activation Pattern).** Assume the same conditions as in Lemma E.7, for all $j ∈ [r], i ∈ S_{D_j, \text{Sure}}$, we have

1. When $x \sim N_j(μ_j, σ_j I_{d×d})$, the activation probability satisfies,

$$\Pr_{x \sim N_j(μ_j, σ_j I_{d×d})} \left[ \left( w_i^{(0)} , x \right) - b_i \geq 0 \right] \geq 1 - O \left( \frac{1}{d^2} \right).$$

(268)

2. For all $j' ≠ j ∈ [r]$, when $x \sim N_j'(μ_j', σ_j' I_{d×d})$, the activation probability satisfies,

$$\Pr_{x \sim N_j'(μ_j', σ_j' I_{d×d})} \left[ \left( w_i^{(0)} , x \right) - b_i \geq 0 \right] \leq O \left( \frac{1}{d^2} \right).$$

(269)

**Proof of Lemma E.9.** In the proof, we need $\tilde{b} = C_b \sqrt{d \log d} log d \sigma_B$, where $C_b$ is a large enough universal constant. For the first statement, when $x \sim N_j(μ_j, σ_j I_{d×d})$, by $C_{\text{Sure},1} ≥ \frac{1}{2}$, we have

$$\Pr_{x \sim N_j(μ_j, σ_j I_{d×d})} \left[ \left( w_i^{(0)} , x \right) - b_i \geq 0 \right] ≥ \frac{1}{2}$$

(270)

(271)

(272)

(273)

(274)

where the third inequality follows the Chernoff bound and symmetricity of the Gaussian vector.

For the second statement, we prove similarly by $0 < C_{\text{Sure},2} ≤ \frac{1}{2}$. □

Then, Lemma E.10 gives gradients of neurons in $S_{D_j, \text{Sure}}$. It shows that these gradients are highly aligned with $D_j$.

**Lemma E.10 (Mixture of Gaussians: Feature Emergence).** Assume the same conditions as in Lemma E.7, for all $j ∈ [r], i ∈ S_{D_j, \text{Sure}}$, we have

$$\left\langle \mathbb{E}_{(x,y)} \left[ y \sigma' \left( \left( w_i^{(0)} , x \right) - b_i \right) , x \right] , y(j) D_j \right\rangle$$

(275)

$$≥ \rho_j B_{μ1} \sqrt{d} \left( 1 - O \left( \frac{1}{d^2} \right) \right) - B_{μ2} O \left( \frac{1}{d^{2.5}} \right) - \sigma_{B+} O \left( \frac{1}{d^{0.97}} \right).$$

(276)
For any unit vector $D_j^\top$ which is orthogonal with $D_j$, we have
\[\left| \langle \mathbb{E}_{(x,y)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right], D_j^\top \rangle \right| \leq B_{\mu_2} O \left( \frac{1}{d^{\tau - \frac{1}{2}}} \right) + \sigma_B O \left( \frac{1}{d^{\beta_9 \tau}} \right). \tag{277}\]

**Proof of Lemma E.10.** For all $j \in [r]$, $i \in S_{D_j, \text{Sure}}$, we have
\[\mathbb{E}_{(x,y)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right] = \sum_{i \in [r]} p_i \mathbb{E}_{\epsilon \sim N_i(x)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right] \tag{278}\]
\[= \sum_{i \in [r]} p_i \mathbb{E}_{\epsilon \sim N_0(x)} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) (x + \mu_j)^{\top} D_j \right] \tag{279}\]
\[= \sum_{i \in [r]} p_i y(i) \mathbb{E}_{\epsilon \sim N_0(0, \sigma I_{d \times d})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_i \rangle - b_i \right) (x + \mu_i)^{\top} D_j \right]. \tag{280}\]

Thus, by Lemma F.7 and Lemma E.9,
\[\left\langle \mathbb{E}_{(x,y)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right], y(j) D_j \right\rangle \tag{281}\]
\[= p_j \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) (x + \mu_j)^{\top} D_j \right] \tag{282}\]
\[+ \sum_{i \in [r], i \neq j} p_i y(i) y(j) \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_i \rangle - b_i \right) (x + \mu_i)^{\top} D_j \right] \tag{283}\]
\[\geq p_j B_{\mu_1} \sqrt{d} \left( 1 - O \left( \frac{1}{d^{\tau}} \right) \right) - B_{\mu_2} O \left( \frac{1}{d^{\tau - \frac{1}{2}}} \right) \tag{284}\]
\[- p_j \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) x^{\top} D_j \right] \tag{285}\]
\[- \sum_{i \in [r], i \neq j} p_i \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_i \rangle - b_i \right) x^{\top} D_j \right] \tag{286}\]
\[\geq p_j B_{\mu_1} \sqrt{d} \left( 1 - O \left( \frac{1}{d^{\tau}} \right) \right) - B_{\mu_2} O \left( \frac{1}{d^{\tau - \frac{1}{2}}} \right) \tag{287}\]
\[- p_j \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \left( 1 - \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) - 1 \right) x^{\top} D_j \right] \tag{288}\]
\[- \sum_{i \in [r], i \neq j} p_i \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_i \rangle - b_i \right) x^{\top} D_j \right] \tag{289}\]
\[= p_j B_{\mu_1} \sqrt{d} \left( 1 - O \left( \frac{1}{d^{\tau}} \right) \right) - B_{\mu_2} O \left( \frac{1}{d^{\tau - \frac{1}{2}}} \right) \tag{290}\]
\[- p_j \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \left( 1 - \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) \right) x^{\top} D_j \right] \tag{291}\]
\[- \sum_{i \in [r], i \neq j} p_i \mathbb{E}_{\epsilon \sim N(0, \sigma I_{d \times d})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_i \rangle - b_i \right) x^{\top} D_j \right] \tag{292}\]
\[\geq p_j B_{\mu_1} \sqrt{d} \left( 1 - O \left( \frac{1}{d^{\tau}} \right) \right) - B_{\mu_2} O \left( \frac{1}{d^{\tau - \frac{1}{2}}} \right) - \sigma_B O \left( \frac{1}{d^{\beta_9 \tau}} \right). \tag{293}\]
For any unit vector $D_j^+$ which is orthogonal with $D_j$, similarly, we have
\[
\left| \left\langle \mathbb{E}_{(x,y)} \left[ y \sigma' \left( \left\langle w_i^0, x \right\rangle - b_i \right) x, D_j^+ \right] \right\rangle \right| 
\leq p_j \mathbb{E}_{x \sim N(0, \sigma_1 I)} \left[ \sigma' \left( \left\langle w_i^0, x + \mu_j \right\rangle - b_i \right) x^T D_j^+ \right] + \sum_{l \in [r], l \neq j} p_l \mathbb{E}_{x \sim N(0, \sigma_1 I)} \left[ \sigma' \left( \left\langle w_i^0, x + \mu_l \right\rangle - b_i \right) (x + \mu_l)^T D_j^+ \right]
\]
\leq B\mu_2 O \left( \frac{1}{d^{1 - \frac{3}{2}}} \right) + p_j \mathbb{E}_{x \sim N(0, \sigma_1 I)} \left[ \sigma' \left( \left\langle w_i^0, x + \mu_j \right\rangle - b_i \right) x^T D_j^+ \right] + \sum_{l \in [r], l \neq j} p_l \mathbb{E}_{x \sim N(0, \sigma_1 I)} \left[ \sigma' \left( \left\langle w_i^0, x + \mu_l \right\rangle - b_i \right) x^T D_j^+ \right]
\leq B\mu_2 O \left( \frac{1}{d^{1 - \frac{3}{2}}} \right) + \sigma_B + O \left( \frac{1}{d^{3/2}} \right).
\]

\[\square\]

E.2.3 Mixture of Gaussians: Final Guarantee

Lemma E.11 (Mixture of Gaussians: Existence of Good Networks. Part statement of Lemma 4.3). Assume the same conditions as in Lemma E.7. Define
\[
f^*(x) = \sum_{j=1}^{r} \frac{y(j)}{\sqrt{\tau \log d} \sigma_B^+} \left[ \sigma \left( \left\langle D_j, x \right\rangle - 2\sqrt{\tau \log d} \sigma_B^+ \right) \right].
\]

For $D_{\text{mixture}}$ setting, we have $f^* \in F_{d,r,B_F,S_{p,\gamma},BG}$, where $B_F = (B_{a1}, B_{a2}, B_b) = \left( \frac{1}{\sqrt{\tau \log d} \sigma_B^+}, \frac{\sqrt{\tau}}{\sqrt{\tau \log d} \sigma_B^+}, 2\sqrt{\tau \log d} \sigma_B^+ \right)$, $p = \Theta \left( \frac{B_{a1}}{\sqrt{\tau \log d} \sigma_B^+ d^{(8n^2 + \sigma_B^+)^2/(2\sigma_B^+)^2}} \right)$, $\gamma = \frac{B_G}{\sqrt{\tau \log d} \sigma_B^+}$, $B_G = p_B B_{a1} \sqrt{d} - O \left( \frac{\sigma_B^+}{\sqrt{\tau \log d} \sigma_B^+} \right)$ and $B_{x2} = (B_{a2} + \sigma_B^+)^2 d$. We also have $\text{OPT}_{d,r,B_F,S_{p,\gamma},BG} \leq \frac{3}{d^{3/2}} + \frac{4}{d^{3/4}} \sqrt{\tau \log d}$.

Proof of Lemma E.11. We can check $B_{x1} = (B_{a2} + \sigma_B^+)^2 d$, $B_{x2} = (B_{a2} + \sigma_B^+)^2 d$ by direct calculation. By Lemma E.7, we have $f^* \in F_{d,r,B_F,S_{p,\gamma},BG}$.

For any $j \in [r]$, by $B_{\mu_1} \geq \Omega \left( \sigma_B^+ \sqrt{\frac{\tau \log d}{d}} \right) \geq 4\sigma_B^+ \sqrt{\frac{\tau \log d}{d}}$, we have
\[
\begin{align*}
\mathbb{P}_{x \sim N_j(\mu_j, I_{d \times d})} \left[ \left\langle D_j, x \right\rangle - 2\sqrt{\tau \log d} \sigma_B^+ \geq \sqrt{\tau \log d} \sigma_B^+ \right] \\
= \mathbb{P}_{x \sim N_j(0, I_{d \times d})} \left[ \left\langle D_j, x \right\rangle + \| \mu_j \|_2 - 2\sqrt{\tau \log d} \sigma_B^+ \geq \sqrt{\tau \log d} \sigma_B^+ \right] \\
\geq \mathbb{P}_{x \sim N_j(0, I_{d \times d})} \left[ \left\langle D_j, x \right\rangle + \sqrt{d} B_{\mu_1} - 2\sqrt{\tau \log d} \sigma_B^+ \geq \sqrt{\tau \log d} \sigma_B^+ \right] \\
\geq \mathbb{P}_{x \sim N_j(0, I_{d \times d})} \left[ \left\langle D_j, x \right\rangle \geq -\sqrt{\tau \log d} \sigma_B^+ \right] \\
\geq 1 - \frac{1}{d^d},
\end{align*}
\]

where the last inequality follows Chernoff bound.
For any \( l \neq j, l \in [r] \), by \( \theta \leq \frac{\pi n^2}{B_2 r^2} \sqrt{\frac{\tau \log d}{d}} \), we have

\[
\Pr_{x \sim N_j(\mu_j, \sigma_j, I_{d \times d})} \left[ \langle D_l, x \rangle - 2 \sqrt{\tau \log d \sigma_{B+}} \geq 0 \right] \leq \Pr_{x \sim N_j(0, \sigma_j, I_{d \times d})} \left[ \langle D_l, x \rangle + \theta B_{i,2} \sqrt{d} - 2 \sqrt{\tau \log d \sigma_{B+}} \geq 0 \right] \leq \Pr_{x \sim N_j(0, \sigma_j, I_{d \times d})} \left[ \langle D_l, x \rangle \geq \sqrt{\tau \log d \sigma_{B+}} \right] \leq \frac{1}{d^r}.
\]

Thus, we have

\[
\Pr_{(x, y) \sim D_{\text{mixture}}} \left[ yf^*(x) > 1 \right] \geq \sum_{j \in [r]} p_j \left( \Pr_{x \sim N_j(\mu_j, \sigma_j, I_{d \times d})} \left[ \langle D_j, x \rangle - 2 \sqrt{\tau \log d \sigma_{B+}} \geq \sqrt{\tau \log d \sigma_{B+}} \right] \right) - \sum_{j \in [r]} p_j \left( \sum_{l \neq j, l \in [r]} \Pr_{x \sim N_j(\mu_j, \sigma_j, I_{d \times d})} \left[ \langle D_l, x \rangle - 2 \sqrt{\tau \log d \sigma_{B+}} < 0 \right] \right) \geq 1 - \frac{2}{d^r}.
\]

We also have

\[
\mathbb{E}_{(x, y) \sim D_{\text{mixture}}} \left[ \mathbb{I}[yf^*(x) \leq 1] \right] \leq \mathbb{E}_{x \sim N_j(\mu_j, \sigma_j, I_{d \times d})} \left[ \left\langle D_l, x \right\rangle - 2 \sqrt{\tau \log d \sigma_{B+}} < \sqrt{\tau \log d \sigma_{B+}} \right] \left[ \frac{y_f^*}{\sqrt{\tau \log d \sigma_{B+}}} \right] + \mathbb{E}_{x \sim N_j(0, \sigma_j, I_{d \times d})} \left[ \sigma' \left( \langle D_l, x \rangle - 2 \sqrt{\tau \log d \sigma_{B+}} > 0 \right) \frac{\langle D_l, x \rangle - 2 \sqrt{\tau \log d \sigma_{B+}}}{\sqrt{\tau \log d \sigma_{B+}}} \right] \leq \frac{1}{d^r} + \frac{1}{\tau \log d} \sum_{j \in [r]} p_j \left( \mathbb{E}_{x \sim N_j(0, \sigma_j, I_{d \times d})} \left[ \sigma' \left( \langle D_l, x \rangle > \sqrt{\tau \log d \sigma_{B+}} \right) \frac{\langle D_l, x \rangle - \sqrt{\tau \log d \sigma_{B+}}}{\sqrt{\tau \log d \sigma_{B+}}} \right] \right) \leq \frac{1}{d^r} + \frac{4}{d^{0.9r - 1} \sqrt{\tau \log d}},
\]

where the second last inequality follows Lemma F.7 and \( r \leq 2d \). Thus, we have

\[
\text{OPT}_{d, r, B, S_p, N_B, \omega} \leq \mathbb{E}_{(x, y) \sim D_{\text{mixture}}} [f(yf^*(x))] = \mathbb{E}_{(x, y) \sim D_{\text{mixture}}} \left[ \mathbb{I}[yf^*(x) \leq 1] \right] \left( 1 - yf^*(x) \right) \leq \mathbb{E}_{(x, y) \sim D_{\text{mixture}}} \left[ \mathbb{I}[yf^*(x) \leq 1] \right] \mathbb{I}[yf^*(x) \leq 1] + \mathbb{E}_{(x, y) \sim D_{\text{mixture}}} \left[ \mathbb{I}[yf^*(x) \leq 1] \right] \leq \frac{3}{d^r} + \frac{4}{d^{0.9r - 1} \sqrt{\tau \log d}}.
\]

**Theorem 4.4 (Mixtures of Gaussians: Main Result).** Assume Assumption 4.1. For any \( \epsilon, \delta \in (0, 1) \), when Algorithm 1 uses hinge loss with

\[
m = \text{poly} \left( \frac{1}{\delta}, \frac{1}{\epsilon}, d^{0.9r - 1}r_B^2, r_B \right) \leq e^d, \quad T = \text{poly} (m), \quad n = \text{poly} (m)
\]

47
and proper hyper-parameters, then with probability at least $1 - \delta$, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f_{\Xi(t)}(x)) \neq y] \leq \frac{\sqrt{2r}}{d^{0.47 - 0.8}} + \epsilon.$$  

**Proof of Theorem 4.4.** Let $\tilde{b} = C_b \sqrt{d \log d} \sigma_w \sigma_B^{+}$, where $C_b$ is a large enough universal constant. By Lemma E.11, we have $f^* \in \mathcal{F}_{d,r,B_F,S_{p,\gamma},B_G}$, where $B_F = (B_{a_1}, B_{a_2}, B_b) = \left(\frac{1}{\sqrt{\log d} \sigma_B^{+}}, \frac{\sqrt{d}}{\sqrt{\log d} \sigma_B^{+}}, 2\sqrt{\log d} \sigma_B^{+}\right)$.  

Let $p = \Theta \left(\frac{B_{a_1}}{\sqrt{\log d} \sigma_B^{+} \log \left[\frac{C_B \sigma_B^{+}}{\epsilon} \right]}\right)$. Choose $\gamma = \frac{1}{\sqrt{d \log d}}$ and $B_G = \frac{D}{\sqrt{\log d}}$. Adjust $\sigma_w$ such that $\tilde{b} = C_b \sqrt{d \log d} \sigma_w \sigma_B^{+} = \Theta \left(\frac{B_{a_1}^2 B_{a_2}^2 \sqrt{d}}{\sigma_B^{+}}\right)$. Injecting above parameters into Theorem 3.12, we have with probability at least $1 - \delta$ over the initialization, with proper hyper-parameters, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f_{\Xi(t)}(x)) \neq y] \leq \frac{3}{d^r} + \frac{4}{d^{0.97 - 1.5} \sqrt{\log d}} + \frac{\sqrt{2r B}\mu_2}{d^{0.97 - 1.5}/\sqrt{\log d}} + O \left(\frac{r B_{a_1} B_{a_2} \sqrt{d}}{B_G \log n}\right) + \epsilon/2$$  

$$\leq \frac{\sqrt{2r}}{d^{0.47 - 0.8}} + \epsilon.$$  

\[\square\]

**E.3 Mixture of Gaussians - XOR**

We consider a special Mixture of Gaussians distribution studied in [99]. Consider the same data distribution in Appendix E.2.1 and Definition E.6 with the following assumptions.

**Assumption E.12 (Mixture of Gaussians in [99]).** Assume four Gaussians cluster with XOR-like pattern, for any $\tau > 0$,

- $r = 4$ and $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$.
- $\mu_1 = -\mu_2$, $\mu_3 = -\mu_4$, and $||\mu_1||_2 = ||\mu_2||_2 = ||\mu_3||_2 = ||\mu_4||_2 = \sqrt{d}$ and $\langle \mu_1, \mu_3 \rangle = 0$.
- For all $j \in [4]$, $\Sigma_j = \sigma_B I_{d \times d}$ and $1 \leq \sigma_B \leq \sqrt{\frac{d}{\tau \log \log d}}$.
- $y(1) = y(2) = 1$ and $y(3) = y(4) = -1$.

We denote this data distribution as $D_{\text{mixture-xor}}$ setting.

**E.3.1 Mixture of Gaussians - XOR: Feature Learning**

**Lemma E.13 (Mixture of Gaussians in [99]: Gradient Feature Set).** Let $C_{\text{Sure},1} = \frac{b}{5}$, $C_{\text{Sure},2} = \frac{\sqrt{\sigma}}{\sqrt{d \log d}}$, $\tilde{b} = \sqrt{d \log d} \sigma_w \sigma_B^{+}$ and $d$ is large enough. For $D_{\text{mixture-xor}}$ setting, we have $(D_j, +1) \in S_{p,\gamma,B_G}$ for all $j \in [4]$, where

$$p = \Theta \left(\frac{1}{\sqrt{\log d} \sigma_B^{+} \log d^{18 + \frac{\gamma}{2\sqrt{d}}}\sigma_B^{+}}\right), \quad \gamma = \sqrt{d}$$  

$$B_G = \frac{\sqrt{d}}{4} \left(1 - O \left(\frac{1}{\log d^{0.018\tau}}\right)\right) - \sigma_B O \left(\frac{1}{\log d^{0.018\tau}}\right).$$  

48
Proof of Lemma E.13. For all $j \in [r]$, by Lemma E.16, for all $i \in S_{D_j,\text{Sure}}$,

\begin{equation}
1 - \frac{\left\langle G(w_i^{(0)}, b_i), D_j \right\rangle}{\|G(w_i^{(0)}, b_i)\|_2} \leq 1 - \frac{\left\langle G(w_i^{(0)}, b_i), D_j \right\rangle}{\sqrt{\left\langle G(w_i^{(0)}, b_i), D_j \right\rangle^2 + \max_D D_j | D_j = 0, D_j = 1 \left\langle G(w_i^{(0)}, b_i), D_j \right\rangle^2}} \tag{324}
\end{equation}

\begin{equation}
\leq 1 - \frac{\left\langle G(w_i^{(0)}, b_i), D_j \right\rangle}{1 + \frac{\sigma_B O(\log d)}{\sqrt{d}} - \sigma_B O(\log d)} \tag{325}
\end{equation}

\begin{equation}
\leq 1 - \frac{\left\langle G(w_i^{(0)}, b_i), D_j \right\rangle}{\frac{1}{\sqrt{d}} \left(1 - O\left(\frac{1}{\log d}\right)\right) - \sigma_B O\left(\frac{1}{\log d}\right)} \tag{326}
\end{equation}

\begin{equation}
\leq \frac{\sigma_B}{\sqrt{d}} = \gamma. \tag{327}
\end{equation}

Thus, we have $G(w_i^{(0)}, b_i) \in C_{D_j,\gamma}$ and $\left\langle G(w_i^{(0)}, b_i), D_j \right\rangle \leq \|G(w_i^{(0)}, b_i)\|_2 \leq B_{x_1}$, $\frac{b_i}{b_i} = +1$. Thus, by Lemma E.14, we have

\begin{equation}
Pr \left[ G(w, b) \in C_{D_j,\gamma} \text{ and } \|G(w, b)\|_2 \geq B_G \text{ and } \frac{b}{|b|} = +1 \right] \tag{330}
\end{equation}

\begin{equation}
\geq Pr \left[ i \in S_{D_j,\text{Sure}} \right] \tag{331}
\end{equation}

\begin{equation}
\geq p. \tag{332}
\end{equation}

Thus, $(D_j, +1) \in S_{p,\gamma,B_G}$. We finish the proof. \qed

Lemma E.14 (Mixture of Gaussians in [99]: Geometry at Initialization). Assume the same conditions as in Lemma E.13. Recall for all $i \in [m]$, $w_i^{(0)} \sim N(0, \sigma_i^2 I_{d \times d})$, over the random initialization, we have for all $i \in [m], j \in [4],$

\begin{equation}
Pr \left[ i \in S_{D_j,\text{Sure}} \right] \geq \Theta \left( \frac{1}{\sqrt{\tau \log \log d} \cdot (\log d)^{\frac{1+\log 2}{2}}} \right). \tag{333}
\end{equation}

Proof of Lemma E.14. WLOG, let $j = 1$. By Assumption E.12, for the first condition in Definition E.6, we have,

\begin{equation}
Pr \left[ \left\langle w_i^{(0)}, \mu_1 \right\rangle \geq C_{\text{Sure},1} b_i \right] \geq \Theta \left( \frac{1}{\sqrt{\tau \log \log d} \cdot (\log d)^{\frac{1+\log 2}{2}}} \right), \tag{334}
\end{equation}

where the last inequality follows Lemma F.6.

For the second condition in Definition E.6, by Lemma F.6, we have,

\begin{equation}
Pr \left[ \left\langle w_i^{(0)}, \mu_2 \right\rangle \leq C_{\text{Sure},2} b_i \right] \geq 1 - \frac{1}{2\sqrt{\tau} \sigma_B} \cdot \frac{1}{e^{\sigma_B}}, \tag{335}
\end{equation}

On the other hand, if $X$ is a $\chi^2(k)$ random variable. Then we have

\begin{equation}
Pr(X \geq k + 2\sqrt{kx} + 2x) \leq e^{-x}. \tag{336}
\end{equation}
Therefore, we have
\[
\Pr \left( \frac{1}{\sigma^2} \| \mathbf{w}^{(0)}_i \|_2^2 \geq d + 2 \sqrt{\left( \frac{18\tau\sigma_B^2}{25} + 2 \right) d \log d} + 2 \left( \frac{18\tau\sigma_B^2}{25} + 2 \right) \log \log d \right) \geq 2 \Rightarrow O \left( \frac{1}{(\log d)^2 \cdot (\log d)^{\frac{18\tau\sigma_B^2}{25}}} \right).
\]
(338)
Thus, by union bound, we have
\[
\Pr [ i \in S_{D_j, \text{Sure}} ] \geq \Theta \left( \frac{1}{\sqrt{\tau \log \log d} \cdot (\log d)^{\frac{18\tau\sigma_B^2}{25}}} \right).
\]
(339)

**Lemma E.15** (Mixture of Gaussians in [99]: Activation Pattern). Assume the same conditions as in Lemma E.13, for all \( j \in [4] \), \( i \in S_{D_j, \text{Sure}} \), we have

1. When \( x \sim N_j(\mu_j, \sigma_B I_{d \times d}) \), the activation probability satisfies,
\[
\Pr_{x \sim N_j(\mu_j, \sigma_B I_{d \times d})} \left[ \langle \mathbf{w}^{(0)}_i, x \rangle - b_i \geq 0 \right] \geq 1 - \frac{1}{(\log d)^{\frac{6}{5}}}. \tag{340}
\]
2. For all \( j' \neq j, j' \in [4] \), when \( x \sim N_{j'}(\mu_{j'}, \sigma_B I_{d \times d}) \), the activation probability satisfies,
\[
\Pr_{x \sim N_{j'}(\mu_{j'}, \sigma_B I_{d \times d})} \left[ \langle \mathbf{w}^{(0)}_i, x \rangle - b_i \geq 0 \right] \leq O \left( \frac{1}{(\log d)^{\frac{1}{2}}} \right). \tag{341}
\]

**Proof of Lemma E.15.** In the proof, we need \( \bar{b} = \sqrt{\tau d \log \log d \sigma_B^2} \). For the first statement, when \( x \sim N_j(\mu_j, \sigma_B I_{d \times d}) \), by \( C_{\text{Sure}, 1} \geq \frac{6}{5} \), we have
\[
\Pr_{x \sim N_j(\mu_j, \sigma_B I_{d \times d})} \left[ \langle \mathbf{w}^{(0)}_i, x \rangle - b_i \geq 0 \right] \geq \Pr_{x \sim N(0, \sigma_B I_{d \times d})} \left[ \langle \mathbf{w}^{(0)}_i, x \rangle \geq (1 - C_{\text{Sure}, 1})b_i \right] \geq \frac{1}{\sigma^2} \mathbf{w}^2_i \cdot \mathbf{b}^2_i \leq \frac{1}{5} \tag{342}
\]
\[
\geq \frac{1}{\sigma^2} \mathbf{w}^2_i \cdot \mathbf{b}^2_i \leq \frac{1}{5} \tag{343}
\]
\[
= 1 - \exp \left( -\frac{b^2_i}{50d\sigma^2_w \sigma_B^2} \right) \geq 1 - \frac{1}{(\log d)^{\frac{6}{5}}}, \tag{344}
\]
where the third inequality follows the Chernoff bound and symmetricity of the Gaussian vector.

For the second statement, we prove similarly by \( 0 < C_{\text{Sure}, 2} \leq \frac{\sqrt{2}}{\sqrt{\tau \log \log d}} \).

Then, Lemma E.16 gives gradients of neurons in \( S_{D_j, \text{Sure}} \). It shows that these gradients are highly aligned with \( D_j \).

**Lemma E.16** (Mixture of Gaussians in [99]: Feature Emergence). Assume the same conditions as in Lemma E.13, for all \( j \in [4] \), \( i \in S_{D_j, \text{Sure}} \), we have
\[
\left\langle E_{(x,y)} \left[ \gamma \sigma' \left( \langle \mathbf{w}^{(0)}_i, x \rangle - b_i \right) x \right], y_{(j)} D_j \right\rangle \geq \frac{1}{4} \sqrt{d} \left( 1 - O \left( \frac{1}{(\log d)^{\frac{6}{5}}} \right) \right) - \sigma_B O \left( \frac{1}{(\log d)^{0.0187}} \right). \tag{347}
\]
(347)
For any unit vector \( D_j^\perp \) which is orthogonal with \( D_j \), we have
\[
\left| \left\langle E_{(x,y)} \left[ \gamma \sigma' \left( \langle \mathbf{w}^{(0)}_i, x \rangle - b_i \right) x \right], D_j^\perp \right\rangle \right| \leq \sigma_B O \left( \frac{1}{(\log d)^{0.0187}} \right). \tag{348}
\]
(348)
Proof of Lemma E.16. For all $j \in [4]$, $i \in S_{D_j, \text{Sure}}$, we have
\[
\mathbb{E}_{(x,y)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right] = \frac{1}{4} \mathbb{E}_{x \sim N_i(x)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right]
\] (350)
\[
= \frac{1}{4} \mathbb{E}_{x \sim N_1(x)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right]
\] (351)
\[
= \frac{1}{4} \mathbb{E}_{x \sim N_2(x)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right]
\] (352)
Thus, by Lemma F.7 and Lemma E.15,
\[
\mathbb{E}_{(x,y)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) x \right] \leq \frac{1}{4} \mu_j^T D_j \left( 1 - O \left( \frac{1}{(\log d)^{\sigma_B}} \right) \right) - \sum_{l \in [4], l \neq j} \frac{1}{4} |\mu_l^T D_j| \circ O \left( \frac{1}{d^{\sigma}} \right)
\] (353)
\[
= \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_B I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) \right] (x + \mu_j)^T D_j
\] (354)
\[
+ \sum_{l \in [4], l \neq j} \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_i I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_l \rangle - b_i \right) \right] (x + \mu_l)^T D_j
\] (355)
\[
\geq \frac{1}{4} \sum_{l \in [4], l \neq j} \mathbb{E}_{x \sim N(0, \sigma_B I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) \right] (x + \mu_j)^T D_j
\] (356)
\[
\geq \frac{1}{4} \sqrt{d} \left( 1 - O \left( \frac{1}{(\log d)^{\sigma_B}} \right) \right) - \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_B I_{4 \times 4})} \left[ \left( 1 - \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) \right) \right] (x + \mu_j)^T D_j
\] (357)
\[
- \sum_{l \in [4], l \neq j} \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_i I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_l \rangle - b_i \right) \right] (x + \mu_l)^T D_j
\] (358)
\[
\geq \frac{1}{4} \sqrt{d} \left( 1 - O \left( \frac{1}{(\log d)^{\sigma_B}} \right) \right) - \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_B I_{4 \times 4})} \left[ \left( 1 - \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) \right) \right] x^T D_j
\] (359)
\[
- \sum_{l \in [4], l \neq j} \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_i I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_l \rangle - b_i \right) \right] x^T D_j
\] (360)
\[
\geq \frac{1}{4} \sqrt{d} \left( 1 - O \left( \frac{1}{(\log d)^{\sigma_B}} \right) \right) - \sigma_B O \left( \frac{1}{(\log d)^{0.018\sigma}} \right)
\] (361)
For any unit vector $D_j^\perp$ which is orthogonal with $D_j$, similarly, we have
\[
\left| \mathbb{E}_{(x,y)} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \right) \right] D_j^\perp \right| \leq \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_B I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) \right] (x + \mu_j)^T D_j^\perp
\] (362)
\[
+ \sum_{l \in [4], l \neq j} \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_i I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_l \rangle - b_i \right) \right] (x + \mu_l)^T D_j^\perp
\] (363)
\[
\leq \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_B I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_j \rangle - b_i \right) \right] (x + \mu_j)^T D_j^\perp
\] (364)
\[
+ \sum_{l \in [4], l \neq j} \frac{1}{4} \mathbb{E}_{x \sim N(0, \sigma_i I_{4 \times 4})} \left[ \sigma' \left( \langle w_i^{(0)}, x + \mu_l \rangle - b_i \right) \right] (x + \mu_l)^T D_j^\perp
\] (365)
\[
\leq \sigma_B O \left( \frac{1}{(\log d)^{0.018\sigma}} \right).
\] (366)
\[
\square
\]
E.3.2 Mixture of Gaussians - XOR: Final Guarantee

Lemma E.17 (Mixture of Gaussians in [99]: Existence of Good Networks). Assume the same conditions as in Lemma E.13 and let $\tau = 1$ and when $0 < \tilde{\tau} \leq O\left(\frac{d}{\sigma_B \log d}\right)$. Define

$$f^*(x) = \sum_{j=1}^{4} \frac{y(j)}{\sqrt{\tilde{\tau} \log d_B}} \left[ \sigma \left( \langle D_j, x \rangle - 2\sqrt{\tilde{\tau} \log d_B} \right) \right].$$

(368)

For $D_{\text{mixture-xor}}$ setting, we have $f^* \in \mathcal{F}_{d,r,B_F,S_p,\gamma,b_G}$, where $B_F = (B_{a1}, B_{a2}, B_b) = \left(\frac{1}{\sqrt{\tau \log d_B}}, \frac{2}{\sqrt{\tau \log d_B}}, 2\sqrt{\tau \log d_B}\right)$, $p = \Omega\left(\frac{1}{\sigma_B \log d_B}\right)$, $\gamma = \frac{\sigma_B}{\sqrt{\tau}}$, $r = 4$, $B_G = \frac{1}{\sqrt{\tau}}$ and $B_{x1} = (1 + \sigma_B)\sqrt{\tau}, B_{x2} = (1 + \sigma_B)^2d$. We also have $\text{OPT}_{d,r,B_F,S_p,\gamma,b_G} \leq \frac{4}{\delta^2} + \sqrt{\frac{\delta}{\sigma_B}} + 4\sqrt{\frac{\delta}{\log d}}$.

Proof of Lemma E.17. We finish the proof by following the proof of Lemma E.11.

Theorem E.18 (Mixture of Gaussians in [99]: Main Result). For $D_{\text{mixture-xor}}$ setting with Assumption E.12, when $d$ is large enough, for any $\delta \in (0, 1)$ and for any $\epsilon \in (0, 1)$ when

$$m = \Omega\left(\sigma_B (\log d)^{\sigma_B^2} \left(\left(\log \left(\frac{1}{\delta}\right)\right)^2 + \frac{1+\sigma_B}{\epsilon^4} + \frac{1}{\epsilon^4}\right) \right) \leq e^d,$$

(369)

$$T = \text{poly}(\sigma_B, 1/\epsilon, 1/\delta, \log d),$$

(370)

$$n = \Omega\left(\frac{m^2(1 + \sigma_B^2)}{\epsilon^2 \max\{\sigma_B \cdot (\log d)^{\sigma_B^2}, 1\}} + \sigma_B \cdot (\log d)^{\sigma_B^2} + \frac{Tm}{\delta}\right),$$

(371)

trained by Algorithm 1 with hinge loss, with probability at least $1 - \delta$ over the initialization and training samples, with proper hyper-parameters, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f_{\mathbb{E}(\cdot)}(x)) \neq y] \leq O\left(1 + \frac{\frac{1}{\sigma_B^2}}{d^4} + \frac{(\log n)^{\frac{1}{2}}}{n^{\frac{1}{4}}}\right) + \epsilon.$$

(372)

Proof of Theorem E.18. Let $\hat{b} = \sqrt{d \log \log d \sigma_w \sigma_B}$. By Lemma E.17, let $\tau = 1$ and when $\tilde{\tau} = O\left(\frac{d}{\sigma_B \log d}\right)$, we have $f^* \in \mathcal{F}_{d,r,B_F,S_p,\gamma,b_G}$, where $B_F = (B_{a1}, B_{a2}, B_b) = \left(\frac{1}{\sqrt{\tau \log d_B}}, \frac{2}{\sqrt{\tau \log d_B}}, 2\sqrt{\tau \log d_B}\right)$, $p = \Omega\left(\frac{1}{\sigma_B \log d_B}\right)$, $\gamma = \frac{\sigma_B}{\sqrt{\tau}}$, $r = 4$, $B_G = \frac{1}{\sqrt{\tau}}$ and $B_{x1} = (1 + \sigma_B)\sqrt{\tau}, B_{x2} = (1 + \sigma_B)^2d$. We also have $\text{OPT}_{d,r,B_F,S_p,\gamma,b_G} \leq \frac{4}{\delta^2} + \sqrt{\frac{\delta}{\sigma_B}} + 4\sqrt{\frac{\delta}{\log d}}$.

Adjust $\sigma_w$ such that $\hat{b} = \sqrt{d \log \log d \sigma_w \sigma_B} = \Theta\left(\frac{B_{a1}^{\frac{1}{2}} B_{a2}^{\frac{1}{2}} B_b^{\frac{1}{2}}}{\sqrt{\tau}}\right)$. Injecting above parameters into Theorem 3.12, we have with probability at least $1 - \delta$ over the initialization, with proper hyper-parameters, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f_{\mathbb{E}(\cdot)}(x)) \neq y] \leq O\left(1 + \frac{\frac{1}{\sigma_B^2}}{d^4} + \frac{(\log n)^{\frac{1}{2}}}{n^{\frac{1}{4}}}\right) + \epsilon.$$

(373)

□

E.4 Parity Functions

We recap the problem setup in Section 4.2 for readers’ convenience.

E.4.1 Problem Setup

Data Distributions. Suppose $\mathbf{M} \in \mathbb{R}^{d \times D}$ is an unknown dictionary with $D$ columns that can be regarded as patterns. For simplicity, assume $d = D$ and $\mathbf{M}$ is orthonormal. Let $\phi \in \mathbb{R}^{d}$ be a hidden representation vector. Let $A \subseteq [D]$ be a subset of size $r \times k$ corresponding to the class relevant patterns.
and \( r \) is an odd number. Then the input is generated by \( \mathbf{M}\phi \), and some function on \( \phi_A \) generates the label. WLOG, let \( A = \{1, \ldots, rk\} \), \( A^\perp = \{rk + 1, \ldots, d\} \). Also, we split \( A \) such that for all \( j \in [r] \), \( A_j = \{(j-1)k+1, \ldots, jk\} \). Then the input \( x \) and the class label \( y \) are given by:

\[
x = \mathbf{M}\phi, \quad y = g^*(\phi_A) = \text{sign} \left( \sum_{j=1}^{r} \text{XOR}(\phi_{A_j}) \right),
\]

(374)

where \( g^* \) is the ground-truth labeling function mapping from \( \mathbb{R}^{rk} \) to \( \mathcal{Y} = \{\pm 1\} \), \( \phi_A \) is the sub-vector of \( \phi \) with indices in \( A \), and \( \text{XOR}(\phi_{A_j}) = \prod_{i \in A_j} \phi_i \) is the parity function.

We still need to specify the distribution of \( \phi \), which determines the structure of the input distribution:

\[
\mathcal{X} := (1 - 2rp_A)\mathcal{X}_U + \sum_{j\in[r]} p_A(\mathcal{X}_{j,+} + \mathcal{X}_{j,-}).
\]

(375)

For all corresponding \( \phi_{A^\perp} \) in \( \mathcal{X} \), we have \( \forall l \in A^\perp \), independently:

\[
\phi_l = \begin{cases} +1, & \text{w.p. } p_o \\ -1, & \text{w.p. } p_o \\ 0, & \text{w.p. } 1 - 2p_o \end{cases},
\]

where \( p_o \) controls the signal noise ratio: if \( p_o \) is large, then there are many nonzero entries in \( A^\perp \) which are noise interfering with the learning of the ground-truth labeling function on \( A \).

For corresponding \( \phi_A \), any \( j \in [r] \), we have

- In \( \mathcal{X}_{j,+} \), \( \phi_{A_j} = [+1, +1, \ldots, +1]^T \) and \( \phi_{A\setminus A_j} \) only have zero elements.
- In \( \mathcal{X}_{j,-} \), \( \phi_{A_j} = [-1, -1, \ldots, -1]^T \) and \( \phi_{A\setminus A_j} \) only have zero elements.
- In \( \mathcal{X}_U \), we have \( \phi_A \) draw from \( \{+1, -1\}^{rk} \) uniformly.

We call this data distribution \( \mathcal{D}_{\text{parity}} \).

**Assumption E.19 (Parity Functions. Recap of Assumption 4.5).** Let \( 8 \leq \tau \leq d \) be a parameter that will control our final error guarantee. Assume \( k \) is an odd number and:

\[
k \geq \Omega(\tau \log d), \quad d \geq rk + \Omega(\tau \tau \log d), \quad p_o = O \left( \frac{rk}{d - rk} \right), \quad p_A \geq \frac{1}{d}.
\]

(376)

**Remark E.20.** The assumptions require \( k, d, \) and \( p_A \) to be sufficiently large so as to provide enough large signals for learning. When \( p_o = \Theta\left( \frac{rk}{d - rk} \right) \) means that the signal noise ratio is constant: the expected norm of \( \phi_A \) and that of \( \phi_{A^\perp} \) are comparable.

To apply our framework, again we only need to compute the parameters in the Gradient Feature set and the corresponding optimal approximation loss. To this end, we first define the gradient features: For all \( j \in [r] \), let

\[
D_j = \frac{\sum_{l \in A_j} \mathbf{M}_l}{\| \sum_{l \in A_j} \mathbf{M}_l \|_2}.
\]

(377)

**Remark E.21.** Our data distribution is symmetric, which means for any \( \phi \in \mathbb{R}^d \):

- \( -y = g^*(-\phi_A) \) and \( -x = \mathbf{M}(-\phi) \),
- \( \mathbb{P}(\phi) = \mathbb{P}(-\phi) \),
- \( \mathbb{E}_{\mathbb{E}(\mathbf{x}, y)|x}\|y\| = 0 \).

Below, we define a sufficient condition that randomly initialized weights will fall in nice gradients set after the first gradient step update.

**Definition E.22 (Parity Functions: Subset of Nice Gradients Set).** Recall \( w_i^{(0)} \) is the weight for the \( i \)-th neuron at initialization. For all \( j \in [r] \), let \( S_{D_j,\text{Sure}} \subseteq [m] \) be those neurons that satisfy

\[
\left\langle w_i^{(0)}, D_j \right\rangle \geq \frac{C_{\text{Sure}} - 1}{\sqrt{r}} \mathbf{b}_i,
\]

53
We show the important Lemma E.23 first and defer other Lemmas after it.

Thus, we have
\[ \left\| w_i^{(0)} \right\|_2 \leq \Theta(\sqrt{r}k\sigma_w), \]
and
\[ \left\| P_A w_i^{(0)} \right\|_2 \leq \Theta(\sqrt{d-r}k\sigma_w), \]
where \( P_A, P_{A^+} \) are the projection operator on the space \( M_A \) and \( M_{A^+} \).

### E.4.2 Parity Functions: Feature Learning

We show the important Lemma E.23 first and defer other Lemmas after it.

**Lemma E.23** (Parity Functions: Gradient Feature Set. Part statement of Lemma 4.7). Let \( C_{\text{Sure},1} = \frac{3}{2}, C_{\text{Sure},2} = \frac{1}{2}, \hat{b} = C_b\sqrt{r}k\log d\sigma_w \), where \( C_b \) is a large enough universal constant. For the parity setting, we have \((D_j, +1), (D_j, -1) \in S_{p,\gamma,B_G}\) for all \( j \in [r] \), where

\[
p = \Theta \left( \frac{1}{\sqrt{tr} \log d \cdot d(9C_b^2 \pi r/s)} \right), \quad \gamma = \frac{1}{d^{r-2}}, \quad B_G = \sqrt{k}p_A - O \left( \frac{\sqrt{k}}{d^r} \right). \tag{378}
\]

**Proof of Lemma E.23.** Note that for all \( l \in [d] \), we have \( M_l^\top x = \phi_l \). For all \( j \in [r] \), by Lemma E.26, for all \( i \in S_{D_j,\text{Sure}} \), when \( \gamma = \frac{1}{d^{r-2}} \),

\[
\begin{align*}
& \left\| G(w_i^{(0)}, b_j), D_j \right\| - (1 - \gamma)\|G(w_i^{(0)}, b_i)\|_2 \\
= & \left\| G(w_i^{(0)}, b_i), \sum_{l \in [d]} M_l \frac{M_l^\top x}{\sqrt{k}} \right\| - (1 - \gamma)\|G(w_i^{(0)}, b_i)\|_2 \\
\geq & \sqrt{k}p_A - O \left( \frac{\sqrt{k}}{d^r} \right) - \left(1 - \frac{1}{d^{r-2}}\right) \left( \sqrt{k}p_A + \sum_{l \in [d]} O \left( \frac{1}{d^r} \right) \right) \\
\geq & \sqrt{k}p_A - O \left( \frac{\sqrt{k}}{d^r} \right) - \left(1 - \frac{1}{d^{r-2}}\right) \left( \sqrt{k}p_A + O \left( \frac{1}{d^r} \right) \right) \\
\geq & \frac{\sqrt{k}p_A}{d^{r-2}} - O \left( \frac{\sqrt{k}}{d^r} \right) - O \left( \frac{1}{d^r} \right) \\
> & 0.
\end{align*}
\]

Thus, we have \( G(w_i^{(0)}, b_i) \in C_{D_j,\gamma} \) and \( \sqrt{k}p_A - O \left( \frac{\sqrt{k}}{d^r} \right) \leq \|G(w_i^{(0)}, b_i)\|_2 \leq \sqrt{k}p_A + O \left( \frac{1}{d^{r-2}} \right) \), \( b_i \mid b \rangle = +1 \). Thus, by Lemma E.24, we have

\[
\Pr_{w,b} \left[ G(w,b) \in C_{D_j,\gamma} \text{ and } \|G(w,b)\|_2 \geq B_G \text{ and } \frac{b_i}{|b|} = +1 \right] \geq \Pr \left[ i \in S_{D_j,\text{Sure}} \right] \geq p. \tag{385}
\]

Thus, \((D_j, +1) \in S_{p,\gamma,B_G} \). Since \( \mathbb{E}_{(x,y)}[y|x] = 0 \), by Lemma E.2 and considering \( i \in [2m] \setminus [m] \), we have \((D_j, -1) \in S_{p,\gamma,B_G} \). We finish the proof.

Below are Lemmas used in the proof of Lemma E.23. In Lemma E.24, we calculate \( p \) used in \( S_{p,\gamma,B_G} \).

**Lemma E.24** (Parity Functions: Geometry at Initialization. Lemma B.2 in \cite{7}). Assume the same conditions as in Lemma E.23, recall for all \( i \in [m] \), \( w_i^{(0)} \sim \mathcal{N}(0, \sigma_w^2 I_{d \times d}) \), over the random initialization, we have for all \( i \in [m], j \in [r] \),

\[
\Pr \left[ i \in S_{D_j,\text{Sure}} \right] \geq \Theta \left( \frac{1}{\sqrt{tr} \log d \cdot d(9C_b^2 \pi r/s)} \right). \tag{388}
\]
Proof of Lemma E.24. For every $i \in [m], j, j' \in [r], j \neq j'$, by Lemma F.6,
\begin{align*}
p_1 &= \Pr \left[ \langle w_i^{(0)}, D_j \rangle \geq \frac{C_{\text{Sure},1}}{\sqrt{k}} b_j \right] = \Theta \left( \frac{1}{\sqrt{\tau r \log d \cdot d(9C_b^2 \tau r/8)} } \right) \quad (389) \\
p_2 &= \Pr \left[ \langle w_i^{(0)}, D_{j'} \rangle \geq \frac{C_{\text{Sure},2}}{\sqrt{k}} b_j \right] = \Theta \left( \frac{1}{\sqrt{\tau r \log d \cdot d(9C_b^2 \tau r/8)} } \right). \quad (390)
\end{align*}

On the other hand, if $X$ is a $\chi^2(k)$ random variable, by Lemma F.5, we have
\[ \Pr(X \geq k + 2\sqrt{kx} + 2x) \leq e^{-x}. \quad (391) \]

Therefore, by assumption $rk \geq \Omega(\tau r \log d), d - rk \geq \Omega(\tau r \log d)$, we have
\begin{align*}
\Pr \left( \frac{1}{\sigma_w^2} \left\| P_{A} w_i^{(0)} \right\|_2^2 \geq rk + 2\sqrt{(9C_b^2 \tau r/8 + 2)rk \log d + 2 (9C_b^2 \tau r/8 + 2) \log d} \right) &\leq O \left( \frac{1}{d^2 \cdot d(9C_b^2 \tau r/8)} \right). \quad (392) \\
\Pr \left( \frac{1}{\sigma_w^2} \left\| P_{A} w_i^{(0)} \right\|_2^2 \geq (d - rk) + 2\sqrt{(9C_b^2 \tau r/8 + 2)(d - rk) \log d + 2 (9C_b^2 \tau r/8 + 2) \log d} \right) &\leq O \left( \frac{1}{d^2 \cdot d(9C_b^2 \tau r/8)} \right). \quad (394)
\end{align*}

Thus, by union bound, and $D_1, \ldots, D_r$ being orthogonal with each other, we have
\begin{align*}
\Pr \left[ i \in S_{D_j, \text{Sure}} \right] &\geq p_1(1 - p_2)^{r-1} - O \left( \frac{1}{d^2 \cdot d(9C_b^2 \tau r/8)} \right) \quad (395) \\
&= \Theta \left( \frac{1}{\sqrt{\tau r \log d \cdot d(9C_b^2 \tau r/8)} } \cdot \left( 1 - \frac{r}{\sqrt{\tau r \log d \cdot d(9C_b^2 \tau r/8)} } \right) \right) \quad (396) \\
&\quad - O \left( \frac{1}{d^2 \cdot d(9C_b^2 \tau r/8)} \right) \quad (397) \\
&= \Theta \left( \frac{1}{\sqrt{\tau r \log d \cdot d(9C_b^2 \tau r/8)} } \right). \quad (398)
\end{align*}

In Lemma E.25, we compute the activation pattern for the neurons in $S_{D_j, \text{Sure}}$.

Lemma E.25 (Parity Functions: Activation Pattern). Assume the same conditions as in Lemma E.23, for all $j \in [r], i \in S_{D_j, \text{Sure}}$, we have

1. When $x \sim X$, we have
   \[ \Pr_{x \sim X} \left[ \sum_{i \in A^+} \langle w_i^{(0)}, M_i \phi_i \rangle \geq t \right] \leq \exp \left( -\frac{t^2}{\Theta(rk\sigma_w^2)} \right). \quad (399) \]

2. When $x \sim X_U$, we have
   \[ \Pr_{x \sim X_U} \left[ \sum_{i \in A^+} \langle w_i^{(0)}, M_i \phi_i \rangle \geq t \right] \leq \exp \left( -\frac{t^2}{\Theta(rk\sigma_w^2)} \right). \quad (400) \]

3. When $x \sim X_U$, the activation probability satisfies,
   \[ \Pr_{x \sim X_U} \left[ \sum_{i \in [d]} \langle w_i^{(0)}, M_i \phi_i \rangle - b_i \geq 0 \right] \leq O \left( \frac{1}{d^r} \right). \quad (401) \]

55
(4) When \( x \sim X_{j,+} \), the activation probability satisfies,

\[
\Pr_{x \sim X_{j,+}} \left[ \sum_{l \in [d]} \langle w_i^{(0)}, M_l \phi_l \rangle - b_i \geq 0 \right] \geq 1 - O \left( \frac{1}{d^\tau} \right). \tag{402}
\]

(5) For all \( j' \neq j, j' \in [r], s \in \{+,-\}, \) when \( x \sim X_{j',s} \) or \( x \sim X_{j,-} \), the activation probability satisfies,

\[
\Pr \left[ \sum_{l \in [d]} \langle w_i^{(0)}, M_l \phi_l \rangle - b_i \geq 0 \right] \leq O \left( \frac{1}{d^\tau} \right). \tag{403}
\]

**Proof of Lemma E.25.** For the first statement, when \( x \sim X \), note that \( \langle w_i^{(0)}, M_l \phi_l \rangle \) is a mean-zero sub-Gaussian random variable with sub-Gaussian norm \( \Theta \left( \| \langle w_i^{(0)}, M_l \phi_l \rangle \|_{\sqrt{p_o}} \right) \).

\[
\Pr_{x \sim X} \left[ \sum_{l \in A^+} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle \geq t \right] = \Pr_{x \sim X} \left[ \sum_{l \in A^+} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle \geq t \right] \leq \exp \left( -\frac{t^2}{\sum_{l \in A^+} \Theta \left( \| \langle w_i^{(0)}, M_l \phi_l \rangle \|_{\sqrt{p_o}} \right) } \right) \tag{404}
\]

\[
\leq \exp \left( -\frac{t^2}{\Theta((d-rk)\sigma_w^2 p_o)} \right) \leq \exp \left( -\frac{t^2}{\Theta(rk\sigma_w^2)} \right), \tag{405}
\]

where the inequality follows general Hoeffding’s inequality.

For the second statement, when \( x \sim X_U \), by Hoeffding’s inequality,

\[
\Pr_{x \sim X_U} \left[ \sum_{l \in A} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle \geq t \right] = \Pr_{x \sim X_U} \left[ \sum_{l \in A} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle \geq t \right] \leq 2 \exp \left( -\frac{t^2}{2 \sum_{l \in A} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle^2 } \right) \leq \exp \left( -\frac{t^2}{\Theta(rk\sigma_w^2)} \right), \tag{406}
\]

In the proof of the third to the last statement, we need \( \tilde{b} = C_b \sqrt{r k \log d \sigma_w} \), where \( C_b \) is a large enough universal constant.

For the third statement, when \( x \sim X_U \), by union bound and previous statements,

\[
\Pr_{x \sim X_U} \left[ \sum_{l \in [d]} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle - b_i \geq 0 \right] \leq \Pr_{x \sim X_U} \left[ \sum_{l \in A} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle \geq \frac{b_i}{2} \right] + \Pr_{x \sim X_U} \left[ \sum_{l \in A^+} \left\langle w_i^{(0)}, M_l \phi_l \right\rangle \geq \frac{b_i}{2} \right] \leq O \left( \frac{1}{d^\tau} \right). \tag{411}
\]
For the forth statement, when \( x \sim X_{j,+} \), by \( C_{\text{Sure},1} \geq \frac{3}{2} \) and previous statements,

\[
\Pr_{x \sim X_{j,+}} \left[ \sum_{l \in [d]} \langle w_i^{(0)}, M_l \phi_l \rangle - b_i \geq 0 \right]
\]

\[
= \Pr_{x \sim X_{j,+}} \left[ \sum_{l \in A_j} \langle w_i^{(0)}, M_l \phi_l \rangle + \sum_{l \in A \setminus A_j} \langle w_i^{(0)}, M_l \phi_l \rangle + \sum_{l \in A^+} \langle w_i^{(0)}, M_l \phi_l \rangle \geq b_i \right]
\]

\[
\geq \Pr_{x \sim X_{j,+}} \left[ \sum_{l \in A^+} \langle w_i^{(0)}, M_l \phi_l \rangle \geq (1 - C_{\text{Sure},1}) b_i \right]
\]

\[
\geq \Pr_{x \sim X_{j,+}} \left[ \sum_{l \in A^+} \langle w_i^{(0)}, M_l \phi_l \rangle \geq -\frac{b_i}{2} \right]
\]

\[
\geq 1 - O \left( \frac{1}{d^r} \right).
\]

For the last statement, we prove similarly by \( 0 < C_{\text{Sure},2} \leq \frac{1}{2} \).

Then, Lemma E.26 gives gradients of neurons in \( S_{D_j, \text{Sure}} \). It shows that these gradients are highly aligned with \( D_j \).

**Lemma E.26** (Parity Functions: Feature Emergence). Assume the same conditions as in Lemma E.23, for all \( j \in [r], i \in S_{D_j, \text{Sure}} \), we have the following holds:

(1) For all \( l \in A_j \), we have

\[
p_A - O \left( \frac{1}{d^r} \right) \leq \mathbb{E}_{(x,y)} \left[ y \sigma'( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] \leq p_A + O \left( \frac{1}{d^r} \right).
\]

(2) For all \( l \in A_{j'} \), any \( j' \neq j, j' \in [r] \), we have

\[
\left| \mathbb{E}_{(x,y)} \left[ y \sigma'( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] \right| \leq O \left( \frac{1}{d^r} \right).
\]

(3) For all \( l \in A^+ \), we have

\[
\left| \mathbb{E}_{(x,y)} \left[ y \sigma'( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] \right| \leq O \left( \frac{1}{d^r} \right).
\]

**Proof of Lemma E.26.** For all \( l \in [d] \), we have

\[
\mathbb{E}_{(x,y)} \left[ y \sigma'( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] = p_A \sum_{l \in [r]} \left( \mathbb{E}_{x \sim X_{j,+}} \left[ \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] - \mathbb{E}_{x \sim X_{j,-}} \left[ \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] \right) \right.
\]

\[
+ (1 - 2p_A) \mathbb{E}_{x \sim X_U} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] \right.
\]

For the first statement, for all \( l \in A_j \), by Lemma E.25 (3) and (4), we have

\[
\mathbb{E}_{(x,y)} \left[ y \sigma'( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] = p_A \left( \mathbb{E}_{x \sim X_{j,+}} \left[ \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] + \mathbb{E}_{x \sim X_{j,-}} \left[ \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] \right) \right.
\]

\[
+ (1 - 2p_A) \mathbb{E}_{x \sim X_U} \left[ y \sigma' \left( \langle w_i^{(0)}, x \rangle - b_i \rangle \phi_l \right] \right.
\]

\[
\geq p_A \left( 1 - O \left( \frac{1}{d^r} \right) \right) - O \left( \frac{1}{d^r} \right)
\]

\[
\geq p_A - O \left( \frac{1}{d^r} \right),
\]
and we also have

\[ \mathbb{E}_{(x,y)} \left[ y\sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] = p_A \left( \mathbb{E}_{X \sim X_{i+}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \right] + \mathbb{E}_{X \sim X_{i-}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \right] \right) + (1 - 2rp_A)\mathbb{E}_{X \sim X_U} \left[ y\sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] \]

\[ \leq p_A + O \left( \frac{1}{d^2} \right). \]

Similarly, for the second statement, for all \( l \in A_j' \), any \( j' \neq j \in [r] \), by Lemma E.25 (3) and (5), we have

\[ \left| \mathbb{E}_{(x,y)} \left[ y\sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] \right| \leq p_A \left( \mathbb{E}_{X \sim X_{j',+}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \right] + \mathbb{E}_{X \sim X_{j',-}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \right] \right) + O \left( \frac{1}{d^2} \right) \]

\[ \leq O \left( \frac{1}{d^2} \right). \]

For the third statement, for all \( l \in A_k' \), by Lemma E.25 (3), (4), (5), we have

\[ \left| \mathbb{E}_{(x,y)} \left[ y\sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] \right| \leq p_A \sum_{l \in [r]} \left[ \mathbb{E}_{X \sim X_{l,+}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] - \mathbb{E}_{X \sim X_{l,-}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] \right] + O \left( \frac{1}{d^2} \right) \]

\[ \leq p_A \left[ \mathbb{E}_{X \sim X_{j,+}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] - \mathbb{E}_{X \sim X_{j,-}} \left[ \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right] \right] + O \left( \frac{1}{d^2} \right) \]

\[ \leq p_A \left[ \left( 1 - \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right) - \mathbb{E}_{X \sim X_{j,-}} \left[ \left( 1 - \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right) \right] \right] + O \left( \frac{1}{d^2} \right) \]

\[ = p_A \left[ \left( 1 - \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right) - \mathbb{E}_{X \sim X_{j,-}} \left[ \left( 1 - \sigma' \left( \left< w_i^{(0)}, x \right> - b_i \right) \phi_i \right) \right] \right] + O \left( \frac{1}{d^2} \right) \]

\[ \leq O \left( \frac{1}{d^2} \right), \]

where the second inequality follows \( 2rp_A \leq 1 \) and the third inequality follows the triangle inequality.

\[ \square \]

**E.4.3 Parity Functions: Final Guarantee**

**Lemma E.27** (Parity Functions: Existence of Good Networks. Part statement of Lemma 4.7). Assume the same conditions as in Lemma E.23. Define

\[ f^*(x) = \sum_{j=1}^{r} \sum_{i=0}^{k} (-1)^{i+1} \sqrt{k} \cdot \left[ \sigma \left( \left< D_j, x \right> - \frac{2i - k - 1}{\sqrt{k}} \right) - 2\sigma \left( \left< D_j, x \right> - \frac{2i - k}{\sqrt{k}} \right) + \sigma \left( \left< D_j, x \right> - \frac{2i - k + 1}{\sqrt{k}} \right) \right]. \]

For \( D_{parity} \) setting, we have \( f^* \in \mathcal{F}_{d,3r(k+1),B_F,S_F,B_G} \), where \( B_F = (B_{a1}, B_{a2}, B_b) = (2\sqrt{k}, 2\sqrt{rk(k+1)}, \frac{k+1}{\sqrt{k}}) \), \( p = \Theta \left( \frac{1}{\sqrt{(r+3)\log d \cdot \frac{\mu_0}{\gamma_0} \gamma}} \right) \), \( \gamma = \frac{1}{\alpha \Theta^2}, B_G = \sqrt{k}p_A - O \left( \frac{\sqrt{k}}{\delta^2} \right) \)

and \( B_{x1} = \sqrt{d}, B_{x2} = d \). We also have \( \text{OPT}_{d,3r(k+1),B_F,S_F,B_G} = 0 \).
Algorithm 1 uses hinge loss with $D$. We denote this data distribution as $\sigma$. We follow the initialization and training dynamic in [15].

Assumption E.28 (Uniform Parity Functions)

Let $f$ be a parity function. We give a comparison. We will provide Theorem E.40 as an alternative version that trains both layers.

To apply our framework, again we only need to compute the parameters in the Gradient Feature set and the corresponding optimal approximation loss. To this end, we first define the gradient features: let

$$D = \frac{\sum_{t \in A} M_t}{\| \sum_{t \in A} M_t \|_2}. \quad (445)$$

We follow the initialization and training dynamic in [15].

Proof of Lemma E.27. We can check $B_{x_1} = \sqrt{d}, B_{x_2} = d$ by direct calculation. By Lemma E.23, we have $f^* \in \mathcal{F}_{d,3r(k+1),B_F,S_{\gamma,B_G}}$. We note that

$$\sigma \left( \langle D_j, x \rangle - \frac{2i-k-1}{\sqrt{k}} \right) - 2\sigma \left( \langle D_j, x \rangle - \frac{2i-k}{\sqrt{k}} \right) + \sigma \left( \langle D_j, x \rangle - \frac{2i-k+1}{\sqrt{k}} \right)$$

is a bump function for $\langle D_j, x \rangle$ at $\frac{2i-k}{\sqrt{k}}$. We can check that $\langle yf^*(x) \rangle \geq 1$. Thus, we have

$$\text{OPT}_{d,3r(k+1),B_F,S_{\gamma,B_G}} \leq \mathcal{L}_{\text{parity}}(f^*) \leq \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{parity}}} \mathcal{L}(x,y)(f^*) = 0. \quad (444)$$

Theorem 4.8 (Parity Functions: Main Result). Assume Assumption 4.5. For any $\epsilon, \delta \in (0,1)$, when Algorithm 1 uses hinge loss with

$$m = \text{poly} \left( \frac{1}{\delta}, \frac{1}{\epsilon}, \epsilon^d \theta(\tau r, \epsilon), \frac{1}{P_a} \right) \leq \epsilon^d, \quad T = \text{poly} \left( m \right), \quad n = \text{poly} \left( m \right)$$

and proper hyper-parameters, then with probability at least $1 - \delta$, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f_{\Xi}(x)) \neq y] \leq \frac{3r \sqrt{k}}{d^{(r-3)/2}} + \epsilon.$$

Proof of Theorem 4.8. Let $\bar{b} = C_b \sqrt{\tau r k \log d} \sigma_w$, where $C_b$ is a large enough universal constant. By Lemma E.27, we have $f^* \in \mathcal{F}_{d,3r(k+1),B_F,S_{\gamma,B_G}}$, where $B_F = (B_{a_1}, B_{a_2}, B_b) = \left( 2\sqrt{k}, 2\sqrt{\tau k(k+1)}, \frac{k+1}{\sqrt{r}} \right)$, $p = \Theta \left( \frac{1}{\sqrt{\tau r k \log d} \frac{1}{\sqrt{k}}} \right)$, $\gamma = \frac{1}{d^{r/2}}$, $B_G = \sqrt{k}p_a - O \left( \frac{\sqrt{k}}{d^r} \right)$ and $B_{x_1} = \sqrt{d}, B_{x_2} = d$. We also have $\text{OPT}_{d,3r(k+1),B_F,S_{\gamma,B_G}} = 0$.

Adjust $\sigma_w$ such that $\bar{b} = C_b \sqrt{\tau r k \log d} \sigma_w = \Theta \left( \frac{b_{a_1} B_{a_1}^{3/2}}{\sqrt{B_{a_1}}} \right)$. Injecting above parameters into Theorem 3.12, we have with probability at least $1 - \delta$ over the initialization, with proper hyper-parameters, there exists $t \in [T]$ such that

$$\Pr[\text{sign}(f_{\Xi}(x)) \neq y] \leq \frac{2\sqrt{2r} \sqrt{k}}{d^{(r-3)/2}} + O \left( \frac{r B_{a_1} B_{a_2}^{3/2} \left( \log n \right)^{1/4}}{\sqrt{B_G n^{1/4}}} \right) + \epsilon/2 \leq \frac{3r \sqrt{k}}{d^{(r-3)/2}} + \epsilon. \quad \square$$

E.5 Uniform Parity Functions

We consider the sparse parity problem over the uniform data distribution studied in [15]. We use the properties of the problem to prove the key lemma (i.e., the existence of good networks) in our framework and then derive the final guarantee from our theorem of the simple setting (Theorem 3.4). We provide Theorem E.31 as (1) use it as a warm-up and (2) follow the original analysis in [15] to give a comparison. We will provide Theorem E.40 as an alternative version that trains both layers.

Consider the same data distribution in Appendix E.4.1 and Definition E.22 with the following assumptions.

Assumption E.28 (Uniform Parity Functions). We follow the data distribution in Appendix E.4.1. Let $r = 1, p_A = 0, p_o = \frac{1}{2}, M = I_{d \times d}$ and $d \geq 2k^2$, and $k$ is an even number.

We denote this data distribution as $\mathcal{D}_{\text{parity-uniform}}$ setting.

To apply our framework, again we only need to compute the parameters in the Gradient Feature set and the corresponding optimal approximation loss. To this end, we first define the gradient features: let

$$D = \frac{\sum_{t \in A} M_t}{\| \sum_{t \in A} M_t \|_2}. \quad (445)$$

We follow the initialization and training dynamic in [15].
**Algorithm 4** Network Training via Gradient Descent [15]. Special case of Algorithm 2

Theorem E.31 (Uniform Parity Functions: Main Result)

By Lemma E.29, w.h.p., we have for properly chosen hyper-parameters, and the training samples, there exists \( t \geq T \), such that

\[
\mathbb{W}^{(1)} = \mathbb{W}^{(0)} - \eta^{(1)} \left( \nabla_{\mathbb{W}} \mathbb{L}_Z(f_{\mathbb{F}}^{(0)}) + \lambda^{(1)} \mathbb{W}^{(0)} \right)
\]

\[
a^{(1)} = a^{(0)} - \eta^{(1)} \left( \nabla_{a} \mathbb{L}_Z(f_{\mathbb{F}}^{(0)}) + \lambda^{(1)} a^{(0)} \right)
\]

for \( t = 2 \) to \( T \) do

\[
a^{(t)} = a^{(t-1)} - \eta^{(t)} \nabla_{a} \mathbb{L}_Z(f_{\mathbb{F}}^{(t-1)})
\]

end for

Use the notation in Section 5.3 of [92], for every \( S \in [n] \), s.t. \( |S| = k \), we define

\[
\xi_k := \text{Maj}(S) = (-1)^{k-1} \frac{d-1}{(k-1)} \cdot 2^{-(d-1)} \left( \frac{d-1}{d-2} \right).
\]

**Lemma E.29** (Uniform Parity Functions: Existence of Good Networks. Rephrase of Lemma 5 in [15]). For every \( \epsilon, \delta \in (0, 1/2) \), denoting \( \tau = \frac{1}{16k^2 \log(k^2 + 4n d^2/\delta)} \), let \( \eta^{(1)} = \frac{1}{\tau^{1/4}} \). \( \mathbf{W}^{(1)} \) satisfies

\[
\mathbb{L}_{\mathbb{F}}(f_{a^{(1)}, \mathbb{W}^{(1)}, b}) \leq \epsilon.
\]

Additionally, it holds that \( \| \sigma \left( \mathbb{W}^{(1)}^\top x - b \right) \|_\infty \leq d + 1 \).

**Remark E.30.** In [15], they update the bias term in the first gradient step. However, if we check the proof carefully, we can see that the fixed bias still goes through all their analysis.

**E.5.1 Uniform Parity Functions: Final Guarantee**

Considering training by Algorithm 4, we have the following results.

**Theorem E.31** (Uniform Parity Functions: Main Result). Fix \( \epsilon \in (0, 1/2) \) and let \( m \geq \Omega \left( k \cdot 2^k \log(k/\epsilon) \right) \), \( n \geq \Omega \left( k^7 / d \left( k^2 \log(k/\epsilon) \log(d/\epsilon) \log(dm/\epsilon) + \frac{k^3 m d^2}{\epsilon} \right) \right) \), \( d \geq \Omega \left( k^4 \log(kd/\epsilon) \right) \). Let \( \eta^{(1)} = \frac{1}{\tau^{1/4}} \), \( \lambda^{(1)} = \frac{1}{\eta^{(1)}} \), and \( \eta = \eta^{(t)} = \Theta \left( \frac{1}{\tau^{d/4}} \right) \), for all \( t \in \{2, 3, \ldots, T\} \). If \( T \geq \Omega \left( \frac{k^3 m d^2}{\epsilon} \right) \), then training by Algorithm 4 with hinge loss, w.h.p. over the initialization and the training samples, there exists \( t \in [T] \) such that

\[
\text{Pr} \left[ \text{sign}(f_{\mathbb{F}}^{(t)})(x) \neq y \right] \leq \mathbb{L}_{\mathbb{F}}(f_{a^{(1)}, \mathbb{W}^{(1)}, b}) \leq \epsilon.
\]

**Proof of Theorem E.31.** By Lemma E.29, w.h.p., we have for properly chosen hyper-parameters,

\[
\text{OPT}_{\mathbb{W}^{(1)}, b, \mathbb{B}_a, 2} \leq \mathbb{L}_{\mathbb{F}}(f_{a^{(1)}, \mathbb{W}^{(1)}, b}) \leq \epsilon.
\]
We compute the $L$-smooth constant of $\tilde{L}_Z(f(a, w^{(1)}, b))$ to $a$.

\[
\left\| \nabla_a \tilde{L}_Z(f(a_1, w^{(1)}), b) - \nabla_a \tilde{L}_Z(f(a_2, w^{(1)}, b)) \right\|_2^n \leq \frac{L}{n} \sum_{x \in \mathcal{Z}} \left[ \ell'(y f(a_1, w^{(1)}, b)(x)) - \ell'(y f(a_2, w^{(1)}, b)(x)) \right] \sigma(W^{(1)}_T x - b) \right\|_2^n \leq \frac{1}{n} \sum_{x \in \mathcal{Z}} \left[ \left| f(a_1, w^{(1)}, b)(x) - f(a_2, w^{(1)}, b)(x) \right| \sigma(W^{(1)}_T x - b) \right\|_2^n \leq \frac{1}{n} \sum_{x \in \mathcal{Z}} \left[ \left\| a_1 - a_2 \right\|_2 \left\| \sigma(W^{(1)}_T x - b) \right\|_2^n \right]^2 .
\]

By the Lemma E.29, we have $\left\| \sigma(W^{(1)}_T x - b) \right\|_\infty \leq d + 1$. Thus, we have,

\[
L = O \left( \frac{1}{n} \sum_{x \in \mathcal{Z}} \left\| \sigma(W^{(1)}_T x - b) \right\|_2^n \right) \leq O(d^2 m).
\]

This means that we can let $T = \Theta \left( \frac{1}{d^2 m} \right)$ and we will get our convergence result. Note that we have $a^{(1)} = 0$ and $\|a\|_2 = O \left( k \sqrt{k} \right)$. So, if we choose $T \geq \Omega \left( \frac{k}{\eta} \right)$, there exists $t \in [T]$ such that

\[
\tilde{L}_Z(f(a^{(1)}, w^{(1)}, b)) - \tilde{L}_Z(f(a, w^{(1)}, b)) \leq O \left( \frac{L^2 \|a^{(1)} - a\|_2^2}{d^2 m^2} \right) \leq \epsilon/3 .
\]

We also have $\sqrt{\frac{\|\omega\|_2^2 \|W^{(1)}\|_2^2 \|B\|_2^2 + \|b\|_2^2}{n}} \leq \frac{\epsilon}{3}$. Then our theorem gets proved by Theorem 3.4. \hspace{1cm} \Box

### E.6 Uniform Parity Functions: Alternative Analysis

It is also possible to unify [15] into our general Gradient Feature Learning Framework by mildly modifying the framework in Theorem 3.12. In order to do that, we first need to use a different metric in the definition of gradient features.

#### E.6.1 Modified General Feature Learning Framework for Uniform Parity Functions

**Definition E.32** (Gradient Feature with Infinity Norm). For a unit vector $D \in \mathbb{R}^d$ with $\|D\|_2 = 1$, and a $\gamma_\infty \in (0, 1)$, a direction neighborhood (cone) $C_{D, \gamma_\infty}$ is defined as: $C_{D, \gamma_\infty} := \left\{ w \ : \ \frac{w}{\|w\|_\infty} - D \right\} \leq \gamma_\infty$. Let $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ be random variables drawn from some distribution $W, B$. A Gradient Feature set with parameters $p, \gamma_\infty, B, B_G, B_{G1}$ is defined as:

\[
S_{p, \gamma_\infty, B, B_G, B_{G1}}(W, B) := \left\{ (D, s) \ : \ \Pr_{w, b} \left[ G(w, b) \in C_{D, \gamma_\infty} \right. \left. , B_{G1} \geq \|G(w, b)\|_2 \geq B_G , s = \frac{b}{|b|} \right] \geq \rho \right\} .
\]

When clear from context, write it as $S_{p, \gamma_\infty, B, B_{G1}}$.

**Definition E.33** (Optimal Approximation via Gradient Features with Infinity Norm). The Optimal Approximation network and loss using gradient feature induced networks $F_{d, r, B, S_{p, \gamma_\infty, B, B_{G1}}}$ are defined as:

\[
f^* := \arg\min_{f \in F_{d, r, B, S_{p, \gamma_\infty, B, B_{G1}}}} \mathcal{L}_D(f), \quad \text{OPT}_{d, r, B, S_{p, \gamma_\infty, B, B_{G1}}} := \min_{f \in F_{d, r, B, S_{p, \gamma_\infty, B, B_{G1}}}} \mathcal{L}_D(f).
\]

We consider the data distribution in Appendix E.4.1 with Assumption E.28, i.e., $D_{\text{parity-uniform}}$ in Appendix E.5. Note that with this dataset, we have $\|x\|_\infty \leq B_{x, \infty}$. We use the following unbiased initialization:

- for $i \in \{1, \ldots, m\}$: $a^{(0)}_i \sim \mathcal{N}(0, \sigma^2_0)$, $w^{(0)}_i \sim \{\pm 1\}^d$, $b_i = \hat{b} \leq 1$,
- for $i \in \{m + 1, \ldots, 2m\}$: $a^{(0)}_i = -a^{(0)}_{i-m}$, $w^{(0)}_i = -w^{(0)}_{i-m}$, $b_i = -b_{i-m}$,
- for $i \in \{2m + 1, \ldots, 4m\}$: $a^{(0)}_i = -a^{(0)}_{i-2m}$, $w^{(0)}_i = w^{(0)}_{i-2m}$, $b_i = b_{i-2m}$.
Let $\nabla_i$ denote the gradient of the $i$-th neuron $\nabla_w \mathcal{L}_D(f_{x_0})$. Denote the subset of neurons with nice gradients approximating feature $(D, s)$ as:

$$G_{(D, s), Nice} := \left\{ i \in [2m] : s = \frac{b_i}{\|b_i\|} \left\| \nabla_i - D \right\|_\infty \leq \gamma, \|a_i^{(0)}\|_{B_G} \geq \|\nabla_i\|_2 \geq \|a_i^{(0)}\|_{B_G} \right\}.$$  

**Lemma E.34** (Existence of Good Networks. Modified Version of Lemma 3.14 Under Uniform Parity Setting). Let $\lambda^{(1)} = \frac{1}{m^2}$. For any $B \in (0, B_b)$, let $\sigma_a = \Theta \left( -\ell'(0)\eta^{(1)}B_GB_b \right)$ and $\delta = 2re^{-\sqrt{mp} + \frac{1}{2}}$. Then, with probability at least $1 - \delta$ over the initialization, there exists $\hat{a}_i$'s such that $f_{(\hat{a}, \hat{w}^{(1)}, \hat{b})}(x) = \sum_{i=1}^{4m} \hat{a}_i \sigma \left( \left\langle \hat{w}^{(1)}_i, x \right\rangle - b_i \right)$ satisfies

$$\mathcal{L}_D(f_{(\hat{a}, \hat{w}^{(1)}, \hat{b})}) \leq \frac{B_{s_1}B_{G_1}B_b}{\sqrt{mp}B_GB_b} + \frac{\sqrt{2\log(d)\gamma}}{B} + \text{OPT}_{d, \epsilon, \mathbb{R}^\infty, \hat{w}^{(1)}, \hat{b}}.$$  

and $\|\hat{a}\|_0 = O \left( r(mp)^{\frac{1}{2}} \right)$, $\|\hat{a}\|_2 = O \left( \frac{B_{s_1}B_{b}}{(mp)^{\frac{1}{2}}} \right)$, $\|\hat{a}\|_\infty = O \left( \frac{B_{s_1}B_{b}}{(b(mp))^{\frac{1}{2}}} \right)$.

**Proof of Lemma E.34.** Recall $f^*(x) = \sum_{j=1}^r a_j^* \sigma(\langle w_j^*, x \rangle - b_j^*)$, where $f^* \in \mathcal{F}_{d, \epsilon, r, B^*, B_{G1}, B_G}$ is defined in Definition E.33 and let $s_j^* = \frac{b_j^*}{\|b_j^*\|}$. By Lemma D.3, with probability at least $1 - \delta_1$, $\delta_1 = 2re^{-\sqrt{mp} + \frac{1}{2}}$, for all $j \in [r]$, we have $G_{(w_j^*, s_j^*), Nice} \geq \frac{mp}{B}$. Then for all $i \in G_{(w_j^*, s_j^*), Nice} \subseteq [2m]$, we have $-\ell'(0)\eta^{(1)}G(w_i^{(0)}, b_i) b_j^*/b_i$ only depend on $w_i^{(0)}$ and $b_i$, which is independent of $a_i^{(0)}$. Given Definition E.32, we have

$$-\ell'(0)\eta^{(1)}G(w_i^{(0)}, b_i) b_j^*/b_i \in \left[ -\ell'(0)\eta^{(1)}B_{s_1}B_b -\ell'(0)\eta^{(1)}B_{s_1}B_b \right].$$  

(460)

We split $[r]$ into $\Gamma = \{ j \in [r] : |b_j^*| < B \}$, $\Gamma_- = \{ j \in [r] : b_j^* \leq -B \}$ and $\Gamma_+ = \{ j \in [r] : b_j^* \geq B \}$. Let $\epsilon_a = \frac{B_{s_1}B_b}{\sqrt{mp}B_GB_b}$. Then we know that for all $j \in \Gamma_+ \cup \Gamma_-$, for all $i \in G_{(w_j^*, s_j^*), Nice}$, we have

$$\Pr_{a_i^{(0)} \sim \mathcal{N}(0, \sigma_a^2)} \left[ | -a_i^{(0)} \ell'(0)\eta^{(1)}G(w_i^{(0)}, b_i) b_j^*/b_i - 1 | \leq \epsilon_a \right] \geq \Pr_{a_i^{(0)} \sim \mathcal{N}(0, \sigma_a^2)} \left[ 1 - \epsilon_a \leq \epsilon_a - a_i^{(0)} \ell'(0)\eta^{(1)}G(w_i^{(0)}, b_i) b_j^*/b_i \leq 1 + \epsilon_a \right] \geq \Pr_{g \sim \mathcal{N}(0, 1)} \left[ (1 - \epsilon_a) \frac{B_{s_1}B_b}{\sqrt{mp}B_GB_b} \right] \leq (1 + \epsilon_a) \frac{B_{s_1}B_b}{\sqrt{mp}B_GB_b} \leq \frac{B_{s_1}B_b}{\sqrt{mp}B_GB_b} \leq \frac{1}{\sqrt{mp}}.$$  

(464)

Thus, with probability $\Omega \left( \frac{1}{\sqrt{mp}} \right)$ over $a_i^{(0)}$, we have

$$\left| -a_i^{(0)} \ell'(0)\eta^{(1)}G(w_i^{(0)}, b_i) b_j^*/b_i - 1 \right| \leq \epsilon_a, \quad a_i^{(0)} = O \left( \frac{B_{s_1}B_b}{\sqrt{mp}B_GB_b} \right).$$  

(467)
Similarly, for \( j \in \Gamma \), for all \( i \in G_{(w_{(j)}^*, s_{(j)})}, \) nice, with probability \( \Omega\left(\frac{1}{\sqrt{mp}}\right) \) over \( a_i^{(0)} \), we have

\[
-\frac{a_i^{(0)}}{b} \ell'(0) \eta^{(1)} \|G(w_i^{(0)}, b_i)\|_2 \frac{B_e}{b} - 1 \leq \epsilon_a, \quad |a_i^{(0)}| = O\left(\frac{\hat{b}}{-\ell'(0) \eta^{(1)} B_e B_e} \right). \tag{468}
\]

For all \( j \in [r] \), let \( \Lambda_j \subseteq G_{(w_{(j)}^*, s_{(j)})}, \) nice be the set of \( i \)'s such that condition Equation (467) or Equation (468) are satisfied. By Chernoff bound and union bound, with probability at least \( 1 - \delta_2 \), \( \delta_2 = re^{-\sqrt{mp}} \), for all \( j \in [r] \) we have \( |\Lambda_j| \geq \Omega(\sqrt{mp}) \). We have for \( \forall j \in \Gamma_+ \cup \Gamma_- \), \( \forall i \in \Lambda_j \),

\[
\frac{|b_j^*|}{b} \left\langle w_i^{(1)}, x \right\rangle - \left\langle w_j^*, x \right\rangle \leq \left\langle -\frac{a_i^{(0)}}{b} \ell'(0) \eta^{(1)} \|G(w_i^{(0)}, b_i)\|_2 \frac{b_j^*}{\|b_j^*\|_2} \left\langle w_i^{(1)}, x \right\rangle - \frac{w_i^{(1)}}{\|w_i^{(1)}\|_2}, x \right\rangle + \left\langle \frac{w_i^{(1)}}{\|w_i^{(1)}\|_2} - w_j^*, x \right\rangle \leq \epsilon_a \|x\|_2 + \sqrt{2 \log(d) d \gamma} \tag{469}
\]

With probability \( 1 - \frac{1}{d^2} \) by Hoeffding’s inequality. Similarly, for \( \forall j \in \Gamma, \forall i \in \Lambda_j \),

\[
\left| \frac{B_e}{b} \left\langle w_i^{(1)}, x \right\rangle - \left\langle w_j^*, x \right\rangle \right| \leq \epsilon_a \|x\|_2 + \sqrt{2 \log(d) d \gamma} \tag{470}
\]

If \( i \in \Lambda_j, j \in \Gamma_+ \cup \Gamma_- \), set \( \tilde{a}_i = a_j^{(0)} \frac{|b_j^*|}{|\Lambda_j|b} \), if \( i \in \Lambda_j, j \in \Gamma \), set \( \tilde{a}_i = a_j^{(0)} \frac{b_j}{|\Lambda_j|b} \), otherwise set \( \tilde{a}_i = 0 \). we have \( \|\tilde{a}\|_0 = O\left( (r(mp)^{\frac{1}{2}}) \right), \|\tilde{a}\|_2 = O\left( \frac{B_{a_2} B_{a_1}}{b(m)p^2} \right), \|\tilde{a}\|_\infty = O\left( \frac{B_{a_1} B_{a_2}}{b(m)p^2} \right) \).
Finally, we have
\[ L_D(f(a, w^{(1)}_j, b)) - L_D(f^*) + L_D(f^*) \]
\[ \leq E_{(x,y)} \left[ |f(a, w^{(1)}_j, b) - f^*(x)| \right] + L_D(f^*) \]
\[ \leq E_{(x,y)} \left[ \sum_{i=1}^{m} a_i \sigma \left( \langle w^{(1)}_i, x \rangle - \hat{b} \right) + \sum_{i=m+1}^{2m} a_i \sigma \left( \langle w^{(1)}_i, x \rangle + \hat{b} \right) - \sum_{j=1}^{r} a_j \sigma \left( \langle w^{(1)}_j, x \rangle - b^*_j \right) \right] + L_D(f^*) \]
\[ \leq E_{(x,y)} \left[ \sum_{j \in \Gamma_+} a_j \frac{1}{|\Lambda_j|} \left| \frac{b_j^*}{b} \right| \sigma \left( \langle w^{(1)}_j, x \rangle - \hat{b} \right) - \sigma \left( \langle w^*_j, x \rangle - b_j^* \right) \right] + L_D(f^*) \]
\[ \leq E_{(x,y)} \left[ \sum_{j \in \Gamma_-} a_j \frac{1}{|\Lambda_j|} \left| \frac{b_j^*}{b} \right| \sigma \left( \langle w^{(1)}_j, x \rangle + \hat{b} \right) - \sigma \left( \langle w^*_j, x \rangle - b_j^* \right) \right] + L_D(f^*) \]
\[ \leq r \|a^*\|_{\infty} (\epsilon B_{\|x\|_2} + \sqrt{2 \log(d) d_{\gamma\infty}} + |\Gamma| \|a^*\|_{\infty} B_{\epsilon}) + L_D(f^*) \]
\[ \leq r B_{\|x\|_2} + \sqrt{2 \log(d) d_{\gamma\infty}} + |\Gamma| B_{\|x\|_2} + L_D(f^*) \]

We finish the proof by union bound and \( \delta \geq \delta_1 + \delta_2 + \frac{1}{\delta} \).

Lemma E.35. (Empirical Gradient Concentration Bound for Single Coordinate). For \( i \in [m] \), when \( n \geq (\log(d))^6 \), with probability at least \( 1 - O \left( \exp \left( -n^2 \right) \right) \) over training samples, we have
\[ \left| \frac{\partial E_Z(f_{\hat{x}})}{\partial w_{i,j}} - \frac{\partial E_D(f_{\hat{x}})}{\partial w_{i,j}} \right| \leq O \left( \frac{|a_i| B_{x_{\infty}}}{n^2} \right), \quad \forall j \in [d]. \]

Proof of Lemma E.35. First, we define,
\[ z_{i,j}^{(l)} = \ell'(y^{(l)} f_{\hat{x}}(x^{(l)})) y^{(l)} \left[ \sigma' \left( \langle w_i, x^{(l)} \rangle - b_i \right) \right] \]
\[ - \ell'(y f_{\hat{x}}(x)) y \sigma' \left( \langle w_i, x \rangle - b_i \right) \]
As \( |\ell'(z)| \leq 1, |y| \leq 1, |\sigma'(z)| \leq 1 \), we have \( z_{i,j}^{(l)} \) is zero-mean random variable with \( \left| z_{i,j}^{(l)} \right| \leq 2B_{x_{\infty}} \) as well as \( E \left[ z_{i,j}^{(l)} \right] \leq 4B_{x_{\infty}}^2 \). Then by Bernstein Inequality, for \( 0 < z < 2B_{x_{\infty}} \), we have
\[ \Pr \left( \left| \frac{\partial E_Z(f_{\hat{x}})}{\partial w_{i,j}} - \frac{\partial E_D(f_{\hat{x}})}{\partial w_{i,j}} \right| \geq |a_i| z \right) \leq \Pr \left( \left| \frac{1}{n} \sum_{l \in [n]} z_{i,j}^{(l)} \right| \geq z \right) \]
\[ \leq \exp \left( -n \cdot \frac{z^2}{8B_{x_{\infty}}} \right). \]
Thus, for some \( i \in [m] \), when \( n \geq (\log(d))^6 \), with probability at least \( 1 - O \left( \exp \Theta \left( -n^{\frac{2}{3}} \right) \right) \), from a union bound over \( j \in [d] \), we have, for all \( j \in [d] \),

\[
\left| \frac{\partial \tilde{G}_Z(f_\xi)}{\partial w_{i,j}} - \frac{\partial L_D(f_\xi)}{\partial w_{i,j}} \right| \leq O \left( \frac{|a_i| B_{z^\infty}}{n^2} \right). \tag{489}
\]

\[ \blacksquare \]

**Lemma E.36** (Existence of Good Networks under Empirical Risk. Modified version of Lemma D.13 Under Uniform Parity Setting). Suppose \( n > \Omega \left( \left( \frac{B}{\sqrt{B_x^2}} + \log \frac{1}{\rho} + \frac{B_{1\infty}}{B_G[l'(0)]} + \frac{B_{L\infty}}{B_{G1[l'(0)]}} \right)^3 + (\log(d))^6 \right) \). Let \( \lambda(1) = \frac{1}{\eta(1)} \). For any \( B_\epsilon \in (0, B_6) \), let \( \sigma_a = \Theta \left( \frac{\rho}{l'(0)[\rho]} B_{G1} B_\epsilon \right) \) and \( \delta = 2 \rho e^{-\sqrt{\rho \epsilon}} + \frac{1}{\eta(1)} \). Then, with probability at least \( 1 - \delta \) over the initialization and training samples, there exists \( \tilde{a}_i \) such that \( f_\tilde{a}(w^{(1)}_{\omega}, b) = \sum_{i=1}^{4m} \tilde{a}_i \sigma \left( \langle w_i^{(1)}, x \rangle - b \right) \) satisfies

\[
L_D(f_\tilde{a}(w^{(1)}_{\omega}, b)) \leq r B_{a1} \left( \frac{2 B_{x2} B_{G1} B_\epsilon}{\sqrt{mp} B_G B_\epsilon} + 2 \log(d) \left( \gamma_{\infty} + O \left( \frac{B_{z^\infty}}{B_G[l'(0)][\rho]} \right) \right) + B_\epsilon \right) + O_{d, r, B_{F}, S_{p, G, a, G, a, G1}},
\]

\[
\text{and} \quad ||\tilde{a}||_0 = O \left( r (mp)^{\frac{1}{2}} \right), \quad ||\tilde{a}||_2 = O \left( \frac{B_{x2} B_{1\infty}}{b(m p)^{\frac{1}{2}}} \right), \quad ||\tilde{a}||_{\infty} = O \left( \frac{B_{a1} B_{b1}}{b(m p)^{\frac{1}{2}}} \right).
\]

**Proof of Lemma E.36.** Denote \( \rho = O \left( \exp \Theta \left( -n^{\frac{2}{3}} \right) \right) \) and \( \beta = O \left( \frac{B_{z^\infty}}{n^2} \right) \). Note that by symmetric initialization, we have \( l'(y f_{\tilde{G}}(x)) = |l'(0)| \) for any \( x \in X \), so that, by Lemma E.35, we have \( |\tilde{G}(w^{(1)\omega}_i, b) - G(w^{(1)\omega}_i, b)| \leq \frac{\beta}{l'(0)} \) with probability at least \( 1 - \rho \). Thus, by union bound, we can see that \( S_{p, G, a, G, a, G1} \subseteq S_{\tilde{G}_{p, G, a, G, a, G1}} \). Consequently, we have \( \text{OPT}_{d, r, B_{F}, S_{p, G, a, G, a, G1}} \leq \text{OPT}_{d, r, B_{F}, S_{\tilde{G}_{p, G, a, G, a, G1}}} \). Exactly follow the proof in Lemma D.4 by replacing \( S_{p, G, a, G, a, G1} \) to \( S_{\tilde{G}_{p, G, a, G, a, G1}} \). Then, we finish the proof by \( \rho \leq \frac{\rho}{(\sqrt{2} - 1) B_{G1}} \).

\[ \blacksquare \]

**Theorem E.37** (Online Convex Optimization under Empirical Risk. Modified version of Theorem D.17 Under Uniform Parity Setting). Consider training by Algorithm 1, and any \( \delta \in (0, 1) \). Assume \( d \log m \geq n \), \( \delta \leq O \left( \frac{1}{\rho^2} \right) \). Set

\[
\sigma_w > 0, \quad \tilde{b} > 0, \quad \eta(1) = \eta, \quad \lambda(1) = \frac{1}{\eta(1)}, \quad \sigma_a = \Theta \left( \frac{\tilde{b}(mp)^{\frac{1}{2}}}{l'(0)[\rho]} \right).
\]

Let \( 0 < T \eta B_{x1} \leq o(1) \) and \( m = \Omega \left( \frac{B}{\sqrt{B_x^2}} + \frac{1}{B_{G1}} + \frac{B_{z^\infty}}{B_G[l'(0)]} \right)^3 \). With probability at least \( 1 - \delta \) over the
Furthermore, for any \( t \in [T] \) such that
\[
\mathcal{L}_D (f_{\mathcal{E}(t)}) \leq \text{OPT} + rB_{a_1} \left( \frac{2\sqrt{2} \sqrt{B_{x1} B_{G1}}}{m p^{\frac{1}{4}}} \sqrt{\frac{B_b}{B_G}} + \sqrt{2 \log (d)} \left( \gamma_{\infty} + O \left( \frac{B_{x_{\infty}}}{B_G |\ell(0)| n^{\frac{1}{2}}} \right) \right) \right)
\]
\[
+ \eta \left( \sqrt{r} B_{a2} B_b T \eta B_{x2} + m \tilde{b} \right) O \left( \frac{\sqrt{\log m B_{x1} (mp)^{\frac{1}{2}}}}{\sqrt{B_b B_G}} + 1 \right) + O \left( \frac{B_{a2}^2 B_b^2}{m \eta T b^2 (mp)^{\frac{1}{2}}} \right)
\]
\[
+ \frac{1}{n^\frac{3}{4}} \eta \left( \eta B_{x1} \left( \frac{\sqrt{r} \log \log \sqrt{B_b B_G}}{\sqrt{B_b B_G}} + \frac{\hat{b}}{B_{x1}} \right) \right)
\cdot \left( \frac{B_x}{\sqrt{B_{x2}} + \tilde{b}} + 2 \right)
\]
\[
+ \frac{1}{n^\frac{3}{4}} \eta \left( m \eta \left( \frac{B_x}{\sqrt{B_{x2}}} \right) + 3 \eta \sqrt{B_{x2}} \right).
\]

Furthermore, for any \( \epsilon \in (0,1) \), set
\[
\hat{b} = \Theta \left( \frac{B_{a2}^2 B_b^2}{\sqrt{r} B_{a1}} \right), \quad m = \Omega \left( \frac{1}{p^{\frac{3}{4}}} \left( r B_{a1} \sqrt{B_{x1} B_{G1}} \sqrt{\frac{B_b}{B_G}} \right)^4 + 1 \right)
\]
\[
\eta = \Theta \left( \frac{\sqrt{r} B_{a2} B_b T \eta B_{x1} \eta B_{x1} \left( \frac{\sqrt{r} \log \log \sqrt{B_b B_G}}{\sqrt{B_b B_G}} + 1 \right) \right)
\]
\[
n = \Omega \left( \frac{m \eta \log m}{\sqrt{B_b B_G}} \right)^3 + \left( \frac{B_x}{\sqrt{B_{x2}}} + \log \frac{T}{p \delta} + 1 \right) \left( \frac{1}{B_G} + \frac{1}{B_{G1}} \right) \left( \frac{B_{x_{\infty}}}{|\ell(0)| n^{\frac{1}{2}}} \right)^{3} \right),
\]

we have there exists \( t \in [T] \) with
\[
\Pr \left[ \text{sign}(f_{\mathcal{E}(t)}(x)) \neq y \right] \leq \mathcal{L}_D (f_{\mathcal{E}(t)}) \leq \text{OPT} + \epsilon.
\]

**Proof of Theorem E.37.** Proof of the theorem and parameter choices remain the same as Theorem D.17 except for setting \( B_x = \sqrt{\frac{B_{a2} B_{b2}}{(mp)^{\frac{1}{2}}} B_{x1} B_{G1}} \) and apply Lemma E.36.

### E.6.2 Feature Learning of Uniform Parity Functions

We denote
\[
g_{i,j} = \mathbb{E}_{(x,y)} \left[ y^{(w_i^{(0)})} \left( \langle w_i^{(0)} , x \rangle \right) - b_i \right] x_j
\]
\[
\xi_k = (-1)^{\frac{k-1}{2}} \frac{n-1}{k} \cdot 2^{-\frac{k-1}{2}} \left( \frac{n-1}{n+1} \right) \left( \frac{n-1}{n+1} \right).
\]

**Lemma E.38** (Uniform Parity Functions: Gradient Feature Learning. Corollary of Lemma 3 in [15]). Assume that \( n \geq 2(k+1)^2 \). Then, the following holds:

If \( j \in A \), then
\[
g_{i,j} = \xi_{k-1} \prod_{l \in A \setminus \{j\}} (w_i^{(0)}).
\]
If \( i \notin A \), then
\[
g_{i,j} = \xi_{k-1} \prod_{l \in A \cup \{j\}} (w_{i,l}^{(0)}).
\]  

Lemma E.39 (Uniform Parity Functions: Existence of Good Networks (Alternative)). Assume the same condition as in Lemma E.38. Define
\[
D = \frac{\sum_{l \in A} M_l}{\|\sum_{l \in A} M_l\|_2}
\]  
and
\[
f^*(x) = \sum_{i=0}^{k} (-1)^i \sqrt{k} \cdot \left[ \sigma \left( (D, x) - \frac{2i - k - 1}{\sqrt{k}} \right) - 2\sigma \left( (D, x) - \frac{2i - k}{\sqrt{k}} \right) + \sigma \left( (D, x) - \frac{2i - k + 1}{\sqrt{k}} \right) \right].
\]

For \( D_{\text{parity-uniform}} \) setting, we have \( f^* \in \mathcal{F}_{d,3(k+1),B_F,S_p,\gamma,\gamma_{G_{f},G_{G_{1}}}} \) where \( B_F = (B_{a_1}, B_{a_2}, B_{b}) = (2\sqrt{k}, 2\sqrt{(k+1)}), \gamma_{\infty} = O\left(\frac{1}{\sqrt{k}}\right), B_{G_{1}} = \Theta(B_{G_{1}}) = \Theta(d^{-k}) \) and \( B_{z_1} = \sqrt{d}, B_{z_2} = d \). We also have \( \OPT_{d,3(k+1),B_F,S_p,\gamma,\gamma_{G_{f},G_{G_{1}}}} = 0 \).

Proof of Lemma E.39. Fix index \( i \), with probability \( p_1 = \Theta(2^{-k}) \), we will have \( w_{i,j}^{(0)} = \sign(a_{i}^{(0)}) \cdot \sign(\xi_{k-1}) \), for all \( j \). For \( w_{i,j}^{(0)} \) that satisfy these conditions, we will have:
\[
\sign(a_{i}^{(0)})g_{i,j} = |\xi_{k-1}|, \quad \forall j \in A
\]
\[
\sign(a_{i}^{(0)})g_{i,j} = |\xi_{k+1}|, \quad \forall j \notin A.
\]

Then by Lemma 4 in [15], we have
\[
\left| \frac{\|\sign(a_{i}^{(0)})G(w_{i}^{(0)}, \hat{b})\| - D}{\|G(w_{i}^{(0)}, \hat{b})\|} \right|_{\infty} \leq \max \left\{ \left| \frac{1}{k} \sqrt{\frac{1}{\pi} + \frac{1}{k}} - \frac{1}{\sqrt{k}} \right|, \left| \frac{1}{(d-k)} \sqrt{\frac{1}{\pi} + \frac{1}{d-k}} \right| \right\}
\]
\[
\leq \frac{\sqrt{k}}{d-k}
\]
and
\[
\left| \|\sign(a_{i}^{(0)})G(w_{i}^{(0)}, \hat{b})\|_{2} - \sqrt{k}|\xi_{k-1}|^2 + (d-k)|\xi_{k+1}|^2 = \Theta(d^{\Theta(k)}) \right.
\]
From here, we can see that if we set \( \gamma_{\infty} = \frac{\sqrt{k}}{d-k}, B_{G_{1}} = B_{G_{1}} = \sqrt{k}|\xi_{k-1}|^2 + (d-k)|\xi_{k+1}|^2, p = p_1 \), we will have \( (D, +1), (D, -1) \in S_p,\gamma_{G_{f},G_{G_{1}}},B_{G_{f},B_{G_{1}}} \) by our symmetric initialization. As a result, we have \( f^* \in \mathcal{F}_{d,3(k+1),B_F,S_p,\gamma,\gamma_{G_{f},G_{G_{1}}}} \). Finally, it is easy to verify that \( f^*(x) = \text{XOR}(x_A) \), thus \( \OPT_{d,3(k+1),B_F,S_p,\gamma,\gamma_{G_{f},G_{G_{1}}}} = 0 \).

Theorem E.40 (Uniform Parity Functions: Main Result (Alternative)). For \( D_{\text{parity-uniform}} \) setting, for any \( \delta \in (0, 1) \) satisfying \( \delta \leq O\left(\frac{1}{\sqrt{k}}\right) \) and for any \( \epsilon \in (0, 1) \) when
\[
m = \text{poly} \left( \log \left( \frac{1}{\delta} \right), \frac{1}{\epsilon}, 2^{\Theta(k)}d \right), T = \Theta\left(d^{\Theta(k)}\right), n = \Theta\left(d^{\Theta(k)}\right)
\]
trained by Algorithm 1 with hinge loss, with probability at least \( 1 - \delta \) over the initialization, with proper hyper-parameters, there exists \( t \in [T] \) such that
\[
\Pr[\text{sign}(f_{\xi(0)}(x)) \neq y] \leq \frac{k^2 \sqrt{\log(d)}}{d-k} + \epsilon.
\]

Proof of Theorem E.40. Plug the values of parameters into Theorem E.37 and directly get the result.
E.7 Multiple Index Model with Low Degree Polynomial

E.7.1 Problem Setup

The multiple-index data problem has been used for studying network learning [18, 32]. We consider proving guarantees for the setting in [32], following our framework. We use the properties of the problem to prove the key lemma (i.e., the existence of good networks) in our framework and then derive the final guarantee from our theorem of the simple setting (Theorem 3.4).

Data Distributions. We draw input from the distribution \( D_X = \mathcal{N}(0, I_{d \times d}) \), and we assume the target function is \( g^*(x) : \mathbb{R}^d \to \mathbb{R} \), where \( g^* \) is a degree \( \tau \) polynomial normalized so that \( \mathbb{E}_{x \sim D_X}[g^*(x)^2] = 1 \).

Assumption E.41. There exists linearly independent vectors \( u_1, \ldots, u_r \) such that \( g^*(x) = g((x, u_1), \ldots, (x, u_r)) \). \( H := \mathbb{E}_{x \sim D_X}[\nabla^2 g^*(x)] \) has rank \( r \), where \( H \) is a Hessian matrix.

Definition E.42. Denote the normalized condition number of \( H \) by

\[
\kappa := \frac{\|H\|}{\sqrt{r}}. \tag{511}
\]

Initialization and Loss. For \( \forall i \in [m] \), we use the following initialization:

\[
a_i^{(0)} \sim \{-1, 1\}, \quad w_i^{(0)} \sim \mathcal{N} \left( 0, \frac{1}{d} I_{d \times d} \right) \quad \text{and} \quad b_i = 0. \tag{512}
\]

For this regression problem, we use mean square loss:

\[
\mathcal{L}_{D_X}(f) = \mathbb{E}_{x \sim D_X} \left[ (f(x) - g^*(x))^2 \right]. \tag{513}
\]

Training Process. We use the following one-step training algorithm for this specific data distribution.

**Algorithm 5** Network Training via Gradient Descent [32]. Special case of Algorithm 2.

Initialize \((a^{(0)}, W^{(0)}, b)\) as in Equation (8) and Equation (512); Sample \( Z \sim D_X^n \)

\[
\rho = \frac{1}{n} \sum_{x \in Z} g^*(x), \quad \beta = \frac{1}{n} \sum_{x \in Z} g^*(x)x
\]

\[
y = g^*(x) - \rho - \beta \cdot x
\]

\[
W^{(1)} = W^{(0)} - \eta^{(1)} \nabla_w \tilde{L}_Z(f_{\Xi^{(0)}}) + \lambda^{(1)} W^{(0)}
\]

Re-initialize \( b_i \sim \mathcal{N}(0, 1) \)

for \( t = 2 \) to \( T \) do

\[
a^{(t)} = a^{(t-1)} - \eta^{(1)} \nabla_a \tilde{L}_Z(f_{\Xi^{(t-1)}})
\]

end for

Lemma E.43 (Multiple Index Model with Low Degree Polynomial: Existence of Good Networks. Rephrase of Lemma 25 in [32]). Assume \( n \geq d^2 r^2 n^2 \left( C_l \log(nmd) \right)^{\tau+1}, d \geq C_d r^{3/2}, \) and \( m \geq r^7 \kappa^{2\tau} (C_l \log(nmd))^{6\tau+1} \) for sufficiently large constants \( C_d, C_l, \) and let \( \eta^{(1)} = \frac{d}{(C_l \log(nmd))^{\tau}} \) and \( \lambda^{(1)} = \frac{1}{\eta^{(1)}}. \) Then with probability \( 1 - \frac{1}{\text{poly}(m,d)} \), there exists \( \tilde{a} \in \mathbb{R}^m \) such that \( f(\tilde{a}, W^{(1)}, b) \) satisfies

\[
\mathcal{L}_{D_X}(f(\tilde{a}, W^{(1)}, b)) \leq O \left( \frac{1}{n} + \frac{r^7 \kappa^{2\tau} (C_l \log(nmd))^{6\tau+1}}{m} \right) \tag{514}
\]

and

\[
\|\tilde{a}\|_2^2 \leq O \left( \frac{r^7 \kappa^{2\tau} (C_l \log(nmd))^{6\tau}}{m} \right). \tag{515}
\]

E.7.2 Multiple Index Model: Final Guarantee

Considering training by Algorithm 5, we have the following results.
Theorem E.44 (Multiple Index Model with Low Degree Polynomial: Main Result). Assume \( n \geq \Omega \left( d^2 r n^2 (C_1 \log(nmd))^r + m \right) \), \( d \geq C_2 k r^{3/4} \), and \( m \geq \Omega \left( \frac{\epsilon}{r^2 n^2 (C_1 \log(nmd))^{r+1}} \right) \) for sufficiently large constants \( C_1, C_2 \). Let \( \eta^{(1)} = \sqrt{\frac{d}{(C_1 \log(nmd))r}} \) and \( \lambda^{(1)} = \frac{1}{\eta^{(1)}} \), and \( \eta = \eta^{(t)} = \Theta(m^{-1}) \), for all \( t \in \{2, 3, \ldots, T\} \). For any \( \epsilon \in (0, 1) \), if \( T \geq \Omega \left( \frac{m^2}{\epsilon^2} \right) \), then with properly set parameters and Algorithm 5, with high probability that there exists \( t \in [T] \) such that

\[
\mathcal{L}_{D,t} f_{(a^{(t)}, W^{(t)}, b)} \leq \epsilon. \tag{516}
\]

Proof of Theorem E.44. By Lemma E.43, we have for properly chosen hyper-parameters,

\[
\text{OPT}_{W^{(t)}, b, B_{a,t}} \leq \mathcal{L}_{D,t} \left( f_{(a^{(t)}, W^{(t)}, b)} \right) \leq O \left( \frac{1}{n} (r^7 k^{2r} (C_1 \log(nmd))^{6r+1}) \right) \tag{517}
\]

\[
\leq \frac{\epsilon}{3}. \tag{518}
\]

We compute the \( L \)-smooth constant of \( \tilde{L}_Z \left( f_{(a^{(t)}, W^{(t)}, b)} \right) \) to \( a \).

\[
\left\| \nabla_a \tilde{L}_Z \left( f_{(a^{(t)}, W^{(t)}, b)} \right) - \nabla_a \tilde{L}_Z \left( f_{(a^{(t)}, W^{(t)}, b)} \right) \right\|_2
\]

\[
= \frac{1}{n} \sum_{x \in Z} \left\| 2 \left( f_{(a^{(t)}, W^{(t)}, b)}(x) - g_* - f_{(a^{(t)}, W^{(t)}, b)}(x) + g_*^+ \sigma(W^{(t)} x - b) \right) \right\|_2 \tag{519}
\]

\[
\leq \frac{1}{n} \sum_{x \in Z} \left\| 2 \left( a_*^+ \sigma(W^{(t)} x - b) - a_*^+ \sigma(W^{(t)} x - b) \right) \right\|_2 \tag{520}
\]

\[
\leq \frac{1}{n} \sum_{x \in Z} \left\| 2 \left| a_1 - a_2 \right|_2 \right\|_2 \left\| \sigma(W^{(t)} x - b) \right\|_2^2 \tag{521}
\]

By the proof of Lemma 25 in [32], we have for \( \forall i \in \{1, 2, \ldots, T\} \), with probability at least \( 1 - \frac{1}{\text{poly}(m,d)} \), \( |\langle w_i, x \rangle| \leq 1 \), with some large polynomial \( \text{poly}(m,d) \). As a result, we have

\[
\frac{1}{n} \sum_{x \in Z} \left\| W^{(t)} x \right\|_2^2 \leq m + \frac{1}{\text{poly}(m,d)} \leq O(m). \tag{523}
\]

Thus, we have,

\[
L = O \left( \frac{1}{n} \sum_{x \in Z} \left\| \sigma(W^{(t)} x - b) \right\|_2^2 \right) \tag{524}
\]

\[
\leq O \left( \frac{1}{n} \sum_{x \in Z} \left\| W^{(t)} x - b \right\|_2^2 \right) \tag{525}
\]

\[
\leq O \left( \frac{1}{n} \sum_{x \in Z} \left\| W^{(t)} x \right\|_2^2 + \|b\|_2^2 \right) \tag{526}
\]

\[
\leq O(m). \tag{527}
\]

This means that we can let \( \eta = \Theta \left( m^{-1} \right) \) and we will get our convergence result. We can bound \( \|a^{(t)}\|_2 \) and \( \|\tilde{a}\|_2 \) by \( \|a^{(t)}\|_2 = O \left( \sqrt{m} \right) \) and \( \|\tilde{a}\|_2 = O \left( \frac{r^7 k^{2r} (C_1 \log(nmd))^{6r}}{m} \right) = O(\epsilon) \). So, if we choose \( T \geq \Omega \left( \frac{m^2}{\epsilon^2} \right) \), there exists \( t \in [T] \) such that \( \tilde{L}_Z \left( f_{(a^{(t)}, W^{(t)}, b)} \right) - \tilde{L}_Z \left( f_{(a^{(t)}, W^{(t)}, b)} \right) \leq O \left( \frac{L\|a^{(t)}\|_2 \|\tilde{a}\|_2}{T} \right) \leq \epsilon/3. \)

We also have \( \sqrt{\frac{\|a^{(t)}\|_2^2 \|W^{(t)}\|_F^2 + \|b\|_2^2}{n}} \leq \frac{\epsilon}{3} \). Then our theorem gets proved by Theorem 3.4. \( \square \)
Discussion. We would like to unify [32], which are very closely related to our framework: their analysis for multiple index data follows the same principle and analysis approach as our general framework, although it does not completely fit into our Theorem 3.12 due to some technical differences. We can cover it with our Theorem 3.4.

Our work and [32] share the same principle and analysis approach. [32] shows that the first layer learns good features by one gradient step update, which can approximate the true labels by a low-degree polynomial function. Then, a classifier (the second layer) is trained on top of the learned first layer which leads to the final guarantees. This is consistent with our framework: we first show that the first layer learns good features by one gradient step update, which can approximate the true labels, and then show a good classifier can be learned on the first layer.

Our work and [32] have technical differences. First, in the second stage, [32] fix the first layer and only update the top layer which is a convex optimization. Our framework allows updates in the first layer and uses online convex learning techniques for the analysis. Second, they consider the square loss (this is used to calculate Hermite coefficients explicitly for gradients, which are useful in the low-degree polynomial function approximation). While in our online convex learning analysis, we need boundedness of the derivative of the loss to show that the first layer weights’ changes are bounded in the second stage. Given the above two technicalities, we analyze their training algorithm (Algorithm 2) which fixes the first layer weights and fits into our Theorem 3.4.

F Auxiliary Lemmas

In this section, we present some Lemmas used frequently.

Lemma F.1 (Lemmas on Gradients).

\[
\nabla_w L_{(x,y)}(f_{\Xi}) = \left[ \frac{\partial L_{(x,y)}(f_{\Xi})}{\partial w_1}, \ldots, \frac{\partial L_{(x,y)}(f_{\Xi})}{\partial w_i}, \ldots, \frac{\partial L_{(x,y)}(f_{\Xi})}{\partial w_{4m}} \right],
\]

\[
\frac{\partial L_{(x,y)}(f_{\Xi})}{\partial w_i} = a_i \ell'(y f_{\Xi}(x)) y \sigma'((w_i, x) - b_i) x,
\]

\[
\nabla_w L_D(f_{\Xi}) = \left[ \frac{\partial L_D(f_{\Xi})}{\partial w_1}, \ldots, \frac{\partial L_D(f_{\Xi})}{\partial w_i}, \ldots, \frac{\partial L_D(f_{\Xi})}{\partial w_{4m}} \right],
\]

\[
\frac{\partial L_D(f_{\Xi})}{\partial w_i} = a_i E_{(x,y)} [\ell'(y f_{\Xi}(x)) y \sigma'((w_i, x) - b_i)] x,
\]

\[
\frac{\partial L_D(f_{\Xi})}{\partial a_i} = E_{(x,y)} [\ell'(y f_{\Xi}(x)) y \sigma((w_i, x) - b_i)].
\]

Proof. These can be verified by direct calculation. \qed

Lemma F.2 (Property of Symmetric Initialization). For any \( x \in \mathbb{R}^d \), we have \( f_{\Xi(0)}(x) = 0 \). For all \( i \in [2m] \), we have \( w_i^{(1)} = -w_{i+2m}^{(1)} \). When input data is symmetric, i.e., \( E_{(x,y)}[y x] = 0 \), for all \( i \in [m] \), we have \( w_i^{(1)} = w_{i+m}^{(1)} \).

Proof of Lemma F.2. By symmetric initialization, we have \( f_{\Xi(0)}(x) = 0 \). For all \( i \in [2m] \), we have

\[
w_i^{(1)} = -\eta^{(1)} \ell'(0) a_i^{(0)} E_{(x,y)} [y \sigma'((w_i^{(0)}, x) - b_i)] x
\]

\[
= \eta^{(1)} \ell'(0) a_i^{(0)} E_{(x,y)} [y \sigma'((w_i^{(0)}, x) - b_i)] x
\]

\[
= -w_{i+2m}^{(1)}.
\]
When \( \mathbb{E}_{(x,y)}[g'] = 0 \), for all \( i \in [m] \), we have

\[
\mathbf{w}_i^{(1)} = -\eta^{(1)} \mathbf{a}_i^{(0)} \mathbb{E}_{(x,y)} \left[ y g' \left( \mathbf{w}_i^{(0)} \cdot \mathbf{x} - \mathbf{b}_i \right) \right] \tag{536}
\]

\[
= \eta^{(1)} \mathbf{a}_i^{(0)} \mathbb{E}_{(x,y)} \left[ y g' \left( -\mathbf{w}_i^{(0)} \cdot \mathbf{x} + \mathbf{b}_{i+m} \right) \right] \tag{537}
\]

\[
= \eta^{(1)} \mathbf{a}_i^{(0)} \mathbb{E}_{(x,y)} \left[ y g' \left( -\mathbf{w}_i^{(0)} \cdot \mathbf{x} + \mathbf{b}_{i+m} \mathbf{x} - y \right) \right] \tag{538}
\]

\[
= \eta^{(1)} \mathbf{a}_i^{(0)} \mathbb{E}_{(x,y)} \left[ -y g' \left( \mathbf{w}_i^{(0)} \cdot \mathbf{x} - \mathbf{b}_{i+m} \mathbf{x} \right) \right] \tag{539}
\]

\[
= \mathbf{w}_i^{(1)}. \tag{540}
\]

**Lemma F.3** (Property of Direction Neighborhood). If \( \mathbf{w} \in \mathcal{C}_{D, \gamma} \), we have \( \rho \mathbf{w} \in \mathcal{C}_{D, \gamma} \) for any \( \rho \neq 0 \). We also have \( \mathbf{0} \notin \mathcal{C}_{D, \gamma} \). Also, if \( (D, s) \in S_{p, \gamma, B_G} \), we have \( (-D, s) \in S_{p, \gamma, B_G} \).

**Proof.** These can be verified by direct calculation.

**Lemma F.4** (Maximum Gaussian Tail Bound). \( M_n \) is the maximum of \( n \) i.i.d. standard normal Gaussian. Then

\[
\Pr \left( M_n \geq \sqrt{2 \log n} + \frac{z}{\sqrt{2 \log n}} \right) \leq e^{-z}. \tag{541}
\]

**Proof.** These can be verified by direct calculation.

**Lemma F.5** (Chi-squared Tail Bound). If \( X \) is a \( \chi^2(k) \) random variable. Then, \( \forall z \in \mathbb{R} \), we have

\[
\Pr(X \geq k + 2\sqrt{k}z + 2z) \leq e^{-z}. \tag{542}
\]

**Proof.** These can be verified by direct calculation.

**Lemma F.6** (Gaussian Tail Bound). If \( g \) is standard Gaussian and \( z > 0 \), we have

\[
\frac{1}{\sqrt{2\pi}} \frac{z}{z^2 + 1} e^{-z^2/2} < \Pr_{g \sim \mathcal{N}(0,1)}[g > z] < \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{543}
\]

**Proof.** These can be verified by direct calculation.

**Lemma F.7** (Gaussian Tail Expectation Bound). If \( g \) is standard Gaussian and \( z \in \mathbb{R} \), we have

\[
|\mathbb{E}_{g \sim \mathcal{N}(0,1)}[g > z]|g > z| < 2 \Pr_{g \sim \mathcal{N}(0,1)}[g > z]^{0.9}. \tag{544}
\]

**Proof of Lemma F.7.** For any \( p \in (0, 1) \), we have

\[
\left| \int_{-\infty}^{\sqrt{2}\text{erf}^{-1}(2p-1)} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx \right| < 2p^{0.9}, \tag{545}
\]

where \( \sqrt{2}\text{erf}^{-1}(2p-1) \) is the quantile function of the standard Gaussian. We finish the proof by replacing \( p \) to be \( \Pr_{g \sim \mathcal{N}(0,1)}[g > z] \).

**Lemma F.8.** If a function \( g \) satisfies \( h(n + 2) = 2h(n + 1) - (1 - \rho^2)h(n) + \beta \) for \( n \in \mathbb{N}_+ \) where \( \rho, \beta > 0 \), then \( h(n) = -\frac{\beta}{\rho^2} + c_1(1 - \rho)^n + c_2(1 + \rho)^n \), where \( c_1, c_2 \) only depends on \( h(1) \) and \( h(2) \).

**Proof.** These can be verified by direct calculation.

**Lemma F.9** (Rademacher Complexity Bounds. Rephrase of Lemma 48 in [32]). For fixed \( \mathbf{W}, \mathbf{b} \), let \( \mathcal{F} = \{ f(\mathbf{a}, \mathbf{w}, \mathbf{b}) : \| \mathbf{a} \| \leq B_{a_2} \} \). Then,

\[
\mathcal{R}(\mathcal{F}) \leq \frac{B_{a_2}^2 \left( \| \mathbf{W} \|^2 + B_{\mathbf{w}}^2 + \| \mathbf{b} \|^2 \right)}{n}. \tag{546}
\]