ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Given a dense shallow neural network, we focus on iteratively creating, training, and combining randomly selected subnetworks (surrogate functions), towards training the full model. By carefully analyzing $i)$ the subnetworks' neural tangent kernel, $ii)$ the surrogate functions' gradient, and $iii)$ how we sample and combine the surrogate functions, we prove linear convergence rate of the training error – within an error region– for an overparameterized single-hidden layer perceptron with ReLU activations for a regression task. Our result implies that, for fixed neuron selection probability, the error term decreases as we increase the number of surrogate models, and increases as we increase the number of local training steps for each selected subnetwork. The considered framework generalizes and provides new insights on dropout training, multi-sample dropout training, as well as Independent Subnet Training; for each case, we provide corresponding convergence results, as corollaries of our main theorem.

1 INTRODUCTION

Overparameterized neural networks have led to unexpected empirical success in deep learning *(Zhang et al., 2021; Goodfellow et al., 2016; Arpit et al., 2017; Recht et al., 2019; Toneva et al., 2018)*, but also have led to new techniques in analyzing neural network training *(Kawaguchi et al., 2017; Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2018; Liang et al., 2019; Arora et al., 2018; Dziugaite & Roy, 2017; Neyshabur et al., 2018; Zhou et al., 2018; Soudry et al., 2018; Shah et al., 2020; Belkin et al., 2019; 2018; Feldman, 2020; Ma et al., 2018; Spigler et al., 2019; Belkin, 2021; Bartlett et al., 2021; Jacot et al., 2018)*. While there is literature that focuses a diverse set of overparameterized neural network architectures *(Frei et al., 2020; Fang et al., 2021; Lu et al., 2020; Huang et al., 2020; Allen-Zhu et al., 2019a; Gu et al., 2020; Cao et al., 2020)* and training algorithms *(Du et al., 2018; Zou et al., 2020; Soltanolkotabi et al., 2018; Oymak & Soltanolkotabi, 2019; Li et al., 2020; Oymak & Soltanolkotabi, 2020)*, most efforts fall under the following scenario: in each iteration, all parameters of the neural network are updated using a version of gradient descent. Yet, advances in regularization techniques *(Srivastava et al., 2014; Wan et al., 2013; Gal & Ghahramani, 2016; Courbariaux et al., 2015; Labach et al., 2019)*, computationally-efficient *(Shazeer et al., 2017; Fedus et al., 2021; Lepikhin et al., 2020; LeJeune et al., 2020; Yao et al., 2021; Yu et al., 2018; Mohtashami et al., 2021; Yuan et al., 2019; Dun et al., 2019; Wolfe et al., 2021)* and communication-efficient distributed training methods *(Vogels et al., 2019; Wang et al., 2021; Yuan et al., 2019; Dun et al., 2019; Wolfe et al., 2021)* deviate from this narrative: One would –explicitly or implicitly– train smaller, randomly-selected, model versions within the large dense network, in an iterative fashion. This raises the question:

"*Can one meaningfully train an overparameterized ML model by iteratively training and combining together smaller versions of it?*"

We provide a positive answer to this question for single-hidden layer perceptrons with ReLU activations. This is a non-trivial, non-convex problem setting, that has been used extensively in proving the behavior of training algorithms on neural networks *(Du et al., 2018; Zou et al., 2020; Soltanolkotabi et al., 2018; Oymak & Soltanolkotabi, 2019; Li et al., 2020; Oymak & Soltanolkotabi, 2020)*.

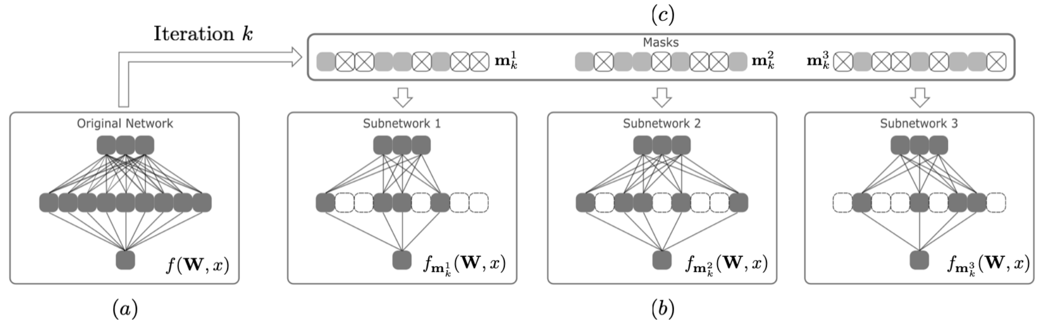


Figure 1: Training a single hidden-layer perceptron using multiple randomly masked subnetworks. Here, $f(\mathbf{W}, \mathbf{x})$ denotes the full model with weight $\mathbf{W}$, and $f_{\mathbf{m}_k^l}(\mathbf{W}, \mathbf{x})$ denotes the surrogate model (subnetwork) with only active neurons dictated by the mask $\mathbf{m}_k^l$ at k-th iteration for subnetwork l.

Let us first describe briefly the scenario considered here; this is depicted in Figure 1. Given a single hidden-layer perceptron Fig.1(a), we sample masks within one training step Fig.1(c), each mask deactivating a subset of the neurons of the original network's hidden layer. In a way, each mask defines a *surrogate model*, as in Fig.1(b), based on the original network, leading to a collection of *subnetworks*. These surrogates independently update their own parameters (possibly on different data shards), by performing a number of (stochastic) gradient descent steps. Lastly, we update the weights of the original network by aggregating the parameters of the subnetworks, before the next iteration starts. Note that multiple masks could share active neurons. When aggregating the updates, we take the mean of the updated values across all subnetworks that have these neurons active.

Motivation and connection to existing methods. Standard gradient descent computes updates over all the weights per iteration, based on a loss function. The present methodology computes updates over weight subsets, based on the generated masks: i.e., the weight updates relate to different loss functions, each depending on the output of a different subnetwork. When the sampled mask leaves all neurons active, the training scheme is the same as in training the whole model. Yet, the same framework connects with existing techniques for training deep neural networks.

Dropout regularization. Dropout *(Srivastava et al., 2014; Wan et al., 2013; Gal & Ghahramani, 2016; Courbariaux et al., 2015)* is a widely-accepted technique against overfitting in deep learning. In each training step, a random mask is generated from some pre-defined distribution, and used to mask-out part of the neurons in the neural network. Later variants of dropout include the drop-connect *(Wan et al., 2013)*, multi-sample dropout *(Inoue, 2019)*, Gaussian dropout *(Wang & Manning, 2013)*, and the variational dropout *(Kingma et al., 2015)*. Here, we restrict our attention to the vanilla dropout, and the multi-sample dropout. The vanilla dropout corresponds to our framework, if in the latter we sample only one mask per iteration, and let the subnetwork perform only one gradient descent update. The multi-sample dropout extends the vanilla dropout in that it samples multiple masks per iteration. *For regression tasks, our theoretical result implies convergence guarantees for these two scenarios on a single hidden-layer perceptron.*

Distributed ML training. Recent advances in distributed model/parallel training have led to variants of distributed gradient descent protocols. Instead of centrally aggregating gradient updates per iteration, distributed local SGD *(Mcdonald et al., 2009; Zinkevich et al., 2010; Zhang & Ré, 2014; Zhang et al., 2016)* updates all the model parameters only after several *local* steps are performed per compute node. This reduces synchronization and thus allows for higher hardware efficiency *(Zhang et al., 2016)*. Yet, all training parameters are updated per outer step, which could be computationally and communication inefficient. The *Independent Subnetwork Training* protocol *(Yuan et al., 2019; Dun et al., 2019; Wolfe et al., 2021)* goes one step further: it combines model- and parallel-training methodologies, with local SGD motions to minimize communication and computational bottlenecks, simultaneously. In particular, IST splits the model vertically, where each machine contains all layers of the neural network, but only with a (non-overlapping) subset of neurons being active in each layer. Multiple local SGD steps can be performed without the workers having to communicate. *Yet, the theoretical understanding of IST is currently missing. Our theoretical result implies convergence guarantees for IST for a single hidden-layer perceptron, and provides insights on how the number of compute nodes affects the performance of the overall protocol.*

Contributions. The present training framework naturally generalizes the approaches above. Yet, current literature –more often than not– omits any theoretical understanding for these scenarios, even for the case of shallow MLPs. While handling multiple layers is a more desirable scenario (and is, indeed, considered as future work), our presented theory illustrates how training and combining multiple randomly masked surrogate models behaves. Our findings can be summarized as follows:

- We provide convergence rate guarantees for i) dropout regularization *(Srivastava et al., 2014)*, ii) multi-sample dropout *(Inoue, 2019)*, iii) and multi-worker IST *(Yuan et al., 2019)*, given a regression task on a single-hidden layer perceptron.
- We show that the Neural Tangent Kernel (NTK) *(Jacot et al., 2018)* of surrogate models stays close to that with initial weights; this implies that it stays close to the infinite width NTK. Consequently, our work shows that training over surrogate models still enjoys linear convergence.
- For subnetworks defined by Bernoulli masks with a fixed distribution parameter, we show that the expectation of aggregated gradient in the first local step is a biased estimator of the gradient computed on the whole network, with the bias term decreasing as the number of subnetworks grows. Moreover, the aggregated gradient starting from the second local step stays close to the aggregated gradient of the first local step. This finding leads to linear convergence of the above training framework with an error term under Bernoulli masks.
- For masks sampled from categorical distribution, we provide tight bounds i) on the average loss increase, when sampling a subnetwork from the whole network; ii) on the loss decrease, when the independently trained subnetworks are combined into the whole model. This finding leads to linear convergence with a more desirable error term than the Bernoulli mask scenario.

Challenges. Much work has been devoted to analyzing the convergence of neural networks based on the NTK perspective *(Jacot et al., 2018)*; see the Related Works section below. The literature in this direction notice that the NTK remains roughly stable throughout training. Therefore, the network output can be approximated by the linearization defined by the NTK. Yet, training with randomly masked neurons poses additional challenges: i) With a randomly generated mask, the NTK changes even with the same set of weights. Thus, analyzing the convergence of neural networks with masked neurons requires careful treatment of the surrogate models' NTKs. ii) Each gradient defined by the subnetwork is a biased estimator of the gradient computed on the whole network, even in the first step of the subnetwork update. iii) From the objective perspective, the non-linear activation function and the gradient aggregation makes the function represented by the updated whole network not necessarily equal to the linear combination of the updated subnetworks. These three challenges complicate analysis, thus driving us to treat the NTK, gradient, and combined network function with special care. We will resolve these difficulties in the proof of the three main theorems.

2 Related Works

The NTK enabled more refined analysis of training overparameterized neural networks *(Jacot et al., 2018; Du et al., 2018; Oymak & Soltanolkotabi, 2020; Song & Yang, 2020; Ji & Telgarsky, 2020; Su & Yang, 2019; Arora et al., 2019; Mianjy & Arora, 2020; Huang et al., 2021)*. NTKs can be viewed as the reproducing kernels of the function space defined by the neural network structure, and are constructed using the inner product between gradients of pairs of data points. With the observation that the NTK stays roughly the same with sufficient overparameterization, recent work has shown that (stochastic) gradient descent achieves zero training loss on shallow neural networks for regression task, even if when the data-points are randomly labeled *(Du et al., 2018; Oymak & Soltanolkotabi, 2020; Song & Yang, 2020)*. Later work characterizes the loss update in terms of the NTK-induced inner-product of the label vector, and notices that, when the label vector aligns with the top eigenvectors of the NTK, training achieves a faster convergence rate *(Arora et al., 2019)*.

A different line of work explores the structure of the data-distribution in classification tasks, by assuming separability when mapped to the Hilbert space induced by the partial application of the NTK *(Ji & Telgarsky, 2020; Mianjy & Arora, 2020)*. Rather than depending on the stability of NTK, the crux of these works relies on the small change in the linearization of the network function. This line of work requires milder overparameterization, and can be easily extended to training stochastic gradient descent without changing the overparameterization requirement. *The above literature assumes all the parameters are updated per iteration.*

There is literature devoted to the analysis of dropout training. For shallow linear neural networks, *(Senen-Cerda & Sanders, 2020a)* give asymptotic convergence rate by carefully characterizing the local minimas. For deep neural networks with ReLU activations, *(Senen-Cerda & Sanders, 2020b)* shows that the training dynamics of dropout converge to a unique stationary set of a projected system of differential equations. Under NTK assumptions, *(Mianjy & Arora, 2020)* shows sublinear convergence rate for an online version for dropout in classification tasks. *Our main theorem implies linear convergence rate of the training loss dynamic for the regression task on a shallow neural network with ReLU activations.*

Theoretical work on Federated Learning (FL) also focuses on the convergence of neural network training. FL-NTK *(Huang et al., 2021)* characterize the asymmetry of the NTK matrix due to the partial data knowledge. For non-i.i.d. data distribution, *(Deng & Mahdavi, 2021)* proves convergence for a shallow neural network by analyzing the semi-Lipschitzness of the hidden layer. *Our work differs since we consider training a partial model with the whole dataset.* Lastly, there is recent work on training partially masked neural networks *(Mohtashami et al., 2021)*. The authors consider minimizing a differentiable objective function under Lipschitz assumptions. *We consider the more frequently used setting of a one-hidden layer perceptron with non-differentiable activation.*

3 TRAINING WITH RANDOMLY MASKED NEURONS

We use bold lower-case letters (e.g., $\mathbf{a}$) to denote vectors, bold upper-case letters (e.g., $\mathbf{A}$) to denote matrices, and standard letters (e.g., a) for scalars. $\|\mathbf{a}\|_2$ stands for the ℓ_2 (Euclidean) vector norm, $\|\mathbf{A}\|_2$ stands for the spectral matrix norm, and $\|\mathbf{A}\|_F$ stands for the Frobenius norm. Unless otherwise stated, p denotes the number of subnetworks, and $l \in [p]$ its index; K denotes the number of global iterations and $k \in [K]$ its index; τ is used for the number of local iterations and $t \in [\tau]$ its index. We use $\mathbf{M}_k$ to denote the mask at iteration k, and $\mathbb{E}_{[\mathbf{M}_k]}[\cdot] = \mathbb{E}_{\mathbf{M}_0,\ldots,\mathbf{M}_k}[\cdot]$ to denote the total expectation over masks $\mathbf{M}_0,\ldots,\mathbf{M}_k$. We use $\mathbb{P}(\cdot)$ to denote the probability of an event, and $\mathbb{I}\{\cdot\}$ the indicator function. For distributions, we use $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ to denote the Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. We use $\texttt{Bern}(\xi)$ to denote the Bernoulli distribution with $\mathbb{P}(x=1)=\xi$ for $x \sim \texttt{Bern}(\xi)$. The categorical distribution over p categories is determined by parameter $\boldsymbol{\xi} \in \mathbb{R}^p$ and produces random variables $x \in [p]$ such that $\mathbb{P}(x=i)=\xi_i$. In this paper, we use a one-hot output of categorical distribution by treating random variables as one-hot vectors $\mathbf{x} \sim \texttt{Categorical}(\boldsymbol{\xi})$, with $\mathbb{P}(x_i=1)=\xi_i$. For a complete clarification of the notations, please refer to Table 1.

3.1 SINGLE HIDDEN-LAYER NEURAL NETWORK WITH RELU ACTIVATIONS

We consider the single hidden-layer neural network with ReLU activations, as in:

$$f(\mathbf{W},\mathbf{a},\mathbf{x}) = \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r\sigma(\langle \mathbf{w}_r,\mathbf{x}\rangle) := f(\mathbf{W},\mathbf{x}).$$

Here, $\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_m]^\top \in \mathbb{R}^{m\times d}$ is the weight matrix of the first layer, and $\mathbf{a} = [a_1,\ldots,a_r]^\top \in \mathbb{R}^m$ is the weight vector of the second layer. We assume that each $\mathbf{w}_r$ is initialized based on $\mathcal{N}(0,\kappa^2\mathbf{I})$. Each weight entry a_r in the second layer is initialized uniformly at random from $\{-1,1\}$. As in *(Du et al., 2018; Zou et al., 2020; Soltanolkotabi et al., 2018; Oymak & Soltanolkotabi, 2019; Li et al., 2020; Oymak & Soltanolkotabi, 2020)*, $\mathbf{a}$ is fixed.

Algorithm 1 RandomlyMaskedTraining

```
Initialize W_0, a
for k = 0, ..., K − 1 do
    Sample mask M_k ∼ D
    for l = 1, ..., p do
        W^l_{k,0} ← W_k
        for t = 0, ..., τ − 1 do
            W^l_{k,t+1} ← W^l_{k,t} − η ∂L_{m^l_k}(W)/∂w_r
        end for
        ΔW^l_k ← W^l_{k,τ} − W_k
    end for
    for r = 1, ..., m do
        w_{k+1,r} ← w_{k,r} + η_{k,r} Σ_{l=1}^{p} Δw^l_{k,r}
    end for
end for
```

Consider a p-subnetwork computing scheme. In the k-th global iteration, we consider each binary mask $\mathbf{M}_k \in \{0,1\}^{m\times p}$ to be composed of subnetwork masks $\mathbf{m}_k^l \in \{0,1\}^m$ for $l \in [p]$. The r-th entry of $\mathbf{m}_k^l$ is denoted as $m_{k,r}^l$. Let $X_{k,r} = \sum_{l=1}^p m_{k,r}^l$ denote the number of subnetworks that update neuron r in global iteration k. Let $N_{k,r} = \max\{X_{k,r},1\}$ to be the normalizer of the aggregated gradient, $N_{k,r}^\perp = \min\{X_{k,r},1\}$ to be the indicator of whether a neuron is selected by at least one subnetwork. The surrogate function defined by a subnetwork mask $\mathbf{m}_k^l$ is given by:

$$f_{\mathbf{m}_k^l}(\mathbf{W}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r m_{k,r}^l \sigma(\langle \mathbf{w}_r, \mathbf{x} \rangle).$$

With colored text, we highlight the differences between the full model and the surrogate functions. ξ is a constant to be defined later on, conveniently selected by our theory. We consider training the neural network using the regression loss. Given a dataset $(\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the function output on the whole dataset is denoted as $f(\mathbf{W}, \mathbf{X}) = [f(\mathbf{W}, \mathbf{x}_1), \dots, f(\mathbf{W}, \mathbf{x}_n)]$. Then, the regression loss of a surrogate model is given by:

$$L_{\mathbf{m}_k^l}(\mathbf{W}) = \frac{1}{2} \left\| \mathbf{y} - f_{\mathbf{m}_k^l}(\mathbf{W}, \mathbf{X}) \right\|_2^2.$$

The surrogate gradient is computed as:

$$\frac{\partial L_{\mathbf{m}_k^l}(\mathbf{W})}{\partial \mathbf{w}_r} = \frac{1}{\sqrt{m}} \sum_{i=1}^{n} a_r m_{k,r}^l \left(f_{\mathbf{m}_k^l}(\mathbf{W}, \mathbf{x}_i) - y_i \right) \mathbf{x}_i \mathbb{I}\{\langle \mathbf{w}_r, \mathbf{x}_i \rangle \geq 0\}.$$

Moreover, we define the global aggregation step size as $\eta_{k,r} = \frac{N_{k,r}^{\perp}}{N_{k,r}}$. Within this setting, the general training algorithm is given by Algorithm 1.

4 Convergence on Two-Layer ReLU Neural Network

We often assume that each entry in the mask $m_{k,r}^l \sim \texttt{Bern}(\xi)$ is sampled independently, unless otherwise stated. Here, $\xi \in (0,1)$ represents the Bernoulli distribution parameter. Since the forward pass of the surrogate function is a linear combination of ξ-proportion of the neurons' output, we multiply each neuron output by a factor of ξ to keep the overall output scaling consistent, as in *(Mianjy & Arora, 2020)*. For notation clarity, we define:

$$u_k^{(i)} = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \xi \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle) = \frac{\xi}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle).$$

We focus on the behavior of the following loss, computed on the whole network over iterations k:

$$L_k = \|\mathbf{y} - \mathbf{u}_k\|_2^2, \text{ where } \mathbf{u}_k = \left[u_k^{(1)}, \dots, u_k^{(n)} \right].$$

This is the regression loss over iterations k between observations $\mathbf{y}$ and the learned model $\mathbf{u}_k$.

Properties of subnetwork NTK. Recent works on analyzing the convergence of gradient descent for neural networks consider approximating the function output $\mathbf{u}_k$ with the first order Taylor expansion *(Du et al., 2018; Arora et al., 2019; Song & Yang, 2020)*. For constant step size η, taking the gradient descent's –$\mathbf{W}_{k+1} = \mathbf{W}_k - \eta \nabla_{\mathbf{W}} L(\mathbf{W}_k)$– first-order Taylor expansion, we get:

$$u_{k+1}^{(i)} \approx u_k^{(i)} + \left\langle \nabla_{\mathbf{W}} u_k^{(i)}, \mathbf{W}_{k+1} - \mathbf{W}_k \right\rangle \approx u_k^{(i)} - \xi\eta \sum_{j=1}^{n} \mathbf{H}(k)_{ij} (u_k^{(j)} - y_j), \tag{1}$$

where $\mathbf{H}(k) \in \mathbb{R}^{n \times n}$ is the finite-width NTK matrix of iteration k, given by

$$\mathbf{H}(k)_{ij} = \frac{\xi}{m} \langle \mathbf{x}_i, \mathbf{x}_j \rangle \sum_{r=1}^{m} \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle \geq 0, \langle \mathbf{w}_{k,r}, \mathbf{x}_j \rangle \geq 0\}.$$

Compared with previous definition of finite-width NTK, we have an additional scaling factor ξ. This is because based on our later definition of masked-NTK, we would like the masked-NTK to be an unbiased estimator of the finite-width NTK. In the overparameterized regime, the change of the network's weights is controlled in a small region around initialization. Therefore, the change of $\mathbf{H}(k)$ is small, staying close to the NTK at initialization. Moreover, the latter can be well approximated by the infinite-width NTK:

$$\mathbf{H}_{ij}^{\infty} = \xi \cdot \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})} \left[\langle \mathbf{x}_i, \mathbf{x}_j \rangle \mathbb{I}\{\langle \mathbf{w}, \mathbf{x}_i \rangle \geq 0, \langle \mathbf{w}, \mathbf{x}_j \rangle \geq 0\} \right].$$

(Du et al., 2018) shows that $\mathbf{H}^{\infty}$ is positive definite.

Theorem 1. *(Du et al., 2018) Denote $\lambda_0 := \lambda_{\min}(\mathbf{H}^{\infty})$, the minimum eigenvalue of $\mathbf{H}^{\infty}$. If for any $i \neq j$ it holds that the points $\mathbf{x}_i, \mathbf{x}_j$ are not co-aligned, i.e., $\mathbf{x}_i \not\parallel \mathbf{x}_j$, then we have $\lambda_0 > 0$.*

With $\mathbf{H}(k)$ staying sufficiently close to $\mathbf{H}^\infty$, (Du et al., 2018; Arora et al., 2019; Song & Yang, 2020) show that $\lambda_{\min}(\mathbf{H}(k)) \geq \frac{\lambda_0}{2} > 0$. Moreover, Equation 1 implies that

$$\mathbf{u}_{k+1} - \mathbf{u}_k \approx -\xi\eta\mathbf{H}(k)(\mathbf{u}_k - \mathbf{y}),$$

that further leads to linear convergence rate:

$$L_{k+1} \approx L_k + \langle \nabla_{\mathbf{u}_k} L_k, \mathbf{u}_{k+1} - \mathbf{u}_k \rangle \approx L_k - \xi\eta\langle \mathbf{u}_k - \mathbf{y}, \mathbf{H}(k)(\mathbf{u}_k - \mathbf{y})\rangle \approx (1 - \xi\eta\lambda_0)\, L_k.$$

In rigorous NTK analysis, the Taylor expansion for both $\mathbf{u}_k$ and L_k produces error term that reduces the convergence rate from $\eta\lambda_0$ to $\gamma\eta\lambda_0$ with $\gamma \in (0, 1)$ being a constant.

Yet, for neural networks with randomly masked neurons, the situation is trickier: in each iteration, due to the different masks, the NTK changes even when the weights stay the same. Since, the NTK of the masked network changes as $i)$ the mask changes, and $ii)$ the network weight changes, we define the masked-NTK induced by the network weight in k-th iteration and the mask in the k'-th iteration as follows:

Definition 1. *Let $\mathbf{m}^l_{k'}$ be the mask of subnetwork l in iteration k'. We define the masked-NTK in global iteration k and local iteration t induced by $\mathbf{m}^l_{k'}$ as:*

$$\left(\mathbf{m}^l_{k'} \circ \mathbf{H}(k,t)\right)_{ij} = \tfrac{1}{m}\langle \mathbf{x}_i, \mathbf{x}_j\rangle \sum_{r=1}^{m} m^l_{k',r} \mathbb{I}\{\langle \mathbf{w}^l_{k,t,r}, \mathbf{x}_i\rangle \geq 0, \langle \mathbf{w}^l_{k,t,r}, \mathbf{x}_j\rangle \geq 0\}.$$

with $\mathbf{w}^l_{k,t,r}$ denoting the rth weight vector in the kth global iteration and tth local iteration.

Here, with colored text we highlight the main differences to the common NTK definition. Note that although we are only interested in the masked-NTK with $k = k'$, to facilitate our analysis on the minimum eigenvalue of masked-NTK, we also allow $k \neq k'$. Note that we have $\mathbb{E}_{\mathbf{M}_k}\left[\mathbf{m}^l_k \circ \mathbf{H}(k,0)\right] = \mathbf{H}(k)$. Throughout iterations of the algorithm, the following theorem shows that all masked-NTKs stay sufficiently close to the infinite-width NTK.

Theorem 2. *Suppose the number of hidden nodes satisfies $m = \Omega\left(n^2\log(Kpn/\delta)/\xi\lambda_0^2\right)$. If for all k, t it holds that $\|\mathbf{w}_{k,t,r} - \mathbf{w}_{0,r}\|_2 \leq R := \frac{\kappa\lambda_0}{8n}$, then with probability at least $1-\delta$, for all $k, k' \in [K]$ and all $l \in [p]$, we have:*

$$\lambda_{\min}(\mathbf{m}^l_{k'} \circ \mathbf{H}(k,t)) \geq \tfrac{\lambda_0}{2}.$$

Note that the above theorem relies on the small weight change in iteration (k, t). In order to guarantee each subnetwork's loss decrease, we need to ensure that the $i)$ the weight change is bounded up to global iteration k (this implies that when a subnetwork is sampled from the whole network, its weights do not deviate much from the initialization); $ii)$ the weight change during the local training of the subnetwork is also bounded. This requires an upper bound on the sampled subnetwork's loss before the local training starts. The following hypothesis sets up the "skeleton" to construct different theorems, based on different problems considered.

Hypothesis 1. *Assume that for all $i \in [n]$ we have $\|\mathbf{x}_i\|_2 = 1$. Fix the number of global iterations K. Suppose the number of hidden nodes satisfies $m = \Omega\left(n^2\log(Kpn/\delta)/\xi\lambda_0^2\right)$, and suppose we use a constant step size $\eta = O\left(\lambda_0/n^2\right)$. Then, if for all $k' \leq k$ the following convergence holds with convergence rate $\alpha \in (0,1)$ and error term $\alpha B_1 > 0$:*

$$\mathbb{E}_{[\mathbf{M}_{k'}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y} - \mathbf{u}_{k'+1}\|_2^2\right] \leq (1-\alpha)\mathbb{E}_{[\mathbf{M}_{k'-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y} - \mathbf{u}_{k'}\|_2^2\right] + \alpha B_1. \tag{2}$$

and, further, the weight perturbation before iteration k is bounded by

$$\|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 + 2\eta\tau\sqrt{\tfrac{2nK}{m\delta}}\left(\frac{1}{\alpha}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y} - \mathbf{u}_k\|_2\right] + (K-k)B\right) \leq R \tag{3}$$

with $B = \sqrt{\frac{B_1}{\alpha}} + \kappa\sqrt{\xi(1-\xi)pn}$, then, with probability at least $1 - 4\delta$, we have:

$$\|\mathbf{y} - \hat{\mathbf{u}}^l_{k,t+1}\|_2^2 \leq \left(1 - \frac{\eta\lambda_0}{2}\right)\|\mathbf{y} - \hat{\mathbf{u}}^l_{k,t}\|_2^2, \tag{4}$$

with the surrogate function of the subnetwork function defined by mask l in global iteration k and local iteration t denoted as

$$\hat{u}^{l(i)}_{k,t} = f_{\mathbf{m}^l_k}(\mathbf{W}^l_{k,t}, \mathbf{x}_i) \quad \hat{\mathbf{u}}^l_{k,t} = \left[\hat{u}^{l(1)}_{k,t}, \ldots, \hat{u}^{l(n)}_{k,t}\right]$$

and the local weight perturbation satisfies that for all $t \in [\tau]$:

$$\|\mathbf{w}_{k,t,r} - \mathbf{w}_{k,r}\| \leq \tfrac{\eta\tau\sqrt{2nK}}{\sqrt{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y} - \mathbf{u}_k\|_2\right] + 2\eta\kappa n\sqrt{\tfrac{2\xi(1-\xi)pK}{m\delta}} \tag{5}$$

The hypothesis above states that, in a given global step k, given the linear convergence of the loss (Equation 2) and a small weight perturbation guarantee (Equation 3) in previous iterations $k' \leq k$, each subnetwork's local loss also decreases linearly (Equation 4), as well as the weight perturbation remains bounded (Equation 5). Yet, the above hypothesis does not connect the subnetwork's loss with the whole network's loss through the sampling and aggregation process.

This is the goal in the sections that follow. Our aim is to turn Hypothesis 1 into a series of specific theorems that cover different cases. In particular, we provide a convergence result for Algorithm 1 for i) masks with i.i.d. Bernoulli entries, where we leverage the aggregation of gradients, and prove the non-decrease property of subnetwork's loss to bound the difference between the aggregated gradient at local iteration t and local iteration 1; and ii) masks with i.i.d Categorical rows, where we leverage the loss change in sampling and aggregating subnetworks, and depend on the linear convergence to show loss decrease. In both settings, we prove the bound on weight perturbation of each subnetwork, as hypothesized above, to bound the aggregated weight perturbation. Also, in both settings, we show that the required condition in Hypothesis 1 can be satisfied.

4.1 Main Results

While the local gradient descent for each subnetwork makes progress with high probability, when a large network is split into small subnetworks, the expected error on the dataset increases. Since the masks are sampled i.i.d. from the joint Bernoulli distribution, the function of each sub-network is an unbiased estimation of the function represented by the large network. Therefore, we have:

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] = \|\mathbf{y}-\mathbf{u}_k\|_2^2 + \mathbb{E}_{\mathbf{M}_k}\left[\|\hat{\mathbf{u}}_k^l-\mathbf{u}_k\|_2^2\right].$$

When analyzing the convergence, the last term needs to be carefully dealt with. Moreover, it is not trivial to show that, when combining the updated network of the local steps, the loss computed on the whole network is smaller than or equal to the error of each sub-network. Resolving these technical difficulties, we present our general result for masks sampled from a joint Bernoulli distribution:

Theorem 3. *Let the assumption of Theorem 1 hold, i.e., $\lambda_0 \geq 0$, and for all $i \in [n]$, we have $\|\mathbf{x}_i\|_2 = 1$ and $|y_i| \leq C-1$ for some $C \geq 0$. Fix the number of global iterations to K and the number of local iterations to τ. Assume for each global iteration k, the mask $\mathbf{M}_k$ is generated such that $m_{k,r}^l \sim \texttt{Bern}(\xi)$ with $\xi \in (0,1]$. Let the number of hidden neurons satisfy*

$$m = \Omega\left(\frac{K}{\delta}\max\left\{\frac{n^4}{\kappa^2\xi\theta\lambda_0^4}, \frac{nK^2B_1}{\kappa^2\theta\lambda_0^2}, K^2p\right\}\right), \tag{6}$$

and fix the local stepsize to $\eta = O\left(\frac{\lambda_0}{\max\{n,p\}n\tau}\right)$. Then with probability at least $1-\delta$ we have

$$\mathbb{E}_{[\mathbf{M}_{k'-1}]}\left[\|\mathbf{y}-\mathbf{u}_{k'}\|_2^2\right] \leq \left(1-\tfrac{1}{4}\eta\theta\tau\lambda_0\right)^{k'}\|\mathbf{y}-\mathbf{u}_0\|_2^2 + B_1, \tag{7}$$

with $\theta = 1-(1-\xi)^p$ and

$$\begin{aligned} B_1 &= \tfrac{64(1-\xi)^2n^3d}{m\lambda_0^2} + \tfrac{68\eta\tau(\theta-\xi^2)n^3\kappa^2}{p\lambda_0} + \tfrac{(\theta-\xi^2)n\kappa^2}{6\tau^2p} + \tfrac{(\tau-1)^2pC_1}{24\tau^2} \\ &\quad + \tfrac{8\eta^2(\tau-1)^2n^4pC_1}{\lambda_0^2} + 4\eta(\tau-1)pnC_1, \\ C_1 &= \frac{4\theta^2(1-\xi)n\kappa^2}{p}. \end{aligned}$$

Overall, given the overparameterization requirement in Equation 9, the neural network training error, as expressed in Equation 7, drops linearly within an error region, defined by B_1 in Equation 7. While this theorem seems complicated, we can make mild assumptions and simplify the key messages of this form. In the following sections, we assume that $\max\{K,d,p\} \leq n$ and $\kappa = n^{-\frac{1}{2}}$.

Dropout. The dropout algorithm (Srivastava et al., 2014) corresponds to the case $\tau = 1, p = 1$. For this assignment, we arrive at the following corollary.

Corollary 1. *Under the assumptions in Theorem 3, and the additional assumption that $\max\{K,d\} \leq n$, fix the number of iterations to K, the step size to $\eta = O\left(\lambda_0/n^2\right)$, and the number of hidden neurons satisfies $m = \Theta\left(n^5K/\xi^2\lambda_0^4\delta\right)$. Suppose we run the dropout algorithm on a two-layer ReLU neural network. Then, with probability at least $1-\delta$, we have:*

$$\mathbb{E}_{[\mathbf{M}_{k'-1}]}\left[\|\mathbf{y}-\mathbf{u}_{k'}\|_2^2\right] \leq \left(1-\tfrac{1}{4}\eta\xi\lambda_0\right)^{k'}\|\mathbf{y}-\mathbf{u}_0\|_2^2 + O\left(1-\xi\right)$$

Typically, $1-\xi$ is usually referred to as the "dropout rate". In our result, as ξ approaches 0, which corresponds to the scenario that no neurons are selected, the convergence rate approaches 1, meaning that the loss hardly decreases. In the mean time, the error term remains constant. On the contrary, as ξ approaches 1, which corresponds to the scenario that all neurons are selected, we get the same convergence rate of $1-O(\eta\lambda_0)$ as in previous literature, and the error term decreases to 0.

Multi-Sample Dropout. The multi-sample dropout *(Inoue, 2019)* corresponds to the scenario where $\tau = 1, p \geq 1$. Our corollary below indicates how increasing p helps the convergence.

Corollary 2. *Under the assumptions in Theorem 3, and the additional assumption that* $\max\{K, d\} \leq n$*, fix the number of iterations to* K*, the step size to* $\eta = O(\lambda_0/n^2)$*, and let the number of hidden neurons satisfy* $m = \Theta\left(n^5 K/\xi\theta\lambda_0^4\delta\right)$*. Suppose we run the* p*-sample dropout algorithm on a two-layer ReLU neural network. Then, with probability at least* $1-\delta$*, we have:*

$$\mathbb{E}_{[\mathbf{M}_{k'-1}]}\left[\|\mathbf{y}-\mathbf{u}_{k'}\|_2^2\right] \leq \left(1-\tfrac{1}{4}\eta\theta\lambda_0\right)^{k'}\|\mathbf{y}-\mathbf{u}_0\|_2^2 + O\left(\tfrac{(1-\xi)^2}{nK} + \tfrac{\theta-\xi^2}{p}\right).$$

Based on this corollary, increasing the number of subnetworks p improve the convergence rate since θ increases as p increases, due to the increasing coverage probability of each neuron. Moreover, increasing the number of subnetworks help decreasing the error term even when the dropout rate ξ is fixed. After p is as large as nK, the error term become dominated by the term $O\left((1-\xi)^2/nK\right)$.

Multi-Worker IST. The multi-worker IST algorithm *(Yuan et al., 2019)* is very similar to the general scheme with $p \geq 1$ and $\tau \geq 1$, but with the additional assumption that $\max\{K, d, p\} \leq n$, and a special choice of initialization $\kappa = n^{-\frac{1}{2}}$.

Corollary 3. *Under the assumptions in Theorem 3, fix the number of iterations to* K*, the step size to* $\eta = O\left(\lambda_0/n\tau\max\{n,p\}\right)$*, and let the number of hidden neurons satisfy* $m = \Theta\left(n^5K/\xi\theta\lambda_0^4\delta\right)$*. Suppose we run the IST algorithm on a two-layer ReLU neural network. Then, with probability at least* $1-\delta$*, we have*

$$\mathbb{E}_{[\mathbf{M}_{k'-1}]}\left[\|\mathbf{y}-\mathbf{u}_{k'}\|_2^2\right] \leq \left(1-\tfrac{1}{4}\eta\theta\tau\lambda_0\right)^{k'}\|\mathbf{y}-\mathbf{u}_0\|_2^2 + O\left(\tfrac{(1-\xi)^2}{nK} + \tfrac{\theta-\xi}{p} + \tfrac{(\tau-1)^2\theta^2(1-\xi)}{\tau^2}\right).$$

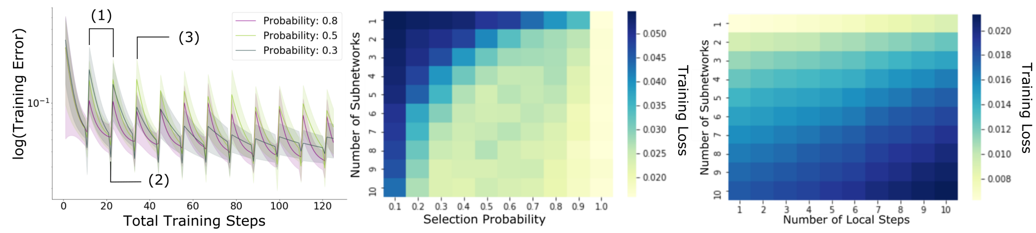
While this corollary presents a convergence result for the multi-worker IST, which is missing in the current literature, it also bears the problem that, when $\tau > 1$, the last error term $\frac{(\tau-1)^2\theta^2(1-\xi)}{\tau^2}$ potientially increase as the number of workers (subnetworks) increase. This could again be due to the fact that some neurons are shared among workers and the gradients from the individual workers are normalized during aggregation. We present an alternative theorem that resolves this issue, by making another assumption on the masks generated in each global iteration. In particular, we assume that, for a p-worker scenario, for each neuron, *the mask is generated from a categorical distribution*, where each worker uses the same probability. In this way, the masks endorsed by each worker are non-overlapping (as stated in *(Yuan et al., 2019)*), and the union of the masks covers the whole set of hidden neurons. The following theorem presents the convergence result under this setting.

Theorem 4. *Suppose the assumption of Theorem 1 holds, i.e.* $\lambda_0 > 0$*, and for all* $i \in [n]$*, we have* $\|\mathbf{x}_i\|_2 = 1$ *and* $|y_i| \leq C-1$ *for some* $C \geq 1$*. Fix the number of global iterations to* K*, and the number of local iterations to* τ*. Let* $\boldsymbol{\xi} = 1/p \cdot \mathbf{1}_p$*, and suppose that for each* $r \in [m], k \in [K]$*, the mask is generated such that* $\mathbf{m}_{k,r} \sim$ ***Categorical***$(\boldsymbol{\xi})$*. Let the number of hidden neurons satisfy* $m = \Omega\left(\tau^2 n^2 K \max\{n^2, d\}/\lambda_0^4\delta\right)$*. Then, with constant local step-size* $\eta = \left(\lambda_0/n^2\right)$ *we have that:*

$$\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] \leq \left(\tfrac{1}{2}+\tfrac{1}{2}\left(1-\tfrac{\eta\lambda_0}{2}\right)^{\tau}\right)^k \cdot C^2 n + O\left(\tfrac{\tau(p-1)^2n^3d}{\lambda_0^2 m}\right),$$

holds with probability at least $1-\delta$*.*

This theorem has a couple noticeable properties. First, when the number of workers $p = 1$, i.e., the scenario of multi-worker IST reducing to the full-network training, the error term disappears, driving further connections between regular and IST training. Second, one can arbitrarily increase the overparameterization parameter m, which further leads to error term decrease: this suggests that IST training could be benefited by wider models, an observation made in *(Yuan et al., 2019; Dun et al., 2019; Wolfe et al., 2021)*.



(a) (b) (c)

Figure 2: Validation experiments on a single hidden layer perceptron.

5 EXPERIMENTS

We validate our theory with the Communities and Crime Data Set *(US Department of Commerce, Bureau of the Census, 1992; Dua & Graff, 2017; Redmond & Baveja, 2002)*. This dataset has 128 features, and we take a subset of 100 samples from the dataset. We run Algorithm 1 on a one hidden layer perceptron with 4000 hidden neurons. We fix the number of global iteration to $K = 200$ and use a constant step size.

In Figure 2a, we plot the logarithm of the mean and variance of the training error dynamic with respect to the K (for clarity, we only plot the first 125 iterations), which includes the sampling step, local training steps, as well as the gradient aggregation step. Notice that there are three types of dynamics, as annotated in the figure: (1) *A smooth decrease of training error:* This corresponds to subnetworks' local training, which is supported by Hypothesis 1 that each subnetwork makes local progress. (2) *The sudden decrease of training error:* This corresponds to the aggregation of locally-trained subnetworks, and is consistent with our proof in Theorem 4. (3) *The sudden increase of training error:* This corresponds to re-sampling subnetworks; according to our theory, the expected average training error increases after sampling.

Figure 2b and Figure 2c provide heatmap results that demonstrate the change of the error term as we vary the number of subnetworks, number of local steps, and the selection probability. In Figure 2b, the subnetworks are generated using Bernoulli masks, and training process with a fixed number of local steps. Note that, as we fix the number of subnetworks and increase the selection probability, the error decreases (lighter colors in heatmap). Moreover, if we fix the number of selection probability and increase the number of subnetworks, the training error also decreases. This is consistent with Theorem 3. In Figure 2c, the subnetworks are generated using categorical masks. Since for categorical masks, the selection probability is determined by the number of subnetworks, we instead vary the number of local steps. Note that the training error increases both when we increase the number of subnetworks (and thus resulting in a smaller subnetwork for each worker) and increasing the number of local steps. This is consistent with our theoretical result in Theorem 4.

6 CONCLUSION

We prove linear convergence up to an error region when training and combining subnetworks in a single hidden-layer perceptron scenario. Our work extends results on dropout, multi-sample dropout, and the Independent Subnet Training, and has broad implications on how the sampling method, the number of subnetworks, and the number of local steps affect the convergence rate and the error region. While our work focus on the single hidden-layer perceptron, we consider multi-layer perceptrons as an interesting direction: we conjecture that a more refined analysis of each layer's output is required *(Du et al., 2019; Allen-Zhu et al., 2019b)*. Moreover, focusing on the convergence of a stochastic algorithm for our framework is a different research direction.

7 Reproducibility Statement

In the theorectical results of this paper, we make assumptions on the dataset similar to previous work *(Du et al., 2018; Song & Yang, 2020; Arora et al., 2019)*, where $i)$ no two input data points are co-aligned, $ii)$ the input data points are normalized, and $iii)$ the value of the labels are bounded. We point out that $i)$ is a standard assumption in machine learning settings, since co-aligned data points with different labels cannot be fitted to zero training error, and $ii)$ and $iii)$ can ba achieved by normalizing the dataset. Moreover, the additional assumption of fixing the number of global iterations K is also common in previous literature *(Su & Yang, 2019; Allen-Zhu et al., 2019a)*. Although increasing K requires a larger overparameterization, we note that, since the learning rate is independent of K, achieving an ϵ-error near the error region requires $K = \Omega\left(\log n/\epsilon / \log(1 - \eta\theta\tau\lambda_0/4)^{-1}\right)$. This shows that K typically do not affect the overparameterization much. Lastly, we provide proof of Theorem 2, Hypothesis 1, Theorem 3 and Theorem 4 in section B, C D, and F in the appendix, respectively. We also provide code for running the experiments as supplementary material.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019b. URL `https://proceedings.mlr.press/v97/allen-zhu19a.html`.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, 2019.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242. PMLR, 2017.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30:6240–6249, 2017.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv preprint arXiv:2105.14368*, 2021.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in Neural Information Processing Systems*, 31:2300–2311, 2018.

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

Jinming Cao, Yangyan Li, Mingchao Sun, Ying Chen, Dani Lischinski, Daniel Cohen-Or, Baoquan Chen, and Changhe Tu. Do-conv: Depthwise over-parameterized convolutional layer. *arXiv preprint arXiv:2006.12030*, 2020.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pp. 3123–3131, 2015.

Yuyang Deng and Mehrdad Mahdavi. Local sgd optimizes overparameterized neural networks in polynomial time. *arXiv preprint arXiv:2107.10868*, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/du19c.html`.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

C. Dun, C. Wolfe, A. Kyrillidis, and C. Jermaine. Resist: Layer-wise decomposition of resnets for distributedtraining. *Preprint*, 2019.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on Learning Theory*, pp. 1887–1936. PMLR, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Spencer Frei, Yuan Cao, and Quanquan Gu. Algorithm-dependent generalization bounds for over-parameterized deep residual networks. *Advances in neural information processing systems*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Yihong Gu, Weizhong Zhang, Cong Fang, Jason D Lee, and Tong Zhang. How to characterize the landscape of overparameterized convolutional neural networks. 2020.

Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning convergence analysis, 2021.

Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feedforward networks?—a neural tangent kernel perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

Hiroshi Inoue. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8580–8589, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks, 2020.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28:2575–2583, 2015.

Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. Survey of dropout methods for deep neural networks. *arXiv preprint arXiv:1904.13310*, 2019.

Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 3525–3535. PMLR, 2020.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pp. 4313–4324. PMLR, 2020.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-Rao metric, geometry, and complexity of neural networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 888–896. PMLR, 2019.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pp. 6426–6436. PMLR, 2020.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pp. 1231–1239, 2009.

Poorya Mianjy and Raman Arora. On convergence and generalization of dropout training, 2020.

Amirkeivan Mohtashami, Martin Jaggi, and Sebastian U Stich. Simultaneous training of partially masked neural networks. *arXiv preprint arXiv:2106.08895*, 2021.

Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, 30:5947–5956, 2017.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pp. 4951–4960. PMLR, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002. ISSN 0377-2217. doi: https://doi.org/10.1016/S0377-2217(01)00264-8. URL https://www.sciencedirect.com/science/article/pii/S0377221701002648.

Albert Senen-Cerda and Jaron Sanders. Asymptotic convergence rate of dropout on shallow linear neural networks, 2020a.

Albert Senen-Cerda and Jaron Sanders. Almost sure convergence of dropout algorithms for neural networks, 2020b.

Vatsal Shah, Soumya Basu, Anastasios Kyrillidis, and Sujay Sanghavi. On generalization of adaptive methods for over-parameterized linear regression. *arXiv preprint arXiv:2011.14066*, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective, 2019.

US Department of Commerce, Bureau of the Census. *Census of Population and Housing 1990 United States: Summary Tape File 1a 3a (Computer Files)*, volume 9575. US Department of Commerce, Bureau of the Census Producer, Washington, DC and Inter-University Consortium for Political and Social Research, Ann Arbor, MI, 1992.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2018.

Thijs Vogels, Sai Praneeth Karinireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. *Advances In Neural Information Processing Systems 32 (Nips 2019)*, 32(CONF), 2019.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pp. 1058–1066. PMLR, 2013.

Hongyi Wang, Saurabh Agarwal, and Dimitris Papailiopoulos. Pufferfish: Communication-efficient models at no extra cost. *Proceedings of Machine Learning and Systems*, 3, 2021.

Sida Wang and Christopher Manning. Fast dropout training. In *international conference on machine learning*, pp. 118–126. PMLR, 2013.

Cameron R Wolfe, Jingkang Yang, Arindam Chowdhury, Chen Dun, Artun Bayer, Santiago Segarra, and Anastasios Kyrillidis. GIST: Distributed training for large-scale graph convolutional networks. *arXiv preprint arXiv:2102.10424*, 2021.

Tianyi Yao, Daniel LeJeune, Hamid Javadi, Richard G Baraniuk, and Genevera I Allen. Minipatch learning as implicit ridge-like regularization. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 65–68. IEEE, 2021.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In *International Conference on Learning Representations*, 2018.

B. Yuan, C. Wolfe, C. Dun, Y. Tang, A. Kyrillidis, and C. Jermaine. Distributed learning of deep neural networks using independent subnet training. *arXiv preprint arXiv:1910.02120*, 2019.

Ce Zhang and Christopher Ré. Dimmwitted: A study of main-memory statistical analytics. *Proceedings of the VLDB Endowment*, 7(12):1283–1294, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel SGD: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations*, 2018.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pp. 2595–2603, 2010.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.

A PRELIMINARY AND DEFINITION

In the proofs of our theorems, we use extensively the following tools.

Definition 2. *(Sub-Gaussian Random Variable) A random variable X is κ^2-sub-Gaussian if*

$$\mathbb{E}[e^{tx}] \leq e^{\frac{\kappa^2 t^2}{2}}$$

Definition 3. *(Sub-Exponential Random Variable) A random variable with mean $\mathbb{E}[X] = \mu$ is (κ', α)-sub-exponential if there exists non-negative (κ', α) such that for all $t \leq \alpha^{-1}$*

$$\mathbb{E}[e^{t(X-\mu)}] \leq e^{-\frac{t^2\kappa'^2}{2}}$$

Property 1. *(Sub-Exponential Tail Bound) For a (κ', α)-sub-exponential random variable X with $\mathbb{E}[X] = \mu$, then we have*

$$\mathbb{P}(X > \mu + t) \leq \begin{cases} e^{-\frac{t^2}{2\kappa'^2}} & \text{if } 0 \leq t \leq \frac{\kappa'^2}{\alpha} \\ e^{-\frac{t^2}{2\alpha}} & \text{if } t > \frac{\kappa'^2}{\alpha} \end{cases}$$

Property 2. *(Markov's Inequality) For a non-negative random variable X, we have*

$$\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}[X]$$

Property 3. *(Hoeffding's Inequality for Bounded Random Variables) Let $X_1, \ldots, X_n$ be independent random variables bounded by $|X_i| \leq 1$ for all $i \in [n]$. Then we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right) \leq e^{-2nt^2}$$

Property 4. *(Berstein's Inequality) Let $X_1, \ldots, X_n$ be random variables with $\mathbb{E}[X_i] = 0$ for all $i \in [n]$. If $|X_i| \leq M$ almost surely, then*

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leq e^{-\frac{t^2/2}{\sum_{j=1}^n \mathbb{E}[X_j^2] + Mt/3}}$$

Property 5. *(Jensen's Inequality for Expectation) For a non-negative random variable X, we have*

$$\mathbb{E}\left[X^{\frac{1}{2}}\right] \leq (\mathbb{E}[X])^{\frac{1}{2}}$$

Apart from the properties above, we also need the following definitions to facilitate our analysis. First, we note that, in the following proofs, we let $R = \frac{\kappa\lambda_0}{192n}$. Define

$$A_{ir} = \{\exists \mathbf{w} \in \mathcal{B}(\mathbf{w}_{0,r}, R) : \mathbb{I}\{\langle \mathbf{w}, \mathbf{x}_i \rangle \geq 0\} \neq \mathbb{I}\{\langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle \geq 0\}\}$$

to denote the event that for sample $\mathbf{x}_i$, the activation pattern of neuron r may change through training if the weight vector change is bounded in the R-ball centered at initialization. Moreover, let

$$\begin{aligned} S_i &= \{r \in [m] : \neg A_{ir}\} \\ S_i^{\perp} &= [m] \setminus S_i \end{aligned} \tag{8}$$

to be the set of neurons whose activation pattern does not change for sample $\mathbf{x}_i$ if the weight vector change is bounded by the R-ball centered at initialization. Moreover, since we are interested in the loss dynamic computed on the following function

$$u_k^{(i)} = \frac{\xi}{\sqrt{m}}\sum_{r=1}^m a_r \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle).$$

we denote the full gradient of loss with respect to each weight vector $\mathbf{w}_r$ as

$$\frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r} = \frac{\xi}{\sqrt{m}}\sum_{i=1}^n a_r \mathbf{x}_i (u_k^{(i)} - y_i)\mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle \geq 0\}$$

B Proof of Theorem 2

Recall the definition of masked-NTK

$$(\mathbf{m}_{k'}^l \circ \mathbf{H}(k,t))_{ij} = \frac{1}{m} \langle \mathbf{x}_i, \mathbf{x}_j \rangle \sum_{r=1}^m m_{k',r}^l \mathbb{I}\{\langle \mathbf{w}_{k,t,r}^l, \mathbf{x}_i \rangle \geq 0, \langle \mathbf{w}_{k,t,r}^l, \mathbf{x}_j \rangle \geq 0\}$$

To start with, we fix $k' \in [K], l \in [p]$, and $i, j \in [n]$. In this case, let

$$h_r = m_{k',r}^l \langle \mathbf{x}_i, \mathbf{x}_j \rangle \mathbb{I}\{\langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle \geq 0, \langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle \geq 0\}$$

Then we have

$$(\mathbf{m}_{k'}^l \circ \mathbf{H}(0,0))_{ij} = \frac{1}{m} \sum_{r=1}^m h_r$$

Also we have

$$\mathbb{E}_{\mathbf{M}_k, \mathbf{W}} [h_r] = \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})} [\mathbb{E}_{\mathbf{M}_k} [h_r]] = \mathbf{H}_{ij}^\infty$$

Note that for all r we have $|h_r| \leq 1$. Thus we apply Hoeffding's inequality for bounded random variables and get

$$\mathbb{P}\left(\left|(\mathbf{m}_{k'}^l \circ \mathbf{H}(0,0))_{ij} - \mathbf{H}_{ij}^\infty\right| \geq t\right) = \mathbb{P}\left(\left|\frac{1}{m}\sum_{r=1}^m h_r - \mathbf{H}_{ij}^\infty\right| \geq t\right) \leq 2e^{-2mt^2}$$

Apply a union bound over i, j gives that with probability at least $1 - 2n^2 e^{-2mt^2}$ it holds that

$$\left|(\mathbf{m}_{k'}^l \circ \mathbf{H}(0,0))_{ij} - \mathbf{H}_{ij}^\infty\right| \leq t$$

for all $i, j \in [n]$. Therefore,

$$\begin{aligned}
\|\mathbf{m}_{k'}^l \circ \mathbf{H}(0,0) - \mathbf{H}^\infty\|_2^2 &\leq \|\mathbf{m}_{k'}^l \circ \mathbf{H}(0,0) - \mathbf{H}^\infty\|_F^2 \\
&\leq \sum_{i,j=1}^n \left|(\mathbf{m}_{k'}^l \circ \mathbf{H}(0,0))_{ij} - \mathbf{H}_{ij}^\infty\right|^2 \\
&\leq n^2 t^2
\end{aligned}$$

Let $t = \frac{\lambda_0}{4n}$ gives

$$\|\mathbf{m}_{k'}^l \circ \mathbf{H}(0,0) - \mathbf{H}^\infty\|_2 \leq \frac{\lambda_0}{4}$$

holds with probability at least $1 - 2n^2 e^{-\frac{m\lambda_0^2}{8n^2}}$. Next we show that for all $k \in [k]$ and $t \in [\tau]$, as long as $\|\mathbf{w}_{k,t,r} - \mathbf{w}_{0,r}\|_2 \leq R$ for all $r \in [m]$, then it holds that

$$\|\mathbf{m}_{k'}^l \circ \mathbf{H}(k,t) - \mathbf{m}_{k'}^l \circ \mathbf{H}(0,0)\|_2 \leq 2n\kappa^{-1} R$$

Following the argument of *(Song & Yang, 2020)*, lemma 3.2, we have

$$\|\mathbf{m}_{k'}^l \circ \mathbf{H}(k,t) - \mathbf{m}_{k'}^l \circ \mathbf{H}(0,0)\|_F^2 \leq \frac{1}{m^2} \sum_{i,j=1}^n \left(\sum_{r=1}^m s_{r,i,j}\right)^2$$

with

$$s_{r,i,j} = m_{k',r}^l \left(\mathbb{I}\{\langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle \geq 0; \langle \mathbf{w}_{0,r}, \mathbf{x}_j \rangle \geq 0\} - \mathbb{I}\{\langle \mathbf{w}_{r,k,t}^l, \mathbf{x}_i \rangle \geq 0; \langle \mathbf{w}_{r,k,t}^l, \mathbf{x}_j \rangle \geq 0\}\right)$$

Then $s_{r,i,j} = 0$ if $\neg A_{ir}$ and $\neg A_{jr}$ happend. In other cases we have $|s_{r,i,j}| \leq 1$. Thus we have that for all $i, j \in [n]$

$$\mathbb{E}_{\mathbf{M}_k, \mathbf{W}_0}[s_{r,i,j}] = \xi \mathbb{P}(A_{ir} \cup A_{jr}) \leq \frac{4\xi R}{\kappa\sqrt{2\pi}} \leq 2\xi\kappa^{-1} R$$

and

$$\mathbb{E}_{\mathbf{M}_k,\mathbf{W}_0}\left[(s_{r,i,j}-\mathbb{E}_{\mathbf{M}_k,\mathbf{W}_0}[s_{r,i,j}])^2\right]\leq\mathbb{E}_{\mathbf{M}_k,\mathbf{w}_{0,r}}[s_{r,i,j}^2]\leq\frac{4\xi R}{\kappa\sqrt{2\pi}}\leq 2\xi\kappa^{-1}R$$

Thus applying Bernstein inequality with $t=\xi\kappa^{-1}R$ gives

$$\mathbb{P}\left(\frac{1}{m^2}\sum_{r=1}^{m}s_{r,i,j}\geq 3\xi\kappa^{-1}R\right)\leq\exp\left(-\frac{m\xi R}{10\kappa}\right)$$

Therefore, taking a union bound gives that, with probability at least $1-n^2e^{-\frac{m\xi R}{10\kappa}}$ we have that

$$\|\mathbf{m}_{k'}^l\circ\mathbf{H}(k,t)-\mathbf{m}_{k'}^l\circ\mathbf{H}(0,0)\|_2\leq\|\mathbf{m}_{k'}^l\circ\mathbf{H}(k,t)-\mathbf{m}_{k'}^l\circ\mathbf{H}(0,0)\|_F\leq 3\xi n\kappa^{-1}R$$

Using $R\leq\frac{\kappa\lambda_0}{12n}$ gives $\|\mathbf{m}_{k'}^l\circ\mathbf{H}(k,t)-\mathbf{m}_{k'}^l\circ\mathbf{H}(0,0)\|_2\leq\frac{\xi\lambda_0}{4}\leq\frac{\lambda_0}{4}$ with probability at least $1-n^2e^{-\frac{m\xi\lambda_0}{12n}}$. Therefore, we have

$$\|\mathbf{m}_{k'}^l\circ\mathbf{H}(k,t)-\mathbf{H}^\infty\|_2\leq\frac{\lambda_0}{2}$$

which implies that $\lambda_{\min}\left(\mathbf{m}_{k'}^l\circ\mathbf{H}(k,t)\right)\geq\frac{\lambda_0}{2}$ holds with probability at least $1-n^2\left(e^{-\frac{m\xi\lambda_0}{12n}}-2e^{-\frac{m\lambda_0^2}{8n^2}}\right)$ for a fixed $k'\in[K]$ and $l\in[p]$. Taking a union bound over all k' and l and plugging in the requirement $m=\Omega\left(\frac{n^2\log\frac{Kpn}{\delta}}{\xi\lambda_0}\right)$ gives the desired result.

C Proof of Hypothesis 1

In this proof, we follow the idea of *(Du et al., 2018)*. However, the difference is that i) we use our masked-NTK during the analysis, and ii) we use a different technique for bounding the weight perturbation. We repeat the requirement stated in the theorem here:

Suppose that, for all $k'<k$, the expected network loss enjoys a linear convergence with and additional error term

$$\mathbb{E}_{[\mathbf{M}_k],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_{k'+1}\|_2^2\right]\leq(1-\alpha)\,\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_{k'}\|_2^2\right]+\alpha B_1$$

for some $\alpha\leq 1$, and

$$\|\mathbf{w}_{k,r}-\mathbf{w}_{0,r}\|_2+2\eta\tau\sqrt{\frac{2nK}{m\delta}}\left(\frac{1}{\alpha}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right]+(K-k)B\right)\leq R$$

for all $r\in[m]$ with

$$B=\sqrt{\frac{B_1}{\alpha}}+\kappa\sqrt{\xi(1-\xi)pn}$$

We start focusing on local iteration $t\in[\tau]$. We also assume that $\|\mathbf{w}_{k,t',r}-\mathbf{w}_{0,r}\|_2\leq R$, and the probability bound of Lemma 23 holds, and show the local convergence. As in previous works, we will show that this assumption actually holds by induction after we show the local convergence. For the local convergence, we are interested in

$$\|\mathbf{y}-\hat{\mathbf{u}}_{k,t+1}^l\|_2^2=\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^l\|_2^2-2\left\langle\mathbf{y}-\mathbf{u}_{k,t}^l,\hat{\mathbf{u}}_{k,t+1}^l-\hat{\mathbf{u}}_{k,t}^l\right\rangle+\|\hat{\mathbf{u}}_{k,t+1}^l-\hat{\mathbf{u}}_{k,t}^l\|_2^2$$

Again we define $\hat{u}_{k,t+1}^{l(i)}-\hat{u}_{k,t}^{l(i)}=I_{1,k,t}^{l(i)}+I_{2,k,t}^{l(i)}$ with

$$I_{1,k,t}^{l(i)}=\frac{1}{\sqrt{m}}\sum_{r\in S_i}a_r m_{k,r}^l\left(\sigma\left(\left\langle\mathbf{w}_{k,t+1,r}^l,\mathbf{x}_i\right\rangle\right)-\sigma\left(\left\langle\mathbf{w}_{k,t,r}^l,\mathbf{x}_i\right\rangle\right)\right)$$

$$I_{2,k,t}^{l(i)}=\frac{1}{\sqrt{m}}\sum_{r\in S_i^\perp}a_r m_{k,r}^l\left(\sigma\left(\left\langle\mathbf{w}_{k,t+1,r}^l,\mathbf{x}_i\right\rangle\right)-\sigma\left(\left\langle\mathbf{w}_{k,t,r}^l,\mathbf{x}_i\right\rangle\right)\right)$$

with the definition of S_i as defined in equation (8), and notice that, with the 1-Lipchitzness of ReLU,

$$\begin{aligned}
\left|I_{2,k,t}^{l(i)}\right| &\leq \frac{1}{\sqrt{m}} \sum_{r\in S_i^{\perp}} \left|\sigma\left(\left\langle \mathbf{w}_{k,t+1,r}^l, \mathbf{x}_i\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{k,t,r}^l, \mathbf{x}_i\right\rangle\right)\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r\in S_i^{\perp}} \left\|\mathbf{w}_{k,t+1,r}^l - \mathbf{w}_{k,t,r}\right\|_2 \\
&\leq \frac{\eta}{\sqrt{m}} \sum_{r\in S_i^{\perp}} \left\|\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2 \\
&\leq \frac{\eta\sqrt{n}}{m} \sum_{r\in S_i^{\perp}} \left\|\mathbf{y} - \hat{\mathbf{u}}_{k,t}^l\right\|_2 \\
&\leq 4\eta\kappa^{-1}\sqrt{n}R\|\mathbf{y} - \hat{\mathbf{u}}_{k,t}^l\|_2
\end{aligned}$$

where the last inequality uses $|S_i^{\perp}| \leq 4m\kappa^{-1}R$ from Lemma 16. Therefore,

$$\left|\left\langle \mathbf{y} - \hat{\mathbf{u}}_{k,t}^l, \mathbf{I}_{2,k,t}^l\right\rangle\right| \leq \sqrt{n}\max_{i\in[n]}\left|I_{2,k,t}^{l(i)}\right| \cdot \|\mathbf{y} - \hat{\mathbf{u}}_{k,t}^l\|_2 \leq 4\eta\kappa^{-1}nR\|\mathbf{y} - \hat{\mathbf{u}}_{k,t}^l\|_2^2$$

Similarly, we have

$$\begin{aligned}
\left(\hat{u}_{k,t+1}^{l(i)} - \hat{u}_{k,t}^{l(i)}\right)^2 &\leq \frac{1}{m}\left(\sum_{r=1}^m \left\|\mathbf{w}_{k,t+1,r}^l - \mathbf{w}_{k,t,r}\right\|_2\right)^2 \\
&\leq \sum_{r=1}^m \left\|\mathbf{w}_{k,t+1,r}^l - \mathbf{w}_{k,t,r}\right\|_2^2 \\
&\leq \eta^2 \sum_{r=1}^m \left\|\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2^2 \\
&\leq \eta^2 n^2 \|\mathbf{y} - \hat{\mathbf{u}}_{k,t}^l\|_2^2
\end{aligned}$$

Lastly, we define $\mathbf{m}_k^l \circ \mathbf{H}(k,t)^{\perp}$ with

$$\left(\mathbf{m}_k^l \circ \mathbf{H}(k,t)^{\perp}\right)_{ij} = \frac{1}{m}\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle \sum_{r\in S_i^{\perp}} m_{k,r}^l \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0, \langle \mathbf{w}_{k,r}, \mathbf{x}_j\rangle \geq 0\}$$

and we have

$$\begin{aligned}
I_{1,k,t}^{l(i)} &= \frac{1}{\sqrt{m}} \sum_{r\in S_i} a_r m_{k,r}^l \left\langle \mathbf{w}_{k,t+1,r}^l - \mathbf{w}_{k,r}, \mathbf{x}_i\right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} \\
&= -\frac{\eta}{\sqrt{m}} \sum_{r\in S_i} a_r m_{k,r}^l \left\langle \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,r}\right)}{\partial \mathbf{w}_r}, \mathbf{x}_i\right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} \\
&= \frac{\eta}{m} \sum_{r\in S_i} \sum_{j=1}^n m_{k,r}^l \left(y_j - \hat{u}_{k,t}^{l(j)}\right) \langle \mathbf{x}_i, \mathbf{x}_j\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0, \langle \mathbf{w}_{k,r}, \mathbf{x}_j\rangle \geq 0\} \\
&= \eta \sum_{j=1}^n \left(\mathbf{m}_k^l \circ \mathbf{H}(k,t) - \mathbf{m}_k^l \circ \mathbf{H}(k,t)^{\perp}\right)_{ij} \left(y_j - \hat{u}_{k,t}^{l(j)}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}\left\langle \mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l},\mathbf{I}_{1,k,t}^{l}\right\rangle &= \eta\sum_{i,j=1}^{n}\left(y_i-\hat{u}_{k,t}^{l(i)}\right)\left(\mathbf{m}_k^l\circ\mathbf{H}(k,t)-\mathbf{m}_k^l\circ\mathbf{H}(k,t)^{\perp}\right)_{ij}\left(y_j-\hat{u}_{k,t}^{l(j)}\right)\\ &= \eta\left\langle \mathbf{y}-\hat{\mathbf{u}}_{k,t},\left(\mathbf{m}_k^l\circ\mathbf{H}(k,t)-\mathbf{m}_k^l\circ\mathbf{H}(k,t)^{\perp}\right)(\mathbf{y}-\hat{\mathbf{u}}_{k,t})\right\rangle\\ &\geq \frac{\eta\lambda_0}{2}\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2-\eta\|\mathbf{m}_k^l\circ\mathbf{H}(k,t)^{\perp}\|_2\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2\\ &\geq \left(\frac{\eta\lambda_0}{2}-4\eta\kappa^{-1}nR\right)\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2\end{aligned}$$

where the last inequality follows from the fact that

$$\begin{aligned}\|\mathbf{m}_k^l\circ\mathbf{H}(k,t)^{\perp}\|_2^2 &\leq \|\mathbf{m}_k^l\circ\mathbf{H}(k,t)^{\perp}\|_F^2\\ &= \frac{1}{m^2}\sum_{i,j=1}^{n}\left(\langle\mathbf{x}_i,\mathbf{x}_j\rangle\sum_{r\in S_i^{\perp}}\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0,\langle\mathbf{w}_{k,r},\mathbf{x}_j\rangle\geq 0\}\right)^2\\ &\leq \frac{n^2}{m^2}|S_i^{\perp}|^2\\ &= 16n^2\kappa^{-2}R^2\end{aligned}$$

Putting things together gives

$$\|\mathbf{y}-\hat{\mathbf{u}}_{k,t+1}^{l}\|_2^2\leq\left(1-\eta\lambda_0+16\eta\kappa^{-1}nR+\eta^2n^2\right)\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2$$

Choose $R\leq\frac{\kappa\lambda_0}{64n}$ and $\eta\leq\frac{\lambda_0}{4n^2}$ gives

$$\|\mathbf{y}-\hat{\mathbf{u}}_{k,t+1}^{l}\|_2^2\leq\left(1-\frac{\eta\lambda_0}{2}\right)\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2$$

Therefore, for all $t\in[\tau]$ we have

$$\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2\leq\left(1-\frac{\eta\lambda_0}{2}\right)^t\|\mathbf{y}-\hat{\mathbf{u}}_{k}^{l}\|_2^2$$

Next we bound the weight change during the local steps. We have

$$\begin{aligned}\|\mathbf{w}_{k,t,r}-\mathbf{w}_{k,r}\|_2 &\leq \eta\sum_{t'=0}^{t-1}\left\|\frac{\partial L_{\mathbf{m}_k}\left(\mathbf{W}_{k,t}^{l}\right)}{\partial\mathbf{w}_r}\right\|_2\\ &\leq \eta\frac{\sqrt{n}}{\sqrt{m}}\sum_{t'=0}^{t-1}\|\mathbf{y}-\hat{\mathbf{u}}_{k,t'}^{l}\|_2\\ &\leq \eta\tau\frac{\sqrt{n}}{\sqrt{m}}\|\mathbf{y}-\hat{\mathbf{u}}_{k}^{l}\|_2\\ &\leq \eta\tau\frac{\sqrt{n}}{\sqrt{m}}\left(\|\mathbf{y}-\mathbf{u}_k\|_2+\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2\right)\end{aligned}$$

Applying Markov's inequality to the global convergence, with probabiltiy at least $1-\frac{\delta}{2K}$, it holds that

$$\|\mathbf{y}-\mathbf{u}_k\|_2\leq\sqrt{2K/\delta}\,\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right]$$

By Lemma 25, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2\right]\leq 4\xi(1-\xi)n\kappa^2$$

Thus with probability at least $1-\frac{\delta}{2pK}$ it holds that

$$\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2\leq 2\kappa\sqrt{2\xi(1-\xi)npK/\delta}$$

Plugging in gives

$$\|\mathbf{w}_{k,t,r} - \mathbf{w}_{k,r}\| \leq \frac{\eta\tau\sqrt{2nK}}{\sqrt{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y} - \mathbf{u}_k\|_2\right] + 2\eta\tau\kappa n\sqrt{\frac{2\xi(1-\xi)pK}{m\delta}}$$

Since for iterations $k' \leq k$, we have

$$\|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 + 2\eta\tau\sqrt{\frac{2nK}{m\delta}}\left(\frac{1}{\alpha}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y} - \mathbf{u}_k\|_2\right] + (K-k)B\right) \leq R$$

Since $k < K$ and $\alpha \leq 1$, and

$$B = \sqrt{\frac{B_1}{\alpha}} + \kappa\sqrt{\xi(1-\xi)pn} \geq \kappa\sqrt{\xi(1-\xi)pn}$$

we have that

$$\|\mathbf{w}_{k,t,r} - \mathbf{w}_{0,r}\|_2 \leq \|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 + \|\mathbf{w}_{k,t,r} - \mathbf{w}_{k,r}\|_2 \leq R$$

which completes the proof.

D Proof of Theorem 3

Before we start the proof, we introduce several notations. Define

$$I_{1,k}^{(i)} = \frac{\xi}{\sqrt{m}}\sum_{r\in S_i} a_r\left(\sigma(\langle\mathbf{w}_{k+1,r},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)$$

$$I_{2,k}^{(i)} = \frac{\xi}{\sqrt{m}}\sum_{r\in S_i^\perp} a_r\left(\sigma(\langle\mathbf{w}_{k+1,r},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)$$

Let

$$\mathbf{I}_{1,k} = \left[I_{1,k}^{(1)},\dots,I_{1,k}^{(n)}\right]$$

and similarly,

$$\mathbf{I}_{2,k} = \left[I_{2,k}^{(1)},\dots,I_{2,k}^{(n)}\right]$$

Then we have $u_{k+1}^{(i)} - u_k^{(i)} = I_1^{(i)} + I_2^{(i)}$ and $\mathbf{u}_{k+1} - \mathbf{u}_k = \mathbf{I}_{1,k} + \mathbf{I}_{2,k}$. Also, we define $\mathbf{H}(k)^\perp$ to be

$$\mathbf{H}(k)_{ij}^\perp = \frac{\xi}{m}\sum_{r\in S_i^\perp}\langle\mathbf{x}_i,\mathbf{x}_j\rangle\,\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0, \langle\mathbf{w}_{k,r},\mathbf{x}_j\rangle \geq 0\}$$

For k-th global iteration, first local iteration, we define the mixing gradient as

$$\begin{aligned}
\mathbf{g}_{k,r} &= \eta_{k,r}\sum_{l=1}^p \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,0}^l\right)}{\partial\mathbf{w}_r}\\
&= \eta_{k,r}\sum_{l=1}^p \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r}\\
&= \frac{\eta_{k,r}}{\sqrt{m}}\sum_{l=1}^p\sum_{i=1}^n m_{k,r}^l(\hat{u}_k^{l(i)} - y_i)a_r\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0\}\\
&= \frac{1}{\sqrt{m}}\sum_{i=1}^m(f_{k,r}^{(i)} - N_{k,r}^\perp y_i)a_r\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0\}
\end{aligned}$$

where we define the mixing function as

$$f_{k,r}^{(i)} = \eta_{k,r}\sum_{l=1}^p m_{k,r}^l\hat{u}_k^{l(i)}$$

As usual, we let $\mathbf{f}_{k,r} = \left[f_{k,r}^{(1)}, \dots, f_{k,r}^{(n)}\right]$. We note that $f_{k,r}^{(i)}$ has the form

$$\begin{aligned} f_{k,r}^{(i)} &= \eta_{k,r} \sum_{l=1}^{p} m_{k,r}^{l} \hat{u}_k^{(l(i)} \\ &= \frac{1}{\sqrt{m}} \sum_{r'=1}^{m} a_r \left(\eta_{k,r} \sum_{l=1}^{p} m_{k,r}^{l} m_{k,r'}^{l} \right) \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle) \end{aligned}$$

Let $\nu_{k,r,r'} = \eta_{k,r} \sum_{l=1}^{p} m_{k,r}^{l} m_{k,r'}^{l}$. The mixing function reduce to the form

$$f_{k,r}^{(i)} = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \nu_{k,r,r'} \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle)$$

Also, note that if $N_{k,r}^{\perp} = 0$, we have $\nu_{k,r,r'} = 0$.

We prove Theorem 3 by a fashion of induction. Assume that for $k' < k$, we have

$$\mathbb{E}_{[\mathbf{M}_{k'}],\mathbf{W}_0,\mathbf{a}} \left[\|\mathbf{y} - \mathbf{u}_{k'+1}\|_2^2\right] \leq (1-\alpha)\mathbb{E}_{[\mathbf{M}_{k'-1}],\mathbf{W}_0,\mathbf{a}} \left[\|\mathbf{y} - \mathbf{u}_{k'}\|_2^2\right] + B_1$$

for some $\alpha \leq 1$, and the weight perturbation before iteration k is bounded by

$$\|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 + 2\eta\tau\sqrt{\frac{2nK}{m\delta}} \left(\frac{1}{\alpha} \mathbb{E}_{[\mathbf{M}_{k-1}]} \left[\|\mathbf{y} - \mathbf{u}_k\|_2\right] + (K-k)B \right) \leq R$$

with $\alpha = \frac{1}{4}\eta\theta\tau\lambda_0$, $R = O\left(\frac{\kappa\lambda_0}{n}\right)$ and

$$\begin{aligned} B_1 = & \frac{64(1-\xi)^2 n^3 d}{m\lambda_0^2} + \frac{68\eta\tau(\theta-\xi^2)n^3\kappa^2}{p\lambda_0} + \frac{(\theta-\xi^2)n\kappa^2}{6\tau^2 p} + \frac{(\tau-1)^2 pC_1}{24\tau^2} + \\ & \frac{8\eta^2(\tau-1)^2 n^4 pC_1}{\lambda_0^2} + 4\eta(\tau-1)pnC_1 \end{aligned}$$

and

$$B = \sqrt{\frac{B_1}{\alpha}} + \kappa\sqrt{\xi(1-\xi)pn}$$

With such conditions, we have that with probability at least $1 - 4\delta$, Hypothesis 1 holds. We start by showing

$$\mathbb{E}_{\mathbf{M}_k}[\|\mathbf{y} - \mathbf{u}_{k+1}\|_2^2] \leq (1-\alpha)\|\mathbf{y} - \mathbf{u}_k\|_2^2 + \alpha B_1$$

After that, we show that

$$\|\mathbf{w}_{k+1,r} - \mathbf{w}_{0,r}\|_2 + 2\eta\tau\sqrt{\frac{2nK}{m\delta}} \left(\frac{1}{\alpha} \mathbb{E}_{[\mathbf{M}_k]} \left[\|\mathbf{y} - \mathbf{u}_{k+1}\|_2\right] + (K-k)B \right) \leq R$$

to complete the proof. Throughout this proof, we assume that

$$\sum_{i=1}^{n} \sum_{r'=1}^{m} \langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle^2 \leq 2mn\kappa^2 - mnR^2$$

and that

$$\|W_0\|_F \leq \sqrt{2md} - \sqrt{m}R$$

Note that Lemma 22 and Lemma 23 shows that, as long as $m = \Omega\left(\log\frac{n}{\delta}\right)$, the above assumption holds with probability at least $1 - \delta$ over initialization. Moreover, Lemma 16 shows that as long as $m = \left(\frac{n\log\frac{n}{\delta}}{\xi\lambda_0}\right)$, with probability at least $1 - \delta$ over initialization we have

$$|S_i^{\perp}| \leq 4m\kappa^{-1}R$$

To start, expanding the loss at iteration $k+1$ gives

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right] &= \|\mathbf{y}-\mathbf{u}_k\|_2^2 - 2\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{u}_{k+1}-\mathbf{u}_k\right]\right\rangle + \mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right] \\ &= \|\mathbf{y}-\mathbf{u}_k\|_2^2 - 2\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{1,k}\right]\right\rangle - 2\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{2,k}\right]\right\rangle + \\ &\quad \mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right]\end{aligned}$$

Following previous work, we bound the second, third, and fourth term separately. However, the second term requires a more detailed analysis. In particular, we let

$$\mathbf{I}'_{1,k} = \mathbf{I}_{1,k} - \eta\theta\tau\mathbf{H}(k)(\mathbf{y}-\mathbf{u}_k)$$

Then the loss at iteration $k+1$ has the form

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right] &= \|\mathbf{y}-\mathbf{u}_k\|_2^2 - 2\eta\theta\tau\left\langle \mathbf{y}-\mathbf{u}_k, \mathbf{H}(k)(\mathbf{y}-\mathbf{u}_k)\right\rangle + \mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right] - \\ &\quad 2\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}'_{1,k}\right]\right\rangle - 2\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{2,k}\right]\right\rangle \\ &\leq (1-\eta\theta\tau\lambda_0)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + 2\left|\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}'_{1,k}\right]\right\rangle\right| + \\ &\quad 2\left|\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{2,k}\right]\right\rangle\right| + \mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right]\end{aligned}$$

where in the last inequality we use $\lambda_{\min}(\mathbf{H}(k)) \geq \frac{\lambda_0}{2}$ from *(Du et al., 2018)*, Assumption 3.1. Moreover, Lemma 6, Lemma 9, and Lemma 10 shows that under the given assumption, with $\eta = O\left(\frac{\lambda_0}{n\tau\max\{n,p\}}\right)$ we have

$$\left|\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}'_{1,k}\right]\right\rangle\right| \leq \frac{1}{8}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{16\eta\theta\tau\xi^2(1-\xi)^2\kappa^2n^3d}{m\lambda_0} + \frac{2\eta^3\xi^2\tau(\tau-1)^2n^4pC_1}{\theta\lambda_0}$$

$$\left|\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{2,k}\right]\right\rangle\right| \leq \frac{1}{8}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{\eta\lambda_0\xi^2(\theta-\xi^2)n\kappa^2}{24p\tau} + \frac{\eta\lambda_0\xi^2(\tau-1)^2pC_1}{96\tau\theta}$$

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right] \leq \frac{1}{4}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{17\eta^2\xi^2\tau^2\theta(\theta-\xi^2)n^3\kappa^2}{p} + \eta^2\xi^2\lambda_0(\tau-1)^2pnC_1$$

Putting things together gives

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right] &\leq \left(1-\frac{1}{4}\eta\theta\tau\lambda_0\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{16\eta\theta\tau\xi^2(1-\xi)^2n^3d}{m\lambda_0} + \frac{2\eta^3\xi^2\tau(\tau-1)^2n^4pC_1}{\theta\lambda_0} \\ &\quad \frac{\eta\lambda_0\xi^2(\theta-\xi^2)n\kappa^2}{24\tau p} + \frac{\eta\lambda_0\xi^2(\tau-1)^2pC_1}{96\tau\theta} + \frac{17\eta^2\xi^2\tau^2\theta(\theta-\xi^2)n^3\kappa^2}{p} + \\ &\quad \eta^2\xi^2\lambda_0\tau(\tau-1)pnC_1 \\ &\leq \left(1-\frac{1}{4}\eta\theta\tau\lambda_0\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{1}{4}\eta\theta\tau\lambda_0B_1\end{aligned}$$

Therefore, we have

$$\mathbb{E}_{[\mathbf{M}_{k-1}]}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] \leq \left(1-\frac{1}{4}\eta\theta\tau\lambda_0\right)^k\|\mathbf{y}-\mathbf{u}_0\|_2^2 + B_1$$

This completes the first part of the proof. To show the second part, using Jensen's inequality, we have $\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2\right] \leq E_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right]^{\frac{1}{2}}$. With $\alpha = \frac{1}{4}\eta\theta\tau\lambda_0$, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2\right] \leq \left(1-\frac{\alpha}{2}\right)\|\mathbf{y}-\mathbf{u}_k\|_2 + \sqrt{\alpha B_1}$$

and thus by taking total expectation, we have

$$\mathbb{E}_{[\mathbf{M}_k],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2\right] \leq \left(1-\frac{\alpha}{2}\right)\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right] + \sqrt{\alpha B_1}$$

Next, we bound the weight perturbation using induction. Recall our induction hypothesis

$$\|\mathbf{w}_{k,r}-\mathbf{w}_{0,r}\|_2 + 2\eta\tau\sqrt{\frac{2nK}{m\delta}}\left(\frac{1}{\alpha}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right] + (K-k)B\right) \leq R$$

We would like to show that

$$\|\mathbf{w}_{k+1,r}-\mathbf{w}_{0,r}\|_2+2\eta\tau\sqrt{\frac{2nK}{m\delta}}\left(\frac{1}{\alpha}\mathbb{E}_{[\mathbf{M}_k],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2\right]+(K-k-1)B\right)\leq R$$

Using the convergence above and the fact that

$$\|\mathbf{w}_{k+1,r}-\mathbf{w}_{0,r}\|_2\leq\|\mathbf{w}_{k,r}-\mathbf{w}_{0,r}\|_2+\|\mathbf{w}_{k+1,r}-\mathbf{w}_{0,r}\|_2$$

it suffice to show that

$$\|\mathbf{w}_{k+1,r}-\mathbf{w}_{k,r}\|_2\leq\eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right]+2\eta\tau\sqrt{\frac{2nK}{m\delta}}B-2\eta\tau\sqrt{\frac{2nKB_1}{m\delta\alpha}}$$

Recall the definition of B as

$$B=\sqrt{\frac{B_1}{\alpha}}+\kappa\sqrt{\xi(1-\xi)pn}$$

It then suffice to show that

$$\|\mathbf{w}_{k+1,r}-\mathbf{w}_{0,r}\|_2\leq\eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right]+2\eta\tau\kappa n\sqrt{\frac{2\xi(1-\xi)pK}{m\delta}}$$

By Hypothesis 1 we have

$$\|\mathbf{w}_{k,t,r}^l-\mathbf{w}_{k,r}\|_2\leq\eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right]+2\eta\tau\kappa n\sqrt{\frac{2\xi(1-\xi)pK}{m\delta}}$$

for all $l\in[p]$ and $t\in[\tau]$. Then we have

$$\begin{aligned}\|\mathbf{w}_{k+1,r}-\mathbf{w}_{k,r}\|_2&\leq\eta_{k,r}\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{w}_{k,\tau,r}^l-\mathbf{w}_{k,r}\|_2\\&\leq\eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right]+2\eta\tau\kappa n\sqrt{\frac{2\xi(1-\xi)pK}{m\delta}}\end{aligned}$$

which completes the proof of the second part. Next, we use over-paramterization to show that the base case also satisfies the condition above. In particular, we want to show that

$$2\eta\tau\sqrt{\frac{2nK}{m\delta}}\left(\frac{1}{\alpha}\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2\right]+KB\right)\leq R\leq\frac{\kappa\lambda_0}{144n}$$

Equivalently, we want

$$\frac{\kappa\lambda_0}{\eta\tau}\sqrt{\frac{m\delta}{n^3K}}=\Omega\left(\max\left\{\frac{1}{\alpha}\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2\right]+KB\right\}\right)$$

We use Lemma 26 to get that

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2^2\right]\leq C^2n$$

Plugging in the bound above and $\alpha=\frac{1}{4}\eta\theta\tau\lambda_0$, and solving for

$$\frac{\kappa\lambda_0}{\eta\tau}\sqrt{\frac{m\delta}{n^3K}}=\Omega\left(\frac{1}{\alpha}\|\mathbf{y}-\mathbf{u}_0\|_2\right)$$

gives that

$$m_1=\Omega\left(\frac{n^4K}{\kappa^2\theta^2\lambda_0^4\delta}\right)$$

Solving for

$$\frac{\kappa\lambda_0}{\eta\tau}\sqrt{\frac{m\delta}{n^3K}}=\Omega(KB)$$

gives that

$$m_2 = \Omega\left(\frac{n^3 K^3 B^2 \eta^2 \tau^2}{\kappa^2 \lambda_0^2 \delta}\right)$$

Using the requirement of η gives

$$m_2 = \Omega\left(\frac{\eta n K^3 B^2 \tau}{\kappa^2 \lambda_0 \delta}\right)$$

Plugging in the definition of B gives

$$m_2 = \Omega\left(\frac{n K^3 B_1}{\theta \kappa^2 \lambda_0^2 \delta} + \frac{\xi(1-\xi) K^3 p}{\delta}\right)$$

Combining the three requirements above gives

$$m = \Omega\left(\frac{K}{\delta} \max\left\{\frac{n^4}{\kappa^2 \xi \theta \lambda_0^4}, \frac{n K^2 B_1}{\kappa^2 \theta \lambda_0^2}, K^2 p\right\}\right)$$

E LEMMAS FOR THEOREM 3

Lemma 1. *The expectation of the mixing function satisfies*

$$\mathbb{E}_{\mathbf{M}_k}\left[f_{k,r}^{(i)}\right] = \theta u_k^{(i)} + \frac{\theta(1-\xi)}{\sqrt{m}} a_r \sigma(\langle \mathbf{w}_{k,r'}, \mathbf{x}_i \rangle)$$

Proof. Note that if $N_{k,r}^{\perp} = 0$, then we have $f_{k,r}^{(i)} = 0$ for all $i \in [n]$. Thus

$$\mathbb{E}_{\mathbf{M}_k}\left[f_{k,r}^{(i)} \mid N_{k,r}^{\perp} = 0\right] = 0$$

Moreover, if $N_{k,r}^{\perp} = 1$, the expectation can be computed as

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[f_{k,r}^{(i)} \mid N_{k,r}^{\perp} = 1\right] &= \mathbb{E}_{\mathbf{M}_k}\left[\frac{\eta_{k,r}}{\sqrt{m}} \sum_{l=1}^{p} \sum_{r'=1}^{m} m_{k,r}^{l} m_{k,r'}^{l} a_r \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle) \mid N_{k,r}^{\perp} = 1\right] \\
&= \frac{1}{\sqrt{m}} \sum_{r'=1}^{m} \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r} \sum_{l=1}^{p} m_{k,r}^{l} m_{k,r'}^{l} \mid N_{k,r}^{\perp} = 1\right] a_{r'} \sigma(\langle \mathbf{w}_{k,r'}, \mathbf{x}_i \rangle) \\
&= \frac{\xi}{\sqrt{m}} \sum_{r'=1}^{m} a_{r'} \sigma(\langle \mathbf{w}_{k,r'}, \mathbf{x}_i \rangle) + \frac{1-\xi}{\sqrt{m}} a_r \sigma(\langle \mathbf{w}_{k,r'}, \mathbf{x}_i \rangle)
\end{aligned}$$

by using Lemma 20. Combining the two conditions above gives that

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[f_{k,r}^{(i)}\right] &= \mathbb{P}(N_{k,r}^{\perp} = 1) \mathbb{E}_{\mathbf{M}_k}\left[f_{k,r}^{(i)} \mid N_{k,r}^{\perp} = 1\right] + \mathbb{P}(N_{k,r}^{\perp} = 0) \mathbb{E}_{\mathbf{M}_k}\left[f_{k,r}^{(i)} \mid N_{k,r}^{\perp} = 0\right] \\
&= \frac{\theta \xi}{\sqrt{m}} \sum_{r'=1}^{m} a_{r'} \sigma(\langle \mathbf{w}_{k,r'}, \mathbf{x}_i \rangle) + \frac{\theta(1-\xi)}{\sqrt{m}} a_r \sigma(\langle \mathbf{w}_{k,r'}, \mathbf{x}_i \rangle) \\
&= \theta u_k^{(i)} + \frac{\theta(1-\xi)}{\sqrt{m}} a_r \sigma(\langle \mathbf{w}_{k,r'}, \mathbf{x}_i \rangle)
\end{aligned}$$

□

Lemma 2. *The expectation of the mixing gradient satisfies*

$$\mathbb{E}_{\mathbf{M}_k}\left[\mathbf{g}_{k,r}\right] = \frac{\theta}{\xi} \frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r} + \frac{\theta(1-\xi)}{m} \sum_{i=1}^{n} \mathbf{x}_i \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle)$$

Proof. With the result from Lemma 1, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}[\mathbf{g}_{k,r}] &= \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\left(\mathbb{E}_{\mathbf{M}_k}\left[f_{k,r}^{(i)}\right] - y_i\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\right]\right)a_r\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle \geq 0\}\\ &= \frac{\theta}{\sqrt{m}}\sum_{i=1}^{n}\left(u_k^{(i)} - y_i + \frac{1-\xi}{\sqrt{m}}a_r\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)a_r\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle \geq 0\}\\ &= \frac{\theta}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r} + \frac{\theta(1-\xi)}{m}\sum_{i=1}^{n}\mathbf{x}_i\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\end{aligned}$$

□

Lemma 3. *Suppose $m \geq p$. If for some $R > 0$ and all $r \in [m]$ the initialization satisfies*

$$\sum_{r=1}^{m}\langle\mathbf{w}_{0,r},\mathbf{x}_i\rangle^2 \leq 2mn\kappa^2 - mnR^2$$

and for all $r \in [m]$, it holds that $\|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 \leq R$, the expected norm of the difference between the mixing function and $u_k^{(i)}$ satisfies

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{f}_{k,r} - \mathbf{u}_k\|_2^2 \mid N_{k,r}^{\perp} = 1\right] \leq \frac{8(\theta-\xi^2)n\kappa^2}{p}$$

Proof. Since $\mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r'} \mid N_{k,r}^{\perp} = 1\right] = \xi$ for $r' \neq r$, we have for $r_1 \neq r_2$, there is at least one of r_1, r_2 that is not r. Thus

$$\mathbb{E}_{\mathbf{M}_k}\left[(\nu_{k,r,r_1} - \xi)(\nu_{k,r,r_2} - \xi) \mid N_{k,r}^{\perp} = 1\right] = 0$$

and for $r \neq r'$

$$\mathrm{Var}_{\mathbf{M}_k}\left(\nu_{k,r,r'} \mid N_{k,r}^{\perp} = 1\right) = \mathbb{E}_{\mathbf{M}_k}\left[(\nu_{k,r,r'} - \xi)^2 \mid N_{k,r}^{\perp} = 1\right]$$

Moreover, for $r = r'$, Lemma 19

$$\mathbb{E}_{\mathbf{M}_k}\left[(\nu_{k,r,r,} - \xi)^2\right] \leq \theta - \xi^2$$

Therefore, using Lemma 21 we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\left(f_{k,r}^{(i)} - u_k^{(i)}\right)^2 \mid N_{k,r}^{\perp} = 1\right] &= \frac{1}{m}\mathbb{E}_{\mathbf{M}_k}\left[\left(\sum_{r'\neq r}^{m}a_r(\nu_{k,r,r'} - \xi)\sigma(\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle)\right)^2 \mid N_{k,r}^{\perp} = 1\right]\\ &= \frac{1}{m}\sum_{r'=1}^{m}\mathrm{Var}_{\mathbf{M}_k}\left(\nu_{k,r,r'} \mid N_{k,r}^{\perp} = 1\right)\sigma(\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle)^2 +\\ &\qquad \frac{1}{m}\mathbb{E}_{\mathbf{M}_k}\left[(\nu_{k,r,r,} - \xi)^2\right]\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)^2\\ &\leq \frac{\theta-\xi^2}{pm}\sum_{r'\neq r}^{m}\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle^2 + \frac{\theta-\xi^2}{m}\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)^2\\ &\leq \frac{2(\theta-\xi^2)}{pm}\left(\sum_{r'=1}^{m}\langle\mathbf{w}_{0,r},\mathbf{x}_i\rangle^2 + mR^2\right) + \frac{2(\theta-\xi^2)}{p}\left(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle + R^2\right)\\ &\leq \frac{8(\theta-\xi^2)\kappa^2}{p}\end{aligned}$$

Plugging this in gives

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{f}_{k,r} - \mathbf{u}_k\|_2^2 \mid N_{k,r}^{\perp} = 1\right] \leq \frac{8(\theta-\xi^2)n\kappa^2}{p}$$

□

Lemma 4. *Under the condition of Lemma 3, the expected norm and squared-norm of the mixing gradient is bounded by*

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2^2\right] \leq \frac{2n\theta}{m}\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{16\theta(\theta-\xi^2)n^2\kappa^2}{pm}$$
$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2\right] \leq \frac{\sqrt{n}\theta}{\sqrt{m}}\|\mathbf{y}-\mathbf{u}_k\|_2 + 4n\kappa\sqrt{\frac{\theta(\theta-\xi^2)}{pm}}$$

Proof. Using Lemma 3, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\|\mathbf{f}_{k,r}-\mathbf{u}_k\|_2^2\right] = P(N_{k,r}^{\perp}=1)\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{f}_{k,r}-\mathbf{u}_k\|_2^2\right] \leq \frac{8\theta(\theta-\xi^2)n\kappa^2}{p}$$

According to Jensen's inequality, we also have

$$\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\|\mathbf{f}_{k,r}-\mathbf{u}_k\|_2\right] \leq 2\kappa\cdot\sqrt{\frac{2\theta(\theta-\xi^2)n}{p}}$$

Moreover, we have

$$\begin{aligned}
\mathbf{g}_{k,r} &= \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\left(f_{k,r}^{(i)}-y_i\right)a_r N_{k,r}^{\perp}\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0\} \\
&= \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\left(f_{k,r}^{(i)}-u_k^{(i)}\right)a_r N_{k,r}^{\perp}\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0\} + \frac{N_{k,r}^{\perp}}{\xi}\cdot\frac{\partial L(\mathbf{W}_k)}{\partial\mathbf{w}_r}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2^2\right] &\leq \frac{2\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\right]}{\xi^2}\left\|\frac{\partial L(\mathbf{W}_k)}{\partial\mathbf{w}_r}\right\|_2^2 + \frac{2}{m}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\sum_{i=1}^{n}\left(f_{k,r}^{(i)}-u_k^{(i)}\right)a_r N_{k,r}^{\perp}\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0\}\right\|_2^2\right] \\
&\leq \frac{2n}{m}\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\sum_{i=1}^{n}\left(f_{k,r}^{(i)}-u_k^{(i)}\right)^2\right] + \frac{2n\theta}{m}\|\mathbf{y}-\mathbf{u}_k\|_2^2 \\
&\leq \frac{2n}{m}\left(\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\|\mathbf{f}_{k,r}-\mathbf{u}_k\|_2^2\right] + \theta\|\mathbf{y}-\mathbf{u}_k\|_2^2\right) \\
&\leq \frac{2n\theta}{m}\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{16\theta(\theta-\xi^2)n^2\kappa^2}{pm}
\end{aligned}$$

This shows the first inequality. To show the second, similarly we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2\right] &\leq \frac{\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\right]}{\xi}\left\|\frac{\partial L(\mathbf{W}_k)}{\partial\mathbf{w}_r}\right\|_2 + \frac{1}{\sqrt{m}}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\sum_{i=1}^{n}\left(f_{k,r}^{(i)}-u_k^{(i)}\right)a_r N_{k,r}^{\perp}\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\geq 0\}\right\|_2\right] \\
&\leq \frac{1}{\sqrt{m}}\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\sum_{i=1}^{n}\left|f_{k,r}^{(i)}-u_k^{(i)}\right|\right] + \frac{\sqrt{n}\theta}{\sqrt{m}}\|\mathbf{y}-\mathbf{u}_k\|_2 \\
&\leq \sqrt{\frac{n}{m}}\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\|\mathbf{f}_{k,r}-\mathbf{u}_k\|_2\right] + \frac{\sqrt{n}\theta}{\sqrt{m}}\|\mathbf{y}-\mathbf{u}_k\|_2 \\
&\leq \frac{\sqrt{n}\theta}{\sqrt{m}}\|\mathbf{y}-\mathbf{u}_k\|_2 + 4n\kappa\sqrt{\frac{\theta(\theta-\xi^2)}{pm}}
\end{aligned}$$

□

Lemma 5. *Under the condition of Theorem 3, we have*

$$\left|\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right| \leq \eta t\sqrt{n}\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2$$

and therefore,

$$\left|\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right| \leq \eta t \sqrt{n} \left(\|\mathbf{y} - \mathbf{u}_k\|_2 + \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right)$$
$$\left(\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right)^2 \leq 2\eta^2 t^2 n \left(\|\mathbf{y} - \mathbf{u}_k\|_2^2 + \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2\right)$$

Proof. We have

$$\begin{aligned}
\left|\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right| &= \frac{1}{\sqrt{m}} \left|\sum_{r=1}^{m} a_r m_{k,r}^l \left(\sigma\left(\left\langle \mathbf{w}_{k,t,r}^l, \mathbf{x}_i\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{k,r}, \mathbf{x}_i\right\rangle\right)\right)\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \left|\sigma\left(\left\langle \mathbf{w}_{k,t,r}^l, \mathbf{x}_i\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{k,r}, \mathbf{x}_i\right\rangle\right)\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \left\|\mathbf{w}_{k,t,r}^l - \mathbf{w}_{k,r}\right\|_2 \\
&\leq \frac{\eta}{\sqrt{m}} \sum_{r=1}^{m} \sum_{t'=0}^{t-1} \left\|\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2 \\
&\leq \eta\sqrt{n} \sum_{t'=0}^{t-1} \|\mathbf{y} - \hat{\mathbf{u}}_{k,t}^l\|_2 \\
&\leq \eta t \sqrt{n} \|\mathbf{y} - \hat{\mathbf{u}}_k^l\|_2
\end{aligned}$$

Therefore,

$$\left|\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right| \leq \eta t \sqrt{n} \left(\|\mathbf{y} - \mathbf{u}_k\|_2 + \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right)$$

Moreover,

$$\left(\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right)^2 = \left|\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right|^2 \leq 2\eta^2 t^2 n \left(\|\mathbf{y} - \mathbf{u}_k\|_2^2 + \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2\right)$$

□

Lemma 6. *Under the condition of Lemma 3, with* $\eta \leq \frac{\lambda_0}{16(\tau-1)n^2}$*, we have*

$$\left|\left\langle \mathbf{y} - \mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{1,k}'\right]\right\rangle\right| \leq \frac{1}{8}\eta\theta\tau\lambda_0 \|\mathbf{y} - \mathbf{u}_k\|_2^2 + \frac{16\eta\theta\tau\xi^2(1-\xi)^2\kappa^2 n^3 d}{m\lambda_0} + \frac{2\eta^3\xi^2\tau(\tau-1)^2 n^4 p C_1}{\theta\lambda_0}$$

Proof. We start by analyzing $\mathbf{w}_{k+1,r} - \mathbf{w}_{k,r}$. Taking expectation, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\mathbf{w}_{k+1,r} - \mathbf{w}_{k,r}\right] &= -\eta \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r} \sum_{l=1}^{p} \sum_{t=0}^{\tau-1} \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right] \\
&= -\eta \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\tau \sum_{l=1}^{p} \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r} + \eta_{k,r} \sum_{l=1}^{p} \sum_{t=1}^{\tau-1} \left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right] \\
&= -\eta\tau \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{g}_{k,r}\right] - \eta \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r} \sum_{l=1}^{p} \sum_{t=1}^{\tau-1} \left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right] \\
&= -\frac{\eta\theta\tau}{\xi} \frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r} - \frac{\eta\theta(1-\xi)\tau}{m} \sum_{i=1}^{n} \mathbf{x}_i \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle) - \\
&\qquad \eta \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r} \sum_{l=1}^{p} \sum_{t=1}^{\tau-1} \left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right]
\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[I_{1,k}^{(i)}\right] &= \frac{\xi}{\sqrt{m}}\sum_{r\in S_i} a_r \mathbb{E}_{\mathbf{M}_k}\left[\sigma(\langle \mathbf{w}_{k+1,r}, \mathbf{x}_i\rangle) - \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle)\right] \\ &= \frac{\xi}{\sqrt{m}}\sum_{r\in S_i} a_r \left\langle \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{w}_{k+1,r} - \mathbf{w}_{k,r}\right], \mathbf{x}_i\right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} \\ &= -\frac{\eta\theta\tau}{\sqrt{m}}\sum_{r\in S_i} a_r \left\langle \frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r}, \mathbf{x}_i\right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} - \eta\left(\mathcal{E}_{1,k}^{(i)} + \mathcal{E}_{2,k}^{(i)}\right) \\ &= \frac{\eta\xi\theta\tau}{m}\sum_{r\in S_i}\sum_{j=1}^{n}(y_i - u_k^{(i)})\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0, \langle \mathbf{w}_{k,r}, \mathbf{x}_j\rangle \geq 0\} - \eta\left(\mathcal{E}_{1,k}^{(i)} + \mathcal{E}_{2,k}^{(i)}\right) \\ &= \eta\theta\tau\sum_{j=1}^{n}\left(\mathbf{H}(k)_{ij} - \mathbf{H}(k)_{ij}^{\perp}\right)(y_j - u_k^{(j)}) - \eta\left(\mathcal{E}_{1,k}^{(i)} + \mathcal{E}_{2,k}^{(i)}\right)\end{aligned}$$

where

$$\mathcal{E}_{1,k}^{(i)} = \frac{\theta\xi(1-\xi)\tau}{m^{\frac{3}{2}}}\sum_{r\in S_i}\sum_{j=1}^{n} a_r \left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_j\rangle)$$

$$\mathcal{E}_{2,k}^{(i)} = \frac{\xi}{\sqrt{m}}\sum_{r\in S_i} a_r \left\langle \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^{p}\sum_{t=1}^{\tau-1}\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right], \mathbf{x}_i\right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\}$$

Let $\mathcal{E}_{1,k} = \left[\mathcal{E}_{1,k}^{(1)}, \ldots, \mathcal{E}_{1,k}^{(n)}\right]$, and $\mathcal{E}_{2,k} = \left[\mathcal{E}_{2,k}^{(1)}, \ldots, \mathcal{E}_{2,k}^{(n)}\right]$. Then we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{1,k}\right] = \eta\theta\tau\left(\mathbf{H}(k) - \mathbf{H}(k)^{\perp}\right)(\mathbf{y} - \mathbf{u}_k) - \eta\left(\mathcal{E}_{1,k} + \mathcal{E}_{2,k}\right)$$

Thus,

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{1,k}'\right] &= \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{1,k}\right] - \eta\theta\tau\mathbf{H}(k)(\mathbf{y} - \mathbf{u}_k) \\ &= \eta\theta\tau\mathbf{H}(k)^{\perp}(\mathbf{y} - \mathbf{u}_k) + \eta\left(\mathcal{E}_{1,k} + \mathcal{E}_{2,k}\right)\end{aligned}$$

According to Lemma 7 and Lemma 8, we have the bound of $\mathcal{E}_{1,k}^{(i)}$ and $\mathcal{E}_{2,k}^{(i)}$ as

$$\begin{aligned}\left|\mathcal{E}_{1,k}^{(i)}\right| &\leq \theta\xi(1-\xi)\tau n\kappa\sqrt{\frac{2d}{m}} \\ \left|\mathcal{E}_{2,k}^{(i)}\right| &\leq \frac{\eta\xi\tau(\tau-1)n^{\frac{3}{2}}}{2}\left(\theta\|\mathbf{y} - \mathbf{u}_k\|_2 + \sqrt{pC_1}\right)\end{aligned}$$

Moreover, according to Lemma 17, we have

$$\|\mathbf{H}(k)^{\perp}\|_2 \leq 4\xi n\kappa^{-1}R$$

Let $R \leq \frac{\kappa\lambda_0}{128n}$, we have

$$\|\mathbf{H}(k)^{\perp}\|_2 \leq \frac{\lambda_0}{32}$$

Therefore, we have

$$\begin{aligned}\left|\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}'_{1,k}\right]\right\rangle\right| &\leq \eta\theta\tau\left|\left\langle \mathbf{y}-\mathbf{u}_k, \mathbf{H}(k)^{\perp}(\mathbf{y}-\mathbf{u}_k)\right\rangle\right| + \eta\sum_{i=1}^{n}\left(\left|\left(y_i-u_k^{(i)}\right)\mathcal{E}_{1,k}^{(i)}\right| + \left|\left(y_i-u_k^{(i)}\right)\mathcal{E}_{2,k}^{(i)}\right|\right)\\
&= \eta\theta\tau\|\mathbf{H}(k)^{\perp}\|_2\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \max_{i\in[n]}\left(\left|\mathcal{E}_{1,k}^{(i)}\right| + \left|\mathcal{E}_{2,k}^{(i)}\right|\right)\eta\sum_{i=1}^{n}\left|y_i-u_k^{(i)}\right|\\
&\leq \frac{1}{32}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \max_{i\in[n]}\left(\left|\mathcal{E}_{1,k}^{(i)}\right| + \left|\mathcal{E}_{2,k}^{(i)}\right|\right)\eta\sqrt{n}\|\mathbf{y}-\mathbf{u}_k\|_2\\
&\leq \frac{1}{32}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{\eta^2\theta\xi\tau(\tau-1)n^2}{2}\|\mathbf{y}-\mathbf{u}_k\|_2^2 +\\
&\qquad \left(\eta\theta\xi(1-\xi)\tau\kappa\sqrt{\frac{2n^3d}{m}} + \frac{\eta^2\xi\tau(\tau-1)n^2}{2}\sqrt{pC_1}\right)\|\mathbf{y}-\mathbf{u}_k\|_2\end{aligned}$$

Using the general inequality that $ab \leq \frac{1}{2}(a^2+b^2)$, and $\eta \leq \frac{\lambda_0}{16(\tau-1)n^2}$, we get

$$\left|\left\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}'_{1,k}\right]\right\rangle\right| \leq \frac{1}{8}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{16\eta\theta\tau\xi^2(1-\xi)^2\kappa^2n^3d}{m\lambda_0} + \frac{2\eta^3\xi^2\tau(\tau-1)^2n^4pC_1}{\theta\lambda_0}$$

□

Lemma 7. *Under the assumption of Theorem 3 we have that for all $k \in [K], i \in [n]$, it holds that*

$$\left|\mathcal{E}_{1,k}^{(i)}\right| \leq \theta\xi(1-\xi)\tau n\kappa\sqrt{\frac{2d}{m}}$$

Proof. We have

$$\begin{aligned}\left|\mathcal{E}_{1,k}^{(i)}\right| &\leq \frac{\theta\xi(1-\xi)\tau}{m^{\frac{3}{2}}}\left|\sum_{r\in S_i}\sum_{j=1}^{n} a_r\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle\sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_j\rangle)\right|\\
&\leq \frac{\theta\xi(1-\xi)\tau}{m^{\frac{3}{2}}}\sum_{r\in S_i}\sum_{j=1}^{n}|\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle|\\
&\leq \frac{\theta\xi(1-\xi)\tau n}{m^{\frac{3}{2}}}\sum_{r\in S_i}\|\mathbf{w}_{k,r}\|_2\\
&\leq \frac{\theta\xi(1-\xi)\tau n}{m^{\frac{3}{2}}}\sum_{r\in S_i}(\|\mathbf{w}_{0,r}\|_2 + R)\\
&\leq \frac{\theta\xi(1-\xi)\tau n}{m}\|\mathbf{W}_0\|_F + \frac{\theta\xi(1-\xi)\tau nR}{\sqrt{m}}\\
&\leq \theta\xi(1-\xi)\tau n\kappa\sqrt{\frac{2d}{m}}\end{aligned}$$

where for the bound of $\|\mathbf{W}_0\|_F$ we use Lemma 22. □

Lemma 8. *Suppose $\|\mathbf{w}_{k,t,r} - \mathbf{w}_{0,r}\|_2 \leq R$ for all $r \in [m]$. Then we have*

$$\left|\mathcal{E}_{2,t}^{(i)}\right| \leq \frac{\eta\xi\tau(\tau-1)n^{\frac{3}{2}}}{2}\left(\theta\|\mathbf{y}-\mathbf{u}_k\|_2 + \sqrt{pC_1}\right)$$

Proof. Since $r \in S_i$, the difference between the surrogate gradients of a sub-network has the form

$$\begin{aligned}\left\|\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right\|_2 &= \frac{1}{\sqrt{m}}\left\|\sum_{j=1}^n a_r m_{k,r}^l \mathbf{x}_j\left(\hat{u}_{k,t}^{l(j)} - \hat{u}_k^{l(j)}\right)\mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_j\rangle \geq 0\}\right\|_2 \\ &\leq \frac{m_{k,r}^l}{\sqrt{m}}\sum_{j=1}^n\left|\hat{u}_{k,t}^{l(j)} - \hat{u}_k^{l(j)}\right|\end{aligned}$$

Therefore, using the convexity of ℓ_2-norm,

$$\begin{aligned}\left\|\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^p\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right]\right\|_2 &\leq \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\left\|\sum_{l=1}^p\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right\|_2\right] \\ &\leq \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^p\left\|\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right\|_2\right] \\ &\leq \mathbb{E}_{\mathbf{M}_k}\left[\frac{\eta_{k,r}}{\sqrt{m}}\sum_{l=1}^p m_{k,r}\sum_{j=1}^n\left|\hat{u}_{k,t}^{l(j)} - \hat{u}_k^{l(j)}\right|\right]\end{aligned}$$

By Lemma 5, we have

$$\left|\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right| \leq \eta t\sqrt{n}\left(\|\mathbf{y} - \mathbf{u}_k\|_2 + \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right)$$

Therefore,

$$\begin{aligned}\left|\mathcal{E}_{2,t}^{(i)}\right| &\leq \frac{\xi}{\sqrt{m}}\sum_{r\in S_i}\left\|\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^p\sum_{t=1}^{\tau-1}\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right]\right\|_2 \\ &\leq \frac{\xi}{\sqrt{m}}\sum_{t=1}^{\tau-1}\sum_{r\in S_i}\left\|\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^p\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial \mathbf{w}_r}\right)\right]\right\|_2 \\ &\leq \frac{\xi}{\sqrt{m}}\sum_{t=1}^{\tau-1}\sum_{r\in S_i}\mathbb{E}_{\mathbf{M}_k}\left[\frac{\eta_{k,r}}{\sqrt{m}}\sum_{l=1}^p m_{k,r}\sum_{j=1}^n\left|\hat{u}_{k,t}^{l(j)} - \hat{u}_k^{l(j)}\right|\right] \\ &\leq \frac{\eta\xi n^{\frac{3}{2}}}{m}\sum_{t=1}^{\tau-1}t\sum_{r\in S_i}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^p m_{k,r}\left(\|\mathbf{y} - \mathbf{u}_k\|_2 + \|\mathbf{y} - \hat{\mathbf{u}}_k^l\|_2\right)\right] \\ &\leq \frac{\eta\xi\tau(\tau-1)n^{\frac{3}{2}}}{2}\left(\theta\|\mathbf{y} - \mathbf{u}_k\|_2 + \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^p m_{k,r}^l\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right]\right) \\ &\leq \frac{\eta\xi\tau(\tau-1)n^{\frac{3}{2}}}{2}\left(\theta\|\mathbf{y} - \mathbf{u}_k\|_2 + \sqrt{pC_1}\right)\end{aligned}$$

where the last inequality follows from Lemma 24. ☐

Lemma 9. *Under the condition of Theorem 3, we have*

$$|\langle \mathbf{y} - \mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_2\right]\rangle| \leq \frac{1}{8}\eta\theta\tau\lambda_0\|\mathbf{y} - \mathbf{u}_k\|_2^2 + \frac{\eta\lambda_0\xi^2(\theta-\xi^2)n\kappa^2}{24p\tau} + \frac{\eta\lambda_0\xi^2(\tau-1)^2pC_1}{96\tau\theta}$$

Proof. To start, we notice that Using the 1-Lipschitzness of ReLU, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\left|I_{2,k}^{(i)}\right|\right] &= \frac{\xi}{\sqrt{m}}\mathbb{E}_{\mathbf{M}_k}\left[\left|\sum_{r\in S_i^\perp} a_r\left(\sigma(\langle \mathbf{w}_{k+1,r},\mathbf{x}_i\rangle - \sigma(\langle \mathbf{w}_{k+1,r},\mathbf{x}_i\rangle)\right)\right|\right] \\
&\leq \frac{\xi}{\sqrt{m}}\sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[|\sigma(\langle \mathbf{w}_{k+1,r},\mathbf{x}_i\rangle - \sigma(\langle \mathbf{w}_{k+1,r},\mathbf{x}_i\rangle)|\right] \\
&\leq \frac{\xi}{\sqrt{m}}\sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[|\langle \mathbf{w}_{k+1,r}-\mathbf{w}_{k,r},\mathbf{x}_i\rangle|\right] \\
&\leq \frac{\xi}{\sqrt{m}}\sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{w}_{k+1,r}-\mathbf{w}_{k,r}\|_2\right] \\
&\leq \frac{\eta\xi}{\sqrt{m}}\sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\eta_{k,r}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2\right] \\
&\leq \frac{\eta\xi}{\sqrt{m}}\sum_{r\in S_i^\perp}\left(\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2\right] + \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{t=1}^{\tau-1}\sum_{l=1}^{p}\left\|\frac{\partial L_{\mathbf{m}_k}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2\right]\right) \\
&\leq \frac{\eta\theta\xi\sqrt{n}}{m}|S_i^\perp|\|\mathbf{y}-\mathbf{u}_k\|_2 + \frac{4\eta\xi\kappa n|S_i^\perp|}{m}\sqrt{\frac{\theta(\theta-\xi^2)}{p}} + \\
&\qquad \frac{\eta\xi\sqrt{n}}{m}\sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{t=1}^{\tau-1}\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^l\|_2\right] \\
&\leq \frac{\eta\theta\xi\sqrt{n}}{m}|S_i^\perp|\|\mathbf{y}-\mathbf{u}_k\|_2 + \frac{4\eta\xi\kappa n|S_i^\perp|}{m}\sqrt{\frac{\theta(\theta-\xi^2)}{p}} + \\
&\qquad \frac{\eta\xi\sqrt{n}}{m}\sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{t=1}^{\tau-1}\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{y}-\hat{\mathbf{u}}_{k}^l\|_2\right] \\
&\leq \frac{\eta\theta\xi\tau\sqrt{n}}{m}|S_i^\perp|\|\mathbf{y}-\mathbf{u}_k\|_2 + \frac{4\eta\xi\kappa n|S_i^\perp|}{m}\sqrt{\frac{\theta(\theta-\xi^2)}{p}} + \\
&\qquad \frac{\eta\xi\sqrt{n}}{m}\sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{t=1}^{\tau-1}\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{u}_k-\hat{\mathbf{u}}_{k}^l\|_2\right]
\end{aligned}$$

where in the seventh inequality we use the bound on $\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2\right]$ from Lemma 4. Moreover, using Lemma 24 we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2\right] \leq \sqrt{pC_1}$$

Then we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\left|I_{2,k}^{(i)}\right|\right] &\leq \frac{\eta\theta\xi\tau\sqrt{n}}{m}|S_i^\perp|\|\mathbf{y}-\mathbf{u}_k\|_2 + \frac{4\eta\xi\kappa n|S_i^\perp|}{m}\sqrt{\frac{\theta(\theta-\xi^2)}{p}} + \frac{\eta\xi(\tau-1)}{m}\sqrt{npC_1}|S_i^\perp| \\
&\leq 8\eta\theta\xi\tau\sqrt{n}\kappa^{-1}R\|\mathbf{y}-\mathbf{u}_k\|_2 + 16\eta\xi nR\sqrt{\frac{\theta(\theta-\xi^2)}{p}} + 4\eta\xi(\tau-1)\kappa^{-1}R\sqrt{npC_1}
\end{aligned}$$

where in the last inequality we use $|S_i^{\perp}| \leq 4m\kappa^{-1}R$. Therefore,

$$\begin{aligned}
|\langle \mathbf{y}-\mathbf{u}_k, \mathbb{E}_{\mathbf{M}_k}\left[\mathbf{I}_{2,k}\right]\rangle| &= \left|\sum_{i=1}^{n}(y_i - u_k^{(i)})\mathbb{E}_{\mathbf{M}_k}\left[I_{2,k}^{(i)}\right]\right| \\
&\leq \sum_{i=1}^{n}\left|y_i - u_k^{(i)}\right| \cdot \left|\mathbb{E}_{\mathbf{M}_k}\left[I_{2,k}^{(i)}\right]\right| \\
&\leq \max_{i\in[n]}\left|\mathbb{E}_{\mathbf{M}_k}\left[I_{2,k}^{(i)}\right]\right|\sum_{i=1}^{n}\left|y_i - u_k^{(i)}\right| \\
&\leq \sqrt{n}\max_{i\in[n]}\left|\mathbb{E}_{\mathbf{M}_k}\left[I_{2,k}^{(i)}\right]\right|\|\mathbf{y}-\mathbf{u}_k\|_2 \\
&\leq 8\eta\theta\xi\tau\kappa^{-1}nR\|\mathbf{y}-\mathbf{u}_k\|_2^2 + 16\eta\xi R\sqrt{\frac{\theta(\theta-\xi^2)n^3}{p}}\|\mathbf{y}-\mathbf{u}_k\|_2 + \\
&\qquad 4\eta\xi(\tau-1)\kappa^{-1}nR\sqrt{pC_1}\|\mathbf{y}-\mathbf{u}_k\|_2 \\
&\leq \frac{1}{8}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{\eta\lambda_0\xi^2(\theta-\xi^2)n\kappa^2}{24p\tau} + \frac{\eta\lambda_0\xi^2(\tau-1)^2pC_1}{96\tau\theta}
\end{aligned}$$

where in the last inequality we use $R \leq \frac{\kappa\lambda_0}{192n}$ and $ab \leq \frac{1}{2}(a^2+b^2)$. □

Lemma 10. *Under the condition of Theorem 3, with $\eta \leq \frac{\lambda_0}{48n\tau\max\{n,p\}}$, we have*

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right] \leq \frac{1}{4}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{17\eta^2\xi^2\tau^2\theta(\theta-\xi^2)n^3\kappa^2}{p} + \eta^2\xi^2\lambda_0(\tau-1)^2pnC_1$$

Proof. As in previous lemma, we use the Lipschitzness of ReLU to get

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\left(u_{k+1}^{(i)} - u_k^{(i)}\right)^2\right] &\leq \frac{\xi^2}{m}\mathbb{E}_{\mathbf{M}_k}\left[\left(\sum_{r=1}^{m}a_r\left(\sigma(\langle\mathbf{w}_{k+1,r},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)\right)^2\right] \\
&\leq \xi^2\sum_{r=1}^{m}\mathbb{E}_{\mathbf{M}_k}\left[\left(\sigma(\langle\mathbf{w}_{k+1,r},\mathbf{x}_i\rangle) - \sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)^2\right] \\
&\leq \xi^2\sum_{r=1}^{m}\mathbb{E}_{\mathbf{M}_k}\left[\langle\mathbf{w}_{k+1,r}-\mathbf{w}_{k,r},\mathbf{x}_i\rangle^2\right] \\
&\leq \xi^2\sum_{r=1}^{m}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{w}_{k+1,r}-\mathbf{w}_{k,r}\|_2^2\right] \\
&= \xi^2\left(D_{1,k} + D_{2,k}\right)
\end{aligned}$$

where

$$\begin{aligned}
D_{1,k} &= \sum_{r\in S_i}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{w}_{k+1,r}-\mathbf{w}_{k,r}\|_2^2\right] \\
D_{2,k} &= \sum_{r\in S_i^{\perp}}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{w}_{k+1,r}-\mathbf{w}_{k,r}\|_2^2\right]
\end{aligned}$$

Using Lemma 11 and Lemma 12 we have

$$\begin{aligned}
D_{1,k} &\leq \left(4\eta^2\tau^2n\theta + 4\eta^4\theta n^3\tau^3(\tau-1)p\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{16\eta^2\tau^2\theta(\theta-\xi^2)n^2\kappa^2}{p} + 4\eta^4n^3\tau^2(\tau-1)^2pC_1 \\
D_{2,k} &\leq \frac{\eta^2\theta\tau\lambda_0}{18}\left(1+(\tau-1)p\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{4\eta^2\lambda_0\theta(\theta-\xi^2)\tau n\kappa^2}{9p} + \frac{\eta^2\tau(\tau-1)\lambda_0pC_1}{18}
\end{aligned}$$

Therefore we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right] &\leq \xi^2 n\left(D_{1,k}+D_{2,k}\right)\\ &\leq \left(4\eta^2\xi^2\tau^2n^2\theta+4\eta^4\theta\xi^2n^4\tau^3(\tau-1)p+\frac{\eta^2\theta\tau n\lambda_0}{18}\left(1+(\tau-1)p\right)\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2+\\ &\qquad \frac{16\eta^2\xi^2\tau^2\theta(\theta-\xi^2)n^3\kappa^2}{p}+4\eta^4\xi^2n^4\tau^2(\tau-1)^2pC_1+\frac{4\eta^2\lambda_0\xi^2\theta(\theta-\xi^2)\tau n^2\kappa^2}{9p}+\\ &\qquad \frac{\eta^2\xi^2\tau(\tau-1)n\lambda_0 pC_1}{18}\end{aligned}$$

With $\eta \leq \frac{\lambda_0}{48n\tau\max\{n,p\}}$, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2\right] \leq \frac{1}{4}\eta\theta\tau\lambda_0\|\mathbf{y}-\mathbf{u}_k\|_2^2+\frac{17\eta^2\xi^2\tau^2\theta(\theta-\xi^2)n^3\kappa^2}{p}+\eta^2\xi^2\lambda_0(\tau-1)^2pnC_1$$

□

Lemma 11.

$$D_{1,k} \leq \left(4\eta^2\tau^2n\theta+4\eta^4\theta n^3\tau^3(\tau-1)p\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2+\frac{16\eta^2\tau^2\theta(\theta-\xi^2)n^2\kappa^2}{p}+4\eta^4n^3\tau^2(\tau-1)^2pC_1$$

Proof. We have

$$\begin{aligned}D_{1,k} &= \eta^2\sum_{r\in S_i}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\eta_{k,r}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial\mathbf{w}_r}\right\|_2^2\right]\\ &\leq \eta^2\sum_{r\in S_i}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\tau\mathbf{g}_{k,r}+\eta_{k,r}\sum_{t=1}^{\tau-1}\sum_{l=1}^{p}\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial\mathbf{w}_r}-\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r}\right)\right\|_2^2\right]\\ &\leq 2\eta^2\tau^2\sum_{r\in S_i}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2^2\right]+2\eta^2(\tau-1)p\sum_{r\in S_i}\sum_{t=1}^{\tau-1}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p}\left\|\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial\mathbf{w}_r}-\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r}\right\|_2^2\right]\end{aligned}$$

Note that for $r\in S_i$, we have

$$\begin{aligned}\left\|\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial\mathbf{w}_r}-\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r}\right\|_2^2 &= \frac{m_{k,r}}{m}\left\|\sum_{i=1}^{n}a_r\mathbf{x}_i\left(\hat{u}_{k,t}^{l(i)}-\hat{u}_k^{l(i)}\right)\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle\}\geq 0\right\|_2^2\\ &\leq \frac{nm_{k,r}}{m}\sum_{i=1}^{n}\left(\hat{u}_{k,t}^{l(i)}-\hat{u}_k^{l(i)}\right)^2\\ &\leq \frac{n^2m_{k,r}^l}{m}\left(2\eta^2t^2n\|\mathbf{y}-\mathbf{u}_k\|_2^2+2\eta^2t^2n\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2\right)\end{aligned}$$

where in the last inequality we use Lemma 5. Plugging in the bound above and the bound on $\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_k\|_2^2\right]$ from Lemma 4 gives

$$\begin{aligned}D_{1,k} &\leq 4\eta^2\tau^2n\theta\|\mathbf{y}-\mathbf{u}_k\|_2^2+\frac{16\eta^2\tau^2\theta(\theta-\xi^2)n^2\kappa^2}{p}+4\eta^4\theta n^3\tau^3(\tau-1)p\|\mathbf{y}-\mathbf{u}_k\|_2^2+\\ &\qquad 4\eta^4n^3\tau^3(\tau-1)p\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2\right]\\ &\leq \left(4\eta^2\tau^2n\theta+4\eta^4\theta n^3\tau^3(\tau-1)p\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2+\frac{16\eta^2\tau^2\theta(\theta-\xi^2)n^2\kappa^2}{p}+4\eta^4n^3\tau^2(\tau-1)^2pC_1\end{aligned}$$

□

Lemma 12.

$$D_{2,k} \leq \frac{\eta^2\theta\tau\lambda_0}{18}\left(1+(\tau-1)p\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{4\eta^2\lambda_0\theta(\theta-\xi^2)\tau n\kappa^2}{9p} + \frac{\eta^2\tau(\tau-1)\lambda_0 pC_1}{18}$$

Proof.

$$\begin{aligned}
D_{2,k} &= \eta^2 \sum_{r\in S_i^\perp}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\eta_{k,r}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2^2\right] \\
&\leq \eta^2\tau\sum_{r\in S_i^\perp}\left(\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2^2\right] + \sum_{t=1}^{\tau-1}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\eta_{k,r}\sum_{l=1}^{p}\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2^2\right]\right) \\
&\leq \eta^2\tau\sum_{r\in S_i^\perp}\left(\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{g}_{k,r}\|_2^2\right] + p\sum_{t=1}^{\tau-1}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p}\left\|\frac{\partial L_{\mathbf{m}_k}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2^2\right]\right) \\
&\leq \frac{2\eta^2\tau n\theta}{m}|S_i^\perp|\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{8\eta^2\theta(\theta-\xi^2)\tau n^2\kappa^2}{pm}|S_i^\perp| + \frac{\eta^2\tau np}{m}|S_i^\perp|\sum_{t=1}^{\tau-1}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^l\|_2^2\right] \\
&\leq \frac{2\eta^2\tau n\theta}{m}|S_i^\perp|\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{8\eta^2\theta(\theta-\xi^2)\tau n^2\kappa^2}{pm}|S_i^\perp| + \frac{2\eta^2\theta\tau(\tau-1)np}{m}|S_i^\perp|\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \\
&\qquad \frac{2\eta^2\tau(\tau-1)np}{m}|S_i^\perp|\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2\right]
\end{aligned}$$

Using $|S_i^\perp| \leq 4m\kappa^{-1}R$ with $R \leq \frac{\xi\kappa\lambda_0}{144n}$ gives

$$\begin{aligned}
D_{2,k} &\leq \frac{\eta^2\theta\tau\lambda_0}{18}\left(1+(\tau-1)p\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{4\eta^2\lambda_0\theta(\theta-\xi^2)\tau n\kappa^2}{9p} + \\
&\qquad \frac{\eta^2\tau(\tau-1)\lambda_0 p}{18}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p}m_{k,r}^l\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2\right] \\
&\leq \frac{\eta^2\theta\tau\lambda_0}{18}\left(1+(\tau-1)p\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{4\eta^2\lambda_0\theta(\theta-\xi^2)\tau n\kappa^2}{9p} + \frac{\eta^2\tau(\tau-1)\lambda_0 pC_1}{18}
\end{aligned}$$

□

F PROOF OF THEOREM 4

As in previous theorem, we start by studying $\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2$. In the case of a categorical mask, we have the nice property that the average of the sub-networks equals to the full network

$$\mathbf{u}_{k+1} = \frac{1}{p}\sum_{l=1}^{p}\hat{\mathbf{u}}_{k,\tau}^l$$

using this property, Lemma 13 characterize $\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2$ as

$$\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2 = \frac{1}{p}\sum_{l=1}^{p}\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^l\|_2^2 - \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\|\hat{\mathbf{u}}_{k,\tau}^l-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2$$

We start by assuming the condition of Hypothesis 1 holds. We proceed by proving the convergence, then we prove the weight perturbation bound with a fashion of induction. Hypothesis 1 implies that

$$\begin{aligned}
\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^l\|_2^2 &\leq \left(1-\frac{\eta\lambda_0}{2}\right)^\tau\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2 \\
&= \|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2 - \frac{\eta\lambda_0}{2}\sum_{t=0}^{\tau-1}\left(1-\frac{\eta\lambda_0}{2}\right)^t\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2
\end{aligned}$$

Using the fact that $\mathbb{E}_{\mathbf{M}_k}\left[\hat{\mathbf{u}}_k^l\right] = \mathbf{u}_k$, we have that

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k\|_2^2\right] = \|\mathbf{y}-\mathbf{u}_k\|_2^2 + \mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2\right]$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right] = \frac{1}{p}\sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] - \frac{\eta\lambda_0}{2p}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\left(1-\frac{\eta\lambda_0}{2}\right)^t\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] - \\ \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l}\mathbb{E}_{\mathbf{M}_k}\left[\|\hat{\mathbf{u}}_{k,\tau}^l-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2\right] \\ = \|\mathbf{y}-\mathbf{u}_k\|_2^2 - \frac{\eta\lambda_0}{2p}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\left(1-\frac{\eta\lambda_0}{2}\right)^t\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] + \\ \frac{1}{p}\sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2\right] - \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l}\mathbb{E}_{\mathbf{M}_k}\left[\|\hat{\mathbf{u}}_{k,\tau}^l-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2\right]\end{aligned}$$

Lemma 14 studies the error term $\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2$ and gives

$$\sum_{l=1}^{p}\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2^2 = \frac{1}{p}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\|\hat{\mathbf{u}}_k^l-\hat{\mathbf{u}}_k^{l'}\|_2^2$$

Plugging in we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right] \leq \|\mathbf{y}-\mathbf{u}_k\|_2^2 - -\frac{\eta\lambda_0}{2p}\sum_{t=0}^{\tau-1}\left(1-\frac{\eta\lambda_0}{2}\right)^t\sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] + \\ \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l}\mathbb{E}_{\mathbf{M}_k}\left[\|\hat{\mathbf{u}}_k^l-\hat{\mathbf{u}}_k^{l'}\|_2^2 - \|\hat{\mathbf{u}}_{k,\tau}^l-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2\right]\end{aligned}$$

Let

$$\iota_k = \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l}\mathbb{E}_{\mathbf{M}_k}\left[\|\hat{\mathbf{u}}_k^l-\hat{\mathbf{u}}_k^{l'}\|_2^2 - \|\hat{\mathbf{u}}_{k,\tau}^l-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2\right]$$

Lemma 15 shows bound of the expectation of ι_k with respect to the initialization. In particular, under the assumption that the network initialization satisfies $\|\mathbf{W}_{0,r}\|_2 \leq \kappa\sqrt{2md} - \sqrt{m}R$ for some $R \geq 0$, and the weight perturbation is bounded by $\|\mathbf{w}_{k,r}^l - \mathbf{w}_{0,r}\|_2 \leq R$ for all $r \in [m]$ then we have for all $\gamma > 0$

$$\iota_k \leq \frac{\gamma\eta\lambda_0}{2p}\sum_{l=1}^{p}\sum_{t=0}^{\tau-1}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k,t}^l\|_2^2\right] + \iota_k'$$

with

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\iota_k'\right] \leq \frac{32\eta\tau(p-1)^2\kappa^2 n^3 d}{\gamma\lambda_0 m}$$

Using this result, we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right] &= \|\mathbf{y}-\mathbf{u}_k\|_2^2 - -\frac{\eta\lambda_0}{2p}\sum_{t=0}^{\tau-1}\left(1-\frac{\eta\lambda_0}{2}\right)^t\sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] + \iota_k \\
&= \|\mathbf{y}-\mathbf{u}_k\|_2^2 - \frac{\eta\lambda_0}{2p}\sum_{l=1}^{p}\sum_{t=0}^{\tau-1}\left(1-\frac{\eta\lambda_0}{2}\right)^t\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] + \\
&\qquad \frac{\gamma\eta\lambda_0}{2p}\sum_{l=1}^{p}\sum_{t=0}^{\tau-1}\left(1-\frac{\eta\lambda_0}{2}\right)^t\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] + \frac{32\eta\tau(p-1)^2\kappa^2n^3d}{\gamma\lambda_0 m} + \\
&\qquad \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\mathbb{E}_{\mathbf{M}_k}\left[\iota_k^{l,l'}\right] \\
&= \|\mathbf{y}-\mathbf{u}_k\|_2^2 - \frac{(1-\gamma)\eta\lambda_0}{2p}\sum_{l=1}^{p}\sum_{t=0}^{\tau-1}\left(1-\frac{\eta\lambda_0}{2}\right)^t\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_k^l\|_2^2\right] + \\
&\qquad \frac{32\eta\tau(p-1)^2\kappa^2n^3d}{\gamma\lambda_0 m} + \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\mathbb{E}_{\mathbf{M}_k}\left[\iota_k^{l,l'}\right] \\
&\leq \|\mathbf{y}-\mathbf{u}_k\|_2^2 - \frac{(1-\gamma)\eta\lambda_0}{2}\sum_{t=0}^{\tau-1}\left(1-\frac{\eta\lambda_0}{2}\right)^t\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \\
&\qquad \frac{32\eta\tau(p-1)^2\kappa^2n^3d}{\gamma\lambda_0 m} + \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\mathbb{E}_{\mathbf{M}_k}\left[\iota_k^{l,l'}\right] \\
&= \left(\gamma + (1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{32\eta\tau(p-1)^2\kappa^2n^3d}{\gamma\lambda_0 m} + \\
&\qquad \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\mathbb{E}_{\mathbf{M}_k}\left[\iota_k^{l,l'}\right]
\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathbf{M}_k,\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2\right] \leq \left(\gamma + (1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)\|\mathbf{y}-\mathbf{u}_k\|_2^2 + \frac{32\eta\tau(p-1)^2\kappa^2n^3d}{\gamma\lambda_0 m}$$

Also,

$$\begin{aligned}
\mathbb{E}_{[\mathbf{M}_k]}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] \leq \left(\gamma + (1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)^k\|\mathbf{y}-\mathbf{u}_0\|_2^2 + \frac{32\eta\tau(p-1)^2\kappa^2n^3d}{\left(1-\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)\gamma(1-\gamma)\lambda_0 m} + \\
\frac{1}{p^2}\sum_{l=1}^{p}\sum_{k'=0}^{k}\sum_{l'=1}^{l-1}\mathbb{E}_{[\mathbf{M}_{k'}]}\left[\iota_k^{l,l'}\right]
\end{aligned}$$

and therefore,

$$\begin{aligned}
\mathbb{E}_{[\mathbf{M}_k],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] &\leq \left(\gamma + (1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)^k\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\|\mathbf{y}-\mathbf{u}_0\|_2^2 + \frac{32\eta\tau(p-1)^2\kappa^2n^3d}{\left(1-\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)\gamma(1-\gamma)\lambda_0 m} \\
&\leq \left(\gamma + (1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)^k C^2 n + \frac{64\tau(p-1)^2\kappa^2n^3d}{\lambda_0^2\gamma(1-\gamma)m}
\end{aligned}$$

Next we bound the weight perturbation. Similar to the proof of Theorem 3, it suffice to show that

$$\|\mathbf{w}_{k+1,r}-\mathbf{w}_{0,r}\|_2 \leq \eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right] + 2\eta\tau\sqrt{\frac{2nK}{m\delta}}B - 2\eta\tau\sqrt{\frac{2nKB_1}{m\delta\alpha}}$$

where in our circumstance, we have $\xi = \frac{1}{p}$. Again, we recall the definition of B as

$$B = \sqrt{\frac{B_1}{\alpha}} + \kappa\sqrt{\xi(1-\xi)pn}$$

with

$$B_1 = \frac{32\eta\tau(p-1)^2\kappa^2 n^3 d}{\left(1-\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)\gamma(1-\gamma)m\lambda_0}$$

It then suffice to show that

$$\|\mathbf{w}_{k+1,r} - \mathbf{w}_{0,r}\|_2 \le \eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}]h,\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right] + 2\eta\tau\kappa n\sqrt{\frac{2\xi(1-\xi)pK}{m\delta}}$$

By Hypothesis 1 we have

$$\|\mathbf{w}^l_{k,t,r} - \mathbf{w}_{k,r}\|_2 \le \eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right] + 2\eta\tau\kappa n\sqrt{\frac{2\xi(1-\xi)pK}{m\delta}}$$

for all $l \in [p]$ and $t \in [\tau]$. Then we have

$$\begin{aligned}
\|\mathbf{w}_{k+1,r} - \mathbf{w}_{k,r}\|_2 &\le \eta_{k,r}\sum_{l=1}^{p} m^l_{k,r}\|\mathbf{w}^l_{k,\tau,r} - \mathbf{w}_{k,r}\|_2 \\
&\le \eta\tau\sqrt{\frac{2nK}{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right] + 2\eta\tau\kappa n\sqrt{\frac{\xi(1-\xi)pK}{m\delta}}
\end{aligned}$$

In the end we use overparameterization to show for $k = 0$. In particular, we show

$$2\eta\tau\sqrt{\frac{2nK}{m\delta}}\left(\frac{1}{\alpha}\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2\right] + KB\right) \le R = O\left(\frac{\kappa\lambda_0}{n}\right)$$

With

$$\alpha = (1-\gamma)\left(1-\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right) \ge \frac{(1-\gamma)\eta\lambda_0}{2}$$

As before, using Lemma 26, we have

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2\right] \le \left(\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2^2\right]\right)^{\frac{1}{2}} = C\sqrt{n}$$

Using the same technique as in Theorem 3, we have

$$m = \Omega\left(\frac{\tau^2 K\max\{n^4, n^2 d, pK^2\}}{(1-\gamma)^2\kappa^2\lambda_0^4\delta\sqrt{\gamma}}\right)$$

G Lemmas for Theorem 4

Lemma 13. *The kth global step produce the squared error satisfying*

$$\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2 = \frac{1}{p}\sum_{l=1}^{p}\|\mathbf{y}-\hat{\mathbf{u}}^l_{k,\tau}\|_2^2 - \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\|\hat{\mathbf{u}}^l_{k,\tau} - \hat{\mathbf{u}}^{l'}_{k,\tau}\|_2^2$$

Proof. We have

$$
\begin{aligned}
\|\mathbf{y}-\mathbf{u}_{k+1}\|_2^2 &= \left\|\mathbf{y}-\frac{1}{p}\sum_{l=1}^{p}\hat{\mathbf{u}}_{k,\tau}^{l}\right\|_2^2 \\
&= \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{p}\left\langle \mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}, \mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l'}\right\rangle \\
&= \frac{1}{p}\sum_{l=1}^{p}\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}\|_2^2 - \frac{1}{p}\sum_{l=1}^{p}\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}\|_2^2 + \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{p}\left\langle \mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}, \mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l'}\right\rangle \\
&= \frac{1}{p}\sum_{l=1}^{p}\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}\|_2^2 - \\
&\qquad \frac{1}{2p^2}\left(\sum_{l=1}^{p}\sum_{l'=1}^{p}\left(\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}\|_2^2+\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2\right)-\sum_{l=1}^{p}\sum_{l'=1}^{p}\left\langle \mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}, \mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l'}\right\rangle\right) \\
&= \frac{1}{p}\sum_{l=1}^{p}\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}\|_2^2 - \frac{1}{2p^2}\sum_{l=1}^{p}\sum_{l'=1}^{p}\|\hat{\mathbf{u}}_{k,\tau}^{l}-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2 \\
&= \frac{1}{p}\sum_{l=1}^{p}\|\mathbf{y}-\hat{\mathbf{u}}_{k,\tau}^{l}\|_2^2 - \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\|\hat{\mathbf{u}}_{k,\tau}^{l}-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2
\end{aligned}
$$

□

Lemma 14. *We have*

$$
\sum_{l=1}^{p}\|\mathbf{u}_k-\hat{\mathbf{u}}_k^{l}\|_2^2 = \frac{1}{p}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\|\hat{\mathbf{u}}_k^{l}-\hat{\mathbf{u}}_k^{l'}\|_2^2
$$

Proof. Using $\mathbf{u}_k = \frac{1}{p}\sum_{l=1}^{p}\hat{\mathbf{u}}_k^l$ we have

$$
\begin{aligned}
\sum_{l=1}^{p}\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2 &= \sum_{l=1}^{p}\left\|\frac{1}{p}\sum_{l'=1}^{p}\hat{\mathbf{u}}_k^l - \hat{\mathbf{u}}_k^l\right\|_2^2 \\
&= \frac{1}{p^2}\sum_{l=1}^{p}\left\|\sum_{l'=1}^{p}\left(\hat{\mathbf{u}}_k^{l'} - \hat{\mathbf{u}}_k^l\right)\right\|_2^2 \\
&= \frac{1}{p^2}\sum_{l=1}^{p}\sum_{l_1=1}^{p}\sum_{l_2=1}^{p}\left\langle \hat{\mathbf{u}}_k^{l_1} - \hat{\mathbf{u}}_k^l, \hat{\mathbf{u}}_k^{l_2} - \hat{\mathbf{u}}_k^l\right\rangle \\
&= \frac{1}{p}\sum_{l_1=1}^{p}\sum_{l_2=1}^{p}\left\langle \hat{\mathbf{u}}_k^{l_1}, \hat{\mathbf{u}}_k^{l_2}\right\rangle - \frac{1}{p}\sum_{l=1}^{p}\sum_{l_1=1}^{p}\left\langle \hat{\mathbf{u}}_k^{l_1}, \hat{\mathbf{u}}_k^{l}\right\rangle \\
&\quad - \frac{1}{p}\sum_{l=1}^{p}\sum_{l_2=1}^{p}\left\langle \hat{\mathbf{u}}_k^{l}, \hat{\mathbf{u}}_k^{l_2}\right\rangle + \sum_{l=1}^{p}\|\hat{\mathbf{u}}_k^l\|_2^2 \\
&= \sum_{l=1}^{p}\|\hat{\mathbf{u}}_k^l\|_2^2 - \frac{1}{p}\sum_{l=1}^{p}\sum_{l'=1}^{p}\left\langle \hat{\mathbf{u}}_k^{l}, \hat{\mathbf{u}}_k^{l'}\right\rangle \\
&= \frac{1}{2p}\left(\sum_{l=1}^{p}\sum_{l'=1}^{p}\left(\|\hat{\mathbf{u}}_k^l\|_2^2 + \|\hat{\mathbf{u}}_k^{l'}\|_2^2\right) - \sum_{l=1}^{p}\sum_{l'=1}^{p}2\left\langle \hat{\mathbf{u}}_k^{l}, \hat{\mathbf{u}}_k^{l'}\right\rangle\right) \\
&= \frac{1}{2p}\sum_{l=1}^{p}\sum_{l'=1}^{p}\|\hat{\mathbf{u}}_k^l - \hat{\mathbf{u}}_k^{l'}\|_2^2 \\
&= \frac{1}{p}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\|\hat{\mathbf{u}}_k^l - \hat{\mathbf{u}}_k^{l'}\|_2^2
\end{aligned}
$$

□

Lemma 15. *If the network initialization satisfies $\|\mathbf{W}_{0,r}\|_2 \leq \kappa\sqrt{2md} - \sqrt{m}R$ for some $R \geq 0$, and the weight perturbation is bounded by $\|\mathbf{w}_{k,r}^l - \mathbf{w}_{0,r}\|_2 \leq R$ for all $r \in [m]$ then for all $\gamma > 0$ we have*

$$
\iota_k \leq \frac{\eta\gamma\lambda_0}{2p}\sum_{l=1}^{p}\sum_{t=0}^{\tau-1}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y} - \mathbf{u}_{k,t}^l\|_2^2\right] + \iota_k'
$$

with

$$
\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\iota_k'\right] \leq \frac{32\eta\kappa^2\tau(p-1)^2n^3d}{\gamma\lambda_0 m}
$$

Proof. First, we have

$$
\begin{aligned}
\|\mathbf{W}_k^l\|_F &= \|\mathbf{W}_0\|_F + \|\mathbf{W}_k^l - \mathbf{W}_0\|_F \\
&= \kappa\sqrt{2md} - \kappa R\sqrt{m} + \left(\sum_{r=1}^{m}\|\mathbf{w}_{k,r}^l - \mathbf{w}_{0,r}\|_2^2\right)^{\frac{1}{2}} \\
&= \kappa\sqrt{2md} - \kappa R\sqrt{m} + R\sqrt{m} \\
&= \kappa\sqrt{2md}
\end{aligned}
$$

For convenience, denote

$$
\sigma_{k,r}^{l,(i)} = \sigma\left(\left\langle \mathbf{w}_{k,r}^l, \mathbf{x}_i\right\rangle\right); \quad \sigma_{k+1,r}^{l,(i)} = \sigma(\left\langle \mathbf{w}_{k,\tau,r}^l, \mathbf{x}_i\right\rangle)
$$

Note that

$$\begin{aligned}
\|\hat{\mathbf{u}}_k^l - \hat{\mathbf{u}}_k^{l'}\|_2^2 - \|\hat{\mathbf{u}}_{k,\tau}^l - \hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2 &= \sum_{i=1}^n \left(\left(u_k^{l,(i)} - \hat{u}_k^{l',(i)}\right)^2 - \left(\hat{u}_{k,\tau}^{l,(i)} - \hat{u}_{k,\tau}^{l',(i)}\right)^2\right) \\
&= \sum_{i=1}^n \left(\hat{u}_k^{l,(i)} - \hat{u}_k^{l',(i)} + \hat{u}_{k,\tau}^{l,(i)} - \hat{u}_{k,\tau}^{l',(i)}\right)\left(\hat{u}_k^{l,(i)} - \hat{u}_k^{l',(i)} - \hat{u}_{k,\tau}^{l,(i)} + \hat{u}_{k,\tau}^{l',(i)}\right) \\
&= \frac{p}{m}\sum_{i=1}^n \left(\sum_{r=1}^m a_r \left(m_{k,r}^l \left(\sigma_{k,r}^{l,(i)} - \sigma_{k+1,r}^{l,(i)}\right) - m_{k,r}^{l'}\left(\sigma_{k,r}^{l',(i)} - \sigma_{k,r}^{l',(i)}\right)\right)\right) \cdot \\
&\qquad \left(\sum_{r=1}^m a_r \left(m_{k,r}^l \left(\sigma_{k,r}^{l,(i)} + \sigma_{k+1,r}^{l,(i)}\right) - m_{k,r}^{l'}\left(\sigma_{k,r}^{l',(i)} + \sigma_{k,r}^{l',(i)}\right)\right)\right) \\
&= \frac{p}{m}\sum_{i=1}^n \sum_{r=1}^m \left(m_{k,r}^l \left(\sigma_{k,r}^{l,(i)} - \sigma_{k+1,r}^{l,(i)}\right) - m_{k,r}^{l'}\left(\sigma_{k,r}^{l',(i)} - \sigma_{k,r}^{l',(i)}\right)\right) \cdot \\
&\qquad \left(m_{k,r}^l \left(\sigma_{k,r}^{l,(i)} + \sigma_{k+1,r}^{l,(i)}\right) - m_{k,r}^{l'}\left(\sigma_{k,r}^{l',(i)} + \sigma_{k,r}^{l',(i)}\right)\right) + \iota_k^{l,l'}
\end{aligned}$$

where

$$\begin{aligned}
\iota_k^{l,l'} = \frac{p}{m}\sum_{i=1}^n \sum_{r=1}^m \sum_{r'\neq r} a_r a_{r'} &\left(m_{k,r}^l \left(\sigma_{k,r}^{l,(i)} - \sigma_{k+1,r}^{l,(i)}\right) - m_{k,r}^{l'}\left(\sigma_{k,r}^{l',(i)} - \sigma_{k,r}^{l',(i)}\right)\right) \cdot \\
&\left(m_{k,r'}^l \left(\sigma_{k,r'}^{l,(i)} + \sigma_{k+1,r'}^{l,(i)}\right) - m_{k,r'}^{l'}\left(\sigma_{k,r'}^{l',(i)} + \sigma_{k,r'}^{l',(i)}\right)\right)
\end{aligned}$$

Using independence between a_r and $a_{r'}$ we can see that $\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\iota_k^{l,l'}\right] = 0$. Moreover, since for $l \neq l'$ we have $m_{k,r}^l \neq m_{k,r}^{l'}$, it is obvious that $m_{k,r}^l m_{k,r}^{l'} = 0$ for $l \neq l'$. Therefore,

$$\begin{aligned}
\|\hat{\mathbf{u}}_k^l - \hat{\mathbf{u}}_k^{l'}\|_2^2 - \|\hat{\mathbf{u}}_{k,\tau}^l - \hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2 &= \frac{p}{m}\sum_{i=1}^n \sum_{r=1}^m \left(m_{k,r}^l \left(\sigma_{k,r}^{l,(i)2} - \sigma_{k+1,r}^{l,(i)2}\right) - m_{k,r}^{l'}\left(\sigma_{k,r}^{l',(i)2} - \sigma_{k+1,r}^{l',(i)2}\right)\right) + \iota_k^{l,l'} \\
&\leq \frac{p}{m}\sum_{i=1}^n \sum_{r=1}^m \left(\left|\sigma_{k,r}^{l,(i)2} - \sigma_{k+1,r}^{l,(i)2}\right| + \left|\sigma_{k,r}^{l',(i)2} - \sigma_{k+1,r}^{l',(i)2}\right|\right) + \iota_k^{l,l'}
\end{aligned}$$

For all $l \in [p], r \in [m]$, and all $q > 0$ we have

$$\begin{aligned}
\sum_{r=1}^m \left|\sigma_{k,r}^{l,(i)2} - \sigma_{k+1,r}^{l,(i)2}\right| &\leq \sum_{r=1}^m \left|\sigma_{k,r}^{l,(i)} - \sigma_{k+1,r}^{l,(i)}\right| \cdot \left|\sigma_{k,r}^{l,(i)} + \sigma_{k+1,r}^{l,(i)}\right| \\
&\leq \sum_{r=1}^m \left\|\mathbf{w}_{k,r}^l - \mathbf{w}_{k+1,r}^l\right\|_2 \cdot \left(\left\|\mathbf{w}_{k,r}^l\right\|_2 + \left\|\mathbf{w}_{k+1,r}^l\right\|_2\right)
\end{aligned}$$

where the second inequality uses the 1-Lipschitzness of ReLU and the fact that $\|\mathbf{x}_i\|_2 = 1$. Since we have for all $r \in [m]$, it holds that

$$\left\|\mathbf{w}_{k,r}^l - \mathbf{w}_{k+1,r}^l\right\|_2 \leq \eta \sum_{t=0}^{\tau-1} \left\|\frac{\partial L_{\mathbf{M}_k}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r}\right\|_2 \leq \frac{\eta\sqrt{n}}{\sqrt{m}} \sum_{t=0}^{\tau-1} \|\mathbf{y} - \hat{\mathbf{u}}_{k,t}^l\|_2$$

Then

$$\sum_{r=1}^{m}\left|\sigma_{k,r}^{l,(i)2}-\sigma_{k+1,r}^{l,(i)2}\right| \leq \frac{\sqrt{n}}{\sqrt{m}}\sum_{t=0}^{\tau-1}\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2\left(\sum_{r=1}^{m}\left(\left\|\mathbf{w}_{k,r}^{l}\right\|_2+\left\|\mathbf{w}_{k+1,r}^{l}\right\|_2\right)\right)$$
$$\leq \eta\sqrt{n}\sum_{t=0}^{\tau-1}\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2\left(\|\mathbf{W}_{k,r}^{l}\|_F+\|\mathbf{W}_{k+1,r}^{l}\|_F\right)$$
$$\leq 4\eta\kappa\sqrt{mnd}\sum_{t=0}^{\tau-1}\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2$$
$$\leq \frac{4\eta\kappa\sqrt{mnd}}{q}\sum_{t=0}^{\tau-1}\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2+4\eta\kappa q\tau\sqrt{mnd}$$

Also, in the last inequality we use $2ab \leq \frac{a^2}{q}+qb^2$ for all $q>0$. Plugging in the choice

$$q=\frac{8(p-1)\kappa\sqrt{n^3d}}{\gamma\lambda_0 m}$$

with some $\gamma>0$ gives

$$\sum_{r=1}^{m}\left|\sigma_{k,r}^{l,(i)2}-\sigma_{k+1,r}^{l,(i)2}\right| \leq \frac{\eta\gamma\lambda_0 m}{4n(p-1)}\sum_{t=0}^{\tau-1}\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2+\frac{16\eta\kappa^2\tau(p-1)n^2d}{\gamma\lambda_0}$$

Therefore

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\hat{\mathbf{u}}_k^l-\hat{\mathbf{u}}_k^{l'}\|_2^2-\|\hat{\mathbf{u}}_{k,\tau}^l-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2\right] \leq \frac{\eta\gamma\lambda_0 p}{4(p-1)}\sum_{t=0}^{\tau-1}\left(\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2\right]+\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l'}\|_2^2\right]\right)+$$
$$\frac{16\eta\kappa^2\tau p(p-1)n^3d}{\gamma\lambda_0 m}+\mathbb{E}_{\mathbf{M}_k}\left[\iota_k^{l,l'}\right]$$

Therefore

$$\iota_k=\frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\mathbb{E}_{\mathbf{M}_k}\left[\|\hat{\mathbf{u}}_k^l-\hat{\mathbf{u}}_k^{l'}\|_2^2-\|\hat{\mathbf{u}}_{k,\tau}^l-\hat{\mathbf{u}}_{k,\tau}^{l'}\|_2^2\right]$$
$$=\leq \frac{\eta\gamma\lambda_0}{4p(p-1)}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\left(\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l}\|_2^2\right]+\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\hat{\mathbf{u}}_{k,t}^{l'}\|_2^2\right]\right)+\frac{32\eta\kappa^2\tau(p-1)^2n^3d}{\gamma\lambda_0 m}+$$
$$\frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\mathbb{E}_{\mathbf{M}_k}\left[\iota_k^{l,l'}\right]$$
$$=\frac{\eta\gamma\lambda_0}{2p}\sum_{l=1}^{p}\sum_{t=0}^{\tau-1}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{y}-\mathbf{u}_{k,t}^{l}\|_2^2\right]+\frac{32\eta\kappa^2\tau(p-1)^2n^3d}{\gamma\lambda_0 m}+\frac{1}{p^2}\sum_{l=1}^{p}\sum_{l'=1}^{l-1}\mathbb{E}_{\mathbf{M}_k}\left[\iota_k^{l,l'}\right]$$

□

H AUXILIARY RESULTS

Lemma 16. *With probability at least* $1-ne^{-m\kappa^{-1}R}$ *we have* $|S_i|\leq 4m\kappa^{-1}R$ *for all* $i\in[n]$.

Proof. Note that $\mathbb{I}\{r\in S_i^{\perp}\}=\mathbb{I}\{\mathbb{I}\{A_{ir}\}\neq 0\}=\mathbb{I}\{A_{ir}\}$. Therefore, we have

$$|S_i^{\perp}|=\sum_{r=1}^{m}\mathbb{I}\{r\in S_i^{\perp}\}=\mathbb{I}\{A_{ir}\}.$$

Since $\mathbb{E}_{\mathbf{w}_{0,r}}\left[\mathbb{I}\{A_{ir}\}\right] = \mathbb{P}\left(A_{ir}\right) \leq \frac{2R}{\kappa\sqrt{2\pi}} \leq \kappa^{-1}R$, we also have

$$\mathbb{E}_{\mathbf{w}_{0,r}}\left[\left(\mathbb{I}\{A_{ir}\} - \mathbb{E}_{\mathbf{w}_{0,r}}\left[\mathbb{I}\{A_{ir}\}\right]\right)^2\right] \leq \mathbb{E}_{\mathbf{w}_{0,r}}\left[\mathbb{I}\{A_{ir}\}^2\right] = \frac{2R}{\kappa\sqrt{2\pi}} \leq \kappa^{-1}R$$

Again apply Bernstein inequality over the random variable $\mathbb{I}\{A_{ir}\} - \mathbb{E}_{\mathbf{w}_{0,r}}\left[\mathbb{I}\{A_{ir}\}\right]$ with $t = 3m\kappa^{-1}R$ gives

$$\mathbb{P}\left(|S_i^{\perp}| \leq 4m\kappa^{-1}R\right) = \mathbb{P}\left(\sum_{r=1}^{m}\mathbb{I}\{A_{ir}\} \geq 4m\kappa^{-1}R\right) \leq \exp\left(-m\kappa^{-1}R\right)$$

□

Lemma 17. *Define $\mathbf{H}^{\perp} \in \mathbb{R}^{n\times n}$ such that*

$$\mathbf{H}^{\perp}_{ij} = \frac{\xi}{m}\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle \sum_{r\in S_i^{\perp}} \mathbb{I}\{\langle \mathbf{w}_r, \mathbf{x}_i\rangle \geq 0; \langle \mathbf{w}_r, \mathbf{x}_i\rangle \geq 0\}$$

If $|S_i^{\perp}| \leq 4m\kappa^{-1}R$, then we have

$$\|\mathbf{H}^{\perp}\|_2 \leq 4n\xi\kappa^{-1}R$$

Proof. We note that

$$\|\mathbf{H}^{\perp}\|_2^2 \leq \|\mathbf{H}^{\perp}\|_F^2 = \sum_{i,j=1}^{n}|\mathbf{H}^{\perp}_{ij}|^2$$

For each i, j pair we have

$$|\mathbf{H}^{\perp}_{ij}| \leq \frac{\xi}{m}|S_i^{\perp}| = 4\xi\kappa^{-1}R$$

Thus

$$\|\mathbf{H}^{\perp}\|_2 \leq \left(\|\mathbf{H}^{\perp}\|_F^2\right)^{-\frac{1}{2}} \leq \left(16n^2\kappa^{-2}\Delta^2\right)^{-\frac{1}{2}} = 4n\xi\kappa^{-1}R$$

□

Lemma 18. *For i.i.d Bernoulli masks with parameter ξ, $N_{k,r}^{\perp} \sim \texttt{Bern}(\theta)$ with*

$$\theta = \mathbb{P}(N_{k,r}^{\perp} = 1) = 1 - (1-\xi)^p$$

Proof. We have

$$\begin{aligned}
\mathbb{P}(N_{k,r}^{\perp} = 1) &= 1 - \mathbb{P}(N_{k,r}^{\perp} = 0) \\
&= 1 - \prod_{l=1}^{p}\mathbb{P}(m_{k,r}^{l} = 0) \\
&= 1 - (1-\xi)^p
\end{aligned}$$

□

Lemma 19. *We have*

$$\mathbb{E}_{\mathbf{M}_k}\left[(\nu_{k,r,r} - \xi)^2\right] \leq \theta - \xi^2$$

Proof. To start, we notice that $\nu_{k,r,r} = \eta_{k,r}\sum_{l=1}^{p} m_{k,r}^{l2} = \eta_{k,r}\sum_{l=1}^{p} m_{k,r}^{l} = N_{k,r}^{\perp}$. Therefore $\mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r}\right] = \mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^{\perp}\right] = \theta$. Moreover, since $N_{k,r}^{\perp 2} = N_{k,r}^{\perp}$, we have $\mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r}^2\right] = \theta$. Thus, using $\theta \geq \xi$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[(\nu_{k,r,r} - \xi)^2\right] &= \mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r}^2\right] - 2\xi\mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r}\right] + \xi^2 \\
&= \theta - 2\xi\theta + \xi^2 \\
&\leq \theta - \xi^2
\end{aligned}$$

□

Lemma 20. *For i.i.d Bernoulli masks with parameter ξ, we have*

$$\mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r'} \mid N^{\perp}_{k,r} = 1\right] = \begin{cases} \xi & \text{if } r \neq r' \\ 1 & \text{if } r = r' \end{cases}$$

Proof. If $r = r'$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r'} \mid N^{\perp}_{k,r} = 1\right] &= \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^{p} m^l_{k,r} \mid N^{\perp}_{k,r} = 1\right] \\
&= \mathbb{E}_{\mathbf{M}_k}\left[\frac{X_{k,r}}{N_{k,r}} \mid N^{\perp}_{k,r} = 1\right] \\
&= \mathbb{E}_{\mathbf{M}_k}\left[N^{\perp}_{k,r} \mid N^{\perp}_{k,r} = 1\right] \\
&= 1
\end{aligned}$$

If $r' \neq r$, then we have that $m^l_{k,r'}$ is independent from $m^l_{k,r}$ and $N_{k,r}$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r'} \mid N^{\perp}_{k,r} = 1\right] &= \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^{p} m^l_{k,r} \mid N^{\perp}_{k,r} = 1\right]\mathbb{E}_{\mathbf{M}_k}\left[m^l_{k,r'}\right] \\
&= \xi\mathbb{E}_{\mathbf{M}_k}\left[\frac{X_{k,r}}{N_{k,r}} \mid N^{\perp}_{k,r} = 1\right] \\
&= \xi
\end{aligned}$$

□

Lemma 21. *The variance follows*

$$Var_{\mathbf{M}_k}\left(\nu_{k,r,r'} \mid N^{\perp}_{k,r} = 1\right) = \begin{cases} \frac{\theta - \xi^2}{p} & \text{if } r \neq r' \\ 0 & \text{if } r = r' \end{cases}$$

Proof. For $r \neq r'$, the expectation of $\nu^2_{k,r,r'}$ given $N^{\perp}_{k,r} = 1$ is

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\nu^2_{k,r,r'} \mid N^{\perp}_{k,r} = 1\right] &= \mathbb{E}_{\mathbf{M}_k}\left[\frac{\sum_{l=1}^{p}\sum_{l'=1}^{p} m^l_{k,r} m^{l'}_{k,r} m^l_{k,r'} m^{l'}_{k,r'}}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right] \\
&= \sum_{l=1}^{p}\sum_{l' \neq l}\mathbb{E}_{\mathbf{M}_k}[m^l_{k,r'}]\mathbb{E}_{\mathbf{M}_k}[m^{l'}_{k,r'}]\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^l_{k,r}}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right]\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^{l'}_{k,r}}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right] + \\
&\qquad \sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}[m^l_{k,r'}]\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^l_{k,r}}{X^2_{k,r}} \mid N^{\perp}_{k,r} = 1\right] \\
&= \xi^2\sum_{l=1}^{p}\sum_{l' \neq l}\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^l_{k,r}}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right]\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^{l'}_{k,r}}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right] + \\
&\qquad \xi\sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^l_{k,r}}{X^2_{k,r}} \mid N^{\perp}_{k,r} = 1\right] \\
&= \xi^2 + \xi\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right] - \xi^2\sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^l_{k,r}}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right]^2
\end{aligned}$$

Therefore, the variance of $\nu_{k,r,r'}$ given $N^{\perp}_{k,r} = 1$ has the form

$$\begin{aligned}
\mathrm{Var}\left(\nu_{k,r,r'} \mid gN^{\perp}_{k,r} = 1\right) &= \mathbb{E}_{\mathbf{M}_k}\left[\nu^2_{k,r,r'} \mid N^{\perp}_{k,r} = 1\right] - \mathbb{E}_{\mathbf{M}_k}\left[\nu_{k,r,r'} \mid N^{\perp}_{k,r} = 1\right]^2 \\
&= \xi\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right] - \xi^2\sum_{l=1}^{p}\mathbb{E}_{\mathbf{M}_k}\left[\frac{m^l_{k,r}}{X_{k,r}} \mid N^{\perp}_{k,r} = 1\right]^2
\end{aligned}$$

Let $X(p) = \sum_{l=1}^{p} m_{\cdot}^{l} \sim \mathcal{B}(p, \xi)$, then we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{X_{k,r}} \mid N_{k,r}^{\perp} = 1\right] = \mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{1 + X(p-1)}\right]$$

$$\mathbb{E}_{\mathbf{M}_k}\left[\frac{m_{k,r}^{l}}{X_r} \mid N_{k,r}^{\perp} = 1\right] = \mathbb{P}(m_{k,r}^{l} = 1 \mid N_{k,r}^{\perp} = 1)\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{1 + X(p-1)}\right] = \frac{\xi}{\theta}\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{1 + X(p-1)}\right]$$

Moreover, using reciprocal moments we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{1 + X(p-1)}\right] = \frac{\theta}{p\xi}$$

Therefore

$$\begin{aligned}\mathrm{Var}_{\mathbf{M}_k}\left(\nu_{k,r,r'} \mid N_{k,r}^{\perp} = 1\right) &= \xi\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{1 + X(p-1)}\right] - \frac{\xi^4}{\theta^2}p\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{1 + X(p-1)}\right]^2 \\ &= \frac{\theta - \xi^2}{p}\end{aligned}$$

If $r = r'$, the variance is

$$\mathrm{Var}_{\mathbf{M}_k}\left(\nu_{r,r} \mid g_r = 1\right) = \mathrm{Var}_{\mathbf{M}_k}\left(g_r \mid g_r = 1\right) = 0$$

□

Lemma 22. *Suppose $\kappa \leq 1, R \leq \kappa\sqrt{\frac{d}{32}}$. With probability at least $1 - e^{md/32}$ we have that*

$$\|W_0\|_F \leq \kappa\sqrt{2md} - \sqrt{m}R$$

Proof. For all $r \in [m], d_1 \in [d]$, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[w_{rd_1}^2\right] = \kappa^2$$

Moreover, each $w_{rd_1}^2$ is a $(2\kappa^2, 2\kappa^2)$-sub-exponential random variable

$$\begin{aligned}\mathbb{E}\left[e^{t(w_{rd_1}^2 - \kappa^2)}\right] &= \frac{1}{\kappa\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{t(w_{rd_1}^2 - \kappa^2} e^{-\frac{w_{rd_1}^2}{2\kappa^2}} dw_{rd_1} \\ &= \frac{1}{\kappa\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(\frac{1}{2\kappa^2} - t)w_{rd_1}^2 - t\kappa^2} dw_{rd_1} \\ &= \frac{1}{\kappa\sqrt{2\pi}} \cdot \sqrt{\frac{\pi}{(2\kappa)^{-1} - t}} \cdot e^{-t\kappa^2} \\ &= \frac{e^{-t\kappa^2}}{\sqrt{1 - 2t\kappa^2}} \leq e^{2t^2\kappa^4}\end{aligned}$$

with $t \leq \frac{1}{2\kappa^2}$. Thus, using independence between entries of $\mathbf{W}_0$ gives

$$\mathbb{E}\left[e^{t(\|\mathbf{W}_0\|_F^2 - md\kappa^2)}\right] \leq \prod_{r=1}^{m}\prod_{d_1=1}^{d}\mathbb{E}\left[e^{t(w_{rd_1}^2 - \kappa^2)}\right] \leq e^{2mdt^2\kappa^4}$$

Invoking the tail bound of sub-exponential random variable gives

$$\mathbb{P}\left(\|\mathbf{W}_0\|_F^2 \geq md\kappa^2 + t\right) \leq \begin{cases} e^{-\frac{t^2}{8md\kappa^4}} & \text{if } 0 \leq t \leq 2md\kappa^2 \\ e^{-\frac{t^2}{4\kappa^2}} & \text{if } t > 2md\kappa^2 \end{cases}$$

Let $t = md\kappa^2 - 2m\kappa R\sqrt{2d} + mR^2$. Then

$$\|\mathbf{W}_0\|_F^2 \leq 2md\kappa^2 + mR^2 - 2m\kappa R\sqrt{2d} = (\kappa\sqrt{2md} - \sqrt{m}R)^2$$

with probability at least $1 - e^{-\frac{t^2}{8md\kappa^4}}$. Using $R \leq \kappa\sqrt{\frac{d}{32}}$ we have $t \geq \frac{1}{2}md\kappa^2$. Thus with probability at least $1 - e^{-\frac{md}{32}}$ we have

$$\|\mathbf{W}_0\|_F \leq \kappa\sqrt{2md} - \sqrt{m}R$$

□

Lemma 23. *Assume $\kappa \leq 1$ and $R \leq \frac{\kappa}{\sqrt{2}}$. With probability at least $1 - ne^{-\frac{m}{32}}$ over initialization, it holds for all $i \in [n]$ that*

$$\sum_{r=1}^{m} \langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle^2 \leq 2m\kappa^2 - mR^2$$

and thus

$$\sum_{i=1}^{n} \sum_{r=1}^{m} \langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle^2 \leq 2mn\kappa^2 - mnR^2$$

Proof. To begin, we show that each $\langle \mathbf{w}_0, \mathbf{x}_i \rangle$ are Gaussian with zero mean and variance κ^2. Using independence between entries of $\mathbf{w}_{0,r}$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{-t\langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle}\right] &= \mathbb{E}\left[\prod_{j=1}^{d} e^{-tw_{0,r,j}x_{i,j}}\right] \\
&= \prod_{j=1}^{d} \mathbb{E}\left[e^{-tw_{0,r,j}x_{i,j}}\right] \\
&= \prod_{j=1}^{d} e^{-t^2 x_{i,j}^2 \kappa^2} \\
&= e^{-t^2\kappa^2 \sum_{j=1}^{d} x_{i,j}^2} \\
&= e^{-t^2\kappa^2}
\end{aligned}$$

where the last equality follows from our assumption that $\|\mathbf{x}_i\|_2 = 1$. Next, we treat each $\omega_{r,i} = \langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle^2$ as a random variable. First, we compute the mean of $\omega_{r,i}$

$$\begin{aligned}
\mathbb{E}[\omega_{r,i}] &= \mathbb{E}_{\mathbf{W}_0}\left[\langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle^2\right] \\
&= \mathbb{E}_{\mathbf{W}_0}\left[\left(\sum_{d_1=1}^{d} w_{0,r,d}x_{i,d}\right)^2\right] \\
&= \sum_{d_1=1}^{d} \mathbb{E}_{\mathbf{W}_0}\left[w_{0,r,d}^2\right] x_{i,d}^2 \\
&= \kappa^2 \sum_{d_1=1}^{d} x_{i,d}^2 \\
&= \kappa^2
\end{aligned}$$

Then, we show that each $\omega_{r,i}$ is sub-exponential with parameter $(2\kappa^2, 2\kappa^2)$.

$$\begin{aligned}
\mathbb{E}\left[e^{t(\omega_{r,i}-\kappa^2)}\right] &= \frac{1}{\kappa\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(\omega_{r,i}-\kappa^2)} e^{-\frac{\omega_{r,i}}{2\kappa^2}} d\sqrt{\omega_{r,i}} \\
&= \frac{1}{\kappa\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\frac{1}{2\kappa^2}-t)(\sqrt{\omega_{r,i}})^2 - t\kappa^2} d\sqrt{\omega_{r,i}} \\
&= \frac{1}{\kappa\sqrt{2\pi}} \cdot \sqrt{\frac{\pi}{(2\kappa^2)^{-1} - t}} \cdot e^{-t\kappa^2} \\
&= \frac{e^{-t\kappa^2}}{1 - 2t\kappa^2} \\
&\leq e^{2t\kappa^4}
\end{aligned}$$

for $t \leq \frac{1}{2\kappa^2}$. Since each $\mathbf{w}_{0,r}$ is independent, we have that each $\omega_{r,i}$ is independent for a fixed i. Thus

$$\mathbb{E}\left[e^{t\sum_{r=1}^m(\omega_{r,i}-\kappa^2)}\right] = \prod_{r=1}^m \mathbb{E}\left[e^{t(\omega_{r,i}-\kappa^2)}\right] \leq e^{2mt\kappa^4}$$

Thus we have

$$\mathbb{P}\left(\sum_{r=1}^m \omega_{r,i} \geq m\kappa^2 + t\right) \leq \begin{cases} e^{-\frac{t^2}{8m\kappa^4}} & \text{if } 0 \leq t \leq 2m\kappa^2 \\ e^{-\frac{t^2}{2\kappa^2}} & \text{if } t \geq 2m\kappa^2 \end{cases}$$

We choose $t = m\kappa^2 - mR^2$. Since $R \leq \frac{\kappa}{\sqrt{2}}$, we have that $\frac{m\kappa^2}{2} \leq t \leq m\kappa^2$. Thus

$$\mathbb{P}\left(\sum_{r=1}^m \omega_{r,i} \geq 2m\kappa^2 - mR^2\right) \leq e^{-\frac{m}{8}}$$

Apply a union bound over all $i \in [n]$ gives that with probability at least $1 - ne^{-\frac{m}{32}}$, it holds for all $i \in [n]$ that

$$\sum_{r=1}^m \omega_{r,i} \leq 2m\kappa^2 - mR^2$$

Sum over n gives that

$$\sum_{i=1}^n \sum_{r=1}^m \langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle^2 \leq 2mn\kappa^2 - mnR^2$$

with probability at least $1 - ne^{-\frac{m}{32}}$. □

Lemma 24. *If for some $R > 0$ and all $r \in [m]$ the initialization satisfies*

$$\sum_{r=1}^m \langle \mathbf{w}_{0,r}, \mathbf{x}_i \rangle^2 \leq 2mn\kappa^2 - mnR^2$$

and for all $r \in [m]$, it holds that $\|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 \leq R$. Then we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2 \sum_{l=1}^p m_{k,r}^l \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2\right] \leq C_1$$

and

$$\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r} \sum_{l=1}^p m_{k,r}^l \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right] \leq \sqrt{pC_1}$$

with

$$C_1 = \frac{4\theta^2(1-\xi)n\kappa^2}{p}$$

Proof. Using reciprocal moments, we have

$$\mathbb{E}_{\mathbf{M}_k}[\eta_{k,r}] = \mathbb{P}\left(N_{k,r}^\perp = 1\right)\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r} \mid N_{k,r}^\perp = 1\right] = \frac{\theta^2}{p\xi}$$

To start, we compute that for $r \neq r'$. Using the independence of $m_{k,r}^l$ and $m_{k,r'}^l$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2 \sum_{l=1}^p m_{k,r}^l(\xi - m_{k,r'}^l)^2\right] &= \sum_{l=1}^p \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2 m_{k,r}(\xi - m_{k,r'})^2\right] \\
&= \sum_{l=1}^p \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2 m_{k,r}^l\right]\mathbb{E}_{\mathbf{M}_k}\left[(\xi - m_{k,r'}^l)^2\right] \\
&= \xi(1-\xi)\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2 \sum_{l=1}^p m_{k,r}^l\right] \\
&= \xi(1-\xi)\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\right]
\end{aligned}$$

For $r = r'$, we use the idempotent

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l(\xi - m_{k,r}^l)^2\right] &= (1-\xi)^2\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l\right]\\ &= (1-\xi)^2\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\right]\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l\left(u_k^{(i)} - \hat{u}_k^{l(i)}\right)^2\right] &\leq \frac{1}{m}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l\left(\sum_{r'=1}^{m} a_r(\xi - m_{k,r'}^l)\sigma(\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle)\right)^2\right]\\ &\leq \frac{1}{m}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l\sum_{r'=1}^{m}(\xi - m_{k,r'}^l)^2\sigma(\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle)^2\right]\\ &\leq \frac{1}{m}\sum_{r'=1}^{m}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l(\xi - m_{k,r'}^l)^2\right]\sigma(\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle)^2\\ &\leq \frac{\xi(1-\xi)}{m}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\right]\sum_{r'=1}^{m}\sigma(\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle)^2 +\\ &\qquad \frac{(1-\xi)(1-2\xi)}{m}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\right]\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)^2\\ &\leq \frac{\theta^2(1-\xi)}{mp}\sum_{r'=1}^{m}\langle\mathbf{w}_{k,r'},\mathbf{x}_i\rangle^2 + \frac{(1-\xi)^2\theta^2}{mp\xi}\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle^2\\ &\leq \frac{2\theta^2(1-\xi)\kappa^2}{p} + \frac{2\theta^2(1-\xi)^2\kappa^2}{mp\xi}\\ &\leq \frac{4\theta^2(1-\xi)\kappa^2}{p}\end{aligned}$$

where in the last inequality we use $m \geq \xi^{-1}$. Thus, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2\right] = \sum_{i=1}^{n}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l\left(u_k^{(i)} - \hat{u}_k^{l(i)}\right)^2\right] \leq C_1$$

Also, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^{p} m_{k,r}^l\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right] &\leq \mathbb{E}_{\mathbf{M}_k}\left[\left(\eta_{k,r}\sum_{l=1}^{p} m_{k,r}^l\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right)^2\right]^{\frac{1}{2}}\\ &\leq \sqrt{p}\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}^2\sum_{l=1}^{p} m_{k,r}^l\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2\right]^{\frac{1}{2}}\end{aligned}$$

Plugging in the previous bound gives the desired result. □

Lemma 25. *If for some $R > 0$ and all $r \in [m]$ the initialization satisfies*

$$\sum_{r=1}^{m}\langle\mathbf{w}_{0,r},\mathbf{x}_i\rangle^2 \leq 2mn\kappa^2 - mnR^2$$

and for all $r \in [m]$, it holds that $\|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 \leq R$. Then we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2\right] \leq 4\xi(1-\xi)n\kappa^2$$

Proof. To start, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\left(u_k^{(i)}-\hat{u}_k^{l(i)}\right)^2\right] &= \frac{1}{m}\mathbb{E}_{\mathbf{M}_k}\left[\left(\sum_{r=1}^m a_r(\xi - m_{k,r}^l)\sigma(\langle \mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)^2\right] \\ &\leq \frac{1}{m}\sum_{r=1}^m \mathbb{E}_{\mathbf{M}_k}\left[(\xi - m_{k,r}^l)^2\right]\sigma(\langle \mathbf{w}_{k,r},\mathbf{x}_i\rangle)^2 \\ &\leq \frac{\xi(1-\xi)}{m}\sum_{r=1}^m \langle \mathbf{w}_{k,r},\mathbf{x}_i\rangle^2 \\ &\leq \frac{2\xi(1-\xi)}{m}\sum_{r=1}^m \langle \mathbf{w}_{0,r},\mathbf{x}_i\rangle^2 + 2\xi(1-\xi)R^2 \\ &\leq 4\xi(1-\xi)\kappa^2\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2^2\right] = \sum_{i=1}^n \mathbb{E}_{\mathbf{M}_k}\left[\left(u_k^{(i)}-\hat{u}_k^{l(i)}\right)^2\right] \leq 4\xi(1-\xi)n\kappa^2$$

□

Lemma 26. *Assume that for all $i \in [n]$, y_i satisfies $|y_i| \leq C-1$ for some $C \geq 1$. Then, we have*

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2^2\right] \leq C^2 n$$

Proof. It is easy to see that

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[u_0^{(i)}\right] = 0$$

Note that

$$\begin{aligned}\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\left(u_0^{(i)}\right)^2\right] &= \frac{\xi^2}{m}\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\left(\sum_{r=1}^m a_r\sigma(\langle \mathbf{w}_{0,r},\mathbf{x}_i\rangle\right)^2\right] \\ &= \frac{\xi^2}{m}\sum_{r=1}^m \mathbb{E}_{\mathbf{W}_0}\left[\langle \mathbf{w}_{0,r},\mathbf{x}_i\rangle^2\right] \\ &= \frac{\xi^2}{m}\sum_{r=1}^m \mathbb{E}_{\mathbf{W}_0}\left[\left(\sum_{d'=1}^d w_{0,r,d'}x_{i,d'}\right)^2\right] \\ &= \frac{\xi^2}{m}\sum_{r=1}^m\sum_{d'=1}^d \mathbb{E}_{\mathbf{W}_0}\left[w_{0,r,d'}^2 x_{i,d'}^2\right] \\ &= \frac{\xi^2}{m}\sum_{r=1}^m\sum_{d'=1}^d x_{i,d'}^2 \\ &= \xi^2\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\left(y_i - u_0^{(i)}\right)^2\right] = y_i^2 - 2y_i\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[u_0^{(i)}\right] + \mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\left(u_0^{(i)}\right)^2\right] = y_i^2 + \xi^2$$

Thus,

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_0\|_2^2\right] = \sum_{i=1}^n \mathbb{E}_{\mathbf{W}_0,\mathbf{a}}\left[\left(y_i - u_0^{(i)}\right)^2\right] = \sum_{i=1}^n y_i^2 + \xi^2 n \leq C^2 n$$

□

Table 1: Notations

SYMBOL	DESCRIPTION	MATHEMATICAL DEFINITION
K	Number of global iterations	$K \in \mathbb{N}_+$
k	Index of global iterations	$k \in [K]$
τ	Number of local iterations	$\tau \in \mathbb{N}_+$
t	Index of local iterations	$t \in [\tau]$
p	Number of subnetworks	$p \in \mathbb{N}_+$
l	Index of subnetworks	$l \in [p]$
ξ	Probability of selecting a neuron	$\xi \in (0, 1]$
$\boldsymbol{\xi}$	Vector probability of selection a neuron by each worker	$\boldsymbol{\xi} \in (0, 1]^p$
η	Constant step size for local gradient update	$\eta \in \mathbb{R}$
$\mathbf{M}_k$	Binary mask in iteration k	$\mathbf{M}_k \in \{0, 1\}^{p \times m}$
$\mathbf{m}_{k,r}$	Binary mask for neuron r in iteration k	$\mathbf{m}_{k,r} \in \{0, 1\}^p$, the vector of rth column of $\mathbf{M}_k$
$\mathbf{m}_k^l$	Binary mask for subnetwork l in iteration k	$\mathbf{m}_k^l \in \{0, 1\}^m$, the vector of lth row of $\mathbf{M}_k$
$m_{k,r}^l$	Binary mask for neuron r in subnetwork l in iteration k	$m_{k,r}^l \in \{0, 1\}$ the (l, r)th entry of $\mathbf{M}_k$
$X_{k,r}$	Number of subnetworks selecting neuron r in iteration k	$X_{k,r} = \sum_{l=1}^p m_{k,r}^l$
$N_{k,r}$	Aggregated gradient normalizer for neuron r in iteration k	$N_{k,r} = \max\{X_{k,r}, 1\}$
$N_{k,r}^{\perp}$	Indicator of gradient existing for neuron r in iteration k	$N_{k,r}^{\perp} = \min\{X_{k,r}, 1\}$
$\eta_{k,r}$	Global gradient aggregation step size for neuron r in iteration k	$\eta_{k,r} = N_{k,r}^{\perp}/N_{k,r}$
$u_k^{(i)}$	Output of the whole network at global iteration k for sample i	$u_k^{(i)} = \frac{\xi}{\sqrt{m}} \sum_{r=1}^m a_r \sigma(\langle \mathbf{w}_{k,r}, \mathbf{x}_i \rangle$
$\mathbf{u}_k$	Output of the whole network at global iteration k for all $\mathbf{X}$	$\mathbf{u}_k = \left[u_k^{(1)}, \ldots, u_k^{(n)}\right]$
$\hat{u}_{k,t}^{l(i)}$	Output of subnetwork l at iteration (k, t) for sample i	$\hat{u}_{k,t}^{l(i)} = \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r m_{k,r}^l \sigma(\left\langle \mathbf{w}_{k,t,r}^l, \mathbf{x}_i \right\rangle$
$\hat{\mathbf{u}}_{k,t}^l$	Output of subnetwork l at iteration (k, t) for all $\mathbf{X}$	$\hat{\mathbf{u}}_{k,t}^l = \left[\hat{u}_{k,t}^{l(1)}, \ldots, \hat{u}_{k,t}^{l(n)}\right]$
$\hat{u}_k^{l(i)}$	Output of subnetwork l at iteration $(k, 0)$ for sample i	$\hat{u}_k^{l(i)} = \hat{u}_{k,0}^{l(i)}$
$\hat{\mathbf{u}}_k^l$	Output of subnetwork l at iteration $(k, 0)$ for all $\mathbf{X}$	$\hat{\mathbf{u}}_k^l = \hat{\mathbf{u}}_{k,0}^l$
L_k	Global loss at iteration k	$L_k = \Vert\mathbf{y} - \mathbf{u}_k\Vert_2^2$
$L_{\mathbf{m}_k^l}(\mathbf{W}_{k,t}^l)$	Local loss for subnetwork l at iteration (k, t)	$L_{\mathbf{m}_k^l}(\mathbf{W}_{k,t}^l) = \Vert\mathbf{y} - f_{\mathbf{m}_k^l}(\mathbf{W}_{k,t}^l)\Vert_2^2$

I GENERALIZATION ERROR

In this section, we provide bound on the generalization error for the scenario in **Theorem** 3.

Theorem 5. *Suppose the assumption of the dataset in **Theorem** 3 holds, and suppose $p \leq n$. Fix some failure probability δ, total number of global iterations $K = \Omega\left(\log \frac{n}{\delta}\right)$, and use the initialization scale $\kappa = O(\frac{\sqrt{\delta}}{n})$ and step size $\eta = O\left(\frac{\lambda_0\sqrt{\delta}}{\tau\sqrt{n^3\max\{n,K^2\}}}\right)$. If the number of hidden neurons satisfies*

$$m = \Omega\left(\frac{K}{\delta}\max\left\{\frac{n^4}{\kappa^2\lambda_0^4}, \frac{nK^2}{\kappa^2\lambda_0^2}\right\}\cdot poly(\theta,\xi)\right), \tag{9}$$

Then with probability at least $1-\delta$, we have that for any 1-Lipschitze loss function ℓ such that $\ell(y_i,y_i)=0$, it holds that

$$L_{\mathcal{D}}(f(\mathbf{W}_K,\cdot)) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(f(\mathbf{W}_k,\mathbf{x}),y)\right] \leq O\left(\delta^{-1}\sqrt{\frac{\mathbf{y}(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + \sqrt{\frac{\log\frac{n}{\lambda_0\delta}}{n\delta}}\right)$$

I.1 GENERAL STRUCTURE

For the simplicity of the proof, we treat ξ as constant and use $O(\cdot)$ analysis. For a third order tensor $\mathbf{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$, we denote its mode-$i$ matricization as $\mathtt{mat}_i(\mathbf{A}) \in \mathbb{R}^{n_{j_1}n_{j_2}\times n_i}$ with $i, j_1, j_2 \in \{1,2,3\}$ being different elements. For a matrix $\mathbf{A}' \in \mathbb{R}^{n_1\times n_2}$, we denote its vectorization as $\mathtt{vec}(\mathbf{A}') \in \mathbb{R}^{n_1\times n_2}$. To start, we fix $K = \Omega\left(\log\frac{n}{\delta}\right)$. In this way, based on our learning rate, we have that$\|\mathbf{u}_K - \mathbf{y}\|_2 = O\left(\frac{1}{\sqrt{\delta}}\right)$. Thus, for a 1-Lipschitz loss function $\ell(\cdot)$, we have

$$L_{\mathcal{S}}(f(\mathbf{W}_k,\cdot)) = \frac{1}{n}\sum_{i=1}^n\left(\ell(u_K^{(i)},y_i) - \ell(y_i,y_i)\right) \leq \frac{1}{\sqrt{n}}\|\mathbf{u}_K - \mathbf{y}\|_2 = O\left(\frac{1}{\sqrt{n\delta}}\right)$$

We define the partial derivative tensor $\mathbf{Z}(k)$ and the masked partial derivative tensor $\mathbf{m}_k^l \circ \mathbf{Z}(k,t) \in \mathbb{R}^{m\times d\times n}$ as

$$\mathbf{Z}(k)_{r,:,i} = \frac{1}{\sqrt{m}}a_r\mathbf{x}_i\mathbb{I}\left\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle \geq 0\right\}$$
$$\left(\mathbf{m}_k^l\circ\mathbf{Z}(k,t)\right)_{r,:,i} = \frac{1}{\sqrt{m}}a_r m_{k,r}^l\mathbf{x}_i\mathbb{I}\left\{\langle\mathbf{w}_{k,t,r}^l,\mathbf{x}_i\rangle \geq 0\right\}$$

Therefore, we have that

$$\mathbf{W}_{k,t+1}^l - \mathbf{W}_{k,t}^l = -\eta\left(\mathbf{m}_k^l\circ(\mathbf{Z}(k,t)\right)(\hat{\mathbf{u}}_{k,t}^l - \mathbf{y})$$

Let

$$\boldsymbol{\eta}_k = \begin{bmatrix}\eta_{k,1} & & \\ & \ddots & \\ & & \eta_{k,m}\end{bmatrix} \in \mathbb{R}^{m\times m}$$

then we have

$$\begin{aligned}
\mathbf{W}_{k+1} - \mathbf{W}_k &= \sum_{t=0}^{\tau-1}\sum_{l=1}^{p} \boldsymbol{\eta}_k \left(\mathbf{W}_{k,t+1}^l - \mathbf{W}_{k,t}^l\right) \\
&= -\eta \sum_{t=0}^{\tau-1}\sum_{l=1}^{p} \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ (\mathbf{Z}(k,t)\right) (\hat{\mathbf{u}}_{k,t}^l - \mathbf{y}) \\
&= -\eta \sum_{t=0}^{\tau-1}\left(\sum_{l=1}^{p} \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ (\mathbf{Z}(k,t)\right)\right)(\mathbf{u}_k - \mathbf{y}) - \eta \sum_{t=0}^{\tau-1}\sum_{l=1}^{p} \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ (\mathbf{Z}(k,t)\right) (\hat{\mathbf{u}}_{k,t}^l - \mathbf{u}_k) \\
&= -\eta\theta\tau\mathbf{Z}(0)(\mathbf{u}_k - \mathbf{y}) - \eta \sum_{t=0}^{\tau-1}\left(\sum_{l=1}^{p} \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ (\mathbf{Z}(k,t)\right) - \theta\mathbf{Z}(0)\right)(\mathbf{u}_k - \mathbf{y}) - \\
&\qquad \eta \sum_{t=0}^{\tau-1}\sum_{l=1}^{p} \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ (\mathbf{Z}(k,t)\right) (\hat{\mathbf{u}}_{k,t}^l - \mathbf{u}_k)
\end{aligned}$$

We denote

$$\Delta\mathbf{g}_{k,r} = \eta_{k,r} \sum_{t=0}^{\tau-1}\sum_{l=1}^{p} m_{k,r}^l \left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} - \frac{\theta}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r}\right)$$

With this definition, the update of each weight vector can be written as

$$\mathbf{w}_{k+1,r} - \mathbf{w}_{k,r} = -\eta\eta_{k,r} \sum_{t=0}^{\tau-1}\sum_{l=1}^{p} \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial \mathbf{w}_r} = -\frac{\eta\theta\tau}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r} - \eta\Delta\mathbf{g}_{k,r}$$

Then we have

$$\begin{aligned}
u_{k+1}^{(i)} - u_k^{(i)} &= I_{1,k}^{(i)} + I_{2,k}^{(i)} \\
&= \frac{\xi}{\sqrt{m}} \sum_{r\in S_i} a_r \left\langle \mathbf{w}_{k+1,r} - \mathbf{w}_{k,r}, \mathbf{x}_i \right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} + I_{2,k}^{(i)} \\
&= -\frac{\eta\xi}{\sqrt{m}} \sum_{r\in S_i} a_r \left\langle \frac{\theta\tau}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial \mathbf{w}_r} + \Delta\mathbf{g}_{k,r}, \mathbf{x}_i \right\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} + I_{2,k}^{(i)} \\
&= -\frac{\eta\theta\xi\tau}{m} \sum_{r\in S_i}\sum_{j=1}^{n} (u_k^{(i)} - y_i) \langle \mathbf{x}_i, \mathbf{x}_j\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0, \langle \mathbf{w}_{k,r}, \mathbf{x}_j\rangle \geq 0\} \\
&\qquad - \frac{\eta\xi}{\sqrt{m}} \sum_{r\in S_i} a_r \langle \Delta\mathbf{g}_{k,r}, \mathbf{x}_i\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} + I_{2,k}^{(i)} \\
&= -\eta\theta\tau \sum_{j=1}^{n} \left(\mathbf{H}(k)_{ij} - \mathbf{H}(k)_{ij}^{\perp}\right)(u_k^{(i)} - y_i) \\
&\qquad - \frac{\eta\xi}{\sqrt{m}} \sum_{r\in S_i} a_r \langle \Delta\mathbf{g}_{k,r}, \mathbf{x}_i\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\} + I_{2,k}^{(i)}
\end{aligned}$$

Letting $\boldsymbol{\epsilon}_k \in \mathbb{R}^n$ be defined as

$$\boldsymbol{\epsilon}_{k,i} = -\frac{\eta\xi}{\sqrt{m}} \sum_{r\in S_i} a_r \langle \Delta\mathbf{g}_{k,r}, \mathbf{x}_i\rangle \mathbb{I}\{\langle \mathbf{w}_{k,r}, \mathbf{x}_i\rangle \geq 0\}$$

Then we have

$$\begin{aligned}
\mathbf{u}_{k+1} - \mathbf{u}_k &= -\eta\theta\tau\mathbf{H}(k)(\mathbf{u}_k - \mathbf{y}) + \eta\theta\tau\mathbf{H}(k)^{\perp}(\mathbf{u}_k - \mathbf{y}) - \boldsymbol{\epsilon}_k + \mathbf{I}_{2,k} \\
&= -\eta\theta\tau\mathbf{H}^{\infty}(\mathbf{u}_k - \mathbf{y}) - \eta\theta\tau(\mathbf{H}(k) - \mathbf{H}^{\infty})(\mathbf{u}_k - \mathbf{y}) + \eta\theta\tau\mathbf{H}(k)^{\perp}(\mathbf{u}_k - \mathbf{y}) - \boldsymbol{\epsilon}_k + \mathbf{I}_{2,k}
\end{aligned}$$

and thus

$$\mathbf{u}_{k+1} - \mathbf{y} = \left(\mathbf{I} - \eta\theta\tau\mathbf{H}^{\infty}\right)(\mathbf{u}_k - \mathbf{y}) - \eta\theta\tau(\mathbf{H}(k) - \mathbf{H}^{\infty})(\mathbf{u}_k - \mathbf{y}) + \eta\theta\tau\mathbf{H}(k)^{\perp}(\mathbf{u}_k - \mathbf{y}) - \boldsymbol{\epsilon}_k + \mathbf{I}_{2,k}$$
$$= \left(\mathbf{I} - \eta\theta\tau\mathbf{H}(0)\right)(\mathbf{u}_k - \mathbf{y}) + \boldsymbol{\epsilon}'_k$$

by denoting

$$\boldsymbol{\epsilon}'_k = -\eta\theta\tau(\mathbf{H}(k) - \mathbf{H}(0))(\mathbf{u}_k - \mathbf{y}) + \eta\theta\tau\mathbf{H}(k)^{\perp}(\mathbf{u}_k - \mathbf{y}) - \boldsymbol{\epsilon}_k + \mathbf{I}_{2,k}$$

Then

$$\mathbf{u}_k - \mathbf{y} = (I - \eta\theta\tau\mathbf{H}^{\infty})^k(\mathbf{u}_0 - \mathbf{y}) + \sum_{k'=0}^{k-1}(I - \eta\theta\tau\mathbf{H}^{\infty})^{k-k'+1}\boldsymbol{\epsilon}'_{k'}$$

Thus we have

$$\mathbf{W}_{k+1} - \mathbf{W}_k = \eta\theta\tau\mathbf{Z}(0)(I - \eta\theta\tau\mathbf{H}^{\infty})^k\mathbf{y} - \eta\theta\tau\mathbf{Z}(0)(I - \eta\theta\tau\mathbf{H}^{\infty})^k\mathbf{u}_0 -$$
$$\eta\theta\tau\mathbf{Z}(0)\sum_{k'=0}^{k-1}(I - \eta\theta\tau\mathbf{H}^{\infty})^{k-k'+1}\boldsymbol{\epsilon}'_{k'} -$$
$$\eta\sum_{t=0}^{\tau-1}\left(\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l \circ (\mathbf{Z}(k,t)\right) - \theta\mathbf{Z}(0)\right)(\mathbf{u}_k - \mathbf{y}) -$$
$$\eta\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l \circ (\mathbf{Z}(k,t)\right)(\hat{\mathbf{u}}_{k,t}^l - \mathbf{u}_k)$$

Therefore, the weight matrix difference can be bounded as

$$\|\mathbf{W}_k - \mathbf{W}_0\|_F \le Q_1 + Q_2 + Q_3 + Q_4 + Q_5$$

with

$$Q_1 = \left\|\eta\theta\tau\sum_{k'=0}^{k-1}\mathbf{Z}(0)(I - \eta\theta\tau\mathbf{H}^{\infty})^{k'}\mathbf{y}\right\|_F$$
$$Q_2 = \left\|\eta\theta\tau\sum_{k'=0}^{k-1}\mathbf{Z}(0)(I - \eta\theta\tau\mathbf{H}^{\infty})^{k'}\mathbf{u}_0\right\|_F$$
$$Q_3 = \left\|\eta\theta\tau\sum_{k_1=0}^{k-1}\sum_{k_2=0}^{k_1-1}\mathbf{Z}(0)(I - \eta\theta\tau\mathbf{H}^{\infty})^{k_1-k_2+1}\boldsymbol{\epsilon}'_{k_2}\right\|_F$$
$$Q_4 = \left\|\eta\sum_{k'=0}^{k-1}\sum_{t=0}^{\tau-1}\left(\sum_{l=1}^{p}\boldsymbol{\eta}_{k'}\left(\mathbf{m}_{k'}^l \circ (\mathbf{Z}(k',t)\right) - \theta\mathbf{Z}(0)\right)(\mathbf{u}_{k'} - \mathbf{y})\right\|_F$$
$$Q_5 = \left\|\eta\sum_{k'=0}^{k-1}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\boldsymbol{\eta}_{k'}\left(\mathbf{m}_{k'}^l \circ (\mathbf{Z}(k',t)\right)(\hat{\mathbf{u}}_{k',t}^l - \mathbf{u}_{k'})\right\|_F$$

Based on next section, we have

$$\mathbb{E}_{[\mathbf{M}_k],\mathbf{a}}\left[Q_1\right] = O\left(\sqrt{y^{\top}(\mathbf{H}^{\infty})^{-1}\mathbf{y}} + \frac{n\left(\log(n/\delta)\right)^{\frac{1}{4}}}{m^{\frac{1}{4}}\lambda_0}\right)$$
$$\mathbb{E}_{[\mathbf{M}_k],\mathbf{a}}\left[Q_2\right] = O\left(\frac{n\kappa\lambda_0}{\sqrt{\delta}}\right)$$
$$\mathbb{E}_{[\mathbf{M}_k],\mathbf{a}}\left[Q_3\right] = O\left(\frac{\eta\tau}{\lambda_0}\left(\sqrt{\frac{n^3}{\delta}} + \frac{\kappa n^2 K}{p} + K\sqrt{npC_1}\right)\right)$$
$$\mathbb{E}_{[\mathbf{M}_k],\mathbf{a}}\left[Q_4\right] = O\left(\frac{\eta\tau n K}{\lambda_0\sqrt{\delta}}\right)$$
$$\mathbb{E}_{[\mathbf{M}_k],\mathbf{a}}\left[Q_5\right] = O(\eta\tau\sqrt{n}\kappa)$$

Then, as long as $\kappa = O\left(\frac{\sqrt{\delta}}{n}\right)$ and $\eta = O\left(\frac{\lambda_0\sqrt{\delta}}{\tau\sqrt{n^3\max\{n,K^2\}}}\right)$, we have that

$$\mathbb{E}_{\mathbf{M}_k,\mathbf{a}}\left[\|\mathbf{W}_k - \mathbf{W}_0\|_F\right] = O\left(\sqrt{\mathbf{y}^\top(\mathbf{H}^\infty)^{-1}\mathbf{y}} + 1\right)$$

Thus, with probability at least $1-\delta$ we have

$$\|\mathbf{W}_k - \mathbf{W}_0\|_F = O\left(\delta^{-1}\left(\sqrt{\mathbf{y}^\top(\mathbf{H}^\infty)^{-1}\mathbf{y}} + 1\right)\right)$$

Moreover, the row-wise weight perturbation is bounded as

$$\begin{aligned}
\|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 &\leq \sum_{k'=0}^{k-1}\|\mathbf{w}_{k,\tau} - \mathbf{w}_{k,0}\|_2 \\
&= O\left(\frac{\eta\tau n\sqrt{K}}{\delta\sqrt{m}}\right)\sum_{k'=0}^{k-1}(1-\frac{1}{8}\eta\theta\tau\lambda_0)^{k'} + O\left(\eta\tau\kappa n\sqrt{\frac{pK^3}{m\delta}}\right) \\
&\leq O\left(\frac{n\sqrt{K}}{\lambda_0\delta\sqrt{m}}\right)
\end{aligned}$$

Consider the functional class

$$\mathcal{F} = \{f(\mathbf{W},\cdot) \mid \|\mathbf{w}_{k,r} - \mathbf{w}_{0,r}\|_2 \leq R; \|\mathbf{W}_k - \mathbf{W}_0\|_F \leq \mathcal{R}\}$$

with

$$\begin{aligned}
R &= O\left(\frac{n\sqrt{K}}{\lambda_0\delta\sqrt{m}}\right) \\
\mathcal{R} &= O\left(\delta^{-1}\left(\sqrt{\mathbf{y}^\top(\mathbf{H}^\infty)^{-1}\mathbf{y}} + 1\right)\right)
\end{aligned}$$

Then the Rademacher complexity of this functional class is

$$\begin{aligned}
\texttt{Rad}(\mathcal{F}) &\leq \frac{\mathcal{R}}{\sqrt{2n}}\left(1 + \left(\frac{2\log\frac{2}{\delta}}{m}\right)^{\frac{1}{4}}\right) + \frac{2R^2\sqrt{m}}{\kappa} + R\sqrt{2\log\frac{2}{\delta}} \\
&\leq O\left(\delta^{-1}\left(\sqrt{\frac{\mathbf{y}(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + \frac{1}{\sqrt{n}}\right)\right)
\end{aligned}$$

Thus, with probability at least $1-\delta$ we have that

$$\sup_{f\in\mathcal{F}}\left(L_{\mathcal{D}}(f) - L_{\mathcal{S}}(f)\right) \leq 2\texttt{Rad}(\mathcal{F}) + \sqrt{\frac{\log\frac{n}{\lambda_0\delta}}{n}} = O\left(\delta^{-1}\sqrt{\frac{\mathbf{y}(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + \sqrt{\frac{\log\frac{n}{\lambda_0\delta}}{n}}\right)$$

which implies that

$$L_{\mathcal{D}}(f) \leq L_{\mathcal{S}}(f) + O\left(\delta^{-1}\sqrt{\frac{\mathbf{y}(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + \sqrt{\frac{\log\frac{n}{\lambda_0\delta}}{n}}\right) = O\left(\delta^{-1}\sqrt{\frac{\mathbf{y}(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + \sqrt{\frac{\log\frac{n}{\lambda_0\delta}}{n\delta}}\right)$$

I.2 Bound of Q_1, Q_2, Q_3, Q_4 and Q_5

We bound each of the terms separately. We first note that

$$\texttt{mat}_3(\mathbf{Z}(0)) = \frac{1}{\sqrt{m}}\begin{bmatrix} a_1\mathbf{x}_1\mathbb{I}\{\langle\mathbf{w}_{0,1},\mathbf{x}_1\rangle\} & \dots & a_1\mathbf{x}_n\mathbb{I}\{\langle\mathbf{w}_{0,1},\mathbf{x}_n\rangle\} \\ \dots & & \dots \\ a_m\mathbf{x}_1\mathbb{I}\{\langle\mathbf{w}_{0,m},\mathbf{x}_1\rangle\} & \dots & a_m\mathbf{x}_n\mathbb{I}\{\langle\mathbf{w}_{0,m},\mathbf{x}_n\rangle\}\end{bmatrix} \in \mathbb{R}^{md\times n}$$

and therefore $\texttt{mat}_3(\mathbf{Z}(0))^\top\texttt{mat}_3(\mathbf{Z}(0)) = \frac{1}{\xi}\mathbf{H}(0)$. Let $\mathbf{T} = \eta\theta\tau\sum_{k'=0}^{k-1}(\mathbf{I} - \eta\theta\tau\mathbf{H}^\infty)^{k'}$.

I.2.1 BOUND OF Q_1

For Q_1, we have

$$\begin{aligned}
Q_1^2 &= \left\| \eta\theta\tau \sum_{k'=0}^{k-1} \mathbf{Z}(0)(I - \eta\theta\tau\mathbf{H}^\infty)^{k'}\mathbf{y} \right\|_F^2 \\
&= \left\| \eta\theta\tau \texttt{vec}\left(\sum_{k'=0}^{k-1} \mathbf{Z}(0)(I - \eta\theta\tau\mathbf{H}^\infty)^{k'}\mathbf{y} \right) \right\|_2^2 \\
&= \left\| \eta\theta\tau \sum_{k'=0}^{k-1} \texttt{mat}_3(\mathbf{Z}(0))(I - \eta\theta\tau\mathbf{H}^\infty)^{k'}\mathbf{y} \right\|_2^2 \\
&= \mathbf{y}^\top \mathbf{T}^\top \texttt{mat}_3(\mathbf{Z}(0))^\top \texttt{mat}(\mathbf{Z}(0))\mathbf{T}\mathbf{y} \\
&= \frac{1}{\xi}\mathbf{y}^\top \mathbf{T}^\top \mathbf{H}(0)\mathbf{T}\mathbf{y} \\
&\leq \frac{1}{\xi}\mathbf{y}^\top \mathbf{T}^\top \mathbf{H}^\infty\mathbf{T}\mathbf{y} + \frac{1}{\xi}\left\|\mathbf{H}(0) - \mathbf{H}^\infty\right\|_2 \|\mathbf{T}\|_2^2 \|\mathbf{y}\|_2^2 \\
&\leq \frac{1}{\xi}\mathbf{y}^\top \mathbf{T}^\top \mathbf{H}^\infty\mathbf{T}\mathbf{y} + O\left(\frac{n^2\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2} \right)
\end{aligned}$$

where the last inequality follows from the fact that

$$\|\mathbf{T}\|_2 \leq \eta\theta\tau\sum_{k'=0}^{k-1}(1-\eta\theta\tau\lambda_0)^k = \lambda_0; \quad \|\mathbf{H}(0) - \mathbf{H}^\infty\|_2 \leq \frac{\xi n\sqrt{\log(n/\delta)}}{\sqrt{m}}; \quad \|\mathbf{y}\|_2 = n$$

Also, consider the eigen-decomposition $\mathbf{H}^\infty = \sum_{i=1}^n \lambda_i \mathbf{v}_i\mathbf{v}_i^\top$. Note that $\mathbf{T}$ and $\mathbf{H}^\infty$ has the same eigenvectors. Therefore,

$$\mathbf{T} = \eta\theta\tau\sum_{i=1}^n\sum_{k'=0}^{k-1}(1-\eta\theta\tau\lambda_i)^{k'}\mathbf{v}\mathbf{v}^\top \preceq \sum_{i=1}^n \lambda_i^{-1}\mathbf{v}_i\mathbf{v}_i^\top = (\mathbf{H}^\infty)^{-1}$$

Thus

$$Q_1^2 \leq \xi^{-1}\mathbf{y}^\top(\mathbf{H}^\infty)\mathbf{y} + O\left(\frac{n^2\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2} \right)$$

which implies that

$$Q_1 \leq \xi^{-\frac{1}{2}}\sqrt{y^\top(\mathbf{H}^\infty)^{-1}\mathbf{y}} + O\left(\frac{n\left(\log(n/\delta)\right)^{\frac{1}{4}}}{m^{\frac{1}{4}}\lambda_0} \right)$$

I.2.2 BOUND OF Q_2

For Q_2, we have

$$Q_2 = \|\mathbf{Z}(0)\mathbf{T}\mathbf{u}_0\|_F \leq \|\mathbf{Z}(0)\|_F\|\mathbf{T}\|_2\|\mathbf{u}_0\|_2$$

Note that with probability at least $1-\delta$ we have that $\|\mathbf{u}_0\|_2 \leq \sqrt{\frac{n}{\delta}}\kappa$. Also, we can bound $\|\mathbf{Z}(0)\|_F$ as

$$\begin{aligned}
\|\mathbf{Z}(0)\|_F^2 &= \sum_{i=1}^n\sum_{r=1}^m \|\mathbf{Z}(0)_{r,:,i}\|_2^2 \\
&= \frac{1}{m}\sum_{i=1}^n\sum_{r=1}^m \|a_r\mathbf{x}_i\mathbb{I}\{\langle \mathbf{w}_{0,r}, \mathbf{x}_i\rangle \geq 0\} \\
&\leq n
\end{aligned}$$

Thus we have

$$Q_2 \leq \sqrt{n}\lambda_0 \cdot \sqrt{\frac{n}{\delta}}\kappa = \frac{n\kappa\lambda_0}{\sqrt{\delta}}$$

I.2.3 BOUND OF Q_3

We start with giving a bound on $\|\boldsymbol{\epsilon}'_k\|_2$. We have that

$$\|\boldsymbol{\epsilon}'_k\|_2 \leq \eta\theta\tau \left(\|\mathbf{H}(k) - \mathbf{H}(0)\|_2\|\mathbf{u}_k - \mathbf{y}\|_2 + \|\mathbf{H}(k)^{\perp}\|_2\|\mathbf{u}_k - \mathbf{y}\|_2\right) + \|\mathbf{I}_{2,k}\|_2 + \|\boldsymbol{\epsilon}_k\|_2$$

Among all the terms, $\|\mathbf{I}_{2,k}\|_2$ can be bounded using the bound on $\mathbb{E}_{\mathbf{M}_k}\left[\left|I_{2,k}^{(i)}\right|\right]$

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{I}_{2,k}\|_2\right] &\leq \sqrt{n}\max_{i\in[n]}\mathbb{E}_{\mathbf{M}_k}\left[\left|I_{2,k}^{(i)}\right|\right] \\ &\leq O\left(\eta\tau n\kappa^{-1}R\right)\|\mathbf{u}_k - \mathbf{y}\|_2 + O\left(\frac{\eta n^{\frac{3}{2}}R}{\sqrt{p}} + \eta\tau\kappa^{-1}nR\sqrt{pC_1}\right)\end{aligned}$$

Moreover, since we have that

$$\|\mathbf{H}(k)^{\perp}\|_2 = O\left(n\kappa^{-1}R\right); \quad \|\mathbf{H}(k) - \mathbf{H}(0)\|_2 = O\left(n\kappa^{-1}R\right)$$

Therefore

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\|\boldsymbol{\epsilon}'_k\|_2\right] \leq O\left(\eta\tau n\kappa^{-1}R\right)\|\mathbf{u}_k - \mathbf{y}\|_2 + \mathbb{E}_{\mathbf{M}_k}\left[\|\boldsymbol{\epsilon}_k\|_2\right] + \\ O\left(\frac{\eta n^{\frac{3}{2}}R}{\sqrt{p}} + \eta\tau\kappa^{-1}nR\sqrt{pC_1}\right)\end{aligned}$$

What remains is to bound $\mathbb{E}_{\mathbf{M}_k}\left[\|\boldsymbol{\epsilon}_k\|_2\right]$. To do this, we first bound the norm of $\Delta\mathbf{g}_{k,r}$. We write

$$\begin{aligned}\Delta\mathbf{g}_{k,r} &= \eta_{k,r}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p} m_{k,r}^{l}\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial\mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r}\right) + \\ &\quad \eta_{k,r}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r} - \frac{\tau}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial\mathbf{w}_r} \\ &= \eta_{k,r}\sum_{t=0}^{\tau-1}\sum_{l=1}^{p}\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial\mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r}\right) + \\ &\quad \tau\left(\mathbf{g}_{k,r} - \frac{1}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial\mathbf{w}_r}\right)\end{aligned}$$

The second term has norm

$$\begin{aligned}\mathbb{E}_{\mathbf{M}_k}\left[\left\|\mathbf{g}_{k,r} - \frac{1}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial\mathbf{w}_r}\right\|_2\right] &= \theta\mathbb{E}_{\mathbf{M}_k}\left[\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\left(f_{k,r}^{(i)} - u_k^{(i)}\right)a_r\mathbf{x}_i\mathbb{I}\{\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle \geq 0\} \mid N_{k,r}^{\perp} = 1\right] + \\ &\quad P(N_{k,r}^{\perp} = 0)\left\|\frac{1}{\xi}\frac{\partial L(\mathbf{W}_k)}{\partial\mathbf{w}_r}\right\|_2 \\ &\leq \frac{\theta\sqrt{n}}{\sqrt{m}}\mathbb{E}_{\mathbf{M}_k}\left[\|\mathbf{f}_{k,r} - \mathbf{u}_k\|_2 \mid N_{k,r}^{\perp} = 1\right] + \frac{(1-\theta)\sqrt{n}}{\sqrt{m}}\|\mathbf{u}_k - \mathbf{y}\|_2 \\ &= O\left(\sqrt{\frac{n}{m}}\right)\|\mathbf{u}_k - \mathbf{y}\|_2 + O\left(\frac{n\kappa}{\sqrt{mp}}\right)\end{aligned}$$

For the first term, in previous proof, we have

$$\left\|\mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^{p}\left(\frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_{k,t}^l\right)}{\partial\mathbf{w}_r} - \frac{\partial L_{\mathbf{m}_k^l}\left(\mathbf{W}_k\right)}{\partial\mathbf{w}_r}\right)\right]\right\|_2 \leq \mathbb{E}_{\mathbf{M}_k}\left[\frac{\eta_{k,r}}{\sqrt{m}}\sum_{l=1}^{p}m_{k,r}\sum_{j=1}^{n}\left|\hat{u}_{k,t}^{l(j)} - \hat{u}_k^{l(j)}\right|\right]$$

with

$$\left|\hat{u}_{k,t}^{l(i)} - \hat{u}_k^{l(i)}\right| \leq \eta t\sqrt{n}\left(\|\mathbf{y} - \mathbf{u}_k\|_2 + \|\mathbf{u}_k - \hat{\mathbf{u}}_k^l\|_2\right)$$

Combining the result above, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\Delta\mathbf{g}_{k,r}\|_2\right] \leq \eta\sqrt{\frac{n}{m}}\sum_{t=0}^{\tau-1} t\left(\|\mathbf{u}_k-\mathbf{y}\|_2 + \mathbb{E}_{\mathbf{M}_k}\left[\eta_{k,r}\sum_{l=1}^{p} m_{k,r}^l\|\mathbf{u}_k-\hat{\mathbf{u}}_k^l\|_2\right]\right) + O\left(\tau\sqrt{\frac{n}{m}}\right)\|\mathbf{u}_k-\mathbf{y}\|_2 + O\left(\frac{\tau n\kappa}{\sqrt{mp}}\right)$$
$$\leq O\left(\tau\sqrt{\frac{n}{m}}(\eta\tau+1)\right)\|\mathbf{u}_k-\mathbf{y}\|_2 + O\left(\frac{\tau n\kappa}{\sqrt{mp}} + \frac{\eta\tau^2\sqrt{pC_1 n}}{\sqrt{m}}\right)$$

Now we bound $\mathbb{E}_{\mathbf{M}_k}\left[\|\boldsymbol{\epsilon}_k\|_2\right]$

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\boldsymbol{\epsilon}_k\|_2\right] \leq \frac{\eta\xi}{\sqrt{m}}\sum_{r\in S_i}\mathbb{E}_{\mathbf{M}_k}\left[\|\Delta\mathbf{g}_{k,r}\|_2\right]$$
$$\leq O\left(\eta\tau\sqrt{n}(\eta\tau+1)\right)\|\mathbf{u}_k-\mathbf{y}\|_2 + O\left(\eta\tau\kappa\sqrt{\frac{n^3}{p}} + \eta^2\tau^2 n\sqrt{pC_1}\right)$$

Combining the bound above, with the assumption that $R \leq \frac{\kappa}{n}$ and $\eta \leq \frac{1}{\tau n}$, we have

$$\mathbb{E}_{\mathbf{M}_k}\left[\|\boldsymbol{\epsilon}_k'\|_2\right] = O(\eta\tau\sqrt{n})\|\mathbf{u}_k-\mathbf{y}\|_2 + O\left(\eta\tau\left(\kappa\sqrt{\frac{n^3}{p}} + \sqrt{pC_1}\right)\right)$$

Using the convergence rate of $\mathbb{E}_{[\mathbf{M}_{k-1}]}\left[\|\mathbf{u}_k-\mathbf{y}\|_2\right]$, since $B_1 = O(1)$ we have that

$$\mathbb{E}_{[\mathbf{M}_k]}\left[\|\boldsymbol{\epsilon}_k'\|_2\right] \leq O(\eta\tau\sqrt{n})(1-\frac{1}{8}\eta\theta\tau\lambda_0)^k\|\mathbf{u}_0-\mathbf{y}\|_2 + O(\eta\tau\sqrt{n})(1-\frac{1}{8}\eta\theta\tau\lambda_0)^k + O\left(\eta\tau\left(\kappa\sqrt{\frac{n^3}{p}} + \sqrt{pC_1}\right)\right)$$
$$\leq (1-\frac{1}{8}\eta\theta\tau\lambda_0)^k O\left(\frac{\eta\tau n}{\sqrt{\delta}}\right) + O\left(\eta\tau\left(\kappa\sqrt{\frac{n^3}{p}} + \sqrt{pC_1}\right)\right)$$

As in the bound of Q_2, we have

$$\mathbb{E}_{[\mathbf{M}_k]}\left[Q_3\right] = \mathbb{E}_{[\mathbf{M}_k]}\left[\left\|\eta\theta\tau\sum_{k_1=0}^{k-1}\sum_{k_2=0}^{k_1-1}\mathbf{Z}(0)(I-\eta\theta\tau\mathbf{H}^{\infty})^{k_1-k_2+1}\boldsymbol{\epsilon}_{k_2}'\right\|_F\right]$$
$$\leq \eta\theta\tau\cdot\sqrt{n}\sum_{k_1=0}^{k-1}\sum_{k_2=0}^{k_1-1}(1-\eta\theta\tau\lambda_0)^{k_1-k_2+1}\mathbb{E}_{[\mathbf{M}_{k_2}]}\left[\|\boldsymbol{\epsilon}_{k_2}'\|_2\right]$$
$$\leq O\left(\eta^2\tau^2\sqrt{\frac{n^3}{\delta}}\right)\theta\sum_{k_1=1}^{k}(1-\frac{1}{8}\eta\theta\tau\lambda_0)^{k_1} + O\left(\eta^2\tau^2\left(\frac{\kappa n^2}{p} + \sqrt{npC_1}\right)\right)\theta\sum_{k_1=0}^{k-1}\sum_{k_2=0}^{k_1-1}(1-\eta\theta\tau\lambda_0)^{k_1-k_2+1}$$
$$\leq O\left(\frac{\eta\tau}{\lambda_0}\left(\sqrt{\frac{n^3}{\delta}} + \frac{\kappa n^2 K}{p} + K\sqrt{npC_1}\right)\right)$$

I.2.4 BOUND ON Q_4

First, we bound $\left\|\sum_{l=1}^p \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ \mathbf{Z}(k,t) - \mathbf{m}_k^l \circ \mathbf{Z}(0,0)\right)\right\|_F$. In particular, we have

$$\begin{aligned}\left\|\sum_{l=1}^p \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ \mathbf{Z}(k,t) - \mathbf{m}_k^l \circ \mathbf{Z}(0,0)\right)\right\|_F^2 &= \frac{1}{m}\sum_{r=1}^m\sum_{i=1}^n\sum_{l=1}^p \left\|\eta_{k,r} a_r \mathbf{x}_i m_{k,r}^l s_{r,i}\right\|_2^2 \\ &\leq \frac{1}{m}\sum_{r=1}^m\sum_{i=1}^n \eta_{k,r}^l \sum_{l=1}^p m_{k,r}^l \left|s_{r,i}\right| \\ &= \frac{1}{m}\sum_{r=1}^m\sum_{i=1}^n N_{k,r}^\perp |s_{r,i}|\end{aligned}$$

with

$$s_{r,i} = \mathbb{I}\left\{\left\langle \mathbf{w}_{k,t,r}^l, \mathbf{x}_i\right\rangle\right\} - \mathbb{I}\{\langle \mathbf{w}_{0,r}, \mathbf{x}_i\rangle\}$$

Then $s_{r,i} = 0$ if $\neg A_{ir}$, and $|s_{r,i}| \leq 1$ otherwise. Thus

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{M}_k}\left[N_{k,r}^\perp |s_{r,i}|\right] = \theta P(A_{ir}) \leq \frac{2\theta R}{\kappa\sqrt{2\pi}} \leq \theta\kappa^{-1}R$$

Also, its variance follows

$$Var_{\mathbf{W}_0,\mathbf{M}_k}\left(s_{r,i}\right) \leq \mathbb{E}_{\mathbf{W}_0,\mathbf{M}_k}\left[s_{r,i}^2\right] = \theta\kappa^{-1}R$$

Thus, applying Berstein inequality gives

$$\mathbb{P}\left(\sum_{r=1}^m\sum_{i=1}^n N_{k,r}^\perp |s_{r,i}| \geq 2mn\theta\kappa^{-1}R\right) \leq \exp\left(-\frac{mn\theta R}{10\kappa}\right)$$

Thus, with probability at least $1 - \exp\left(-\frac{mn\theta R}{10\kappa}\right)$ we have that

$$\left\|\sum_{l=1}^p \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ \mathbf{Z}(k,t) - \mathbf{m}_k^l \circ \mathbf{Z}(0,0)\right)\right\|_F^2 \leq n\theta\kappa^{-1}R$$

Take a union bound over all k, t, and apply overparameterization gives that with probability at least $1-\delta$, it holds that

$$\left\|\sum_{l=1}^p \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ \mathbf{Z}(k,t) - \mathbf{m}_k^l \circ \mathbf{Z}(0,0)\right)\right\|_F \leq \sqrt{n\theta\kappa^{-1}R}$$

Then we bound $\left\|\sum_{l=1}^p \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ \mathbf{Z}(0,0)\right) - \theta\mathbf{Z}(0)\right\|_F$. We have

$$\begin{aligned}\left\|\sum_{l=1}^p \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ \mathbf{Z}(0,0)\right) - \theta\mathbf{Z}(0)\right\|_F^2 &= \frac{1}{m}\sum_{r=1}^m\sum_{i=1}^n \left\|\left(\eta_{k,r}\sum_{l=1}^p m_{k,r}^l - \theta\right) a_r \mathbf{x}_i \mathbb{I}\{\langle \mathbf{w}_{0,r}, \mathbf{x}_i\rangle \geq 0\}\right\|_2^2 \\ &= \frac{1}{m}\sum_{r=1}^m\sum_{i=1}^n \left(N_{k,r}^\perp - \theta\right)^2\end{aligned}$$

Since we have $\mathbb{E}_{\mathbf{M}_k}\left[N_{k,r}^\perp\right] = \theta$, and $Var_{\mathbf{M}_k}(N_{k,r}^\perp) = \theta - \theta^2$ we can again apply Berstein inequality to get that

$$\mathbb{P}\left(\sum_{r=1}^m\sum_{i=1}^n (N_{k,r}^\perp - \theta)^2 \geq 2mn\theta(1-\theta)\right) \leq \exp\left(-\frac{mn\theta(1-\theta)}{10}\right)$$

Thus, with probability at least $1 - \exp\left(-\frac{mn\theta(1-\theta)}{10}\right)$ we have that

$$\left\|\sum_{l=1}^p \boldsymbol{\eta}_k \left(\mathbf{m}_k^l \circ \mathbf{Z}(0,0)\right) - \theta\mathbf{Z}(0)\right\|_F^2 \leq 2n\theta(1-\theta)$$

Take a union bound over all k and apply overparameterization gives that with probability at least $1-\delta$, it holds that

$$\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(0,0)\right)-\theta\mathbf{Z}(0)\right\|_F\leq\sqrt{n\theta(1-\theta)}$$

Thus, we have that

$$\begin{aligned}\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(k,t)\right)-\theta\mathbf{Z}(0)\right\|_F&\leq\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(0,0)\right)-\theta\mathbf{Z}(0)\right\|_F+\\&\quad\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(k,t)-\mathbf{m}_k^l\circ\mathbf{Z}(0,0)\right)\right\|_F\\&\leq O\left(\sqrt{n}\right)\end{aligned}$$

with $R\leq\frac{\kappa}{n}$. Therefore,

$$\begin{aligned}Q_4&\leq\eta\sum_{k'=0}^{k-1}\sum_{t=0}^{\tau-1}\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(k,t)\right)-\theta\mathbf{Z}(0)\right\|_F\|\mathbf{u}_{k'}-\mathbf{y}\|_2\\&\leq O(\eta\tau\sqrt{n})\sum_{k'=0}^{k-1}\left(\left(1-\frac{1}{8}\eta\theta\tau\lambda_0\right)^{k'}O\left(\sqrt{\frac{n}{\delta}}\right)+O(1)\right)\\&=O\left(\frac{\eta\tau nK}{\lambda_0\sqrt{\delta}}\right)\end{aligned}$$

I.2.5 BOUND ON Q_5

To bound Q_5, we first note that

$$\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(k,t)\right)(\hat{\mathbf{u}}_{k,t}-\mathbf{u}_k)\right\|_F^2\leq\frac{1}{m}\sum_{r=1}^{m}\sum_{i=1}^{n}\sum_{l=1}^{p}\eta_{k,r}m_{k,r}^l(\hat{u}_{k,t}^{l(i)}-u_k^{(i)})^2$$

And, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{a}}\left[(\hat{u}_{k,t}^{l(i)}-u_k^{(i)})^2\right]&=\frac{1}{m}\sum_{r=1}^{m}\left(m_{k,r}^l\sigma\left(\langle\mathbf{w}_{k,t,r}^l,\mathbf{x}_i\rangle\right)-\xi\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)^2\\&\leq\frac{1}{m}\sum_{r=1}^{m}\left(\left\|\mathbf{w}_{k,t,r}^l-\mathbf{w}_{k,r}\right\|_2+\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle^2\right)\end{aligned}$$

since

$$\begin{aligned}\left(m_{k,r}^l\sigma\left(\langle\mathbf{w}_{k,t,r}^l,\mathbf{x}_i\rangle\right)-\xi\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)^2&\leq\left(\sigma\left(\langle\mathbf{w}_{k,t,r}^l,\mathbf{x}_i\rangle\right)-\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)\right)^2+\\&\quad(m_{k,r}^l-\xi)^2\sigma(\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle)^2\\&=\left\|\mathbf{w}_{k,t,r}^l-\mathbf{w}_{k,r}\right\|_2^2+\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle^2\end{aligned}$$

With

$$\begin{aligned}\left\|\mathbf{w}_{k,t,r}^l-\mathbf{w}_{k,r}\right\|_2&\leq\frac{\eta\tau\sqrt{2nK}}{\sqrt{m\delta}}\mathbb{E}_{[\mathbf{M}_{k-1}],\mathbf{W}_0,\mathbf{a}}\left[\|\mathbf{y}-\mathbf{u}_k\|_2\right]+2\eta\tau\kappa n\sqrt{\frac{2\xi(1-\xi)pK}{m\delta}}\\&\leq\left(1-\frac{1}{8}\eta\theta\tau\lambda_0\right)^k O\left(\frac{\eta\tau n\sqrt{K}}{\delta\sqrt{m}}\right)+O\left(\eta\tau\kappa n\sqrt{\frac{pK}{m\delta}}\right):=B_3\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{a}}\left[\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(k,t)\right)\left(\hat{\mathbf{u}}_{k,t}-\mathbf{u}_k\right)\right\|_F^2\right] &\leq \frac{1}{m^2}\sum_{r,r'=1}^{m}\sum_{i=1}^{n}\sum_{l=1}^{p}\eta_{k,r}m_{k,r}^l(B_3^2+\|\mathbf{w}_{k,r}\|_2^2) \\
&\leq nB_3^2+\frac{1}{m}\sum_{r=1}^{n}\sum_{i=1}^{n}\langle\mathbf{w}_{k,r},\mathbf{x}_i\rangle^2 \\
&\leq nB_3^2+2n\kappa^2
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}_{\mathbf{a}}\left[Q_5\right] &\leq \eta\sum_{k'=0}^{k-1}\sum_{t=0}^{\tau-1}\mathbb{E}_{\mathbf{a}}\left[\left\|\sum_{l=1}^{p}\boldsymbol{\eta}_k\left(\mathbf{m}_k^l\circ\mathbf{Z}(k,t)\right)\left(\hat{\mathbf{u}}_{k,t}-\mathbf{u}_k\right)\right\|_F\right] \\
&\leq \eta\tau\sqrt{n}\left(\sum_{k'=0}^{k-1}B_3+\kappa\right) \\
&\leq O(\eta\tau\sqrt{n}\kappa)
\end{aligned}$$