Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks

Yuanzhi Li
Machine Learning Department
Carnegie Mellon University
yuanzhil@andrew.cmu.edu

Colin Wei
Computer Science Department
Stanford University
colinwei@stanford.edu

Tengyu Ma
Computer Science Department
Stanford University
tengyuma@stanford.edu

Abstract

Stochastic gradient descent with a large initial learning rate is widely used for training modern neural net architectures. Although a small initial learning rate allows for faster training and better test performance initially, the large learning rate achieves better generalization soon after the learning rate is annealed. Towards explaining this phenomenon, we devise a setting in which we can prove that a two layer network trained with large initial learning rate and annealing provably generalizes better than the same network trained with a small learning rate from the start. The key insight in our analysis is that the order of learning different types of patterns is crucial: because the small learning rate model first memorizes easy-to-generalize, hard-to-fit patterns, it generalizes worse on hard-to-generalize, easier-to-fit patterns than its large learning rate counterpart. This concept translates to a larger-scale setting: we demonstrate that one can add a small patch to CIFAR-10 images that is immediately memorizable by a model with small initial learning rate, but ignored by the model with large learning rate until after annealing. Our experiments show that this causes the small learning rate model's accuracy on unmodified images to suffer, as it relies too much on the patch early on.

1 Introduction

It is a commonly accepted fact that a large initial learning rate is required to successfully train a deep network even though it slows down optimization of the train loss. Modern state-of-the-art architectures typically start with a large learning rate and anneal it at a point when the model's fit to the training data plateaus [25, 32, 17, 42]. Meanwhile, models trained using only small learning rates have been found to generalize poorly despite enjoying faster optimization of the training loss.

A number of papers have proposed explanations for this phenomenon, such as sharpness of the local minima [22, 20, 24], the time it takes to move from initialization [18, 40], and the scale of SGD noise [38]. However, we still have a limited understanding of a surprising and striking part of the large learning rate phenomenon: from looking at the section of the accuracy curve before annealing, it would appear that a small learning rate model should outperform the large learning rate model in both training and test error. Concretely, in Fig. 1, the model trained with small learning rate outperforms

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

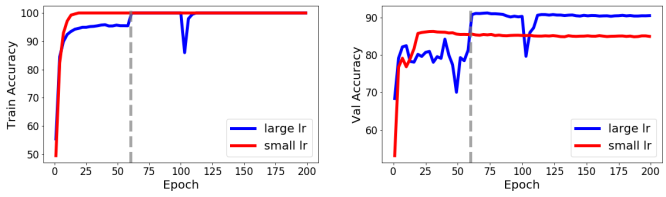

Figure 1: CIFAR-10 accuracy vs. epoch for WideResNet with weight decay, no data augmentation, and initial lr of 0.1 vs. 0.01. Gray represents the annealing time. **Left:** Train. **Right:** Validation.

the large learning rate until epoch 60 when the learning rate is first annealed. Only after annealing does the large learning rate visibly outperform the small learning rate in terms of generalization.

In this paper, we propose to theoretically explain this phenomenon via the concept of *learning order* of the model, i.e., the rates at which it learns different types of examples. This is not a typical concept in the generalization literature — learning order is a training-time property of the model, but most analyses only consider post-training properties such as the classifier's complexity [8], or the algorithm's output stability [9]. We will construct a simple distribution for which the learning order of a two-layer network trained under large and small initial learning rates determines its generalization.

Informally, consider a distribution over training examples consisting of two types of patterns ("pattern" refers to a grouping of features). The first type consists of a set of easy-to-generalize (i.e., discrete) patterns of low cardinality that is difficult to fit using a low-complexity classifier, but easily learnable via complex classifiers such as neural networks. The second type of pattern will be learnable by a low-complexity classifier, but are inherently noisy so it is difficult for the classifier to generalize. In our case, the second type of pattern requires *more samples* to correctly learn than the first type. Suppose we have the following split of examples in our dataset:

$$\begin{aligned} &20\% \text{ containing only easy-to-generalize and hard-to-fit patterns} \\ &20\% \text{ containing only hard-to-generalize and easy-to-fit patterns} \\ &60\% \text{ containing both pattern types} \end{aligned} \tag{1.1}$$

The following informal theorems characterize the learning order and generalization of the large and small initial learning rate models. They are a dramatic simplification of our Theorems 3.4 and 3.5 meant only to highlight the intuitions behind our results.

Theorem 1.1 (Informal, large initial LR + anneal)**.** *There is a dataset with size N of the form (1.1) such that with a large initial learning rate and noisy gradient updates, a two layer network will:*

1) initially only learn hard-to-generalize, easy-to-fit patterns from the $0.8N$ examples containing such patterns.

2) learn easy-to-generalize, hard-to-fit patterns only after the learning rate is annealed.

Thus, the model learns hard-to-generalize, easily fit patterns with an effective sample size of $0.8N$ and still learns all easy-to-generalize, hard to fit patterns correctly with $0.2N$ samples.

Theorem 1.2 (Informal, small initial LR)**.** *In the same setting as above, with small initial learning rate the network will:*

1) quickly learn all easy-to-generalize, hard-to-fit patterns.

2) ignore hard-to-generalize, easily fit patterns from the $0.6N$ examples containing both pattern types, and only learn them from the $0.2N$ examples containing only hard-to-generalize patterns.

Thus, the model learns hard-to-generalize, easily fit patterns with a smaller effective sample size of $0.2N$ and will perform relatively worse on these patterns at test time.

Together, these two theorems can justify the phenomenon observed in Figure 1 as follows: in a real-world network, the large learning rate model first learns hard-to-generalize, easier-to-fit patterns and is unable to memorize easy-to-generalize, hard-to-fit patterns, leading to a plateau in accuracy. Once the learning rate is annealed, it is able to fit these patterns, explaining the sudden spike in both train and test accuracy. On the other hand, because of the low amount of SGD noise present in easy-to-generalize, hard-to-fit patterns, the small learning rate model quickly overfits to them before fully learning the hard-to-generalize patterns, resulting in poor test error on the latter type of pattern.

Both intuitively and in our analysis, the non-convexity of neural nets is crucial for the learning-order effect to occur. Strongly convex problems have a unique minimum, so what happens during training does not affect the final result. On the other hand, we show the non-convexity causes the learning order to highly influence the characteristics of the solutions found by the algorithm.

In Section F.1, we propose a mitigation strategy inspired by our analysis. In the same setting as Theorems 1.1 and 1.2, we consider training a model with small initial learning rate while adding noise before the activations which gets reduced by some constant factor at some particular epoch in training. We show that this algorithm provides the same theoretical guarantees as the large initial learning rate, and we empirically demonstrate the effectiveness of this strategy in Section 6. In Section 6 we also empirically validate Theorems 1.1 and 1.2 by adding an artificial memorizable patch to CIFAR-10 images, in a manner inspired by (1.1).

1.1 Related Work

The question of training with larger batch sizes is closely tied with learning rate, and many papers have empirically studied large batch/small LR phenomena [22, 18, 35, 34, 11, 41, 16, 38], particularly focusing on vision tasks using SGD as the optimizer.[1] Keskar et al. [22] argue that training with a large batch size or small learning rate results in sharp local minima. Hoffer et al. [18] propose training the network for longer and with larger learning rate as a way to train with a larger batch size. Wen et al. [38] propose adding Fisher noise to simulate the regularization effect of small batch size.

Adaptive gradient methods are a popular method for deep learning [14, 43, 37, 23, 29] that adaptively choose different step sizes for different parameters. One motivation for these methods is reducing the need to tune learning rates [43, 29]. However, these methods have been observed to hurt generalization performance [21, 10], and modern architectures often achieve the best results via SGD and hand-tuned learning rates [17, 42]. Wilson et al. [39] construct a toy example for which ADAM [23] generalizes provably worse than SGD. Additionally, there are several alternative learning rate schedules proposed for SGD, such as warm-restarts [28] and [33]. Ge et al. [15] analyze the exponentially decaying learning rate and show that its final iterate achieves optimal error in stochastic optimization settings, but they only analyze convex settings.

There are also several recent works on implicit regularization of gradient descent that establish convergence to some idealized solution under particular choices of learning rate [27, 36, 1, 7, 26]. In contrast to our analysis, the generalization guarantees from these works would depend only on the complexity of the final output and not on the order of learning.

Other recent papers have also studied the order in which deep networks learn certain types of examples. Mangalam and Prabhu [30] and Nakkiran et al. [31] experimentally demonstrate that deep networks may first fit examples learnable by "simpler" classifiers. For our construction, we prove that the neural net with large learning rate follows this behavior, initially learning a classifier on linearly separable examples and learning the remaining examples after annealing. However, the phenomenon that we analyze is also more nuanced: with a small learning rate, we prove that the model first learns a complex classifier on low-noise examples which are not linearly separable.

Finally, our proof techniques and intuitions are related to recent literature on global convergence of gradient descent for over-parametrized networks [6, 12, 13, 1, 5, 7, 4, 26, 2]. These works show that gradient descent learns a *fixed* kernel related to the initialization under sufficient over-parameterization. In our analysis, the underlying kernel is *changing* over time. The amount of noise due to SGD governs the space of possible learned kernels, and as a result, regularizes the order of learning.

2 Setup and Notations

Data distribution. We formally introduce our data distribution, which contains examples supported on two types of components: a $\mathcal{P}$ component meant to model hard-to-generalize, easier-to-fit patterns, and a $\mathcal{Q}$ component meant to model easy-to-generalize, hard-to-fit patterns (see the discussion in our introduction). Formally, we assume that the label y has a uniform distribution over $\{-1, 1\}$, and the

[1] While these papers are framed as a study of large-batch training, a number of them explicitly acknowledge the connection between large batch size and small learning rate.

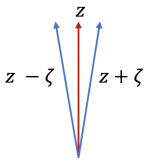

Figure 2: A visualization of the vectors z, $z-\zeta$, and $z+\zeta$ used to define the distribution $\mathcal{Q}$ in 2 dimensions. $z \pm \zeta$ will have label -1 and z has label $+1$. Note that the norm of ζ is much smaller than the norm of z.

data x is generated as

$$\text{Conditioned on the label } y \tag{2.1}$$
$$\text{with probability } p_0, \quad x_1 \sim \mathcal{P}_y, \text{ and } x_2 = 0 \tag{2.2}$$
$$\text{with probability } q_0, \quad x_1 = 0, \text{ and } x_2 \sim \mathcal{Q}_y \tag{2.3}$$
$$\text{with probability } 1-p_0-q_0, \quad x_1 \sim \mathcal{P}_y, \text{ and } x_2 \sim \mathcal{Q}_y \tag{2.4}$$

where $\mathcal{P}_{-1}, \mathcal{P}_1$ are assumed to be two half Gaussian distributions with a margin γ_0 between them:

$$x_1 \sim \mathcal{P}_1 \Leftrightarrow x_1 = \gamma_0 w^\star + z | \langle w^\star, z \rangle \geq 0, \text{ where } z \sim \mathcal{N}(0, I_{d\times d}/d)$$
$$x_1 \sim \mathcal{P}_{-1} \Leftrightarrow x_1 = -\gamma_0 w^\star + z | \langle w^\star, z \rangle \leq 0, \text{ where } z \sim \mathcal{N}(0, I_{d\times d}/d)$$

Therefore, we see that when x_1 is present, the linear classifier $\text{sign}({w^\star}^\top x_1)$ can classify the example correctly with a margin of γ_0. To simplify the notation, we assume that $\gamma_0 = 1/\sqrt{d}$ and $w^\star \in \mathbb{R}^d$ has a unit ℓ_2 norm. Intuitively, $\mathcal{P}$ is linearly separable, thus learnable by *low complexity* (e.g. linear) classifiers. However, because of the dimensionality, $\mathcal{P}$ has high noise and requires a relatively large sample complexity to learn. The distribution $\mathcal{Q}_{-1}$ and $\mathcal{Q}_1$ are supported only on three distinct directions $z-\zeta, z$ and $z+\zeta$ with some random scaling α, and are thus low-noise and memorizable. Concretely, $z-\zeta$ and $z+\zeta$ have negative labels and z has positive labels.

$$\begin{aligned} x_2 \sim \mathcal{Q}_1 &\Leftrightarrow x_2 = \alpha z \text{ with } \alpha \sim [0,1] \text{ uniformly} \\ x_2 \sim \mathcal{Q}_{-1} &\Leftrightarrow x_2 = \alpha(z + b\zeta) \text{ with } \alpha \sim [0,1], b \sim \{-1,1\} \text{ uniformly} \end{aligned} \tag{2.5}$$

Here for simplicity, we take z to be a unit vector in $\mathbb{R}^d$. We assume $\zeta \in \mathbb{R}^d$ has norm $\|\zeta\|_2 = r$ and $\langle z, \zeta \rangle = 0$. We will assume $r \ll 1$ so that $z+\zeta, z, z-\zeta$ are fairly close to each other. We depict $z-\zeta, z, z+\zeta$ in Figure 2. We choose this type of $\mathcal{Q}$ to be the easy-to-generalize, hard-to-fit pattern. Note that z is not linearly separable from $z+\zeta, z-\zeta$, so non-linearity is necessary to learn $\mathcal{Q}$. On the other hand, it is also easy for *high-complexity* models such as neural networks to memorize $\mathcal{Q}$ with relatively small sample complexity.

Memorizing $\mathcal{Q}$ with a two-layer net. It is easy for a two-layer relu network to memorize the labels of x_2 using two neurons with weights w, v such that $\langle w, z \rangle < 0$, $\langle w, z-\zeta \rangle > 0$ an $\langle v, z \rangle < 0$, $\langle v, z+\zeta \rangle > 0$. In particular, we can verify that $-\langle w, x_2 \rangle_+ - \langle v, x_2 \rangle_+$ will output a negative value for $x_2 \in \{z-\zeta, z+\zeta\}$ and a zero value for $x_2 = z$. Thus choosing a small enough $\rho > 0$, the classifier $-\langle w, x_2 \rangle_+ - \langle v, x_2 \rangle_+ + \rho$ gives the correct sign for the label y.

We assume that we have a training dataset with N examples $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(N)}, y^{(N)})\}$ drawn i.i.d from the distribution described above. We use p and q to denote the empirical fraction of data points that are drawn from equation (2.2) and (2.3).

Two-layer neural network model. We will use a two-layer neural network with relu activation to learn the data distribution described above. The first layer weights are denoted by $U \in \mathbb{R}^{m\times 2d}$ and the second layer weight is denoted by $u \in \mathbb{R}^m$. With relu activation, the output of the neural network is $u^\top(\mathbb{1}(Ux) \odot Ux)$ where $\odot$ denotes the element-wise dot product of two vectors and $\mathbb{1}(z)$ is the binary vector that contains $\mathbf{1}(z_i \geq 0)$ as entries. It turns out that we will often be concerned with the object that disentangles the two occurrences of U in the formula $u^\top(\mathbb{1}(Ux) \odot Ux)$. We define the following notation to facilitate the reference to such an object. Let

$$N_A(u, U; x) \triangleq w^\top (\mathbb{1}(Ax) \odot Ux) \tag{2.6}$$

That is, $N_A(w, W; x)$ denotes the function where we compute the activation pattern $\mathbb{1}(Ax)$ by the matrix A instead of U. When u is clear from the context, with slight abuse of notation, we write $N_A(U; x) \triangleq u^\top (\mathbb{1}(Ax) \odot Ux)$. In this notation, our model is defined as $f(u, U; x) = N_U(u, U; x)$. We consider several different structures regarding the weight matrices U. The simplest version

which we consider in the main body of this paper is that U can be decomposed into two $U = \begin{bmatrix} W \\ V \end{bmatrix}$ where W only operates on the first d coordinates (that is, the last d columns of W are zero), and V only operates on the last d coordinates (those coordinates of x_2.) Note that W operates on the $\mathcal{P}$ component of examples, and V operates on the $\mathcal{Q}$ component of examples. In this case, the model can be decomposed into

$$f(u, U; x) = N_U(u, U; x) = N_W(w, W; x) + N_V(v, V; x) = N_W(w, W; x_1) + N_V(v, V; x_2)$$

Here we slightly abuse the notation to use W to denote both a matrix of $2d$ columns with last d columns being zero, or a matrix of d columns. We also extend our theorem to other U such as a two layer convolution network in Section F.

Training objective. Let $\ell(f; (x, y))$ be the loss of the example (x, y) under model f. Throughout the paper we use the logistic loss $\ell(f; (x, y)) = -\log \frac{1}{1+e^{-yf(x)}}$. We use the standard training loss function $\widehat{L}$ defined as: $\widehat{L}(u, U) = \frac{1}{N} \sum_{i \in [N]} \ell\left(f(u, U; \cdot); (x^{(i)}, y^{(i)})\right)$ and let $\widehat{L}_{\mathcal{S}}(u, U)$ denote the average over some subset $\mathcal{S}$ of examples instead of the entire dataset.

We consider a regularized training objective $\widehat{L}_\lambda(u, U) = \widehat{L}(u, U) + \frac{\lambda}{2}\|U\|_F^2$. For the simplicity of derivation, the second layer weight vector u is random initialized and fixed throughout this paper. Thus with slight abuse of notation the training objective can be written as $\widehat{L}_\lambda(U) = \widehat{L}(u, U) + \frac{\lambda}{2}\|U\|_F^2$.

Notations. Here we collect additional notations that will be useful throughout our proofs. The symbol $\oplus$ will refer to the symmetric difference of two sets or two binary vectors. The symbol $\setminus$ refers to the set difference. Let us define $\mathcal{M}_1$ to be the set of all $i \in [N]$ such that $x_1^{(i)} \neq 0$, let $\bar{\mathcal{M}}_1 = [N] \setminus \mathcal{M}_1$. Let $\mathcal{M}_2$ to be the set of all $i \in [N]$ such that $x_2^{(i)} \neq 0$, let $\bar{\mathcal{M}}_2 = [N] \setminus \mathcal{M}_2$. We define $q = \frac{|\bar{\mathcal{M}}_1|}{N}$ and $p = \frac{|\bar{\mathcal{M}}_2|}{N}$ to be the empirical fraction of data containing patterns only from $\mathcal{Q}$ and $\mathcal{P}$, respectively. We will sometimes use $\widehat{\mathbb{E}}$ to denote an empirical expectation over the training samples. For a vector or matrix v, we use $\mathrm{supp}(v)$ to denote the set of indices of the non-zero entries of v. For $U \in \mathbb{R}^{m \times d}$ and $R \subset [m]$, let U^R be the restriction of U to the subset of rows indexed by R. We use $[U]_i$ to denote the i-th row of U as a row vector in $\mathbb{R}^{1 \times d}$. Let the symbol $\odot$ denote the element-wise product between two vectors or matrices. The notation $I_{n \times n}$ will denote the $n \times n$ identity matrix, and $\mathbf{1}$ the all 1's vector where dimension will be clear from context. We define "with high probability" to mean with probability at least $1 - e^{-C \log^2(d)}$ for a sufficiently large constant C. $\tilde{O}, \tilde{\Omega}$ will be used to hide polylog factors of d.

3 Main Results

The training algorithm that we consider is stochastic gradient descent with spherical Gaussian noise. We remark that we analyze this algorithm as a simplification of the minibatch SGD noise encountered when training real-world networks. There are a number of works theoretically characterizing this particular noise distribution [19, 18, 38], and we leave analysis of this setting to future work.

We initialize U_0 to have i.i.d. entries from a Gaussian distribution with variance τ_0^2, and at each iteration of gradient descent we add spherical Gaussian noise with coordinate-wise variance τ_ξ^2 to the gradient updates. That is, the learning algorithm for the model is

$$U_0 \sim \mathcal{N}(0, \tau_0^2 I_{m \times m} \otimes I_{d \times d})$$

$$U_{t+1} = U_t - \gamma_t \nabla_U (\widehat{L}_\lambda(u, U_t) + \xi_t) = (1 - \gamma_t \lambda) U_t - \gamma_t (\nabla_U \widehat{L}(u, U_t) + \xi_t) \tag{3.1}$$

$$\text{where } \xi_t \sim \mathcal{N}(0, \tau_\xi^2 I_{m \times m} \otimes I_{d \times d}) \tag{3.2}$$

where γ_t denotes the learning rate at time t. We will analyze two algorithms:

> **Algorithm 1** (L-S): The learning rate is η_1 for t_0 iterations until the training loss drops below the threshold $\varepsilon_1 + q \log 2$. Then we anneal the learning rate to $\gamma_t = \eta_2$ (which is assumed to be much smaller than η_1) and run until the training loss drops to ε_2.
>
> **Algorithm 2** (S): We used a fixed learning rate of η_2 and stop at training loss $\varepsilon_2' \leq \varepsilon_2$.

For the convenience of the analysis, we make the following assumption that we choose τ_0 in a way such that the contribution of the noises in the system stabilize at the initialization:[2]

Assumption 3.1. *After fixing λ and τ_ξ, we choose initialization τ_0 and large learning rate η_1 so that*

$$(1-\eta_1\lambda)^2\tau_0^2 + \eta_1^2\tau_\xi^2 = \tau_0^2 \tag{3.3}$$

As a technical assumption for our proofs, we will also require $\eta_1 \lesssim \varepsilon_1$.

We also require sufficient over-parametrization.

Assumption 3.2 (Over-parameterization)**.** *We assume throughout the paper that $\tau_0 = 1/\text{poly}\left(\frac{d}{\varepsilon}\right)$ and $m \geq \text{poly}\left(\frac{d}{\varepsilon\tau_0}\right)$ where* poly *is a sufficiently large constant degree polynomial. We note that we can choose τ_0 arbitrarily small, so long as it is fixed before we choose m.*

As we will see soon, the precise relation between N, d implies that the level of over-parameterization is polynomial in N, ϵ, which fits with the conditions assumed in prior works, such as [26, 13].

Assumption 3.3. *Throughout this paper, we assume the following dependencies between the parameters. We assume that $N, d \to \infty$ with a relationship $\frac{N}{d} = \frac{1}{\kappa^2}$ where $\kappa \in (0,1)$ is a* small value.[3] *We set $r = d^{-3/4}$, $p_0 = \kappa^2/2$, and $q_0 = \Theta(1)$. The regularizer will be chosen to be $\lambda = d^{-5/4}$. All of these choices of hyper-parameters can be relaxed, but for simplicity of exposition we only work this setting.*

We note that under our assumptions, for sufficiently large N, $p \approx p_0$ and $q \approx q_0$ up to constant multiplicative factors. Thus we will mostly work with p and q (the empirical fractions) in the rest of the paper. We also note that our parameter choice satisfies $(rd)^{-1}, d\lambda, \lambda/r \leq \kappa^{O(1)}$ and $\lambda \leq r^2/(\kappa^2 q^3 p^2)$, which are a few conditions that we frequently use in the technical part of the paper.

Now we present our main theorems regarding the generalization of models trained with the L-S and S algorithms. The final generalization error of the model trained with the L-S algorithm will end up a factor $O(\kappa) = O(p^{1/2})$ smaller than the generalization error of the model trained with S algorithm.

Theorem 3.4 (Analysis of Algorithm L-S)**.** *Under Assumption 3.1, 3.2, and 3.3, there exists a universal constant $0 < c < 1/16$ such that Algorithm 1* (L-S) *with annealing at loss $\varepsilon_1 + q\log 2$ for $\varepsilon_1 \in \left(d^{-c}, \kappa^2 p^2 q^3\right)$ and stopping criterion $\varepsilon_2 = \sqrt{\varepsilon_1/q}$ satisfies the following:*

1. *It anneals the learning rate within $\widetilde{O}\left(\frac{d}{\eta_1\varepsilon_1}\right)$ iterations.*
2. *It stops at at most $t = \widetilde{O}\left(\frac{d}{\eta_1\varepsilon_1} + \frac{1}{\eta_2 r\varepsilon_1^3}\right)$. With probability at least 0.99, the solution U_t has test (classification) error and test loss* ***at most*** *$O\left(p\kappa\log\frac{1}{\varepsilon_1}\right)$.*

Roughly, the learning order and generalization of the L-S model is as follows: before annealing the learning rate, the model only learns an effective classifier for $\mathcal{P}$ on the $\approx (1-q)N$ samples in $\mathcal{M}_1$ as the large learning rate creates too much noise to effectively learn $\mathcal{Q}$ (Lemma 4.1 and Lemma 4.2). After the learning rate is annealed, the model memorizes $\mathcal{Q}$ and correctly classifies examples with only a $\mathcal{Q}$ component during test time (formally shown in Lemmas 4.3 and 4.4). For examples with only $\mathcal{P}$ component, the generalization error is (ignoring log factors and other technicalities) $p\sqrt{\frac{d}{N}} = O(p\kappa)$ via standard Rademacher complexity. The full analysis of the L-S algorithm is clarified in Section 4.

Theorem 3.5 (Lower bound for Algorithm S)**.** *Let ε_2 be chosen in Theorem 3.4. Under Assumption 3.1, 3.2 and 3.3, there exists a universal constant $c > 0$ such that w.h.p, Algorithm 2 with any $\eta_2 \leq \eta_1 d^{-c}$ and any stopping criterion $\varepsilon_2' \in (d^{-c}, \varepsilon_2]$, achieves training loss ε_2' in at most $\widetilde{O}\left(\frac{d}{\eta_2\varepsilon_2'}\right)$ iterations, and both the test error and the test loss of the obtained solution are* ***at least*** *$\Omega(p)$.*

[2] Let τ_0' be the solution to (3.3) holding $\tau_\xi, \eta_1, \lambda$ fixed. If the standard deviation of the initialization is chosen to be smaller than τ_0', then standard deviation of the noise will grow to τ_0'. Otherwise if the initialization is chosen to be larger, the contribution of the noise will decrease to the level of τ_0' due to regularization. In typical analysis of SGD with spherical noises, often as long as either the noise or the learning rate is small enough, the proof goes through. However, here we will make explicit use of the large learning rate or the large noise to show better generalization performance.

[3] Or in a non-asymptotic language, we assume that N, d are sufficiently large compared to κ: $N, d \gg \text{poly}(\kappa)$

We explain this lower bound as follows: the S algorithm will quickly memorize the $\mathcal{Q}$ component which is low noise and ignore the $\mathcal{P}$ component for the $\approx 1-p-q$ examples with both $\mathcal{P}$ and $\mathcal{Q}$ components (shown in Lemma 5.2). Thus, it only learns $\mathcal{P}$ on $\approx pN$ examples. It obtains a small margin on these examples and therefore misclassifies a constant fraction of $\mathcal{P}$-only examples at test time. This results in the lower bound of $\Omega(p)$. We formalize the analysis in Section 5.

Decoupling the Iterates. It will be fruitful for our analysis to separately consider the gradient signal and Gaussian noise components of the weight matrix U_t. We will decompose the weight matrix U_t as follows: $U_t = \overline{U}_t + \widetilde{U}_t$. In this formula, $\overline{U}_t$ denotes the signals from all the gradient updates accumulated over time, and $\widetilde{U}_t$ refers to the noise accumulated over time:

$$\begin{aligned}\overline{U}_t &= -\sum_{s=1}^{t}\gamma_{s-1}\left(\prod_{i=s}^{t-1}(1-\gamma_i\lambda)\right)\nabla\widehat{L}(U_{s-1}) \\ \widetilde{U}_t &= \left(\prod_{i=0}^{t-1}(1-\gamma_i\lambda)\right)U_0 - \sum_{s=1}^{t}\gamma_{s-1}\left(\prod_{i=s}^{t-1}(1-\gamma_i\lambda)\right)\xi_{s-1}\end{aligned} \tag{3.4}$$

Note that when the learning rate γ_t is always η, the formula simplifies to $\overline{U}_t = \sum_{s=1}^{t}\eta(1-\eta\lambda)^{t-s}\nabla\widehat{L}(U_{s-1})$ and $\widetilde{U}_t = (1-\eta\lambda)^t U_0 + \sum_{s=1}^{t}\eta(1-\eta\lambda)^{t-s}\xi_{s-1}$. The decoupling and our particular choice of initialization satisfies that the noise updates in the system stabilize at initialization, so the marginal distribution of $\widetilde{U}_t$ is always the same as the initialization. Another nice aspect of the signal-noise decomposition is as follows: we use tools from [6] to show that if the signal term $\overline{U}$ is small, then using only the noise component $\widetilde{U}$ to compute the activations roughly preserves the output of the network. This facilitates our analysis of the network dynamics. See Section A.1 for full details.

Decomposition of Network Outputs. For convenience, we will explicitly decompose the model prediction at each time into two components, each of which operates on one pattern: we have $N_{U_t}(u, U_t; x) = g_t(x) + r_t(x)$,

$$\text{where } g_t(x) = g_t(x_2) \triangleq N_{V_t}(v, V_t; x) = N_{V_t}(v, V_t; x_2) \tag{3.5}$$

$$r_t(x) = r_t(x_1) \triangleq N_{W_t}(w, W_t; x) = N_{W_t}(w, W_t; x_1) \tag{3.6}$$

In other words, the network g_t acts on the $\mathcal{Q}$ component of examples, and the network r_t acts on the $\mathcal{P}$ component of examples.

4 Characterization of Algorithm 1 (L-S)

We characterize the behavior of algorithm L-S with large initial learning rate. We provide proof sketches in Section B.1 with full proofs in Section D.

Phase I: initial learning rate η_1. The following lemma bounds the rate of convergence to the point where the loss gets annealed. It also bounds the total gradient signal accumulated by this point.

Lemma 4.1. *In the setting of Theorem 3.4, at some time step $t_0 \le \widetilde{O}\left(\frac{d}{\eta_1\varepsilon_1}\right)$, the training loss $\widehat{L}(U_{t_0})$ becomes smaller than $q\log 2 + \epsilon_1$. Moreover, we have $\|\overline{U}_{t_0}\|_F^2 = O\left(d\log^2\frac{1}{\varepsilon_1}\right)$.*

Our proof of Lemma 4.1 views the SGD dynamics as optimization with respect to the neural tangent kernel induced by the activation patterns where the kernel is rapidly changing due to the noise terms ξ. This is in contrast to the standard NTK regime, where the activation patterns are assumed to be stable [13, 26]. Our analysis extends the NTK techniques to deal with a sequence of changing kernels which share a common optimal classifier (see Section B.1 and Theorem B.2 for additional details).

The next lemma says that with large initial learning rate, the function g_t does not learn anything meaningful for the $\mathcal{Q}$ component before the $\frac{1}{\eta_1\lambda}$-timestep. Note that by our choice of parameters $1/\lambda \gg d$ and Lemma 4.1, we anneal at the time step $\widetilde{O}\left(\frac{d}{\eta_1\varepsilon_1}\right) \le \frac{1}{\eta_1\lambda}$. Therefore, the function has not learned anything meaningful about the memorizable pattern on distribution $\mathcal{Q}$ before we anneal.

Lemma 4.2. *In the setting of Theorem 3.4, w.h.p., for every $t \le \frac{1}{\eta_1\lambda}$,*

$$|g_t(z+\zeta) + g_t(z-\zeta) - 2g_t(z)| \le \widetilde{O}\left(\frac{r^2}{\lambda}\right) = \widetilde{O}(d^{-1/4}) \tag{4.1}$$

Phase II: after annealing the learning rate to η_2. After iteration t_0, we decrease the learning rate to η_2. The following lemma bounds how fast the loss converges after annealing.

Lemma 4.3. *In the setting of Theorem 3.4, there exists $t = \widetilde{O}\left(\frac{1}{\varepsilon_1^3 \eta_2 r}\right)$, such that after $t_0 + t$ iterations, we have that*

$$\widehat{L}(U_t) = O\left(\sqrt{\varepsilon_1/q}\right)$$

Moreover, $\|\overline{U}_{t_0+t} - \overline{U}_{t_0}\|_F^2 \leq \widetilde{O}\left(\frac{1}{\varepsilon_1^2 r}\right) \leq O(d)$.

The following lemma bounds the training loss on the example subsets $\mathcal{M}_1, \bar{\mathcal{M}}_1$.

Lemma 4.4. *In the setting of Lemma 4.3 using the same $t = \widetilde{O}\left(\frac{1}{\varepsilon_1^3 \eta_2 r}\right)$, the average training losses on the subsets $\mathcal{M}_1$ and $\bar{\mathcal{M}}_1$ are both good in the sense that*

$$\widehat{L}_{\mathcal{M}_1}(r_{t_0+t}) = O(\sqrt{\varepsilon_1/q}) \text{ and } \widehat{L}_{\bar{\mathcal{M}}_1}(g_{t_0+t}) = O(\sqrt{\varepsilon_1/q^3}) \tag{4.2}$$

Intuitively, low training loss of g_{t_0+t} on $\bar{\mathcal{M}}_1$ immediately implies good generalization on examples containing patterns from $\mathcal{Q}$. Meanwhile, the classifier for $\mathcal{P}$, r_{t_0+t}, has low loss on $(1-q)N$ examples. Then the test error bound follows from standard Rademacher complexity tools applied to these $(1-q)N$ examples.

5 Characterization of Algorithm 2 (S)

We present our small learning rate lemmas, with proofs sketches in Section B.2 and full proofs in Section E.

Training loss convergence. The below lemma shows that the algorithm will converge to small training error too quickly. In particular, the norm of W_t is not large enough to produce a large margin solution for those x such that $x_2 = 0$.

Lemma 5.1. *In the setting of Theorem 3.5, there exists a time $t' = \tilde{O}\left(\frac{1}{\eta_2 \varepsilon_2'^3 r}\right)$ such that $\widehat{L}_{\mathcal{M}_2}(U_{t'}) \leq \varepsilon_2'$. Moreover, there exists t with $t = \tilde{O}\left(\frac{1}{\eta_2 \varepsilon_2'^3 r} + \frac{Np}{\eta_2 \varepsilon_2'}\right)$ such that $\widehat{L}(U_t) \leq \varepsilon_2'$ after t iterations. Moreover, we have that $\|\overline{U}_t\|_F^2 \leq \tilde{O}\left(\frac{1}{\varepsilon_2'^2 r} + Np\right)$.*

Lower bound on the generalization error. The following important lemma states that our classifier for $\mathcal{P}$ does not learn much from the examples in $\mathcal{M}_2$. Intuitively, under a small learning rate, the classifier will already learn so quickly from the $\mathcal{Q}$ component of these examples that it will not learn from the $\mathcal{P}$ component of examples in $\mathcal{M}_1 \cap \mathcal{M}_2$. We make this precise by showing that the magnitude of the gradients on $\mathcal{M}_2$ is small.

Lemma 5.2. *In the setting of theorem 3.5, let*

$$\overline{W}_t^{(2)} = \frac{1}{N}\eta_2 \sum_{s \leq t} (1 - \eta_2 \lambda)^{t-s} \sum_{i \in \mathcal{M}_2} \nabla_W \widehat{L}_{\{i\}}(U_s) \tag{5.1}$$

be the (accumulated) gradient of the weight W, restricted to the subset $\mathcal{M}_2$. Then, for every $t = O\left(d/\eta_2 \varepsilon_2'\right)$, we have: $\left\|\overline{W}_t^{(2)}\right\|_F \leq \tilde{O}\left(d^{15/32}/\varepsilon_2'^2\right)$. For notation simplicity, we will define $\varepsilon_3 = d^{-1/32}\frac{1}{\varepsilon_2'^2}$. Then, $\left\|\overline{W}_t^{(2)}\right\|_F \leq \tilde{O}\left(\sqrt{d}\varepsilon_3\right)$.

The above lemma implies that W does not learn much from examples in $\mathcal{M}_2$, and therefore must overfit to the pN examples in $\bar{\mathcal{M}}_2$. As $pN \leq d/2$ by our choice of parameters, we will not have enough samples to learn the d-dimensional distribution $\mathcal{P}$. The following lemma formalizes the intuition that the margin will be poor on samples from $\mathcal{P}$.

Lemma 5.3. *There exists $\alpha \in \mathbb{R}^d$ such that $\alpha \in span\{x_1^{(i)}\}_{i \in \bar{\mathcal{M}}_2}$ and $\|\alpha\|_2 = \tilde{\Omega}(\sqrt{Np})$ such that w.h.p. over a randomly chosen x_1, we have that*

$$r_t(x_1) - r_t(-x_1) = 2\langle \alpha, x_1 \rangle \pm \tilde{O}\left(\varepsilon_3\right) \tag{5.2}$$

As the margin is poor, the predictions will be heavily influenced by noise. We use this intuition to prove the classification lower bound for Theorem 3.5.

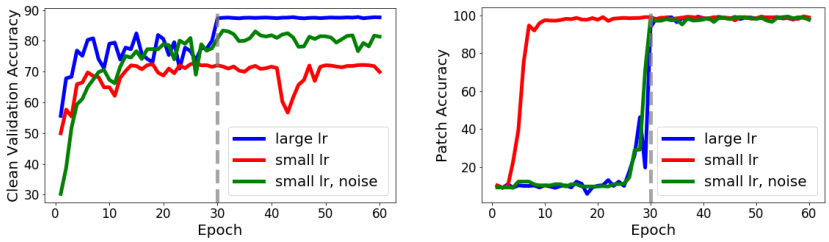

Figure 3: Accuracy vs. epoch on patch-augmented CIFAR-10. The gray line indicates annealing of activation noise and learning rate. **Left:** Clean validation set. **Right:** Images containing only the patch.

6 Experiments

Our theory suggests that adding noise to the network could be an effective strategy to regularize a small learning rate in practice. We test this empirically by adding small Gaussian noise during training *before* every activation layer in a WideResNet16 [42] architecture, as our analysis highlights pre-activation noise as a key regularization mechanism of SGD. The noise level is annealed over time. We demonstrate on CIFAR-10 images without data augmentation that this regularization can indeed counteract the negative effects of small learning rate, as we report a 4.72% increase in validation accuracy when adding noise to a small learning rate. Full details are in Section H.1.

We will also empirically demonstrate that the choice of large vs. small initial learning rate can indeed invert the learning order of different example types. We add a memorizable 7×7 pixel patch to a subset of CIFAR-10 images following the scenario presented in (1.1), such that around 20% of images have no patch, 16% of images contain only a patch, and 64% contain both CIFAR-10 data and patch. We generate the patches so that they are not easily separable, as in our constructed $\mathcal{Q}$, but they are low in variation and therefore easy to memorize. Precise details on producing the data, including a visualization of the patch, are in Section H.2. We train on the modified dataset using WideResNet16 using 3 methods: large learning rate with annealing at the 30th epoch, small initial learning rate, and small learning rate with noise annealed at the 30th epoch.

Figure 3 depicts the validation accuracy vs. epoch on clean (no patch) and patch-only images. From the plots, it is apparent that the small learning rate picks up the signal in the patch very quickly, whereas the other two methods only memorize the patch after annealing.

From the validation accuracy on clean images, we can deduce that the small learning rate method is indeed learning the CIFAR images using a small fraction of all the available data, as the validation accuracy of a small LR model when training on the full dataset is around 83%, but the validation on clean data after training with the patch is 70%. We provide additional arguments in Section H.2.

7 Conclusion

In this work, we show that the order in which a neural net learns to fit different types of patterns plays a crucial role in generalization. To demonstrate this, we construct a distribution on which models trained with large learning rates generalize provably better than those trained with small learning rates due to learning order. Our analysis reveals that more SGD noise, or larger learning rate, biases the model towards learning "generalizing" kernels rather than "memorizing" kernels. We confirm on artificially modified CIFAR-10 data that the scale of the learning rate can indeed influence learning order and generalization. Inspired by these findings, we propose a mitigation strategy that injects noise before the activations and works both theoretically for our construction and empirically. The design of better algorithms for regularizing learning order is an exciting question for future work.

Acknowledgements

CW acknowledges support from a NSF Graduate Research Fellowship.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Can SGD learn recurrent neural networks with provable generalization? *CoRR*, abs/1902.01028, 2019. URL http://arxiv.org/abs/1902.01028.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *CoRR*, abs/1905.10337, 2019. URL http://arxiv.org/abs/1905.10337.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. *arXiv preprint arXiv:1811.04918*, November 2018.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *arXiv preprint arXiv:1810.12065*, 2018.

[5] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *arXiv preprint arXiv:1810.12065*, 2018.

[6] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, November 2018.

[7] 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. *CoRR*, abs/1901.08584, 2019. URL http://arxiv.org/abs/1901.08584.

[8] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[9] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[10] Jinghui Chen and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. *arXiv preprint arXiv:1806.06763*, 2018.

[11] Xiaowu Dai and Yuhua Zhu. Towards theoretical understanding of large batch training in stochastic gradient descent. *arXiv preprint arXiv:1812.00542*, 2018.

[12] Simon S. Du, Jason D. Lee, Yuandong Tian, Barnabás Póczos, and Aarti Singh. Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. In *International Conference on Machine Learning (ICML)*. http://arxiv.org/abs/1712.00779, 2018.

[13] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. *ArXiv e-prints*, 2018.

[14] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[15] Rong Ge, Sham M. Kakade, Rahul Kidambi, and Praneeth Netrapalli. The Step Decay Schedule: A Near Optimal, Geometrically Decaying Learning Rate Procedure. *arXiv e-prints*, art. arXiv:1904.12838, Apr 2019.

[16] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

[19] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

[20] Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Dnn's sharpest directions along the sgd trajectory. *arXiv preprint arXiv:1807.05031*, 2018.

[21] Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

[22] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? *CoRR*, abs/1802.06175, 2018. URL http://arxiv.org/abs/1802.06175.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[27] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix recovery. *CoRR*, abs/1712.09203, 2017. URL http://arxiv.org/abs/1712.09203.

[28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[29] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

[30] Karttikeya Mangalam and Vinay Prabhu. Do deep neural networks learn shallow learnable examples first? June 2019.

[31] Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L. Edelman, Fred Zhang, and Boaz Barak. SGD on Neural Networks Learns Functions of Increasing Complexity. *arXiv e-prints*, art. arXiv:1905.11604, May 2019.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

[34] Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

[35] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

[36] 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.

[37] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 2012.

[38] Yeming Wen, Kevin Luk, Maxime Gazeau, Guodong Zhang, Harris Chan, and Jimmy Ba. Interplay between optimization and generalization of stochastic gradient descent with covariance noise. *arXiv preprint arXiv:1902.08234*, 2019.

[39] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

[40] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

[41] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[42] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[43] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

A Basic Properties and Toolbox

In this section, we collect a few basic properties of the neural networks we are studying. In section G, we provide two lemmas on Gaussian random variables and perturbation theory of the matrices.

Proposition A.1.

$$[\nabla \widehat{L}(U)]_i = \widehat{\mathbb{E}}\left[\ell'(f(u,U;(x,y)))\mathbb{1}([U]_i x)x\right] \tag{A.1}$$

Proposition A.2. *Let $[\nabla \widehat{L}(U)]_i$ be the i-th row of $\nabla \widehat{L}(U)$. We have that $\|[\nabla \widehat{L}(U)]_i\|_2 \lesssim 1/\sqrt{m}$.*

Proposition A.3. *For any t, if $\gamma_s = \eta$ for every $s \le t$, then we have that $\|[\overline{U}_t]_i\|_2 \lesssim \min\{\frac{1}{\sqrt{m}\lambda}, \eta t/\sqrt{m}\}$ and $\|\overline{U}_t\|_F \lesssim \frac{1}{\lambda}$.*

Proof. By equation (3.4) and Proposition A.2, we have that

$$\|[\overline{U}_t]_i\|_2 = \sum_s \eta(1-\eta\lambda)^{t-s}\|[\nabla \widehat{L}(U_s)]_i\|_2 \le \frac{1}{\sqrt{m}}\sum_s \eta(1-\eta\lambda)^{t-s} \lesssim \min\left\{\frac{1}{\sqrt{m}\lambda}, \frac{\eta t}{\sqrt{m}}\right\}$$

□

Proposition A.4. *Suppose that matrix $\widetilde{U} \in \mathbb{R}^{m\times d}$ is a random variable whose columns have i.i.d distribution $\mathcal{N}(0,\tau^2 I_{m\times m})$ and $u \in \mathbb{R}^m$ such that each entry of u is i.i.d. uniform in $\{-m^{-1/2}, m^{1/2}\}$.For every x, we have that w.h.p. over the randomness of $\widetilde{U}$ and u that*

$$\left|N_{\widetilde{U}}(u,\widetilde{U};x)\right| \lesssim \tau\|x\|_2 \log d \tag{A.2}$$

Proof of Proposition A.4. By definition, we have that

$$N_{\widetilde{U}}(u,\widetilde{U};x) = \sum_{i\in[m]} u_i[[\widetilde{U}]_i x]_+ \tag{A.3}$$

By definition, $\widetilde{U} \in \mathbb{R}^{m\times d}$ where each entry is i.i.d. $\mathcal{N}(0,\tau^2)$, which implies that when $m \ge d$, w.h.p. $\|\widetilde{U}\|_2 = O(\tau\sqrt{m})$.

Hence $\|[\widetilde{U}x]_+\|_2 \le \|\widetilde{U}x\|_2 \lesssim \tau\sqrt{m}\|x\|_2$. Now, since each u_i is i.i.d. uniform $\{-m^{-1/2}, m^{1/2}\}$, using the randomness of u_i we know that w.h.p.

$$\left|\sum_{i\in[m]} u_i[[\widetilde{U}]_i x]_+\right| \lesssim \frac{\log m}{\sqrt{m}}\|[\widetilde{U}x]_+\|_2 \lesssim \tau\|x\|_2 \log d \tag{A.4}$$

□

Proposition A.5. *Under the same setting as Lemma A.8, we will also have w.h.p over the randomness of $\widetilde{U}$ and u, $\forall \overline{U} \in \mathbb{R}^{d\times m}$,*

$$\left|N_U(u,\widetilde{U};x) - N_{\widetilde{U}}(u,\widetilde{U};x)\right| \lesssim B\|\overline{U}\|_F^{5/3}\tau^{-2/3}m^{-1/6} \tag{A.5}$$

Thus, it also follows that

$$|N_U(u,\widetilde{U};x)| \lesssim B\|\overline{U}\|_F^{5/3}\tau^{-2/3}m^{-1/6} + \tau B\log d \tag{A.6}$$

Proof. We know that for every i where $\mathbb{1}([U]_i x) \ne \mathbb{1}([\widetilde{U}]_i x)$, it holds that $|[\widetilde{U}]_i x| \le |[\overline{U}]_i x|$. This implies that

$$\left|N_U(u,\widetilde{U};x) - N_{\widetilde{U}}(u,\widetilde{U};x)\right| \le \frac{1}{\sqrt{m}}\sum_{i\in[m]} |\mathbb{1}([U]_i x) - \mathbb{1}([\widetilde{U}]_i x)||[\overline{U}]_i x| \tag{A.7}$$

$$\le \frac{1}{\sqrt{m}}\|\mathbb{1}(Ux) - \mathbb{1}(\widetilde{U}x)\|_1 \max_i |[\overline{U}]_i x| \tag{A.8}$$

$$\lesssim B\|\overline{U}\|_F^{4/3}\tau^{-4/3}m^{1/6}\max_i \|[\overline{U}]_i\|_2 \tag{A.9}$$

Here in the last inequality we applied Lemma A.8. The second statement follows from Proposition A.4 and triangle inequality. □

We have the following Rademacher complexity bound:

Lemma A.6 (Lemma G5 and 5.9 of [3]). *Let $U = \overline{U} + \widetilde{U}$, where $\widetilde{U} \in \mathbb{R}^{m\times d}$ is a random variable whose columns have i.i.d distribution $\mathcal{N}(0, \tau_0^2 I_{m\times m})$ and $u \in \mathbb{R}^m$ such that each entry of u is i.i.d. uniform in $\{-m^{-1/2}, m^{1/2}\}$. W.h.p. over the samples $\{x^{(i)}\}$ and the randomness of $u, \widetilde{U}$, we have that for every $\rho \in [0, 1/\lambda]$:*

$$\mathcal{R} := \frac{1}{\sqrt{N}} \sum_{i\in[N]} \mathbb{E}_\sigma \left[\left| \sup_{\|\bar{U}\|_F^2 \le \rho^2} \sigma_i N_U(u, \bar{U}; x^{(i)}) \right| \right] \le O(\rho + \varepsilon_s) \tag{A.10}$$

A.1 Preliminaries on Decoupling the Iterates

In this section, we collect useful statements which will help with decoupling the signal $\overline{U}$ from the noise $\widetilde{U}$ in our analysis. First, we observe that if the noise updates in the system stabilize at initialization, the marginal distribution of U_t is always the same as the initialization.

Proposition A.7. *Under Assumption 3.1, suppose we run Algorithm 1. Then for any t before annealing the learning rate, $\widetilde{U}_t$ has marginal distribution $\mathcal{N}(0, \tau_0^2 I_{m\times m} \otimes I_{d\times d})$. In other words, each entry of $\widetilde{U}_t$ follows $\mathcal{N}(0, \tau_0^2)$ and they are independent with each others.*

One nice aspect of the signal-noise decomposition is as follows: we use tools from [6] to show that if the signal term $\overline{U}$ is small, then using only the noise component $\widetilde{U}$ to compute the activations roughly preserves the output of the network. This facilitates our analysis of the network dynamics.

Lemma A.8. *[Lemma 5.2 of [6]] Let $x \in \mathbb{R}^d$ be a fixed example with $\|x\|_2 \le B$. For every $\tau > 0$, let $U = \overline{U} + \widetilde{U}$ where $\widetilde{U} \in \mathbb{R}^{m\times d}$ is a random variable whose columns have i.i.d distribution $\mathcal{N}(0, \tau^2 I_{m\times m})$ and $u \in \mathbb{R}^m$ such that each entry of u is i.i.d. uniform in $\{-m^{-1/2}, m^{1/2}\}$. We have that, w.h.p over the randomness of $\widetilde{U}$ and u, $\forall \overline{U} \in \mathbb{R}^{d\times m}$,*

$$\left| N_U(u, \overline{U}; x) - N_{\widetilde{U}}(u, \overline{U}; x) \right| \lesssim B\|\overline{U}\|_F \tau^{-2} m^{-1/6} \tag{A.11}$$

Moreover, we have that $\|\mathbb{1}(Ux) - \mathbb{1}(\widetilde{U}x)\|_1 \lesssim \|\overline{U}\|_F^{4/3} \tau^{-4/3} m^{2/3}$.

As we will often apply (A.11) with $\|\overline{U}\|_F \lesssim \frac{1}{\lambda}$, for notational simplicity we denote throughout the paper $\varepsilon_s = \left(\frac{1}{\lambda\tau_0}\right)^{4/3} m^{-1/3}$. By our choice of $m \ge \text{poly}(d/\tau_0)$ we know that $\varepsilon_s \le d^{-\Theta(1)}$.

B Proof Sketches

B.1 Proof Sketches for Large Learning Rate

We first introduce notations that will be useful in these proofs. We will explicitly decouple the noise in the weights from the signal by abstracting the loss as a function of only the signal portion $\overline{U}_t$ of the weights. Let us define the following:

$$f_t(B; x) = N_{U_t}(u, B + \widetilde{U}_t; x) \tag{B.1}$$

Moreover, we define

$$K_t(B) \triangleq \frac{1}{N} \sum_{i=1}^{N} \ell(f_t(B; \cdot); (x^{(i)}, y^{(i)})) \tag{B.2}$$

By definition, we know that

$$L_t = \widehat{L}(U_t) = K_t(\overline{U}_t) \tag{B.3}$$

$$\nabla_U \widehat{L}(U_t) = \nabla K_t(\overline{U}_t) \tag{B.4}$$

Now the proof of Lemma 4.1 relies on the following two results, which we state below and prove in Section D.1. The first says that there is a common target for the signal part of the network that is a good solution for all of the K_t.

Lemma B.1. *In the setting of Lemma 4.1, there exists a solution $U^\star$ satisfying a) $\|U^\star\|_F^2 \leq O\left(d \log^2 \frac{1}{\varepsilon_1}\right)$ and b) for every $t \geq 0$*

$$K_t(U^\star) \leq q \log 2 + \epsilon_1/2 \tag{B.5}$$

Now the second statement is a general one proving that gradient descent on a sequence of convex, but changing, functions will still find a optimum provided these functions share the same solution.

Theorem B.2. *Suppose $K_1, \ldots, K_T : \mathbb{R}^d \to \mathbb{R}^*$ is a sequence of differentiable convex functions satisfying*

1. *$\exists z^\star$ and a constant $c^\star \in \mathbb{R}^*$ such that $K_t(z^\star) \leq c^\star, \forall t = 1, \ldots, T$, and that $\|z_0 - z^\star\|_2 \leq R$, $\|z^\star\|_2 \leq R$.*
2. *K_t's are L-Lipschitz, i.e., $\|\nabla K_t(z)\|_2 \leq L, \forall z, t$*

Let $K_t^\lambda(z) \triangleq K_t(z) + \frac{\lambda}{2}\|z\|_2^2$. Consider the following iterative algorithm that starts from $z_0 \in \mathbb{R}^d$,

$$\forall t \geq 0, \quad z_{t+1} = z_t - \eta \nabla K_t^\lambda(z_t) \tag{B.6}$$

For every $\mu > 0$, we have that for $\lambda R^2 \leq \frac{1}{100}\mu$ and $\eta \leq \frac{\mu}{100(\lambda^2 R^2 + L^2)}$, $\eta T > \frac{R^2}{\mu}$, there is a $t^\star \in [T]$ such that:

$$K_{t^\star}(z_{t^\star}) \leq c^\star + \mu \tag{B.7}$$

Furthermore, the iterates satisfy $\|z_t - z^\star\|_2 \leq R$ for all $t \leq t^\star$.

Combining these two statements leads to the proof of Lemma 4.1.

Proof of Lemma 4.1. We can apply Theorem B.2 with K_t defined in (B.2) and $z^\star = U^\star$ defined in Lemma B.1, using $R = O\left(d \log^2 \frac{1}{\varepsilon_1}\right)$. We note that η_1 satisfies the conditions of Theorem B.2 by our parameter choices, which completes the proof. □

To prove Lemma 4.2, we will essentially argue in Section D.2 that the change in activations caused by the noise will prevent the model from learning $\mathcal{Q}$ with a large learning rate. This is because the examples in $\mathcal{Q}$ require a very specific configuration of activation patterns to learn correctly, and the noise will prevent the model from maintaining this configuration.

Now after we anneal the learning rate, in order to conclude Lemmas 4.3 and 4.4, the following must hold: 1) the network learns the $\mathcal{Q}$ component of the distribution and 2) the network does not forget the $\mathcal{P}$ component that it previously learned. To prove the latter, we rely on the following lemma stating that the activations do not change much with a small learning rate:

Lemma B.3. *The activation patterns do not change much after annealing the learning rate: for every $t_0, t \leq \frac{1}{\eta_2 \lambda}$, for any x and for any row $[U_t]_i$ of the weight matrix U, we have that*

$$\|\mathbb{1}([U_{t_0+t}]x) - \mathbb{1}([U_{t_0}]x)\|_1 \lesssim \sqrt{\frac{\eta_2}{\eta_1}} m + \varepsilon_s m \tag{B.8}$$

Moreover, for all $i \in [m]$, $\left\|[\overline{U}_t]_i\right\|_2 \leq \frac{1}{\lambda\sqrt{m}}$, it holds that w.h.p. for every x:

$$\left|N_{U_{t_0+t}}(u, U_{t_0+t}; x) - N_{U_{t_0}}(u, \overline{U}_{t_0+t}; x)\right| \lesssim \frac{1}{\lambda} \times \left(\sqrt{\frac{\eta_2}{\eta_1}} + \varepsilon_s\right) + \tau_0 \log d \tag{B.9}$$

We prove the above lemma in Section D.3. Now to complete the proof of Lemma 4.3, we will construct a target solution for all timesteps after annealing the learning rate based on the activations at time t_0 (as they do not change by much in subsequent time steps because of Lemma B.3) and reapply Theorem B.2. Finally, to prove Lemma 4.4, we use the fact that the W_t component of the solution does not change by much, and therefore the loss on $\mathcal{M}_1$ is still low.

B.2 Proof Sketches for Small Learning Rate

The proof of Lemma 5.1 proceeds similarly as the proof of Lemma 4.3: we will show the existence of a target solution of K_t for all iterations, and use Theorem B.2 to prove convergence to this target solution.

Now to sketch the proof of Lemma 5.2, we will first define the following notation: define $\ell'_{j,t} = \ell'(-y^{(j)} N_{U_t}(u, U_t; x^{(j)})$ to be the derivative of the loss at time t on example j. Let ρ_t be the average of the absolute value of the derivative.

$$\rho_t = \frac{1}{N} \sum_{j \in \mathcal{M}_2} \left| \ell'_{i,t} \right| \tag{B.10}$$

The next two statements argue that ρ_t can be large only in a limited number of time steps. As the training loss converges quickly with small learning rate, this will be used to argue that the $\mathcal{P}$ components of examples in $\mathcal{M}_2$ provide a very limited signal to W_t. The proofs of these statements are in Section E.2.

We first show the following lemma that says that if ρ_t is large (which means the loss is large as well), then the total gradient norm has to be big. This lemma holds because there is little noise in the $\mathcal{Q}$ component of the distribution, and therefore the gradient of V_t will be large if ρ_t is large.

Lemma B.4. *For every $t \leq \frac{1}{\eta_2 \lambda}$, we have that if $\rho_t = \Omega\left(\frac{1}{N}\right)$, then w.h.p.*

$$\|\nabla \widehat{L}(U_t)\|_F^2 \geq \Omega\left(r \rho_t^4\right) \tag{B.11}$$

Now we use the above lemma to bound the number of times when ρ_t is large.

Proposition B.5. *In the setting of Lemma 5.2, let $\mathcal{T}$ be the set of iterations where $\rho_t \geq \varepsilon_2'^2 \varepsilon_3^2$, where ε_3 is defined in Lemma 5.2. Then w.h.p, $|\mathcal{T}| \lesssim \frac{1}{r \varepsilon_2'^8 \varepsilon_3^8 \eta_2}$.*

Now if ρ_t is small, the gradient accumulated on W_t from examples in $\mathcal{M}_2$ must be small. We formalize this argument in our proof of Lemma 5.2 in Section E.2.

Lemma 5.3 will then follow by explicitly decomposing $\overline{W}_t$ into a component in $\text{span}\{x_1^{(i)}\}_{i \in \bar{\mathcal{M}}_2}$ and some remainder, which is shown to be small by Lemma 5.2. This is presented in the below lemma, which is proved in Section E.3.

Lemma B.6. *There exists real numbers $\{\alpha_k\}_{k \in \bar{\mathcal{M}}_2}$ such that for every $j \in [m]$, we have*

$$[\overline{W}_t]_j = w_j \sum_{k \in \bar{\mathcal{M}}_2} \alpha_k x_1^{(k)} \mathbb{1}([W_0]_j x_1^{(k)}) + [\overline{W}'_t]_j$$

with $\|\overline{W}'_t\|_F \leq \widetilde{O}\left(\varepsilon_3 \sqrt{d}\right)$.

This allows us to conclude Lemma 5.3 via computations carried out in Section E.3.

Finally, to complete the proof of Theorem 3.5, we will argue in Section C.2 that a classifier r_t of the form given by (5.2) cannot have small generalization error because it will be too heavily influenced by the noise in x_1.

C Proof of Main Theorems

C.1 Proof of Theorem 3.4

We start with the following lemma that shows that if g has small training error on $\bar{\mathcal{M}}_1$, then the output of g on x_2 is large compared to $\|x_2\|$. This is because for the loss to be low, g must have a good margin on x_2. However, as the norm of x_2 is roughly uniform in $[0, 1]$, the examples with small norm will force g to have larger output.

Lemma C.1 (Signal of g)**.** *W.h.p. for every $t \geq 0$ and every $\delta \geq \frac{1}{\sqrt{qN}}$, as long as $\widehat{L}_{\bar{\mathcal{M}}_1}(g_{t_0+t}) \leq \delta$, we have that: for every (x, y),*

$$y g_{t_0+t}(x_2) \gtrsim \frac{\|x\|_2}{\delta} \tag{C.1}$$

Proof of Lemma C.1. We use $\bar{\mathcal{M}}_1^{(1)}$ to denote the set of all $x_2^{(i)} \in \bar{\mathcal{M}}_1$ such that $x_2^{(i)} = \alpha(z-\zeta)$. Similarly, we use $\bar{\mathcal{M}}_1^{(2)}$ to denote the set of all $x_2^{(i)} \in \bar{\mathcal{M}}_1$ such that $x_2^{(i)} = \alpha(z+\zeta)$, and use $\bar{\mathcal{M}}_1^{(3)}$ to denote the set of all $x_2^{(i)} \in \bar{\mathcal{M}}_1$ such that $x_2^{(i)} = \alpha z$.

Let $g_{t_0+t}(z+\zeta) = \rho_1, g_{t_0+t}(z-\zeta) = \rho_2, g_{t+t_0}(z) = \rho_3$. By the positive homogeneity of ReLU, we know that for every $x_2 \in \bar{\mathcal{M}}_1^{(i)}$, it holds:

$$g_{t_0+t}(x_2) = \|x_2\|_2 \rho_i \tag{C.2}$$

Since $\widehat{L}_{\bar{\mathcal{M}}_1}(g_{t_0+t}) \le \delta$, it holds that w.h.p. for every $i \in [3]$,

$$\widehat{L}_{\bar{\mathcal{M}}_1^{(i)}}(g_{t_0+t}) \le 4\delta \tag{C.3}$$

Hence, at most 40δ fraction of $x_2 \in \bar{\mathcal{M}}_1^{(i)}$ satisfies $\ell(g_{t_0+t}; (x_2, y)) \ge \frac{1}{10}$. Since $\|x_2\|_2$ is uniform on $[0,1]$, this implies that as long as $\delta \ge \frac{1}{\sqrt{qN}}$, w.h.p., 80δ fraction of the $x_2 \in \bar{\mathcal{M}}_1^{(i)}$ satisfies that $\|x_2\|_2 = O(\delta)$. Among of these examples, at least 40δ fraction of them should satisfy $\ell(g_{t_0+t}; (x_2, y)) \le \frac{1}{10}$, which implies that $\|x_2\|\rho_i \gtrsim 1$. This implies that $\rho_i \gtrsim 1/\delta$ and the conclusion follows from equality (C.2).

□

Our proof of Theorem 3.4 now amounts to carefully checking that all examples in $\mathcal{M}_2$ are classified correctly, and the classifier r_{t_0+t} will generalize well on $\bar{\mathcal{M}}_2$.

Proof of Theorem 3.4. By Lemma 4.4, we know that for $t = \widetilde{O}\left(\frac{1}{\varepsilon_1^3 \eta_2 r}\right)$ we have $\widehat{L}_{\bar{\mathcal{M}}_1}(g_{t_0+t}) = O(\sqrt{\varepsilon_1/q^3})$. Thus applying Lemma C.1, we obtain that as long as $\varepsilon_1 \ge \frac{1}{\sqrt{N}}$ (which is implied by Assumption 3.3)

$$y g_{t_0+t}(x_2) \ge \Omega\left(\frac{\|x\|_2 \sqrt{q^3}}{\sqrt{\varepsilon_1}}\right) \tag{C.4}$$

On the other hand for r_{t_0+t}, by Lemma 4.1 and Lemma 4.3 we know that $\|\overline{W}_{t_0+t}\|_F = \widetilde{O}(\sqrt{d})$. Let us define $\mathcal{D}_{x_1}$ to be the marginal distribution of x_1. We know that $x_1 = \alpha w^\star + \beta$ where w.h.p. $|\alpha| = \tilde{O}(d^{-1/2})$ and $\beta \sim \mathcal{N}(0, 1/d \times (I - w^\star (w^\star)^\top))$. Hence we have that w.h.p. over $x_1 \sim \mathcal{D}_{x_1}$, $\|\overline{W}_{t_0+t} x_1\|_2 \le |\alpha| \|\overline{W}_{t_0+t}\|_F + d^{-1/2}\|\beta\|_2 \|\overline{W}_{t_0+t}\|_F \le \widetilde{O}(d^{-1/2}) \|\overline{W}_{t_0+t}\|_F \le \widetilde{O}(1)$.

This implies that for $x_1 \sim \mathcal{D}_{x_1}$, applying Lemma A.8 gives us

$$\begin{aligned}
|r_{t_0+t}(x_1)| &= |N_{U_{t_0+t}}(u, U_{t_0+t}; x_1)| && \text{(C.5)}\\
&\lesssim |N_{U_{t_0+t}}(u, \overline{U}_{t_0+t}; x_1)| + \frac{\varepsilon_s}{\lambda} + \tau_0 \log d && \text{(by Proposition A.5)}\\
&\lesssim \|u\|_2 \|\overline{W}_{t_0+t} x_1\|_2 + \frac{\varepsilon_s}{\lambda} + \tau_0 \log d = \widetilde{O}(1) && \text{(by our choice of } \tau_0, m)
\end{aligned}$$

Hence as long as $\|x_2\|_2 = \tilde{\Omega}(\sqrt{\varepsilon_1/q^3} \log \frac{1}{\varepsilon_1})$, it holds that

$$y(r_{t_0+t}(x_1) + g_{t_0+t}(x_2)) = \tilde{\Omega}(1) \times \log \frac{1}{\varepsilon_1} \tag{C.6}$$

This implies that $\ell(r_{t_0+t} + g_{t_0+t}; (x, y)) \le \varepsilon_1$. Otherwise, when $\|x_2\|_2 = \widetilde{O}\left(\sqrt{\varepsilon_1/q^3}\right)$, we also know that w.h.p. $\ell(r_{t_0+t} + g_{t_0+t}; (x, y)) \le \ell(r_{t_0+t}; (x, y)) = \tilde{O}(1)$, since $y g_{t_0+t}(x_2) \ge 0$. On the other hand by Lemma 4.4, we also know that

$$\widehat{L}_{\mathcal{M}_1}(r_{t_0+t}) = O(\sqrt{\varepsilon_1/q}) \tag{C.7}$$

Moreover, applying Lemma A.6 on r_{t_0+t} with $\|W_{t_0+t}\|_F^2 \le \|W_{t_0}\|_F^2 + \|W_{t_0+t} - W_{t_0}\|_F^2 \lesssim \left(d \log^2 \frac{1}{\varepsilon}\right)$ by Lemma 4.2 and Lemma 4.3, we have that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(r_{t_0+t}; (x,y)) \mid x_1 \neq 0\right] \lesssim \sqrt{\varepsilon_1/q} + \kappa \log \frac{1}{\varepsilon_1} \lesssim \kappa \log \frac{1}{\varepsilon_1} \tag{C.8}$$

where we used the fact that $\varepsilon_1 \le \kappa^2 p^2 q^3$.

It follows thats

$$\mathbb{E}\left[\ell(r_{t_0+t} + g_{t_0+t}; (x,y))\right] \tag{C.9}$$

$$\le \Pr[x_2 = 0]\mathbb{E}\left[\ell(r_{t_0+t}; (x,y))\right] + \Pr[x_2 \neq 0]\mathbb{E}\left[\ell(r_{t_0+t} + g_{t_0+t}; (x,y))\right] \tag{C.10}$$

$$\le \mathbb{E}\left[\ell(r_{t_0+t}; (x,y)) \mid x_1 \neq 0\right] \Pr[x_2 = 0] + \widetilde{O}(1) \Pr\left[x_2 \neq 0, \|x_2\|_2 = O\left(\sqrt{\varepsilon_1/q^3}\right)\right] + \varepsilon_1 \tag{C.11}$$

$$\le \widetilde{O}\left(\sqrt{\varepsilon_1/q^3}\right) + \varepsilon_1 \le O\left(p\kappa \log \frac{1}{\varepsilon_1}\right) \tag{C.12}$$

Here the last step uses the definition of ε_1 that $\varepsilon_1 \le \kappa^2 p^2 q^3$. □

C.2 Proof of Theorem 3.5

We will prove Theorem 3.5 using Lemma 5.3 by roughly arguing that the predictions made by r_t will be heavily influenced by a vector α in the low rank span of examples from $\bar{\mathcal{M}}_2$. With high probability, this vector α will be noisy and not align well with the ground truth $w^\star$, leading to mispredictions.

Proof of Theorem 3.5. Recall that ε_2' denotes the stopping criterion used in Theorem 3.5 and $\varepsilon_3 = d^{-1/32}\frac{1}{\varepsilon_2'^2}$. Using Lemma 5.3, we know that w.h.p.

$$r_t(x_1) - r_t(-x_1) = 2\langle \alpha, x_1\rangle \pm \widetilde{O}\left(\varepsilon_3\right) \tag{C.13}$$

Consider the matrix $M = (x_1^{(i)})_{i\in\bar{\mathcal{M}}_2} \in \mathbb{R}^{d\times Np}$. By definition, we know that $M = M_0 + M_1$ where $M_0 = w^\star\beta^\top$ where $\beta_i \in \{-d^{-1/2}, d^{-1/2}\}$ and M_1 is a Gaussian random matrix with each entry i.i.d. $\mathcal{N}(0, 1/d)$.

By Lemma G.2 we know that w.h.p. over the randomness of $x_1^{(i)}$'s, for $\alpha \in \text{span}\{x_1^{(i)}\}_{i\in\bar{\mathcal{M}}_2}$ we have as long as $Np \le d/2$: $\frac{\langle \alpha, w^\star\rangle}{\|\alpha\|_2\|w^\star\|_2} \le 0.9$. For every randomly chosen x_1, we can also write $x_1 = \gamma w^\star + \beta$ where $\beta \perp w^\star$ so β is independent of γ, hence

$$\langle \alpha, x_1\rangle = \gamma\langle \alpha, w^\star\rangle + \langle \alpha, \beta\rangle \tag{C.14}$$

Note that $\langle \alpha, \beta\rangle \sim \mathcal{N}(0, \sigma^2\|\alpha\|_2^2/d)$ with $\sigma \ge 0.1$, and with probability at least 0.1, $\gamma \le 2\|\alpha\|_2/\sqrt{d}$. This implies that with probability at least $\Omega(1)$ over a randomly chosen x_1 we can have:

$$\langle w^\star, x_1\rangle = \gamma < 0, \quad |\gamma| \le 2\|\alpha\|_2/\sqrt{d} \tag{C.15}$$

For β, we know that with probability at least $\Omega(1)$, we have:

$$\langle \alpha, \beta\rangle \ge 3\|\alpha\|_2/\sqrt{d} \tag{C.16}$$

Moreover, since β is independent of γ, we know that with probability $\Omega(1)$ both events can happen, in which case:

$$\langle w^\star, x_1\rangle < 0, \quad \langle \alpha, x_1\rangle = \gamma\langle \alpha, w^\star\rangle + \langle \alpha, \beta\rangle \ge \|\alpha\|_2/\sqrt{d} \tag{C.17}$$

Thus, since $\|\alpha\|_2 = \Omega(\sqrt{Np})$ by Lemma 5.3, we know that as long as

$$\frac{\sqrt{p}}{\kappa} = \frac{\sqrt{Np}}{\sqrt{d}} = \tilde{\Omega}\left(\varepsilon_3\right) \tag{C.18}$$

which is implied by $\varepsilon_3 = \widetilde{O}\left(\frac{\sqrt{p}}{\kappa}\right)$, it holds that $\langle \alpha, x_1 \rangle \geq \tilde{\Omega}(\varepsilon_3)$. This implies that

$$r_t(x_1) = r_t(-x_1) + 2\langle \alpha, x_1 \rangle \pm \widetilde{O}(\varepsilon_3) \tag{C.19}$$

$$\geq r_t(-x_1) \tag{C.20}$$

However, since $\langle w^\star, x_1 \rangle < 0$, we know that either $r_t(x_1) < 0$, which results in $r_t(-x_1) < 0$ but $\langle w^\star, -x_1 \rangle > 0$. So when $x_2 = 0$, the network classifies $(-x_1, 0)$ incorrectly. On the other hand, we have when $r_t(x_1) > 0$ the network will classify $(x_1, 0)$ incorrectly. Since $\langle w^\star, x_1 \rangle < 0$ and $r_t(x_1) \geq r_t(-x_1)$ holds with probability $\Omega(1)$, this shows that the test error is at least $\Omega(p)$. □

D Proofs for Large Learning Rate Lemmas

D.1 Proofs for Lemma 4.1

To prove Lemma 4.1, we will show that the network will learn all examples with $\mathcal{P}$ component while the learning rate is large. The key to the proof is that although the large learning rate noise only allows the network to search over coarse kernels, $\mathcal{P}$ is still learnable by these kernels because of its linearly-separable structure. To make this precise, we decompose the weights U_t Into the signal and noise components, and show that there exists a fixed "target" signal matrix which will classify $\mathcal{P}$ correctly no matter the noise matrix.

Recall our definitions of $f_t(B; x)$, $K_t(B)$ in (B.1) and (B.2), and that

$$L_t = \widehat{L}(U_t) = K_t(\overline{U}_t) \tag{D.1}$$

$$\nabla_U \widehat{L}(U_t) = \nabla K_t(\overline{U}_t) \tag{D.2}$$

Recall that Lemma B.1 leverages the linearly-separable structure of $\mathcal{P}$ to find a "target" signal matrix that correctly classifies $\mathcal{P}$ w.h.p over the noise matrix. We state its proof below.

Proof of Lemma B.1. By proposition A.3, $\|\overline{U}_t\|_F \leq O\left(\frac{1}{\lambda}\right)$. We apply Lemma A.8 as follows: by Proposition A.7, $\widetilde{U}_t$'s entry has marginal distribution $\mathcal{N}(0, \tau_0^2)$ and therefore the column of $\widetilde{U}_t$ has distribution $\mathcal{N}(0, \tau_0^2 I_{m\times m})$. Since w.h.p. $\|x\|_2 \lesssim \sqrt{\log d}$, the coupling Lemma A.8 gives

$$\|\mathbb{1}(U_t x) - \mathbb{1}(\widetilde{U}_t x)\|_0 \leq \varepsilon_s m \tag{D.3}$$

On the other hand, we also have by Proposition A.5, using the fact that $\max_i \|[\overline{U}_i]\|_2 \lesssim \frac{1}{\sqrt{m\lambda}}$, w.h.p.

$$\left| N_{U_t}(u, \widetilde{U}_t; x) \right| \lesssim \tau_0 \log d + \frac{\varepsilon_s}{\lambda} \lesssim \tau_0 \log d \tag{D.4}$$

Here in the last inequality we used the fact that the network is sufficiently over-parameterized so that $\varepsilon_s = \widetilde{O}(\tau_0 \lambda)$.

Using (D.4), noting that our choice of m, λ, τ_0 satisfies $\tau_0 \log d = o(\varepsilon_1)$, we conclude

$$\left| N_{U_t}(u, \widetilde{U}_t; x) \right| \leq \varepsilon_1/20 \tag{D.5}$$

Now, let us consider $U^* = (W^*, V^*)$ given by $V^* = 0$ and an $W^* \in \mathbb{R}^{m\times d}$ defined as: for all $i \in [m]$, $W_i^* = 20 w_i \sqrt{d} w^\star \log \frac{1}{\varepsilon_1} \in \mathbb{R}^d$. We will have $\|U^*\|_F^2 = O\left(d^2 \log \frac{1}{\varepsilon_1}\right)$. We first decompose $f_t(U^*; x)$ into

$$f_t(U^*, x) = N_{U_t}(u, U^* + \widetilde{U}_t; x) \tag{D.6}$$

$$= N_{U_t}(u, \widetilde{U}_t; x) + N_{U_t}(u, U^*; x) \tag{D.7}$$

For the term $N_{U_t}(u, U^*; x)$, we know that

$$N_{U_t}(u, U^*; x) = N_{W_t}(w, W^*; x) = 20\langle w^\star, x_1 \rangle \sqrt{d} \log \frac{1}{\varepsilon_1} \times \sum_{i=1}^{m/2} w_i^2 \mathbb{1}([W_t]_i x_1) \tag{D.8}$$

$$= 20\langle w^\star, x_1 \rangle \sqrt{d} \log \frac{1}{\varepsilon_1} \times \frac{1}{m} \|\mathbb{1}(W_t x_1)\|_1 \tag{D.9}$$

By Lemma A.8, we know that $\left|\mathbb{1}(W_t x) - \mathbb{1}(\widetilde{W}_t x)\right|_1 \le O(\varepsilon_s m)$ and that $20\langle w^\star, x_1\rangle\sqrt{d}\log\frac{1}{\varepsilon_1} \lesssim \sqrt{d}\log d$, which implies that

$$N_{U_t}(u, U^*; x) = 20\langle w^\star, x_1\rangle\sqrt{d}\log\frac{1}{\varepsilon_1} \times \frac{1}{m}\|\mathbb{1}(\widetilde{W}_t x_1)\|_1 \pm O\left(\sqrt{d}\varepsilon_s \log d\right) \tag{D.10}$$

Note that entries of $\widetilde{W}_t x_1$ are i.i.d. random Bernoulli$(1/2)$, thus we know that w.h.p.

$$\frac{2}{m}\|\mathbb{1}(\widetilde{W}_t x_1)\|_1 = \frac{1}{2} \pm O(m^{-1/2}\sqrt{\log d}) = \frac{1}{2} \pm O(m^{-1/3}) \tag{D.11}$$

Thus, by our choice that $m^{-1/3} = O(\varepsilon_1)$ and $\sqrt{d}\varepsilon_s = O(\varepsilon_1)$,

$$\left|N_{U_t}(u, U^*; x) - 5\langle w^\star, x_1\rangle \log\frac{1}{\varepsilon_1}\right| \le \frac{\varepsilon_1}{20} \tag{D.12}$$

By (D.5), this also implies that

$$\left|N_{U_t}(u, \widetilde{U}_t + U^*; x) - 5\langle w^\star, x_1\rangle \log\frac{1}{\varepsilon_1}\right| \le \frac{\varepsilon_1}{10} \tag{D.13}$$

By definition of $w^\star$, we know that

$$\frac{1}{N}\sum_{i=1}^{N} \ell\left(5\langle w^\star, x_1^{(i)}\rangle \log\frac{1}{\varepsilon_1}; (x^{(i)}, y^{(i)})\right) \le q\log 2 + \varepsilon_1/5 \tag{D.14}$$

Thus, from the fact that ℓ is 1-Lipschitz, it follows that

$$K_t(U^*) \le q\log 2 + \varepsilon_1/2 \tag{D.15}$$

□

Now we wish to argue that even though the noise matrix is changing, gradient descent will still find the fixed target signal matrix $U^\star$. This leverages the fact that once we fix the activation patterns, we can view each step of the optimization as gradient descent with respect to a convex, but changing, function. Below we provide a proof of Theorem B.2, which allows for optimization of this changing function.

Proof of Theorem B.2. For the sake of contradiction, we assume that $K_t(z_t) \ge c^\star + \mu$ for all $t \le T$. Using the definition of K_t^λ, we have that the update rule of z_t can be written as

$$z_{t+1} = z_t - \eta\nabla K_t(z_t) - \eta\lambda z_t \tag{D.16}$$

$$= (1-\eta\lambda)z_t - \eta\nabla K_t(z_t) \tag{D.17}$$

It follows that

$$\|z_{t+1} - z^\star\|_2^2 = \|(1-\eta\lambda)(z_t - z^\star) - \eta(\lambda z^\star + \nabla K_t)\|_2^2 \tag{D.18}$$

$$= \|(1-\eta\lambda)(z_t - z^\star)\|_2^2 + \|\eta(\lambda z^\star + \nabla K_t)\|_2^2 - 2\eta(1-\eta\lambda)\langle \nabla K_t(z_t), z_t - z^\star\rangle$$
$$- 2\eta\lambda(1-\eta\lambda)\langle z_t - z^\star, z^\star\rangle \qquad \text{(expanding)}$$
$$\le \|(1-\eta\lambda)(z_t - z^\star)\|_2^2 + 2\eta^2(\lambda^2 R^2 + L^2) - 2\eta(1-\eta\lambda)(K_t(z_t) - K_t(z^\star))$$
(by convexity of K_t)

$$+ 2\eta\lambda(1-\eta\lambda)\|z_t\|R + 2\eta\lambda(1-\eta\lambda)R^2 \tag{D.19}$$

Assuming that $\|z_t - z^\star\|_2 \le R$, we have that as long as $\lambda R^2 \le \frac{1}{100}\mu$ and $\eta \le \frac{\mu}{100(\lambda^2 R^2 + L^2)}$, we have:

$$\|z_{t+1} - z^\star\|_2^2 \le \|(z_t - z^\star)\|_2^2 + 2\eta^2(\lambda^2 R^2 + L^2) - 2\eta(1-\eta\lambda)\mu + 6\eta\lambda R^2 \tag{D.20}$$

$$\le \|(z_t - z^\star)\|_2^2 - \eta\mu \tag{D.21}$$

Therefore, by induction,

$$\|z_T - z^\star\|_2^2 \le \|(z_0 - z^\star)\|_2^2 - T\eta\mu \le R^2 - T\eta\mu < 0 \tag{D.22}$$

which is a contradiction.

□

D.2 Proof of Lemma 4.2

We define $\tilde{g}_t$ to be the neural network operating on x_2 with activation pattern computed from $\widetilde{V}_t$ and and weights using $\overline{V}_t$:

$$\tilde{g}_t(x) = \tilde{g}_t(x_2) = N_{\widetilde{V}_t}(v, \overline{V}_t; x) \tag{D.23}$$

In the full proof of Lemma 4.2 at the end of the section, we will show that $\tilde{g}_t$ is very close to g_t and therefore we focus on $\tilde{g}_t$ in most parts of the section, and show that it satisfies the almost-linearity condition in Lemma 4.2.

In this section, we will often consider the activation patterns on the inputs $z, z-\zeta, z+\zeta$ at various time steps. For convenience, we have the following definition:

Definition D.1. *For any s, and vector w, let $\mathcal{E}_s^w \triangleq \{i \in [m] : [\widetilde{V}_s]_i w \geq 0\}$ denote the set of neurons that have positive pre-activation on the input w (with weights $\widetilde{V}_s$), and $\bar{\mathcal{E}}_s^w \triangleq \{i \in [m] : [\widetilde{V}_s]_i w < 0\}$ be the set of neurons with negative pre-activations on the input w. (We will mostly be interested in the quantities $\mathcal{E}^{z-\zeta}, \bar{\mathcal{E}}^{z-\zeta}, \mathcal{E}^{z+\zeta}, \bar{\mathcal{E}}^{z+\zeta}$ and their intersections.)*

For a set $\mathcal{E} \subset [m]$, we will use $\mathbb{1}(\mathcal{E}) \in \{0,1\}^m$ to denote the indicator vector for the set $\mathcal{E}$. With this notation, we have that

$$\mathbb{1}(\mathcal{E}_s^x) = \mathbb{1}(\widetilde{V}_s x) \tag{D.24}$$

We start by providing a decomposition of $\tilde{g}_t(z-\zeta) + \tilde{g}_t(z+\zeta) - 2\tilde{g}_t(z)$, and a bound based on how much the activation of $z, z-\zeta, z+\zeta$ differs.

Lemma D.2. *Let $Q_t \triangleq \mathrm{diag}(v)\overline{V}_t$. Then, we have that*

$$\begin{aligned}&\tilde{g}_t(z-\zeta) + \tilde{g}_t(z+\zeta) - 2\tilde{g}_t(z) \\ &= (\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^\top Q_t z + (\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^\top Q_t \zeta\end{aligned} \tag{D.25}$$

Proof. We fix t and drop the subscript of t throughout the proof. Recall the definition of $\tilde{g}_t$ in equation (D.23), we have

$$\begin{aligned}\tilde{g}(x) := N_{\widetilde{V}}(v, \overline{V}; x) &= v^\top \left(\mathbb{1}(\widetilde{V}x) \odot \overline{V}x\right) \\ &= \mathbb{1}(\widetilde{V}x)^\top Qx \qquad \text{(by the definition of } Q = \mathrm{diag}(v)\overline{V})\end{aligned}$$

Therefore,

$$\begin{aligned}\tilde{g}(z-\zeta) + \tilde{g}(z+\zeta) - 2\tilde{g}(z) &= \mathbb{1}(\mathcal{E}^{z-\zeta})^\top Q(z-\zeta) + \mathbb{1}(\mathcal{E}^{z+\zeta})^\top Q(z+\zeta) - 2\mathbb{1}(\mathcal{E}^z)^\top Qz \\ &= (\mathbb{1}(\mathcal{E}^{z-\zeta}) + \mathbb{1}(\mathcal{E}^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}^z))^\top Qz + (\mathbb{1}(\mathcal{E}^{z+\zeta}) - \mathbb{1}(\mathcal{E}^{z-\zeta}))^\top Q\zeta\end{aligned}$$

□

Towards bounding the terms in equation (D.25), we will need to reason about the activations patterns of $z, z-\zeta, z+\zeta$ at various time steps. We first show that the activation patterns of $z-\zeta$ and $z+\zeta$ have to agree in most of neurons except an $\approx r$ fraction of them. This will be useful to show that the second term of the RHS of equation (D.25) is small.

Proposition D.3. *In the setting of Lemma D.2, w.h.p over the randomness of the initialization and all the randomness in the algorithm, for every $t \leq \mathrm{poly}(d), i \in [m]$, $i \in \mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}$ implies that $|[\widetilde{V}_t]_i z| \lesssim \tau_0 r\sqrt{\log d}$. Moreover, the size of the set $\mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}$ is bounded by*

$$|\mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}| \lesssim rm\sqrt{\log d} \tag{D.26}$$

Proof. Recall that $[\widetilde{V}_t]_i \in \mathbb{R}^{1\times d}$ denote the i-th row of the matrix $\widetilde{V}_t$. Recall that $i \in \mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}$ means that $[\widetilde{V}_t]_i(z-\zeta)$ and $[\widetilde{V}_t]_i(z+\zeta)$ have different signs, which in turn implies that

$$|[\widetilde{V}_t]_i z| \leq |[\widetilde{V}_t]_i \zeta| \tag{D.27}$$

Recall that $\|\zeta\|_2 = r$ and by Proposition A.7 $[\widetilde{V}_t]_i$ has distribution $\mathcal{N}(0, \tau_0^2 I_{d\times d})$. Therefore, by standard Gaussian concentration and union bound, with high probability over the randomness of the initialization and the algorithm, for all $t \le \text{poly}(d)$,

$$|[\widetilde{V}_t]_i \zeta| \lesssim \tau_0 \|\zeta\|_2 \sqrt{\log d} = \tau_0 r \sqrt{\log d}\,. \tag{D.28}$$

This proves the first part of the lemma.

Moreover, note that $\Pr\left[|[\widetilde{V}_t]_i z| \le \tau_0 r\sqrt{\log d}\right] \lesssim r\sqrt{\log d}$. By the independence between $[\widetilde{V}_t]_i$'s and standard concentration inequalities (Bernstein inequality), we have that with high probability, there are at most $rm\sqrt{\log d} + \log d$ entries $i \in [m]$ satisfying $|[\widetilde{V}_t]_i z| \le \tau_0 r\sqrt{\log d}$. Together with the first part of the lemma, and that m is sufficiently large so that $rm\sqrt{\log d} + \log d \lesssim rm\sqrt{\log d}$, we complete the proof of equation (D.26). □

We use the lemma above to conclude that the second term in the decomposition (D.25) is at most on the order of r^2/λ.

Proposition D.4. *In the setting of Lemma D.2, we have that*

$$\|(\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^\top Q_t \zeta\|_2 \lesssim \frac{r^2\sqrt{\log d}}{\lambda}\,. \tag{D.29}$$

Proof.

$$|(\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^\top Q_t \zeta| \le \|(\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^\top Q_t\|_2 \|\zeta\|_2 \tag{D.30}$$

By the definition of our algorithm, before annealing the learning rate, we have

$$[Q_t]_i = v_i \cdot [\overline{V}_t]_i = v_i \sum_{s=1}^{t} \eta_1 (1-\eta_1\lambda)^{t-s} [\nabla_V \widehat{L}(U_{s-1})]_i\,. \tag{D.31}$$

Using Proposition A.3 and that $|v_i| = \frac{1}{\sqrt{m}}$, we have that $\|[Q_t]_i\|_2 \lesssim \frac{1}{\lambda m}$. It follows that

$$\|(\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^\top Q_t\|_2 \le |\mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}| \cdot \max_i \|[Q_t]_i\|_2 \lesssim \frac{r\sqrt{\log d}}{\lambda}\,. \tag{D.32}$$

Equation above and equation (D.30) complete the proof. □

Next we will reason about the first term of the RHS of equation (D.25). Note that this is less obvious than the bound for the second term of RHS because both Q and z don't depend on the scale of r, whereas the norm of $\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z})$ only linearly depends on r. However, it is still the case that the first term of RHS of (D.25) scales in r^2 because of the subtle interactions between $\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z})$ and Q_t, as demonstrated in the proofs below.

The following lemma decomposes Q into a sum of the contribution of the gradient from all the previous steps.

Proposition D.5. *In the setting of Lemma D.2, let $\Delta Q_t \triangleq \text{diag}(v)\nabla_V \widehat{L}(U_t)$. ($\Delta Q_t$ can be viewed as the raw change of Q_t at the time step t without considering the effect of the regularizer.) We have that*

$$|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^\top Q_t z| \le \eta_1 \sum_{s=1}^{t} \|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^\top \Delta Q_{s-1}\|_2$$

Proof. Denote $a = \mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z})$ for notational simplicity. By definition of our algorithm, we have

$$a^\top Q_t = a^\top \text{diag}(v) \sum_{s=1}^{t} \eta_1 (1-\eta_1\lambda)^{t-s} \nabla_V \widehat{L}(U_{s-1}) = a^\top \sum_{s=1}^{t} \eta_1 (1-\eta_1\lambda)^{t-s} \Delta Q_{s-1} \tag{D.33}$$

It follows that

$$\|a^\top Q_t\|_2 \le \eta \sum_{s=1}^{t} \|a^\top \Delta Q_{s-1}\|_2\,.$$

Using the fact that $\|z\|_2 \le 1$ we complete the proof. □

In the sequel, we will bound from above the quantity $\|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^\top \Delta Q_{s-1}\|_2$ for every s. One important fact is that the following proposition which shows that ΔQ_s has a lot of repetitive rows that enable additional cancellation in addition to the cancellation in $\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z)$.

Proposition D.6. *Define the analog of $\mathcal{E}_s^w$ with V_t to compute the activation pattern: for any s, and vector w, let $\mathcal{G}_s^w \triangleq \{i \in [m] : [V_s]_i w \geq 0\}$ and define $\bar{\mathcal{G}}_s^w \triangleq \{i \in [m] : [V_s]_i w < 0\}$ similarly.*

Suppose at some iteration s, $z - \zeta$ and $z + \zeta$ have the same activation pattern at neuron i and j in the sense that $i, j \in \mathcal{G}_s^{z-\zeta} \cap \mathcal{G}_s^{z+\zeta}$, or $i, j \in \bar{\mathcal{G}}_s^{z-\zeta} \cap \bar{\mathcal{G}}_s^{z+\zeta}$. Then the corresponding gradient update at that iteration for the weight vectors associated with i and j are the same up to a potential sign flip:

$$[\Delta Q_s]_i = v_i[\nabla_V \widehat{L}(U_s)]_i = v_j[\nabla_V \widehat{L}(U_s)]_j = [\Delta Q_s]_j \tag{D.34}$$

Moreover, suppose we have that i, j satisfy that $[\widetilde{V}_s]_i x \gtrsim \tau_0 r\sqrt{\log d}$ and $[\widetilde{V}_s]_j x \gtrsim \tau_0 r\sqrt{\log d}$ (or $[\widetilde{V}_s]_i x \lesssim -\tau_0 r\sqrt{\log d}$ and $[\widetilde{V}_s]_j x \lesssim -\tau_0 r\sqrt{\log d}$) for $x \in \{z - \zeta, z + \zeta\}$, then the same conclusion holds for i and j.

Proof. Note that by definition, $[\Delta Q_s]_i = v_i[\nabla_V \widehat{L}(U_s)]_i$, and thus it suffices to prove that $v_i[\nabla_V \widehat{L}(U_s)]_i = v_j[\nabla_V \widehat{L}(U_s)]_j$. By Proposition A.1, we have that

$$[\nabla_V \widehat{L}(U_s)]_i = \widehat{\mathbb{E}}\left[\ell'(f(u, U_s; (x, y)))v_i \mathbb{1}([V_s]_i x_2)x_2\right] \tag{D.35}$$

Note that x_2 can only take (a positive scaling of) four values $z - \zeta, z, z + \zeta, 0$. We claim that for every choice of these four values, for the i, j satisfying the condition of the lemma, we have

$$\ell'(f(u, U_s; (x, y)))\mathbb{1}([V_s]_i x_2)x_2 = \ell'(f(u, U_s; (x, y)))\mathbb{1}([V_s]_j x_2)x_2 \tag{D.36}$$

Note that the equation above together with $v_i^2 = v_j^2 = 1$ suffices to complete the proof.

Equation (D.36) is true for $x_2 = 0$. Suppose without loss of generality, $i, j \in \mathcal{G}_s^{z-\zeta} \cap \mathcal{G}_s^{z+\zeta}$. Then we know that $i, j \in \mathcal{G}_s^z$ because $[V_s]_i(z - \zeta) + [V_s]_i(z + \zeta) = 2[V_s]_i z$. Therefore $\mathbb{1}([V_s]_i x_2) = \mathbb{1}([V_s]_j x_2) = 1$ for all $x_2 \in \{z - \zeta, z, z + \zeta\}$. Thus we proved equation (D.36) and complete the proof of the first part of the lemma.

Now to prove the second part of the lemma, suppose i, j satisfy that $[\widetilde{V}_s]_i x \gtrsim \tau_0 r\sqrt{\log d}$ and $[\widetilde{V}_s]_j x \gtrsim \tau_0 r\sqrt{\log d}$ for $x \in \{z - \zeta, z + \zeta\}$. Using $\|[\widetilde{V}_s]_i\|_2 \leq \frac{1}{\lambda\sqrt{m}}$ from Proposition A.3, we have that $[V_s]_i z \geq [\widetilde{V}_s]_i z - |[\overline{V}_s]_i z| \gtrsim \tau_0 r\sqrt{\log d} - O(\frac{1}{\lambda\sqrt{m}}) \geq \tau_0 r\sqrt{\log d}$ where used the assumption that $1/\lambda = \text{poly}(d)$ and $m = \text{poly}(d/\tau_0)$. Therefore, we conclude that $i, j \in \mathcal{G}_s^{z-\zeta} \cap \mathcal{G}_s^{z+\zeta}$. Now by the first lemma of the lemma we complete the proof. □

Now we are ready to bound the first term on the RHS of equation D.25, which is the crux of the proofs in this section. The key here is to get a bound that scales quadratically in r.

Proposition D.7. *In the setting of Lemma D.2, let ΔQ_s be defined in Proposition D.5. Then, we have that*

$$\|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^\top \Delta Q_s\|_2 \lesssim \frac{r^2\sqrt{\log d}}{\sqrt{\lambda\eta_1(s-t)}} \tag{D.37}$$

As a direct corollary of the equation above and Proposition D.5, we have that

$$|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^\top Q_t z| \lesssim \frac{r^2\sqrt{\log d}}{\lambda} \tag{D.38}$$

Proof. By the set operations and the facts that $\mathcal{E}_t^{z-\zeta} \cap \mathcal{E}_t^{z+\zeta} \subset \mathcal{E}_t^z$ and that $\mathcal{E}_t^z \subset \mathcal{E}_t^{z-\zeta} \cup \mathcal{E}_t^{z+\zeta}$, we have that

$$\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z) = \left(\mathbb{1}(\mathcal{E}_t^{z-\zeta} \backslash \mathcal{E}_t^z) - \mathbb{1}(\mathcal{E}^z \backslash \mathcal{E}_t^{z+\zeta})\right) + \left(\mathbb{1}(\mathcal{E}_t^{z+\zeta} \backslash \mathcal{E}_t^z) - \mathbb{1}(\mathcal{E}_t^z \backslash \mathcal{E}_t^{z-\zeta})\right) \tag{D.39}$$

Define

$$\begin{aligned}\mathcal{F}_s^+ &= \{i \in [m] : [\widetilde{V}_s]_i z \gtrsim \tau_0 r\sqrt{\log d}\}\\ \mathcal{F}_s^- &= \{i \in [m] : [\widetilde{V}_s]_i z \lesssim -\tau_0 r\sqrt{\log d}\}\\ \mathcal{F}_s^c &= \{i \in [m] : |[\widetilde{V}_s]_i z| \lesssim \tau_0 r\sqrt{\log d}\}\end{aligned} \tag{D.40}$$

where the $\lesssim, \gtrsim$ notations hide universal constants that make the first conclusion of Proposition D.3 true. By the second part of Proposition D.3 (or more directly equation (D.28)), we have that $\mathcal{F}_s^+ \subset \mathcal{E}_s^{z-\zeta} \cap \mathcal{E}_s^{z+\zeta}$, and $\mathcal{F}_s^- \subset \mathcal{E}_s^{z-\zeta} \cap \bar{\mathcal{E}}_s^{z+\zeta}$. By Proposition D.6, we have that for any $i, j \in \mathcal{F}_s^-$, $[\Delta Q_s]_i = [\Delta Q_s]_j$. For notational simplicity, let $A = \mathcal{E}_t^{z+\zeta} \backslash \mathcal{E}_t^z$ and $B = \mathcal{E}_t^z \backslash \mathcal{E}_t^{z-\zeta}$. Therefore it follows that

$$\begin{aligned}&\left\| \left(\mathbb{1}(\mathcal{E}_t^{z+\zeta} \backslash \mathcal{E}_t^z) - \mathbb{1}(\mathcal{E}_t^z \backslash \mathcal{E}_t^{z-\zeta}) \right)^\top \Delta Q_s \right\|_2 = \left\| \sum_{i\in A} [\Delta Q_s]_i - \sum_{i \in B} [\Delta Q_s]_i \right\|_2\\ &= \left\| \sum_{i\in A\cap \mathcal{F}_s^+} [\Delta Q_s]_i - \sum_{i\in B\cap \mathcal{F}_s^+} [\Delta Q_s]_i \right\|_2 + \left\| \sum_{i\in A\cap \mathcal{F}_s^-} [\Delta Q_s]_i - \sum_{i\in B\cap \mathcal{F}_s^-} [\Delta Q_s]_i \right\|_2\\ &+ \left\| \sum_{i\in A\cap \mathcal{F}_s^c} [\Delta Q_s]_i - \sum_{i\in B\cap \mathcal{F}_s^c} [\Delta Q_s]_i \right\|_2\\ &\le \frac{1}{m}\left(\big||A\cap \mathcal{F}_s^+| - |B\cap \mathcal{F}_s^+|\big| + \big||A\cap \mathcal{F}_s^-| - |B\cap \mathcal{F}_s^-|\big| + |A\cap \mathcal{F}_s^c| + |B\cap \mathcal{F}_s^c| \right)\end{aligned} \tag{D.41}$$

where in the last inequality we use that for any $i, j \in \mathcal{F}_s^-$, $[\Delta Q_s]_i = [\Delta Q_s]_j$, and the fact that $\|[\Delta Q_s]_i\|_2 = \frac{1}{\sqrt{m}}\|[\nabla_V \widehat{L}(U_s)]_i\|_2 \le 1/m$ (by Proposition A.2.)

Next, we first bound

$$|A\cap \mathcal{F}_s^+| - |B\cap \mathcal{F}_s^+| = \sum_{i\in[m]} \mathbf{1}(i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^z, i \in \mathcal{F}_s^+) - \mathbf{1}(i \in \mathcal{E}_t^z, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+). \tag{D.42}$$

Note that the distribution of $([\widetilde{V}_s]_i, [\widetilde{V}_t]_i$'s are independent across the choice of i. Thus we will compute $\Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^z, i \in \mathcal{F}_s^+] - \Pr[i \in \mathcal{E}_t^z, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+]$ and then apply concentration concentration inequality for the sum. Note that the event here depends on three quantities $[\widetilde{V}_s]_i z$, $[\widetilde{V}_t]_i z$, and $[\widetilde{V}_t]_i \zeta$. First of all, $[\widetilde{V}_t]_i \zeta$ is independent of these other two because ζ is orthogonal to z and $[\widetilde{V}_t]_i$ and $[\widetilde{V}_s]_i$ have spherical covariance matrices.

By the definition of $\widetilde{V}_s, \widetilde{V}_t$, we can express their relationship by writing $[\widetilde{V}_t]_i z = (1-\eta_1\lambda)^{t-s}[\widetilde{V}_s]_i z + [\Xi_{t,s}]_i z$, where $\Xi_{t,s} = \eta_1 \sum_{j\in[t-s]}(1-\eta_1\lambda)^{t-s-j}\xi_{s+j}$. Recall that by proposition A.7, we have $[\widetilde{V}_s]_i z \sim \mathcal{N}(0, \tau_0^2)$ and $[\Xi_{t,s}]_i z$ are two independent Gaussians. Let $\sigma_{t,s}$ be the variance of $[\Xi_{t,s}]_i z$. We compute $\sigma_{t,s}$ by observing that

$$\tau_0^2 = \mathsf{Var}([\widetilde{V}_t]_i z) = \mathsf{Var}((1-\eta_1\lambda)^{t-s}[\widetilde{V}_s]_i z) + \mathsf{Var}([\Xi_{t,s}]_i z) = (1-\eta_1\lambda)^{2(t-s)}\tau_0^2 + \sigma_{s,t}^2$$

Solving the equation we obtain that

$$\sigma_{s,t} = \sqrt{\tau_0^2(1-(1-\eta_1\lambda)^{2(t-s)})} \ge \tau_0\sqrt{\lambda\eta_1(s-t)} \tag{D.43}$$

Note that $\zeta^\top z = 0$, thus $[\widetilde{V}_s]_i z$ is independent of $[\widetilde{V}_t]_i \zeta$ conditioned on $[\widetilde{V}_t]_i z$, for every $s \le t$. For notational simplicity, let $Y_1 = [\widetilde{V}_s]_i z$, $Y_2 = [\widetilde{V}_t]_i z$, and $Y_3 = [\widetilde{V}_t]_i \zeta$, and $\kappa = O(\tau_0 r\sqrt{\log d})$ where the big O notation hide the same constant factor used in defining $\mathcal{F}_s^+$ in equation (D.40). Let $Y_4 = [\Xi_{t,s}]_i z = Y_1 - \beta Y_2$ where $\beta = \eta_1(1-\eta_1\lambda)^{t-s} \gtrsim 1$ (because $t \le 1/(\eta_1\lambda)$). Note that by the calculation above, Y_4 has standard deviation $\sigma_{s,t}$ which is bounded from below by $\tau_0\sqrt{\lambda\eta_1(s-t)}$.

Then, we have that

$$\begin{aligned}
\Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^z, i \in \mathcal{F}_s^+] &= \Pr\left[Y_2 + Y_3 \geq 0, Y_2 \leq 0, Y_1 \geq \kappa\right] && \text{(D.44)}\\
&= \Pr\left[Y_2 + Y_3 \geq 0, Y_2 \leq 0, Y_4 \geq \kappa - \beta Y_2\right] && \text{(D.45)}\\
&= \mathop{\mathbb{E}}_{Y_2}\left[\Pr\left[Y_2 + Y_3 \geq 0, Y_2 \leq 0, Y_4 \geq \kappa - \beta Y_2 \mid Y_2\right]\right]\\
&\qquad \text{(by the law of total expecation)}\\
&= \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0)\Pr\left[Y_3 \geq -Y_2 \mid Y_2\right] \cdot \Pr\left[Y_4 \geq \kappa - \beta Y_2 \mid Y_2\right]\right]\\
&\qquad \text{(because } Y_1, Y_3, Y_4 \text{ are independent conditioned on } Y_2.\text{)}
\end{aligned}$$

Similarly, we have that

$$\begin{aligned}
\Pr[i \in \mathcal{E}_t^z, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+] &= \Pr\left[Y_2 \geq 0, Y_2 - Y_3 \leq 0, Y_1 \geq \kappa\right]\\
&= \Pr\left[-Y_2 \geq 0, -Y_2 - Y_3 \leq 0, -Y_1 \geq \kappa\right]\\
&\qquad ((Y_1, Y_2, Y_3) \text{ has the same distribution as } (-Y_1, -Y_2, Y_3)))\\
&= \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0)\Pr\left[Y_3 \geq -Y_2 \mid Y_2\right] \cdot \Pr\left[Y_4 \leq -\kappa - \beta Y_2 \mid Y_2\right]\right]\\
&\qquad \text{(because } Y_1, Y_3, Y_4 \text{ are independent conditioned on } Y_2.\text{)}\\
&= \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0)\Pr\left[Y_3 \geq -Y_2 \mid Y_2\right] \cdot \Pr\left[Y_4 \geq \kappa + \beta Y_2 \mid Y_2\right]\right]\\
&\qquad \text{(because } (Y_4, Y_2) \text{ has the same distribution as } (-Y_4, Y_2).\text{)}
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\left|\Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^z, i \in \mathcal{F}_s^+] - \Pr[i \in \mathcal{E}_t^z, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+]\right| && \text{(D.46)}\\
&= \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0)\Pr\left[Y_3 \geq -Y_2 \mid Y_2\right]\Pr\left[\kappa - \beta Y_2 \leq Y_4 \leq \kappa + \beta Y_2 \mid Y_2\right]\right] && \text{(D.47)}\\
&\lesssim \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0)\Pr\left[Y_3 \geq -Y_2 \mid Y_2\right]\frac{|Y_2|}{\sigma_{s,t}}\right] \quad \text{(because the density of } Y_4 \text{ is bounded by } O(1/\sigma_{s,t}))\\
&\lesssim \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0)\exp(-|Y_2|^2/2(r^2\tau_0^2))\frac{|Y_2|}{\sigma_{s,t}}\right] \quad \text{(because } Y_3 \text{ has variance } r^2\tau_0^2\text{)}\\
&\lesssim \int_{-\infty}^{0} 1/\tau_0 \cdot \exp(-z^2/(2r^2\tau_0^2))\exp(-z^2/\tau_0^2)|z|/\sigma_{s,t}dz \lesssim r^2\tau_0/\sigma_{s,t}\\
&\lesssim \frac{r^2}{\sqrt{\lambda\eta_1(s-t)}} && \text{(D.48)}
\end{aligned}$$

Now by equation (D.42) and standard concentration inequality, and the fact that m is sufficiently large, we have that with high probability,

$$\left||A \cap \mathcal{F}_s^+| - |B \cap \mathcal{F}_s^+|\right| \lesssim \frac{r^2 m}{\sqrt{\lambda\eta_1(s-t)}} + \log d \lesssim \frac{r^2 m}{\sqrt{\lambda\eta_1(s-t)}} \tag{D.49}$$

Similarly, we can prove that

$$\left||A \cap \mathcal{F}_s^-| - |B \cap \mathcal{F}_s^-|\right| \lesssim \frac{r^2 m}{\sqrt{\lambda\eta_1(s-t)}} \tag{D.50}$$

Finally, we have that

$$\begin{aligned}
\Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^z, i \in \mathcal{F}_s^c] &= \Pr\left[Y_2 + Y_3 \geq 0, Y_2 \leq 0, |Y_1| \leq \kappa\right] && \text{(D.51)}\\
&= \mathbb{E}\left[\Pr\left[Y_2 + Y_3 \geq 0, Y_2 \leq 0, |Y_4 - \beta Y_2| \leq \kappa\right]\right] \\
&\qquad \text{(by the law of total expecation)} \\
&= \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0) \Pr\left[Y_3 \geq -Y_2 \mid Y_2\right] \cdot \kappa/\sigma_{s,t}\right] \\
&\qquad \text{(because the density of } Y_4 \text{ is bounded by } O(1/\sigma_{s,t})) \\
&\lesssim \mathop{\mathbb{E}}_{Y_2}\left[\mathbf{1}(Y_2 \leq 0) \exp(-|Y_2|^2/2(r^2\tau_0^2)) \frac{\kappa}{\sigma_{s,t}}\right] \\
&\qquad \text{(because } Y_3 \text{ has variance } r^2\tau_0^2) \\
&\lesssim \kappa r \tau_0/\sigma_{s,t} \lesssim \frac{r^2\sqrt{\log d}}{\sqrt{\lambda\eta_1(s-t)}} && \text{(D.52)}
\end{aligned}$$

Using standard concentration inequality and the fact that m is sufficiently large, we have that with high probability,

$$|A \cap \mathcal{F}_s^c| \lesssim \frac{r^2 m\sqrt{\log d}}{\sqrt{\lambda\eta_1(s-t)}} + \log d \lesssim \frac{r^2 m\sqrt{\log d}}{\sqrt{\lambda\eta_1(s-t)}} \tag{D.53}$$

We can also prove the same bound for $|B \cap \mathcal{F}_s^c|$ analogously. Using equation (D.41) and the several equations above, we conclude that

$$\left\| \left(\mathbb{1}(\mathcal{E}_t^{z+\zeta} \backslash \mathcal{E}_t^z) - \mathbb{1}(\mathcal{E}_t^z \backslash \mathcal{E}_t^{z-\zeta})\right)^\top \Delta Q_s \right\|_2 \lesssim \frac{r^2\sqrt{\log d}}{\sqrt{\lambda\eta_1(s-t)}} \tag{D.54}$$

Thus equation (D.37) follows from equation (D.39) and proving a bound for $\left(\mathbb{1}(\mathcal{E}_t^{z+\zeta} \backslash \mathcal{E}_t^z) - \mathbb{1}(\mathcal{E}_t^z \backslash \mathcal{E}_t^{z-\zeta})\right)^\top \Delta Q_s$ similarly to the equation above. To prove equation (D.38), we use Proposition D.5, and equation (D.37) to obtain that

$$\begin{aligned}
|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^\top Q_t z| &\leq \eta_1 \sum_{s=1}^t \|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^\top \Delta Q_{s-1}\|_2 \\
&\lesssim \eta_1 \sum_{s=1}^t \frac{r^2\sqrt{\log d}}{\sqrt{\lambda\eta_1(s-t)}} \lesssim r^2\sqrt{\log d}\sqrt{t\eta_1/\lambda} && \text{(D.55)}\\
&\lesssim r^2\sqrt{\log d}/\lambda && \text{(D.56)}
\end{aligned}$$

where the last step uses that the condition that $t \leq 1/(\eta_1\lambda)$.

□

Now combining the Propositions above we are ready to prove Lemma 4.2.

Proof of Lemma 4.2. Using triangle inequality, Proposition A.8, and equation (A.6) of Proposition A.5, we have that for any x of norm $O(1)$,

$$\begin{aligned}
|g_t(x) - \tilde{g}_t(x)| &\leq |N_{V_t}(v, \overline{V}_t; x) - N_{\widetilde{V}_t}(v, \overline{V}_t; x)| + |N_{V_t}(v, \widetilde{V}_t; x)| && \text{(D.57)}\\
&\leq \|\overline{V}_t\|_F \tau_0^{-2} m^{-1/6} + \|\overline{V}_t\|_F^{5/3} \tau_0^{-2/3} m^{-1/6} + \tau_0 \log d \\
&\qquad \text{(by Proposition A.8, and equation (A.6) of Proposition A.5)} \\
&\leq 1/\text{poly}(d) \\
&\text{(because } \tau_0 = 1/\text{poly}\left(\tfrac{d}{\varepsilon}\right) \text{ and } m \geq \text{poly}\left(\tfrac{d}{\varepsilon\tau_0}\right) \text{ and } \|\overline{V}_t\| \lesssim 1/\lambda \text{ by Proposition A.3.)}
\end{aligned}$$

Thus we can only focus on $\tilde{g}_t$. Using Lemma D.2, we have that

$$\begin{aligned}
&|\tilde{g}_t(z-\zeta) + \tilde{g}_t(z+\zeta) - 2\tilde{g}_t(z)| \\
&\leq |(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^\top Q_t z| + |(\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^\top Q_t \zeta| && \text{(D.58)}\\
&\lesssim \frac{r^2\sqrt{\log d}}{\lambda} + \frac{r^2\sqrt{\log d}}{\lambda} && \text{(by equation (D.38) of Proposition D.7 and Proposition D.4)}
\end{aligned}$$

which completes the proof. □

D.3 Proof of Lemma B.3

The proof of Lemma B.3 relies on the fact that a smaller learning rate preserves the noise generated from the timestep before annealing. This allows us to reason that the new activations are similar to the original before reducing the learning rate.

Proof of Lemma B.3. By definition, we have that

$$\begin{aligned}[U_{t_0}]_i &= [\overline{U}_{t_0}]_i + [\widetilde{U}_{t_0}]_i \\ [U_{t_0+t}]_i &= [\overline{U}_{t_0+t}]_i + [\widetilde{U}_{t_0+t}]_i = [\overline{U}_{t_0+t}]_i + (1-\eta_2\lambda)^t[\widetilde{U}_{t_0}]_i + [\Xi_t]_i \end{aligned} \tag{D.59}$$

where $\Xi_t := \eta_2 \sum_{j\le t}(1-\lambda\eta_2)^{t-j}\xi_{t_0+j}$.
By properties of a sum of Independent Gaussians, we have $[\Xi_t]_i \sim \mathcal{N}(0, \sigma_t^2 I)$ where σ_t is the standard deviation of each entry of Ξ_t. We also have that Ξ_t is independent of $\widetilde{U}_{t_0}$. Moreover, for every $t \le \frac{1}{\eta_2\lambda}$, the standard deviation σ_t can be bounded by

$$\begin{aligned}\sigma_t^2 &= \eta_2^2 \sum_{j\le t}(1-\lambda\eta_2)^{2(t-j)}\tau_\xi^2 \le \eta_2^2\tau_\xi^2 t \\ &= \frac{\eta_2^2(\tau_0^2 - (1-\eta_1\lambda)^2\tau_0^2)}{\eta_1^2}t \le \frac{2\eta_2^2\lambda\tau_0^2 t}{\eta_1} \le \frac{2\eta_2\tau_0^2}{\eta_1}\end{aligned} \tag{D.60}$$

(Note that since $\eta_2 \ll \eta_1$, we should expect that the standard deviations satisfy $\sigma_t \ll \sigma_0$. That is, the additional randomness introduced in the pre-activation is small.)

On the other hand, for every $t \le \frac{1}{\eta_2\lambda}$, the contribution of $\widetilde{U}_{t_0}$ to U_{t+t_0} is still present because the entry of $(1-\eta_2\lambda)^t[\widetilde{U}_{t_0}]_i$ has variance at least on the order of the variance of the entries of $[\widetilde{U}_{t_0}]_i$, which is $\gtrsim \tau_0^2$. This also implies that the variance of the entries of $\widetilde{U}_{t_0+t}$ is lower bounded by the variance of $(1-\eta_2\lambda)^t[\widetilde{U}_{t_0}]_i$. This in turn is lower bounded by τ_0^2 up to constant factor.

Therefore, using the decomposition (D.59) and the bounds above, we should expect that the sign of U_{t_0+t} strongly correlates with the the sign of U_{t_0}, which will be formally shown below. Using Lemma A.8, we have that the activation pattern is mostly decided by the noise part ($\widetilde{U}_{t+t_0}$ and $\widetilde{U}_{t_0}$), in the sense that for every x,

$$\|\mathbb{1}(U_{t_0}x) - \mathbb{1}(\widetilde{U}_{t_0}x)\|_1 \lesssim \|\overline{U}_{t_0}\|_F^{4/3}\tau_0^{-4/3}m^{2/3} \le \varepsilon_s m \tag{D.61}$$

This can obtained by setting $\tilde{U} = \widetilde{U}_{t_0}, \overline{U} = \overline{U}_{t_0}, \tau = \tau_0$ in Lemma A.8, and using $\|\overline{U}_{t_0}\|_F \le 1/\lambda$ from Proposition A.3. Similarly, setting $\tilde{U} = \widetilde{U}_{t_0+t}, \overline{U} = \overline{U}_{t_0+t}$, and letting τ be the standard deviation of entries of $\widetilde{U}_{t_0+t}$ (which has been shown to be $\gtrsim \tau_0$), we get

$$\|\mathbb{1}(U_{t_0+t}x) - \mathbb{1}(\widetilde{U}_{t_0+t}x)\|_1 \lesssim \|\overline{U}_{t_0+t}\|_F^{4/3}\tau^{-4/3}m^{2/3} \le \varepsilon_s m \tag{D.62}$$

Fixing x, we can decompose our target to

$$\|\mathbb{1}(U_{t_0+t}x) - \mathbb{1}(U_{t_0}x)\|_1 \le \tag{D.63}$$

$$\|\mathbb{1}(U_{t_0+t}x) - \mathbb{1}(\widetilde{U}_{t_0+t}x)\|_1 + \|\mathbb{1}(\widetilde{U}_{t_0+t}x) - \mathbb{1}(\widetilde{U}_{t_0}x)\|_1 + \|\mathbb{1}(\widetilde{U}_{t_0}x) - \mathbb{1}(U_{t_0}x)\|_1 \tag{D.64}$$

We've bounded the first and third term on the RHS of the equation above. For the middle term, let $\alpha_i = (1-\eta_2\lambda)^t[\widetilde{U}_{t_0}]_i x$ and $\beta_i = [\Xi_{t+t_0}]_i x$. Note that $[\widetilde{U}_{t+t_0}]_i x = \alpha_i + \beta_i$ and that α_i and β_i are zero-mean independent Gaussian random variables with variance $\gtrsim \tau_0^2\|x\|^2$ and variance $\lesssim \eta_2\tau_0^2\|x\|^2/\eta_1$, respectively. The basic property of Gaussian random variable implies that

$$\Pr\left[\mathbb{1}(\alpha_i + \beta_i) \ne \mathbb{1}(\beta_i)\right] \lesssim \sqrt{\frac{\eta_2\tau_0^2\|x\|^2/\eta_1}{\tau_0^2\|x\|^2}} = \sqrt{\eta_2/\eta_1} \tag{D.65}$$

Since α_i, β_i's are independent, by basic concentration inequality (e.g., Bernstein inequality or Hoeffding inequality), we have that with high probability

$$\|\mathbb{1}(\widetilde{U}_t x) - \mathbb{1}(\widetilde{U}_{t_0}x)\|_1 \lesssim \sqrt{\eta_2/\eta_1}m + \sqrt{m\log d} \lesssim \sqrt{\eta_2/\eta_1}m + m^{2/3} \tag{D.66}$$

Combining the equation above with equation (D.61), (D.62),and (D.64) completes the proof for the first part.

For the second part, we can bound

$$\left|N_{U_{t_0+t}}(u, U_{t_0+t}; x) - N_{U_{t_0}}(u, \overline{U}_{t_0+t}; x)\right| \tag{D.67}$$

$$\leq \left|N_{U_{t_0+t}}(u, U_{t_0+t}; x) - N_{U_{t_0+t}}(u, \overline{U}_{t_0+t}; x)\right| + \left|N_{U_{t_0+t}}(u, \overline{U}_{t_0+t}; x) - N_{U_{t_0}}(u, \overline{U}_{t_0+t}; x)\right| \tag{D.68}$$

$$\lesssim \left|N_{U_{t_0+t}}(u, \widetilde{U}_{t_0+t}; x)\right| \tag{D.69}$$

$$+ \frac{1}{\sqrt{m}}\|\mathbb{1}([U_{t_0+t}]x) - \mathbb{1}([U_{t_0}]x)\|_1 \max_i \|[\overline{U}_{t_0+t}]_i\|_2 \tag{D.70}$$

$$\lesssim \left(\sqrt{\frac{\eta_2}{\eta_1}} + \varepsilon_s\right) \times \frac{1}{\lambda} + \tau_0 \log d \tag{D.71}$$

where the last inequality is due to $\max_i \|[\overline{U}_{t_0+t}]_i\|_2 = O(1/\sqrt{m}\lambda)$ by Proposition A.3, and bounding $\left|N_{U_{t_0+t}}(u, \widetilde{U}_{t_0+t}; x)\right| \lesssim \frac{\varepsilon_s}{\lambda} + \tau_0 \log d$ by Proposition A.5.

□

We note that this lemma also applies to the setting when $t_0 = 0$, i.e. we start with an initial small learning rate and compare to the random initialization. This is useful for the proofs in the small initial learning rate setting.

D.4 Proof of Lemma 4.3

We will now show that the network learns patterns from $\mathcal{Q}$ once the learning rate is annealed by constructing a common target for the network at every subsequent time step. We will then use Theorem B.2 to show that the optimization finds this target. Let us define

$$\varepsilon_0 := \frac{1}{N}\sum_{i \in \mathcal{M}_1} \ell(r_{t_0}; (x^{(i)}, y^{(i)})) \tag{D.72}$$

Formally, we first show the following proposition, which proves the existence of a target solution that has good accuracy on $\bar{\mathcal{M}}_1$ and does not unlearn the network's progress on $\mathcal{M}_1$:

Lemma D.8. *In the setting of Lemma 4.3, let $K_t(B)$ be defined in equation (B.2). Then, there exists a solution U^* satisfying $\|U^*\|_F^2 = \widetilde{O}\left(\frac{1}{\varepsilon_1^2 r}\right)$ and*

$$K_{t_0+t}(\overline{U}_{t_0} + U^*) \leq \varepsilon_0 + \varepsilon_1 \tag{D.73}$$

To prove this proposition, we need the following lemma:

Proposition D.9. *Suppose g_t satisfies that $|g_t(z+\zeta) + g_t(z-\zeta) - 2g_t(z)| \leq \delta$ for some $\delta \lesssim 1$. Then, we have that*

$$\widehat{L}_{\bar{\mathcal{M}}_1}(u, U) \geq \log 2 - O(\delta) - O(\log d/\sqrt{qN}) \tag{D.74}$$

And moreover, if $\widehat{L}_{\bar{\mathcal{M}}_1}(u, U) \leq \log 2 + O(\delta')$ for some $\delta' \geq \delta$, then the prediction of g_t on $z-\zeta, z, z+\zeta$ satisfies $|g_t(z-\zeta)|, |g_t(z+\zeta)|, |g_t(z)| = O(\sqrt{\delta' + \log d/\sqrt{qN}})$.

Proof. For convenience, let us denote $g_t(z+\delta) = u, g_t(z-\delta) = v, g_t(z) = (u+v)/2 + \gamma$. By our assumption, we have that $|\gamma| \leq \delta$.

Let $h(z) := -\log \frac{1}{1+e^{-z}}$. We have that w.h.p, for $c = O(\log d/\sqrt{qN})$,

$$4L_{\bar{\mathcal{M}}_1}(u, U) \geq [h(-u) + h(-v) + 2h((u+v)/2 + \gamma)] \cdot (1-c) \tag{D.75}$$

$$= [\Delta + 2h(-(u+v)/2) + 2h((u+v)/2 + \gamma)] \cdot (1-c) \tag{D.76}$$

where Δ is defined as

$$\Delta = h(-u) + h(-v) - 2h(-(u+v)/2) \geq 0 \qquad \text{(by convexity of } h\text{)}$$

and the factor of $1-c$ comes from the fact that the fraction of examples that are $z-\zeta, z+\zeta, z$ will be $1/4 \pm O(\log d/\sqrt{qN})$, $1/4 \pm O(\log d/\sqrt{qN})$, $1/2 \pm O(\log d/\sqrt{qN})$, respectively, w.h.p. Since the function $h(z)$ is a 2-Lip function, we know that

$$|h((u+v)/2+\gamma) - h((u+v)/2)| \le 2\gamma \tag{D.77}$$

It follows that

$$\begin{aligned}
4L_{\bar{\mathcal{M}}_1}(u,U) &\ge (\Delta + 2h(-(u+v)/2) + 2h((u+v)/2+\gamma))(1-c) \\
&\ge (2h(-(u+v)/2) + 2h((u+v)/2) - 4\gamma)(1-c) \\
&\qquad \text{(because } \Delta \ge 0 \text{ and equation (D.77))} \\
&\ge 4\log 2 - 4\gamma - O(\log d/\sqrt{qN}) \qquad \text{(by convexity of } h\text{)} \\
&\ge 4\log 2 - O(\delta) - O(\log d/\sqrt{qN})
\end{aligned}$$

The equation above together with the assumption $\widehat{L}_{\bar{\mathcal{M}}_1}(u,U) \le \log 2 + O(\delta')$ implies that

$$4\log 2 + O(\delta') \ge 4L_{\bar{\mathcal{M}}_1}(u,U) \ge (\Delta + 2h((u+v)/2) + 2h(-(u+v)/2) - O(\delta))(1-c) \tag{D.78}$$

which implies that $h((u+v)/2) + h(-(u+v)/2) - 2h(0) + \Delta \le O(\delta') + O(c)$. It follows that $h((u+v)/2) + h(-(u+v)/2) - 2h(0) \le O(\delta') + O(c)$ and $\Delta \le O(\delta') + O(c)$. Now we note that By the strict convexity of $h(z)$, we
can easily conclude that $|u|, |v| \le O(\sqrt{\delta' + c})$. □

Next, we will bound ε_0 and the value of g_{t_0}. This allows us to conclude that g_{t_0} is small, so that it is easy to "unlearn" once the learning rate is annealed.

Lemma D.10. *Suppose the condition in Lemma 4.1 holds. Then*

$$|g_{t_0}(z)|, |g_{t_0}(z+\zeta)|, |g_{t_0}(z-\zeta)| \le O(\sqrt{\varepsilon_1/q}) \tag{D.79}$$

$$\varepsilon_0 = O(\sqrt{\varepsilon_1/q}) \tag{D.80}$$

Proof of Lemma D.10. Since $L_{t_0} \le q\log 2 + \varepsilon_1$, we know that $\widehat{L}_{\bar{\mathcal{M}}_1}(u, U_{t_0}) \le \log 2 + 2\varepsilon_1/q$. Applying Proposition D.9 with $\delta' = \varepsilon_1$ and $\delta = O(r^2/\lambda) = O(\varepsilon_1)$, we have that $|g_{t_0}(z)|, |g_{t_0}(z+\zeta)|, |g_{t_0}(z-\zeta)| \le O(\sqrt{\varepsilon_1/q})$ and $\widehat{L}_{\bar{\mathcal{M}}_1}(u, U_{t_0}) \ge \log 2 - \varepsilon_1$.

Hence we have that (since ℓ is 2-Lipschitz)

$$\begin{aligned}
\varepsilon_0 &= \frac{1}{N}\sum_{i\in\mathcal{M}_1} \ell(r_{t_0}; (x^{(i)}, y^{(i)})) && \text{(D.81)} \\
&\le \frac{1}{N}\sum_{i\in\mathcal{M}_1} \ell(r_{t_0} + g_{t_0}; (x^{(i)}, y^{(i)})) + \frac{2}{N}\sum_{i\in\mathcal{M}_1} |g_{t_0}(x^{(i)})_2| && \text{(D.82)} \\
&\le \left(L_{t_0} - q\widehat{L}_{\bar{\mathcal{M}}_1}(u, U_{t_0})\right) + O(\sqrt{\varepsilon_1/q}) && \text{(D.83)} \\
&\le O(\sqrt{\varepsilon_1/q}) && \text{(D.84)}
\end{aligned}$$

□

Now we will complete the proof of Proposition D.8.

Proof of Proposition D.8. Let us define sets $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ as the following:

$$\mathcal{E}_1 = \{i \in [m] \mid \langle [V_{t_0}]_i, z-\zeta\rangle \ge 0, \langle [V_{t_0}]_i, z\rangle \ge 0, \langle [V_{t_0}]_i, z+\zeta\rangle < 0\} \tag{D.85}$$

$$\mathcal{E}_2 = \{i \in [m] \mid \langle [V_{t_0}]_i, z-\zeta\rangle \ge 0, \langle [V_{t_0}]_i, z\rangle < 0, \langle [V_{t_0}]_i, z+\zeta\rangle < 0\} \tag{D.86}$$

$$\mathcal{E}_3 = \{i \in [m] \mid \langle [V_{t_0}]_i, z-\zeta\rangle < 0, \langle [V_{t_0}]_i, z\rangle < 0, \langle [V_{t_0}]_i, z+\zeta\rangle \ge 0\} \tag{D.87}$$

Let us define weight matrix $V^* \in \mathbb{R}^{m\times d}$ as:

$$V_i^* = \begin{cases} \frac{20c\log(1/\varepsilon_1)}{r\varepsilon_1} v_i z & \text{if } i \in \mathcal{E}_1; \\ -\frac{40c\log(1/\varepsilon_1)}{r\varepsilon_1} v_i z & \text{if } i \in \mathcal{E}_2; \\ -\frac{20c\log\log(1/\varepsilon_1)}{r\varepsilon_1} v_i z & \text{if } i \in \mathcal{E}_3; \\ 0 & \text{otherwise.} \end{cases} \tag{D.88}$$

for some sufficiently large universal constant c.

Note that the random noise vector $[\widetilde{V}_{t_0}]_i$ will satisfy the condition for set $\mathcal{E}_i$ with probability proportional to the angle between $z-\zeta$ and z, which is $r\pm O(r^2)$ by Taylor approximation of $\arcsin$. Thus, as V_{t_0} and $\widetilde{V}_{t_0}$ differ in at most $\varepsilon_s m$ activations, w.h.p., $|\mathcal{E}_1|, |\mathcal{E}_2|, |\mathcal{E}_3| = \frac{1}{2\pi} rm \pm \widetilde{O}\left(r^2 m + \sqrt{m}\right) \pm \varepsilon_s m$. This implies that

$$\|V^*\|_F^2 = \widetilde{O}\left(\frac{1}{r\varepsilon_1^2}\right) \tag{D.89}$$

Now, for $x_2 = z - \zeta$, we have that

$$N_{V_{t_0}}(v, V^*, z-\zeta) = \frac{1}{m}\left(|\mathcal{E}_1|\frac{20c\log(1/\varepsilon_1)}{r\varepsilon_1} - \frac{40c\log(1/\varepsilon_1)}{r\varepsilon_1}|\mathcal{E}_2|\right) \le -2c\log(1/\varepsilon_1)/\varepsilon_1 \tag{D.90}$$

and for $x_2 = z + \zeta$, we have that

$$N_{V_{t_0}}(v, V^*, z+\zeta) = -\frac{1}{m}|\mathcal{E}_3|\frac{20c\log(1/\varepsilon_1)}{r\varepsilon_1} \le -2c\log(1/\varepsilon_1)/\varepsilon_1 \tag{D.91}$$

Now, for $x_2 = z$, we have that

$$N_{V_{t_0}}(v, V^*, z) = \frac{1}{m}|\mathcal{E}_1|\frac{20c\log(1/\varepsilon_1)}{r\varepsilon_1} \ge 2c\log(1/\varepsilon_1)/\varepsilon_1 \tag{D.92}$$

Hence we can also easily conclude that for every $x_2 \in \{\alpha(z-\zeta), \alpha z, \alpha(z+\zeta)\}$,

$$yN_{V_{t_0}}(v, V^*, x_2) \ge \frac{2c\log(1/\varepsilon_1)\|x_2\|_2}{\varepsilon_1} \tag{D.93}$$

Note that for every $i \in [m]$,

$$|\langle V_i^*, x_2\rangle| \le \frac{1}{\sqrt{m}}\widetilde{O}\left(\frac{1}{\varepsilon_1 r}\right) \tag{D.94}$$

Now applying Lemma B.3, with $\eta_2 = O(\eta_1\lambda^2(\varepsilon_1 r)^2)$, we have that for every x_2, w.h.p. $\|\mathbb{1}([V_{t_0+t}]x_2) - \mathbb{1}([V_{t_0}]x_2)\|_1 \lesssim \lambda\varepsilon_1 rm$. This implies that for every $t \le \frac{1}{\eta_2\lambda}$ and every $x_2 \in \{z-\delta, z+\delta, z\}$, w.h.p.

$$\left|\sum_{i\in[m]} v_i\langle V_i^*, x_2\rangle\left[\mathbb{1}([V_{t_0+t}]_i x_2) - \mathbb{1}([V_{t_0}]_i x_2)\right]\right| \le \frac{1}{m}\widetilde{O}\left(\frac{1}{\varepsilon_1 r}\right) \times O(\lambda\varepsilon_1 rm) \le 1 \tag{D.95}$$

Combining with (D.93), this gives us

$$yN_{V_{t_0+t}}(v, V^*; x_2) = y\left(\sum_{i\in[m]} v_i\langle V_i^*, x_2\rangle \mathbb{1}([V_{t_0+t}]_i x_2)\right) \ge \frac{c\|x_2\|_2}{\varepsilon_1}\log\frac{1}{\varepsilon_1} \tag{D.96}$$

On the other hand we have that by Lemma D.10, it holds that

$$\begin{aligned} |N_{V_{t_0}}(v, \overline{V}_{t_0}; x_2)| &\le |g_{t_0}(x_2)| + |N_{V_{t_0}}(v, \overline{V}_{t_0}; x_2) - N_{V_{t_0}}(v, V_{t_0}; x_2)| && \text{(D.97)} \\ &\le |g_{t_0}(x_2)| + |N_{V_{t_0}}(v, \widetilde{V}_{t_0}; x_2)| && \text{(D.98)} \\ &\lesssim |g_{t_0}(x_2)| + \frac{\varepsilon_s}{\lambda} + \tau_0\log d \le O(1) && \text{(applying Proposition A.5)} \end{aligned}$$

Thus, we also have

$$|yN_{V_{t_0+t}}(v, \overline{V}_{t_0}; x_2)| = \left| \left(\sum_{i\in[m]} v_i \langle [\overline{V}_{t_0}]_i, x_2 \rangle \mathbb{1}([V_{t_0+t}]_i x_2) \right) \right| \tag{D.99}$$

$$\leq \left| \left(\sum_{i\in[m]} v_i \langle [\overline{V}_{t_0}]_i, x_2 \rangle \mathbb{1}([V_{t_0}]_i x_2) \right) \right| + \left| \left(\sum_{i\in[m]} v_i \langle [\overline{V}_{t_0}]_i, x_2 \rangle \left[\mathbb{1}([V_{t_0+t}]_i x_2) - \mathbb{1}([V_{t_0}]_i x_2) \right] \right) \right| \tag{D.100}$$

Now the first term equals $|N_{V_{t_0}}(v, \overline{V}_{t_0}; x_2)| = O(1)$, and the second term is bounded by

$$\left| \left(\sum_{i\in[m]} v_i \langle [\overline{V}_{t_0}]_i, x_2 \rangle \left[\mathbb{1}([V_{t_0+t}]_i x_2) - \mathbb{1}([V_{t_0}]_i x_2) \right] \right) \right| \leq \frac{1}{m} O\left(\frac{1}{\lambda}\right) \times O(\lambda \varepsilon_1 r m)$$

using Proposition A.3 to upper bound $\|[\overline{V}_{t_0}]_i\|_2$. Thus, it follows that $|yN_{V_{t_0+t}}(v, \overline{V}_{t_0}; x_2)| = O(1)$. It follows that for every $x_2 \in \{z - \zeta, z, z + \zeta\}$ and its corresponding label y, as long as $\|x_2\|_2 \geq \varepsilon_1$,

$$\begin{aligned} yN_{V_{t_0+t}}(v, \overline{V}_{t_0} + V^*; x_2) &\geq yN_{V_{t_0+t}}(v, V^*; x_2) - \left| yN_{V_{t_0+t}}(v, \overline{V}_{t_0}; x_2) \right| && \text{(D.101)} \\ &\geq c \log(1/\varepsilon_1) - \left| yN_{V_{t_0+t}}(v, \overline{V}_{t_0}; x_2) \right| && \text{(D.102)} \\ &\geq 3 \log(1/\varepsilon_1) && \text{(choosing } c \text{ sufficiently large)} \end{aligned}$$

Now we can compute

$$\begin{aligned} &\left| N_{W_{t_0+t}}(w, \overline{W}_{t_0}, x_1) - r_{t_0}(x_1) \right| && \text{(D.103)} \\ \leq &\left| N_{W_{t_0+t}}(w, \overline{W}_{t_0}, x_1) - N_{W_{t_0}}(w, \overline{W}_{t_0}, x_1) \right| + \left| N_{W_{t_0}}(w, \widetilde{W}_{t_0}, x_1) \right| && \text{(D.104)} \\ \leq &\frac{1}{\sqrt{m}} \|\mathbb{1}(W_{t_0+t} x_1) - \mathbb{1}(W_{t_0} x_1)\|_1 \max_i \|[\overline{W}_{t_0}]_i\|_2 \|x_1\|_2 + \left| N_{W_{t_0}}(w, \widetilde{W}_{t_0}, x_1) \right| \\ & && \text{(by Lemma B.3 and } \|[\overline{W}_{t_0}]_i\|_2 = O\left(\tfrac{1}{\sqrt{m}} \tfrac{1}{\lambda}\right) \text{ from Proposition A.2)} \\ \lesssim &\frac{\varepsilon_s}{\lambda} + \tau_0 \log d \leq q\varepsilon_1 && \text{(D.105)} \end{aligned}$$

The last inequality follows from our choice of parameters such that $\tau_0 \log d \leq q\varepsilon_1$. Putting together Eq (D.101) and (D.103) and defining $U^* = (0, V^*)$, we have that

$$\begin{aligned} &K_{t_0+t}(\overline{U}_{t_0} + U^*) = K_{t_0+t}((\overline{W}_{t_0}, \overline{V}_{t_0} + V^*)) && \text{(D.106)} \\ &\leq \frac{|\mathcal{M}_1|}{N} \widehat{L}_{\mathcal{M}_1}(r_{t_0}) + O(q\varepsilon_1) + \frac{|\bar{\mathcal{M}}_1|}{N} \widehat{L}_{\bar{\mathcal{M}}_1}(N_{V_{t_0+t}}(v, \overline{V}_{t_0} + V^*; *)) \\ & && \text{(by definition of } \mathcal{M}_1 \text{ and Lipschitz-ness of } \ell) \\ &\leq \varepsilon_0 + \varepsilon_1 && \text{(D.107)} \end{aligned}$$

This completes the proof.

□

Proof of Lemma 4.3. By proposition D.8, there exists V^* with $\|V^*\|_F^2 \leq \widetilde{O}\left(\frac{1}{r\varepsilon_1^2}\right)$ such that for every $t \leq \frac{1}{\eta_2 \lambda}$,

$$K_{t_0+t}((\overline{W}_{t_0}, \overline{V}_{t_0} + V^*)) \leq \varepsilon_0 + \varepsilon_1 \tag{D.108}$$

By Theorem B.2, with $z^* = (\overline{W}_{t_0}, V^*)$, starting from $z_0 = (\overline{W}_{t_0}, \overline{V}_{t_0})$, we can take $R^2 = \widetilde{O}\left(\frac{1}{r\varepsilon_1^2}\right)$, $L = 1$, $\mu = \varepsilon_1$ to conclude that the algorithm converges to $\varepsilon_0 + 2\varepsilon_1$ in $\widetilde{O}\left(\frac{1}{\eta_2 r \varepsilon_1^3}\right)$ iterations. Applying Lemma D.10 to bound ε_0 completes the proof. □

D.5 Proof of Lemma 4.4

By the 1-Lipschitzness of logistic loss, we know that

$$\left|\widehat{L}_{\mathcal{M}_1}(r_{t_0}) - \widehat{L}_{\mathcal{M}_1}(r_{t_0+t})\right| \tag{D.109}$$

$$= \left|\frac{1}{|\mathcal{M}_1|}\sum_{i\in\mathcal{M}_1}\left(\ell(r_{t_0};(x^{(i)},y^{(i)})) - \ell(r_{t_0+t};(x^{(i)},y^{(i)}))\right)\right| \tag{D.110}$$

$$\leq \frac{1}{|\mathcal{M}_1|}\sum_{i\in\mathcal{M}_1}\left|r_{t_0}(x_1^{(i)}) - r_{t_0+t}(x_1^{(i)})\right| \tag{D.111}$$

To bound this term, we can directly use Cauchy-Shwartz and obtain that:

$$\sum_{i\in\mathcal{M}_1}\left|r_{t_0+t}(x_1^{(i)}) - r_{t_0}(x_1^{(i)})\right| \tag{D.112}$$

$$\leq \sqrt{N}\sqrt{\sum_{i\in\mathcal{M}_1}\left(r_{t_0+t}(x_1^{(i)}) - r_{t_0}(x_1^{(i)})\right)^2} \tag{D.113}$$

We can further bound $r_{t_0+t}(x_1^{(i)}) - r_{t_0}(x_1^{(i)})$ by applying Lemma B.3, as from our choice of parameters $\eta_2 \leq \eta_1\varepsilon_1^4\lambda^2, \varepsilon_s/\lambda \leq \varepsilon_1^2, \tau_0 \log d \leq \varepsilon_1^2$:

$$\left|r_{t_0+t}(x_1^{(i)}) - r_{t_0}(x_1^{(i)})\right| \tag{D.114}$$

$$\leq \left|N_{W_{t_0+t}}(w, W_{t_0+t}, x_1^{(i)}) - N_{W_{t_0}}(w, \overline{W}_{t_0+t}, x_1^{(i)})\right| + \tag{D.115}$$

$$\left|N_{W_{t_0}}(w, \overline{W}_{t_0+t}, x_1^{(i)}) - N_{W_{t_0}}(w, \overline{W}_{t_0}, x_1^{(i)})\right| + \left|N_{W_{t_0}}(w, \widetilde{W}_{t_0}, x_1^{(i)})\right| \tag{D.116}$$

$$\leq \left|N_{W_{t_0}}(w, \overline{W}_{t_0+t}, x_1^{(i)}) - N_{W_{t_0}}(w, \overline{W}_{t_0}, x_1^{(i)})\right| + O\left(\frac{1}{\lambda}\times\left(\sqrt{\frac{\eta_2}{\eta_1}} + \varepsilon_s\right) + \tau_0\log d\right)$$

(by Lemma B.3 and Proposition A.5)

$$\leq \left|N_{W_{t_0}}(w, \overline{W}_{t_0+t}, x_1^{(i)}) - N_{W_{t_0}}(w, \overline{W}_{t_0}, x_1^{(i)})\right| + \varepsilon_1^2 \tag{D.117}$$

Now, let us denote $X = (x^{(i)})_{i\in[N]}$ as the data matrix. By the standard Gaussian matrix spectral norm bound we know that w.h.p. $\|X\|_2^2 \leq 10\frac{N}{d}$.

This gives us:

$$\sqrt{N}\sqrt{\sum_{i\in\mathcal{M}_1}\left(N_{W_{t_0}}(w, \overline{W}_{t_0+t}, x_1^{(i)}) - N_{W_{t_0}}(w, \overline{W}_{t_0}, x_1^{(i)})\right)^2}$$

$$\leq \sqrt{N}\sqrt{\|\overline{W}_{t_0+t} - \overline{W}_{t_0}\|_F^2\|X\|_2^2} \qquad \text{(expanding the expression of } N_{W_{t_0}}(w, W_{t_0+t}, x_1^{(i)})\text{)}$$

$$\leq \sqrt{N}\sqrt{10\left(\left\|\overline{W}_{t_0+t} - \overline{W}_{t_0}\right\|_F^2\right)\frac{N}{d}} \tag{D.118}$$

$$\leq N\tilde{O}\left(\frac{1}{\sqrt{dr}\varepsilon_1}\right) \leq N\varepsilon_1 \tag{D.119}$$

Here in (D.119), we use the assumption $dr \geq \tilde{\Omega}\left(\frac{1}{\varepsilon_1^4}\right)$ in Theorem 3.4 along with the fact that by Lemma 4.3, we have that

$$\left\|\overline{W}_{t_0+t} - \overline{W}_{t_0}\right\|_F^2 \leq \tilde{O}\left(\frac{1}{r\varepsilon_1^2}\right) \tag{D.120}$$

Thus, using (D.119), it follows that

$$\sum_{i\in\mathcal{M}_1}\left|r_{t_0+t}(x_1^{(i)})-r_{t_0}(x_1^{(i)})\right|$$
$$\lesssim\sqrt{N}\sqrt{\sum_{i\in\mathcal{M}_1}\left(N_{W_{t_0}}(w,\overline{W}_{t_0+t},x_1^{(i)})-N_{W_{t_0}}(w,\overline{W}_{t_0},x_1^{(i)})\right)^2}+N\varepsilon_1^2\le N\varepsilon_1$$

By (D.109) and our definition of ε_0 as

$$\varepsilon_0:=\frac{|\mathcal{M}_1|}{N}\widehat{L}_{\mathcal{M}_1}(r_{t_0})=(1-q)\widehat{L}_{\mathcal{M}_1}(r_{t_0}) \tag{D.121}$$

we must have

$$\left|\widehat{L}_{\mathcal{M}_1}(r_{t_0+t})-\frac{\varepsilon_0}{1-q}\right|\le\varepsilon_1/2 \tag{D.122}$$

Using the bound on ε_0 that $\varepsilon_0=O(\sqrt{\varepsilon_1/q})$ by Lemma D.10, we conclude the bound on $\widehat{L}_{\mathcal{M}_1}(r_{t_0+t})$. In the end, by $\widehat{L}_{\bar{\mathcal{M}}_1}(g_{t_0+t})\le\widehat{L}_{t_0+t}$ and the assumption that $\widehat{L}_{t_0+t}\le O(\sqrt{\varepsilon_1/q})$, it must hold that (since $|\bar{\mathcal{M}}_1|=qN$)

$$\widehat{L}_{\bar{\mathcal{M}}_1}(g_{t_0+t})\lesssim\sqrt{\frac{\varepsilon_1}{q^3}} \tag{D.123}$$

so we can complete the proof.

E Proofs for Small Learning Rate

E.1 Proof of Lemma 5.1

We first show the following Lemma:

Lemma E.1. *In the setting of theorem 3.5, there exists a solution $U^\star$ satisfying a) $\|U^\star\|_F^2\le\widetilde{O}(\frac{1}{\varepsilon_2'^2r}+Np)$ and b) for every $t\le\frac{1}{\eta_2\lambda}$,*

$$K_t(U^\star)\le\varepsilon_2' \tag{E.1}$$

Proof of Lemma E.1. We can construct the matrix $U^\star$ as follows: let $X=(x_1^i)_{i\in\bar{\mathcal{M}}_2}\in\mathbb{R}^{d\times Np}$ and $Y=(y^{(i)})_{i\in\bar{\mathcal{M}}_2}\in\mathbb{R}^{1\times Np}$. If we define $s=X(X^\top X)^{-1}y^\top\in\mathbb{R}^{d\times1}$, we know that $s^\top X=y$ with $\|s\|_2=O\left(\sqrt{Np}\right)$. Thus, we can define V^* as in Lemma 4.3 with $t_0=0$, and $W_i^*=10\log\frac{1}{\varepsilon_2'}sw_i$, and we can see that for every $t\le\frac{1}{\eta_2\lambda}$, it holds that

$$K_t((W^*,V^*))\le\varepsilon_2' \tag{E.2}$$

□

To prove Lemma 5.1, we can apply an identical analysis as 4.3 to show that for $t'=\widetilde{O}\left(\frac{1}{\eta_2\varepsilon_2'^3r}\right)$, $\widehat{L}_{\mathcal{M}_2}(U_{t'})\le\varepsilon_2'$. The rest of the proof follows from combining Theorem B.2 and Lemma E.1.

E.2 Proof of Lemma 5.2

We will use the following Lemma from [6].

Lemma E.2 (Lemma 6.3 of [6])**.** *For every v_1,v_2,v_3, let $g\sim\mathcal{N}(0,I)$ in $\mathbb{R}^d$, then we have:*

$$\mathbb{E}_g\left[\|v_1\mathbb{1}(\langle g,z-\zeta\rangle)(z-\zeta)+v_2\mathbb{1}(\langle g,z+\zeta\rangle)(z+\zeta)+v_3\mathbb{1}(\langle g,z\rangle)z\|_2^2\right] \tag{E.3}$$
$$\gtrsim r\left(v_1^2+v_2^2+v_3^2\right) \tag{E.4}$$

Recall the expression ρ_t defined in (B.10). We first prove Lemma B.4 here, which says that if ρ_t is large (which means the loss is large as well), then the total gradient norm has to be big.

Proof of Lemma B.4. For notation simplicity, let's fix t and let

$$Q_j = \ell'_{j,t} \tag{E.5}$$

The gradient with respect to V can be computed by

$$\nabla_{[V]_k} \widehat{L}(U_t) = \frac{1}{N} \sum_{j \in \mathcal{M}_2} Q_j v_k \mathbb{1}(\langle [V_t]_k, x_2^{(j)} \rangle) x_2^{(j)} \tag{E.6}$$

Let us denote the set $\mathcal{S}_{2,1}^{(\alpha_0)}, \mathcal{S}_{2,2}^{(\alpha_0)}, \mathcal{S}_{2,3}^{(\alpha_0)}$ as:

$$\mathcal{S}_{2,1}^{(\alpha_0)} = \left\{ j \in [m] \mid x_2^{(j)} = \alpha_j (z - \zeta) \text{ for some } \alpha_j \geq \alpha_0 \right\} \tag{E.7}$$

$$\mathcal{S}_{2,2}^{(\alpha_0)} = \left\{ j \in [m] \mid x_2^{(j)} = \alpha_j (z + \zeta) \text{ for some } \alpha_j \geq \alpha_0 \right\} \tag{E.8}$$

$$\mathcal{S}_{2,3}^{(\alpha_0)} = \left\{ j \in [m] \mid x_2^{(j)} = \alpha_j z \text{ for some } \alpha_j \geq \alpha_0 \right\} \tag{E.9}$$

We then have that

$$N m v_k \nabla_{[V]_k} L_t \tag{E.10}$$

$$= \sum_{j \in \mathcal{S}_{2,1}^{(0)}} \alpha_j Q_j \mathbb{1}(\langle [V_t]_k, z - \zeta \rangle)(z - \zeta) + \sum_{j \in \mathcal{S}_{2,2}^{(0)}} \alpha_j Q_j \mathbb{1}(\langle [V_t]_k, z + \zeta \rangle)(z + \zeta) \tag{E.11}$$

$$+ \sum_{j \in \mathcal{S}_{2,3}^{(0)}} \alpha_j Q_j \mathbb{1}(\langle [V_t]_k, z \rangle) z \tag{E.12}$$

For each $k \in [m]$, let us define

$$\tilde{L}_k \triangleq \sum_{j \in \mathcal{S}_{2,1}^{(0)}} \alpha_j Q_j \mathbb{1}(\langle [\tilde{V}_t]_k, z - \zeta \rangle)(z - \zeta) + \sum_{j \in \mathcal{S}_{2,2}^{(0)}} \alpha_j Q_j \mathbb{1}(\langle [\tilde{V}_t]_k, z + \zeta \rangle)(z + \zeta) \tag{E.13}$$

$$+ \sum_{j \in \mathcal{S}_{2,3}^{(0)}} \alpha_j Q_j \mathbb{1}(\langle [\tilde{V}_t]_k, z \rangle) z \tag{E.14}$$

i.e., the loss gradient using activations computed by the noise component of V_t scaled by a factor of $N m v_k$.

By the Geometry of ReLU Lemma E.2, we have that w.h.p.

$$\mathbb{E}_{[\tilde{V}_t]_k} \left[\left\| \tilde{L}_k \right\|_2^2 \right] \geq r\Omega \left(\left(\sum_{j \in \mathcal{S}_{2,1}^{(0)}} \alpha_j Q_j \right)^2 + \left(\sum_{j \in \mathcal{S}_{2,2}^{(0)}} \alpha_j Q_j \right)^2 + \left(\sum_{j \in \mathcal{S}_{2,3}^{(0)}} \alpha_j Q_j \right)^2 \right) \tag{E.15}$$

$$\geq r\Omega \left(\left(\sum_{j \in \mathcal{M}_2} \alpha_j |Q_j| \right)^2 \right) \tag{E.16}$$

Where the last inequality is obtained since for every $j \in \mathcal{S}_{2,j'}^{(0)}$, Q_j has the same sign.

Since each $[\tilde{V}_t]_k$ are independent and $|\alpha_j Q_j|, \|z\|_2, \|\zeta\|_2 = O(1)$, by concentration, we know that taking a union bound over all choices of Q_j, w.h.p.

$$\|\tilde{L}\|_F^2 \geq mr\Omega \left(\left(\sum_j \alpha_j |Q_j| \right)^2 \right) - \widetilde{O}(m^{1/2} N^4) \tag{E.17}$$

where $\tilde{L}$ denotes the matrix where each $\tilde{L}_k$ is a row. By Coupling Lemma A.8, we note that as

$$\frac{1}{N^2 m}\|\tilde{L}\|_F^2 - \|\nabla \hat{L}(U_t)\|_F^2 \lesssim \frac{1}{Nm}\sum_k \sum_j Q_j^2 |\mathbb{1}(\langle [V_t]_k, x_2^{(j)}\rangle) - \mathbb{1}(\langle [\widetilde{V}_t]_k, x_2^{(j)}\rangle)| \lesssim O(\varepsilon_s)$$

we therefore also have w.h.p.:

$$\|\nabla \hat{L}(U_t)\|_F^2 \geq \frac{1}{N^2 m}\|\tilde{L}\|_F^2 - O(\varepsilon_s) \tag{E.18}$$

$$\geq \frac{r}{N^2}\Omega\left(\left(\sum_j \alpha_j |Q_j|\right)^2\right) - \widetilde{O}(m^{-1/2}N^2) - O(\varepsilon_s) \tag{E.19}$$

Note that $\alpha_j \sim U(0,1)$, and therefore for every fixed $\alpha_0 \geq \frac{1}{\sqrt{N}}$, w.h.p. there are $O(N\alpha_0)$ many α_j such that $\alpha_j \leq \alpha_0$. For each of them, we also know that $|Q_j| \leq 1$, which implies that

$$\left(\sum_j \alpha_j |Q_j|\right)^2 \geq \alpha_0^2 \left(\sum_{j:\alpha_j \geq \alpha_0} |Q_j|\right)^2 \tag{E.20}$$

$$\geq \alpha_0^2 \left(\left(\sum_j |Q_j|\right) - O(N\alpha_0^2)\right)^2 \tag{E.21}$$

$$\geq \alpha_0^2 (N(\rho_t - O(\alpha_0^2)))^2 \tag{E.22}$$

Picking $\alpha_0 = \Theta(\sqrt{\rho_t})$, we complete the proof by our choice of $m \geq N^{10}\frac{1}{(\lambda\tau_0)^4}$. □

Now we prove Proposition B.5, which bounds the number of iterations in which ρ_t can be large.

Proof of Proposition B.5. Consider the function $\mathcal{F}_s(x) := N_{U_0}(u, \overline{U}_s; x)$, and let us define $\mathcal{G}_{s+1}(x) := N_{U_0}(u, \overline{U}_s - \frac{\eta_2}{1-\eta_2\lambda}\nabla \widehat{L}(U_s); x)$. We have that since $\overline{U}_{s+1} = (1-\eta_2\lambda)\overline{U}_s - \eta_2 \nabla \widehat{L}(U_s)$,

$$\widehat{L}(\mathcal{F}_{s+1}) = \widehat{L}((1-\eta_2\lambda)\mathcal{G}_{s+1}) \leq (1+\eta_2\lambda)\widehat{L}(\mathcal{G}_{s+1}) \tag{E.23}$$

Here we use the fact that for logistic loss ℓ, $\ell((1-\alpha)z) \leq (1+\alpha)\ell(z)$ for every $z \in \mathbb{R}, \alpha \in [0, 0.1]$.

Now, by standard gradient descent analysis, we have that (as the logistic loss has Lipschitz derivative and the data have bounded norm):

$$\widehat{L}(\mathcal{G}_{s+1}) \leq \widehat{L}(\mathcal{F}_s) - \frac{\eta_2}{1-\eta_2\lambda}\langle \nabla \widehat{L}(\mathcal{F}_s), \nabla \widehat{L}(U_s)\rangle + 2\eta_2^2 \|\nabla \widehat{L}(U_s)\|_F^2 \tag{E.24}$$

$$\leq \widehat{L}(\mathcal{F}_s) - \frac{\eta_2}{1-\eta_2\lambda}\langle \nabla \widehat{L}(\mathcal{F}_s), \nabla \widehat{L}(U_s)\rangle + O(\eta_2^2) \qquad \text{(by Proposition A.2)}$$

Next, we will bound $\|\nabla \widehat{L}(U_s) - \nabla \widehat{L}(\mathcal{F}_s)\|_F$. We can compute

$$\|\nabla \widehat{L}(U_s) - \nabla \widehat{L}(\mathcal{F}_s)\|_F^2 \tag{E.25}$$

$$\leq \frac{1}{N^2 m}\sum_{k\in[m]} \Big\| \sum_j \big(\ell'(-y^{(j)} N_{U_s}(u, U_s; x^{(j)}))\mathbb{1}([U_s]_k x^{(j)}) - \tag{E.26}$$

$$\ell'(-y^{(j)} N_{U_0}(u, \overline{U}_s; x^{(j)}))\mathbb{1}([U_0]_k x^{(j)})\big) x^{(j)}\Big\|_2^2 \tag{E.27}$$

$$\leq \frac{1}{Nm}\sum_{k\in[m]}\sum_j \Big\| \big(\ell'(-y^{(j)} N_{U_s}(u, U_s; x^{(j)}))\mathbb{1}([U_s]_k x^{(j)}) - \tag{E.28}$$

$$\ell'(-y^{(j)} N_{U_0}(u, \overline{U}_s; x^{(j)}))\mathbb{1}([U_0]_k x^{(j)})\big) x^{(j)}\Big\|_2^2 \tag{E.29}$$

where the last step followed via Cauchy-Schwarz. Now by the Lipschitzness of ℓ', we have the bound

$$\ell'(-y^{(j)}N_{U_s}(u,U_s;x^{(j)}))\mathbb{1}([U_s]_k x^{(j)}) - \ell'(-y^{(j)}N_{U_0}(u,\overline{U}_s;x^{(j)}))\mathbb{1}([U_0]_k x^{(j)}) \lesssim \\ |N_{U_s}(u,U_s;x^{(j)})) - N_{U_0}(u,\overline{U}_s;x^{(j)})| + |\mathbb{1}([U_s]_k x^{(j)}) - \mathbb{1}([U_0]_k x^{(j)})|$$

Plugging this back into (E.29), by the coupling Lemma B.3 we obtain the bound

$$\|\nabla\widehat{L}(U_s) - \nabla\widehat{L}(\mathcal{F}_s)\|_F^2 \lesssim \frac{1}{\lambda}(\varepsilon_s + \sqrt{\eta_2/\eta_1}) + \tau_0 \log d := \varepsilon_c^2$$

This implies that for $\eta_2\lambda < 0.1$,

$$\widehat{L}(\mathcal{G}_{s+1}) \le \widehat{L}(\mathcal{F}_s) - \frac{1}{2}\eta_2\|\nabla\widehat{L}(U_s)\|_F^2 + O(\eta_2^2 + \eta_2\varepsilon_c) \tag{E.30}$$

Hence, we have

$$\widehat{L}(\mathcal{F}_{s+1}) \le (1+\eta_2\lambda)\widehat{L}(\mathcal{F}_s) - \frac{1}{2}(1+\eta_2\lambda)\eta_2\|\nabla\widehat{L}(U_s)\|_F^2 + O(\eta_2^2 + \eta_2\varepsilon_c) \tag{E.31}$$

which implies that for every $t \le \frac{1}{\eta_2\lambda}$, as long as $\eta_2, \varepsilon_c = O(\lambda)$, we have:

$$\eta_2 \sum_{s\le t} \|\nabla\widehat{L}(U_s)\|_F^2 \lesssim \widehat{L}(\mathcal{F}_0) \lesssim 1 \tag{E.32}$$

By Lemma B.4, we have that if $\rho_t \ge {\varepsilon_2'}^2\varepsilon_3^2$, then $\|\nabla\widehat{L}(U_s)\|_F^2 \ge r{\varepsilon_2'}^8\varepsilon_3^8$. It follows that there will be at most $O(\frac{1}{r{\varepsilon_2'}^8\varepsilon_3^8\eta_2})$ such t.

□

Finally, we complete the proof of Lemma 5.2 by noting that ρ_t cannot be large for very many iterations, and therefore W_t will not obtain much signal from the $\mathcal{P}$ component of examples in $\mathcal{M}_2$.

Proof of Lemma 5.2. We have,

$$\left\|\sum_{j\in\mathcal{M}_2} \nabla_W \widehat{L}_j(U_t)\right\|_2^2 = \sum_{k\in[m]} \left\|\sum_{j\in\mathcal{M}_2} \ell'_{j,t} w_k \mathbb{1}(\langle [W_t]_k, x_1^{(j)}\rangle) x_1^{(j)}\right\|_2^2$$

Now we note that the above can be reformulated as a matrix multiplication between the matrix of data X and the vector with entry $\ell'_{j,t}w_k\mathbb{1}(\langle[W_t]_k, x_1^{(j)}\rangle)$ in the j-th coordinate for $j \in \mathcal{M}_2$ and 0 elsewhere. Thus,

$$\left\|\sum_{j\in\mathcal{M}_2} \nabla_W \widehat{L}_j(U_t)\right\|_2^2 \le \sum_{k\in[m]} \|X\|_2^2 \left(\sum_{j\in\mathcal{M}_2} \left(\ell'_{j,t} w_k \mathbb{1}(\langle [W_t]_k, x_1^{(j)}\rangle)\right)^2\right)$$

(definition of spectral norm)

$$\le \sum_{k\in[m]} \|X\|_2^2 \left(\sum_{j\in\mathcal{M}_2} \left(\ell'_{j,t} w_k\right)^2\right)$$

$$= \|X\|_2^2 \sum_{j\in\mathcal{M}_2} \left(\ell'_{j,t}\right)^2 \qquad \text{(because } w_k \in \{\pm 1/\sqrt{m}\}\text{)}$$

$$\lesssim \|X\|_2^2 \sum_{j\in\mathcal{M}_2} \left|\ell'_{j,t}\right| \qquad \text{(because the } \ell \text{ is } O(1)\text{-Lipschitz)}$$

$$\lesssim N/d \cdot N\rho_t \tag{E.33}$$

The last line followed from the spectral norm bound on matrix X. Let $\mathcal{T}$ be defined as in Proposition B.5. It follows that

$$\left\|\overline{W}_t^{(2)}\right\|_F \leq \eta_2 \sum_{s\leq t} \left\| \left(\frac{1}{N} \sum_{j\in\mathcal{M}_2} \nabla_W \ell(f_s; (x^{(j)}, y^{(j)})) \right) \right\|_F \tag{E.34}$$

$$= \eta_2 \sum_{s\in\mathcal{T}} \left\| \left(\frac{1}{N} \sum_{j\in\mathcal{M}_2} \nabla_W \ell(f_s; (x^{(j)}, y^{(j)})) \right) \right\|_F + \tag{E.35}$$

$$\eta_2 \sum_{s\notin\mathcal{T}} \left\| \left(\frac{1}{N} \sum_{j\in\mathcal{M}_2} \nabla_W \ell(f_s; (x^{(j)}, y^{(j)})) \right) \right\|_F$$

$$\leq \eta_2 \sum_{s\in\mathcal{T}} \left\| \left(\frac{1}{N} \sum_{j\in\mathcal{M}_2} \nabla_W \ell(f_s; (x^{(j)}, y^{(j)})) \right) \right\|_F + \eta_2 t O\left(\frac{\varepsilon_2'\varepsilon_3}{\sqrt{d}}\right)$$

(by definition of $\mathcal{T}$ and equation (E.33))

Note that we can additionally bound the first term by $\eta_2|\mathcal{T}|O(\frac{1}{\sqrt{d}})$ as $\rho_t \leq 1$ by the Lipschitzness of ℓ. Thus, applying our bound on $|\mathcal{T}|$, we get

$$\left\|\overline{W}_t^{(2)}\right\|_F \leq O\left(\frac{1}{r\sqrt{d}\varepsilon_2'^8\varepsilon_3^8} + \frac{\eta_2\varepsilon_2'\varepsilon_3 t}{\sqrt{d}}\right) \tag{E.36}$$

Now the conclusion of the lemma follows by the assumption that $t = O(d/\eta_2\varepsilon_2')$ and our choice of η_2 and $\frac{1}{\varepsilon_2'^8\varepsilon_3^8 r} \leq \varepsilon_2' d$ in Theorem 3.5.

□

E.3 Proof of Lemma 5.3

We now prove the decomposition lemma of $\overline{W}_t$, Lemma B.6. Recall our definition of $\overline{W}_t^{(2)}$ as

$$\overline{W}_t^{(2)} = \frac{1}{N}\eta_2 \sum_{s\leq t} (1-\eta_2\lambda)^{t-s} \sum_{i\in\mathcal{M}_2} \nabla_W \widehat{L}_{\{i\}}(U_s) \tag{E.37}$$

Proof of Lemma B.6. For each step, we know that for every $j \in [m]$,

$$\nabla_{W_j} \hat{L}(U_s) = w_j \frac{1}{N} \sum_{i\in[N]} \ell_{i,s}' \mathbb{1}([W_s]_j x_1^{(i)}) x_1^{(i)}$$

Thus, multiplying by $\eta_2(1-\eta_2\lambda)^{t-s}$ and summing, following our definition of $\overline{W}_t^{(2)}$ in (5.1), we get

$$[\overline{W}_t]_j = [\overline{W}_t^{(2)}]_j + w_j \frac{1}{N}\eta_2 \sum_{s\leq t} (1-\eta_2\lambda)^{t-s} \sum_{i\in\bar{\mathcal{M}}_2} \ell_{i,s}' \mathbb{1}([W_s]_j x_1^{(i)}) x_1^{(i)} \tag{E.38}$$

$$= [\overline{W}_t^{(2)}]_j + \tag{E.39}$$

$$w_j \frac{1}{N}\eta_2 \sum_{s\leq t} (1-\eta_2\lambda)^{t-s} \cdot \left(\sum_{i\in\bar{\mathcal{M}}_2} \ell_{i,s}' \mathbb{1}([W_0]_j x_1^{(i)}) x_1^{(i)} + \right. \tag{E.40}$$

$$\left. \sum_{i\in\bar{\mathcal{M}}_2} \ell_{i,s}' \left[\mathbb{1}([W_s]_j x_1^{(i)}) - \mathbb{1}([W_0]_j x_1^{(i)}) \right] x_1^{(i)} \right) \tag{E.41}$$

Now we focus on bounding the bottom term. We can see that

$$\sum_{j\in[m]}\left\|w_j\frac{1}{N}\sum_{i\in\bar{\mathcal{M}}_2}\ell'_{i,s}\left[\mathbb{1}([W_s]_j x_1^{(i)})-\mathbb{1}([W_0]_j x_1^{(i)})\right]x_1^{(i)}\right\|_2^2 \tag{E.42}$$

$$\le \frac{1}{mN}\sum_{j\in[m]}\sum_{i\in\bar{\mathcal{M}}_2}\left\|\ell'_{i,s}\left[\mathbb{1}([W_s]_j x_1^{(i)})-\mathbb{1}([W_0]_j x_1^{(i)})\right]x_1^{(i)}\right\|_2^2$$

(since $w_j=\pm1/\sqrt{m}$ and by Cauchy-Schwarz)

$$\lesssim \frac{1}{mN}\sum_{i\in\bar{\mathcal{M}}_2}\sum_{j\in[m]}\left\|\left[\mathbb{1}([W_s]_j x_1^{(i)})-\mathbb{1}([W_0]_j x_1^{(i)})\right]x_1^{(i)}\right\|_2^2 \qquad \text{(by Lipschitzness of } \ell)$$

By Auxiliary Coupling Lemma B.3 with $t_0=0$, we know that for $s\le\frac{1}{\eta_2\lambda}$, w.h.p.

$$\sum_{j\in[m]}\left\|\left[\mathbb{1}([W_s]_j x_1^{(i)})-\mathbb{1}([W_0]_j x_1^{(i)})\right]x_1^{(i)}\right\|_2^2 \le \left\|\mathbb{1}(W_s x_1^{(i)})-\mathbb{1}(W_0 x_1^{(i)})\right\|_1\|x_1^{(i)}\|_2^2 \tag{E.43}$$

$$\le \widetilde{O}\left(\varepsilon_s m+\sqrt{\frac{\eta_2}{\eta_1}}m\right) \tag{E.44}$$

Thus, we have

$$\sum_{j\in[m]}\left\|w_j\frac{1}{N}\sum_{i\in\bar{\mathcal{M}}_2}\ell'_{i,s}\left[\mathbb{1}([W_s]_j x_1^{(i)})-\mathbb{1}([W_0]_j x_1^{(i)})\right]x_1^{(i)}\right\|_2^2 \tag{E.45}$$

$$\lesssim \frac{1}{mN}\sum_{i\in\bar{\mathcal{M}}_2}\sum_{j\in[m]}\left\|\left[\mathbb{1}([W_s]_j x_1^{(i)})-\mathbb{1}([W_0]_j x_1^{(i)})\right]x_1^{(i)}\right\|_2^2 \tag{E.46}$$

$$\le \widetilde{O}\left(\varepsilon_s+\sqrt{\frac{\eta_2}{\eta_1}}\right) \tag{E.47}$$

Now, we can express the weight

$$[\overline{W}_t]_j = w_j\sum_{k\in\bar{\mathcal{M}}_2}\alpha_k x_1^{(k)}\mathbb{1}([W_0]_j x_1^{(k)})+[\overline{W}'_t]_j \tag{E.48}$$

for some real values $\{\alpha_k\}_{k\in\bar{\mathcal{M}}_2}$ with

$$\alpha_k=\eta_2\sum_{s\le t}(1-\eta_2\lambda)^{t-s}\ell'_{k,s} \tag{E.49}$$

and

$$[\overline{W}'_t]_j=[\overline{W}_t^{(2)}]_j+w_j\frac{1}{N}\eta_2\sum_{s\le t}(1-\eta_2\lambda)^{t-s}\sum_{i\in\bar{\mathcal{M}}_2}\ell'_{i,s}\left[\mathbb{1}([W_t]_j x_1^{(i)})-\mathbb{1}([W_0]_j x_1^{(i)})\right]x_1^{(i)} \tag{E.50}$$

By the above calculation, (E.47), and Lemma 5.2, we have:

$$\|\overline{W}'_t\|_F\le\|\overline{W}_t^{(2)}\|_F+\frac{1}{\lambda}\widetilde{O}\left(\sqrt{\varepsilon_s+\sqrt{\frac{\eta_2}{\eta_1}}}\right)\le\widetilde{O}\left(\varepsilon_3\sqrt{d}\right) \tag{E.51}$$

where the last inequality followed by our choice of parameters. □

Using the decomposition lemma, the conclusion of Lemma 5.3 now follows via computation.

Proof of Lemma 5.3. We first show that the network output on $x_1^{(i)}$ is close to that of some kernel prediction function by applying Lemma B.6. We vector-multiply the equality $[\bar{W}_t]_j = w_j \sum_{k\in\bar{\mathcal{M}}_2} \alpha_k x_1^{(k)} \mathbb{1}([W_0]_j x_1^{(k)}) + [\bar{W}'_t]_j$ on both sides by $w_j \mathbb{1}([W_0]_j x_1^{(i)})$ and sum over all j to get:

$$\left| \sum_{j\in[m]} w_j \langle [\overline{W}_t]_j, x_1^{(i)} \rangle \mathbb{1}([W_0]_j x_1^{(i)}) - \frac{1}{m} \sum_{j\in[m]} \sum_{k\in\bar{\mathcal{M}}_2} \alpha_k \langle x_1^{(k)}, x_1^{(i)} \rangle \mathbb{1}([W_0]_j x_1^{(k)}) \mathbb{1}([W_0]_j x_1^{(i)}) \right| \tag{E.52}$$

$$= \left| \sum_{j\in[m]} w_j \langle [\bar{W}'_t]_j, x_1^{(i)} \rangle \mathbb{1}([W_0]_j x_1^{(i)}) \right| \tag{E.53}$$

$$\leq \sqrt{\sum_{j\in[m]} \langle [\bar{W}'_t]_j, x_1^{(i)} \rangle^2} \qquad \text{(by Cauchy-Schwarz)}$$

$$= \|\overline{W}'_t x_1^{(i)}\|_2 \tag{E.54}$$

Let us define the function $\mathfrak{U}$ as:

$$\mathfrak{U}(x_1) := \frac{1}{m} \sum_{j\in[m]} \sum_{k\in\bar{\mathcal{M}}_2} \alpha_k \langle x_1^{(k)}, x_1 \rangle \mathbb{1}([W_0]_j x_1^{(k)}) \mathbb{1}(\langle [W_0]_j, x_1 \rangle) \tag{E.55}$$

Note that $\mathfrak{U}$ is some kernel prediction function. Since each $[W_0]_j$ is distributed as a vector of i.i.d. spherical Gaussians, we know that for fixed $x_1^{(k)}, x_1$:

$$\mathbb{E}\left[\mathbb{1}([W_0]_j x_1^{(k)}) \mathbb{1}(\langle [W_0]_j, x_1 \rangle) \right] = \frac{1}{2\pi} \arccos \Theta(x_1^{(k)}, x_1) \tag{E.56}$$

In the above equation $\Theta(x_1^{(k)}, x_1^{(i)})$ is the principle angle between $x_1^{(k)}, x_1^{(i)}$. Since each $[W_0]_j$ is i.i.d., with basic concentration bounds, we know that w.h.p.

$$\begin{aligned}
\mathfrak{U}(x_1^{(i)}) &= \sum_{k\in\bar{\mathcal{M}}_2} \alpha_k \langle x_1^{(k)}, x_1^{(i)} \rangle \frac{1}{2\pi} \arccos \Theta(x_1^{(k)}, x_1^{(i)}) \pm O(m^{-1/6}) \\
&= \frac{1}{2} \alpha_i \|x_1^{(i)}\|_2^2 \\
&+ \sum_{k\in\bar{\mathcal{M}}_2, k\neq i} \alpha_k \langle x_1^{(k)}, x_1^{(i)} \rangle \frac{1}{4} \left(1 - \frac{1}{2\pi} \frac{\langle x_1^{(k)}, x_1^{(i)} \rangle}{\|x_1^{(k)}\|_2 \|x_1^{(i)}\|_2} \pm O\left(\frac{\langle x_1^{(k)}, x_1^{(i)} \rangle}{\|x_1^{(k)}\|_2 \|x_1^{(i)}\|_2} \right)^3 \right) \\
&\pm O(m^{-1/6}) \qquad \text{(by Taylor expansion of arccos)} \\
&= \frac{1}{2} \alpha_i \|x_1^{(i)}\|_2^2 \qquad \text{(E.57)} \\
&+ \sum_{k\in\bar{\mathcal{M}}_2, k\neq i} \alpha_k \langle x_1^{(k)}, x_1^{(i)} \rangle \frac{1}{4} \left(1 - \frac{1}{2\pi} \frac{\langle x_1^{(k)}, x_1^{(i)} \rangle}{\|x_1^{(k)}\|_2 \|x_1^{(i)}\|_2} \pm \widetilde{O}\left(d^{-3/2} \right) \right) \pm O(m^{-1/6})
\end{aligned}$$

The last inequality uses the fact that w.h.p. for $k \neq i$, $\frac{\langle x_1^{(k)}, x_1^{(i)} \rangle}{\|x_1^{(k)}\|_2 \|x_1^{(i)}\|_2} = \widetilde{O}(d^{-1/2})$.

Let us define $\alpha = \frac{1}{4}\sum_{k\in\bar{\mathcal{M}}_2}\alpha_k x_1^{(k)}$; then

$$\left|\sum_{k\in\bar{\mathcal{M}}_2}\alpha_k\langle x_1^{(k)},x_1^{(i)}\rangle\frac{1}{4}\left(1-\frac{1}{2\pi}\frac{\langle x_1^{(k)},x_1^{(i)}\rangle}{\|x_1^{(k)}\|_2\|x_1^{(i)}\|_2}\right)\right| \tag{E.58}$$

$$\leq |\langle\alpha,x_1^{(i)}\rangle| + \frac{1}{8\pi}\sum_{k\in\bar{\mathcal{M}}_2}|\alpha_k|\frac{\langle x_1^{(k)},x_1^{(i)}\rangle^2}{\|x_1^{(k)}\|_2\|x_1^{(i)}\|_2} \tag{E.59}$$

$$\leq |\langle\alpha,x_1^{(i)}\rangle| + |\alpha_i\langle x_1^{(i)},x_1^{(i)}\rangle| + \frac{1}{d}\widetilde{O}\left(\sum_{k\in\bar{\mathcal{M}}_2,k\neq i}|\alpha_k|\right) \tag{E.60}$$

Since the training loss is at $\varepsilon_2 \leq p/10$, we know that $\frac{1}{|\bar{\mathcal{M}}_2|}\sum_{i\in\bar{\mathcal{M}}_2}|\mathfrak{U}(x_1^{(i)})| \geq 1$ (or else the loss would not be low).

Since $|\mathfrak{U}(x_1^{(i)})| \leq |\langle\alpha,x_1^{(i)}\rangle| + \frac{3}{2}|\alpha_i|\|x_1^{(i)}\|_2^2 + \frac{1}{d}\widetilde{O}\left(\sum_{k\in\bar{\mathcal{M}}_2,k\neq i}|\alpha_k|\right) + O(m^{-1/6})$, we can get:

$$\frac{1}{|\bar{\mathcal{M}}_2|}\sum_{i\in\bar{\mathcal{M}}_2}\left(|\langle\alpha,x_1^{(i)}\rangle| + |\alpha_i| + \frac{1}{d}\widetilde{O}\left(\sum_{k\in\bar{\mathcal{M}}_2,k\neq i}|\alpha_k|\right)\right) \geq \frac{1}{2} \tag{E.61}$$

Since $Np \leq d$, this implies that

$$\frac{1}{|\bar{\mathcal{M}}_2|}\sum_{i\in\bar{\mathcal{M}}_2}\left(|\langle\alpha,x_1^{(i)}\rangle| + \widetilde{O}\left(|\alpha_i|\right)\right) \geq \frac{1}{2} \tag{E.62}$$

Thus, either $\frac{1}{|\bar{\mathcal{M}}_2|}\sum_{i\in\bar{\mathcal{M}}_2}|\langle\alpha,x_1^{(i)}\rangle| \geq \frac{1}{4}$, which implies that

$$\left\|(x_1^{(i)})_{i\in\bar{\mathcal{M}}_2}\alpha\right\|_2^2 = \sum_{i\in\bar{\mathcal{M}}_2}|\langle\alpha,x_1^{(i)}\rangle|^2 \geq \frac{|\bar{\mathcal{M}}_2|}{16} \tag{E.63}$$

Since w.h.p., $\|(x_1^{(i)})_{i\in\bar{\mathcal{M}}_2}\|_2 \leq O(1)$, we know that $\|\alpha\|_2 = \tilde{\Omega}(\sqrt{|\bar{\mathcal{M}}_2|}) = \tilde{\Omega}(\sqrt{Np})$.

The other possibility is that $\sum_{i\in\bar{\mathcal{M}}_2}\widetilde{O}\left(|\alpha_i|\right) \geq |\bar{\mathcal{M}}_2|/4$, which also implies that $\|\alpha\|_2 = \tilde{\Omega}(\sqrt{|\bar{\mathcal{M}}_2|}) = \tilde{\Omega}(\sqrt{Np})$ from Cauchy-Schwarz.

We now ready to conclude the proof: for randomly chosen x_1, it holds that

$$N_{W_0}(w,\overline{W}_t,x_1) \tag{E.64}$$

$$= \frac{1}{m}\sum_{j\in[m]}\sum_{k\in\bar{\mathcal{M}}_2}\alpha_k\langle x_1^{(k)},x_1\rangle\mathbb{1}([W_0]_j x_1^{(k)})\mathbb{1}([W_0]_j x_1) \pm \|\overline{W}_t' x_1\|_2 \tag{E.65}$$

$$= \frac{1}{m}\sum_{j\in[m]}\sum_{k\in\bar{\mathcal{M}}_2}\alpha_k\langle x_1^{(k)},x_1\rangle\mathbb{1}([W_0]_j x_1^{(k)})\mathbb{1}([W_0]_j x_1) \pm \widetilde{O}\left(\frac{\|\overline{W}_t'\|_F}{\sqrt{d}}\right) \tag{E.66}$$

$$= \frac{1}{m}\sum_{j\in[m]}\sum_{k\in\bar{\mathcal{M}}_2}\alpha_k\langle x_1^{(k)},x_1\rangle\mathbb{1}([W_0]_j x_1^{(k)})\mathbb{1}([W_0]_j x_1) \pm \widetilde{O}\left(\varepsilon_3\right) \tag{E.67}$$

Now using the same expansion of $\mathfrak{U}$ as before gives

$$\mathfrak{U}(x_1) := \frac{1}{m}\sum_{j\in[m]}\sum_{k\in\bar{\mathcal{M}}_2}\alpha_k\langle x_1^{(k)},x_1\rangle\mathbb{1}([W_0]_j x_1^{(k)})\mathbb{1}([W_0]_j x_1) \tag{E.68}$$

$$= \sum_{k\in\bar{\mathcal{M}}_2}\alpha_k\langle x_1^{(k)},x_1\rangle\frac{\arccos(\Theta(x_1^{(k)},x_1))}{2\pi} \pm O(m^{-1/6}) \tag{E.69}$$

Now we note that as the nonzero degrees in the polynomial expansion of arccos are all odd, we have

$$\mathfrak{U}(x_1) - \mathfrak{U}(-x_1) = 2\langle \alpha, x_1 \rangle \pm O(m^{-1/6}) \tag{E.70}$$

The end result is that by Lemma B.3, it will hold that:

$$\begin{aligned} r_t(x_1) &= N_{W_0}(w, \overline{W}_t, x_1) \pm O\left(\frac{1}{\lambda} \times \left(\varepsilon_s + \sqrt{\frac{\eta_2}{\eta_1}}\right) + \tau_0 \log d\right) \\ &= \mathfrak{U}(x_1) \pm \widetilde{O}\left(\varepsilon_3\right) \qquad \text{(by our choice of parameters)} \end{aligned} \tag{E.71}$$

This implies that

$$r_t(x_1) - r_t(-x_1) = 2\langle \alpha, x_1 \rangle \pm \widetilde{O}\left(\varepsilon_3\right) \tag{E.72}$$

□

F General case

F.1 Mitigation strategy

Instead of using large learning rate and annealing to a small learning rate, the regularization effect also exists if we use a small learning rate (η_2) and large pre-activation noise and then decay the noise. Hence the update is given as:

$$U_{t+1} = U_t - \eta_2 \nabla_U (\widehat{L}_\lambda(u, U_t) + \xi_t) \tag{F.1}$$

where $\xi_t \sim N(0, \tau_\xi^2 I_{m\times m} \otimes I_{d\times d})$. However, the output of the network is given as:

$$f_t(x) = u^\top \left(\mathbb{1}(U_t x + \Xi_t) \odot (U_t x + \Xi_t)\right) \tag{F.2}$$

Here $\Xi_t \sim \mathcal{N}(0, \tau_t^2 I_{m\times m})$ is a (freshly random) gaussian variable at each iteration.

The following theorem holds:

Theorem F.1 (General case)**.** *The same conclusion as in Theorem 3.4 holds if we first use noise level $\tau_t = \tau_0$ and then anneal to $\tau_t = 0$ after $\widetilde{O}\left(\frac{d}{\eta_1 \varepsilon_1}\right)$ iterations.*

F.2 Extension to two layer convolution network

We are also able to extend our results to convolutional networks. We consider a convolution network with $\frac{m}{k}$ channels, patch size d and stride d/k for some $k \le d$. Thus, the i-th patch consists of input $x_{(i)} = (x_{(i-1)d/k+1}, \cdots, x_{(i-1)d/k+d})$. Hence for $u \in \mathbb{R}^m, U \in \mathbb{R}^{\frac{m}{k}\times d}$, where $u = (u_1, \cdots, u_k)$ for each $u_i \in \mathbb{R}^{\frac{m}{k}}$, the network is given as:

$$N_U(u, U; x) = \sum_{i\in[k]} u_i^\top [U x_{(i)}]_+ \tag{F.3}$$

For every $A \in \mathbb{R}^{\frac{m}{k}\times d}$, we also use the notation

$$N_A(u, U; x) = \sum_{i\in[k]} u_i^\top \mathbb{1}(A x_{(i)}) U x_{(i)} \tag{F.4}$$

$$N_A(u_i, U; x) = u_i^\top \mathbb{1}(A x_{(i)}) U x_{(i)} \tag{F.5}$$

We make a simplifying assumption that z, ζ are only supported on the last d/k coordinates. The main theorem can be stated as the follows:

Theorem F.2 (General case)**.** *The same conclusions as in Theorem 3.4 and Theorem 3.5 hold if we replace the value of r by r/k and d by dk in both the theorem and in Assumption 3.3.*

Following the notation, we still denote

$$g_t(x) = g_t(x_{(k)}) = N_{U_t}(u, U_t; (0, x_{(k)})) \tag{F.6}$$

$$r_t(x) = r_t(x_{(1)}) = N_{U_t}(u, U_t; (x_{(1)}, 0)) \tag{F.7}$$

We use this definition so that $N_{U_t}(u, U_t; x) = g_t(x) + r_t(x)$ for every $t \geq 0$.

We denote $u = (u_1, \cdots, u_k)$ for the weight of the second layer associated with each convolution.

The main difference between the convolution setting and the simple case is that there is only one hidden weight that is shared across channels. However, since the output layers of these channels have different weights, we can disentangle these channels and think of them as updating "separately", which is given as the following two lemmas.

Lemma F.3 (disentangle convolution 1)**.** *For every fixed $x \in \mathbb{R}^{2d}$ and matrices $U_1, \cdots, U_k : \mathbb{R}^{\frac{m}{k} \times d}$ that can depend on $\widetilde{U}_t$ but not depend on u, with each $\|U_i\|_F \leq O\left(\frac{1}{\lambda}\right)$, we have w.h.p. over the randomness of $u, \widetilde{U}_t$:*

$$\left| N_{U_t}(u, \sum_{i\in[k]} u_i \odot U_i; x) - \sum_{i\in[k]} N_{U_t}(u_i, u_i \odot U_i; x) \right| \leq \widetilde{O}\left(k^2 \frac{\|x\|_2}{\lambda m^{1/2}} + k\varepsilon_s \|x\|_2 \right) \tag{F.8}$$

Here $u_i \odot U_i = ((u_i)_j (U_i)_j)_{j\in[\frac{m}{k}]}$.

Lemma F.4 (disentangle convolution 2)**.** *For every s, t, w.h.p. over the randomness of $u, \widetilde{U}_t, \widetilde{U}_s$, every $i, i' \in [k]$ with $i \neq i'$, and every $x, x' \in \mathbb{R}^d$, if we define $U_i = u_i \odot \mathbb{1}([U_s]x')x'^\top$, then as long as $\|\overline{U}_s\|_F, \|\overline{U}_t\|_F = O\left(\frac{1}{\lambda}\right)$, the following holds:*

$$|N_{U_t}(u_{i'}, U_i; x)| \leq \widetilde{O}\left(\frac{d^2 \|x\|_2 \|x'\|_2}{m^{1/2}} + \|x'\|_2 \|x\|_2 \sqrt{\varepsilon_s} + \|x\|_2 \varepsilon_s \right) \tag{F.9}$$

To apply this lemma, we can see that $u_i \odot \mathbb{1}([U_s]x')x'^\top$ is (a scaling of) the gradient coming from channel i on input x' at iteration s. This lemma says that it will have negligible effect on the output of channel $i' \neq i$ for (any) later iterations t. Hence at each iteration, every channel is updating almost separately.

Proof of Lemma F.3. By Lemma A.8, we know that

$$\left| N_{U_t}(u, \sum_{i\in[k]} u_i \odot U_i; x) - \sum_{i\in[k]} N_{U_t}(u, u_i \odot U_i; x) \right| \tag{F.10}$$

$$\leq \left| N_{\widetilde{U}_t}(u, \sum_{i\in[k]} u_i \odot U_i; x) - \sum_{i\in[k]} N_{\widetilde{U}_t}(u_i, u_i \odot U_i; x) \right| + O\left(k\varepsilon_s \|x\|_2\right) \tag{F.11}$$

Now, we can directly decompose

$$N_{\widetilde{U}_t}(u, \sum_{i\in[k]} u_i \odot U_i; x) = \sum_{i\in[k]} N_{\widetilde{U}_t}(u_i, u_i \odot U_i; x) \tag{F.12}$$

$$+ \sum_{i\in[k]} \sum_{i'\in[k], i'\neq i} N_{\widetilde{U}_t}(u_{i'}, u_i \odot U_i; x) \tag{F.13}$$

Since U_i does not depend on the randomness of $u_{i'}$ but only $\widetilde{U}_t$, fixing $\widetilde{U}_t, U_i$ we know that since each entry of $u_{i'}$ i.i.d. mean zero, we have:

$$\mathbb{E}_{u_{i'}} \left[N_{\widetilde{U}_t}(u_{i'}, u_i \odot U_i; x) \right] = 0 \tag{F.14}$$

Applying basic concentration bounds on $N_{\widetilde{U}_t}(u_{i'}, u_i \odot U_i; x)$, it holds that w.h.p. $|N_{\widetilde{U}_t}(u_{i'}, u_i \odot U_i; x)| \leq \widetilde{O}\left(\frac{\|x\|_2}{\lambda m}\right)$. Putting this back into Eq (F.12), we complete the proof.

□

Proof of Lemma F.4. By Lemma A.8, we know that

$$|N_{U_t}(u_{i'}, U_i; x)| \le \left|N_{\widetilde{U}_t}(u_{i'}, U_i; x)\right| + O(\varepsilon_s) \tag{F.15}$$

Hence, by definition, we have that

$$N_{\widetilde{U}_t}(u_{i'}, U_i; x) = N_{\widetilde{U}_t}(u_{i'}, u_i \odot \mathbb{1}([U_s]x')x'^\top; x) \tag{F.16}$$

Again by Lemma A.8, we know that $\|\mathbb{1}([U_s]) - \mathbb{1}(\widetilde{U}_s)\|_1 \le \varepsilon_s m$, hence we have since the absolute value of each entry of u_i is $m^{-1/2}$:

$$\left|N_{\widetilde{U}_t}(u_{i'}, u_i \odot \mathbb{1}([U_s]x')x'^\top; x)\right| \le \left|N_{\widetilde{U}_t}(u_{i'}, u_i \odot \mathbb{1}([\widetilde{U}_s]x')x'^\top; x)\right| + \|x'\|_2\|x\|_2\sqrt{\varepsilon_s} \tag{F.17}$$

Now for fixed x', x, for $\left|N_{\widetilde{U}_t}(u_{i'}, u_i \odot \mathbb{1}([\widetilde{U}_s]x')x'^\top; x)\right|$, since $\mathbb{1}([\widetilde{U}_s]x')x'^\top$ does not depend on the randomness of $u_{i'}$, following the previous lemma we can show that with probability at least $1 - e^{-d^2}$, $\left|N_{\widetilde{U}_t}(u_{i'}, u_i \odot \mathbb{1}([\widetilde{U}_s]x')x'^\top; x)\right| \le \widetilde{O}\left(\frac{\|x\|_2\|x'\|_2 d^2}{\lambda m}\right)$. Now, taking union bound over an epsilon-net of $x', x \in \mathbb{R}^d$ we conclude that for every x, x', w.h.p. $\left|N_{\widetilde{U}_t}(u_{i'}, u_i \odot \mathbb{1}([\widetilde{U}_s]x')x'^\top; x)\right| \le \widetilde{O}\left(\frac{\|x\|_2\|x'\|_2 d^2}{\lambda m}\right)$. Putting this back to Eq (F.17) we complete the proof. □

We set $\varepsilon_c = \widetilde{O}\left(kd^4\frac{1}{\lambda m^{1/2}}\right)$, and with this lemma, we can restate Lemma B.1, Lemma D.8 and Lemma E.1 in the following way: Suppose $\varepsilon_c \le \min\{\varepsilon_1/10, \varepsilon_2'/10\}$ for every x in the training set. Then the following lemmas hold by directly applying Lemma F.3.

Corollary F.5. *In the setting of Theorem F.2, there exists a solution $U^\star$ satisfying a) $\|U^\star\|_F^2 \le O(dk\log^2(1/\varepsilon))$ and b) for every $t \ge 0$:*

$$K_t(U^\star) \le q\log 2 + \epsilon_1/2 \tag{F.18}$$

Corollary F.6. *In the setting of Theorem F.2, there exists a solution U^* satisfying $\|U^*\|_F^2 = \widetilde{O}\left(\frac{k}{\varepsilon_1^2 r}\right)$ and for every $t \le \frac{1}{\eta_2\lambda}$:*

$$K_{t_0+t}(\overline{U}_{t_0} + U^*) \le \varepsilon_0 + \varepsilon_1 \tag{F.19}$$

Corollary F.7. *In the setting of Theorem F.2, there exists a solution $U^\star$ satisfying a) $\|U^\star\|_F^2 \le \widetilde{O}\left(\frac{k}{\varepsilon_2'^2 r} + Npk\right)$ and b) for every $t \le \frac{1}{\eta_2\lambda}$,*

$$K_t(U^\star) \le \varepsilon_2' \tag{F.20}$$

To prove these Lemmas, we can simply define $U^* = \sqrt{k}W^* + \sqrt{k}V^*$ for W^*, V^* given in the original proof and apply Lemma F.3. The reason we need k here is because there are $\frac{m}{k}$ channels instead of m, so the square norm scales up by a factor of k.

Now the next two convergence theorems follow directly from Lemma 4.1 and Lemma 4.3 and apply with initial learning rate η_1.

Corollary F.8. *In the setting of Theorem F.2 with initial learning rate η_1, at some step $t_0 \le \widetilde{O}\left(\frac{dk}{\eta_1\varepsilon_1}\right)$, the training loss $\widehat{L}(u, U_{t_0})$ becomes smaller than $q\log 2 + \epsilon_1$. Moreover, we have $\|\overline{U}_{t_0}\|_F^2 = O\left(dk\log^2(1/\varepsilon_1)\right)$.*

Corollary F.9. *In the setting of Theorem F.2, with initial learning rate η_1, there exists $t = \widetilde{O}\left(\frac{k}{\varepsilon_1^3\eta_2 r}\right)$, such that after $t_0 + t$ iterations we have that*

$$L_{t_0+t} = O\left(\sqrt{\varepsilon_1/q}\right) \tag{F.21}$$

Moreover, $\|\overline{U}_{t_0+t} - \overline{U}_{t_0}\|_F^2 \le \widetilde{O}\left(\frac{k}{\varepsilon_1^2 r}\right)$

The following statement applies when we use a small initial learning rate and follows from the proof of Lemma 5.1.

Corollary F.10. *In the setting of Theorem F.2, with initial learning rate η_2, there exists t with*

$$t = \widetilde{O}\left(\frac{k}{\eta_2 \varepsilon_2'^3 r} + \frac{Npk}{\eta_2 \varepsilon_2'}\right) \tag{F.22}$$

such that $L_t \le \varepsilon_2'$ after t iterations. Moreover, we have that $\|\overline{U}_t\|_F^2 \le \widetilde{O}\left(\frac{k}{\varepsilon_2'^2 r} + Npk\right)$

Now, the following lemma directly adapts from Lemma 4.2 by applying Lemma F.4:

Lemma F.11. *In the setting of Theorem F.2 with initial learning rate η_1, w.h.p., for every $t \le \frac{1}{\eta_1 \lambda}$,*

$$|g_t(z+\zeta) + g_t(z-\zeta) - 2g_t(z)| \le \widetilde{O}\left(\frac{r^2}{\lambda}\right) \tag{F.23}$$

With these lemmas, we can directly conclude the following:

Corollary F.12. *In the setting of Lemma F.9 with initial learning rate η_1, the following holds:*

$$\widehat{L}_{\mathcal{M}_1}(r_{t_0+t}) = O(\sqrt{\varepsilon_1/q}) \tag{F.24}$$

$$\widehat{L}_{\bar{\mathcal{M}}_1}(g_{t_0+t}) = O(\sqrt{\varepsilon_1/q^3}) \tag{F.25}$$

Corollary F.13. *In the setting with initial learning rate η_2, for every $\varepsilon_3 > 0$ such that $\frac{1}{\varepsilon_2'^8 \varepsilon_3^8 r} \le \varepsilon_2' dk$, there exists $\alpha \in \mathbb{R}^d$ such that $\alpha \in span\{x_1^{(i),(j)}\}_{i\in\bar{\mathcal{M}}_2, j\in[k]}$ and $\alpha = \tilde{\Omega}(\sqrt{Np})$ such that w.h.p. over a randomly chosen $x_1 \sim \mathcal{N}(0, I/d)$, we have that*

$$r_t(x_1) - r_t(-x_1) = 2\langle \alpha, x_1\rangle \pm \widetilde{O}\left(\varepsilon_3 + \frac{Npk}{d^{3/2}}\right) \tag{F.26}$$

Here $x_1^{(i),(j)} = ([x_1^{(i)}]_s)_{s\in\{(j-1)d/k+1,(j-1)d/k+2,\cdots,d\}}$

The final proof of Theorem F.2 follows directly from the proof of Theorem 3.4 and Theorem 3.5.

G Toolbox

Lemma G.1. *Let $X_1, X_2 \sim \mathcal{N}(0,1)$ and $a, b > 0$ such that $a^2 + b^2 = 1$. Then for every $\gamma_1, \gamma_2 \in \mathbb{R}$, we have that*

$$|\Pr[X_1 \ge \gamma_1 \mid aX_1 + bX_2 = \gamma_2] - \Pr[X_1 \ge \gamma_1 \mid aX_1 + bX_2 = 0]| \lesssim \frac{a|\gamma_2|}{b} \tag{G.1}$$

$$\Pr[|X_1| \le \gamma_1 \mid aX_1 + bX_2 = \gamma_2] \lesssim \frac{|\gamma_1|}{b} \tag{G.2}$$

Proof of Lemma G.1. Without loss of generality, we assume $a\gamma_2/b \ge 0$. Let $Y_1 = aX_1 + bX_2$ and $Y_2 = bX_1 - aX_2$. We have that Y_1, Y_2 are independent random Gaussian variables with marginal distribution $\mathcal{N}(0,1)$. Moreover, $X_1 = aY_1 + bY_2$. Thus, $X_1 \mid aX_1 + bX_2 = \gamma_2$ is the same as $aY_1 + bY_2 \mid Y_1 = \gamma_2$, which has distribution $\mathcal{N}(a\gamma_2, b^2)$. Let Z be a standard Gaussian, then

$$\begin{aligned}
&|\Pr[X_1 \ge \gamma_1 \mid aX_1 + bX_2 = \gamma_2] - \Pr[X_1 \ge \gamma_1 \mid aX_1 + bX_2 = 0]| \\
&= |\Pr[bZ + a\gamma_2 \ge \gamma_1] - \Pr[bZ \ge \gamma_1]| = \left|\Pr\left[\frac{\gamma_1}{b} \ge Z \ge \frac{\gamma_1}{b} - \frac{a\gamma_2}{b}\right]\right| \\
&\lesssim \left|\frac{a\gamma_2}{b}\right| \qquad \text{(beacuse the density of } \mathcal{N}(0,1) \text{ is bounded by } O(1))
\end{aligned}$$

Moreover,

$$\Pr[|X_1| \le \gamma_1 \mid aX_1 + bX_2 = \gamma_2] = \Pr[|bZ + a\gamma_2| \le \gamma_1] \lesssim |\gamma_1|/b \tag{G.3}$$

□

Table 1: Validation accuracies for WideResNet16 trained and tested on original CIFAR-10 images without data augmentation.

Method	Val. Acc
Large LR + anneal	90.41%
Small LR + noise	89.65%
Small LR	84.93%

Lemma G.2. *Let $M = M_0 + M_1$ where $M_1 \in \mathbb{R}^{d,d'}$ with $d' \leq d$ is a matrix with each entry i.i.d. $\mathcal{N}(0, 1/d)$ and $M_0 = w^\star \beta^\top$ where $\|\beta\|_2 \leq 1$ can depend on M_1. Then for every vector $z \in \mathbb{R}^{d'}$ we have that:*

$$\frac{\langle w^\star, Mz\rangle}{\|Mz\|_2} \leq 0.9 \tag{G.4}$$

Proof of Lemma G.2. Note that $Mz = w^\star\langle\beta, z\rangle + M_1 z$. Since M_1 is a random gaussian matrix and $d' \leq d$, we know that w.h.p. for every z we have $\frac{\langle w^\star M_1 z\rangle}{\|M_1 z\|_2} \leq \frac{\sqrt{2}}{2}$.

This implies that

$$\begin{aligned}
\|Mz\|_2^2 &= |\langle\beta, z\rangle|^2 + \|M_1 z\|_2^2 + 2\langle\beta, z\rangle\langle w^\star, M_1 z\rangle && \text{(G.5)}\\
&\geq |\langle\beta, z\rangle|^2 + \langle w^\star M_1 z\rangle^2 + 2\langle\beta, z\rangle\langle w^\star, M_1 z\rangle + \frac{1}{2}\|M_1 z\|_2^2 && \text{(G.6)}\\
&= (\langle\beta, z\rangle + \langle w^\star, M_1 z\rangle)^2 + \frac{1}{2}\|M_1 z\|_2^2 && \text{(G.7)}\\
&= \langle w^\star, Mz\rangle^2 + \frac{1}{2}\|M_1 z\|_2^2 && \text{(G.8)}
\end{aligned}$$

This completes the proof.

□

H Additional Details for Experiments

In this section we provide additional details on the experimental results of Section 6. All of our models were trained using a single NVIDIA TitanXp GPU and our code is implemented via PyTorch. We note that for all our experiments, the mean pixel is subtracted from the CIFAR image and then the image is divided by the standard deviation pixel. We use mean and standard deviation values in the PyTorch WideResNet implementation: https://github.com/xternalz/WideResNet-pytorch.

H.1 Additional Details for Noise Mitigation Strategy

In this section, we provide additional details for the mitigation strategy for a small learning rate described in Section 6. In Table 1, we demonstrate on CIFAR-10 images without data augmentation that this regularization can indeed counteract the negative effects of small learning rate, as we report a 4.72% increase in validation accuracy when adding noise to a small learning rate.

We train for all models for 200 epochs, annealing the learning rates by a factor of 0.2 at the 60th, 120th, and 150th epoch for all models. The large learning rate model uses an initial learning rate of 0.1, whereas the small learning rate model uses initial learning rate of 0.01. The large learning rate is a standard hyperparameter setting for the WideResNet16 architecture, and we chose the small learning rate by scaling this value down. The other hyperparameter settings are standard. We remove data augmentation from the training set to isolate the effect of adding noise.

We add noise before every time we apply the relu activation. As it is costly to add i.i.d. noise that is the size of the entire hidden layer, we sample Gaussian noise that has shape equal to the last two dimensions of the 4 dimensional hidden layer, where the first two dimensions are batch size and number of channels, and duplicate this over the first 2 dimensions. We sample different noise for every batch.

Figure 4: Visualizations of CIFAR-10 images with patches added.

Our annealing schedule simply multiplies the noise level by a constant factor at every iteration. We tune the standard deviation of the noise to 0.2 and the annealing rate to 0.995 every iteration. We show results from a single trial as the small LR with noise algorithm already shows substantial improvement over vanilla small LR.

H.2 Additional Details on Patch-Augmented CIFAR-10

We first describe in greater detail our method for producing the patch. First, the split of our data is the following: of the 50000 CIFAR-10 training images, 10000 will contain no patch and 40000 will have a patch. We generate this split randomly before training and keep it fixed. During a single epoch, we iterate through all images, loading the 10000 clean images the same way each time. For the remaining 40000 examples, we use a patch-only image with probability 0.2 and a patch mixed with CIFAR image with probability 0.8. Thus, 20% of the updates are on clean images, 16% of updates are on patches only, and 64% of updates are on mixed images, but the actual split of the data is slightly different because of our implementation.

The patch will be located in the center of the image. We visualize the patches in Figure 4. We generate the patch as follows: before training begins, we sample a random vector z with i.i.d entries from $\mathcal{N}(0, \sigma_z^2)$ as well as $\zeta_i \sim [-\beta, \beta]$ for classes $i = 1, \ldots, 10$. Then to generate patch-only images, we add a scalar multiple of ζ_i to z if the example belongs to class i. This scalar multiple is in the range $[-\alpha, \alpha]$ for some α we tune. We set coordinates not in the patch to 0. To generate images that contain both patch and a CIFAR example, we simply add $z \pm \zeta_i$. In all, the hyperparameters we tune are σ_z, β, α.

We must choose σ, β, α on the correct scale so that large and small learning rates don't both ignore the patch or overfit to the patch. For the experiment shown, $\sigma_z = 1.25, \beta = 0.1, \alpha = 1.75$.

Our large initial learning rate model trains with learning rate 0.1, annealing to 0.004 at the 30th epoch. and the small LR model trains with fixed learning rate 0.004. Our small LR with noise model trains with fixed learning rate 0.004, initial noise 0.4, and decays the noise to 4e-6 after the 30th epoch. We train all models for 60 epochs total, starting from the same dataset and choice of patches. Table 2 demonstrates the final validation accuracy numbers on patch-augmented and clean data.

Now we provide additional evidence that the generalization disparity is indeed due to the learning order effect and not simply because the large learning rate model can already generalize better on clean CIFAR-10 images. To see this, we consider the generalization error of models trained on 10000 clean CIFAR images: the small LR model achieves 65% validation accuracy, and the large LR model achieves 76% validation accuracy. For comparison, on the full clean dataset the small LR model achieves 83% validation accuracy whereas the large LR model achieves 90% accuracy.

We note that the final number of 69.89% clean image accuracy for the small LR model trained on the patch dataset is much closer to 65% than 83%, suggesting that it is indeed using a fraction of the available CIFAR samples because of learning order. On the other hand, the large LR model achieves final clean validation accuracy of 87.61% when trained on the patch dataset, which is very close to the 90% that is achievable training on the full clean dataset. This indicates that the large LR model is still using the majority of the images to learn CIFAR examples before annealing, as it has not yet memorized the patches.

Table 2: Validation accuracies for CIFAR-10 training dataset modified with patch. The mixed validation set similarly contains patches, but the clean set does not.

Method	Mixed Val. Acc.	Clean Val. Acc.
Large LR + anneal	95.35%	87.61%
Small LR	92.83%	69.89%
Small LR + noise	94.43%	81.36%