OPT2023: 15th Annual Workshop on Optimization for Machine Learning

On the Synergy Between Label Noise and Learning Rate Annealing in Neural Network Training

Stanley Wei STANLEY.WEI@PRINCETON.EDU
Princeton University

Tongzheng Ren TONGZHENG@UTEXAS.EDU
UT Austin

Simon S. Du SSDU@CS.WASHINGTON.EDU
University of Washington

Abstract

In the past decade, stochastic gradient descent (SGD) has emerged as one of the most dominant algorithms in neural network training, with enormous success in different application scenarios. However, the implicit bias of SGD with different training techniques is still under-explored. Some of the common heuristics in practice include 1) using large initial learning rates and decaying it as the training progresses, and 2) using mini-batch SGD instead of full-batch gradient descent. In this work, we show that under certain data distributions, these two techniques are *both* necessary to obtain good generalization on neural networks. We consider mini-batch SGD with label noise, and at the heart of our analysis lies the concept of feature learning order, which has previously been characterized theoretically by Li et al. [19] and Abbe et al. [1]. Notably, we use this to give the first concrete separations in generalization guarantees, between training neural networks with both label noise SGD and learning rate annealing and training with one of these elements removed.

1. Introduction

Despite the extreme over-parameterization used in real-world settings, neural networks trained via stochastic gradient descent (SGD) still exhibit remarkable generalization capabilities in practice. However, the exact characterization of the generalization power is still unclear, given that over-parameterized neural networks have a large amount of parameters that can perfectly interpolate the training set, even when it consists of random labels in the training data [31]. Recent work has demonstrated that the implicit bias due to the optimization algorithm is the key to understanding the regularization of such overparameterized models towards well-generalizing parameters [4, 9, 18]. In this paper, we focus on mini-batch SGD, one of the most universal optimization algorithms to train deep neural networks. Specifically, we investigate the impact of batch size and learning rate schedule on the generalization error.

Regarding the role of batch size, Wu et al. [28] Keskar et al. [14], and Smith et al. [24] demonstrate the importance of SGD's mini-batch noise as a source of regularization. Recently, several works have attempted to understand this implicit regularization from the lens of noise structure, especially label noise SGD training [6, 7, 20], i.e. when training labels are independently perturbed at each iteration (e.g. additive or multiplicative noise in regression, random label flipping in classification). Label noise SGD is an intriguing setting because it injects a parameter-dependent noise at each iteration of training that is empirically similar to the mini-batch noise [11], and therefore enables label noise

© S. Wei, T. Ren & S.S. Du.

to be a good approximation of the true mini-batch noise. Specifically, Damian et al. [7] show that label noise SGD will converge to a stationary point of a regularized loss function that penalizes sharp minima in the landscape as measured roughly by the trace of the Hessian. Li et al. [20] introduces a characterization of label noise training by describing the behavior of the algorithm around local minimizers. HaoChen et al. [11] and Vivien et al. [26] analyze the role of label noise SGD when training quadratically parameterized linear models, and show an implicit bias towards sparse features. However, these settings do not immediately have implications regarding neural network training dynamics and generalization ability. Moreover, in these works, explicit separations in generalization guarantees are not shown when label noise is removed from the training procedure. We therefore provide an analysis in the neural network setting that gives such separations for label noise training.

In terms of the regularization effect of learning rate in SGD training, conventional wisdom is to use large initial learning rates followed by annealing once the training loss stagnates [12, 15, 30]. Although the implicit bias of SGD with small learning rate is well understood [10, 25, 27], the large learning rate regime is still mostly unclear. In particular, small learning rate SGD leads to max-margin solutions for classification problems [25], and minimum norm solutions for regression problems [27]. Nevertheless, recent empirical studies [13, 16, 22] have demonstrated that models trained directly with small learning rate have poor generalization. Theoretically, several avenues of exposition have been proposed to interpret this phenomenon [5, 19, 21, 29]. In the convex least-squares linear regression setting, Nakkiran [22] and Wu et al. [29] show that large learning rate SGD biases towards certain convergence directions in the loss landscape, which lead to specific separations in generalization error. For nonconvex problems, Mohtashami et al. [21] analyze the large learning rate regime in the context of the optimization landscape, but they require certain assumptions on the frequency of sharp local minima. Li et al. [19] demonstrate that over certain data distributions, training neural networks with large initial learning provably helps, due to a bias towards a certain learning order of features, and they also provide a separation result between the two learning rate regimes. However, these two works only analyze the case of training with full-batch gradient descent (GD) rather than SGD.

In this work, we investigate the implicit bias of optimization using label noise SGD with learning rate annealing. Under a certain data distribution assumption introduced by [19], we demonstrate that the algorithm regularizes towards a certain feature learning order, and this bias provably leads to good generalization. On the contrary, removing either learning rate annealing *or* label noise will provably hurt generalization. To the best of our knowledge, this is the first result theoretically characterizing the specific benefits of employing *both* label noise structure and an annealed learning rate schedule for generalization in neural network training (a non-convex setting), as previous works have only considered the impact of label noise [6, 7, 11, 20] and learning rate annealing [5, 19, 22, 29] in isolation. Notably, our results allows us to obtain concrete separations in generalization error over our data distribution.

2. Preliminaries

Data Distribution Following the idea of Li et al. [19], we work with the data distribution defined as follows:

$$\begin{aligned} & y \sim \mathcal{U}(\{\pm 1\}) \\ \text{with probability } p_0 \quad & x_1 \sim \mathcal{P}_y \text{ and } x_2 = 0 \\ \text{with probability } q_0 \quad & x_1 = 0 \text{ and } x_2 \sim \mathcal{Q}_y \\ \text{with probability } 1 - p_0 - q_0 \quad & x_1 \sim \mathcal{P}_y \text{ and } x_2 \sim \mathcal{Q}_y \end{aligned}$$

Each observation consists of two d-dimensional features, one that has a low-complexity separating hyperplane (e.g. linear) yet requires high sample complexity to learn, and one that has a high-complexity separating hyperplane (e.g. nonlinear) yet requires low sample complexity to learn.

Here, the $\mathcal{P}$ component is the noisy yet linearly separable component. Essentially, it is composed of two half-Gaussians with a margin of $\gamma_0 = \frac{1}{\sqrt{d}}$, and $w^\star \in \mathbb{R}^d$ with $\|w^\star\|_2 = 1$:

$$\begin{aligned} x_1 \sim \mathcal{P}_1 &\iff x_1 = \gamma_0 w^\star + \beta | \langle w^\star, \beta \rangle \geq 0 \\ x_1 \sim \mathcal{P}_{-1} &\iff x_1 = -\gamma_0 w^\star + \beta | \langle w^\star, \beta \rangle \leq 0 \\ &\text{where } \beta \sim \mathcal{N}(0, I_d/d) \end{aligned}$$

On the other hand, the $\mathcal{Q}$ component is the noiseless, non-linearly separable component. Specifically, we define the $\mathcal{Q}$ component as:

$$\begin{aligned} x_2 \sim \mathcal{Q}_1 &\iff x_2 = \alpha z \\ x_2 \sim \mathcal{Q}_{-1} &\iff x_2 = \alpha(z + b\zeta) \\ &\text{where } \alpha \sim \mathcal{U}([0,1]), b \sim \mathcal{U}(\{\pm 1\}) \end{aligned}$$

Here, $z, \zeta \in \mathbb{R}^d$, $\|z\|_2 = 1$, with $\|\zeta\|_2 = r \ll 1$, $z^T \zeta = 0$. We refer the reader to Appendix A for an illustration of the data distribution.

Our setting consists of a dataset of size N, denoted $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$, where all the data points $(x^{(i)}, y^{(i)}) \sim \mathcal{D}$ are i.i.d. samples. Throughout the paper, we will be interested in several subsets of the N data points. We define the following sets:

$$\mathcal{M}_1 = \{i \in [N] : x_1^{(i)} \neq 0\}, \quad \mathcal{M}_2 = \{i \in [N] : x_2^{(i)} \neq 0\}$$

The natural complements $\bar{\mathcal{M}}_1$ and $\bar{\mathcal{M}}_2$ will be defined with respect to $[N]$. We also define the empirical proportions of the data points using p, q so that $p = \frac{|\bar{\mathcal{M}}_2|}{N}$ and $q = \frac{|\bar{\mathcal{M}}_1|}{N}$.

Parameterization We use two layer neural networks to predict the label y with the observation x; formally, the neural network is defined as $f(u, U; x) = u^T \sigma(Ux)$. Here $u \in \mathbb{R}^m$, $U \in \mathbb{R}^{m \times 2d}$. σ is the ReLU activation (i.e. $\sigma(x) = \max\{x, 0\}$).

Loss Function The empirical risk of the binary classification problem is defined as $\hat{L}(u, U) = \frac{1}{N} \sum_{i \in [N]} \ell(f(u, U; \cdot); (x^{(i)}, y^{(i)}))$, where $\ell(f; (x, y)) \triangleq \ell(yf(x)) = \log(1 + e^{-yf(x)})$ denotes the logistic loss. For ease of presentation, we also denote $\hat{L}_S(u, U) = \frac{1}{|S|} \sum_{i \in S} \ell(f(u, U; \cdot); (x^{(i)}, y^{(i)}))$, where S is a subset of our dataset. Finally, we define the expected risk (test loss) to be $L(u, U) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(f(w, U; \cdot); (x, y))]$.

Additional Notations We use $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to denote the asymptotic order up to logarithmic terms. We also define for a matrix for a d-column matrix P the value $[P]_i \in \mathbb{R}^{1\times d}$ to be the ith row of P.

3. Main Results

In this section, we state our main results on training a two layer neural network using label noise SGD with annealing learning rate and illustrate the necessity of both of these techniques.

3.1. Training Algorithms

To show the necessity of both training techniques, we consider the analysis of the following three algorithms. For each of our algorithms, we initialize the weight matrix U at U_0 so that every entry is i.i.d. with the variance τ_0^2. We also initialize the last layer u to have entries that are i.i.d. from $\mathcal{U}(\{\pm\frac{1}{\sqrt{m}}\})$. We fix u and only update U throughout the training. The algorithms are as follows; α_t denotes the learning rate at iteration t of the training process.

Label noise SGD + learning rate annealing (LNSGD-LS): We train using *label noise SGD* with flipping probability δ, using large initial learning rate of η_1, until the training loss $\hat{L}(U)$ reaches $q\log 2+\epsilon_1$. Then, we *anneal* to small learning rate η_2, and continue to train with label noise SGD until the training loss $\hat{L}(U)$ reaches $\epsilon_2 = \sqrt{\frac{\epsilon_1}{q}}$. Formally, we have an update rule of:

$$
\begin{aligned}
U_{t+1} &= U_t - \alpha_t \nabla_U \ell(\sigma_t y^{(i_t)} f_t(x^{(i_t)})) \\
\text{where } \sigma_t &= \begin{cases} 1 & \text{with probability } 1-\delta \\ -1 & \text{with probability } \delta \end{cases} \\
i_t &\sim \mathcal{U}([N])
\end{aligned}
$$

Label noise SGD + no annealing (LNSGD-S): We train using label noise SGD with flipping probability δ, using small learning rate η_2 *throughout* the training process, until the training loss $\hat{L}(U)$ reaches $\epsilon_2' = O(\epsilon_2)$. Formally, we have an update rule of:

$$
\begin{aligned}
U_{t+1} &= U_t - \alpha_t \nabla_U \ell(\sigma_t y^{(i_t)} f_t(x^{(i_t)})) \\
\text{where } \sigma_t &= \begin{cases} 1 & \text{with probability } 1-\delta \\ -1 & \text{with probability } \delta \end{cases} \\
i_t &\sim \mathcal{U}([N])
\end{aligned}
$$

Full batch GD + annealing (FBGD-LS): We train using *full batch gradient descent (without label noise)* using an initial learning rate of η_1, until the training loss $\hat{L}(U)$ reaches $q\log 2+\epsilon_1$. Then, we *anneal* the learning rate to η_2, and continue to train until the training loss $\hat{L}(U)$ reaches $\epsilon_2 = \sqrt{\frac{\epsilon_1}{q}}$. Formally, we have an update rule of:

$$
U_{t+1} = U_t - \alpha_t \nabla_U \hat{L}(u, U)
$$

Algorithm LNSGD-LS applies both the label noise SGD and learning rate annealing, while Algorithm LNSGD-S and Algorithm FBGD-LS apply only one of label noise SGD and learning rate annealing, respectively. In particular, the latter two algorithms allow us to analyze in isolation the respective individual influences of label noise SGD and learning rate annealing, while the former precisely enables us to analyze the synergy between the two elements in the context of both optimization and generalization.

3.2. Assumptions

We now formally introduce the assumptions used in our analysis.

Assumption 1 (Overparameterization) *We assume the scale of initialization $\tau_0 = \frac{1}{poly(d)}$, label flipping probability $\delta = \frac{1}{poly(d)}$, and hidden-layer width $m \geq poly(d)$.*

As with real-world models, we assume sufficient overparameterization exists in our model; this is also the standard assumption in the literature [2, 8].

Assumption 2 (Data Generation) *We define $\kappa^2 = \frac{d}{N}$, and assume that $\kappa \ll 1$. In particular, we will take d, N to tend towards infinity. In addition, we set the length of Q's orthogonal component $r = d^{-3/4}$, and the data distribution's component-wise probabilities $p_0 = \frac{\kappa^2}{2}$, and $q_0 = \Theta(1)$.*

We remark that by taking N sufficiently large, we will have that $p \approx p_0$ and $q \approx q_0$; as such, we will proceed in our analysis using p and q.

Assumption 3 (Hyperparameters) *We assume $\epsilon_1 \in (d^{-1/8}, \kappa^2 p^2 q^3)$. In addition, we assume $\eta_1 = O(\epsilon_1)$. Moreover, to isolate the influence of annealing, we will consider $\eta_2 = o(\eta_1)$.*

The upper bound assumption on the large initial learning rate η_1 is not unique to our setting; if the learning rate is too large, the value of the loss function will diverge, consistent with classical optimization literature.

3.3. Main Theorems

We now give our main results, corresponding to the three algorithms we use. Proof outlines and intuitions for the theorems are provided in Appendix *D*, and full proofs are given in the subsequent sections.

Theorem 4 (LNSGD-LS) *With high probability over the randomness of initialization and the minibatch sampling in SGD, the classification and test errors at the end of training are both $\tilde{O}(p^{3/2})$; in other words, $L(u, U) = \tilde{O}(p^{3/2})$.*

Theorem 5 (LNSGD-S) *With high probability over the randomness of initialization and the minibatch sampling in SGD, the classification and test errors are both $\Omega(p)$; in other words, $L(u, U) = \tilde{\Omega}(p)$.*

Theorem 6 (FBGD-LS) *With high probability over the randomness of initialization, the classification and test errors are both $\Omega(p)$; in other words, $L(u, U) = \tilde{\Omega}(p)$.*

Note that these theorems imply a separation, because of the fact that $p < 1$. As we will see in our training dynamics, each of these algorithms regularize towards a specific learning order of features. Our analysis will reveal that this has a direct influence on each of the above generalization errors.

4. Discussion and Future Work

In this paper, we give the first concrete separation results for neural network generalization in the presence of an annealed learning rate schedule and/or the presence of label noise in the training process. For an overparameterized one-hidden-layer neural network model, we reveal an interesting picture on the synergy between these two elements, through their respective implicit biases towards a certain learning order. Some exciting directions for future work include giving general distributions and algorithms to analyze the regularization of feature learning order, and recovering broader classes of generalization guarantees related to large initial learning rate and/or label noise SGD.

References

[1] Emmanuel Abbe, Enric Boix-Adserà, Matthew Stewart Brennan, Guy Bresler, and Dheeraj Mysore Nagaraj. The staircase property: How hierarchical structure can guide deep learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *CoRR*, abs/1811.04918, 2018.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. *Implicit Regularization in Deep Matrix Factorization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[5] Gaspard Beugnot, Julien Mairal, and Alessandro Rudi. On the benefits of large learning rates for kernel methods. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 254–282. PMLR, 02–05 Jul 2022.

[6] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 483–513. PMLR, 09–12 Jul 2020.

[7] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

[8] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[9] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[11] Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2315–2357. PMLR, 15–19 Aug 2021.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[13] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho*, and Krzysztof Geras*. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1g87C4KwB.

[14] 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. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[16] Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *CoRR*, abs/2002.10376, 2020. URL https://arxiv.org/abs/2002.10376.

[17] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2–47. PMLR, 06–09 Jul 2018.

[19] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022.

[21] Amirkeivan Mohtashami, Martin Jaggi, and Sebastian Stich. On avoiding local minima using gradient descent with large learning rates, 2022. URL https://arxiv.org/abs/2205.15142.

[22] Preetum Nakkiran. Learning rate annealing can provably help generalization, even for convex problems. *CoRR*, abs/2005.07360, 2020.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[24] Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9058–9067. PMLR, 13–18 Jul 2020.

[25] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

[26] Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2127–2159. PMLR, 02–05 Jul 2022.

[27] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020.

[28] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd, 2019.

[29] Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *International Conference on Learning Representations*, 2021.

[30] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87.

[31] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Appendix A. Data Distribution

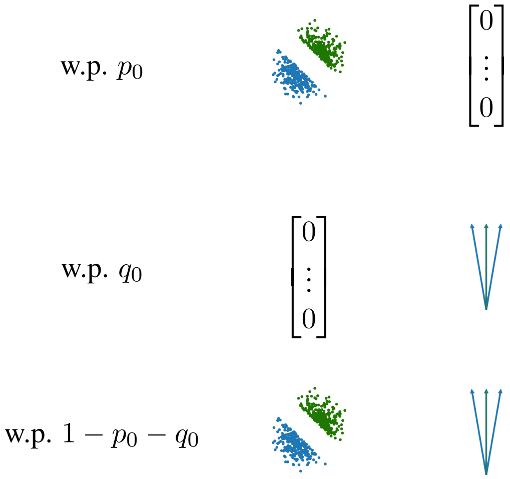
w.p. p_0
$\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$
w.p. q_0
$\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$
w.p. $1 - p_0 - q_0$

Figure 1: Here, the green points represent positive labels, and the blue points represent negative labels.

In the above Figure 1, we include an illustration of the data distribution from Section 2. To reiterate, we follow the idea of Li et al. [19], using the following definition of the data distribution:

$$\begin{aligned} & y \sim \mathcal{U}(\{\pm 1\}) \\ \text{with probability } p_0 \quad & x_1 \sim \mathcal{P}_y \text{ and } x_2 = 0 \\ \text{with probability } q_0 \quad & x_1 = 0 \text{ and } x_2 \sim \mathcal{Q}_y \\ \text{with probability } 1 - p_0 - q_0 \quad & x_1 \sim \mathcal{P}_y \text{ and } x_2 \sim \mathcal{Q}_y \end{aligned}$$

In the figure, the left column represents is $\mathcal{P}$ component, and the right column represents the $\mathcal{Q}$ component. The $\mathcal{P}$ component is defined as:

$$\begin{aligned} x_1 \sim \mathcal{P}_1 & \iff x_1 = \gamma_0 w^\star + \beta | \langle w^\star, \beta \rangle \geq 0 \\ x_1 \sim \mathcal{P}_{-1} & \iff x_1 = -\gamma_0 w^\star + \beta | \langle w^\star, \beta \rangle \leq 0 \\ & \text{where } \beta \sim \mathcal{N}(0, I_d/d) \end{aligned}$$

where $\gamma_0 = \frac{1}{\sqrt{d}}$, and $w^\star \in \mathbb{R}^d$ with $\|w^\star\|_2 = 1$. The $\mathcal{Q}$ component is defined as:

$$\begin{aligned} x_2 \sim \mathcal{Q}_1 & \iff x_2 = \alpha z \\ x_2 \sim \mathcal{Q}_{-1} & \iff x_2 = \alpha(z + b\zeta) \\ & \text{where } \alpha \sim \mathcal{U}([0,1]), b \sim \mathcal{U}(\{\pm 1\}) \end{aligned}$$

where, $z, \zeta \in \mathbb{R}^d$, $\|z\|_2 = 1$, with $\|\zeta\|_2 = r \ll 1$, $z^T \zeta = 0$.

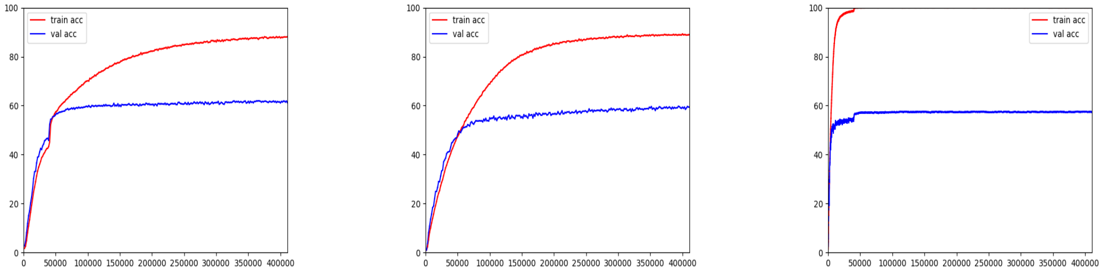

Figure 2: We empirically demonstrate the effect both label noise and learning rate annealing on neural networks training on CIFAR100. The horizontal axis represents the iteration number, and the vertical axis represents the classification accuracy in percentage out of 100. a) Label noise SGD with learning rate annealing (LNSGD-LS) achieves a final validation accuracy of 61.96%. b) Label noise SGD without learning rate annealing (LNSGD-S) achieves a final validation accuracy of 59.40%. c) Large batch GD with learning rate annealing (FBGD-LS) achieves a final validation accuracy of 57.33%.

Appendix B. Experiments

We empirically verify the separations derived from our theoretical analysis of the three algorithms. In particular, we train a VGG19 network architecture [23] on the CIFAR100 dataset using our three algorithms, LNSGD-LS, LNSGD-S, and FBGD-LS. We adapt the experimental setup of HaoChen et al. [11], where label noise means changing the label of an individual image to another random label that is not the true label.

Our label noise SGD algorithms use a fixed flipping probability of 0.1 (even after annealing, if applicable), and a small batch size of 32. Our large batch GD algorithm uses a large batch size of 256. In addition, if the algorithm uses annealing, then our learning rate schedule is 0.01 followed by a single decay to 0.001 at anneal time; otherwise, the learning rate is a constant 0.001 throughout training. Furthermore, we set the annealing time (if applicable) to be at iteration 40000 of training; for all three algorithms, we train for 400000 iterations. Here, we define an iteration as a minibatch; see Figure 2. In the end, label noise SGD achieves the highest validation accuracy, therefore backing up our theoretical analysis. Furthermore, we can see that LNSGD-S and FBGD-LS seem to have quicker convergence in training error, as predicted by our theoretical analysis as well.

Appendix C. Additional Notations

We follow the notation from Li et al. [19], and denote $\mathbb{1}(w)$ for some vector w to be the element-wise indicator function vector defined as $(\mathbb{1}(w))_i = \mathbb{1}(w_i \geq 0)$. Then, we can define $N_A(u, U; x) \triangleq u^T(\mathbb{1}(Ax) \odot Ux)$, where $\odot$ represents the element-wise product between two vectors (or matrices).[1] We decompose the weight of the first layer as $U = \begin{pmatrix} W \\ V \end{pmatrix}$. Here, W only operates on the first d coordinates of an observation x (i.e. the x_1 component of the data point), and V only operates on the last d coordinates of an observation x (i.e. the x_2 component of the data point). For simplicity,

[1] In particular, we have $N_U(u, U; x) = f(u, U; x)$.

we denote W to be both a matrix in $\mathbb{R}^{m \times 2d}$ with the last d columns 0, or a matrix in $\mathbb{R}^{m \times d}$. The notations for V are treated similarly. As such, we can note that:

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

Furthermore, in any training algorithm, we will denote $U_t = \begin{pmatrix} W_t \\ V_t \end{pmatrix}$ to be the weight matrix during training at time t, with similar abuse of notation as above for W_t and V_t. For convenience, we will now define the decomposition of the network prediction into two parts, each corresponding to a component of the data it operates on:

$$\begin{aligned} r_t(x) = r_t(x_1) &\triangleq N_{W_t}(w, W_t; x) = N_{W_t}(w, W_t; x_1) \\ g_t(x) = g_t(x_2) &\triangleq N_{V_t}(v, V_t; x) = N_{V_t}(v, V_t; x_2) \end{aligned}$$

We will also denote $f(u, U_t; (x^{(i)}))$ as $f_t(x^{(i)})$.

Appendix D. Outline of Algorithm Analyses

In this section, we provide the high-level details of the algorithms' analyses. We first start out with a core tool that we use throughout the proofs. Subsequently, we will outline the characterizations for each of the three algorithms. Proof sketches and full proofs are available in the appendix.

D.1. Tools for Label Noise Analysis

Note that the gradient update rule in FBGD-LS is the standard definition of full-batch gradient descent, and hence there are no stochasticity in the training process, except due to initialization. On the other hand, towards reasoning about the label noise update rule in LNSGD-LS and LNSGD-S, we detail an *iterate decoupling* procedure.

First, we define the *expected smoothed loss* to be $\bar{\ell}(z) \triangleq \delta\ell(-z) + (1-\delta)\ell(z)$. As such, the update rule of label noise SGD will be:

$$\begin{aligned} U_{t+1} &= U_t - \alpha_t \nabla_U \ell(\sigma_t y^{(i_t)} f_t(x^{(i_t)})) \\ &= U_t - \alpha_t \nabla_U \bar{\ell}(y^{(i_t)} f_t(x^{(i_t)})) - \underbrace{\alpha_t \epsilon_t y^{(i_t)} \nabla_U f_t(x^{(i_t)})}_{\text{label noise}} \end{aligned}$$

Here, $i_t \sim \mathcal{U}([N])$, and ϵ_t is the random variable defined as:

$$\epsilon_t = \begin{cases} \delta\Gamma & \text{w.p. } 1-\delta \\ -(1-\delta)\Gamma & \text{w.p. } \delta \end{cases} \qquad \text{where } \Gamma \triangleq \ell'(y^{(i_t)} f_t(x^{(i_t)})) + \ell'(-y^{(i_t)} f_t(x^{(i_t)}))$$

Now, consider the following decoupling of the $U_t = \bar{U}_t + \tilde{U}_t$ at time t of training.

$$\begin{aligned} \bar{U}_t &= -\sum_{s=1}^{t} \alpha_{s-1} \nabla_U \left(\bar{\ell}(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \right) \\ \tilde{U}_t &= U_0 - \sum_{s=1}^{t} \alpha_{s-1} \epsilon_{s-1} y^{(i_{s-1})} \cdot \nabla_U f_{s-1}(x^{(i_{s-1})}) \end{aligned}$$

In essence, $\bar{U}_t$ and $\tilde{U}_t$ represent the accumulated signal and accumulated noise at time t. Such signal-noise decoupling is useful, because the weights of the hidden layer at each iteration is a random variable that depends on the historical trajectory. In particular, this means that the accumulated label noise $\tilde{U}_t$ is a martingale, and we can tools from martingale concentration to understand the dynamics of this term at each phase of the training. Likewise, the accumulated signal $\bar{U}_t$ allows us to analyze how much the algorithm actually "learns" on each component (in terms of signal); the scale of this quantity will influence the margins of the separating hyperplanes in our analysis, which we use to show the generalization guarantees all of our algorithms.

D.2. Analysis of Theorem 4 (LNSGD-LS)

We break the analysis of LNSGD-LS into two phases: before annealing and after annealing. For the first phase, pre-annealing, we show that the large learning rate induces a large accumulated noise term in V, and therefore, the activation patterns are too noisy to learn the $\mathcal{Q}$ component. At such, when the loss reaches $q \log 2 + \epsilon_1$, essentially all $(1-q)N$ data points from $\mathcal{M}_1$ have been learned well. Following the learning rate annealing, the remaining qN data points in $\bar{\mathcal{M}}_1$ will be memorized, and once a loss of ϵ_2 is reached, the model will have low training error on both the $\mathcal{P}$ and $\mathcal{Q}$ components. Low training error on the $\mathcal{Q}$ component will then immediately imply good generalization, and low training error on the $\mathcal{P}$ component means that Rademacher complexity bounds can control the generalization error to $\tilde{O}\left(p\sqrt{\frac{d}{N}}\right) = \tilde{O}(p\kappa) = \tilde{O}(p^{3/2})$, since a p fraction of data points contain only noisy d-dimensional features from $\mathcal{P}$.

D.2.1. PHASE 1: LEARNING RATE η_1

We first give the following lemma, which bounds with high probability the time it takes to reach the target loss and start annealing.

Lemma 7 *With high probability, we will anneal at some time $\hat{t}_1 \leq \tilde{O}\left(\frac{d}{\eta_1 \epsilon_1}\right)$. Furthermore, the training loss at the time will satisfy $\hat{L}(U_{\hat{t}_1}) \leq q \log 2 + \epsilon_1$.*

Afterwards, we show that at anneal time, the margin on the $\mathcal{Q}$ component is still poor. Specifically, we have the following lemma to show that $\mathcal{Q}$ is not learned well; by our hyperparameter choices, it holds that $\tilde{O}\left(\frac{d}{\eta_1 \epsilon_1}\right) \leq \tilde{O}\left(\frac{d^{9/8}}{\eta_1}\right)$.

Lemma 8 *With high probability, for all time steps $t \leq \tilde{O}\left(\frac{d^{9/8}}{\eta_1}\right)$, it will hold that*

$$|g_t(z+\zeta) + g_t(z-\zeta) - 2g_t(z)| \leq O(r^2 d^{9/8} \sqrt{\log d})$$

D.2.2. PHASE 2: LEARNING RATE η_2

After annealing at time $\hat{t}_1$ as defined in Lemma 7, we show that with high probability, the loss is small when we stop training (i.e. at time $\hat{t}_1 + \hat{t}_2$).

Lemma 9 *Using the $\hat{t}_1$ from Lemma 7, it holds with high probability that for some $\hat{t}_2 \leq \tilde{O}\left(\frac{1}{\eta_2 \epsilon_2^3 r}\right)$ the training loss at time $\hat{t}_1 + \hat{t}_2$ will satisfy $\hat{L}(U_{\hat{t}_1 + \hat{t}_2}) \leq O\left(\sqrt{\frac{\epsilon_1}{q}}\right)$.*

We can further decompose the loss when we stop the training into $\mathcal{M}_1$ and $\bar{\mathcal{M}}_1$, to get the following.

Lemma 10 *Using $\hat{t}_1$ and $\hat{t}_2$ defined in the Lemma 7 and Lemma 9 respectively, it holds with high probability that $\hat{L}_{\mathcal{M}_1}(r_{\hat{t}_1+\hat{t}_2}) \le O\left(\sqrt{\frac{\epsilon_1}{q}}\right)$ and $\hat{L}_{\bar{\mathcal{M}}_1}(g_{\hat{t}_1+\hat{t}_2}) \le O\left(\sqrt{\frac{\epsilon_1}{q^3}}\right)$.*

By decomposing the final loss into the $\mathcal{M}_1$ and $\bar{\mathcal{M}}_1$ components, we can see that low loss on the $g_{\hat{t}_1+\hat{t}_2}$ component of the network implies good generalization directly, since the $\mathcal{Q}$ component of the distribution has no noise. Furthermore, low loss on the by $r_{\hat{t}_1+\hat{t}_2}$ network component implies low test error by Rademacher complexity generalization bounds in Allen-Zhu et al. [2], hence proving Theorem 4.

D.3. Analysis of Theorem 5 (LNSGD-S)

Towards proving Theorem 5, we will show that the algorithm converges to small training error very fast, due to the memorization of the $\mathcal{Q}$ component under small learning rate. We will then show that as $\mathcal{Q}$ is in the process of being memorized, the $\mathcal{P}$ component does not receive a lot of signal from data in $\mathcal{M}_1 \cap \mathcal{M}_2$, since the $\mathcal{Q}$ feature is being used to learn the labels in this set. Consequently, the weights in W must fit to the pN points in $\bar{\mathcal{M}}_2$. Since $pN \le \frac{d}{2}$ there is insufficient sample complexity to learn the component. This leads to the lower bound in classification error of $\Omega(p)$, as a constant fraction of data points in $\bar{\mathcal{M}}_2$ cannot be predicted correctly.

D.3.1. PHASE 1: MEMORIZING $\mathcal{Q}$

We bound the time it takes for the algorithm to converge. In particular, we first show that the $\mathcal{Q}$ component is memorized very quickly.

Lemma 11 *With high probability, there will be a time $\hat{t}_1 \le \tilde{O}\left(\frac{1}{\eta_2(\epsilon_2')^3 r}\right)$ such that the loss of $\mathcal{M}_2$ at time $\hat{t}_1$ satisfies $\hat{L}_{\mathcal{M}_2}(U_{\hat{t}_1}) \le \epsilon_2'$.*

This above lemma bounds the time it takes to converge to a solution that already achieves low loss on the $\mathcal{M}_2$ portion of the dataset, i.e. all the data points with the $\mathcal{Q}$ component as a feature is mostly classified correctly.

D.3.2. PHASE 2: AFTER MEMORIZING $\mathcal{Q}$

We then proceed to bound the additional time it takes to reach the stopping criterion loss of ϵ_2'.

Lemma 12 *Using $\hat{t}_1$ defined in Lemma 11, we have that with high probability, there will be a time $\hat{t}_2 \le \tilde{O}\left(\frac{pN}{\eta_2 \epsilon_2'}\right)$ so that the total loss at that time satisfies $\hat{L}(U_{\hat{t}_1+\hat{t}_2}) \le \epsilon_2'$.*

We show that when we stop the training process, the accumulated gradient on W is still small, when restricted to the subset of data points $\mathcal{M}_2$.

Lemma 13 *Define $\bar{W}_t^{(2)} = -\eta_2 \sum_{s=1}^{t} \nabla_W \left(\bar{\ell}(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})}))\right) \cdot \mathbb{1}(i_{s-1} \in \mathcal{M}_2)$. Then, with high probability, for $t \le \tilde{O}\left(\frac{d}{\eta_2 \epsilon_2'}\right)$, it holds that that $\|\bar{W}_t^{(2)}\|_F \le \tilde{O}\left(\frac{d^{31/64}}{{\epsilon_2'}^2}\right)$.*

Since the gradient signal is small for the $\mathcal{P}$ component on the data points from $\mathcal{M}_2$, this intuitively implies that the weights W must be mostly learned from the examples in $\bar{\mathcal{M}}_2$. On the other hand, since $|\bar{\mathcal{M}}_2| = pN \leq \frac{d}{2}$ by Assumption 2, there is not enough sample complexity to learn $\mathcal{P}$, since it is a d-dimensional distribution. This leads to the following lemma, which states that the margin on $\mathcal{P}$ remains small.

Lemma 14 *Using $\hat{t}_1$ and $\hat{t}_2$ defined in the Lemma 11 and Lemma 12 respectively, we have that at time $\hat{t}_1 + \hat{t}_2$, there will exist an $\alpha \in span(\{x_1^{(i)}\}_{i \in \bar{\mathcal{M}}_2})$ satisfying $\|\alpha\|_2 \geq \Omega(\sqrt{pN})$, such that with high probability over $x_1 \sim \mathcal{P}$,*

$$r_{\hat{t}_1+\hat{t}_2}(x_1) - r_{\hat{t}_1+\hat{t}_2}(-x_1) = 2\alpha^T x_1 \pm \tilde{O}\left(\frac{1}{d^{1/64}(\epsilon_2')^2}\right)$$

In other words, the margin is on the weights W is still very poor, which means the Gaussian noise from the $\mathcal{P}$ component will be very significant. This therefore leads to the the lower bound in generalization error in Theorem 5.

D.4. Analysis of Theorem 6 (FBGD-LS)

In this case where noise is completely removed from the training process, even with the same annealing schedule as LNSGD-LS, the model will attempt to quickly memorize the examples for the $\mathcal{Q}$ component first, because the activation patterns are not influenced by noise as in the case of LNSGD-LS, and therefore progress is being made in learning $\mathcal{Q}$. As a result, the training loss will quickly fall below the annealing criterion $q \log 2 + \epsilon_1$. Even though neither the $\mathcal{P}$ component nor the $\mathcal{Q}$ component is yet fully learned, after annealing, the $\mathcal{Q}$ component will be quickly memorized. This once again leads to the underfitting effect of $\mathcal{P}$ as in LNSGD-S, and hence the classification error lower bound of $\Omega(p)$. Here, we unravel the impact of learning rate annealing in the absence of label noise: a transition of the training process from a feature exploration stage of all features, to a feature memorization stage of one specific feature.

D.4.1. Phase 1: Exploring Both Features Before Annealing

For the first phase of training, we will argue that the anneal time is reached very quickly.

Lemma 15 *With high probability, there will be a time $\hat{t}_1 \leq \tilde{O}\left(\frac{1}{\eta_1 r}\right)$ such that $\hat{L}_{\mathcal{M}_2}(U_{\hat{t}_1}) \leq q \log 2 + 2\epsilon_1$.*

This previous lemma bounds the time it takes to converge to a *partial* signal for the $\mathcal{Q}$ component. The training is now close to the annealing condition; from this, we can further give an upper bound on the annealing time via a construction of a partial $\mathcal{P}$ signal, formalized via the following lemma.

Lemma 16 *Using $\hat{t}_1$ in Lemma 15, with high probability, there exists a $\hat{t}_2 \leq \tilde{O}\left(\frac{pN}{\eta_1}\right)$ such that $\hat{L}(U_{\hat{t}_1+\hat{t}_2}) \leq q \log 2 + \epsilon_1$.*

D.4.2. PHASE 2: MEMORIZING ONLY $\mathcal{Q}$ AFTER ANNEALING

After annealing, we show that with small learning rate and no label noise, the $\mathcal{Q}$ component will be quickly memorized. We then show that the accumulated signal on W is small on $\mathcal{M}_2$, which allows us to show that overall, the $\mathcal{P}$ component does not receive a lot of signal from the $\mathcal{M}_2$ portion of the dataset.

For convenience, let us denote $\hat{t} = \hat{t}_1 + \hat{t}_2$, where $\hat{t}_1$ and $\hat{t}_2$ are from Lemma 15 and Lemma 16, respectively. Then, the following lemma holds.

Lemma 17 *With high probability, there exists a time* $\hat{t}_3 \leq \tilde{O}\left(\frac{1}{\eta_2 \epsilon_2^3 r}\right)$ *such that* $\hat{L}_{\mathcal{M}_2}(U_{\hat{t}+\hat{t}_3}) \leq \epsilon_2$.

The above lemma bounds the time it takes for $\mathcal{Q}$ to be *fully* memorized. This is similar to the setting of LNSGD-S; even though there is label noise in LNSGD-S, the result that we end up with will be similar, because the analysis of LNSGD-S reveals that the label noise does not influence the activation patterns much.

Finally, we can bound the time until FBGD-LS achieves its stopping criterion.

Lemma 18 *Using* $\hat{t}_3$ *from Lemma 17, we have that with high probability, there exists a time* $\hat{t}_4 \leq \tilde{O}\left(\frac{pN}{\eta_2 \epsilon_2}\right)$ *such that* $\hat{L}(U_{\hat{t}+\hat{t}_3+\hat{t}_4}) \leq \epsilon_2$.

We once again argue that in this phase after annealing, the total accumulated signal is small as well. This is shown by the following bound.

Lemma 19 *Define* $\bar{W}_t^{(2)\prime} = -\eta_2 \sum_{s=1}^{\hat{t}+\hat{t}_3+t} \frac{1}{N} \sum_{i \in \mathcal{M}_2} \nabla_W \ell(y^{(i)} f_{s-1}(x^{(i)}))$. *Using* $\hat{t}_3$ *from Lemma 17, with high probability, for* $t \leq \tilde{O}\left(\frac{d}{\eta_2 \epsilon_2}\right)$, *it holds that that* $\|\bar{W}_t^{(2)\prime}\|_F \leq \tilde{O}\left(\frac{d^{31/64}}{\epsilon_2^2}\right)$.

D.4.3. GENERALIZATION

By Lemma 19, we get that the total signal on the $\mathcal{P}$ component provided by the data points in $\mathcal{M}_2$ is small. Hence, the $\mathcal{P}$ component must once again be mostly learned via the signal from the pN points in $\bar{\mathcal{M}}_2$, leading the insufficient sample complexity. In particular, this can be formalized by the following lemma.

Lemma 20 *Using* $\hat{t}_3$ *and* $\hat{t}_4$ *defined in the Lemma 17 and Lemma 18 respectively, we have that at time* $\hat{t} + \hat{t}_3 + \hat{t}_4$, *there will exist an* $\alpha \in span(\{x_1^{(i)}\}_{i \in \bar{\mathcal{M}}_2})$ *satisfying* $\|\alpha\|_2 \geq \Omega(\sqrt{pN})$, *such that with high probability over* $x_1 \sim \mathcal{P}_y$,

$$r_{\hat{t}+\hat{t}_3+\hat{t}_4}(x_1) - r_{\hat{t}+\hat{t}_3+\hat{t}_4}(-x_1) = 2\alpha^T x_1 \pm \tilde{O}\left(\frac{1}{d^{1/64}\epsilon_2^2}\right)$$

This last lemma implies that the margin on the predictions W is still very poor when training stops, and therefore even when training ends, the $\mathcal{P}$ component remains underfitted. This directly leads to the generalization error lower bound in Theorem 6.

Appendix E. Toolbox of Lemmas and Concentrations

E.1. Decoupling

Recall the following signal-noise decoupling $U_t = \bar{U}_t + \tilde{U}_t$ at time t of training from Appendix $D.1$; we give more detailed decomposition as follows.

$$\begin{aligned}\bar{U}_t &= -\sum_{s=1}^{t-1} \alpha_{s-1} \nabla_U \left(\bar{\ell}(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})}))\right) \\ &= -\sum_{s=1}^{t-1} \alpha_{s-1} \cdot \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})} B_{s-1} \\ \tilde{U}_t &= U_0 - \sum_{s=1}^{t-1} \alpha_{s-1} \epsilon_{s-1} y^{(i_{s-1})} \cdot B_{s-1}\end{aligned}$$

where we define

$$\begin{aligned}B_{s-1} &= \nabla_U f_{s-1}(x^{(i_{s-1})}) \\ &= \begin{pmatrix} -u_1 \mathbb{1}([U_{s-1}]_1 x) \cdot x^{(i_{s-1})} - \\ \vdots \\ -u_m \mathbb{1}([U_{s-1}]_m x) \cdot x^{(i_{s-1})} - \end{pmatrix}\end{aligned}$$

E.2. Lemmas and Concentrations

Proposition 21 *Recall the random variable ϵ_t in the label noise term. Then it holds that $\mathbb{E}[\epsilon_t] = 0$, and $\mathbb{V}[\epsilon_t] = \delta(1-\delta)\Gamma^2 = O(\delta)$.*

Lemma 22 *Let $[\nabla_U \hat{L}(U)]_i$ = denote the i-th row of $\nabla_U \hat{L}(U)$. Then, with high probability $\|[\nabla_U \hat{L}(U)]_i\|_2 \lesssim \frac{1}{\sqrt{m}}$. Furthermore, with high probabiilty, it also holds that $\|[\nabla_U \bar{\ell}(U)]_i\|_2 \lesssim \frac{1}{\sqrt{m}}$.*

Proof For the first part of the lemma, note that this holds because

$$\|[\nabla_U \hat{L}(U)]_i\|_2 = \hat{\mathbb{E}}[\ell'(f(u, U; (x, y)) \cdot u_i \mathbb{1}([U]_i x) x] \lesssim \frac{1}{\sqrt{m}}$$

This follows from the fact that $\ell' \in (-1, 0)$, and $\|x\|_2 = O(1)$ with high probability. For the second part of the lemma, note similarly that

$$\|[\nabla_U \bar{\ell}(U)]_i\|_2 = \|[\bar{\ell}'(f(u, U; (x, y)) \cdot u_i \mathbb{1}([U]_i x) x\|_2 \lesssim \frac{1}{\sqrt{m}}$$

■

Lemma 23 *For any time step t prior to annealing, we have that $\|[\bar{U}_t]_i\|_2 \lesssim \frac{\eta_1 t}{\sqrt{m}}$. The same holds for $\|[\tilde{U}_t]_i\|_2 \lesssim \frac{\eta_1 t}{\sqrt{m}}$.*

Proof This follows from writing out the summation, followed by an application of triangle inequality and Lemma 22. ■

Proposition 24 *With high probability, the matrix B_t satisfies*

$$\|B_t\|_2 \leq 1$$

Proof We have that with high probability, the spectral norm of $\|B_t\|_2$ satisfies

$$\|B_t\|_2 = \sqrt{\|B_t B_t^T\|_2} = \sqrt{\frac{1}{m}\|\mathbb{1}([U_t]x^{(i_t)})\|_1 \|x^{(i_t)}\|_2} \lesssim 1$$

■

Lemma 25 (Freedman's inequality for matrix martingales) *Consider a matrix martingale $\{Y_k : k = 0, 1, \dots\}$, where the $Y_k \in \mathbb{R}^{d_1 \times d_2}$, with difference sequence $\{X_k : k = 1, 2, \dots\}$; furthermore, suppose that $\sqrt{\|X_k X_k^T\|_2} \leq R$ almost surely for $k \geq 1$. Define the column and row quadratic variation, respectively, as follows, for $k \geq 1$:*

$$W_{col,k} \triangleq \sum_{j=1}^{k} \mathbb{E}[X_j X_j^T]$$
$$W_{row,k} \triangleq \sum_{j=1}^{k} \mathbb{E}[X_j^T X_j]$$

Then, for all $t \geq 0$ and $\sigma^2 > 0$, with probability at least $1 - (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right)$, at least one of the following holds for all $k \geq 0$

$$\|Y_k\| \leq t$$
$$\|W_{col,k}\| \geq \sigma^2$$
$$\|W_{row,k}\| \geq \sigma^2$$

Lemma 26 *For all real numbers $|z| = O(poly \log d)$, it holds that $|\ell(z) - \bar{\ell}(z)| \leq O(\frac{1}{poly(d)})$.*

Proof Note that

$$\begin{aligned} |\ell(z) - \bar{\ell}(z)| &= \delta|\ell(z) - \ell(-z)| \\ &\lesssim O\left(\frac{\text{poly}\log d}{\text{poly}(d)}\right) \\ &\lesssim O\left(\frac{1}{\text{poly}(d)}\right) \end{aligned}$$

where the first inequality follows from our Assumption 1. ■

This lemma will be very useful throughout our analysis, because by our data distribution, it implies that with high probability, $|L(U) - \bar{L}(U)| \leq O\left(\frac{1}{\text{poly}(d)}\right)$, where we define

$$\bar{L}(U) \triangleq \frac{1}{N} \sum_{i \in [N]} \bar{\ell}(y^{(i)} f(x^{(i)}))$$

Therefore, we will simply analyze $\bar{L}$ throughout our proofs, as the results for the original L will hold as well, due to the closeness. We also define

$$\bar{L}_t \triangleq \frac{1}{N} \sum_{i \in [N]} \bar{\ell}(y^{(i)} f_t(x^{(i)}))$$

Appendix F. Proof of Theorem 4 (LNSGD-LS)

F.1. Proof of Lemma 7

For the following section, we denote $t_1 = \tilde{O}(\frac{d}{\eta_1 \epsilon_1})$. The proof structure roughly entails a construction of a target signal, and demonstrating convergence to this target. We first give some concentration results derived from Freedman's inequality.

Lemma 27 *With high probability, it holds that for all time steps $s \leq t_1$ (in particular, this includes the time before annealing) that*

$$\left\| \sum_{j=1}^{s} \epsilon_{j-1} B_{j-1} \right\| \leq \tau_0 \sqrt{\log d} \cdot \frac{1}{\eta_1}$$

Proof This is an application of Freedman's inequality on the martingale defined by $\{\sum_{s=1}^{t} \epsilon_{s-1} B_{s-1}\}_{t \geq 0}$. In particular, we have that with high probability,

$$\sqrt{\|X_k X_k^T\|_2} = \sqrt{\|\epsilon_k^2 B_k B_k^T\|_2} \lesssim \delta$$

by Proposition 24. Thus, we can now choose $\sigma^2 = O(t_1 \delta)$ and $t = \frac{\tau_0 \sqrt{\log d}}{\eta_1}$ for Freedman's inequality, to give the high probability bound. In particular, this means that with high probability, Freedman's inequality tells us that because for times $s \leq t_1$, we have

$$\|W_{col,s}\| \lesssim O(s\delta) \lesssim O(t_1 \delta) = \sigma^2$$

it must hold with high probability that

$$\|Y_s\| \leq t = \frac{\tau_0 \sqrt{\log d}}{\eta_1}$$

By definition of Y_s here, we have the desired result. ■

Lemma 28 *With high probability, we have that*

$$\|\tilde{U}_{t_1}\|_2 = O(\tau_0 \sqrt{m})$$

Proof We know that

$$\begin{aligned}\|\tilde{U}_{t_1}\|_2 &\le \|U_0\|_2 + \eta_1 \|\sum_{s=1}^{t_1} \epsilon_{s-1} B_{s-1}\|_2 \\ &\lesssim \tau_0\sqrt{m} + \frac{\eta_1 \tau_0 \sqrt{\log d}}{\eta_1} \\ &\lesssim \tau_0\sqrt{m}\end{aligned}$$

Here, the second inequality follows from the Freedman's inequality guarantee we get in Lemma 27. ■

We first make use of a forward perturbation lemma from Allen-Zhu et al. [3], which tells us the following:

Lemma 29 *With high probability, it holds that for every $\tilde{U}$ satisfying $\|\tilde{U}\| \lesssim \omega$ that*

$$\|\mathbb{1}(U_{t_1}x) - \mathbb{1}(\tilde{U}_{t_1}x)\|_1 \lesssim m\omega^{2/3}$$

As a result, this gives us the bound of

$$|N_{U_{t_1}}(u, \bar{U}_{t_1}; x) - N_{\tilde{U}_{t_1}}(u, \bar{U}_{t_1}; x)| \lesssim (\tau_0\sqrt{m})^{2/3}\eta_1 t_1$$

Proof The first part of the lemma is given by Allen-Zhu et al. [3]. Consequently, the second part follows from the fact that

$$\begin{aligned}|N_{U_{t_1}}(u, \bar{U}_{t_1}; x) - N_{\tilde{U}_{t_1}}(u, \bar{U}_{t_1}; x)| &\le \frac{1}{\sqrt{m}} \sum_{i\in[m]} |\mathbb{1}([U_{t_1}]_i x) - \mathbb{1}([\tilde{U}_{t_1}]_i x)| \cdot |[\bar{U}_{t_1}]_i x| \\ &\lesssim \frac{1}{\sqrt{m}} \cdot m\omega^{2/3} \cdot \frac{\eta_1 t_1}{\sqrt{m}} \\ &\lesssim (\tau_0\sqrt{m})^{2/3}\eta_1 t_1\end{aligned}$$

where the last inequality follows from the choice of $\omega = O(\tau_0\sqrt{m})$ from Lemma 28. ■

We also have the following lemmas, which, as we see, will be useful later.

Lemma 30 *With high probability, it holds that*

$$|N_{U_{t_1}}(u, \tilde{U}_{t_1}; x) - N_{\tilde{U}_{t_1}}(u, \tilde{U}_{t_1}; x)| \lesssim (\tau_0\sqrt{m})^{2/3}\eta_1 t_1$$

Proof Similar to the analysis of the previous lemma, we have that

$$\begin{aligned}|N_{U_{t_1}}(u, \tilde{U}_{t_1}; x) - N_{\tilde{U}_{t_1}}(u, \tilde{U}_{t_1}; x)| &\le \frac{1}{\sqrt{m}} \sum_{i\in[m]} |\mathbb{1}([U_{t_1}]_i x) - \mathbb{1}([\tilde{U}_{t_1}]_i x)| \cdot |[\tilde{U}]_i x| \\ &\lesssim \frac{1}{\sqrt{m}} \cdot m(\tau_0\sqrt{m})^{2/3} \cdot \frac{\eta_1 t_1}{\sqrt{m}} \\ &\lesssim (\tau_0\sqrt{m})^{2/3}\eta_1 t_1\end{aligned}$$

■

Lemma 31 *With high probability, it holds that* $|N_{\tilde{U}_{t_1}}(u, \tilde{U}_{t_1}; x)| \lesssim \tau_0\sqrt{\log d}$.

Proof We have that with high probability

$$\begin{aligned}|N_{\tilde{U}_{t_1}}(u, \tilde{U}_{t_1}; x)| &= \sum_{i\in[m]} u_i\sigma([\tilde{U}_{t_1}]_i x)\\ &\lesssim \frac{1}{\sqrt{m}}\cdot\|\tilde{U}_{t_1}x\|_2\\ &\lesssim \frac{1}{\sqrt{m}}\cdot\tau_0\sqrt{m}\sqrt{\log d}\\ &\lesssim \tau_0\sqrt{\log d}\end{aligned}$$

The second inequality follows from the fact that with high probability, $\|\tilde{U}_{t_1}\|_2 \lesssim \tau_0\sqrt{m}$, which we showed in Lemma 28, and the $\sqrt{\log d}$ factor comes from standard concentration inequality. ■

We can now combine the previous two lemmas to obtain a bound on $|N_{U_{t_1}}(u, \tilde{U}_{t_1}; x)|$, using triangle inequality.

Lemma 32 *With high probability, it holds that*

$$|N_{U_{t_1}}(u, \tilde{U}_{t_1}; x)| \lesssim (\tau_0\sqrt{m})^{2/3}\eta_1 t_1 + \tau_0\sqrt{\log d}$$

We are now able to give a high probability guarantee of what the loss will be at our fixed annealing time t_1. This is captured by the following.

Lemma 33 *With high probability, we have that at some time* $\hat{t}_1 \lesssim t_1$,

$$\frac{1}{N}\sum_{i=1}^{N}\bar{\ell}(y^{(i)}f_{t_1}(x^{(x)})) \lesssim q\log 2 + \epsilon_1$$

First, let us define

$$\begin{aligned}h_t(B; x) &\triangleq N_{U_t}(u, B + \tilde{U}_t; x)\\ K_t(B) &\triangleq \frac{1}{N}\sum_{i=1}^{N}\bar{\ell}(h_t(B; \cdot); (x^{(i)}, y^{(i)}))\end{aligned}$$

These gadgets will faciliate the analysis by isolating the effects of the accumulated signal and noise.
Proof Note that we have $\|\mathbb{1}(U_{t_1}x) - \mathbb{1}(\tilde{U}_{t_1}x))\|_1 \lesssim m(\tau_0\sqrt{m})^{2/3}$ by Lemma 29. Furthermore, we also have that $|N_{U_{t_1}}(u, \tilde{U}_{t_1}; x)| \lesssim \tau_0\sqrt{\log d}$ from the previous lemma. From our choice of parameters that $\tau_0\sqrt{\log d} \le \frac{\epsilon_1}{20}$, we obtain that $|N_{U_{t_1}}(u, \tilde{U}_{t_1}; x)| \le \frac{\epsilon_1}{20}$.

We now consider a target signal $U^\star = \begin{pmatrix} W^\star \\ V^\star \end{pmatrix}$ of $V^\star = 0$ and $W^\star = 20w_i\sqrt{d}\log\frac{1}{\epsilon_1}$ following the analysis of Li et al. [19]. Consider this target signal at time t_1, combined with our noise $\tilde{U}_{t_1}$. Then, we will have that

$$\begin{aligned}h_{t_1}(U^\star) &= N_{U_{t_1}}(u, U^\star + \tilde{U}_{t_1}; x)\\ &= N_{U_{t_1}}(u, U^\star; x) + N_{U_{t_1}}(u, \tilde{U}_{t_1}; x)\end{aligned}$$

We first consider the first term $N_{U_{t_1}}(u, U^\star; x)$. In particular, from the definition of our constructed signal, we know that

$$\begin{aligned}
N_{U_{t_1}}(u, U^\star; x) &= 20w^{\star T}x_1 \cdot \sqrt{d}\log\frac{1}{\epsilon_1}\sum_{i\in[m]} w_i^2 \mathbb{1}([W_{t_1}]_i x_1) \\
&= 20w^{\star T}x_1 \cdot \sqrt{d}\log\frac{1}{\epsilon_1}\cdot\frac{\|\mathbb{1}(W_{t_1}x_1)\|_1}{m} \\
&= 20w^{\star T}x_1 \cdot \sqrt{d}\log\frac{1}{\epsilon_1}\cdot\frac{\|\mathbb{1}(\tilde{W}_{t_1}x_1)\|_1}{m} \pm O\left(\sqrt{d}\log d\cdot(\tau_0\sqrt{m})^{2/3}\right)
\end{aligned}$$

where the last equality follows with high probability from Lemma 29 and our choice of ϵ_1. Now, from the symmetry of the distribution of the martingale $\tilde{W}_{t_1}$, we know that with high probability,

$$\begin{aligned}
\frac{\|\mathbb{1}(\tilde{W}_{t_1}x_1)\|_1}{m} &= \frac{1}{2} \pm O\left(\sqrt{\frac{\log d}{m}}\right) \\
&= \frac{1}{2} \pm O(\epsilon_1)
\end{aligned}$$

where the last equality follows from our choice of parameters. Therefore, substituting this quantity back into the expression for $N_{U_{t_1}}(u, U^\star; x)$, we obtain that

$$|N_{U_{t_1}}(u, U^\star; x_1) - 10w^{\star T}x_1 \log\frac{1}{\epsilon_1}| \leq \frac{\epsilon_1}{20}$$

which furthermore implies, by triangle inequality, that

$$|N_{U_{t_1}}(u, U^\star + \tilde{U}_{t_1}; x_1) - 10w^{\star T}x_1 \log\frac{1}{\epsilon_1}| \leq \frac{\epsilon_1}{10}$$

Therefore, we have that with high probability,

$$\frac{1}{N}\sum_{i\in[N]} \bar{\ell}(y^{(i)}\cdot 10w^{\star T}x_1^{(i)}\log\frac{1}{\epsilon_1}) \leq q\log 2 + \frac{\epsilon_1}{10}$$

where the $q\log 2$ comes from the $\mathcal{Q}$ component that have $x_1 = 0$, and the $\epsilon/10$ comes from standard concentration over the randomness in $\mathcal{P}$. Therefore, since $\bar{\ell}$ is 1-Lipschitz, we obtain that

$$\bar{L}(U^\star + \tilde{U}_{t_1}) = K_t(U^\star) \leq q\log 2 + \frac{\epsilon_1}{2}$$

■

In particular, over the course of training, the function K is essentially a convex but changing function of the matrix B. This property will be useful in the following result, showing that the target in the previous theorem can be reached. We adapt a version of Li et al. [19]'s for stochastic gradient descent and gradient descent, and without dependence on weight decay.

Lemma 34 (Optimization via GD/SGD) *Consider a fixed differentiable convex function K, and a training process with gradient descent starting from z_0 with update rule*

$$z_{t+1} = z_t - \eta \nabla K(z_t)$$

Then, assuming that all the K's is L-Lipschitz, and there exists a common point $z^\star$ such that for all t, $K_t(z^\star) \leq c^\star$ for some fixed $c^\star$ and $z^\star$ satisfying $\|z^\star - z_0\|_2, \|z^\star\|_2 \leq R$, we get that for all $\mu > 0$ and $\frac{R^2}{T\mu} < \eta \leq \frac{\mu}{100}$ with high probability, there exists a $t^\star \in [T]$ satisfying

$$K_{t^\star}(z_{t^\star}) \leq c^\star + \mu$$

Furthermore, for all the $t \leq t^\star$, it holds that $\|z_t - z^\star\|_2 \leq R$.

As a corollary to the optimization lemma, we are able to apply the lemma with $R = O(d \log^2 \frac{1}{\epsilon_1})$, with $z^\star$ being the target signal constructed; this gives the exact convergence time in Lemma 7.

F.2. Proof of Lemma 8

In this section, we reload the notation on t_1 so that $t_1 = O\left(\frac{d^{9/8}}{\eta_1}\right)$. Our overall proof strategy will be to show that activations don't change much in this first phase on this $\mathcal{Q}$ component. In particular, we first show the result using $\tilde{V_t}$ as network activations, and then show that it is "close" to the true network (which uses $\tilde{V}$ directly as activations by definition).

Definition 35 *For a time step t and vector w, define $\mathcal{E}_t^w$ to be $\{i \in [m] : [\tilde{V_t}]_i w \geq 0\}$. In other words, it's the set of activated neurons in the V portion of the accumulated label noise. Define $\bar{\mathcal{E}}_t^w$ to be the nonactivated neurons.*

Definition 36 *For a set $\mathcal{E} \subset [m]$, denote $\mathbb{1}(\mathcal{E})$ to be the indicator vector.*

Definition 37 *Define $\tilde{g}_t(x) = N_{\tilde{V_t}}(v, \bar{V}_t; x)$.*

Lemma C.2 from Li et al. [19] still holds in our setting. Specifically, we have that:

Lemma 38 (Partition into terms) *Let $Q_t = diag(v)\bar{V}_t$. Then it holds that*

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

Proof This is by definition of $\mathbb{1}$ and Q_t. ■

In bounding the two terms from the above lemma, we first give the following notation. For the following section, let

$$B_{s-1} = \nabla_V N_{V_{s-1}}(v, V_{s-1}; x^{(i_{s-1})}) = \begin{pmatrix} -v_1 \mathbb{1}([V_{s-1}]_1 x) \cdot x^{(i_{s-1})} - \\ \vdots \\ -v_m \mathbb{1}([V_{s-1}]_m x) \cdot x^{(i_{s-1})} - \end{pmatrix}$$

From the above, we obtain the following result, which bounds the number of neurons that have different activations between $z + \zeta$ and $z - \zeta$.

Proposition 39 *For $i \in \mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}$ ($\oplus$ means symmetric difference of sets), we have that $|[\tilde{V}_{t_1}]_i \zeta| \lesssim \tau_0 r\sqrt{\log d}$. In addition, we also have that $|\mathcal{E}_{t_1}^{z-\zeta} \oplus \mathcal{E}_{t_1}^{z+\zeta}| \lesssim rm\sqrt{\log d}$.*

Proof Consider an $i \in \mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}$; note that this is equivalent to $|[\tilde{V}_t]_i z| \leq |[\tilde{V}_t]_i \zeta|$. We then note that at initialization, it holds that $\tilde{V}_0$ has iid $\mathcal{N}(0, \tau_0^2)$ entries. This implies that with high probability, $|[\tilde{V}_0]_i \zeta| \leq \tau_0 r\sqrt{\log d}$. Furthermore, since

$$\Pr\left[|[\tilde{V}_0]_i z| \leq |[\tilde{V}_0]_i \zeta|\right] \leq \Pr\left[|[\tilde{V}_0]_i z| \leq \tau_0 r\sqrt{\log d}\right] \lesssim r\sqrt{\log d} \tag{1}$$

we have by Bernstein's that with high probability, at most $rm\sqrt{\log d} + \log d \lesssim rm\sqrt{\log d}$ rows $i \in [m]$ of $\tilde{V}_0$ have its index in $\mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}$.

Now, if we assume that $\frac{\eta_1 t_1 \sqrt{\delta}}{\sqrt{m}} \lesssim \tau_0\sqrt{\log d}$, then it holds with high probability that

$$\begin{aligned}
\left|\sum_{s=1}^{t_1} \eta_1 \epsilon_{s-1}[B_{s-1}]_i \zeta\right| &\leq \eta_1 \sum_{s=1}^{t_1} |\epsilon_{s-1}[B_{s-1}]_i \zeta| \\
&\leq \frac{\eta_1 r}{\sqrt{m}} \sum_{s=1}^{t_1} |\epsilon_{s-1}| \\
&\lesssim \frac{\eta_1 r}{\sqrt{m}} t_1(\delta + \sqrt{\delta}) \\
&\lesssim \frac{\eta_1 r}{\sqrt{m}} t_1 \sqrt{\delta} \\
&\lesssim \tau_0 r\sqrt{\log d}
\end{aligned}$$

Therefore, we have that with high probability,

$$\begin{aligned}
|[\tilde{V}_{t_1}]_i \zeta| &\leq |[\tilde{V}_0]_i \zeta| + \left|\sum_{s=1}^{t_1} \eta_1 \epsilon_{s-1}[B_{s-1}]_i \zeta\right| \\
&\lesssim \tau_0 r\sqrt{\log d}
\end{aligned}$$

and thus the proposition follows. ■

To finish off the bound of the second term in Lemma 38, we have the following result.

Proposition 40 *We have that at time t_1, with high probability,*

$$\|(\mathbb{1}(\mathcal{E}_{t_1}^{z+\zeta}) - \mathbb{1}(\mathcal{E}_{t_1}^{z-\zeta}))^T Q_{t_1} \zeta\| \lesssim \eta_1 r^2 t_1 \sqrt{\log d}$$

Proof

$$\begin{aligned}
\|(\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^T Q_t \zeta\| &\leq \|(\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^T Q_t\| \|\zeta\| \\
&\leq |\mathcal{E}_t^{z-\zeta} \oplus \mathcal{E}_t^{z+\zeta}| \cdot \max_i \|[Q_t]_i\| \|\zeta\| \\
&\lesssim rm\sqrt{\log d} \cdot \frac{1}{\sqrt{m}} \cdot \frac{\eta_1 t_1}{\sqrt{m}} \cdot r \\
&\lesssim \eta_1 r^2 t_1 \sqrt{\log d}
\end{aligned}$$

■

Thus, we have bounded the second term in Lemma 38. We will now proceed to bound the first term.

Proposition 41 *Let $\Delta Q_t = diag(v)\nabla_V \bar{\ell}(y^{(i_t)} f_t(x^{(i_t)}))$, i.e. the change in Q_t gained in time step t. Then, we have that*

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

Proof We simply write out the terms. More specifically, we know that

$$\begin{aligned}\|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T Q_t\| &= \|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T \mathrm{diag}(v) \sum_{s=1}^{t} \eta \nabla_V \bar{\ell}(y^{(i_t)} f_t(x^{(i_t)}))\| \\ &= \|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T \sum_{s=1}^{t} \eta \Delta Q_s\|\end{aligned}$$

Thus, by triangle inequality, it holds that

$$\|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T Q_t\| \leq \eta \sum_{s=1}^{t} \|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T \Delta Q_s\|$$

The result then follows from $\|z\| = 1$. ■

We now attempt to bound $\|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T \Delta Q_s\|$. We give the following set definitions and proposition from Li et al. [19].

Definition 42 *Define the following sets*

$$\begin{aligned}\mathcal{F}_s^+ &= \{i \in [m] : [\tilde{V}_s]_i z \gtrsim \tau_0 r \sqrt{\log d}\} \\ \mathcal{F}_s^- &= \{i \in [m] : [\tilde{V}_s]_i z \lesssim \tau_0 r \sqrt{\log d}\} \\ \mathcal{F}_s^c &= \{i \in [m] : |[\tilde{V}_s]_i z| \lesssim \tau_0 r \sqrt{\log d}\} \\ A &= \mathcal{E}_t^{z+\zeta} \setminus \mathcal{E}_t^{z} \\ B &= \mathcal{E}_t^{z} \setminus \mathcal{E}_t^{z-\zeta}\end{aligned}$$

Proposition 43 *By definitions of the above sets, we have that*

$$\begin{aligned}&\|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T \Delta Q_s\|_2 \\ &\lesssim \frac{1}{m}\left(\left||A \cup \mathcal{F}_s^+| - |B \cup \mathcal{F}_s^+|\right| + \left||A \cup \mathcal{F}_s^-| - |B \cup \mathcal{F}_s^-|\right| + |A \cup \mathcal{F}_s^c| + |B \cup \mathcal{F}_s^c|\right)\end{aligned}$$

Using the above tools, we are able to show the bound lemma, and therefore obtain our desired bound.

Lemma 44 *With high probability, we have that*

$$\|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T \Delta Q_s\|_2 \lesssim \eta_1 r^2 t_1 \sqrt{\log d}$$

Proof By the Li et al. [19] decomposition from the previous proposition, we see that it suffices to bound the absolute value of

$$\begin{aligned}\frac{1}{m}(|A \cap \mathcal{F}_s^+| - |B \cap \mathcal{F}_s^+|) &= \frac{1}{m}\sum_{i\in[m]}[\mathbb{1}(i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^{z}, i \in \mathcal{F}_s^+) - \mathbb{1}(i \in \mathcal{E}_t^{z}, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+)]\\ &= \Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^{z}, i \in \mathcal{F}_s^+] - \Pr[i \in \mathcal{E}_t^{z}, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+]\end{aligned}$$

as well as

$$\frac{1}{m}|A \cap \mathcal{F}_s^c| = \Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^{z}, i \in \mathcal{F}_s^c]$$

We first note the following expression, by definition of $\tilde{V}$.

$$[\tilde{V}_t]_i z = [\tilde{V}_s]_i z - \sum_{j=s}^{t} \eta_1 \epsilon_{j-1}[B_{j-1}]_i z$$

We know that the second term on the right hand side is a martingale (and each term in the summation forms the martingale difference sequence), and therefore we can get the variance of $\sum_{j=s}^{t} \eta_1 \epsilon_{j-1}[B_{j-1}]_i z$ to be $\sigma_{s,t}^2 = O(\eta_1^2(t-s)\delta)$ (follows from the fact that $\|z\| = 1$).

In bounding the difference of cardinalities from above, we first, for the sake of convenience, define the random variables $Y_1 = [\tilde{V}_s]_i z$, $Y_2 = [\tilde{V}_t]_i z$, and $Y_3 = [\tilde{V}_t]_i \zeta$. We also define $Y_4 = \sum_{j=s}^{t} \eta_1 \epsilon_{j-1}[B_{j-1}]_i z = Y_2 - Y_1$. Then, the following holds:

$$\Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^{z}, i \in \mathcal{F}_s^+] = \Pr[Y_2 + Y_3 \geq 0, Y_2 \leq 0, Y_1 \geq O(\tau_0 r \sqrt{\log d})]$$

and

$$\begin{aligned}\Pr[i \in \mathcal{E}_t^{z}, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+] &= \Pr[Y_2 \geq 0, Y_2 - Y_3 \leq 0, Y_1 \geq O(\tau_0 r \sqrt{\log d})]\\ &= \Pr[Y_2 \leq 0, -Y_2 - Y_3 \leq 0, -Y_1 \geq O(\tau_0 r \sqrt{\log d})]\end{aligned}$$

where the second inequality in the latter relation follows from symmetric distribution, since we are dealing with martingales. From this, we obtain the following:

$$\begin{aligned}&\left|\Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^{z}, i \in \mathcal{F}_s^+] - \Pr[i \in \mathcal{E}_t^{z}, i \notin \mathcal{E}_t^{z-\zeta}, i \in \mathcal{F}_s^+]\right|\\ &= \mathbb{E}_{Y_2}\left[\mathbb{1}(Y_2 \leq 0) \cdot \Pr[Y_3 \geq -Y_2 \text{ and } |Y_1| \leq O(\tau_0 r \sqrt{\log d})|Y_2]\right]\\ &= \mathbb{E}_{Y_2}\left[\mathbb{1}(Y_2 \leq 0) \cdot \Pr[Y_3 \geq -Y_2 \text{ and } |Y_4 - Y_2| \leq O(\tau_0 r \sqrt{\log d})|Y_2]\right]\\ &\lesssim \mathbb{E}_{Y_2}\left[\mathbb{1}(Y_2 \leq 0) \cdot \Pr[Y_3 \geq -Y_2|Y_2] \cdot \frac{O(\tau_0 r \sqrt{\log d})}{\sigma_{s,t}}\right]\\ &\lesssim \mathbb{E}_{Y_2}\left[\mathbb{1}(Y_2 \leq 0) \cdot \exp(-Y_2^2/2(r^2\tau_0^2)) \cdot \frac{O(\tau_0 r \sqrt{\log d})}{\sigma_{s,t}}\right]\\ &\lesssim \frac{\tau_0 r \sqrt{\log d}}{\eta_1 \sqrt{(t-s)\delta}}\end{aligned}$$

which immediately implies that

$$|\frac{1}{m}(|A \cap \mathcal{F}_s^+| - |B \cap \mathcal{F}_s^+|)| \lesssim \frac{\tau_0 r\sqrt{\log d}}{\eta_1\sqrt{(t-s)\delta}}$$

By a similar argument, one can see that

$$|\frac{1}{m}(|A \cap \mathcal{F}_s^-| - |B \cap \mathcal{F}_s^-|)| \lesssim \frac{\tau_0 r\sqrt{\log d}}{\eta_1\sqrt{(t-s)\delta}}$$

Now, to bound $\frac{1}{m}|A \cap \mathcal{F}_s^c|$ (and similarly $\frac{1}{m}|B \cap \mathcal{F}_s^c|$), we note the following:

$$\begin{aligned}\Pr[i \in \mathcal{E}_t^{z+\zeta}, i \notin \mathcal{E}_t^z, i \in \mathcal{F}_s^c] &= \Pr[Y_2 + Y_3 \geq 0, Y_2 \leq 0, |Y_1| \leq O(\tau_0 r\sqrt{\log d})] \\ &= \mathbb{E}_{Y_2}\left[\mathbb{1}(Y_2 \leq 0) \cdot \Pr[Y_3 \geq -Y_2 \text{ and } |Y_4 - Y_2| \leq O(\tau_0 r\sqrt{\log d})|Y_2]\right] \\ &\lesssim \frac{\tau_0 r\sqrt{\log d}}{\eta_1\sqrt{(t-s)\delta}}\end{aligned}$$

where the last inequality follows since we already bounded an identical expression above. Therefore, it holds that

$$\frac{1}{m}|A \cap \mathcal{F}_s^c| \lesssim \frac{\tau_0 r\sqrt{\log d}}{\eta_1\sqrt{(t-s)\delta}}$$

We now see that

$$\begin{aligned}&\frac{1}{m}\left(\left||A \cup \mathcal{F}_s^+| - |B \cup \mathcal{F}_s^+|\right| + \left||A \cup \mathcal{F}_s^-| - |B \cup \mathcal{F}_s^-|\right| + |A \cup \mathcal{F}_s^c| + |B \cup \mathcal{F}_s^c|\right) \\ &\lesssim \frac{\tau_0 r\sqrt{\log d}}{\eta_1\sqrt{(t-s)\delta}}\end{aligned}$$

In particular, this means that when we sum over all time steps $s \leq t$, we obtain

$$\begin{aligned}\eta_1 \sum_{s=1}^{t} \|(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^z))^T \Delta Q_s\|_2 &\lesssim \eta_1 \sum_{s=1}^{t} \frac{\tau_0 r\sqrt{\log d}}{\eta_1\sqrt{(t-s)\delta}} \\ &\lesssim \eta_1 r^2 t_1 \sqrt{\log d}\end{aligned}$$

where we used the assumption that $\tau_0 \lesssim \eta_1\sqrt{t_1\delta}$ in the last step. The lemma then follows.

■

Lemma 45 (**$\tilde{g}_t$ is close to g_t**) *With high probability, it holds that*

$$|g_t(x) - \tilde{g}_t(x)| \leq \frac{1}{poly(d)}$$

Proof We note that

$$\begin{aligned}|g_t(x) - \tilde{g}_t(x)| &\leq |N_{V_t}(v, \bar{V}_t; x) - N_{\tilde{V}_t}(v, \bar{V}_t; x)| + |N_{V_t}(v, \tilde{V}_t; x)| \\ &\lesssim (\tau_0\sqrt{m})^{2/3}\eta_1 t_1 + (\tau_0\sqrt{m})^{2/3}\eta_1 t_1 + \tau_0\sqrt{\log d} \\ &\lesssim \tau_0\sqrt{\log d} \\ &\lesssim \frac{1}{\text{poly}(d)}\end{aligned}$$

■

As a corollary to this, we obtain the desired bound for Lemma 8, because we can just consider the dynamics of $\tilde{g}_{t_1}$. This is shown via the following.

Proof [Proof of Lemma 8] We know that

$$\begin{aligned}|g_{t_1}(z-\zeta) + g_{t_1}(z+\zeta) - 2g_{t_1}(z)| &\leq |\tilde{g}_{t_1}(z-\zeta) + \tilde{g}_{t_1}(z+\zeta) - 2\tilde{g}_{t_1}(z)| + \frac{1}{\text{poly}(d)} \\ &= |(\mathbb{1}(\mathcal{E}_t^{z-\zeta}) + \mathbb{1}(\mathcal{E}_t^{z+\zeta}) - 2\mathbb{1}(\mathcal{E}_t^{z}))^T Q_t z + (\mathbb{1}(\mathcal{E}_t^{z+\zeta}) - \mathbb{1}(\mathcal{E}_t^{z-\zeta}))^T Q_t \zeta| \\ &\lesssim \eta_1 r^2 t_1 \sqrt{\log d} \\ &\lesssim r^2 d^{9/8}\sqrt{\log d}\end{aligned}$$

as desired. ■

F.3. Proof of Lemma 9

Let $\hat{t}_1$ denote the annealing time from Lemma 7. Let us also denote throughout this section $t_2 = O(\frac{1}{\eta_2 \epsilon_2^3 r})$. At a high level, the idea of the proof is that the additional variance added by the label noise from annealing time $\hat{t}_1$ to any time $\hat{t}_1 + t$ is small, because of the small learning rate. Intuitively, this means that most of the noise from before annealing is preserved.

We first begin with the following lemma.

Lemma 46 *With high probability, it holds for $t \geq 0$ that*

$$|N_{U_{\hat{t}_1+t}}(u, \tilde{U}_{\hat{t}_1+t}; x)| \lesssim \tau_0\sqrt{\log d} + (\tau_0\sqrt{m})^{2/3}\eta_1\hat{t}_1 + \eta_2^{5/3}t^{5/3}$$

Proof The proof of this is the same analysis as Lemma 32, followed by an application of triangle inequality. ■

Lemma 47 *For any x and any $t = O(t_2 d^{1/8})$, it holds with high probability that*

$$\|\mathbb{1}([U_{\hat{t}_1+t}]x) - \mathbb{1}([U_{\hat{t}_1}]x)\|_1 \lesssim \epsilon_1^2$$

Proof Without loss of generality, let us assume $\|x\|_2 = 1$. First, let us give the following decomposition.

$$\begin{aligned}&\|\mathbb{1}(U_{\hat{t}_1+t}x) - \mathbb{1}(U_{\hat{t}_1}x)\|_1 \\ &= \|\mathbb{1}(U_{\hat{t}_1+t}x) - \mathbb{1}(\tilde{U}_{\hat{t}_1+t}x)\|_1 + \|\mathbb{1}(\tilde{U}_{\hat{t}_1+t}x) - \mathbb{1}(\tilde{U}_{\hat{t}_1}x)\|_1 + \|\mathbb{1}(\tilde{U}_{\hat{t}_1}x) - \mathbb{1}(U_{\hat{t}_1}x)\|_1\end{aligned}$$

We know from Lemma 29 that

$$\|\mathbb{1}(\tilde{U}_{\hat{t}_1}x) - \mathbb{1}(U_{\hat{t}_1}x)\|_1 \lesssim m \cdot (\tau_0\sqrt{m})^{2/3}$$

By similar analysis, we obtain that

$$\|\mathbb{1}(\tilde{U}_{\hat{t}_1+t}x) - \mathbb{1}(U_{\hat{t}_1+t}x)\|_1 \lesssim m \cdot (\tau_0\sqrt{m})^{2/3} + m \cdot \eta_2^{2/3} t^{2/3}$$

We'd now like to bound the middle term. Note that

$$[\tilde{U}_{\hat{t}_1+t}]_i x = [\tilde{U}_{\hat{t}_1}]_i x - \sum_{s=\hat{t}_1}^{\hat{t}_1+t} \eta_2 \epsilon_{s-1}[B_{s-1}]_i x$$

In order to compute $\Pr[\mathbb{1}([\tilde{U}_{\hat{t}_1+t}]_i x) \neq \mathbb{1}([\tilde{U}_{\hat{t}_1}]_i x)]$, we consider the additional $\sum_{s=\hat{t}_1}^{\hat{t}_1+t} \eta_2 \epsilon_{s-1}[B_{s-1}]_i x$ term and the initial $[\tilde{U}_{\hat{t}_1}]_i x$ term. Note that both of these random variables are zero-mean; in particular, we know that $[\tilde{U}_{\hat{t}_1}]_i x$ has variance $\Omega(\tau_0^2)$, and $\sum_{s=\hat{t}_1}^{\hat{t}_1+t} \eta_2 \epsilon_{s-1}[B_{s-1}]_i x$ has variance $O\left(\frac{\eta_2^2 t\delta}{\sqrt{m}}\right)$. Therefore, we obtain that

$$\begin{aligned}\Pr[\mathbb{1}([\tilde{U}_{\hat{t}_1+t}]_i x) \neq \mathbb{1}([\tilde{U}_{\hat{t}_1}]_i x)] &= \Pr[\mathbb{1}([\tilde{U}_{\hat{t}_1}]_i x + \sum_{s=\hat{t}_1}^{\hat{t}_1+t} \eta_2 \epsilon_{s-1}[B_{s-1}]_i x) \neq \mathbb{1}(\tilde{U}_{\hat{t}_1}]_i x)] \\ &\lesssim \sqrt{\frac{\eta_2^2 t\delta/\sqrt{m}}{\tau_0^2}} \\ &\lesssim \frac{\eta_2}{\tau_0}\sqrt{t\delta} \cdot m^{-1/4}\end{aligned}$$

This gives us that with high probability,

$$\|\mathbb{1}(\tilde{U}_{\hat{t}_1+t}x) - \mathbb{1}(\tilde{U}_{\hat{t}_1}x)\|_1 \lesssim m^{3/4} \cdot \frac{\eta_2}{\tau_0}\sqrt{t\delta} + \sqrt{m\log d}$$

Finally, the lemma follows, by summing all three terms together. ■

The result of Lemma 47 allows us to conclude the following:

Lemma 48 *With high probability, it holds for any x that at time $t \leq O(t_2 d^{1/8})$*

$$|N_{U_{\hat{t}_1+t}}(u, U_{\hat{t}_1+t}; x) - N_{U_{\hat{t}_1}}(u, \bar{U}_{\hat{t}_1+t}; x)| \lesssim \epsilon_1^2.$$

Proof This follows by triangle inequality, using previous lemmas. In particular, we have that

$$
\begin{aligned}
&|N_{U_{\hat{t}_1+t}}(u, U_{\hat{t}_1+t}; x) - N_{U_{\hat{t}_1}}(u, \bar{U}_{\hat{t}_1+t}; x)| \\
&\leq |N_{U_{\hat{t}_1+t}}(u, U_{\hat{t}_1+t}; x) - N_{U_{\hat{t}_1+t}}(u, \bar{U}_{\hat{t}_1+t}; x)| + |N_{U_{\hat{t}_1+t}}(u, \bar{U}_{\hat{t}_1+t}; x) - N_{U_{\hat{t}_1}}(u, \bar{U}_{\hat{t}_1+t}; x)| \\
&\lesssim |N_{U_{\hat{t}_1+t}}(u, \tilde{U}_{\hat{t}_1+t}; x)| + \frac{1}{\sqrt{m}} \cdot \|\mathbb{1}(U_{\hat{t}_1+t}x) - \mathbb{1}(U_{\hat{t}_1}x)\|_1 \cdot \max_i \|[\bar{U}_{\hat{t}_1+t'}]_i\|_2 \\
&\lesssim |N_{U_{\hat{t}_1+t}}(u, \tilde{U}_{\hat{t}_1+t}; x)| \\
&+ \frac{1}{\sqrt{m}} \left(m \cdot (\tau_0\sqrt{m}) + m \cdot \eta_2^{2/3} t^{2/3} + m^{3/4} \cdot \frac{\eta_2}{\tau_0} \sqrt{t\delta} + \sqrt{m \log d} \right) \cdot \frac{(\eta_1 \hat{t}_1 + \eta_2 t)}{\sqrt{m}} \\
&\lesssim \tau_0 \sqrt{\log d} + (\tau_0\sqrt{m})^{2/3} \eta_1 \hat{t}_1 + \eta_2^{5/3} t^{5/3} \\
&+ \frac{1}{m} \left(m \cdot (\tau_0\sqrt{m}) + m \cdot \eta_2^{2/3} t^{2/3} + m^{3/4} \cdot \frac{\eta_2}{\tau_0} \sqrt{t\delta} + \sqrt{m \log d} \right) \cdot (\eta_1 \hat{t}_1 + \eta_2 t) \\
&\lesssim \tau_0 \sqrt{\log d} + (\tau_0\sqrt{m})^{2/3} \eta_1 \hat{t}_1 + \eta_2^{5/3} t^{5/3} + (m^{-1/4} \frac{\eta_2 \sqrt{t\delta}}{\tau_0} + m^{-1/2} \sqrt{\log d})(\eta_1 \hat{t}_1 + \eta_2 t) \\
&\lesssim \tau_0 \sqrt{\log d} + (\tau_0\sqrt{m})^{2/3} \eta_1 \hat{t}_1 + \eta_2^{5/3} t^{5/3} \\
&\lesssim \epsilon_1^2
\end{aligned}
$$

where the last line follows from our choice of parameters that $\tau_0\sqrt{\log d} \lesssim \epsilon_1^2$, $(\tau_0\sqrt{m})^{2/3}\eta_1\hat{t}_1 \lesssim \epsilon_1^2$, $\eta_2^{5/3} t_2 d^{1/8} \lesssim \epsilon_1^2$. ■

We now proceed to complete the proof of Lemma 9, via the following series of lemmas. We can then show that there exists a target signal that the weights approaches after the learning rate is annealed. In particular, this target signal analysis will be similar to the spirit of the target signal analysis before the annealing. The difference is that here, we will construct a target $V^\star$ for the nonlinear component of the data.

First, let us recall the following notation. We use $\hat{L}$ to denote the average loss on ℓ, and we use $\bar{L}$ to denote the average loss on $\bar{\ell}$. By Lemma 26 we know that in general, $\hat{L}$ and $\bar{L}$ will be close, because of our choice of scaling for δ.

Let us now give the following Lemma C.9 from Li et al. [19], as adapted for our expected smoothed loss $\bar{\ell}$.

Lemma 49 *Consider a time t, where it holds that $|g_t(z+\zeta) + g_t(z-\zeta) - 2g_t(z)| \leq \kappa$ for some small κ. Then, it holds that*

$$\bar{L}_{\bar{\mathcal{M}}_1}(u, U_t) \geq \log 2 - O(\kappa) - O\left(\frac{\log d}{\sqrt{qN}}\right)$$

In addition, if $\bar{L}_{\bar{\mathcal{M}}_1} \leq \log 2 + O(\kappa')$ for $\kappa' \gtrsim \kappa$, then

$$|g_t(z+\zeta)|, |g_t(z-\zeta)|, |g_t(z)| \leq O\left(\sqrt{\kappa' + \frac{\log d}{\sqrt{qN}}}\right)$$

We can then continue to show the following, allowing us to bound the key term ϵ_0.

Lemma 50 *With high probability, it holds that*

$$|g_{\hat{t}_1}(z+\zeta)|, |g_{\hat{t}_1}(z-\zeta)|, |g_{\hat{t}_1}(z)|, \epsilon_0 \leq O\left(\sqrt{\frac{\epsilon_1}{q}}\right)$$

where we define

$$\epsilon_0 \triangleq \frac{1}{N}\sum_{i\in\mathcal{M}_1} \bar{\ell}(y^{(i)} r_{t_1}(x^{(i)}))$$

to be a notion of pseudo-loss that is computed via just the W weights.

Proof To show this, we first recall that with high probability, $\bar{L}_{\hat{t}_1} \leq q\log 2 + \epsilon_1$, which implies that $\bar{L}_{\mathcal{M}_1}(u, U_{\hat{t}_1}) \leq \log 2 + \frac{\epsilon_1}{q}$. In the setting of the previous lemma, let $\kappa = \kappa' = O(\epsilon_1) = O(r^2 d^{9/8}\sqrt{\log d})$. Then this implies that $|g_{\hat{t}_1}(z+\zeta)|, |g_{\hat{t}_1}(z-\zeta)|, |g_{\hat{t}_1}(z)| \leq O\left(\sqrt{\frac{\epsilon_1}{q}}\right)$.

Thus, we obtain that

$$\begin{aligned}
\epsilon_0 &= \frac{1}{N}\sum_{i\in\mathcal{M}_1} \bar{\ell}(y^{(i)} r_{t_1}(x^{(i)})) \\
&\leq \frac{1}{N}\sum_{i\in\mathcal{M}_1} \bar{\ell}(y^{(i)}(r_{t_1}(x_1^{(i)}) + g_{t_1}(x_2^{(i)}))) + \frac{1}{N}\sum_{i\in\mathcal{M}_1} |g_{t_1}(x_2^{(i)})| \\
&\leq (\bar{L}_{t_1} - q\bar{\ell}_{\bar{\mathcal{M}}_1}(u, U_{t_1})) + O\left(\frac{\sqrt{\epsilon_1/q}}{N}\right) \\
&\leq (q\log 2 + \epsilon_1 - q\log 2 + q\epsilon_1) + O\left(\frac{\sqrt{\epsilon_1/q}}{N}\right) \\
&\leq O\left(\sqrt{\frac{\epsilon_1}{q}}\right)
\end{aligned}$$

■

We are now able to construct a target signal for $V^\star$ that we can show convergence to in the second phase of training. This will be formalized via the following lemma.

Lemma 51 *At some time $\hat{t}_2 \lesssim t_2$ time steps after annealing, we have that there exists a target signal $U^\star$ such that*

$$K_{\hat{t}_1+t}(\bar{U}_{\hat{t}_1} + U^\star) \leq \epsilon_0 + \epsilon_1.$$

Proof We follow the construction of Li et al. [19] for the additional target signal. In particular, first define the sets

$$\begin{aligned}
\mathcal{E}_1 &= \{i \in [m] : [V_{\hat{t}_1}]_i(z-\zeta) \geq 0, [V_{\hat{t}_1}]_i z \geq 0, [V_{\hat{t}_1}]_i(z+\zeta) < 0\} \\
\mathcal{E}_2 &= \{i \in [m] : [V_{\hat{t}_1}]_i(z-\zeta) \geq 0, [V_{\hat{t}_1}]_i z < 0, [V_{\hat{t}_1}]_i(z+\zeta) < 0\} \\
\mathcal{E}_3 &= \{i \in [m] : [V_{\hat{t}_1}]_i(z-\zeta) < 0, [V_{\hat{t}_1}]_i z < 0, [V_{\hat{t}_1}]_i(z+\zeta) \geq 0\}
\end{aligned}$$

Then, define the target additional signal matrix after annealing as $U^\star$ with $W^\star = 0$ and $V^\star$ to be

$$V_i^\star = \begin{cases} \frac{20c\log(1/\epsilon_1)v_i}{r\epsilon_1} z & \text{if } i \in \mathcal{E}_1 \\ -\frac{40c\log(1/\epsilon_1)v_i}{r\epsilon_1} z & \text{if } i \in \mathcal{E}_2 \\ -\frac{20c\log(1/\epsilon_1)v_i}{r\epsilon_1} z & \text{if } i \in \mathcal{E}_3 \\ 0 & \text{otherwise} \end{cases}$$

Now, let us consider $[\tilde{V}_{\hat{t}_1}]_i$; in particular, i will be in the $\mathcal{E}$ sets with probability $O(r)$[2], the angle between $z - \zeta$ and z, because at time t_1 the network still has a bad margin on $\mathcal{Q}$, and hence hasn't learned much.

Next, we note that for $x_2 = \alpha(z - \zeta)$, with high probability it holds that, by definition of $V^\star$,

$$N_{V_{\hat{t}_1}}(v, V^\star; \alpha(z-\zeta)) = \alpha \frac{1}{m}\left(|\mathcal{E}_1|\frac{20c\log(1/\epsilon_1)}{r\epsilon_1} - |\mathcal{E}_2|\frac{40c\log(1/\epsilon_1)}{r\epsilon_1}\right) \leq \alpha\frac{-2c\log(1/\epsilon_1)}{\epsilon_1}$$

Similarly, we get high probability bounds of the cases of $x_2 = \alpha(z + \zeta)$ and $x_2 = \alpha z$ as follows.

$$N_{V_{\hat{t}_1}}(v, V^\star; \alpha(z+\zeta)) = -\alpha\frac{1}{m}|\mathcal{E}_3|\frac{20c\log(1/\epsilon_1)}{r\epsilon_1} \leq \alpha\frac{-2c\log(1/\epsilon_1)}{\epsilon_1}$$

and

$$N_{V_{\hat{t}_1}}(v, V^\star; \alpha z) = \alpha\frac{1}{m}|\mathcal{E}_1|\frac{20c\log(1/\epsilon_1)}{r\epsilon_1} \geq \alpha\frac{2c\log(1/\epsilon_1)}{\epsilon_1}$$

In particular, by definitions of the labels, we obtain that for all $i \in [N]$,

$$y^{(i)} N_{V_{\hat{t}_1}}(v, V^\star, x_2^{(i)}) \geq \frac{2c\log(1/\epsilon_1)}{\epsilon_1}\|x_2^{(i)}\|_2$$

Now, recall that we want to bound $K_{\hat{t}_1+t'}(\bar{U}_{\hat{t}_1} + U^\star) = K_{\hat{t}_1+t_2}((\bar{W}_{\hat{t}_1}, \bar{V}_{\hat{t}_1} + V^\star))$. For the signal $\bar{W}$, it suffices to note that

$$\begin{aligned} &|N_{W_{\hat{t}_1+t_2}}(w, \bar{W}_{\hat{t}_1}; x_1) - N_{W_{\hat{t}_1+t_2}}(w, W_{\hat{t}_1}; x_1)| \\ &\leq |N_{W_{\hat{t}_1+t_2}}(w, \bar{W}_{\hat{t}_1}; x_1) - N_{W_{\hat{t}_1}}(w, \bar{W}_{\hat{t}_1}; x_1)| + |N_{W_{\hat{t}_1}}(w, \tilde{W}_{\hat{t}_1}; x_1)| \\ &\lesssim \frac{1}{\sqrt{m}}\|\mathbb{1}(W_{\hat{t}_1+t_2}x_1) - \mathbb{1}(W_{\hat{t}_1}x_1)\|_1 \max_i \|[\bar{W}_{\hat{t}_1}]_i\|_2 + |N_{W_{\hat{t}_1}}(w, \tilde{W}_{\hat{t}_1}; x_1)| \\ &\lesssim \frac{1}{m}\epsilon_1^2 \cdot \eta_1 \hat{t}_1 + \tau_0\sqrt{\log d} \\ &\lesssim q\epsilon_1 \end{aligned}$$

Now, we will bound the loss accounted for by the $\bar{V}_{\hat{t}_1} + V^\star$ term. In particular, we note that by triangle inequality, $yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1} + V^\star; x_2) \geq yN_{V_{\hat{t}_1+t_2}}(v, V^\star; x_2) - |yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1}; x_2)|$, where the first term on the right hand side is always positive as shown before (which allows us to remove the absolute value). We will proceed to bound the terms on the right hand side.

[2] For $\|\zeta\|_2 \ll \|z\|_2$, the small-angle approximation holds.

First, note that for all i,

$$|V_i^\star x_2| \lesssim \frac{1}{\sqrt{m}\epsilon_1 r}$$

by definition of our $V^\star$. As such we can give the following bound:

$$\begin{aligned}
&\left|\sum_{i\in[m]} v_i(V_i^\star x_2)\cdot(\mathbb{1}([V_{\hat{t}_1+t_2}]_i x_2) - \mathbb{1}([V_{\hat{t}_1}]_i x_2]))\right| \\
&\lesssim \frac{1}{m\epsilon_1 r}\cdot \|\mathbb{1}([V_{\hat{t}_1+t_2}]x_2)) - \mathbb{1}([V_{\hat{t}_1}]x_2])\|_1 \\
&\lesssim \frac{1}{m\epsilon_1 r}\cdot \epsilon_1^2 \\
&\lesssim O(1)
\end{aligned}$$

From this, we obtain that

$$\begin{aligned}
yN_{V_{\hat{t}_1+t_2}}(v, V^\star; x_2) &= y\left(\sum_{i\in[m]} v_i(V_i^\star x_2)\cdot \mathbb{1}([V_{\hat{t}_1+t_2}]_i x_2)\right) \\
&\geq y(\sum_{i\in[m]} v_i(V_i^\star x_2)\cdot(\mathbb{1}([V_{\hat{t}_1+t_2}]_i x_2) - \mathbb{1}([V_{\hat{t}_1}]_i x_2]))) + y(\sum_{i\in[m]} v_i(V_i^\star x_2)\cdot(\mathbb{1}([V_{\hat{t}_1}]_i x_2])) \\
&\gtrsim \frac{c\|x_2\|_2}{\epsilon_1}\log(1/\epsilon_1)
\end{aligned}$$

We now upper bound the $|yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1}; x_2)|$ term. To do so, first note that

$$\begin{aligned}
|yN_{V_{\hat{t}_1}}(v, \bar{V}_{\hat{t}_1}; x_2)| &\leq |N_{V_{\hat{t}_1}}(v, \bar{V}_{\hat{t}_1}; x_2) - N_{V_{\hat{t}_1}}(v, V_{\hat{t}_1}; x_2)| + |N_{V_{\hat{t}_1}}(v, V_{\hat{t}_1}; x_2)| \\
&= |N_{V_{\hat{t}_1}}(v, \tilde{V}_{\hat{t}_1}; x_2)| + |g_{\hat{t}_1}(x_2)| \\
&\lesssim \tau_0\sqrt{\log d} + |g_{\hat{t}_1}(x_2)| \\
&\lesssim O(1)
\end{aligned}$$

In particular, this gives us

$$\begin{aligned}
|yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1}; x_2)| &= \left|\left(\sum_{i\in[m]} v_i([\bar{V}_{\hat{t}_1}]_i x_2)\mathbb{1}([V_{\hat{t}_1+t_2}]_i x_2)\right)\right| \\
&\leq \left|\left(\sum_{i\in[m]} v_i([\bar{V}_{\hat{t}_1}]_i x_2)\mathbb{1}([V_{\hat{t}_1}]_i x_2)\right)\right| + \left|\left(\sum_{i\in[m]} v_i([\bar{V}_{\hat{t}_1}]_i x_2)(\mathbb{1}([V_{\hat{t}_1+t_2}]_i x_2) - \mathbb{1}([V_{\hat{t}_1}]_i x_2))\right)\right| \\
&= |yN_{V_{\hat{t}_1}}(v, \bar{V}_{\hat{t}_1}; x_2)| + \left|\left(\sum_{i\in[m]} v_i([\bar{V}_{\hat{t}_1}]_i x_2)(\mathbb{1}([V_{\hat{t}_1+t_2}]_i x_2) - \mathbb{1}([V_{\hat{t}_1}]_i x_2))\right)\right| \\
&\lesssim O(1)
\end{aligned}$$

As noted earlier, since $yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1} + V^\star; x_2) \geq yN_{V_{\hat{t}_1+t_2}}(v, V^\star; x_2) - |yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1}; x_2)|$, we can prove the following:

$$\begin{aligned} yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1} + V^\star; x_2) &\geq yN_{V_{\hat{t}_1+t_2}}(v, V^\star; x_2) - |yN_{V_{\hat{t}_1+t_2}}(v, \bar{V}_{\hat{t}_1}; x_2)| \\ &\gtrsim \frac{c}{\epsilon_1} \log(1/\epsilon_1) - O(1) \\ &\gtrsim \log\left(\frac{1}{\epsilon_1}\right) \end{aligned}$$

From this point, we obtain that

$$\begin{aligned} K_{t_1+t'}(\bar{U}_{t_1} + U^\star) &= K_{t_1+t'}((\bar{W}_{t_1}, \bar{V}_{t_1} + V^\star)) \\ &\leq \frac{|\mathcal{M}_1|}{N} \bar{L}_{\mathcal{M}_1}(r_{t_1}) + O(q\epsilon_1) + \frac{|\bar{M}_1|}{N} \bar{L}_{\bar{\mathcal{M}}_1}(N_{t_1+t'}(v, \bar{V}_{t_1} + V^\star); \cdot) \\ &\leq \epsilon_0 + \epsilon_1 \end{aligned}$$

Here, convergence using the Lemma 34 will give the desired upper bound for number of time steps.
■

F.4. Proof of Lemma 10

We can then show that our label noise SGD algorithm converges to a low training loss on *both* the $\mathcal{P}$ and $\mathcal{Q}$ component, which is good. In particular, for the $\mathcal{P}$ component, the low loss is on the entire $\mathcal{M}_1$, which implies generalization via standard Rademacher complexity bounds. This will be formalized by the following analysis.

Proof Consider the difference between losses at time $\hat{t}_1$ and time $\hat{t}_1 + \hat{t}_2$ on the $\mathcal{P}$ component, restricted to the set $\mathcal{M}_1$. Then, we have that

$$\begin{aligned} &|\bar{L}_{\mathcal{M}_1}(r_{\hat{t}_1}) - \bar{L}_{\mathcal{M}_1}(r_{\hat{t}_1+\hat{t}_2})| \\ &= |\frac{1}{|\mathcal{M}_1|} \sum_{i \in \mathcal{M}_1} (\bar{\ell}(y^{(i)} r_{t_1}(x^{(i)})) - \bar{\ell}(y^{(i)} r_{t_1+t'}(x^{(i)})))| \\ &\leq \frac{1}{|\mathcal{M}_1|} \sum_{i \in \mathcal{M}_1} |r_{t_1}(x^{(i)}) - r_{\hat{t}_1+\hat{t}_2}(x^{(i)})| \\ &\leq \frac{1}{|\mathcal{M}_1|} \cdot \sqrt{N} \cdot \sqrt{\sum_{i \in \mathcal{M}_1} (r_{\hat{t}_1}(x^{(i)}) - r_{\hat{t}_1+\hat{t}_2}(x^{(i)}))^2} \end{aligned}$$

where the third line follows from $\bar{\ell}$ being 1-Lipschitz, and the fourth line follows from Cauchy-Schwartz. In particular, by our choice of η_2, we will bound the term inside the square root as follows.

$$\begin{aligned} &|r_{\hat{t}_1}(x^{(i)}) - r_{\hat{t}_1+\hat{t}_2}(x^{(i)})| \\ &\leq |N_{W_{\hat{t}_1+\hat{t}_2}}(w, W_{\hat{t}_1+\hat{t}_2}; x_1^{(i)}) - N_{W_{\hat{t}_1}}(w, \bar{W}_{\hat{t}_1+\hat{t}_2}; x_1^{(i)})| + |N_{W_{\hat{t}_1}}(w, \bar{W}_{\hat{t}_1+\hat{t}_2}; x_1^{(i)}) - N_{W_{\hat{t}_1}}(w, \bar{W}_{\hat{t}_1}; x_1^{(i)})| \\ &+ |N_{W_{\hat{t}_1}}(w, \tilde{W}_{\hat{t}_1}; x_1^{(i)})| \\ &\lesssim |N_{W_{\hat{t}_1}}(w, \bar{W}_{\hat{t}_1+\hat{t}_2}; x_1^{(i)}) - N_{W_{\hat{t}_1}}(w, \bar{W}_{\hat{t}_1}; x_1^{(i)})| + \epsilon_1^2 \end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\sqrt{N}\cdot\sqrt{\sum_{i\in\mathcal{M}_1}(r_{\hat{t}_1}(x^{(i)})-r_{\hat{t}_1+\hat{t}_2}(x^{(i)}))^2}\\
&\le\sqrt{N}\sqrt{\sum_{i\in\mathcal{M}_1}|N_{W_{\hat{t}_1}}(w,\bar{W}_{\hat{t}_1+\hat{t}_2};x_1^{(i)})-N_{W_{\hat{t}_1}}(w,\bar{W}_{\hat{t}_1};x_1^{(i)})|^2}\\
&\lesssim\sqrt{N}\sqrt{\|\bar{W}_{\hat{t}_1+\hat{t}_2}-\bar{W}_{\hat{t}_1}\|_F^2\|X\|_2^2}\\
&\lesssim\sqrt{N}\sqrt{\|\bar{W}_{\hat{t}_1+\hat{t}_2}-\bar{W}_{\hat{t}_1}\|_F^2\frac{N}{d}}\\
&\lesssim\frac{N}{\sqrt{d}}\cdot\frac{1}{\epsilon_1\sqrt{r}}\\
&\lesssim N\epsilon_1
\end{aligned}$$

where the last line follows from

$$\frac{1}{\epsilon_1^4}\lesssim dr^2\lesssim dr$$

Now, we note that simply $\epsilon_0=(1-q)\bar{L}_{\mathcal{M}_1}(r_{\hat{t}_1})$. Therefore, substituting everything back in, we obtain that

$$\left|\bar{L}_{\mathcal{M}_1}(r_{\hat{t}_1+\hat{t}_2})-\frac{\epsilon_0}{1-q}\right|\le\frac{\epsilon_1}{q}$$

Hence, $|\bar{L}_{\mathcal{M}_1}(r_{\hat{t}_1+\hat{t}_2})|\lesssim O\left(\sqrt{\frac{\epsilon_1}{q}}\right)$ follows directly from our previous upper bound of $\epsilon_0\lesssim\sqrt{\frac{\epsilon_1}{q}}$. Furthermore, since $\bar{L}_{\mathcal{M}_1}\le\bar{L}_{\hat{t}_1+\hat{t}_2}\le O\left(\sqrt{\frac{\epsilon_1}{q}}\right)$, we have that

$$\bar{L}_{\bar{\mathcal{M}}_1}(g_{\hat{t}_1+\hat{t}_2})\lesssim\sqrt{\frac{\epsilon_1}{q^3}}$$

as desired.

■

F.5. Proof of Theorem 4

We now make use of the lemmas we have proved, to finally prove the theorem statement of algorithm LNSGD-LS. Essentially, the fine grained loss from Lemma 10 allows us to conclude the theorem, because $\mathcal{P}$ has enough sample complexity, and $\mathcal{Q}$ naturally generalizes well.

Proof First, we give a lemma due to Li et al. [19], which essentially lower bounds the magnitude of the output at stopping time of training. We adapt it as follows.

Lemma 52 *With high probability for $\hat{t}_1,\hat{t}_2$ from earlier, for every $\kappa\ge\frac{1}{\sqrt{qN}}$, it holds that $\bar{L}_{\bar{\mathcal{M}}_1}(g_{\hat{t}_1+\hat{t}_2})\le\delta$ implies $yg_{\hat{t}_1+\hat{t}_2}(x_2)\gtrsim\frac{\|x\|_2}{\kappa}$.*

This lemma is useful in our analysis as it allows us to reason that the low-norm x_2 observations of the $\mathcal{Q}$ component will still have large output in g.

First, note that by our choice of hyperparameters, $\epsilon_1 \geq \frac{1}{\sqrt{N}}$. Therefore, it holds that $yg_{\hat{t}_1+\hat{t}_2}(x_2) \gtrsim \frac{\|x\|_2}{\sqrt{\epsilon_1/q^3}}$. To look at the $\mathcal{P}$ component, we note that with high probability,

$$|r_{\hat{t}_1+\hat{t}_2}| = |N_{\hat{t}_1+\hat{t}_2}(u, U_{\hat{t}_1+\hat{t}_2}; x_1)| \lesssim 1$$

This implies the quantity $\|x_2\|$ will determine the magnitude of $y(r_{\hat{t}_1+\hat{t}_2} + g_{\hat{t}_1+\hat{t}_2})$. In particular, if $\|x_2\| \gtrsim \sqrt{\frac{\epsilon_1}{q^3}\log\frac{1}{\epsilon_1}}$, then

$$y(r_{\hat{t}_1+\hat{t}_2} + g_{\hat{t}_1+\hat{t}_2}) \gtrsim \log\frac{1}{\epsilon_1}$$

Otherwise, if $\|x_2\|_2 \lesssim \sqrt{\frac{\epsilon_1}{q^3}}$, then

$$y(r_{\hat{t}_1+\hat{t}_2} + g_{\hat{t}_1+\hat{t}_2}) \lesssim 1$$

Recall from the construction of the target signal of W that $\|W_{\hat{t}_1+\hat{t}_2}\|_F^2 \lesssim d\log^2\frac{1}{\epsilon_1}$. Then, it holds that

$$\begin{aligned}
&\mathbb{E}_{(x,y)}[\bar{\ell}(r_{\hat{t}_1+\hat{t}_2} + g_{\hat{t}_1+\hat{t}_2})] \\
&\leq \Pr[x_2 = 0]\mathbb{E}[\bar{\ell}(r_{\hat{t}_1+\hat{t}_2})] + \Pr[x_2 \neq 0]\mathbb{E}[\bar{\ell}(r_{\hat{t}_1+\hat{t}_2} + g_{\hat{t}_1+\hat{t}_2})] \\
&\leq p\kappa\log\frac{1}{\epsilon_1} + \epsilon_1 \\
&\leq O(p\kappa\log\frac{1}{\epsilon_1})
\end{aligned}$$

where the second to last inequality comes from our choice of ϵ_1, and the last equality follows from our previously conditional expectations.

■

Appendix G. Proof of Theorem 5 (LNSGD-S)

G.1. Proof of Lemma 11

Proof Essentially, we construct a target signal for the $\mathcal{Q}$, and show convergence to this target. The proof follows the same target signal construction as Lemma 51. In particular, here the upper bound on the time steps needed is $O(\frac{1}{\eta_2(\epsilon_2')^3 r})$, and the SGD optimization Lemma 34 can be used to show the target signal is reached. ■

G.2. Proof of Lemma 12

For this second phase of training after $\mathcal{Q}$ is memorized, we follow Li et al. [19]'s construction of the target signal, which roughly states that the target signal has a dependency on the linearly separating hyperplane fit via only the pN observations in $\bar{\mathcal{M}}_2$.

Proof The target matrix $U^\star$ is constructed by considering the vector $\beta = X(X^TX)^{-1}y^T \in \mathbb{R}^{d\times 1}$, where $X \in \mathbb{R}^{d\times pN}$ and $y \in \mathbb{R}^{1\times pN}$. Then, construct $W_i^\star = 10w_i \log\frac{1}{\epsilon_2'}\beta$, and $V^\star$ as in Lemma 51. It suffices here to note that $\|\beta\|_2 \lesssim \sqrt{pN}$; by following the SGD optimization Lemma 34, we obtain the time to be $O\left(\frac{pN}{\eta_2\epsilon_2'}\right)$.

$\blacksquare$

G.3. Proof of Lemma 13 and Lemma 14

Essentially, this lemma tells us that the accumulated gradient signal $\bar{W}$ on the $\mathcal{M}_2$ component is small when we stop the training.

In proving this lemma, we first denote $t_3 = O\left(\frac{d}{\eta_2\epsilon_2'}\right)$. Furthermore, we define the following.

$$\rho_t = \frac{1}{N}\sum_{i\in\mathcal{M}_2} |\bar{\ell}'(y^{(i)}f_t(x^{(i)}))|$$

Intuitively, this quantity represents the average absolute value of the logistic loss's derivative over all the observations in $\mathcal{M}_2$. We basically argue that ρ_t will decrease very quickly.

Lemma 53 *For $t \le t_3$, it holds that if $\rho_t = \Omega\left(\frac{1}{N}\right)$, then with high probability,*

$$\|\nabla\bar{L}_t\|_F^2 \gtrsim r\rho_t^4$$

We use the following ReLU geometry lemma due to Li and Liang [17], which we use to show that with high probability, the gradient as calculated by the noise activations is large.

Lemma 54 *For any $v_1, v_2, v_3 \in \mathcal{R}$, it holds that*

$$\begin{aligned}&\mathbb{E}_w[\|v_1\mathbb{1}(w^T(z-\zeta))(z-\zeta) + v_2\mathbb{1}(w^T(z+\zeta))(z+\zeta) + v_3\mathbb{1}(w^Tz)z\|_2^2]\\ &\gtrsim r(v_1^2+v_2^2+v_3^2)\end{aligned}$$

where $w \sim \mathcal{N}(0, \tau_0^2 I)$.

Proof [Proof of Lemma 53] Fix a time t. We first note that, for a fixed $j \in [m]$,

$$\nabla_{[V]_j}\bar{L}_t = \frac{1}{N}\sum_{i\in\mathcal{M}_2} \bar{\ell}'(y^{(i)}f_t(x^{(i)}))\cdot v_i\cdot\mathbb{1}([V_t]_jx_2^{(i)})x_2^{(i)}$$

Following Li et al. [19]'s definitions of the following sets, we denote

$$\mathcal{S}_{\alpha_0}^w = \{i \in [N] : x_2^{(i)} = \alpha_i w \text{ for some } \alpha_i \ge \alpha_0\}$$

where w is a vector.

Then, note that

$$\begin{aligned}
&Nmv_j \nabla_{[V]_j} \bar{L}_t \\
&= \sum_{i \in \mathcal{S}_0^{z-\zeta}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \mathbb{1}([V_t]_j(z-\zeta))(z-\zeta) + \sum_{i \in \mathcal{S}_0^{z+\zeta}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \mathbb{1}([V_t]_j(z+\zeta))(z+\zeta) \\
&+ \sum_{i \in \mathcal{S}_0^{z}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \mathbb{1}([V_t]_j z) z
\end{aligned}$$

We can then define the matrix $\tilde{L}$ so that

$$\begin{aligned}
[\tilde{L}]_j = &\sum_{i \in \mathcal{M}_0^{z-\zeta}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \mathbb{1}([\tilde{V}_t]_j(z-\zeta))(z-\zeta) + \sum_{i \in \mathcal{M}_0^{z+\zeta}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \mathbb{1}([\tilde{V}_t]_j(z+\zeta))(z+\zeta) \\
&+ \sum_{i \in \mathcal{M}_0^{z}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \mathbb{1}([\tilde{V}_t]_j z) z
\end{aligned}$$

Thus, we obtain that with high probability,

$$\begin{aligned}
\mathbb{E}[\|[\tilde{L}]_j\|_2^2] &\gtrsim r \left(\left[\sum_{i \in \mathcal{M}_0^{z-\zeta}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \right]^2 + \left[\sum_{i \in \mathcal{M}_0^{z+\zeta}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \right]^2 + \left[\sum_{i \in \mathcal{M}_0^{z}} \alpha_i \bar{\ell}'(y^{(i)} f_t(x^{(i)})) \right]^2 \right) \\
&\gtrsim r \left(\sum_{i \in \mathcal{M}_2} \alpha_i |\bar{\ell}'(y^{(i)} f_t(x^{(i)}))| \right)^2
\end{aligned}$$

Since this holds for all neurons, we get that

$$\|\tilde{L}\|_F^2 \gtrsim mr \left(\sum_{i \in \mathcal{M}_2} \alpha_i |\bar{\ell}'(y^{(i)} f_t(x^{(i)}))| \right)^2 - O(m^{1/2} N^4)$$

by concentration inequalities.

We can now reason that $\frac{1}{N^2 m}\|\tilde{L}\|_F^2$ and $\|\nabla L(\bar{U}_t)\|_F^2$ are close. In particular, we can note that

$$\begin{aligned}
\left| \frac{1}{N^2 m} \|\tilde{L}\|_F^2 - \|\nabla \bar{L}(U_t)\|_F^2 \right| &\lesssim \frac{1}{Nm} \sum_j \sum_i \bar{\ell}'(y^{(i)} f_t(x^{(i)}))^2 |\mathbb{1}([V_t]_j x_2^{(i)}) - \mathbb{1}([\tilde{V}_t]_j x_2^{(i)})| \\
&\lesssim \eta_2^{2/3} t^{2/3}
\end{aligned}$$

Now, to bound the $\left(\sum_{i \in \mathcal{M}_2} \alpha_i |\bar{\ell}'(y^{(i)} f_t(x^{(i)}))|\right)^2$ term, we note that

$$\begin{aligned}
\left(\sum_{i \in \mathcal{M}_2} \alpha_i |\bar{\ell}'(y^{(i)} f_t(x^{(i)}))| \right)^2 &\gtrsim \rho_t \left(\sum_{i \in \mathcal{M}_{\Theta(\sqrt{\rho_t})}} |\bar{\ell}'(y^{(i)} f_t(x^{(i)}))| \right)^2 \\
&\gtrsim \rho_t (N\rho_t)^2
\end{aligned}$$

Therefore, this tells us that

$$\begin{aligned}
\|\nabla \bar{L}(U_t)\|_F^2 &\gtrsim \frac{1}{N^2 m}\|\tilde{L}\|_F^2 - \eta_2^{2/3} t^{2/3} \qquad (2) \\
&\gtrsim \frac{1}{N^2 m}[mr\left(\sum_{i\in\mathcal{M}_2} \alpha_i |\bar{\ell}'(y^{(i)} f_t(x^{(i)}))|\right)^2 - O(m^{1/2}N^4)] \qquad (3) \\
&\gtrsim r\rho_t^4 \qquad (4)
\end{aligned}$$

as desired. ■

At this point, we can bound the number of time steps where ρ_t is large (and hence where the gradient norm is large). This is formalized by the following lemma.

Lemma 55 *With high probability, the number of time steps t where $\rho_t \gtrsim \frac{d^{-1/32}}{(\epsilon_2')^2}$ is smaller than $\frac{d^{1/8}(\epsilon_2')^8}{\eta_2 r}$.*

Proof To bound this, we will analyze the gradient dynamics on the loss function itself. Consider a pseudo-network activated on the initialization weights; we define such to be $\mathcal{F}_s(x) \triangleq N_{U_0}(u, \bar{U}_s; x)$. Then, we obtain that

$$\bar{L}(\mathcal{F}_{s+1}) \leq \bar{L}(\mathcal{F}_s) - \eta_2 \langle \nabla_U \bar{\ell}(y^{(i_s)} f_s(x^{(i_s)})), \nabla_U \bar{L}(\mathcal{F}_s)\rangle + O(\eta^2)$$

This implies that

$$\begin{aligned}
\bar{L}(\mathcal{F}_0) + O(\eta_2^2 t) &\geq \eta_2 \sum_{s=1}^t \langle \nabla_U \bar{\ell}(y^{(i_s)} f_s(x^{(i_s)})), \nabla_U \bar{L}(\mathcal{F}_s)\rangle \\
&\gtrsim \eta_2 t \|\nabla_U \bar{L}(\mathcal{F}_s)\|_F^2
\end{aligned}$$

Therefore, we see that for a given scalar β, if $\rho \geq \beta$, then $\|\nabla_U \bar{L}(\mathcal{F}_s)\|_F^2 \geq r\beta^4$ by the previous lemma. In particular, the last inequality implies that there are $\lesssim \frac{1}{\eta_2 r \beta^4}$ time steps where ρ_t is big; here, we have $\beta = O\left(\frac{d^{-1/32}}{(\epsilon_2')^2}\right)$.

■

Now, we can finally go about proving Lemma 13.

Proof [Proof of Lemma 13] Recall $\bar{W}_t^{(2)} = -\eta_2 \sum_{s=1}^t \cdot \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})} B_{s-1} \cdot \mathbb{1}(i_{s-1} \in \mathcal{M}_2)$. Then, we have that

$$\begin{aligned}
\left\|\sum_{i\in\mathcal{M}_2} \nabla_W \bar{\ell}(y^{(i)} f_t(x^{(i)}))\right\|_F^2 &= \sum_{j\in[m]} \left\|\sum_{i\in\mathcal{M}_2} \bar{\ell}'(y^{(i)} f_t(x^{(i)})) w_j \mathbb{1}([W_t]_j x_1^{(i)}) x_1^{(i)}\right\|_2^2 \\
&\leq \sum_{j\in[m]} \|X\|_2^2 \left(\sum_{i\in\mathcal{M}_2} |\bar{\ell}'(y^{(i)} f_t(x^{(i)}))|\right) \\
&\lesssim \frac{N^2}{d} \cdot \rho_t
\end{aligned}$$

We can now bound $\|\bar{W}_t^{(2)}\|_F$. In particular, this gives us

$$\begin{aligned}
\|\bar{W}_t^{(2)}\|_F &\leq \eta_2 \sum_{s\leq t:\rho_s \text{ small}} \|\mathbb{1}(i_s \in \mathcal{M}_2)\nabla_W \bar{\ell}'(y^{(i_s)} f_s(x^{(i_s)}))\|_F \\
&\quad + \eta_2 \sum_{s\leq t:\rho_s \text{ large}} \|\mathbb{1}(i_s \in \mathcal{M}_2)\nabla_W \bar{\ell}'(y^{(i_s)} f_s(x^{(i_s)}))\|_F \\
&\lesssim \frac{\eta_2\sqrt{\beta}t}{\sqrt{d}} + \eta_2 \sum_{s\leq t:\rho_s \text{ large}} \|\mathbb{1}(i_s \in \mathcal{M}_2)\nabla_W \bar{\ell}'(y^{(i_s)} f_s(x^{(i_s)}))\|_F \\
&\lesssim \frac{\eta_2\sqrt{\beta}t}{\sqrt{d}} + \frac{\eta_2(r\beta^4)}{\sqrt{d}}
\end{aligned}$$

Again, substituting in $\beta = O\left(\frac{d^{-1/32}}{(\epsilon_2')^2}\right)$ and recalling that $t \lesssim t_3 = O\left(\frac{d}{\eta_2\epsilon_2'}\right)$, we get the desired bound in Lemma 13.

$\blacksquare$

We now give a lemma, which states that the signal $\bar{W}_t$ must mostly lie in the span of the data points in $\bar{\mathcal{M}}_2$. We've already shown that the remainder is small in some sense from proving Lemma 13, so we formalize this argument below.

Lemma 56 *For $t \lesssim t_3$, there exists scalars α_i for every $i \in \bar{\mathcal{M}}_2$ such that with high probability, for all $j \in [m]$,*

$$[\bar{W}_t]_j = w_j \sum_{i\in\bar{\mathcal{M}}_2} \alpha_i x_1^{(i)} \mathbb{1}([\tilde{W}]_0 x_1^{(k)}) + [P_t]_j$$

it holds that $\|P_t\|_F \lesssim \frac{d^{31/64}}{{\epsilon_2'}^2}$

Proof Note that for $j \in [m]$,

$$\begin{aligned}
[\bar{W}_t^{(2)}]_j &= -\eta_2 \sum_{s=1}^{t} \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})}[B_{s-1}]_j \cdot \mathbb{1}(i_{s-1} \in \mathcal{M}_2) \\
&= -\eta_2 \sum_{s=1}^{t} \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})} w_j \mathbb{1}([W_{s-1}]_j x_1^{(i_{s-1})}) x_1^{(i_{s-1})} \cdot \mathbb{1}(i_{s-1} \in \mathcal{M}_2)
\end{aligned}$$

This implies that

$$\begin{aligned}
[\bar{W}_t]_j &= [\bar{W}_t^{(2)}]_j - \eta_2 \sum_{s=1}^{t} \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})} w_j \mathbb{1}([W_{s-1}]_j x_1^{(i_{s-1})}) x_1^{(i_{s-1})} \cdot \mathbb{1}(i_{s-1} \in \bar{\mathcal{M}}_2) \\
&= [\bar{W}_t^{(2)}]_j \\
&\quad - \eta_2 \sum_{s=1}^{t} \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})} w_j \mathbb{1}([W_0]_j x_1^{(i_{s-1})}) x_1^{(i_{s-1})} \cdot \mathbb{1}(i_{s-1} \in \bar{\mathcal{M}}_2) \\
&\quad - \eta_2 \sum_{s=1}^{t} \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})} w_j (\mathbb{1}([W_{s-1}]_j x_1^{(i_{s-1})}) - \mathbb{1}([W_0]_j x_1^{(i_{s-1})})) x_1^{(i_{s-1})} \cdot \mathbb{1}(i_{s-1} \in \bar{\mathcal{M}}_2)
\end{aligned}$$

We note that for large t, because of our choice of small η_2, we get this this is roughly

$$\begin{aligned}
[\bar{W}_t^{(2)}]_j - \eta_2 \sum_{s=1}^{t} \nabla_W \bar{L}_{\bar{\mathcal{M}}_2} &\le [\bar{W}_t^{(2)}]_j - \eta_2 w_j \sum_{s=1}^{t} \frac{1}{N} \sum_{i \in \bar{\mathcal{M}}_2} \nabla_W \bar{\ell}(y^{(i)} f_s(x^{(i)})) \\
&\le [\bar{W}_t^{(2)}]_j - \eta_2 w_j \sum_{s=1}^{t} \frac{1}{N} \Big(\sum_{i \in \bar{\mathcal{M}}_2} \bar{\ell}'(y^{(i)} f_s(x^{(i)})) \cdot \mathbb{1}([W_0]_j x_1^{(i)}) x_1^{(i)} \\
&+ \sum_{i \in \bar{\mathcal{M}}_2} \bar{\ell}'(y^{(i)} f_s(x^{(i)})) \cdot [\mathbb{1}([W_s]_j x_1^{(i)}) - \mathbb{1}([W_0]_j x_1^{(i)})] x_1^{(i)} \Big)
\end{aligned}$$

We note that for large t, because of our choice of small η_2, we can bound the last term as follows:

$$\begin{aligned}
&\left\| \sum_{s=1}^{t} \bar{\ell}'(y^{(i_{s-1})} f_{s-1}(x^{(i_{s-1})})) \cdot y^{(i_{s-1})} w_j (\mathbb{1}([W_{s-1}]_j x_1^{(i_{s-1})}) - \mathbb{1}([W_0]_j x_1^{(i_{s-1})})) x_1^{(i_{s-1})} \cdot \mathbb{1}(i_{s-1} \in \bar{\mathcal{M}}_2) \right\|_2 \\
&\lesssim pt \|(\mathbb{1}([W_{t-1}]_j x_1^{(i_{t-1})}) - \mathbb{1}([W_0]_j x_1^{(i_{t-1})})) \cdot x_1^{(i_{t-1})}\|_2 \\
&\lesssim pt\epsilon_1^2
\end{aligned}$$

Combining everything together, we obtain

$$\|P_t\|_F \le \|\bar{W}_t^{(2)}\|_F + \epsilon_1^2 \lesssim \frac{d^{31/64}}{{\epsilon_2'}^2}$$

■

Finally, we can simply apply Lemma 5.3 from Li et al. [19], which exactly recovers Lemma 14. For clarity of exposition, we include the lemma statement here, adapted as follows.

Lemma 57 *For $t \lesssim t_3$, there exists some $\|\alpha\|_2 = \Omega(\sqrt{pN})$ in the span of the $\bar{\mathcal{M}}_2$ datapoints. Then, it holds that with high probability,*

$$r_t(x_1) - r_t(-x_1) = 2\alpha^T x_1 \pm \tilde{O}\left(\frac{d^{-1/64}}{{\epsilon_2'}^2}\right)$$

G.4. Proof of Theorem 5

Using the lemmas we have showed, we can now prove the theorem for LNSGD-S. In particular, this last lemma tells us that the predictions at the time training stops are still not good on the $\mathcal{P}$ component, because the margin is heavily influenced by noise.

Proof Refer to Theorem 3.5 in Li et al. [19] for the classification lower bound. ■

Appendix H. Proof of Theorem 6 (FBGD-LS)

H.1. Proof of Lemma 15

Proof Here, we construct a partial target solution of $\mathcal{Q}$. In particular, we use the notation from Lemma 51 to construct $V^\star$. Here, we take $W^\star = 0$ and

$$V_i^\star = \begin{cases} -\frac{40c\log(1/\epsilon_1)v_i}{r}z & \text{if } i \in \mathcal{E}_2 \\ 0 & \text{otherwise} \end{cases}$$

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

$$N_{V_t}(v, V^\star; x_2) = \frac{1}{m}\left(-\frac{40c\log(1/\epsilon_1)}{r}|\mathcal{E}_2|\right) \leq -40c\log(1/\epsilon_1)$$

for some constant c. Essentially, only the $z - \zeta$ vectors in $\mathcal{Q}$ are currently classified correctly. For the remainder of this lemma, a similar analysis can be performed as Lemma 51 to obtain the desired upper bound on time steps of $O\left(\frac{1}{\eta_1 r}\right)$, by considering Lemma 34 for full batch gradient descent. ■

H.2. Proof of Lemma 16

We now proceed to show convergence to a partial solution on the $\mathcal{P}$ component as well.
Proof Here, we construct our target signal $W^\star$ to be the similar to Lemma 12; the difference is that here, it is a partial solution. In particular, we define β similarly to Lemma 12, but here $W^\star = w_i\beta$; $V^\star = 0$ here. It can be verified that by following Lemma 34, we get the upper bound on iterations to be $O\left(\frac{pN}{\eta_1}\right)$. ■

As a result, we observe that at annealing time, both $\mathcal{P}$ and $\mathcal{Q}$ have a partial signal. Contrasting this case with LNSGD-LS, we see that we enter phase 2 of the algorithm with indeed a much bigger signal in $\mathcal{Q}$ (as constructed by $V^\star$), due to the lack of noise in the gradients.

H.3. Proof of Phase 2 of FBGD-LS

We refer readers to the proofs of Lemma 11, Lemma 12, Lemma 13, and Lemma 14 for algorithm LNSGD-S; the results and target signals for the post-annealing phase of FBGD-LS are identical to those lemmas, including the proof of the classification lower bound.