Simplicity Bias via Global Convergence of Sharpness Minimization

Khashayar Gatmiry [1] **Zhiyuan Li** [2] **Sashank Reddi** [3] **Stefanie Jegelka** [1]

Abstract

The remarkable generalization ability of neural networks is usually attributed to the implicit bias of SGD, which often yields models with lower complexity using simpler (e.g. linear) and low-rank features (Huh et al., 2021). Recent works have provided empirical and theoretical evidence for the bias of particular variants of SGD (such as label noise SGD) toward flatter regions of the loss landscape. Despite the folklore intuition that flat solutions are 'simple', the connection with the simplicity of the final trained model (e.g. low-rank) is not well understood. In this work, we take a step toward bridging this gap by studying the *simplicity structure* that arises from minimizers of the sharpness for a class of two-layer neural networks. We show that, for any high dimensional training data and certain activations, with small enough step size, label noise SGD always converges to a network that replicates a single linear feature across all neurons; thereby, implying a simple rank one feature matrix. To obtain this result, our main technical contribution is to show that label noise SGD always minimizes the sharpness on the manifold of models with zero loss for two-layer networks. Along the way, we discover a novel property — a local geodesic convexity — of the trace of Hessian of the loss at approximate stationary points on the manifold of zero loss, which links sharpness to the geometry of the manifold. This tool may be of independent interest.

1. Introduction

Overparameterized neural networks trained by stochastic gradient descent (SGD) have demonstrated remarkable generalization ability. The emergence of this ability, even when the network perfectly fits the data and without any explicit regularization, still remains a mystery (Zhang et al., 2017). A line of recent works have attempted to provide a theoretical basis for such observations by showing that *SGD has an implicit bias toward functions with "low complexity" among all possible networks with zero training loss* (Neyshabur et al., 2017; Soudry et al., 2018; Gunasekar et al., 2017; Li et al., 2018; Gunasekar et al., 2018; Woodworth et al., 2020; Li et al., 2019; HaoChen et al., 2021; Pesme et al., 2021). In particular, it is noted that SGD tends to pick simpler features to fit the data over more complex ones (Hermann & Lampinen, 2020; Kalimeris et al., 2019; Neyshabur et al., 2014; Pezeshki et al., 2021). This tendency is typically referred to as *simplicity bias*. Notably, Huh et al. (2021) empirically observes that the feature matrix on the training set tends to become low-rank and that this property is amplified in deeper networks.

On the other hand, recently, a different form of implicit bias of SGD based on the sharpness of loss landscape has garnered interest. The study of flatness of the loss landscape and its positive correlation with generalization is not new (e.g. (Hochreiter & Schmidhuber, 1997)), and has been supported by a plethora of empirical and theoretical evidence (Keskar et al., 2016; Dziugaite & Roy, 2017; Jastrzebski et al., 2017; Neyshabur et al., 2017; Wu et al., 2018; Jiang et al., 2019; Blanc et al., 2019; Wei & Ma, 2019a;b; HaoChen et al., 2020; Foret et al., 2021; Damian et al., 2021; Li et al., 2021; Ma & Ying, 2021; Ding et al., 2022; Nacson et al., 2022; Wei et al., 2022; Lyu et al., 2022; Norton & Royset, 2021). It has been suggested that, after reaching zero loss, neural networks trained using some variants of SGD are biased toward regions of the loss landscape with low sharpness (Blanc et al., 2019; Damian et al., 2021; Li et al., 2021; Arora et al., 2022; Lyu et al., 2022; Li et al., 2022a; Wen et al., 2022; Bartlett et al., 2022). In particular, Blanc et al. (2020) discovered the sharpness-reduction implicit bias of a variant of SGD called *label noise SGD*, which explicitly adds noise to the labels. They proved that SGD locally diverges from points that are not stationary for the trace of the Hessian of the loss after fitting the training data. Along this line, Li et al. (2021) show that in the limit of step size going to zero, label noise SGD converges to a gradient flow according to the trace of Hessian of the loss after reaching almost zero loss. This flow can be viewed as a Riemannian gradient flow with respect to the sharpness

[1]Massachusetts Institute of Technology [2]Toyota Technological Institute at Chicago [3]Google Research. Correspondence to: Khashayar Gatmiry <gatmiry@mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

on the manifold of zero loss, *i.e.*, the flow following the gradient of sharpness projected onto the tangent space of the manifold. Furthermore, a standard Bayesian argument based on minimum description length suggests such flat minimizers correspond to "simple" models.

From a conceptual point of view, both simplicity bias and sharpness minimization bias suggest that SGD learns *simple* solutions; however, the nature of "simplicity" is different. Simplicity bias is defined in the feature space, meaning the model learns simple features, *e.g.*, low-rank features, but low sharpness is defined in the parameter space, meaning the loss landscape around the learned parameter is flat, such that the model is resilient to random parameter perturbations and admits a short and simple description from a Bayesian perspective. It is thus an interesting question whether these two notions of "simplicity" for neural networks have any causal relationship. Existing works that attempt to study this connection are restricted to somewhat contrived settings (e.g. (Blanc et al., 2020; Shallue et al., 2018; Szegedy et al., 2016; Shallue et al., 2018; HaoChen et al., 2021)). For instance, Blanc et al. (2020) study 1-D two-layer ReLU networks and two-layer sigmoid networks trained on a single data point, where they show the stationary points of the trace of Hessian on the manifold of zero loss are solutions that are simple in nature. They further prove that label noise SGD locally diverges from non-stationary points. As another example, Li et al. (2021) and Vivien et al. (2022) independently show that label noise SGD can recover the sparse ground truth, but only for a simple overparameterized quadratic linear model. In this paper, we study the following fundamental question:

> *(1) Is there a non-linear general setting where the sharpness-reduction implicit bias* provably *implies simplicity bias?*

However, even if we can show that a model with minimum flatness is a simple model, a key open question is whether the sharpness-reduction implicit bias, characterized by the Riemannian gradient flow of sharpness in (Li et al., 2021), even successfully minimizes the sharpness in the first place. The convergence of the gradient flow so far is only known in very simple settings, *i.e.*, quadratically overparametrized linear nets (Li et al., 2021). In this regard, a central open question in the study of sharpness minimization for label noise SGD is:

> *(2) Does label noise SGD converge to the global minimizers of the sharpness on the manifold of zero loss, and if it does, how fast does it converge? In particular, for what class of activations does global convergence happen?*

In this paper, we show sharpness regularization provably leads to simplicity bias for two-layer neural networks with certain activation functions. We further prove that the Riemannian gradient flow on the manifold linearly converges to the global minimizer of the sharpness *for all initialization*. To our best knowledge, this is the first global linear convergence result for the gradient flow of trace of Hessian on manifolds of minimizers.

More formally, we consider the mean squared loss

$$\mathcal{L}(\theta) = \sum\nolimits_{i=1}^{n} (\sum\nolimits_{j=1}^{m} \phi(\theta_j^\top x_i) - y_i)^2 \tag{1}$$

for a two-layer network model where the weights of the second layer are fixed to one and $\{x_i, y_i\}_{i=1}^n$ are the training dataset. [1] We consider label noise SGD on $\mathcal{L}$, that is, running gradient descent on the loss $\sum_{i=1}^n (\sum_{j=1}^m \phi(\theta_j^\top x_i) - y_i + \zeta_{t,i})^2$ at each step t, in which independent Gaussian noise $\zeta_{t,i} \sim \mathcal{N}(0, \sigma^2)$ is added to the labels. The following is a corollary of the main results of our paper, which proves the asymptotic convergence of label noise SGD to the global minimizer of the sharpness. This result holds under Assumptions 3.1, 3.2, 3.4) and for sufficiently small learning rate.

Theorem 1.1 (Convergence to simple solutions from any initialization)**.** *Given any $\epsilon > 0$ and arbitrary initial parameter $\theta[0]$, for label noise SGD with any noise variance $\sigma^2 > 0$ initialized at $\theta[0]$, there exists a threshold η_0 such that for any learning rate $\eta \le \eta_0$, running label noise SGD with step size η and number of iterations $T = \tilde{O}(\log(1/\epsilon) \cdot \eta^{-2}\sigma^{-2})$, with probability at least $1 - \epsilon$,*

1. *$\mathcal{L}(\theta[T]) \le \epsilon$;*
2. *$\mathrm{Tr}(D^2\mathcal{L}(\theta[T])) \le \inf_{\theta:\mathcal{L}(\theta)=0} \mathrm{Tr}(D^2\mathcal{L}(\theta)) + \epsilon$;*
3. *For all neurons $\theta_j[T]$ and all data points x_i, $\left|\theta_j[T]^\top x_i - \phi^{-1}(y_i/m)\right| \le \epsilon$,*

where $D^2\mathcal{L}$ is the Hessian of the loss and we hide constants about loss $\mathcal{L}$ and initialization $\theta[0]$ in $O(\cdot)$.

Informally, Theorem 1.1 states that, first, label noise SGD reaches zero loss and then (1) successfully minimizes the sharpness to its global optima (2) and more importantly, it recovers the simplest possible solution that fits the data, in which the feature matrix is rank one (see Appendix B.3 for proof details). In particular, Theorem 1.1 holds for pure label noise SGD without any additional projection. Additionally, if we initialize the weights at zero, then because the update of label noise SGD is always in the span of the data points, the weights remain in the data span as well. On the other hand, property 3 in Theorem 1.1 states that

[1] We discuss how our result generalizes to the case of unequal weights at the end of Section 4.2.

Results	Assumptions	Basic Description
Theorem 3.3	$\phi', \phi''' > 0$ or $\phi = x^3$ X is full-rank	The global minimizer of the squared loss, all neurons collapse into the same vector in the data subspace
Theorem 3.5	$\phi', \phi''' > 0$ $X^\top X \geq \mu I$ β-normality [3.4]	The Riemannian gradient flow of sharpness converges to the global minimizer of sharpness on the manifold of zero loss
Theorem 1.1	$\phi', \phi''' > 0$ or $\phi = x^3$ $X^\top X \geq \mu I$ β-normality [3.4]	label noise SGD with small enough step size converges to the global minimizer of the sharpness on the manifold of zero loss

Table 1. Summary of main theoretical results and key assumptions

the dot products of all the neurons with each single data point converge to each other, which means all the neurons collapse into the same vector. Finally note that Theorem 1.1 is obtained by combining Theorem 4.6 in (Li et al., 2021) with our main result in this paper, which concerns the global convergence of the Riemannian gradient flow of sharpness (Theorems 3.3 and 3.5). See section A for a detailed derivation of Theorem 1.1 from these results. To summarize, **our technical contributions cover the following two aspects:**

- **Simplicity bias:** We prove that at every minimizer of the trace of Hessian, the pre-activation of different neurons becomes the same for every training data. In other words, all of the feature embeddings of the dataset with respect to different neurons collapse into a single vector, *e.g.*, see Figure 1. This combined with our convergence result mathematically proves that the low-rank feature observation by (Huh et al., 2021) holds in our two-layer network setting trained by sharpness minimization induced regularizer algorithms, such as label noise SGD and 1-SAM.

- **Convergence analysis:** Under an additional normality assumption on the activation (Assumption 3.2), we show that the gradient flow of the trace of Hessian converges exponentially fast to the global optima on the manifold of zero loss, and as a result, label noise SGD with a sufficiently small step size successfully converges to the global minimizer of trace of Hessian among all the models achieving zero loss. Importantly, we do not assume additional technical conditions (such as PL inequality and non-strict saddle property used in prior works) about the landscape of the loss or its sharpness, and our constants are explicit and only depend on the choice of the activation and coherence of the data matrix. Moreover, our convergence results hold in the strong sense i.e., the convergence holds for the last iterations of label noise SGD.

  We have summarized our main results in Table 1. The novelty of our approach (after that we show it converges to zero loss) is that we characterize the convergence on the manifold of zero loss in two phases (1) the first phase where the algorithm is far from stationary points, trace of Hessian decreases with a at least constant rate and there is no convexity in trace of Hessian, (2) the second phase where the algorithm gets close to stationarity, we prove a novel g-convexity property of trace of Hessian which holds only locally on the manifold, but we show implies an exponentially fast convergence rate by changing the Lyapunov potential from value of trace of Hessian to the norm squared of its gradient on the manifold. We further prove that approximate stationary points are close to a global optimum via a semi-monotonicity property.

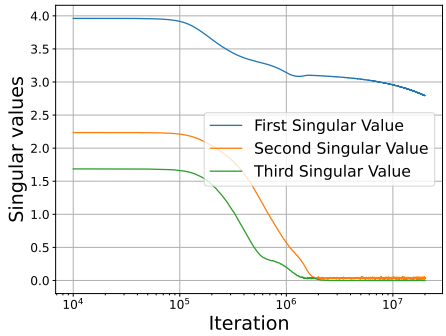

Figure 1. The rank of the feature and weight matrices go to one as Theorem 1.1 predicts in the high dimensional setting $d > n$.

Interestingly, we observe this simplicity bias even beyond the regime of our theory; namely when the number of data points exceeds the ambient dimension, instead of a single vector, the feature embeddings of the neurons cluster and collapse into a small set of vectors (see Figure 2).

2. Related Work

Implicit Bias of Sharpness Minimization. Recent theoretical investigations, including those by Blanc et al. (2019), Damian et al. (2021), and Li et al. (2021), have indicated that Stochastic Gradient Descent (SGD) with label noise intrinsically favors local minimizers with smaller Hessian traces. This is under the assumption that these minimizers are connected locally as a manifold and these analyses focus on the final phase of training when iterates are close to the manifold of minimizes. The effect of label noise or mini-

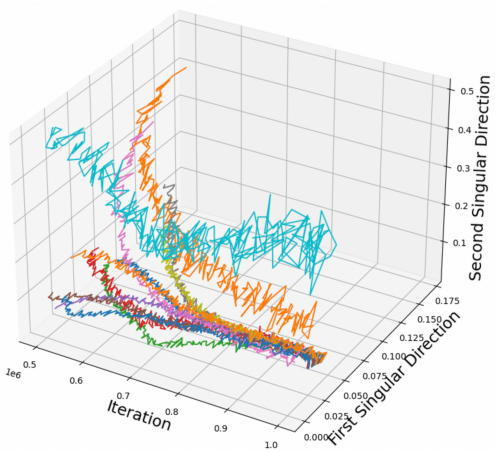

Figure 2. The feature embeddings of neurons collapse into clusters in the low dimensional regime $d < n$. The plot visualizes this phenomenon using the first and second principal components of the feature embedding $\Theta^* X$.

batch noise in the central phase of training, *i.e.*, when the loss is still substantially above zero is more difficult and is only known for simple quadratically overparametrized linear models (Vivien et al., 2022; Andriushchenko et al., 2023b). Further, Arora et al. (2022) demonstrated that normalized Gradient Descent (GD) intrinsically penalizes the Hessian's largest eigenvalue. Ma et al. (2022) proposes that such sharpness reduction phenomena could also be triggered by a multi-scale loss landscape. In the context of scale-invariant loss functions, Lyu et al. (2022) found that GD with weight decay implicitly reduces the spherical sharpness, defined as the Hessian's largest eigenvalue evaluated at the normalized parameter.

Another line of research examines the sharpness minimization effect of large learning rates, assuming that (stochastic) gradient descent converges at the end of training. This analysis primarily hinges on the concept of linear stability, as referenced in works by Wu et al. (2018), Cohen et al. (2021), Ma & Ying (2021), and Cohen et al. (2022). More recent theoretical analyses, such as those by Damian et al. (2022) and Li et al. (2022b), suggest that the sharpness minimization effect of large learning rates in gradient descent does not necessarily depend on the convergence assumption and linear stability. Instead, they propose a four-phase characterization of the dynamics in the so-called Edge of Stability regime, as detailed in Cohen et al. (2021).

Sharpness-aware Minimization (SAM, Foret et al. (2021)) is a highly effective regularization method that improves generalization by penalizing the sharpness of the landscape. SAM was independently developed in (Zheng et al., 2021; Norton & Royset, 2021). Recently, Wen et al. (2022); Bartlett et al. (2022) proved that SAM successfully minimizes the worst-direction sharpness under the assumption that the minimizers of training loss connect as a smooth manifold. Wen et al. (2022) also analyze SAM with batch size 1 for regression problems and show that the implicit bias of SAM becomes penalizing average-direction sharpness, which is approximately equal to the trace of the Hessian of the training loss. Andriushchenko et al. (2023a) observe that the SAM update rule for ReLU networks is locally biased toward sparsifying the features leading to low-rank features. However, their analysis is local and does not provide a convergence proof. They further confirm empirically that deep networks trained by SAM are biased to produce low-rank features.

3. Main Results

Problem Setup. In this paper, we focus on the following two-layer neural network model:

$$r_{\theta,\mathsf{NN}}(x) \triangleq \sum_{j=1}^{m} \phi(\theta_j^\top x). \tag{2}$$

Here $\theta = (\theta_1, \dots, \theta_m)$ and ϕ are the set of parameters and activation of the neural network, respectively. As illustrated in Equation (2), we assume that all the second-layer weights are equal to one. At the end of section 4.2 we discuss how our approach in Theorem 3.3 generalizes to any fixed choice of weights in the second layer. We denote the matrix of the weights by $\Theta \in \mathbb{R}^{m\times d}$, whose j^{th} row is θ_j. Given a training dataset $\{(x_i, y_i)\}_{i=1}^n$, we study the mean squared loss

$$\mathcal{L}(\theta) \triangleq \sum_{i=1}^{n} (r_{\theta,\mathsf{NN}}(x_i) - y_i)^2. \tag{3}$$

For simplicity, we assume that the data points have unit norm, i.e. $\forall i \in [n]$, $\|x_i\| = 1$. Let $\mathcal{M}$ denote the zero level set of $\mathcal{L}$: $\mathcal{M} \triangleq \left\{\theta \in \mathbb{R}^{md} \middle| \mathcal{L}(\theta) = 0\right\}$. As we show in Lemma G.2 in the Appendix, $\mathcal{M}$ is well-defined as a manifold due to a mild non-degeneracy condition implied by Assumption 3.1. We equip $\mathcal{M}$ with the standard Euclidean metric in $\mathbb{R}^{md}$.

The starting point of our investigation is the work (Li et al., 2021), which characterizes the minimizer of the trace of the Hessian of the loss, $\mathrm{Tr} D^2\mathcal{L}$, as the implicit bias of label noise SGD by showing that, in the limit of step size going to zero, label noise SGD evolves on the manifold of zero loss according to the following deterministic gradient flow:

$$\frac{d}{dt}\theta(t) \triangleq -\nabla \mathrm{Tr} D^2\mathcal{L}(\theta(t)). \tag{4}$$

Here, ∇ is the Riemannian gradient operator on the manifold $\mathcal{M}$, which is the projection of the normal Euclidean gradient

onto the tangent space of $\mathcal{M}$ at $\theta(t)$. We use this result of (Li et al., 2021) to relate the convergence of label noise SGD to the convergence of the flow (4). Our main object of study in this paper is then the global convergence of this flow. Note that starting from a point $\theta(0)$ on $\mathcal{M}$, $\theta(t)$ remains on $\mathcal{M}$ for all times t (for a quick recap on basic notions in Differential Geometry such as covariant derivative and gradient on a manifold, we refer the reader to Appendix I.) On a manifold, similar to Euclidean space, having a zero gradient $\|\nabla \mathrm{Tr} D^2\mathcal{L}(\theta^*)\| = 0$ (the gradient defined on the manifold) at some points θ^* is a necessary condition for being a local minimizer, or equivalently the projection of Euclidean gradient onto the tangent space at θ^* has to vanish. More generally, we call θ an ϵ-stationary point on $\mathcal{M}$ if $\|\nabla \mathrm{Tr} D^2\mathcal{L}(\theta)\| \leq \epsilon$.

3.1. Characterization of Stationary Points

Before characterizing the stationary points of $\mathrm{Tr} D^2\mathcal{L}$, we introduce an important assumption on the activation function, namely the convexity and positivity of its derivative.

Assumption 3.1 (Strict positivity and convexity of ϕ')**.** For all $z \in \mathbb{R}$, $\phi'(z) \geq \varrho_1 > 0, \phi'''(z) \geq \varrho_2 > 0$.

For example, Assumption 3.1 is satisfied for $\phi(x) = x^3 + \nu x$ with constants $\varrho_1 = \nu, \varrho_2 = 1$. In Lemma G.1, given the condition $\phi''' > 0$ we show that $\theta(t)$ remains in a bounded domain along the gradient flow, which means having the weaker assumption that $\phi', \phi''' > 0$ automatically implies Assumption 3.1 for some positive constants ϱ_1 and ϱ_2. Another activation of interest to us that happens to be extensively effective in NLP applications is the cube (x^3) activation (Chen & Manning, 2014). Although x^3 does not satisfy the strict positivity required in Assumption 3.1, we separately prove Theorem 3.3 for x^3 (see Lemma G.4 in the Appendix.)

Next, let $X \triangleq \left(x_1 \middle| \dots \middle| x_n\right)$ be the data matrix. We make a coherence assumption on X which requires the input dimension d to be at least as large as the number of data points, denoted by n.

Assumption 3.2 (Data matrix coherence)**.** The data matrix X satisfies $X^\top X \geq \mu I$.

Assumption 3.1 implies that $\mathcal{M}$ is well-defined as a manifold (see Lemma G.2 for a proof). Under aforementioned Assumptions 3.1 and 3.2, we show that the trace of the Hessian regularizer has a unique stationary point on the manifold.

Theorem 3.3 (First-order optimal points)**.** *Under Assumptions 3.1 and 3.2, the first order optimal points and global optimums of* $\mathrm{Tr} D^2\mathcal{L}$ *on* $\mathcal{M}$ *coincide and are equal to the set of all* $\theta^* = \left[\theta_j^*\right]_{j=1}^m$ *such that for all* $i \in [n]$ *and* $j \in [m]$ *satisfy the following:*

$$\theta_j^{*\top} x_i = \phi^{-1}(y_i/m). \tag{5}$$

Note that instead of Assumption 3.2, just having the linear independence of $\{x_i\}_{i=1}^n$ suffices to prove Theorem 3.3, as pointed out in section 4.2. In particular, Theorem 3.3 means that the feature matrix is rank one at a global optimum θ^*, which together with our main convergence result Theorem 1.1 proves the low-rank bias conjecture proposed by (Huh et al., 2021) in the two-layer network setting for label noise SGD and 1-SAM. Notice the $1/m$ effect of the number of neurons on the global minimizers of the implicit bias in Equation (5). While Theorem 3.3 concerns networks with equal second-layer weights, the case of unequal weights can be analyzed similarly, as we discuss after the proof of Theorem 3.3 in Section 4.2.

3.2. Convergence Rate

Next, to state our convergence result, we introduce the key β-normality assumption, under which we bound the rate of convergence of the trace of Hessian to its global minimizer.

Assumption 3.4 (β-normality)**.** For all $z \in \mathbb{R}$ the second derivative of $\phi(z)$ can be bounded by the first and third derivatives as:

$$\beta\phi''(z) \leq \phi'^2(z)\phi'''(z).$$

An example class of activations that satisfies both Assumptions 3.1 and 3.4 is of the form $\phi(z) = z^{2k+1} + \nu z$ for $\nu > 0$, which is well-known in the deep learning theory literature (Li et al., 2018; Jelassi et al., 2022; Allen-Zhu & Li, 2020; Woodworth et al., 2020). Notably, these activations satisfy Assumption 3.4 with normality coefficient $\beta = \min\{\frac{1}{(2k+1)^2(2k-1)}, \nu^2\}$.

Under Assumptions 3.1, 3.2, and 3.4, we show that $\theta(t)$ converges to θ^* exponentially fast in Theorem 3.5. This results from strong local g-convexity of trace of Hessian for approximate stationary points (see Lemmas 4.4 and B.4), plus a semi-monotonicity property for the trace of Hessian (Lemma 4.6).

Theorem 3.5 (Convergence of the gradient flow)**.** *Consider the limiting flow of Label noise SGD on the manifold of zero loss, which is the gradient flow in* (4)*. Then, under Assumptions 3.1,3.2, and 3.4,*

(C.1) *The gradient flow reaches* ϵ*-stationarity for all times*

$$t \geq \frac{\mathrm{Tr} D^2\mathcal{L}(\theta(0))}{\mu\beta^2} + \frac{\log(\beta^2/(\varrho_1^2\varrho_2^2\epsilon^2) \vee 1)}{\varrho_1\varrho_2\mu}.$$

(C.2) *For all $j \in [m]$ and $i \in [n]$, the dot product of the jth neuron to the ith data point becomes ϵ-close to that of any global optimum θ^*:*

$$\left|\theta_j(t)^\top x_i - \theta_j^{*\top} x_i\right| \leq \epsilon, \tag{6}$$

where θ^* is defined in Theorem 3.3.

The formal proof of Theorem 3.5 is given in Section B.3. We sketch its proof in Section 4.3. Note that the rate is logarithmic in the accuracy parameter ϵ, i.e., the flow has exponentially fast convergence to the global minimizer.

4. Proof Sketches

Before delving into the technical details of the proofs, we discuss some necessary background and additional notation used in the proofs.

4.1. Additional Notation

We use $f_i(\theta)$ to denote the output of the network on the ith input x_i, i.e.,

$$f_i(\theta) \triangleq r_{\theta,\mathsf{NN}}(x_i). \tag{7}$$

We use $f(\theta) \triangleq (f_1(\theta), \ldots, f_n(\theta))$ to denote the array of outputs of the network on $\{x_i\}_{i=1}^n$. We use D for the Euclidean directional derivative or Euclidean gradient, and ∇ for the covariant derivative on the manifold or the gradient on the manifold. We denote the Jacobian of f at point θ by $Df(\theta)$ whose ith row is $Df_i(\theta)$, the Euclidean gradient of f_i at θ. Recall the definition of the manifold of zero loss as the zero level set of the loss $\mathcal{L}$, which is the intersection of zero level sets of functions $f_i, \forall i \in [m]$: $\mathcal{M} = \left\{\theta \in \mathbb{R}^{md} | \, \forall i \in [n],\ f_i(\theta) = y_i\right\}$. The tangent space $\mathcal{T}_\theta(\mathcal{M})$ of $\mathcal{M}$ at point θ can be identified by the tangents to all curves on $\mathcal{M}$ passing through θ, and the normal space $\mathcal{T}_\theta^N(\mathcal{M})$ in this setting is just the orthogonal subspace to $\mathcal{T}_\theta(\mathcal{M})$. We denote the projection operators onto $\mathcal{T}_\theta(\mathcal{M})$ and $\mathcal{T}_\theta^N(\mathcal{M})$ by P_θ and P_θ^N, respectively. For a set of vectors $\{v_i\}_{i=1}^n$ where for all $i \in [n]$, $v_i \in \mathbb{R}^d$, we use the notation $\left[v_i\right]_{i=1}^n$ to denote the vector in $\mathbb{R}^{nd}$ obtained by stacking vectors $\{v_i\}_{i=1}^n$.

4.2. Proof Sketch of Theorem 3.3

In this section, we describe the high-level proof idea of Theorem 3.3. Note that Theorem 3.3 characterizes the first order stationary points of $\mathrm{Tr}D^2\mathcal{L}$ on $\mathcal{M}$, i.e., points θ^* on $\mathcal{M}$ for which $\nabla \mathrm{Tr}D^2\mathcal{L}(\theta^*) = 0$. The starting point of the proof is the observation that the gradients $Df_i(\theta)$ form a basis for the normal space at θ.

Lemma 4.1 (Basis for the normal space)**.** *For every θ, the set of vectors $\{Df_i(\theta)\}_{i=1}^n$ form a basis for the normal space $\mathcal{T}_\theta^N(\mathcal{M})$ of $\mathcal{M}$ at θ.*

For a stationary point θ^* with $\nabla F(\theta^*) = 0$, the Euclidean gradient at θ^* should be in the normal space $\mathcal{T}_{\theta^*}^N(\mathcal{M})$. But by Lemma 4.1, this means there exist coefficients $\{\alpha_i\}_{i=1}^n$ such that

$$D(\mathrm{Tr}D^2\mathcal{L})(\theta^*) = \sum_{i=1}^n 2\alpha_i Df_i(\theta^*). \tag{8}$$

To further understand what condition (8) means for our two-layer network (2), we first state a result from prior work in Lemma 4.2 (see e.g. (Blanc et al., 2020)) regarding the trace of the Hessian of the mean squared loss $\mathcal{L}$ for some parameter $\theta \in \mathcal{M}$ (see Section C.2 for a proof).

Lemma 4.2. *For the loss on the dataset defined in Equation (3) and parameters θ with $\mathcal{L}(\theta) = 0$, we have*

$$\mathrm{Tr}D^2\mathcal{L}(\theta) = \sum_{i=1}^n \|Df_i(\theta)\|^2.$$

Using Lemma 4.2, we can explicitly calculate the trace of Hessian for our two-layer network in (2).

Lemma 4.3 (Trace of Hessian in two-layer networks)**.** *For the neural network model (2) and the mean squared loss $\mathcal{L}$ in Equation (3), for any θ with $\mathcal{L}(\theta) = 0$ we have*

$$\mathrm{Tr}D^2\mathcal{L}(\theta) = \sum_{i=1}^n \sum_{j=1}^m \phi'(\theta_j^\top x_i)^2. \tag{9}$$

Now we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. From Equation (8) and by explicitly calculating the gradients $\{Df_i(\theta^*)\}_{i=1}^n$ using Lemma 4.3, we have

$$\Big[\sum_{i=1}^n 2\phi'(\theta_j^{*\top} x_i)\phi''(\theta_j^{*\top} x_i)x_i\Big]_{j=1}^m = \sum_{i=1}^n 2\alpha_i \Big[\phi'(\theta_j^{*\top} x_i)x_i\Big]_{j=1}^m.$$

But using our assumption that the data points $\{x_i\}_{i=1}^n$ are linearly independent, we have for all $i \in [n]$ and $j \in [m]$, $\phi''(\theta_j^{*\top} x_i) = \alpha_i$. Now, because ϕ''' is positive and ϕ'' is strictly monotone, its inverse is well-defined:

$$\theta_j^{*\top} x_i = \nu_i \triangleq \phi''^{-1}(\alpha_i), \tag{10}$$

This implies $\phi(\theta_j^{*\top} x_i) = \phi(\nu_i)$. Therefore,

$$y_i = \sum_{j=1}^m \phi(\theta_j^{*\top} x_i) = m\phi(\nu_i). \tag{11}$$

But note that from strict monotonicity of ϕ from Assumption 3.1, we get that it is invertible. Therefore, Equation (11) implies

$$\nu_i = \phi^{-1}(y_i/m), \quad \text{and} \quad \alpha_i = \phi''(\phi^{-1}(y_i/m)).$$

This characterizes the first order optimal points of $\mathrm{Tr}D^2\mathcal{L}$.

On the other hand, note that $\mathrm{Tr}D^2\mathcal{L} \geq 0$, so its infimum over $\mathcal{M}$ is well-defined. Moreover, from Lemma G.1 the set of θ's with $\mathrm{Tr}D^2\mathcal{L}(\theta) \leq c$ for arbitrary c is bounded in $\mathbb{R}^{md}$. Therefore $\mathrm{Tr}D^2\mathcal{L}$ has a global minimizer on $\mathcal{M}$. But the global minimizers are also stationary points and from Equation (10) we see that all of the first order optimal points have the same value of $\mathrm{Tr}D^2\mathcal{L}$. Therefore, all first order optimal points are global optima and satisfy Equation (10). □

Results for General Second Layer Weights: Note that if the weights of the second layer are fixed arbitrarily to $(w_j)_{j=1}^m$, then Equation (10) simply turns into $\theta_j^{*\top} x_i = \phi''^{-1}(\alpha_i/w_j)$; namely, the ϕ'' image of the feature matrix $\Theta^* X$ is rank one. Interestingly, this shows that in the general case, the right matrix to look for obtaining a low-rank structure is the embedding of the feature matrix by the second derivative of the activation (for the cube activation this embedding is the feature matrix itself.) Next, we investigate the convergence of the gradient flow in Equation (4).

4.3. Proof Sketch of Theorem 3.5

The first claim **(C.1)** of Theorem 3.5 is that after time $t \geq \tilde{\Omega}(\log(1/\epsilon))$, the gradient will always be smaller than ϵ. It is actually trivial to prove that there exists some time at most $t = \tilde{O}(1/\epsilon^2)$ such that $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\| \leq \epsilon$ without any assumption, based on a standard argument which is a continuous version of descent lemma. Namely, if the gradient of $\mathrm{Tr}D^2\mathcal{L}$ remains larger than some $\epsilon > 0$ along the flow until time $t > \mathrm{Tr}D^2\mathcal{L}(\theta(0))/\epsilon^2$, then we get a contradiction:

$$\begin{aligned}&\mathrm{Tr}D^2\mathcal{L}(\theta(0)) - \mathrm{Tr}D^2\mathcal{L}(\theta(t))\\ &= \int_0^t \left\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\right\|^2 > \mathrm{Tr}D^2\mathcal{L}(\theta(0)).\end{aligned}$$

Our novelty here is a new characterization of the loss landscape under Assumptions 3.1, 3.2 (Lemma 4.4 proved in Appendix B.1).

Lemma 4.4 (PSD Hessian when gradient vanishes)**.** *Suppose that activation ϕ satisfies Assumptions 3.1, 3.2, and 3.4. Consider a point θ on the manifold where the gradient is small, namely $\|\nabla_\theta \mathrm{Tr}D^2\mathcal{L}(\theta)\| \leq \sqrt{\mu}\beta$. Then, the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ on the manifold is PSD at point θ, or equivalently $\mathrm{Tr}D^2\mathcal{L}$ is locally g-convex on $\mathcal{M}$ around θ.*

Lemma 4.4 implies that whenever the gradient is sufficiently small, the time derivative of the squared gradient norm will also be non-positive:

$$\begin{aligned}&\frac{d}{dt}\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2\\ &= -\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))^\top \nabla^2 \mathrm{Tr}D^2\mathcal{L}(\theta(t)) \nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t)) \leq 0,\end{aligned}$$

Therefore, once the gradient is sufficiently small at some t, it will always remain small for all $t' > t$. In fact, we show that the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ on the manifold is strictly positive with a uniform $\varrho_1\varrho_2\mu$ lower bound in a suitable subspace which includes the gradient (Lemma B.4). Then

$$\begin{aligned}&\frac{d\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2}{dt}\\ &= -\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))^\top \nabla^2 \mathrm{Tr}D^2\mathcal{L}(\theta(t)) \nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\\ &\leq -\varrho_1\varrho_2\mu \|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2,\end{aligned}$$

which further implies a linear convergence of gradient norm:

$$\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2 \leq \|\nabla \mathrm{Tr}D^2\mathcal{L}(\gamma(\theta(t_0)))\|^2 e^{-(t-t_0)\varrho_1\varrho_2\mu}.$$

Next, we state the high level ideas that we use to prove Lemma 4.4, which is the key to proving Theorem 3.5. Before that, we review some necessary background.

Computing the Hessian on $\mathcal{M}$. To prove Lemma 4.4, we calculate the Hessian of $F \triangleq \mathrm{Tr}D^2\mathcal{L}$ on the manifold in Lemma 4.5. The Hessian of a function F on $\mathbb{R}^{md}$ is defined as $D^2F[u,w] = \langle D(DF(\theta))[u], w\rangle$, where $DF(\theta)$ is the usual Euclidean gradient of F at θ, and $D(.)[u]$ denotes directional derivative. To calculate the Hessian on the manifold, one needs to substitute the Euclidean gradient $DF(\theta)$ by the gradient $\nabla F(\theta)$ on the manifold. Moreover, $D(DF(\theta))[u]$ has to be substituted by the covariant derivative $\nabla_u \nabla F(\theta)$, a different differential operator than the usual derivative. We recall the characterization of the covariant derivative as the projection of the Euclidean directional derivative onto the tangent space. For more background on covariant differentiation, we refer the reader to Appendix I.

Fact 1. For vector fields V, W on $\mathcal{M}$, we have $\nabla_V W(\theta) = P_\theta(DW(\theta))[V]$.

We also recall the definition of Hessian $\nabla^2 F$ on $\mathcal{M}$ based on covariant derivative.

Fact 2. The Hessian of F at point θ on $\mathcal{M}$ is given by $\nabla^2 F(w,u) = \langle \nabla_w \nabla F, u\rangle$.

Next, we state Lemma 4.5, which is proved in Appendix B.1.1.

Lemma 4.5 (Hessian of the implicit regularizer on the manifold)**.** *Recall that $\{Df_i(\theta)\}_{i=1}^n$ is a basis for the normal space $\mathcal{T}_\theta^N(\mathcal{M})$ according to Lemma 4.1. Let $\alpha' = (\alpha_i')_{i=1}^n$*

be the coefficients representing $P_\theta^N(D(\mathrm{Tr}D^2\mathcal{L})(\theta)) \in \mathcal{T}_\theta^N(\mathcal{M})$ *in the basis* $\{Df_i(\theta)\}_{i=1}^n$*, i.e.*

$$P_\theta^N\Big(D(\mathrm{Tr}D^2\mathcal{L})(\theta)\Big) = \sum\nolimits_{i=1}^n \alpha_i' Df_i(\theta).$$

Then, the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ *on* $\mathcal{M}$ *can be explicitly written (in the Euclidean chart* $\mathbb{R}^{md}$*) using* α' *as*

$$\begin{aligned}&\nabla^2\mathrm{Tr}D^2\mathcal{L}(\theta)[u,w]\\&= D^2\mathrm{Tr}D^2\mathcal{L}(\theta)[u,w] - \sum\nolimits_{i=1}^n \alpha_i' D^2 f_i(\theta)[u,w], \quad (12)\end{aligned}$$

for arbitrary $u, w \in \mathcal{T}_\theta(\mathcal{M})$*, where* D^2 *and* ∇^2 *denote the normal Euclidean Hessian and Hessian over the manifold.*

Observe in the formula of $\nabla^2(\mathrm{Tr}D^2\mathcal{L})$ in Equation (12) the first term is just the normal Euclidean Hessian of $\mathrm{Tr}D^2\mathcal{L}$ while we get the second "projection term" due to the difference of covariant differentiation with the usual derivative. Finally, to prove the second claim **(C.2)** of Theorem 3.5, we prove Lemma 4.6 in Appendix C.5 which shows that the small gradient norm of $\|\nabla\mathrm{Tr}D^2\mathcal{L}(\theta)\| \le \delta$ implies the (approximate) alignment of features.

Lemma 4.6 (Small gradient implies close to optimum)**.** *Suppose* $\|\nabla\mathrm{Tr}D^2\mathcal{L}(\theta)\| \le \delta$*. Then, for all* $i \in [n], j \in [m]$*,* $\left|\theta_j^\top x_i - \theta_j^{*\top} x_i\right| \le \delta/(\sqrt{\mu}\varrho_1\varrho_2)$*, for* θ^* *in Equation* (5)*.*

5. Experimental Setup

We examine our theory on the convergence to a rank one feature matrix in Figure 1 via a synthetic experiment by considering a network with $m = 10$ neurons on ambient dimension $d = 3$ and $n = 3$ data points. We further pick learning rate $\eta = 0.05$ and noise variance $\sigma = 0.03$ for implementing label noise SGD. Each entry of the data points is generated uniformly on $[0, 1]$, which is the same data generating process in all the experiments. As Figure 1 shows, the second and third eigenvalues converge to zero which is predicted by Theorem 3.3. On the other hand, in a similar setting with the same learning rate $\eta = 0.05$ but larger $\sigma = 0.2$ we run a synthetic experiment with $d = 5$ this time smaller than the number of data points $n = 6$. In this case, the feature matrix cannot be rank one anymore. However, one might guess that it still converges to something low-rank even though our theoretical result is shown only for the high dimensional case; in this case, the smallest possible rank to fit the data is indeed two, and surprisingly we see in Figure 3 that except the first and second eigenvalues of the feature matrix, the rest go to zero. To further test extrapolating our theory to the low dimensional case ($d < n$), this time we use a higher number of data points $n = 15$ with ambient dimension $d = 10$ and 30 neurons, with learning rate $\eta = 0,1$ and noise variance $\sigma = 0.2$. We observe a very interesting property: even though the feature matrix is high rank in this case, the neurons tend to converge to one another, especially when they are close, as if there is a hidden attraction force between them. To visualize our experimental discovery, in Figure 2 we plot the first two principal components (with the largest singular values) of the feature vectors for each neuron, and observe how it changes while running label noise SGD. We further pick 12 neurons for which the first and second components are bounded in $[0, 0.1]$ and $[0, 0.5]$, respectively. As one can see in Figure 2, most of the principal components of the neurons except possibly one outlier are converging to the same point. This is indeed the case for the feature vectors of neurons as well, namely, they collapse into clusters, but unfortunately, we only have two dimensions to visualize this, and we have picked the most two effective indicators (the first and second principal components.)

6. Conclusion

In this paper, we take an important step toward understanding the implicit bias of label noise SGD with trace of Hessian induced regularizer for a class of two layer networks. We discover an intriguing relationship between this induced regularizer and the low-rank simplicity bias conjecture of neural networks proposed in (Huh et al., 2021): we show that by initializing the neurons in the subspace of high-dimensional input data, all of the neurons converge into a single vector. Consequently, for the final model, (i) the rank of the feature matrix is effectively one, and (ii) the functional representation of the final model is *very simple*; specifically, its sub-level sets are half-spaces. To prove this, in spite of the lack of convexity, we uncover a novel structure in the landscape of the loss regularizer: the trace of Hessian regularizer becomes locally geodetically convex at points that are approximately stationary. Furthermore, in the limit of step size going to zero, we prove that label noise SGD or 1-SAM converge *exponentially fast* to a global minimizer.

Furthermore, generalizing the class of activations that enjoy fast convergence or proving the existence of fundamental barriers and handling the case of low-dimensional input data are interesting future directions. Based on the compelling empirical evidence from (Huh et al., 2021) and our results for two-layer networks, we hypothesize that a low-rank simplicity bias can also be shown for deeper networks, possibly using the sharpness minimization framework.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This research was supported in part by Office of Naval Research grant N00014-20-1-2023 (MURI ML-SCOPE), NSF AI Institute TILOS (NSF CCF-2112665), NSF award 2134108, and the Alexander von Humboldt Foundation.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Andriushchenko, M., Bahri, D., Mobahi, H., and Flammarion, N. Sharpness-aware minimization leads to low-rank features. *arXiv preprint arXiv:2305.16292*, 2023a.

Andriushchenko, M., Varre, A. V., Pillaud-Vivien, L., and Flammarion, N. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pp. 903–925. PMLR, 2023b.

Arora, S., Li, Z., and Panigrahi, A. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024. PMLR, 2022.

Bartlett, P. L., Long, P. M., and Bousquet, O. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *arXiv preprint arXiv:2210.01513*, 2022.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *arXiv preprint arXiv:1904.09080*, 2019.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020.

Chen, D. and Manning, C. D. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 740–750, 2014.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability, 2021.

Cohen, J. M., Ghorbani, B., Krishnan, S., Agarwal, N., Medapati, S., Badura, M., Suo, D., Cardoze, D., Nado, Z., Dahl, G. E., et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Damian, A., Ma, T., and Lee, J. D. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Damian, A., Nichani, E., and Lee, J. D. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

Ding, L., Drusvyatskiy, D., and Fazel, M. Flat minima generalize for low-rank matrix recovery. *arXiv preprint arXiv:2203.03756*, 2022.

Do Carmo, M. P. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

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

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Gatmiry, K. and Vempala, S. S. Convergence of the riemannian langevin algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

Gatmiry, K., Kelner, J., and Vempala, S. S. Sampling with barriers: Faster mixing via lewis weights. *arXiv preprint arXiv:2303.00480*, 2023.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

HaoChen, J. Z., Wei, C., Lee, J. D., and Ma, T. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

HaoChen, J. Z., Wei, C., Lee, J., and Ma, T. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pp. 2315–2357. PMLR, 2021.

Hermann, K. and Lampinen, A. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33: 9995–10006, 2020.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Huh, M., Mobahi, H., Zhang, R., Cheung, B., Agrawal, P., and Isola, P. The low-rank simplicity bias in deep networks. *arXiv preprint arXiv:2103.10427*, 2021.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Jelassi, S., Sander, M., and Li, Y. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., and Zhang, H. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47. PMLR, 2018.

Li, Y., Wei, C., and Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pp. 11669–11680, 2019.

Li, Z., Wang, T., and Arora, S. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

Li, Z., Wang, T., and Yu, D. Fast mixing of stochastic gradient descent with normalization and weight decay. *Advances in Neural Information Processing Systems*, 35: 9233–9248, 2022a.

Li, Z., Wang, Z., and Li, J. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *arXiv preprint arXiv:2207.12678*, 2022b.

Lyu, K., Li, Z., and Arora, S. Understanding the generalization benefit of normalization layers: Sharpness reduction. *arXiv preprint arXiv:2206.07085*, 2022.

Ma, C. and Ying, L. On linear stability of sgd and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021.

Ma, C., Wu, L., and Ying, L. The multiscale structure of neural network loss functions: The effect on optimization and origin. *arXiv preprint arXiv:2204.11326*, 2022.

Nacson, M. S., Ravichandran, K., Srebro, N., and Soudry, D. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pp. 16270–16295. PMLR, 2022.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Norton, M. D. and Royset, J. O. Diametrical risk minimization: Theory and computations. *Machine Learning*, pp. 1–19, 2021.

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Pezeshki, M., Kaba, O., Bengio, Y., Courville, A. C., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., and Dahl, G. E. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600*, 2018.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Vivien, L. P., Reygner, J., and Flammarion, N. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Conference on Learning Theory*, pp. 2127–2159. PMLR, 2022.

Wei, C. and Ma, T. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *Advances in Neural Information Processing Systems*, pp. 9722–9733, 2019a.

Wei, C. and Ma, T. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019b.

Wei, C., Chen, Y., and Ma, T. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022.

Wen, K., Ma, T., and Li, Z. How does sharpness-aware minimization minimize sharpness? *arXiv preprint arXiv:2211.05729*, 2022.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Wu, L., Ma, C., and E, W. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Zheng, Y., Zhang, R., and Mao, Y. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8156–8165, 2021.

A. Proof of Theorem 1.1

Theorem 1.1 is a direct combination of Theorem 4.6 of Li et al. (2021), Theorem 3.3, and Theorem 3.5. For the convenience of the reader, we restate Theorem 4.6 of Li et al. (2021) here. To avoid any confusions, we have specialized Theorem 4.6 of Li et al. (2021) into our case for label noise SGD with our notation. We remind the reader that we use $\theta(t)$ for the gradient flow at time t and θ_t for the tth iteration of label noise SGD (Note that we do not necessarily set $\theta(0)$ equal to θ_0, see the theorem below). Before restating Theorem 4.6, we first restate its two prerequisite assumptions.

Assumption A.1 (Assumption 3.1. in Li et al. (2021))**.** Assume that the loss $L : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^3$ function, and that $\mathcal{M}$ is a $(md-n)$-dimensional $\mathcal{C}^2$-submanifold of $\mathbb{R}^{md}$, where for all $\theta \in \mathcal{M}$, θ is a local minimizer of $\mathcal{L}$ and $\text{rank}(\nabla^2\mathcal{L}(\theta)) = n$.

Assumption A.2 (Assumption 3.2. in Li et al. (2021))**.** For arbitrary $\theta \in \mathbb{R}^{md}$, let $\psi(\theta) \in \mathcal{M}$ be the point of convergence of the gradient flow according to $\mathcal{L}$ starting from θ, assuming it exists. Assume that U is an open neighborhood of $\mathcal{M}$ satisfying that gradient flow of $\mathcal{L}$ starting in U converges to some point in $\mathcal{M}$, i.e., $\forall\theta \in U$, $\psi(x) \in \mathcal{M}$.

Theorem A.3 (Theorem 4.6 in Li et al. (2021))**.** *Suppose the loss function $\mathcal{L}$, the manifold of local minimizer $\mathcal{M}$ and the open neighborhood U satisfy Assumptions 3.1 and 3.2 stated above, and $\theta_0 \in U$ is fixed for all choices of step size $\eta > 0$. If the gradient flow of trace of hessian (4) has a global solution $\theta(t)$ with $\theta(0) = \psi(\theta_0)$ and $\theta(t)$ never leaves U, i.e. $\mathbb{P}(\theta(t) \in U, \forall t \geq 0) = 1$. Then for any $T > 0$, $\theta_{\lfloor T/(\eta^2\sigma^2)\rfloor}$ converges in distribution to $\theta(T)$ as $\eta \to 0$.*

First note that from Lemma D.1, the gradient flow in the limit of step size going to zero always converges to the manifold of zero, independent of the initialization. Therefore, we can set the neighborhood U in Theorem 4.6 of (Li et al., 2021) to be the whole $\mathbb{R}^{md}$. Now we need to check Assumptions 3.1. and 3.2. in (Li et al., 2021).

Checking Assumption A.1 The smoothness of the squared loss $\mathcal{L}$ defined in Equation (3) follows from the smoothness of the activation function ϕ. Furthermore, as we derived in Equation (22), we have $\nabla^2\mathcal{L} = 2\sum_{i=1}^n Df_i(\theta)Df_i(\theta)^\top$, and from the non-singularity of the Jacobian of f stated in Lemma G.3 and the linear independence of the inputs x_i's due to Assumption (3.2), we get that Df_i's are always linearly independent, and as a result, the Hessian $\nabla^2\mathcal{L}$ is always rank n. Combining these with the fact that $\mathcal{M}$ is a well-defined smooth submanifold of $\mathbb{R}^{md}$ due to Lemma G.2 fully checks Assumption A.1.

Checking Assumption A.2 We take U to be the entire $\mathbb{R}^{md}$. Then according to Lemma D.1, starting from any θ_0, the gradient flow of the loss converges to a point on the manifold $\mathcal{M}$.

Since Assumptions A.1 and A.2 are satisfied, we can apply Theorem A.3: namely, Theorem A.3 implies there exists a threshold η_0 such that for step size $\eta \leq \eta_0$, we have $\theta_{\lceil t/\eta^2\rceil}$ is δ to $\theta(t)$ in distribution, where recall that $\theta(t)$ is the Riemannian gradient flow according to the trace of Hessian defined in Equation (4), initialized at $\psi(\theta_0)$.

In order to obtain Lipschitz bounds for $\theta(t)$ which evolves according to (4), we need to obtain information about the initial point $\theta(0)$ on the manifold, which is the limiting point $\tilde{\theta}$ of the trace of hessian gradient flow. This limiting point is defined in Lemma D.1. In particular, in that Lemma we show that the trace of Hessian of the final point is bounded as

$$\text{Tr}D^2\mathcal{L}(\theta(0)) = \text{Tr}D^2\mathcal{L}(\tilde{\theta}) \leq 2\text{Tr}D^2\mathcal{L}(\theta_0) + \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\theta_0)},$$

for $C = 4m\mu\varrho_1^2$. On the other hand, note that from Theorem 3.5, $\theta(t)$ converges to $\tilde{\epsilon}$-proximity of a global minimizer θ^* of trace of Hessian on $\mathcal{M}$ in the sense of Equation (6) after time at most

$$\frac{\text{Tr}D^2\mathcal{L}(\theta(0))}{\mu\beta^2} + \frac{\log(\beta^2/(\varrho_1^2\varrho_2^2\tilde{\epsilon}^2)\vee 1)}{\varrho_1\varrho_2\mu} \leq \frac{2\text{Tr}D^2\mathcal{L}(\theta_0) + \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\theta_0)}}{\mu\beta^2} + \frac{\log(\beta^2/(\varrho_1^2\varrho_2^2\tilde{\epsilon}^2)\vee 1)}{\varrho_1\varrho_2\mu} := \tilde{T}.$$

Now using Theorem A.3 for $T = \tilde{T}$, we get that $\theta_{\lfloor\tilde{T}/(\eta^2\sigma^2)\rfloor}$ converges to $\theta(\tilde{T})$ in distribution. But $\theta(\tilde{T})$ is a deterministic point mass, so convergence in distribution implies convergence in probability. Namely, for every $\tilde{\epsilon} > 0$, there exists a threshold η_0 such that for step size $\eta \leq \eta_0$, with probability $1-\epsilon$,

$$\left|\theta(\tilde{T})_j^\top x_i - \theta_{\lfloor\tilde{T}/(\eta^2\sigma^2)\rfloor}{}_j^\top x_i\right| \leq \tilde{\epsilon}. \tag{13}$$

On the other hand, from Theorem 3.5, after $\lfloor\tilde{T}/(\eta^2\sigma^2)\rfloor$ iterations of label noise SGD, with probability at least $1-\epsilon$, $\theta_{\tilde{T}j}{}^\top x_i$ is close to $\theta_j^{*\top} x_i$ for all $j \in [m]$ and $i \in [n]$:

$$\left|\theta(\tilde{T})_j^\top x_i - \theta^*{}_j^\top x_i\right| \leq \tilde{\epsilon}. \tag{14}$$

Combining Equations (14) and (13):

$$\left|{\theta^*}_j^\top x_i - \theta_{\lfloor \tilde{T}/(\eta^2\sigma^2)\rfloor_j}^\top x_i\right| \leq 2\tilde{\epsilon}. \tag{15}$$

Moreover, note that from Lemma G.1, there exists δ', ϵ' depending on the activation ϕ such that for all $i \in [n], j \in [m]$ and times $t \geq 0$:

$$\left|\theta(t)_j^\top x_i - z^*\right| \leq \mathrm{Tr}D^2\mathcal{L}(\theta(0))/(\epsilon'\delta') + \epsilon'/2 \leq \left(2\mathrm{Tr}D^2\mathcal{L}(\theta_0) + \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\theta_0)}\right)/(\epsilon'\delta') + \epsilon'/2 := R,$$

where z^* is the global minimizer of ϕ' and we defined the RHS to be parameter R. This means that $\theta(t)_j^\top x_i$ always remains within distance R from z^* along the gradient flow on the manifold. Combining this with Equation (13) implies that the points $\theta(\tilde{T})_j^\top x_i$, $\theta_{\lfloor \tilde{T}/(\eta^2\sigma^2)\rfloor_j}^\top x_i$, and ${\theta^*}_j^\top x_i$ are within distance at most $R + \tilde{\epsilon}$ of z^*. On the other hand, the functions $\mathrm{Tr}D^2\mathcal{L}$ and $\mathcal{L}$ are smooth, and in particular have bounded Lipschitz constants L_1, L_2 in the hypercube $[-(R+\tilde{\epsilon}), (R+\tilde{\epsilon})]^{mn}$. Using these Lipschitz constants with Equation (15) implies

$$\left|\mathrm{Tr}D^2\mathcal{L}(\theta_{\lfloor \tilde{T}/\eta^2\rfloor}) - \mathrm{Tr}D^2\mathcal{L}(\theta^*)\right| \leq 2L_1\tilde{\epsilon},$$
$$\left|\mathcal{L}(\theta_{\lfloor \tilde{T}/\eta^2\rfloor}) - \mathcal{L}(\theta^*)\right| \leq 2L_2\tilde{\epsilon}.$$

Picking $\tilde{\epsilon} \leq \min(\epsilon/(2L_1), \epsilon/(2L_2))$ completes the proof.

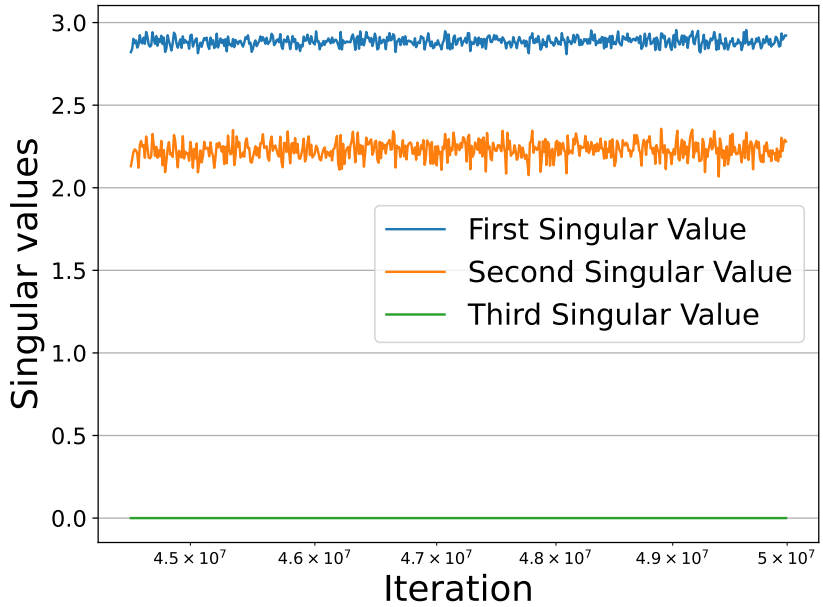

Figure 3. The feature embeddings of neurons collapse into clusters in the low dimensional regime $d < n$. The plot visualizes this phenomenon using the first and second principal components of the feature embedding $\Theta^* X$.

B. Full proof of Theorem 3.5

In order to prove Theorem 3.5, we take these steps:

- In Section B.1, we show that the norm of the gradient of $\mathrm{Tr}D^2\mathcal{L}$ becomes decreasing at approximate stationary points in Lemma B.1. To this end, we show that under Assumptions 3.1, 3.2, 3.4, at each point $\theta \in \mathcal{M}$, we either have a large enough gradient or we are in a neighborhood of the global optima of the trace of Hessian in which trace of Hessian becomes g-convex as stated in Lemma 4.4.

- In Section B.2, Lemma B.3, we show that the norm of the gradient of $\mathrm{Tr}D^2\mathcal{L}$ decays exponentially fast under Assumptions 3.1, 3.2, and 3.4. To show this, we prove Lemma B.4 in which we show under Assumptions 3.1, 3.2,and 3.4 the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ on $\mathcal{M}$ becomes positive definite at approximate stationary points.

- Finally, we combine Lemmas B.1, B.3, and 4.6 to show Theorem 3.5 in Appendix B.3.

B.1. Local g-convexity at approximate stationary points

In this section we prove Lemma B.1.

Lemma B.1 (PSD Hessian when gradient vanishes)**.** *The norm of the gradient $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|$ on the gradient flow (4) becomes decreasing with respect to t at $\sqrt{\mu}\beta$-approximate stationary points, i.e. when $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\| < \sqrt{\mu}\beta$.*

To prove Lemma B.1, we prove Lemma 4.4 in the rest of this section. Note that from the continuity of $\left\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta)\right\|$, we see that the Hessian of the implicit regularizer $\mathrm{Tr}D^2\mathcal{L}$ is PSD in a neighborhood of a $\sqrt{\mu}\beta/2$-stationary point $\theta \in \mathcal{M}$. This implies that $\mathrm{Tr}D^2\mathcal{L}$ is g-convex in a neighborhood of such θ. We state a short proof of this in Lemma H.1 in the appendix. Even though under Assumption 3.1 we get that $\mathrm{Tr}D^2\mathcal{L}$ has a convex extension in $\mathbb{R}^{md}$, it is not necessarily g-convex over the manifold. This is because the Hessian over the manifold is calculated based on covariant differentiation, which is different from the normal Euclidean derivative. In particular, for a hypersurface $\mathcal{M}$ in $\mathbb{R}^{md}$, the covaraint derivative is obtained by taking the normal Euclidean derivative from the tangent of a curve on $\mathcal{M}$, then projecting the resulting vector back onto the tangent space. This projecting back operation in the definition of Covariant derivative results in an additional "projection term" in the algebraic formula of the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ on $\mathcal{M}$, in addition to its normal Euclidean Hessian. The inversion in the projection operator makes it fundamentally difficult to derive good estimates on the projection term that are relatable to the Euclidean Hessian of $\mathrm{Tr}D^2\mathcal{L}$. Nonetheless, we discover a novel structure in $\mathcal{M}$ which is that even though we cannot control the projection term in the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ globally, but for ϵ-stationary points θ on the manifold, we can relate it to the approximate stationarity condition (equivalently approximate KKT condition) at θ. This relation enables us to control the projection term and show that the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ on $\mathcal{M}$ is indeed PSD at approximate stationary points. To prove Lemma 4.4, first we calculate the formula for the Hessian of an arbitrary function F on a hypersurface $\mathcal{M}$ embedded in $\mathbb{R}^{md}$, in Lemma B.2. We then use this formula to calculate the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ on $\mathcal{M}$ in Lemma 4.5. Finally we use this formula to prove Lemma 4.4 in Section 4.3.

Now we derive a general formula for the Hessian of any smooth function F on the manifold in Lemma B.2, proved in Appendix C.3.

Lemma B.2 (Hessian formula on a submanifold of $\mathbb{R}^{md}$)**.** *For any smooth function F defined on $\mathbb{R}^{md}$, the Hessian of F on $\mathcal{M}$ is given by:*

$$\nabla^2 F(\theta)[u,w] = D^2F[u,w] - D^2f(\theta)[u,w]^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)(DF(\theta)),$$

for $u, w \in \mathcal{T}_\theta(\mathcal{M})$.

Next, we prove Lemma 4.5, exploiting the formula that we derived in Lemma B.2 for the Hessian of a general smooth function F over $\mathcal{M}$.

B.1.1. Proof of Lemma 4.5

Proof. Using the formula for the Hessian of a function over the manifold in Lemma B.2, we have

$$\begin{aligned}
&\nabla^2 \mathrm{Tr} D^2\mathcal{L}(\theta)[u,w] \\
&= D^2\mathrm{Tr}D^2\mathcal{L}(\theta)[u,w] - D^2 f(\theta)[u,w]^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)(D\mathrm{Tr}D^2\mathcal{L}(\theta)) \\
&= D^2\mathrm{Tr}D^2\mathcal{L}(\theta)[u,w] \\
&- D^2 f(\theta)[u,w]^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)\Big(\nabla_\theta \mathrm{Tr}D^2 L(\theta) + 2\sum_{i=1}^n \alpha_i' Df_i(\theta)\Big) \\
&\overset{(16)}{=} D^2\mathrm{Tr}D^2\mathcal{L}(\theta)[u,w] - 2D^2 f(\theta)[u,w]^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)\Big(\sum_{i=1}^n \alpha_i' Df_i(\theta)\Big) \qquad (16) \\
&= D^2\mathrm{Tr}D^2\mathcal{L}[u,w] - 2D^2 f(\theta)[u,w]^\top (Df(\theta)Df(\theta)^\top)^{-1} (Df(\theta)Df(\theta)^\top)\alpha' \\
&= D^2\mathrm{Tr}D^2\mathcal{L}(\theta)[u,w] - 2\sum_{i=1}^n \alpha_i' D^2 f_i(\theta)[u,w],
\end{aligned}$$

where equality (16) follows from the fact that $\nabla_\theta(\mathrm{Tr}D^2\mathcal{L}) \in \mathcal{T}_\theta(\mathcal{M})$, which means it is orthogonal to $Df_i(\theta)$ for all $i \in [n]$. □

Using Lemma 4.5, we prove Lemma 4.4. We restate Lemma 4.4 here for the convenience of the reader.

Lemma (see Lemma 4.4). Suppose that activation ϕ satisfies Assumptions 3.1, 3.2, 3.4. Consider a point θ on the manifold where the gradient is small, namely $\|\nabla_\theta \mathrm{Tr}D^2\mathcal{L}(\theta)\| \le \sqrt{\mu}\beta$. Then, the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ on the manifold is PSD at point θ, or equivalently $\mathrm{Tr}D^2\mathcal{L}$ is locally g-convex on $\mathcal{M}$ around θ.

B.1.2. Proof of Lemma 4.4

Proof. Suppose $\|\nabla_\theta \mathrm{Tr}D^2\mathcal{L}(\theta)\| = \delta$. Write the projection of the Euclidean gradient $D(\mathrm{Tr}D^2\mathcal{L})(\theta)$ onto the normal space $\mathcal{T}_\theta^N(\mathcal{M})$ in the row space of $Df(\theta)$, i.e. there exists a vector $\alpha' = (\alpha_i')_{i=1}^n$ such that

$$P_\theta^N[D(\mathrm{Tr}D^2\mathcal{L})(\theta)] \triangleq 2\sum_{i=1}^n \alpha_i' Df_i(\theta),$$

where P_θ^N is the projection onto the normal space at point θ on the manifold. Then, if we decompose $D(\mathrm{Tr}D^2\mathcal{L})(\theta)$ into the part in the tangent space and the part in the normal space, because the part on the tangent space is the gradient of $\mathrm{Tr}D^2\mathcal{L}$ on the manifold, we have

$$D(\mathrm{Tr}D^2\mathcal{L})(\theta) = \nabla_\theta \mathrm{Tr}D^2 L(\theta) + 2\sum_{i=1}^n \alpha_i' Df_i(\theta).$$

Now from the assumption that the norm of $\nabla \mathrm{Tr}D^2\mathcal{L}(\theta)$ is at most δ and using the formula for $\mathrm{Tr}D^2\mathcal{L}$ in Lemma 4.3, we have

$$\begin{aligned}
\delta^2 = \Big\|\nabla_\theta \mathrm{Tr}D^2\mathcal{L}(\theta)\Big\|^2 &= \Big\|D(\mathrm{Tr}D^2\mathcal{L})(\theta) - 2\sum_{i=1}^n \alpha_i' \Big[\phi'(\theta_j^\top x_i)x_i\Big]_{j=1}^m\Big\|^2 \\
&= 4\sum_{j=1}^m \Big\|\sum_{i=1}^n (\phi''(\theta_j^\top x_i) - \alpha_i')\phi'(\theta_j^\top x_i)x_i\Big\|^2,
\end{aligned}$$

which implies

$$4\mu\sum_{i=1}^n (\phi''(\theta_j^\top x_i) - \alpha_i')^2 \phi'(\theta_j^\top x_i)^2 \le 4\Big\|\sum_{i=1}^n (\phi''(\theta_j^\top x_i) - \alpha_i')\phi'(\theta_j^\top x_i)x_i\Big\|^2 \le \delta^2.$$

This further implies $\forall i \in [n], j \in [m]$:

$$|(\phi''(\theta_j^\top x_i) - \alpha_i')\phi'(\theta_j^\top x_i)| \leq \delta/(2\sqrt{\mu}). \tag{17}$$

Now using the β-normality of ϕ, the fact that $\delta/\sqrt{\mu} \leq \beta$, and the positivity of ϕ' and ϕ''', we have for all $i \in [n]$ and $j \in [m]$,

$$2\Big|\phi''(\theta_j^\top x_i)(\alpha_i' - \phi''(\theta_j^\top x_i))\Big| \leq \phi'(\theta_j^\top x_i)\phi'''(\theta_j^\top x_i). \tag{18}$$

On the other hand, using Lemma 4.5, we can explicitly calculate the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ as

$$\nabla^2 \mathrm{Tr}D^2\mathcal{L}(\theta)[u,u] = D^2\mathrm{Tr}D^2\mathcal{L}(\theta)[u,u] - 2\sum_{i=1}^n \alpha_i' D^2 f_i(\theta)[u,u].$$

Now using the formula for the Euclidean Hessian of $\mathrm{Tr}D^2\mathcal{L}$ and f_i's in Lemmas F.1 and F.2, we get

$$\begin{aligned}
\nabla^2 \mathrm{Tr}D^2\mathcal{L}(\theta)[u,u] &= \sum_{j=1}^m \sum_{i=1}^n \Big(2\phi'(\theta_j^\top x_i)\phi'''(\theta_j^\top x_i) + 2\phi''(\theta_j^\top x_i)^2\Big)(x_i^\top u_j)^2 \\
&\quad - \sum_{j=1}^m \sum_{i=1}^n 2\alpha_i'\phi''(\theta_j^\top x_i)(x_i^\top u_j)^2 \\
&= \sum_{j=1}^m \sum_{i=1}^n \Big(2\phi'(\theta_j^\top x_i)\phi'''(\theta_j^\top x_i) + 2\phi''(\theta_j^\top x_i)(\phi''(\theta_j^\top x_i) - \alpha_i')\Big)(x_i^\top u_j)^2 \\
&\geq \sum_{j=1}^m \sum_{i=1}^n \phi'(\theta_j^\top x_i)\phi'''(\theta_j^\top x_i)(x_i^\top u_j)^2 \geq 0,
\end{aligned} \tag{19}$$

where the second to last inequality follows from Equation (18). Hence, the Hessian over the manifold is PSD at $\sqrt{\mu}\beta$-stationary points. □

Based on Lemma 4.4, the proof of Lemma B.1 is immediate.

B.1.3. Proof of Lemma B.1

Proof. To prove Lemma B.1 simply note that the derivative of the norm squared of the gradient becomes negative:

$$\frac{d}{dt}\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2 = -\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))^\top \nabla^2 \mathrm{Tr}D^2\mathcal{L}(\theta(t))\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t)) \leq 0.$$

□

Next, we show a strict positive definite property for the hessian of approximate stationary points under our Assumptions 3.1, 3.2, and 3.4.

B.2. Strong local g-convexity

It turns out we can obtain an exponential decay on the norm of the gradient at approximate stationary points, given explicit positive lower bounds on ϕ' and ϕ''' stated in Assumption 3.1. As we mentioned, from Lemma G.1 the gradient flow does not exit a bounded region around θ^* on $\mathcal{M}$, which means Assumption 3.1 automatically holds for some positive constants ϱ_1, ϱ_2 for the flow in Equation (4), given that we have the weaker condition $\phi', \phi''' > 0$ everywhere; this is because of compactnesss of bounded regions in the Euclidean space.

We state the exponential decay in Lemma B.3.

Lemma B.3 (Decay of norm of gradient)**.** *Under assumptions 3.1, 3.1, and 3.2, given time t_0 such that $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t_0))\| \leq \sqrt{\mu}\beta$, then the norm of gradient of $\mathrm{Tr}D^2\mathcal{L}$ is exponentially decreasing for times $t \geq t_0$. Namely,*

$$\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2 \leq \|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t_0))\|^2 e^{-(t-t_0)\varrho_1\varrho_2\mu}.$$

The key idea to obtain this exponential decay in Lemma B.3 is that at approximate stationary points θ, under Assumption 3.1, the Hessian of $\mathrm{Tr}D^2\mathcal{L}$ becomes strictly positive definite on a subspace of $\mathcal{T}_\theta(\mathcal{M})$ which is spanned by $\{x_i\}_{i=1}^n$ in the coordinates corresponding to θ_j. Fortunately, the gradient $\nabla \mathrm{Tr}D^2\mathcal{L}(\theta)$ is in this subspace, hence we can obtain the following Lemma. We show this formally in Lemma B.4, proved in Appendix C.4.

Lemma B.4 (Strong convexity)**.** *Under Assumptions 3.1, 3.1, and 3.2 we have for $\sqrt{\mu}\beta$-stationary θ, i.e. when $\|\nabla_\theta \mathrm{Tr}D^2\mathcal{L}\| \leq \sqrt{\mu}\beta$,*

$$\nabla \mathrm{Tr}D^2\mathcal{L}(\theta)^\top \Big(\nabla^2 \mathrm{Tr}D^2\mathcal{L}(\theta)\Big)\nabla \mathrm{Tr}D^2\mathcal{L}(\theta) \geq \varrho_1\varrho_2\mu\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta)\|^2.$$

B.2.1. Proof of Lemma B.3

Proof. Using Lemma B.4, we get positive definite property of the Hessian $\nabla^2 \mathrm{Tr}D^2\mathcal{L}$ on the manifold by Lemma B.4, which implies

$$\begin{aligned}\frac{d\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2}{dt} &= -\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))^\top \nabla^2 \mathrm{Tr}D^2\mathcal{L}(\theta(t))\nabla g(\theta(t)) \\ &\leq -\varrho_1\varrho_2\mu\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2.\end{aligned}$$

This inequality implies an exponential decay in the size of the gradient, namely for any $t \geq t_0$

$$\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2 \leq \|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t_0))\|^2 e^{-(t-t_0)\varrho_1\varrho_2\mu}.$$

□

Next, we prove Theorem 3.5.

B.3. Proof of Theorem 3.5

Using the non-decreasing property and exponential decay of the gradient at approximate stationary points that we showed in Lemmas B.1 and B.3 respectively, we bound the rate of convergence of the gradient flow (4) stated in Theorem 3.5. The idea is that once the flow enters the approximate stationary state, it gets trapped there by the local g-convexity. Finally, under Assumption 3.1, we show a semi-monotonicity property of $\mathrm{Tr}D^2\mathcal{L}$ in Lemma 4.6, proved in Appendix C.5; namely, having small gradient at a point on the manifold implies that the neurons are close to an optimal point θ^*, as defined in Equation (5), in the subspace spanned by data. This automatically translates our rate of decay of gradient in Lemma B.3 to a convergence rate of $\theta(t)$ to θ^*.
First, note that once the norm of gradient satisfies

$$\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\| \leq \epsilon$$

for any $\epsilon \leq \sqrt{\mu}\beta$, it remains bounded by ϵ due to the non-increasing property of the norm of gradient guaranteed by Lemma B.1. Therefore, if for some time t we have $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\| \leq \epsilon$ then $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t'))\| \leq \epsilon$ for all $t' \geq t$. On the other hand, observe that the implicit regularizer decays with rate of squared norm of its gradient along the flow:

$$\frac{d\mathrm{Tr}D^2\mathcal{L}(\theta(t))}{dt} = -\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\|^2.$$

Therefore, if for some time t we have $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t))\| \geq \epsilon$ we can lower bound the value of $\mathrm{Tr}D^2\mathcal{L}(\theta)$ at time t as

$$\begin{aligned}0 \leq \mathrm{Tr}D^2\mathcal{L}(\theta(t)) &\leq \mathrm{Tr}D^2\mathcal{L}(\theta(0)) - \int_{t'=0}^{t} \|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta(t'))\|^2 dt' \\ &\leq \mathrm{Tr}D^2\mathcal{L}(\theta(0)) - \int_{t'=0}^{t} \epsilon^2 dt' = \mathrm{Tr}D^2\mathcal{L}(\theta(0)) - t\epsilon^2,\end{aligned}$$

which implies

$$t \leq \mathrm{Tr}D^2\mathcal{L}(\theta(0))/\epsilon^2.$$

To further strengthen this into an exponential convergence rate, we know once the norm of the gradient goes below $\sqrt{\mu}\beta$ it will decay exponentially fast due to Lemma B.3. From the first part, it takes at most $\mathrm{Tr}D^2\mathcal{L}(\theta(0))/(\mu\beta^2)$ time for the norm of gradient to go under threshold $\sqrt{\mu}\beta$, then after time $\log((\mu\beta^2/\epsilon'^2)\vee 1)/(\varrho_1\varrho_2\mu)$ it goes below ϵ' using the exponential decay. Hence, the overall time $t_{\epsilon'}$ it takes so that norm of the gradient goes below threshold ϵ' is bounded by

$$t_{\epsilon'} \leq \frac{\mathrm{Tr}D^2\mathcal{L}(\theta(0))}{\mu\beta^2} + \frac{\log((\mu\beta^2/\epsilon'^2)\vee 1)}{\varrho_1\varrho_2\mu}. \tag{20}$$

Now we want to combine this result with Lemma 4.6. For the convenience of the reader, we restate Lemma 4.6.

Lemma (see Lemma 4.6). Suppose $\|\nabla \mathrm{Tr}D^2\mathcal{L}(\theta)\| \leq \delta$. Then, for all $i \in [n], j \in [m]$,

$$\left|\theta_j^\top x_i - \theta_j^{*\top} x_i\right| \leq \delta/(\sqrt{\mu}\varrho_1\varrho_2),$$

for a θ^* as defined in Equation (5).

Combining Equation (20) with Lemma 4.6, we find that to guarantee the property

$$|\theta^\top(t)_j x_i - \theta_j^{*\top} x_i| \leq \epsilon, \forall i \in [n], j \in [m], \tag{21}$$

we need to pick $\epsilon' = \epsilon\sqrt{\mu}\varrho_1\varrho_2$, which implies that we have condition (21) guaranteed for all times

$$t \geq \frac{\mathrm{Tr}D^2\mathcal{L}(\theta(0))}{\mu\beta^2} + \frac{\log(\beta^2/(\varrho_1^2\varrho_2^2\epsilon^2)\vee 1)}{\varrho_1\varrho_2\mu}.$$

C. Remaining proofs for characterizing the stationary points and convergence

C.1. Proof of Lemma 4.2

For any $\theta \in \mathbb{R}^{md}$ we can write

$$D^2\mathcal{L}(\theta) = 2\sum_{i=1}^n Df_i(\theta)Df_i(\theta)^\top + 2\sum_{i=1}^n D^2 f_i(\theta)(f_i(\theta) - y_i).$$

But from $\mathcal{L}(\theta) = y_i$ we get for every $i \in [n]$, $f_i(\theta) = 0$, which implies

$$D^2\mathcal{L} = 2\sum_{i=1}^n Df_i(\theta)Df_i(\theta)^\top. \tag{22}$$

Taking trace from both sides completes the proof.

C.2. Proof of Lemma 4.3

Note that the gradient of the neural network function calculated on x_i with respect to the parameter θ_j of the jth neuron is

$$D_{\theta_j} f_i(\theta) = \phi'(\theta_j^\top x_i)x_i.$$

This implies

$$\|Df_i(\theta)\|^2 = \sum_{j=1}^m \phi'(\theta_j^\top x_i)^2. \tag{23}$$

Summing Equations (23) for all $j \in [m]$ completes the proof.

C.3. Proof of Lemma 4.5

Combining Facts 1 and 2 implies

$$\begin{aligned}\nabla^2 F(\theta)[u,w] &= \Big\langle \nabla_u(\nabla F(\theta)), w\Big\rangle \\ &= \Big\langle P_\theta\Big(D(P_\theta(DF(\theta)))[u]\Big), w\Big\rangle \\ &\overset{(24)}{=} \Big\langle D(P_\theta(DF(\theta)))[u], w\Big\rangle \qquad (24)\\ &= \Big\langle D\Big((I-P_\theta^N)(DF(\theta))\Big)[u], w\Big\rangle, \qquad (25)\end{aligned}$$

where Equality (24) follows because $w \in \mathcal{T}_\theta(\mathcal{M})$, the fact that for any vector v,

$$v = P_\theta(v) + P_\theta^N(v),$$

and that the part $P_\theta^N\Big(D(P_\theta(DF(\theta)))(u)\Big)$ has zero dot product with w. Now note that from Lemma 4.1, the rows of the Jacobian matrix $Df(\theta)$ spans the normal space $\mathcal{T}_\theta^N(\mathcal{M})$ for any $\theta \in \mathcal{M}$. Therefore, the projection matrix onto the normal space $\mathcal{T}_\theta^N(\mathcal{M})$ at point $\theta \in \mathcal{M}$ regarding the operator P_θ^N is given by

$$P_\theta^N(v) = Df(\theta)^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)v. \qquad (26)$$

Plugging Equation (26) into Equation (25), we get:

$$\begin{aligned}\nabla^2 F(\theta)[u,w] &= \Big\langle D\Big((I - Df(\theta)^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta))DF(\theta)\Big)[u], w\Big\rangle \\ &= w^\top D\Big((I - Df^\top (DfDf^\top)^{-1} Df)DF\Big)[u] \\ &= w^\top (I - Df(\theta)^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta))D^2F(\theta)u \\ &+ w^\top D\Big(I - Df(\theta)^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)\Big)[u]DF(\theta), \qquad (27)\end{aligned}$$

where in the last line we used the chain rule. But note that $w \in \mathcal{T}_\theta(\mathcal{M})$, which implies $P_\theta(w) = w$. This means

$$(I - Df(\theta)^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta))w = w.$$

Plugging this back into Equation (27) and noting the fact that $D(I)[u] = 0$,

$$\nabla^2 F(\theta)[u,w] = w^\top D^2F(\theta)u - w^\top D\Big(Df(\theta)^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)\Big)[u]DF(\theta). \qquad (28)$$

Now regarding the second term in Equation (28), note that the directional derivative in direction u can either hit the $Df(\theta)$ terms or the middle part $(Df(\theta)Df(\theta)^\top)^{-1}$, i.e. we get

$$\begin{aligned}w^\top D\Big(Df(\theta)^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)\Big)[u] &= w^\top D^2 f(\theta)[u]^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta) \\ &+ w^\top Df(\theta)^\top D\Big((Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)\Big)[u]. \qquad (29)\end{aligned}$$

Now the key observation here is that because $w \in \mathcal{T}_\theta(\mathcal{M})$, we have

$$Df(\theta)w = 0,$$

which means the second term in Equation (29) is zero. Plugging this back into Equation (28) implies

$$\nabla^2 F(\theta)[u,w] = w^\top D^2F(\theta)u - D^2 f(\theta)[u,w]^\top (Df(\theta)Df(\theta)^\top)^{-1} Df(\theta)DF(\theta),$$

which completes the proof.

C.4. Proof of Lemma B.4

Let $\mathcal{N}$ be the subspace of $\mathcal{T}_\theta(\mathcal{M})$ which can be represented by

$$v = \Big[\sum_{i=1}^{n} \nu_i^j x_i\Big]_{j=1}^{m}, \tag{30}$$

for arbitrary coefficients $(\nu_i^j)_{i=1,\dots,n,j=1,\dots,m}$. First, we show that $\nabla \mathrm{Tr} D^2\mathcal{L}(\theta) \in \mathcal{N}$. Recall the definition of the gradient of $\mathrm{Tr}D^2\mathcal{L}$ on $\mathcal{M}$, i.e.

$$\nabla \mathrm{Tr} D^2\mathcal{L}(\theta) = P_\theta(D\mathrm{Tr}D^2\mathcal{L}(\theta)) = D\mathrm{Tr}D^2\mathcal{L}(\theta) - P_\theta^N(D\mathrm{Tr}D^2\mathcal{L}(\theta)).$$

Note that $P_\theta^N(D\mathrm{Tr}D^2\mathcal{L}(\theta)) \in \mathcal{N}$. This is because from Lemma 4.1 we know that $\{Df_i(\theta)\}_{i=1}^n$ is a basis for $\mathcal{T}_\theta^N(\mathcal{M})$ and each $Df_i(\theta)$ is clearly in the form (30). Now we argue for any vector $v \in \mathcal{N}$, we have

$$\nabla^2 \mathrm{Tr} D^2\mathcal{L}(\theta)[v,v] \geq \mu\varrho_1\varrho_2\|v\|^2. \tag{31}$$

Let $v \triangleq \Big[\sum_{i=1}^n \nu_i^j x_i\Big]_{j=1}^m$. Note that for all $j \in [m]$, from Assumptions 3.1 and 3.2:

$$\begin{aligned}
\nabla^2 \mathrm{Tr} D^2\mathcal{L}(\theta)[v,v] &= \sum_{j=1}^m\sum_{i=1}^n \phi'(\theta_j^\top x_i)\phi'''(\theta_j^\top x_i)\Big((\sum_{i'=1}^n \nu_{i'}^j x_{i'})^\top x_i\Big)^2 \\
&\geq \varrho_1\varrho_2 \sum_{j=1}^m\sum_{i=1}^n \Big((\sum_{i'=1}^n \nu_{i'}^j x_{i'})^\top x_i\Big)^2 \\
&\geq \varrho_1\varrho_2\mu \sum_{j=1}^m \|\sum_{i'=1}^n \nu_{i'}^j x_{i'}\|^2,
\end{aligned}$$

where in the last inequality, we used the fact that from Assumption 3.2, for any vector x in the subspace spanned by $\{x_i\}_{i=1}^n$ we have $x^\top XX^\top x \geq \mu\|x\|^2$. Combining this with Equation (19) in the proof of Lemma 4.4 completes the proof.

Next, we prove Lemma 4.6 which translates the approximate stationary property into topological closeness to a global optimum.

C.5. Proof of Lemma 4.6

Similar to the proof of Lemma 4.4, we have (17):

$$|\phi''(\theta_j^\top x_i) - \alpha_i'|\varrho_1 \leq |(\phi''(\theta_j^\top x_i) - \alpha_i')\phi'(\theta_j^\top x_i)| \leq \delta/(2\sqrt{\mu}),$$

which implies

$$|\phi''(\theta_j^\top x_i) - \alpha_i'| \leq \delta/(2\sqrt{\mu}\varrho_1). \tag{32}$$

Now since $\phi''' > 0$, we have that ϕ'' is strictly monotone and invertible. Define

$$\nu_i' = \phi''^{-1}(\alpha_i'). \tag{33}$$

The fact that $\phi''' \geq \varrho_2$ provides us with a Lipschitz constant of $1/\varrho_2$ for ϕ''^{-1}. Using this with Equations (32) and (33) implies

$$\Big|\theta_j^\top x_i - \nu_i'\Big| \leq \delta/(2\sqrt{\mu}\varrho_1\varrho_2). \tag{34}$$

From Equation (34) we want to show $|\nu_i - \nu_i'| \leq \delta/(2\sqrt{\mu}\varrho_1\varrho_2)$. Suppose this is not true. Then either $\nu_i > \nu_i' + \delta/(2\sqrt{\mu}\varrho_1\varrho_2)$ or $\nu_i < \nu_i' - \delta/(2\sqrt{\mu}\varrho_1\varrho_2)$. In the first case, using Equation (34) we get for all $j \in [m]$:

$$\nu_i > \theta_j^\top x_i,$$

which from the strict monotonicity of ϕ implies

$$y_i/m = \phi(\nu_i) > \phi(\theta_j^\top x_i),$$

which means $f_i(\theta) < y_i$. But this clearly contradicts with the fact that θ is on the manifold of zero loss, i.e. $f_i(\theta) = y_i$. In the other case $\nu_i < \nu_i' - \delta/(2\sqrt{\mu}\varrho_1\varrho_2)$ we can get a similar contradiction. Hence, overall we proved

$$|\nu_i - \nu_i'| \leq \delta/(2\sqrt{\mu}\varrho_1\varrho_2).$$

Combining this with Equation (34) completes the proof.

D. Proof of convergence of the flow to $\mathcal{M}$

In this section, we analyze the behavior of label noise SGD starting from θ_0 with positive loss, when step size goes to zero. To this end, we study a different gradient flow than the one in Equation (4).

D.1. Gradient flow regarding label noise SGD in the limit

In Lemma D.1, we prove that under Assumptions 3.1, 3.2, the gradient flow with respect to the gradient $-D\mathcal{L}(\theta)$ converges to the manifold $\mathcal{M}$ exponentially fast. This is in particular important to in the proof of Theorem 1.1, in particular in proving that SGD with small enough step size will reach zero loss, because SGD in the limit of step size going to zero will converge to this gradient flow outside of $\mathcal{M}$. The key to show this result is a PL inequality that we prove in this setting for $\mathcal{L}$, in Lemma D.2. Here we obtain explicit constants which is not necessary for proving Theorem 1.1.

Lemma D.1 (Convergence of the gradient flow to the manifold)**.** *In the limit of step size going to zero, label noise SGD initialized at a point θ_0 converges to the following gradient flow*

$$\bar{\theta}(0) = \theta_0, \tag{35}$$

$$\bar{\theta}'(t) = -D\mathcal{L}(\bar{\theta}(t)). \tag{36}$$

Under Assumption 3.1 and 3.2, after time

$$t \geq \log(1/\epsilon)$$

we have,

$$\mathcal{L}(\bar{\theta}(t)) \leq e^{-Ct}\mathcal{L}(\bar{\theta}_0),$$

for constant $C = 4m\mu\varrho_1^2$. Furthermore, $\theta(t)$ converges to $\tilde{\theta} \in \mathcal{M}$ such that for all $t > 0$

$$\|\bar{\theta}(t) - \tilde{\theta}\| \leq \frac{2}{\sqrt{C}} e^{-Ct/2}\sqrt{\mathcal{L}(\bar{\theta}_0)}, \tag{37}$$

and

$$\begin{aligned}\mathrm{Tr}D^2\mathcal{L}(\tilde{\theta}) &\leq \sum_{i=1}^{n}\sum_{j=1}^{m}\phi'^2({\theta_0}_j{}^\top x_i) + \sum_{i=1}^{n}\max_{j=1}^{m}\{\phi'^2({\theta_0}_j{}^\top x_i) \pm \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\theta_0)}\} \\ &\leq 2\mathrm{Tr}D^2\mathcal{L}(\theta_0) + \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\theta_0)}.\end{aligned}$$

D.2. Proof of Lemma D.1

First, taking derivative from $\mathcal{L}(\bar{\theta}(t))$ and using the PL inequality from Lemma D.2:

$$\frac{d}{dt}\mathcal{L}(\bar{\theta}(t)) = -\|D\mathcal{L}(\bar{\theta})\|^2 \leq -4m\mu\varrho_1^2\mathcal{L}(\bar{\theta}),$$

which implies

$$\mathcal{L}(\bar{\theta}(t)) \leq e^{-4m\mu\varrho_1^2 t}\mathcal{L}(\theta_0). \tag{38}$$

Moreover, similar to page 38 in (Li et al., 2021), given the PL constant $C = 4m\mu\varrho_1^2$ we have

$$\left\|\frac{d\bar{\theta}(t)}{dt}\right\| = \left\|\nabla\mathcal{L}(\bar{\theta}(t))\right\| \leq \frac{\left\|\nabla\mathcal{L}(\bar{\theta}(t))\right\|^2}{\sqrt{C\mathcal{L}(\bar{\theta}(t))}} = \frac{-\frac{d\mathcal{L}(\bar{\theta}(t))}{dt}}{\sqrt{C\mathcal{L}(\bar{\theta}(t))}} = -\frac{2}{\sqrt{C}}\frac{d\sqrt{\mathcal{L}(\bar{\theta})}}{dt}. \tag{39}$$

This implies

$$\begin{aligned}\|\bar{\theta}(t) - \bar{\theta}(0)\| &\leq \int_0^t \left\|\frac{d\bar{\theta}(t)}{dt}\right\| dt \leq -\frac{2}{\sqrt{C}}\int_0^t \frac{d\sqrt{\mathcal{L}(\bar{\theta}(t))}}{dt} \\ &= \frac{2}{\sqrt{C}}(\sqrt{\mathcal{L}(\bar{\theta}(0))} - \sqrt{\mathcal{L}(\bar{\theta}(t))}) \leq \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\bar{\theta}(0))}.\end{aligned} \tag{40}$$

Additionally, from Equation (39) we get for $t_1 \geq t_2$:

$$\|\bar{\theta}(t_1) - \bar{\theta}(t_2)\| \leq \frac{2}{\sqrt{C}}(\sqrt{\mathcal{L}(\bar{\theta}(t_1))}), \tag{41}$$

which combined with Equation (38) implies that $\theta(t)$ converges to some $\tilde{\theta}$. On the other hand, since $\mathcal{L}(\theta(t))$ converges to $\mathcal{L}(\tilde{\theta})$ by continuity of $\mathcal{L}$ and because $\mathcal{L}(\bar{\theta}(t))$ is going to zero, we conclude that the limit point $\tilde{\theta}$ should have zero loss, i.e. $\tilde{\theta} \in \mathcal{M}$. Now applying Equation (41) for $t_2 = t$ and sending t_1 to infinity implies for all $t > 0$:

$$\|\bar{\theta}(t) - \tilde{\theta}\| \leq \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\bar{\theta}(t))} \leq \frac{2}{\sqrt{C}}e^{-Ct/2}\sqrt{\mathcal{L}(\theta_0)}.$$

On the other hand, recall that from Equation, trace of Hessian on the manifold is equal to

$$\mathrm{Tr}D^2\mathcal{L}(\theta) = \sum_{i=1}^n \left\|\nabla f_i(\bar{\theta})\right\|^2 = \sum_{i=1}^n\sum_{j=1}^m \phi'(\bar{\theta}_j^\top x_i)^2. \tag{42}$$

But under Assumption 3.1 we have that ϕ'^2 is convex, because

$$\frac{d^2}{dz^2}\phi'(z)^2 = 2\phi''(z)^2 + 2\phi'(z)\phi'''(z) \geq 0.$$

But from Equation (40) we have

$$\sum_{j=1}^m |\bar{\theta}_j^\top(t)x_i - \bar{\theta}_j^\top(0)x_i| \leq \|\bar{\theta}(t) - \bar{\theta}(0)\| \leq \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\bar{\theta}(0))}.$$

Combining the convexity and positivity of ϕ'^2 with Equation (40) and the formula for $\left\|\nabla f_i(\theta)\right\|^2$:

$$\begin{aligned}\left|\left\|\nabla f_i(\bar{\theta}(t))\right\|^2 - \left\|\nabla f_i(\bar{\theta}(0))\right\|^2\right| &\leq \max_{j=1}^m\{\phi'^2(\bar{\theta}_j^\top(0)x_i \pm \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\bar{\theta}(0))}) - \phi'^2(\bar{\theta}_j^\top(0)x_i)\} \\ &\leq \max_{j=1}^m\{\phi'^2(\theta_j^\top(0)x_i \pm \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\bar{\theta}(0))})\},\end{aligned} \tag{43}$$

where $\pm$ above in the argument of ϕ' means we take maximum with respect to both $+$ and $-$. Summing Equation (43) for all i, we have for constant

$$C_2 := \sum_{i=1}^n \max_{j=1}^m\{\phi'^2(\bar{\theta}_j^\top(0)x_i \pm \frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\bar{\theta}(0))})\},$$

we have

$$\sum_{i=1}^n \left\|\nabla f_i(\bar{\theta}(t))\right\|^2 \leq \sum_{i=1}^n \left\|\nabla f_i(\theta_0)\right\|^2 + C_2.$$

Sending t to infinity implies

$$\sum_{i=1}^{n} \left\|\nabla f_i(\tilde{\theta})\right\|^2 \leq \sum_{i=1}^{n} \left\|\nabla f_i(\theta_0)\right\|^2 + C_2.$$

But since $\tilde{\theta} \in \mathcal{M}$, from (42) we have

$$\mathrm{Tr} D^2\mathcal{L}(\tilde{\theta}) \leq \sum_{i=1}^{n}\sum_{j=1}^{m} \phi'({\theta_0}_j^\top x_i)^2 + C_2,$$

which completes the proof.

D.3. A PL inequality for $\mathcal{L}$

Next, in Lemma D.2 we show a PL inequality for $\mathcal{L}$ under Assumptions 3.1 and 3.2, which we then use to show Lemma D.1.

Lemma D.2 (PL inequality outside of manifold)**.** *Under Assumptions 3.1 and 3.2, the loss $\mathcal{L}(\theta)$ satisfies a PL inequality of the form*

$$\|D\mathcal{L}(\bar{\theta})\|^2 \geq 4m\mu\varrho_1^2\mathcal{L}(\bar{\theta}).$$

Proof. Note that for all $j \in [m]$, $D_{\theta_j}\mathcal{L}(\bar{\theta})$ is given by

$$D_{\bar{\theta}_j}\mathcal{L}(\bar{\theta}) = 2\sum_{i=1}^{n}(r_{\bar{\theta},NN}(x_i) - y_i)\phi'(\bar{\theta}_j^\top x_i)x_i.$$

Hence

$$\begin{aligned}
\|D_{\bar{\theta}_j}\mathcal{L}(\bar{\theta})\|^2 &= 4\Big\|\sum_{i=1}^{n}(r_{\bar{\theta},NN}(x_i) - y_i)\phi'(\bar{\theta}_j^\top x_i)x_i\Big\|^2 \\
&\geq 4\mu\sum_{i=1}^{n}\phi'(\bar{\theta}_j^\top x_i)^2(r_{\bar{\theta},NN}(x_i) - y_i)^2 \\
&\geq 4\mu\varrho_1^2\mathcal{L}(\bar{\theta}).
\end{aligned} \tag{44}$$

summing Equation (44) for all neurons completes the proof. □

E. Proof of the final guarantee for Label noise SGD

In this section we prove Theorem 1.1.

E.1. Proof of Theorem 1.1

Our main theorem (Theorem 1.1) is a direct combination of Theorem 4.6 of Li et al. (2021), and Theorem 3.3 and Theorem 3.5. First note that from Lemma D.1, the gradient flow in the limit of step size going to zero always converges to the manifold of zero, independent of the initialization. Therefore, the neighborhood U in Theorem 4.6 of (Li et al., 2021) is the whole $\mathbb{R}^{md}$, which implies that for step size $\eta < \eta_0$ less than a certain threshold η_0, $\theta_{\lceil t/\eta^2 \rceil}$ is close enough to $\theta(t)$ in distribution, where $\theta(t)$ is the Riemannian gradient flow defined in Theorem 3.5 which is initialized as $\theta(0) = \tilde{\theta}$. Recall that $\tilde{\theta}$ is the limit point of the first gradient flow (before reaching the manifold $\mathcal{M}$) defined in Lemma D.1.

Moreover, again from Lemma D.1, the gradient flow $\bar{\theta}(t)$ defined in that Lemma remains within distance of $O(\frac{2}{\sqrt{C}}\sqrt{\mathcal{L}(\bar{\theta}_0)})$ of initialization. Therefore, trace of Hessian of the loss at $\tilde{\theta}$ is bounded and only depends on the initialization θ_0. This enables us to have a bounded initial value for the second ODE in (4), depending only on the primary initialization $\bar{\theta}(0)$ (which can be outside the manifold). In particular, note that from Theorems 3.3 and 3.5, this flow converges to ϵ'-proximity of a global minimizer θ^* of trace of Hessian on $\mathcal{M}$ in the sense of (6) after time at most $\tilde{O}(\log(1/\epsilon))$.khasha:fix the dependency here from ϵ to ϵ', also make the dependence on lipschitz constants more explicit.

Now using the unit norm assumption $\|x_i\| = 1$ and picking the step size threshold η_0 small enough, we see that after at most $K = \Theta((1 + \log(1/\epsilon'))/\eta^2)$ iterations of label noise SGD for $\eta \leq \eta_0$, for all $j \in [m]$ and $i \in [n]$ we have with high probability

$$\left|{\theta_T}_j^\top x_i - {\theta^*}_j^\top x_i\right| \leq 2\epsilon'. \tag{45}$$

Therefore, picking $\epsilon' \leq \epsilon/2$, we get for all $i \in [n]$ and $j \in [m]$:

$$\left|{\theta_T}_j^\top x_i - \phi^{-1}(y_i/m)\right| \leq \epsilon.$$

Moreover, note that from Equation (37) and Lemma G.1, θ remains in a ball of bounded radius depending only on θ_0, hence $\mathcal{L}$ and $\mathrm{Tr}D^2\mathcal{L}$ both have a bounded Lipschitz constant only depending on the initialization as well. Therefore, we pick ϵ' small enough so that we get $\mathcal{L}(\theta_K) \leq \epsilon$ and $\left|\mathrm{Tr}D^2\mathcal{L}(\theta_K) - \mathrm{Tr}D^2\mathcal{L}(\theta^*)\right| \leq \epsilon$ for the final iteration K of label noise SGD. This completes the proof.

F. Derivative formulas

F.1. Derivative tensors

In Lemma F.1 we calculate the Jacobian of f and the Hessian of f_i's.

Lemma F.1. *The Jacobian of f is given by:*

$$Df = \begin{bmatrix} \phi'(\theta_1^\top x_1)x_1^\top & \dots & \dots \\ \dots & \phi'(\theta_j^\top x_i)x_i^\top,\ \phi'(\theta_{j+1}^\top x_i)x_i^\top & \dots \\ \dots & \dots & \dots. \end{bmatrix}$$

Moreover, the Hessian of f_i's is given by

$$D^2 f_i(.,.) = \begin{bmatrix} \phi''(x_i^\top\theta_1)x_i x_i^\top & 0 & \dots \\ 0\dots & \dots & \\ 0\dots & \phi''(x_i^\top\theta_j)x_i x_i^\top & 0\dots \\ & \dots & . \end{bmatrix}$$

As a result

$$D^2 f_i(u,u) = \sum_{j=1}^N \phi''(x_i^\top\theta_j)(x_i^\top u_j)^2.$$

In this Lemma F.2, we calculate the Hessian of the trace of Hessian regularizer in our two-layer setting.

Lemma F.2. *For the implicit regularizer $F \triangleq \mathrm{Tr}D^2\mathcal{L}$ regarding our two-layer network with mean squared loss as in (3), we have*

$$F(\theta) = \sum_{i=1}^n \sum_{j=1}^m \phi'^2(\theta_j^\top x_i)\|x_i\|^2 = \sum_i \sum_j \phi'^2(\theta_j^\top x_i),$$

and

$$D^2F(.,.) \tag{46}$$

$$= \sum_i \begin{bmatrix} (2\phi''^2(x_i^\top\theta_1) + \phi'''(x_i^\top\theta_1)\phi'(x_i^\top\theta_1))x_i x_i^\top & 0 & \dots \\ 0\dots & \dots 0\dots & (2\phi''^2(x_i^\top\theta_j) + \phi'''(x_i^\top\theta_j)\phi'(x_i^\top\theta_j))x_i x_i^\top & 0\dots \\ & \dots & . \end{bmatrix} \tag{47}$$

Hence

$$D^2F(\theta)[u,u] = \sum_{i=1}^n \sum_{j=1}^m (2\phi''^2(x_i^\top\theta_j) + 2\phi'''(x_i^\top\theta_j)\phi'(x_i^\top\theta_j))\langle x_i, u_j\rangle^2. \tag{48}$$

G. Other proofs

In Lemma G.1 we show that the gradient flow in Equation (4) remains in a bounded region.

Lemma G.1 (Bounded region)**.** *Under the positivity assumption $\phi''' > 0$, the gradient flow $\frac{d}{dt}\theta(t) = -\nabla \mathrm{Tr} D^2\mathcal{L}(\theta(t))$ remains in a bounded region for all times t.*

Proof. Note that the value of $\mathrm{Tr}D^2\mathcal{L}$ is decreasing along the flow, so for all $t \geq 0$:

$$\mathrm{Tr}D^2\mathcal{L}(\theta(t)) \leq \mathrm{Tr}D^2\mathcal{L}(\theta_0).$$

But noting the formula of $\mathrm{Tr}D^2\mathcal{L}$ in Lemma 4.3, we get that for all $j \in [m]$ and $i \in [n]$:

$$\phi'(\theta_j^\top x_i)^2 \leq \mathrm{Tr}D^2\mathcal{L}(\theta_0). \tag{49}$$

Now let z^* be the minimizer of ϕ'. Now from the assumption of the Lemma we have $\phi'''(z^*) > 0$, which implies from continuity of ϕ''' that for $\epsilon', \delta > 0$ we have for all $z \in (z^* - \epsilon', z^* + \epsilon')$,

$$\phi'''(z) \geq \delta'.$$

This means for all $z \leq z^* - \epsilon'$ we have $\phi''(z) \leq -\epsilon'^2\delta'/2$ and for $z \geq z^* + \epsilon'$ we have $\phi'(z) \geq \epsilon'^2\delta'/2$. Even more, for any $|z - z^*| \geq \epsilon'$ we have

$$\phi'(z) \geq (|z - z^*| - \epsilon')\epsilon'\delta' + \epsilon'^2\delta'/2 = (|z - z^*| - \epsilon'/2)\epsilon'\delta'.$$

Therefore, Equation (49) implies for $z = \theta_j^\top x_i$:

$$(|z - z^*| - \epsilon'/2)\epsilon'\delta' \leq \mathrm{Tr}D^2\mathcal{L}(\theta_0),$$

or

$$|z - z^*| \leq \mathrm{Tr}D^2\mathcal{L}(\theta_0)/(\epsilon'\delta') + \epsilon'/2.$$

This implies the boundedness of the arguments of all the activations for all data points, which completes the proof. □

Lemma G.2 ($\mathcal{M}$ is well-defined)**.** *The intersection of the sub-level sets $f_i(\theta) = y_i$, defined in Section 4 as $\mathcal{M}$, is well-defined as a differentiable manifold.*

Proof. Due to a classical application of the implicit function theorem after a linear change of coordinates using a basis for the kernel of the Jacobian, it turns out that there exists a local chart for $\mathcal{M}$ around each point $\theta \in \mathcal{M}$ for which the Jacobian matrix is non-degenerate. But due to Lemma G.3 the Jacobian is always non-degenerate, so $\mathcal{M}$ is indeed well-defined as a manifold. □

Lemma G.3 (Non-singularity of the Jacobian)**.** *Under assumption 3.1, the Jacobian Df is non-singular, i.e. its rank is equal to the number of its rows.*

Proof. For an arbitrary data point θ_j, we show the non-singularity of the submatrix of Df whose columns corresponds to θ_j, namely $D_{\theta_j} f$. Any linear combination of the rows of this matrix is of the form

$$\sum_{i=1}^{n} \alpha_i \phi'(\theta_j^\top x_i) x_i. \tag{50}$$

But from the coherence assumption in 3.1, we see that $\{x_i\}_{i=1}^n$ are linearly independent, so the linear combination in Equation (50) is also non-zero, which implies the non-singularity of $D_{\theta_i} f$ and completes the proof. □

Lemma G.4 (Cube activation)**.** *If the distribution of the labels does not have a point mass on zero (i.e. if $\mathbb{P}(y = 0) = 0$), then for the cube activation the manifold $\mathcal{M}$ is well-defined, and Theorem 2 also holds.*

Proof. To show that $\mathcal{M}$ is well-defined, similar to Lemma G.2 we show the jacobian $Df(\theta)$ is non-degenerate for all θ. To this end, it is enough to show that for each x_i, there is a j such that $D_{\theta_j} f_i$ is non zero. This is because $D_{\theta_j} f_i$ is always a scaling of x_i, so if it is non-zero, then it is a non-zero scaling of x_i, and since x_i's are linearly independent from Assumption 3.1 the non-degeneracy of Df follows. To show the aformentioned claim, note that for each $\theta \in \mathcal{M}$, from the assumption on the distribution of the label, we have with probability one:

$$\sum_{j=1}^{m} \phi({\theta_j}^\top x_i) = y_i \neq 0.$$

Therefore, there exists $\tilde{j}$ such that

$$\phi({\theta_j}^\top x_i) \neq 0,$$

which implies

$${\theta_j}^\top x_i \neq 0,$$

which means

$$\phi'({\theta_j}^\top x_i) \neq 0.$$

Therefore,

$$D_{\theta_j} f_i(\theta) = \phi'({\theta_j}^\top x_i) x_i \neq 0,$$

and the proof of the non-degeneracy of Df is complete.

Next, we show that proof of Theorem 3.3 is still valid even though Assumption 3.1 does not hold for $\phi(x) = x^3$. This is because the only argument that we use in the proof of Theorem 3.3 which depends on Assumption 3.1 is the invertibility of ϕ and ϕ'' which holds for the cube activation. Hence, the proof of complete.

□

H. G-convexity

In the following Lemma H.1, we show that PSD property of the Hessian of an arbitrary smooth function F on $\mathcal{M}$ at point θ implies that its locally g-convex at θ.

Lemma H.1 (Local g-convexity)**.** *For a function F on manifold $\mathcal{M}$, given that its Hessian is PSD in an open neighborhood $\mathcal{C} \subseteq \mathcal{M}$, then F is g-convex in $\mathcal{C}$ on $\mathcal{M}$.*

Proof. Consider a geodesic $\gamma(t) \subseteq \mathcal{M}$. Then taking derivative of F on γ with respect to t:

$$\frac{d}{dt} F(\gamma(t)) = \langle \nabla F(\gamma(t)), \gamma'(t) \rangle,$$

and for the second derivative

$$\begin{aligned} \frac{d^2}{dt^2} F(\gamma(t)) &= \langle \nabla F(\gamma(t)), \nabla_{\gamma'(t)} \gamma'(t) \rangle + \langle \nabla^2 F(\gamma(t)) \gamma'(t), \gamma'(t) \rangle \\ &= \langle \nabla F(\gamma(t)), \nabla_{\gamma'(t)} \gamma'(t) \rangle, \end{aligned}$$

where we used the fact that $\nabla_{\gamma'(t)} \gamma'(t) = 0$ for a geodesic, and the fact that the Hessian of F is PSD. Therefore, F is convex in the normal sense on any geodesic in $\mathcal{C}$, which means it is g-convex in $\mathcal{C}$. □

I. Background on Riemannian Geometry

In this section, we recall some of basic concepts in differential geometry, related to our analysis in this work. The material here is directly borrowed from (Gatmiry & Vempala, 2022; Gatmiry et al., 2023). For a more detailed treatment on Differential Geometry, we refer the reader to (Do Carmo, 2016).

Abstract manifold. A manifold $\mathcal{M}$ is a topological space such that for an integer $n \in \mathbb{N}$ and given each point $p \in \mathcal{M}$, there exists an open set U around p such that U is a homeomorphism to an open set of $\mathbb{R}^n$. We call a function f defined on $\mathcal{M}$ smooth if for every $p \in \mathcal{M}$ and open set U around p which is homeomorphism to an open set $V \subseteq \mathbb{R}^n$ with mapping ϕ, then $f \circ \phi$ is a smooth function of V. This is how we define vector fields on $\mathcal{M}$ later on.

Note that the manifold we consider in this work is a sub-manifold embedded in the Euclidean space, given by the intersection of sublevel sets of $f_i(\theta)$'s. Given that the Jacobian of $f = (f_i)_{i=1}^n$ is non-singular(Lemma G.3), the manifold $\mathcal{M} = \{x \mid \ \forall i \in [n] : \ f_i(x) = y_i\}$ is well-defined.

Tangent space. For any point $p \in \mathcal{M}$, one can define the notion of tangent space for p, $T_p(\mathcal{M})$, as the set of equivalence class of differentiable curves γ starting from p ($\gamma(0) = p$), where we define two such curves γ_0 and γ_1 to be equivalent if for any function f on the manifold:

$$\frac{d}{dt} f(\gamma_0(t))\big|_{t=0} = \frac{d}{dt} f(\gamma_1(t))\big|_{t=0}.$$

When the manifold is embedded in $\mathbb{R}^d$, the tangent space at p can be identified by the subspace of $\mathbb{R}^d$ obtained from the tangent vectors to curves on $\mathcal{M}$ passing through p.

Vector field. A vector field V is a smooth choice of a vector $V(p) \in T_p(\mathcal{M})$ in the tangent space for all $p \in \mathcal{M}$.

Metric and inner product. A metric $\langle ., . \rangle$ is a tensor on the manifold $\mathcal{M}$ that is simply a smooth choice of a symmetric bilinear map on the tangent space of each point $p \in \mathcal{M}$. Namely, for all $w, v, z \in T_p(\mathcal{M})$:

$$\begin{aligned} \langle v + w, z \rangle &= \langle v, z \rangle + \langle w, z \rangle, \\ \langle \alpha v, \beta w \rangle &= \alpha\beta \langle v, w \rangle. \end{aligned}$$

Covariant derivative. Given two vector fields V and W, the covariant derivative, also called the Levi-Civita connection $\nabla_V W$ is a bilinear operator with the following properties:

$$\begin{aligned} \nabla_{\alpha_1 V_1 + \alpha_2 V_2} W &= \alpha_1 \nabla_{V_1} W + \alpha_2 \nabla_{V_2} W, \\ \nabla_V (W_1 + W_2) &= \nabla_V (W_1) + \nabla_V (W_2), \\ \nabla_V (\alpha W_1) &= \alpha \nabla_V (W_1) + V(\alpha) W_1 \end{aligned}$$

where $V(\alpha)$ is the action of vector field V on scalar function α. Importantly, the property that differentiates the covariant derivative from other kinds of derivaties over the manifold is that the covariant derivative of the metric is zero, i.e., $\nabla_V g = 0$ for any vector field V. In other words, we have the following intuitive rule:

$$\nabla_V \langle W_1, W_2 \rangle = \langle \nabla_V W_1, W_2 \rangle + \langle W_1, \nabla_V W_2 \rangle.$$

Moreover, the covariant derivative has the property of being torsion free, meaning that for vector fields W_1, W_2:

$$\nabla_{W_1} W_2 - \nabla_{W_2} W_1 = [W_1, W_2],$$

where $[W_1, W_2]$ is the Lie bracket of W_1, W_2 defined as the unique vector field that satisfies

$$[W_1, W_2] f = W_1(W_2(f)) - W_2(W_1(f))$$

for every smooth function f.

In a local chart with variable x, if one represent $V = \sum V^i \partial x_i$, where ∂x_i are the basis vector fields, and $W = \sum W^i \partial x_i$,

the covariant derivative is given by

$$\begin{aligned}\nabla_V W &= \sum_i V^i \nabla_i W = \sum_i V^i \sum_j \nabla_i (W^j \partial x_j) \\ &= \sum_i V^i \sum_j \partial_i (W^j) \partial x_j + \sum_i V^i \sum_j W^j \nabla_i \partial x_j \\ &= \sum_j V(W^j) \partial x_j + \sum_i \sum_j V^i W^j \sum_k \Gamma_{ij}^k \partial x_k = \\ &= \sum_k \big(V(W^k) + \sum_i \sum_j V^i W^j \Gamma_{ij}^k \big) \partial x_k.\end{aligned}$$

The Christoffel symbols Γ_{ij}^k are the representations of the Levi-Cevita derivatives of the basis $\{\partial x_i\}$:

$$\nabla_{\partial x_j} \partial x_i = \sum_k \Gamma_{ij}^k \partial x_k$$

and are given by the following formula:

$$\Gamma_{ij}^k = \frac{1}{2} \sum_m g^{km} (\partial_j g_{mi} + \partial_i g_{mj} - \partial_m g_{ij}).$$

Above, g^{ij} refers to the (i, j) entry of the inverse of the metric.

Gradient. For a function F over the manifold, it naturally acts linearly over the space of vector fields: given a vector field V, the mapping $V(F)$ is linear in V. Therefore, by Riesz representation theorem there is a vector field W where

$$V(F) = \langle W, V \rangle$$

for any vector field V. This vector field W is defined as the gradient of F, which we denote here by ∇F. To explicitly derive the form of gradient in a local chart, note that

$$V(F) = (\sum_i \partial x_i V_i)(F) = \sum_i V_i \partial_i F = \sum_{j,m,i} V_j g_{jm} g^{mi} \partial_i F = \langle V, \big(\sum_i g^{mi} \partial_i F \big)_m \rangle.$$

The above calculation implies

$$\nabla F = \sum_m (\sum g^{mi} \partial x_i F) \partial x_m.$$

So $g^{-1} DF$ is the representation of ∇F in the Euclidean chart with basis $\{\partial x_i\}$.

Given a smooth function F on $\mathbb{R}^{md}$ and sub-manifold $\mathcal{M} \subseteq \mathbb{R}^{md}$, the gradient ∇F on $\mathcal{M}$ at any point $\theta \in \mathcal{M}$ is given by the Euclidean projection of the normal Euclidean gradient of F at θ onto the tangent space $\mathcal{T}_\theta(\mathcal{M})$.