Appendix

The appendix consists of two parts. Section A contains the details for our proof. Section B provides more detailed descriptions of our experiments and attaches additional experiments.

A Omitted Proofs

We provide a sketch of omitted proofs in this part. For future convenience, we state the expression of the following quantities for f:

$$f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x}) = \sum_{r=1}^{m} a_r \sigma(\boldsymbol{w}_r^T(\boldsymbol{x} + \boldsymbol{\delta})).$$

$$\frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}} = \sum_{r=1}^{m} a_r \sigma'(\boldsymbol{w}_r^T(\boldsymbol{x} + \boldsymbol{\delta}))\boldsymbol{w}_r,$$

where $\sigma'(x) = e^x/(1+e^x)$.

$$\frac{\partial^2 f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}^2} = \sum_{r=1}^{m} a_r \sigma''(\boldsymbol{w}_r^T(\boldsymbol{x} + \boldsymbol{\delta}))\boldsymbol{w}_r \boldsymbol{w}_r^T,$$

where $\sigma''(x) = e^x/(1+e^x)^2$.

A.1 Proof of Lemma 4.1

Lemma A.1. *There exists a threshold $m_{\min} = \Omega(d^{5/2})$, so that for each $m > m_{\min}$, there exists $\varepsilon_{\max}(m) = \Theta((\log m)^{-1/2})$, if $\varepsilon < \varepsilon_{\max}$, then with high probability,*

$$B_l \leqslant \Big\| \frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}} \Big\| \leqslant B_u \sqrt{\log d},$$

for all $\boldsymbol{\delta} \in \mathbb{B}(\boldsymbol{0}, \varepsilon)$, where B_l, B_u is of order $\Theta(1)$.

Proof. We denote $a_r \sigma'(\boldsymbol{w}_r^T(\boldsymbol{x} + \boldsymbol{\delta}))\boldsymbol{w}_r$ as $\boldsymbol{\xi}_r(\boldsymbol{\delta})$. For a threshold $T > 0$,

$$\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x}) = \sum_{r=1}^{m} \boldsymbol{\xi}_r(\boldsymbol{\delta}) I\{\|\boldsymbol{w}_r\| \leqslant T\} + \sum_{r=1}^{m} \boldsymbol{\xi}_r(\boldsymbol{\delta}) I\{\|\boldsymbol{w}_r\| > T\}.$$

For $t > 0$,

$$\Big\{ \max_{\boldsymbol{\delta} \in \mathbb{B}(\boldsymbol{0}, \varepsilon)} \|\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})\| > t \Big\} \subseteq \Big\{ \max_{\boldsymbol{\delta} \in \mathbb{B}(\boldsymbol{0}, \varepsilon)} \Big\| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \boldsymbol{\xi}_r(\boldsymbol{\delta}) I\{\|\boldsymbol{w}_r\| \leqslant T\} \Big\| > t/2 \ \text{ or } \max_r \|\boldsymbol{w}_r\| > T \Big\}$$

Let us take $T = c_1\sqrt{\log m}$, as long as c_1 is large enough, we know that we can control $\max \mathbb{P}(\max_r \|\boldsymbol{w}_r\| > T)$ to be small. Thus, in order to bound

$$\mathbb{P}\big(\max_{\boldsymbol{\delta} \in \mathbb{B}(0, \varepsilon)} \|\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})\| > t \big),$$

we only need to bound

$$\mathbb{P}\Big(\max_{\boldsymbol{\delta} \in \mathbb{B}(\boldsymbol{0}, 1), \|\boldsymbol{s}\| = 1} \Big| \sum_{r=1}^{m} \boldsymbol{s}^T [\boldsymbol{\xi}_r(\varepsilon\boldsymbol{\delta}) - \boldsymbol{\xi}_r(\boldsymbol{0})] I\{\|\boldsymbol{w}_r\| \leqslant T\} \Big| > t/2 \Big).$$

We prove $\sqrt{d}\Upsilon(\boldsymbol{\delta},\boldsymbol{s})$ is a ψ_2-process, where

$$\Upsilon(\boldsymbol{\delta},\boldsymbol{s}) := \Big|\sum_{r=1}^{m} a_r[\sigma'(\boldsymbol{w}_r^T(\boldsymbol{x}+\varepsilon\boldsymbol{\delta}))-\sigma'(\boldsymbol{w}_r^T(\boldsymbol{x}))]\boldsymbol{w}_r^T s I\{\|\boldsymbol{w}_r\|\leqslant T\}\Big|.$$

We only need to prove

$$\mathbb{E}\exp(d\frac{|\Upsilon(\boldsymbol{\delta},\boldsymbol{s})-\Upsilon(\boldsymbol{\delta}',\boldsymbol{s}')|^2}{\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|^2+\|\boldsymbol{s}-\boldsymbol{s}'\|^2})\leqslant 2.$$

Since

$$\begin{aligned}\mathbb{E}\exp(d\frac{|\Upsilon(\boldsymbol{\delta},\boldsymbol{s})-\Upsilon(\boldsymbol{\delta}',\boldsymbol{s}')|^2}{\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|^2+\|\boldsymbol{s}-\boldsymbol{s}'\|^2}) &= \int_0^\infty e^t\mathbb{P}\Big(d\frac{|\Upsilon(\boldsymbol{\delta},\boldsymbol{s})-\Upsilon(\boldsymbol{\delta}',\boldsymbol{s}')|^2}{\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|^2+\|\boldsymbol{s}-\boldsymbol{s}'\|^2}>t\Big)dt\\ &\leqslant \int_0^\infty e^t\mathbb{P}\Big(|\sum_{r=1}^m\sqrt{d}(u_r+v_r)|>\sqrt{t}\Big)dt\\ &\leqslant \int_0^\infty e^t\mathbb{P}\Big(|\sum_{r=1}^m\sqrt{d}u_r|>\sqrt{t}/2\Big)dt+\int_0^\infty e^t\mathbb{P}\Big(|\sum_{r=1}^m\sqrt{d}v_r|>\sqrt{t}/2\Big)dt,\end{aligned}$$

where

$$\begin{aligned}u_r :&= a_r[\sigma'(\boldsymbol{w}_r^T(\boldsymbol{x}+\varepsilon\boldsymbol{\delta}))-\sigma'(\boldsymbol{w}_r^T(\boldsymbol{x}+\varepsilon\boldsymbol{\delta}'))]/\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|(\boldsymbol{w}_r^T\boldsymbol{s})I\{\|\boldsymbol{w}_r\|\leqslant T\}\\ &= a_r l(\boldsymbol{\delta},\boldsymbol{\delta}',\boldsymbol{w}_r,\boldsymbol{x})\varepsilon\boldsymbol{w}_r^T(\boldsymbol{\delta}-\boldsymbol{\delta}')/\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|\boldsymbol{w}_r^T s I\{\|\boldsymbol{w}_r\|\leqslant T\}\end{aligned}$$

and $|l|$ is bounded by 1.

$$\begin{aligned}v_r :&= a_r[\sigma'(\boldsymbol{w}_r^T(\boldsymbol{x}+\varepsilon\boldsymbol{\delta}'))-\sigma'(\boldsymbol{w}_r^T(\boldsymbol{x}))]\boldsymbol{w}_r^T(\boldsymbol{s}-\boldsymbol{s}')/\|\boldsymbol{s}-\boldsymbol{s}'\|I\{\|\boldsymbol{w}_r\|\leqslant T\}\\ &= a_r l(\boldsymbol{\delta}',\boldsymbol{0},\boldsymbol{w}_r,\boldsymbol{x})\varepsilon\boldsymbol{w}_r^T\boldsymbol{\delta}'\boldsymbol{w}_r^T(\boldsymbol{s}-\boldsymbol{s}')/\|\boldsymbol{s}-\boldsymbol{s}'\|I\{\|\boldsymbol{w}_r\|\leqslant T\}\end{aligned}$$

Since we take $T=c_1\sqrt{\ln m}$, for $\varepsilon\leqslant\lambda(\log m)^{-1/2}$, as long as λ is small enough, it is easy to see u_r and v_r are sub-gaussian and can ensure

$$\int_0^\infty e^t\mathbb{P}\Big(|\sum_{r=1}^m\sqrt{d}u_r|>\sqrt{t}/2\Big)dt+\int_0^\infty e^t\mathbb{P}\Big(|\sum_{r=1}^m\sqrt{d}v_r|>\sqrt{t}/2\Big)dt\leqslant 2.$$

Thus, by chaining, we know,

$$\max_{\boldsymbol{\delta},\boldsymbol{\delta}'\in\mathbb{B}(\boldsymbol{0},\varepsilon)}\Big\|\frac{\partial f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial\boldsymbol{\delta}}-\frac{\partial f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}'+\boldsymbol{x})}{\partial\boldsymbol{\delta}}\Big\|\leqslant B, \tag{7}$$

and as $\lambda\to 0$, $B\to 0$ with high probability.

Now, we prove that

$$\mathbb{P}(\|\frac{1}{\sqrt{m}}\sum_{r=1}^m\boldsymbol{\xi}_r(\boldsymbol{0})I\{\|\boldsymbol{w}_r\|\leqslant T\}\|\geqslant c_3\sqrt{\log d})$$

with small probability for some constant $c_3>0$.

$$\mathbb{P}(\|\sum_{r=1}^m\boldsymbol{\xi}_r(\boldsymbol{0})I\{\|\boldsymbol{w}_r\|\leqslant T\}\|\geqslant t)\leqslant d\mathbb{P}(|\sum_{r=1}^m\boldsymbol{\xi}_r^i(\boldsymbol{0})I\{\|\boldsymbol{w}_r\|\leqslant T\}|\geqslant\frac{t}{\sqrt{d}})$$

where $\boldsymbol{\xi}_r^i$ is the i-th coordinate of $\boldsymbol{\xi}_r$. By concentration of sub-gaussian, we know when $t=O(\sqrt{\log d})$, the probability is small.

At last, let us provide the lower bound, in which we use central limit theorem. We denote $\boldsymbol{\xi}_r(\boldsymbol{0})$ as $\boldsymbol{\varsigma}_r$. The covariance matrix of $\boldsymbol{\varsigma}_r$

$$\Sigma = E\boldsymbol{\varsigma}_r\boldsymbol{\varsigma}_r^T.$$

It can be derived directly by using the multi-variate Berry Esseen bound: for any convex set $\mathcal{C}$,

$$|\mathbb{P}(\Sigma^{-1/2}\frac{\partial f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}} \in \mathcal{C}) - \mathcal{N}(\mathbf{0}, I_d)\{\mathcal{C}\}| \leqslant O(d^{1/4})\sum_{r=1}^{m}\mathbb{E}\Big\|\frac{\Sigma^{-1/2}}{\sqrt{m}}\varsigma_r\Big\|^3.$$

Let us take $\mathcal{C}$ as $\mathbb{B}(\mathbf{0}, \sqrt{cd})$ for $0 < c < 1$. Then, by Bernstein inequality, we can obtain

$$\begin{aligned}\mathcal{N}(\mathbf{0}, I_d)\{\mathbb{B}(\mathbf{0}, \sqrt{(1-c)d})\} &\leqslant \mathbb{P}(d - \chi^2(d) \geqslant cd)\\ &\leqslant 2e^{\frac{-dc^2}{8}}.\end{aligned}$$

Thus, plugging in the expression for the gradient

$$\mathbb{P}(\|\Sigma^{-1/2}\frac{\partial f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}}\| \in [\sqrt{(1-c)d}, \sqrt{(1+c)d}]) \geqslant 1 - 4e^{\frac{-dc^2}{8}} - O(d^{1/4})\sum_{r=1}^{m}\mathbb{E}\Big\|\frac{\Sigma^{-1/2}}{\sqrt{m}}\varsigma_r\Big\|^3.$$

□

Lemma A.2. *Under the assumptions of Lemma A.1, then with high probability,*

$$\Big\|\frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\| \leqslant M_u \log m,\ \Big\|\frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}^2} - \frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}'+\boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\| \leqslant K_u(\log m)^{3/2}\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|,$$

for all $\boldsymbol{\delta} \in \mathbb{B}(\mathbf{0}, \varepsilon)$, where M_u, K_u are of order $\Theta(1)$.

Proof. The proof is almost the same as Lemma A.1. We will not reiterate it here. □

[Proof of Lemma 4.1] Under the result of Lemma A.1, Lemma A.2 and $\mathcal{L} < |y - f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})| < \mathcal{U}$, notice that when $\boldsymbol{\delta}$ is in the interior of the ball,

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \eta\partial L(\boldsymbol{\delta}_t),$$

$$L(\boldsymbol{\delta}_{t+1}) = L(\boldsymbol{\delta}_t) - \eta\Big\|\frac{\partial L(\boldsymbol{\delta}_t)}{\partial \boldsymbol{\delta}}\Big\|_2^2 + \frac{1}{2}\eta^2\Big(\frac{\partial L(\boldsymbol{\delta}_t)}{\partial \boldsymbol{\delta}}\Big)^T\frac{\partial^2 L(\tilde{\boldsymbol{\delta}}_t)}{\partial \boldsymbol{\delta}^2}\frac{\partial L(\boldsymbol{\delta}_t)}{\partial \boldsymbol{\delta}},$$

for $\tilde{\boldsymbol{\delta}}_t \in \mathbb{B}(\mathbf{0}, \varepsilon)$.

Since

$$\frac{\partial^2 L(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^2} = 2(u-y)\frac{\partial^2 f}{\partial \boldsymbol{\delta}^2} + 2\frac{\partial f}{\partial \boldsymbol{\delta}}\Big(\frac{\partial f}{\partial \boldsymbol{\delta}}\Big)^T,$$

the dominating term will be

$$\eta\Big\|\frac{\partial L(\boldsymbol{\delta}_t)}{\partial \boldsymbol{\delta}}\Big\|_2^2,$$

as long as

$$\frac{1}{\eta} \geqslant M_u\mathcal{U}\log m + 2B_u^2\log d \geqslant \max_{\boldsymbol{\delta}\in\mathbb{B}(\mathbf{0},\varepsilon)}\Big\|\frac{\partial^2 L(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\|.$$

Since $m = \Omega(d^{5/2})$, $\eta_{\max}(m) = \Theta\big((\log m)^{-1}\big)$, we can obtain the final result easily.

A.2 Proof of Lemma 4.2

We first provide lemmas for the proof of main theorems of the behavior of projected gradient method on the sphere. The techniques are mainly adopted from [10], we include them here for completeness.

Lemma A.3. *For any $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_0$ on the sphere with radius ε, denoted as $\varepsilon\mathbb{S}^{d-1}$, let $\mathcal{T}_0 = \mathcal{T}(\boldsymbol{\delta}_0)$, then*

$$\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta} - \boldsymbol{\delta}_0)\| \leqslant \frac{\|\boldsymbol{\delta} - \boldsymbol{\delta}_0\|^2}{2\varepsilon}.$$

Furthermore, if $\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|<\varepsilon$, we will also have

$$\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\| \leqslant \frac{\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\|^2}{\varepsilon}$$

Proof. Recall that $c(\boldsymbol{\delta})=\|\boldsymbol{\delta}\|^2-\varepsilon^2$, $\nabla c(\boldsymbol{\delta})=2\boldsymbol{\delta}\in\mathcal{T}^c(\boldsymbol{\delta})$, so we can obtain

$$|\nabla c(\boldsymbol{\delta}_0)^T(\boldsymbol{\delta}-\boldsymbol{\delta}_0)|^2=|2\boldsymbol{\delta}_0^T\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)|^2=4\varepsilon^2\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\|^2. \tag{8}$$

On the other hand, since $c(\boldsymbol{\delta})=c(\boldsymbol{\delta})=0$, besides, for all $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_0$

$$|c(\boldsymbol{\delta})-c(\boldsymbol{\delta}_0)-\nabla c(\boldsymbol{\delta}_0)^T(\boldsymbol{\delta}-\boldsymbol{\delta}_0)|=\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|^2,$$

thus, it results to

$$|\nabla c(\boldsymbol{\delta}_0)^T(\boldsymbol{\delta}-\boldsymbol{\delta}_0)|^2=\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|^4. \tag{9}$$

Combine Eq. (8) with (9), it gives

$$\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\|^2 \leqslant \frac{\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|^4}{4\varepsilon^2}. \tag{10}$$

Notice $\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|^2=\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\|^2+\|\mathcal{P}_{\mathcal{T}_0}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\|^2$, plugging into Eq. (9), we can obtain

$$\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\| \leqslant \frac{\|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\|^2}{\varepsilon}, \text{ or } \|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\| \geqslant \varepsilon \text{ (abandoned).}$$

□

Lemma A.4. *For all $\hat{v}\in\mathcal{T}(\boldsymbol{\delta})$ and $\hat{w}\in\mathcal{T}^c(\boldsymbol{\delta})$, so that $\|\hat{v}\|=\|\hat{w}\|=1$, we have*

$$\max\left\{\|\mathcal{P}_{\mathcal{T}_0^c}\cdot\hat{v}\|,\ \|\mathcal{P}_{\mathcal{T}_0}\cdot\hat{w}\|\right\} \leqslant \frac{\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|}{\varepsilon}.$$

Proof. By Lemma A.3, we know

$$|\mathcal{P}_{\mathcal{T}_0^c}\cdot\hat{v}\|=\frac{|2\boldsymbol{\delta}_0^T\hat{v}|}{2\varepsilon}.$$

Besides, since $\hat{v}\in\mathcal{T}(\boldsymbol{\delta})$, $2\boldsymbol{\delta}^T\hat{v}=0$, thus,

$$|2\boldsymbol{\delta}_0^T\hat{v}|=|2(\boldsymbol{\delta}_0-\boldsymbol{\delta})^T\hat{v}| \leqslant 2\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|,$$

which gives

$$\|\mathcal{P}_{\mathcal{T}_0^c}\cdot\hat{v}\| \leqslant \frac{\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|}{\varepsilon}.$$

Meanwhile, since $\hat{w}\in\mathcal{T}^c(\boldsymbol{\delta})$, $\boldsymbol{\delta}^T\hat{w}/\varepsilon=\hat{w}$,

$$\|\mathcal{P}_{\mathcal{T}_0}\cdot\hat{w}\|=\|\mathcal{P}_{\mathcal{T}_0^c}\cdot\hat{w}-\hat{w}\|=\|\mathcal{P}_{\mathcal{T}_0^c}\cdot\hat{w}-\mathcal{P}_{\mathcal{T}^c}\cdot\hat{w}\| \leqslant \|\hat{w}\|\cdot\frac{\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|}{\varepsilon}=\frac{\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|}{\varepsilon}.$$

□

[Proof of Lemma 4.2] For any $\hat{v}\in\varepsilon\mathbb{S}^{d-1}$, let $\tilde{\boldsymbol{\delta}}_1=\boldsymbol{\delta}_0+\eta\hat{v}$ and $\tilde{\boldsymbol{\delta}}_2=\boldsymbol{\delta}_0+\eta\mathcal{P}_{\mathcal{T}_0}\cdot\hat{v}$

$$\|\Pi_{\mathbb{B}(\mathbf{0},\varepsilon)}(\tilde{\boldsymbol{\delta}}_1)-\boldsymbol{\delta}_2\| \leqslant \frac{4\eta^2}{\varepsilon}.$$

Let $\boldsymbol{z}_1=\Pi_{\mathbb{B}(\mathbf{0},\varepsilon)}(\tilde{\boldsymbol{\delta}}_1)$, we know that $\|\boldsymbol{z}_1-\tilde{\boldsymbol{\delta}}_1\| \leqslant \eta$, $(\tilde{\boldsymbol{\delta}}_1-\boldsymbol{z}_1)\in\mathcal{T}^c(\boldsymbol{z}_1)$ and $\|\boldsymbol{\delta}_0-\boldsymbol{z}_1\| \leqslant 2\eta$. Thus, by Lemma A.4

$$\|\mathcal{P}_{\mathcal{T}_0}(\tilde{\boldsymbol{\delta}}_1-\boldsymbol{z}_1)\|=\frac{\|\mathcal{P}_{\mathcal{T}_0}(\tilde{\boldsymbol{\delta}}_1-\boldsymbol{z}_1)\|}{\|\tilde{\boldsymbol{\delta}}_1-\boldsymbol{z}_1\|}\cdot\|\tilde{\boldsymbol{\delta}}_1-\boldsymbol{z}_1\| \leqslant \frac{\|\boldsymbol{\delta}_0-\boldsymbol{z}_1\|\cdot\|\tilde{\boldsymbol{\delta}}_1-\boldsymbol{z}_1\|}{\varepsilon} \leqslant \frac{2\eta^2}{\varepsilon}.$$

Let $\boldsymbol{v}_1 = \boldsymbol{\delta}_0 + \mathcal{P}_{\mathcal{T}_0}(\boldsymbol{z}_1 - \boldsymbol{\delta}_0)$, then

$$\|\boldsymbol{v}_1 - \tilde{\boldsymbol{\delta}}_2\| = \|\boldsymbol{v}_1 - \boldsymbol{\delta}_0 - \tilde{\boldsymbol{\delta}}_2 + \boldsymbol{\delta}_0\| = \|\mathcal{P}_{\mathcal{T}_0}(\boldsymbol{z}_1 - \boldsymbol{\delta}_0) - \mathcal{P}_{\mathcal{T}_0}(\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_0)\| \leqslant \frac{2\eta^2}{\varepsilon}.$$

Meanwhile, by Lemma A.3,

$$\|\boldsymbol{z}_1 - \boldsymbol{v}_1\| = \|\mathcal{P}_{\mathcal{T}_0^c}(\boldsymbol{z}_1 - \boldsymbol{\delta}_0)\| \leqslant \frac{\|\boldsymbol{z}_1 - \boldsymbol{\delta}_0\|}{2\varepsilon} \leqslant \frac{2\eta^2}{\varepsilon}.$$

Then, we can obtain

$$\|\boldsymbol{z}_1 - \tilde{\boldsymbol{\delta}}_2\| \leqslant \|\boldsymbol{v}_1 - \tilde{\boldsymbol{\delta}}_2\| + \|\boldsymbol{z}_1 - \boldsymbol{v}_1\| \leqslant \frac{4\eta^2}{\varepsilon}.$$

A.3 Proof of Lemma 4.3 and 4.4

Since

$$\Gamma(\boldsymbol{\delta}) = 2(u-y)\Big[\frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}} - \varepsilon^{-2}\boldsymbol{\delta}^T \frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}}\boldsymbol{\delta}\Big].$$

$$\Xi(\boldsymbol{\delta}) = 2(u-y)\Big[\frac{\partial^2 f}{\partial \boldsymbol{\delta}^2} - \varepsilon^{-2}\boldsymbol{\delta}^T \frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}} I\Big] + 2\frac{\partial f}{\partial \boldsymbol{\delta}}\Big(\frac{\partial f}{\partial \boldsymbol{\delta}}\Big)^T,$$

Recall with high probability,

$$\Big\|\frac{\partial^2 f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\| \leqslant M_u \log m, \ \Big\|\frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}}\Big\| \leqslant B_u\sqrt{\log d},$$

as long as $\angle[\partial f(\boldsymbol{\delta}), \boldsymbol{\delta}] \geqslant \phi$ for some constant ϕ and ε is small enough, such that

$$\Big|\varepsilon^{-2}\boldsymbol{\delta}^T \frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}}\Big| \geqslant \max\Big\{\Big\|\frac{\partial^2 f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\|, \frac{1}{\mathcal{L}}\Big\|\frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}}\Big\|^2\Big\}$$

the dominating term is

$$2(u-y)\varepsilon^{-2}\boldsymbol{\delta}^T \frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}} I.$$

Besides, for any constant $\angle[\partial f(\boldsymbol{\delta}), \boldsymbol{\delta}] \leqslant \beta$, notice that $\boldsymbol{\delta}/\varepsilon \in \mathbb{S}^{d-1}$

$$\|\Gamma(\boldsymbol{\delta})\| \geqslant \sqrt{1-\beta^2}\|\partial f(\boldsymbol{\delta})\|.$$

Combined with the lower bound obtained in Lemma A.1 $\|f(\boldsymbol{\delta})\| \geqslant \mathcal{L}B_l$, we can obtain the result.

A.4 Proof of Lemma 4.5

By Lemma 4.3, we know if $\angle[\partial f(\boldsymbol{\delta}), \boldsymbol{\delta}] \leqslant \beta$,

$$\|\Gamma(\boldsymbol{\delta})\| \geqslant \sqrt{1-\beta^2}\|\partial f(\boldsymbol{\delta})\| \geqslant \mathcal{L}B_l\sqrt{1-\beta^2}.$$

Thus, if we choose $\beta = \sqrt{1 - \eta/(\mathcal{L}B_l)^2}$,

$$\|\Gamma(\boldsymbol{\delta})\| \geqslant \sqrt{\eta}.$$

Notice the corresponding solid angle with respect to Δ_η^- and Δ_η^+ will be less or equal to $\pi/2$ if we have

$$\arccos\big(\min_{\boldsymbol{\delta}, \boldsymbol{\delta}'} \angle[\partial f(\boldsymbol{\delta}), \partial f(\boldsymbol{\delta}')]\big) + \arccos\Big(\sqrt{1 - \frac{\eta}{(\mathcal{L}B_l)^2}}\Big) \leqslant \frac{\pi}{4}.$$

That is due to the following fact: $\boldsymbol{\delta}^\star \in \Delta_\eta^+ = \{\boldsymbol{\delta} : \angle[\partial f(\boldsymbol{\delta}), \boldsymbol{\delta}] \geqslant 1 - \sqrt{\eta}/(\mathcal{L}B_l), \boldsymbol{\delta} \in \varepsilon\mathbb{S}^{d-1}\}$, where $\boldsymbol{\delta}^\star \parallel \partial f(\boldsymbol{\delta}^\star)$. Then, for any $\boldsymbol{\delta} \in \varepsilon\mathbb{S}^{d-1}$,

$$\arccos(\angle[\boldsymbol{\delta}, \boldsymbol{\delta}^\star]) \leqslant \arccos(\angle[\partial f(\boldsymbol{\delta}^\star), \boldsymbol{\delta}^\star]) + \arccos(\angle[\boldsymbol{\delta}, \partial f(\boldsymbol{\delta})]) + \arccos(\min_{\boldsymbol{\delta}} \angle[\partial f(\boldsymbol{\delta}), \partial f(\boldsymbol{\delta}^\star)]).$$

Combined with the fact $\arccos(\angle[\partial f(\boldsymbol{\delta}^\star), \boldsymbol{\delta}^\star]) = 1$, we know the corresponding solid angle with respect to Δ_η^+ will be less or equal to $\pi/2$. Similar proof can be obtained for Δ_η^-.

Notice if a point $\boldsymbol{\delta}$ in the ball reaches the sphere at $\tilde{\boldsymbol{\delta}}$ by gradient descent, then the tangent direction along longitude at $\tilde{\boldsymbol{\delta}}$ and the direction of $\boldsymbol{\delta}^\star$ should be smaller than the angle between the $\partial f(\boldsymbol{\delta})$ and $\partial f(\boldsymbol{\delta}^\star)$. Then, by basic geometry, we know if

$$\arccos(\min_{\boldsymbol{\delta}, \boldsymbol{\delta}'} \angle[\partial f(\boldsymbol{\delta}), \partial f(\boldsymbol{\delta}')]) < \pi/4,$$

the trajectory initialized by drawing from a uniform distribution over $\mathbb{B}^\circ(\mathbf{0}, \varepsilon)$ will never reach Δ_η^-. Meanwhile, if there exists t^* such that $\boldsymbol{\delta}_{t^*} \in \Delta_\eta^+$, then for all $t \geqslant t^*$, $\boldsymbol{\delta}_t \in \Delta_\eta^+$.

A.5 Proof of Theorem 3.1 and Corollary 3.1

With the previous results, we are ready to state our main results. Recall for $\boldsymbol{\delta}, \boldsymbol{\delta}_0 \in \varepsilon\mathbb{S}^{d-1}$, if $\partial^2 L(\boldsymbol{\delta}, \lambda^*)$ is ρ-Lipschitz, that is $\|\partial^2 L(\boldsymbol{\delta}_1, \lambda^*) - \partial^2 L(\boldsymbol{\delta}_2, \lambda^*)\| \leqslant \rho\|\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\|$, we can obtain for any $\boldsymbol{\delta}, \boldsymbol{\delta}_0$ are on the sphere, we have

$$L(\boldsymbol{\delta}) \leqslant L(\boldsymbol{\delta}_0) + \Gamma(\boldsymbol{\delta}_0)^T(\boldsymbol{\delta} - \boldsymbol{\delta}_0) + \frac{1}{2}(\boldsymbol{\delta} - \boldsymbol{\delta}_0)^T\Xi(\boldsymbol{\delta}_0)(\boldsymbol{\delta} - \boldsymbol{\delta}_0) + \frac{\rho}{6}\|\boldsymbol{\delta} - \boldsymbol{\delta}_0\|^3.$$

Meanwhile, by Lemma 4.2, there exists approximation of PGD:

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \eta\Gamma(\boldsymbol{\delta}_t) + \tau_t,$$

where $\|\tau_t\| \leqslant O(\eta^2/\varepsilon)$ for all t. Combine the above two formulas, if we further have $\|\Xi\| \leqslant \nu$, it gives

$$\begin{aligned}
L(\boldsymbol{\delta}_{t+1}) - L(\boldsymbol{\delta}_t) &\leqslant \Gamma(\boldsymbol{\delta}_t)^T(\boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}_t) + \frac{1}{2}(\boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}_t)^T\Xi(\boldsymbol{\delta}_t)(\boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}_t) + \frac{\rho}{6}\|\boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}_t\|^3 \\
&= \Gamma(\boldsymbol{\delta}_t)^T(-\eta\Gamma(\boldsymbol{\delta}_t) + \tau_t) + \frac{1}{2}(-\eta\Gamma(\boldsymbol{\delta}_t) + \tau_t)^T\Xi(\boldsymbol{\delta}_t)(-\eta\Gamma(\boldsymbol{\delta}_t) + \tau_t) + \frac{\rho}{6}\|-\eta\Gamma(\boldsymbol{\delta}_t) + \tau_t\|^3 \\
&\leqslant -\eta\|\Gamma(\boldsymbol{\delta}_t)\|^2 + \Gamma(\boldsymbol{\delta}_t)^T\tau_t + \frac{1}{2}\eta^2\Gamma(\boldsymbol{\delta}_t)^T\Xi(\boldsymbol{\delta}_t)\Gamma(\boldsymbol{\delta}_t) + \frac{1}{2}\tau_t^T\Xi(\boldsymbol{\delta}_t)\tau_t - \eta\Gamma(\boldsymbol{\delta}_t)^T\Xi(\boldsymbol{\delta}_t)\tau_t \\
&+ \frac{\rho}{6}\big(\eta^3\|\Gamma(\boldsymbol{\delta}_t)\|^3 + \|\tau_t\|^3 + 3\eta^2\|\Gamma(\boldsymbol{\delta}_t)\|^2 \cdot \|\tau_t\| + 3\eta\|\Gamma(\boldsymbol{\delta}_t)\| \cdot \|\tau_t\|^2\big) \\
&\leqslant -\eta\|\Gamma(\boldsymbol{\delta}_t)\|^2 + \|\Gamma(\boldsymbol{\delta}_t)\|\frac{4\eta^2}{\varepsilon} + \frac{1}{2}\eta^2\nu\|\Gamma(\boldsymbol{\delta}_t)\|^2 + \frac{8\eta^4\nu}{\varepsilon^2} + \frac{4\eta^3\nu}{\varepsilon}\|\Gamma(\boldsymbol{\delta}_t)\| \\
&+ \frac{\rho}{6}\big(\eta^3\|\Gamma(\boldsymbol{\delta}_t)\|^3 + \frac{64\eta^6}{\varepsilon^3} + \frac{12\eta^4}{\varepsilon}\|\Gamma(\boldsymbol{\delta}_t)\|^2 + \frac{48\eta^5}{\varepsilon^2}\|\Gamma(\boldsymbol{\delta}_t)\|\big).
\end{aligned}$$

By Lemma A.2, we have $\nu \leqslant M_u \log m$, $\rho \leqslant K_u(\log m)^{3/2}$ for some M_u, K_u of order $\Theta(1)$ with high probability under the conditions given. Thus, there exists a threshold $\eta_{\max}(m, \varepsilon) = \min\{\Theta\big((\log m)^{-2}\big), \varepsilon^2\}$, if $\eta \leqslant \eta_{\max}(m, \varepsilon)$, whenever $\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t+1} \in \varepsilon\mathbb{S}^{d-1}$ and $\|\Gamma(\boldsymbol{\delta}_t)\| \geqslant \sqrt{\eta}$,

$$L(\boldsymbol{\delta}_{t+1}) - L(\boldsymbol{\delta}_t) \leqslant -\Omega(\eta^2).$$

We conclude the above statements by the following lemma.

Lemma A.5. *If $m = \Omega(d^{5/2})$, there exists $\varepsilon_{\max}(m) = \Theta\big((\log m)^{-1/2}\big)$ and $\eta_{\max}(m, \varepsilon) = \min\{\Theta\big((\log m)^{-2}\big), \varepsilon^2\}$, if $\varepsilon < \varepsilon_{\max}(m)$, $\eta < \eta_{\max}(m, \varepsilon)$*

$$\Big\|\frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}}\Big\| \leqslant B_u\sqrt{\log d}, \ \Big\|\frac{\partial^2 f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta} + \boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\| \leqslant M_u \log m,$$

$$\Big\|\frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial\boldsymbol{\delta}^2}-\frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}'+\boldsymbol{x})}{\partial\boldsymbol{\delta}^2}\Big\|\leqslant K_u(\log m)^{3/2}\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|,$$

for all $\boldsymbol{\delta}\in\mathbb{B}(\mathbf{0},\varepsilon)$*, where* M_u, K_u *are of order* $\Theta(1)$ *with high probability, besides whenever* $\boldsymbol{\delta}_t\in\varepsilon\mathbb{S}^{d-1}$ *and* $\|\Gamma(\boldsymbol{\delta}_t)\|\geqslant\sqrt{\eta}$*,*

$$L(\boldsymbol{\delta}_{t+1})-L(\boldsymbol{\delta}_t)\leqslant-\Omega(\eta^2).$$

Proof. By the nature of PGD, we know $\boldsymbol{\delta}_{t+1}$ is either in the ball $\mathbb{B}(\mathbf{0},\varepsilon)$ or on the sphere $\varepsilon\mathbb{S}^{d-1}$. When $\boldsymbol{\delta}_{t+1}\in\varepsilon\mathbb{S}^{d-1}$, by the analysis above, we have $L(\boldsymbol{\delta}_{t+1})-L(\boldsymbol{\delta}_t)\leqslant-\Omega(\eta^2)$. When $\boldsymbol{\delta}_{t+1}\in\mathbb{B}^\circ(\mathbf{0},\varepsilon)$, by Lemma 4.1,

$$L(\boldsymbol{\delta}_{t+1})-L(\boldsymbol{\delta}_t)\leqslant-\Omega(\eta).$$

Thus, to sum up, we have for both cases

$$L(\boldsymbol{\delta}_{t+1})-L(\boldsymbol{\delta}_t)\leqslant-\Omega(\eta^2).$$

□

Then, we need to deal with the case when $\|\Gamma(\boldsymbol{\delta}_t)\|\leqslant\sqrt{\eta}$. If we further denote the region $\Lambda^+=\{\boldsymbol{\delta}:\boldsymbol{v}^T\Xi(\boldsymbol{\delta})\boldsymbol{v}\geqslant\gamma,\boldsymbol{\delta}\in\varepsilon\mathbb{S}^{d-1}\}$ and $\Lambda^-=\{\boldsymbol{\delta}:\boldsymbol{v}^T\Xi(\boldsymbol{\delta})\boldsymbol{v}\leqslant-\gamma,\boldsymbol{\delta}\in\varepsilon\mathbb{S}^{d-1}\}$, where γ is the universal constant specified in Lemma 4.4. By Lemma 4.3 and 4.4, we know if $\sqrt{\eta}\leqslant\mathcal{L}B_l\sqrt{1-\phi^2}$, we have $\Delta_\eta^+\subseteq\Lambda^+$ and $\Delta_\eta^-\subseteq\Lambda^-$. Δ_η^+ are those points near local optimums, which we try to avoid being stuck at. Lemma 4.5 provides insights how can the trajectory avoids being stuck near the local optimums.

The following corollary can help us realize Lemma 4.5. Specifically, by zooming into the proof of Lemma 4.1, it is straightforward to obtain the following corollary.

Corollary A.1. *For any* $\boldsymbol{\delta},\boldsymbol{\delta}'\in\mathbb{B}(\mathbf{0},\varepsilon)$

$$\angle[\partial f(\boldsymbol{\delta}),\partial f(\boldsymbol{\delta}')]\to1,$$

as $\varepsilon\to0$ *in Eq. 7 under the setting of Lemma A.1 .*

Proof. Once we notice that

$$\begin{aligned}\angle[\partial f(\boldsymbol{\delta}),\partial f(\boldsymbol{\delta}')]-1&=\Big\langle\frac{\partial f(\boldsymbol{\delta})}{\|\partial f(\boldsymbol{\delta})\|}-\frac{\partial f(\boldsymbol{\delta}')}{\|\partial f(\boldsymbol{\delta}')\|},\frac{\partial f(\boldsymbol{\delta}')}{\|\partial f(\boldsymbol{\delta}')\|}\Big\rangle\\&\leqslant\frac{2\|\partial f(\boldsymbol{\delta})-\partial f(\boldsymbol{\delta}')\|}{\|\partial f(\boldsymbol{\delta})\|}\end{aligned}$$

and $\|\partial f(\boldsymbol{\delta})\|\geqslant\mathcal{L}B_l$, the proof is straightforward. □

Thus, if η,ε are smaller than some constant thresholds, then

$$\arccos\big(\min_{\boldsymbol{\delta},\boldsymbol{\delta}'}\angle[\partial f(\boldsymbol{\delta}),\partial f(\boldsymbol{\delta}')]\big)+\arccos\Big(\sqrt{1-\frac{\eta}{(\mathcal{L}B_l)^2}}\Big)\leqslant\frac{\pi}{4}$$

stands. As a result, the trajectory can successfully avoid being stuck near local maximums.

The only case left is when $\|\Gamma(\boldsymbol{\delta}_t)\|\leqslant\sqrt{\eta}$ and $\boldsymbol{\delta}_t\in\Delta_\eta^-$. If $\boldsymbol{\delta}^*$ is one of the local minimums and we focus on studying the case when $\boldsymbol{\delta}_t$ falls at the local neighborhood belongs to Δ_η^- corresponding to $\boldsymbol{\delta}^*$.

Notice that

$$\Gamma(\boldsymbol{\delta}_t)=\Gamma(\boldsymbol{\delta}^*)+\int_0^1\nabla\Gamma(\boldsymbol{\delta}^*+t(\boldsymbol{\delta}_t-\boldsymbol{\delta}^*))dt\cdot(\boldsymbol{\delta}_t-\boldsymbol{\delta}^*).$$

By looking up the derivative of $\Gamma(\boldsymbol{\delta})$, we have the following characterization:

$$\nabla\Gamma(\boldsymbol{\delta})=\Xi(\boldsymbol{\delta})-\nabla c(\boldsymbol{\delta})\nabla\lambda^*(\boldsymbol{\delta})^T.\tag{11}$$

Denote

$$N(\boldsymbol{\delta}) = -\nabla c(\boldsymbol{\delta})\nabla\lambda^*(\boldsymbol{\delta})^T.$$

If $\boldsymbol{\delta} \in \varepsilon\mathbb{S}^{d-1}$, $\nabla c(\boldsymbol{\delta}) = 2\boldsymbol{\delta}$ is parallel to the normal vector of the tangent space at $\boldsymbol{\delta}$, thus, we have for any $\boldsymbol{v}$, $N(\boldsymbol{\delta})\boldsymbol{v} \in \mathcal{T}^c(\boldsymbol{\delta})$.

The fact that $\nabla\Gamma(\boldsymbol{\delta}) = \Xi(\boldsymbol{\delta}) + N(\boldsymbol{\delta})$ is very important, since it gives the following extra characterization of $\boldsymbol{\delta}_t$: for small enough η, if $\|\Gamma(\boldsymbol{\delta}_t)\| \leqslant \sqrt{\eta}$, then $\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\| = O(\sqrt{\eta})$. Besides, since $\Gamma(\boldsymbol{\delta}^*) = 0$, if $\|\Gamma(\boldsymbol{\delta}_t)\| \leqslant \sqrt{\eta}$, $\boldsymbol{\delta}_t$ falls into a neighborhood of $\boldsymbol{\delta}^*$ such that Ξ has smallest eigenvalue larger or equal than $\gamma > 0$ as long as η is small enough. Besides, we have when $\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\| = \Omega(\sqrt{\eta})$, then $\|\Gamma(\boldsymbol{\delta}_t)\| = \Omega(\sqrt{\eta})$. The previous discussion can be formalized as the following lemma.

Lemma A.6. *For small enough η, if $\|\Gamma(\boldsymbol{\delta}_t)\| \leqslant \sqrt{\eta}$, and $\boldsymbol{\delta}_t$ is in the neighborhood of $\boldsymbol{\delta}^*$ such that Ξ has smallest eigenvalue larger or equal than γ, then*

$$\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\| = O(\sqrt{\eta}).$$

Furthermore, we have

$$\Gamma(\boldsymbol{\delta}_t)^T(\boldsymbol{\delta}_t - \boldsymbol{\delta}^*) \geqslant \frac{\gamma}{2}\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2.$$

Proof. By Lemma A.3, we know that for $\boldsymbol{\delta}_t, \boldsymbol{\delta}^* \in \varepsilon\mathbb{S}^{d-1}$

$$\|\mathcal{P}_{\mathcal{T}^c_{\boldsymbol{\delta}_t}}(\boldsymbol{\delta}_t - \boldsymbol{\delta}^*)\| \leqslant \frac{\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2}{2\varepsilon}.$$

As $\boldsymbol{\delta}^*$ is one of the minimizer on the sphere, we must have $\Gamma(\boldsymbol{\delta}^*) = 0$. Thus, for small enough α, if we denote $\boldsymbol{\delta}^{\mathcal{P}}_{\boldsymbol{v}} = \Pi_{\mathbb{B}(\mathbf{0},\varepsilon)}(\boldsymbol{\delta} + \alpha\boldsymbol{v})$, for any $\boldsymbol{v} \in \mathcal{T}(\boldsymbol{\delta})$ with norm 1

$$\begin{aligned}
\|\Gamma(\boldsymbol{\delta}^{\mathcal{P}}_{\boldsymbol{v}}) - \Gamma(\boldsymbol{\delta}^*)\| &\approx \|\nabla\Gamma(\boldsymbol{\delta}^*)(\boldsymbol{\delta}^{\mathcal{P}}_{\boldsymbol{v}} - \boldsymbol{\delta}^*)\| \\
&\geqslant \|\Xi(\boldsymbol{\delta}^*)(\boldsymbol{\delta}^{\mathcal{P}}_{\boldsymbol{v}} - \boldsymbol{\delta}^*)\| - \|N(\boldsymbol{\delta}^*)(\boldsymbol{\delta}^{\mathcal{P}}_{\boldsymbol{v}} - \boldsymbol{\delta}^*)\| \\
&\geqslant \gamma\alpha - \frac{4\alpha^2}{\varepsilon}(\|\Xi\| + \|N\|) \\
&\geqslant \frac{\gamma\alpha}{2}.
\end{aligned}$$

If we further denote $\mathcal{R}$ as the region where $\boldsymbol{\delta}$ has the following properties:

- smallest eigenvalue of $\Xi(\boldsymbol{\delta})$ larger or equal to γ;
- the distance from $\boldsymbol{\delta}$ to one of the minimizers is at least $\Omega(\eta)$,

with abuse of notations, we want to prove for any $\boldsymbol{\delta}$ belongs to $\mathcal{R}$, there is a path $\{\boldsymbol{\delta}_t\}$ to the region that the distance from $\boldsymbol{\delta}$ to one of the minimizers is at most $O(\eta)$, where $\boldsymbol{\delta}_0 = \boldsymbol{\delta}$, such that $\|\Gamma(\boldsymbol{\delta}_t)\|$ is decreasing along the path. If that statement is true let $\alpha = c\sqrt{\eta}$ for some constant c, and η/ε^4 is small enough, then we know $\|\Gamma(\boldsymbol{\delta})\| \leqslant \sqrt{\eta}$ implies $\boldsymbol{\delta}$ being very close to one of the minimizers, distance up to $O(\sqrt{\eta})$. So

$$\begin{aligned}
\Gamma(\boldsymbol{\delta}_t)^T(\boldsymbol{\delta}_t - \boldsymbol{\delta}) &= (\boldsymbol{\delta}_t - \boldsymbol{\delta}^*)^T\int_0^1 \nabla\Gamma(\boldsymbol{\delta}^* + t(\boldsymbol{\delta}_t - \boldsymbol{\delta}^*))dt \cdot (\boldsymbol{\delta}_t - \boldsymbol{\delta}^*) \\
&\geqslant \gamma\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2 - O(\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^3) \\
&\geqslant \frac{\gamma}{2}\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2.
\end{aligned}$$

Finally, we show we can always find such path that of decreasing norm of Γ. Notice

$$\frac{d\|\Gamma(\boldsymbol{\delta})\|^2}{2dt} = \langle\Gamma(\boldsymbol{\delta}_t), \Xi\frac{d\boldsymbol{\delta}_t}{dt}\rangle.$$

If $d\boldsymbol{\delta}_t/dt = -\varpi\Gamma(\boldsymbol{\delta}_t)$ for some constant ϖ, then $\langle \Gamma(\boldsymbol{\delta}_t), \Xi\frac{d\boldsymbol{\delta}_t}{dt}\rangle \leqslant -\varpi\gamma\|\Gamma(\boldsymbol{\delta}_t)\|^2$, which implies the norm is decreasing along the path. Thus, for discrete version $\boldsymbol{\delta}_{t+1} = \Pi_{\mathbb{B}(\mathbf{0},\varepsilon)}(\boldsymbol{\delta}_t - \varpi\nabla L(\boldsymbol{\delta}_t))$, as long as ϖ is small enough (can be much smaller than η), combine with Lemma 4.2, the dominating term of $d\|\Gamma(\boldsymbol{\delta}_t))\|^2/dt$ will be negative, thus, the proof is complete.

□

Now we are ready to state the convergence result when $\boldsymbol{\delta}_t \in \varepsilon\mathbb{S}^{d-1}$ is in a neighborhood of one of the minimizers $\boldsymbol{\delta}^*$. We have shown that eventually for some t, $\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\| \leqslant O(\sqrt{\eta})$. We need further to show that the trajectory will remain near $\boldsymbol{\delta}^*$ ever since.

Lemma A.7. *For small enough η, for a $T > 0$ such that $\|\Gamma(\boldsymbol{\delta}_T)\| \leqslant \sqrt{\eta}$, and is in the neighborhood of a minimizer $\boldsymbol{\delta}^*$ and Ξ has smallest eigenvalue larger or equal than γ, for any $t \geqslant T$,*

$$\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\| \leqslant O(\sqrt{\eta}).$$

Proof. Notice that if $\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\| \leqslant c\sqrt{\eta}$ for some constant $c > 0$,

$$\begin{aligned}
\|\boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}^*\|^2 &= \|\boldsymbol{\delta}_t - \eta\Gamma(\boldsymbol{\delta}_t) + \iota_t - \boldsymbol{\delta}^*\|^2 \\
&= \|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2 - 2\eta\Gamma(\boldsymbol{\delta}_t)^T(\boldsymbol{\delta}_t - \boldsymbol{\delta}^*) + 2\iota_t^T(\boldsymbol{\delta}_t - \boldsymbol{\delta}^*) + \|\eta\Gamma(\boldsymbol{\delta}_t) - \iota_t\|^2 \\
&\leqslant (1 - \gamma\eta)\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2 + \|2\iota_t\|\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\| + 2c\eta^3 + \frac{32\eta^4}{\varepsilon^2} \\
&\leqslant (1 - \gamma\eta)\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2 + \frac{8\eta^{2.5}}{\varepsilon} + o(\eta^2).
\end{aligned}$$

Then, $\|\boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}^*\| \leqslant \sqrt{\eta}$ for small enough η and $\eta^{0.5}/\varepsilon = o(1)$. Further,

$$\|\boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}^*\|^2 - \frac{9\eta}{\gamma} \leqslant (1 - \gamma\eta)(\|\boldsymbol{\delta}_t - \boldsymbol{\delta}^*\|^2 - \frac{9\eta}{\gamma}).$$

Then, the proof is straightforward. □

Shrinking step size η_t The above discussions are all about constant η. Now, we further discuss about shrinking step size. Specifically, after $\|\Gamma(\boldsymbol{\delta}_t)\|$ reaches $\sqrt{\eta}$, we can shrink the learning rate with suitable $\eta_0 < \eta$, and $\{\eta_s\}_{s\geqslant 0}$ are strictly decreasing with respect to $s \geqslant 0$, such that

$$\boldsymbol{\delta}_{t+s+1} = \Pi_{\mathbb{B}(\mathbf{0},\varepsilon)}\left[\boldsymbol{\delta}_{t+s} - \eta_s\nabla L(\boldsymbol{\delta}_{t+s})\right]$$

By Lemma A.7, for small enough η and $\eta^{0.5}/\varepsilon = o(1)$, notice that $\|\boldsymbol{\delta}_{t+s} - \boldsymbol{\delta}^*\| \leqslant \varepsilon$, we can still have

$$\|\boldsymbol{\delta}_{t+1+s} - \boldsymbol{\delta}^*\|^2 \leqslant (1 - \gamma\eta_s)\|\boldsymbol{\delta}_{t+s} - \boldsymbol{\delta}^*\|^2 + 9\eta_s^2.$$

For simplicity, we denote $\|\boldsymbol{\delta}_{t+s} - \boldsymbol{\delta}^*\|^2$ as D_s. We would show if $\eta_s \to 0$, $D_s \to 0$.

Lemma A.8. *As long as $\eta_s \to 0$ and $\Pi_{i=0}^k(1-\gamma\eta_i/2) \to 0$, we can obtain $D_s \to 0$. Furthermore, if*

$$\eta_s\Pi_{i=0}^k(1 - \frac{\gamma\eta_{s+i}}{2}) \leqslant \eta_{s+k+1}$$

for all $s, k \in \mathbb{N}$, then for all $s \in \mathbb{N}$,

$$D_s \leqslant O(\eta_s).$$

Specifically, if $\eta_s = 1/(s + z)$ for large enough integer z,

$$D_s \leqslant O(\frac{1}{z + s}).$$

Proof. First, if there exists $S \in \mathbb{N}^+$, such that for all $s \geqslant S$, we have

$$D_s \geqslant \frac{18\eta_s}{\gamma},$$

then we have

$$D_{s+1} \leqslant (1-\gamma\eta_s)D_s + 9\eta_s^2 \leqslant (1-\frac{\gamma\eta_s}{2})D_s.$$

As a result, $\forall s \geqslant S$, and $k \in \mathbb{N}^+$

$$D_{s+k+1} \leqslant \Pi_{i=s}^{s+k}(1-\frac{\gamma\eta_i}{2})D_s.$$

On the other hand, if there does not exist such S, there exists infinitely many s, such that

$$D_s < \frac{18\eta_s}{\gamma}.$$

Moreover, if $D_s < 18\eta_s/\gamma$

$$\begin{aligned} D_{s+1} &\leqslant (1-\gamma\eta_s)D_s + 9\eta_s^2 \\ &\leqslant (1-\gamma\eta_s)\frac{18\eta_s}{\gamma} + 9\eta_s^2 \\ &\leqslant \frac{18\eta_s}{\gamma}. \end{aligned}$$

So, for any $\varepsilon > 0$, we can choose large enough s, such that $D_s < \frac{18\eta_s}{\gamma}$, and $D_{s+1+i} < \varepsilon$ for any $i \geqslant 0$. So, we will always have $D_t \to 0$ as $t \to \infty$ as long as $\Pi_{i=0}^{k}(1-\gamma\eta_i/2) \to_{k\to\infty} 0$ and $\eta_k \to_{k\to\infty} 0$.

Besides, we have

$$D_{s+1} \leqslant \max\Big\{(1-\frac{\gamma\eta_s}{2})D_s, \frac{18\eta_s}{\gamma}\Big\}.$$

Notice if

$$\eta_s\Pi_{i=0}^{k}(1-\frac{\gamma\eta_{s+i}}{2}) \leqslant \eta_{s+k+1} \tag{12}$$

for all $s, k \in \mathbb{N}$, then for all $s \in \mathbb{N}$, we can obtain

$$D_s \leqslant \frac{18\eta_s}{\gamma}.$$

For example, if $\eta_s = 2/(\gamma s + \gamma z)$ for large enough integer z, Eq. 12 is satisfied by simple algebra and we have

$$D_s \leqslant O(\frac{1}{z+s}).$$

□

Now we are ready to state the proof of our main theorem.

[Proof of Theorem 3.1 and and Corollary 3.1] Under the assumptions, we have the following properties hold simultaneously:

a .

$$B_l \leqslant \Big\|\frac{\partial f(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}}\Big\| \leqslant B_u\sqrt{\log d},$$

for all $\boldsymbol{\delta} \in \mathbb{B}(\mathbf{0}, \varepsilon)$, where B_l, B_u is of order $\Theta(1)$.

b .

$$\Big\|\frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\| \leqslant M_u \log m\,, \Big\|\frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})}{\partial \boldsymbol{\delta}^2}-\frac{\partial^2 f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}'+\boldsymbol{x})}{\partial \boldsymbol{\delta}^2}\Big\| \leqslant K_u(\log m)^{3/2}\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|,$$

for all $\boldsymbol{\delta} \in \mathbb{B}(\mathbf{0},\varepsilon)$, where M_u, K_u are of order $\Theta(1)$.

c.

$$\arccos\big(\min_{\boldsymbol{\delta},\boldsymbol{\delta}'} \angle[\partial f(\boldsymbol{\delta}),\partial f(\boldsymbol{\delta}')]\big) + \arccos\Big(\sqrt{1-\frac{\eta}{(\mathcal{L}B_l)^2}}\Big) \leqslant \frac{\pi}{4}.$$

d. There exists a threshold $\varepsilon_{\max}(m) = \Theta((\log m)^{-1})$, if $\varepsilon < \varepsilon_{\max}$, with high probability, there exists universal constants $\phi, \gamma > 0$, for any $\boldsymbol{\delta} \in \varepsilon\mathbb{S}^{d-1}$, such that

$$\angle[\partial f(\boldsymbol{\delta}),\boldsymbol{\delta}] \geqslant \phi,$$

then for all $\|\boldsymbol{v}\| = 1$,

$$\operatorname{sgn}\Big((y-u)\boldsymbol{\delta}^T\partial f(\boldsymbol{\delta})\Big)\cdot \boldsymbol{v}^T\Xi\boldsymbol{v} \geqslant \gamma.$$

As a result, whenever $\boldsymbol{\delta}_{t+1} \in \mathbb{B}^\circ(\mathbf{0},\varepsilon)$

$$L(\boldsymbol{\delta}_{t+1}) - L(\boldsymbol{\delta}_t) \leqslant -\Omega(\eta).$$

Whenever $\boldsymbol{\delta}_t \in \varepsilon\mathbb{S}^{d-1}$, condition c will ensure $\boldsymbol{\delta}_t \notin \Delta_\eta^-$ so that he trajectory will not stuck near local maximums. Besides, we have $\|\Gamma(\boldsymbol{\delta})\| \geqslant \sqrt{\eta}$ if $\boldsymbol{\delta} \notin \Delta_\eta^+$ for $\boldsymbol{\delta} \in \varepsilon\mathbb{S}^{d-1}$. That can ensure $L(\boldsymbol{\delta}_{t+1}) - L(\boldsymbol{\delta}_t) \leqslant -\Omega(\eta^2)$ for $\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t+1} \in \varepsilon\mathbb{S}^{d-1}$.

From the above discussions, we divide the ball into three regions. Let $\mathcal{R}_1 = \mathbb{B}^\circ(\mathbf{0},\varepsilon)$ be the interior of the ball. Let $\mathcal{R}_2 = \Delta_\eta^+$ and $\mathcal{R}_3 = \mathbb{B}(\mathbf{0},\varepsilon) \cap (\mathcal{R}_1 \cup \mathcal{R}_2)^c$. Since there exists $\mathcal{L},\mathcal{U} > 0$ such that $\mathcal{L} < |y - f(\boldsymbol{a},\boldsymbol{W},\boldsymbol{\delta}+\boldsymbol{x})| < \mathcal{U}$ for all $\boldsymbol{\delta} \in \mathbb{B}(\mathbf{0},\varepsilon)$, we claim at most $O(\eta^{-2})$ iterations, the trajectory will arrive at $\mathcal{R}_2$. That is because each step will have at least $O(\eta^2)$ progress in decreasing the value of loss if $\boldsymbol{\delta}_t \notin \mathcal{R}_2$.

Lastly, when $\|\Gamma(\boldsymbol{\delta})\| \leqslant \sqrt{\eta}$, the results follow by applying Lemma A.7 and A.8.

A.6 Proof of Theorem 5.1

Formal Statement of Theorem 5.1 Recall from Lemma 4.4, for $m = \Omega(d^{5/2})$, there exists a threshold $\varepsilon_{\max}(m) = \Theta((\log m)^{-1})$, if $\varepsilon < \varepsilon_{\max}$, with high probability, there exists universal constants $\phi, \gamma > 0$, for any $\boldsymbol{\delta} \in \varepsilon\mathbb{S}^{d-1}$, such that

$$\angle[\partial f(\boldsymbol{\delta}),\boldsymbol{\delta}] \geqslant \phi,$$

then for all $\|\boldsymbol{v}\| = 1$,

$$\operatorname{sgn}\Big((y-u)\boldsymbol{\delta}^T\partial f(\boldsymbol{\delta})\Big)\cdot \boldsymbol{v}^T\Xi\boldsymbol{v} \geqslant \gamma.$$

Based on that, there also exists a threshold $\tau_\varepsilon > 0$, such that when $\varepsilon < \tau_\varepsilon$,

$$\min_{\boldsymbol{\delta},\boldsymbol{\delta}'} \angle[\partial f(\boldsymbol{\delta}),\partial f(\boldsymbol{\delta}')]\big) \geqslant \phi,$$

and there is only one minimum on the sphere in that case.

Proof. Notice $\min_{\boldsymbol{\delta},\boldsymbol{\delta}'} \angle[\partial f(\boldsymbol{\delta}),\partial f(\boldsymbol{\delta}')]\big) \to 1$ as $\varepsilon \to 0$, so we know if ε is small enough, we will have $\min_{\boldsymbol{\delta},\boldsymbol{\delta}'} \angle[\partial f(\boldsymbol{\delta}),\partial f(\boldsymbol{\delta}')]\big) \geqslant \phi$. That means the solid cone formed by $\partial f(\boldsymbol{\delta})$ is included in the corresponding solid cone of $\Lambda^+ = \{\boldsymbol{\delta} : \boldsymbol{v}^T\Xi(\boldsymbol{\delta})\boldsymbol{v} \geqslant \gamma, \boldsymbol{\delta} \in \varepsilon\mathbb{S}^{d-1}\}$.

Assume there are two local minimums, actually in Λ^+ the local minimums are strict , then there exists a path on the sphere such that there is a local maximum on this path. However, that is impossible since the Hessian approximate is positive definite. □

B More About Experiments

B.1 Implementation Details

Loss landscapes on simulated data. In our experiments, we use a two-layer neural network with the hidden size of 16 and the initialization is as Sec. 2. We first randomly choose an two-dimensional input x with a norm smaller than the input scale. The epsilon ϵ is the product of the perturbation ratio r and the input scale. We then randomly choose 10000 perturbations in the epsilon ball. The adversarial losses of these perturbations on the input x are shown in Figure 1. The choice of the input x is not important in our experiments and the landscapes based on another random choice is shown in Sec. B.2. The impact of the width of the hidden layer is also shown in Sec. B.2.

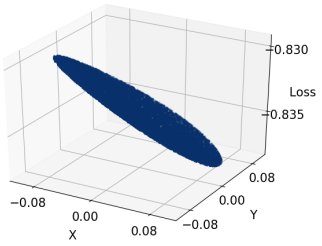
Trajectories on simulated data. We use the same settings for neural networks as those in the landscapes. We choose the perturbation with the maximal loss among the 10000 random sampled perturbations as our local maxima. To show the trajectories, we conduct PGD 10 times with the best learning rate from 1e−6, 1e−5, 1e−4, 1e−3, 1e−2, 1e−1, 1, 1e 1, 1e2, 1e3.

Trajectories on real-world data. Our experiments in Fig. 2 are based on a real-world dataset MNIST. We use the same multiple-layer CNN architecture except the dropout in `https://github.com/pytorch/examples/tree/master/mnist`. We change the original ten-class classification to binary classification to distinguish odd and even numbers. Because the inputs are high-dimensional (28×28), we instead show the loss of PGD from the local maxima to the local minima. We first randomly sample an image from MNIST as our input x. We then start with a random perturbation and use PGD to find the local maxima. After tuning the hyperparameters, we find that running 1000 epochs of PGD with a learning rate of 1.0 can achieve good enough local maxima. After that, we run 1000 epochs of PGD with a learning rate of 1.0 to show the adversarial loss of each point on the trajectory from the local maxima to the local minima.

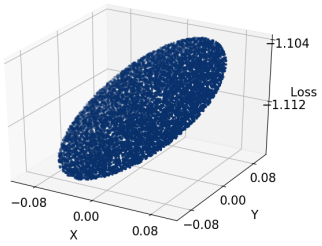
Dynamics of trajectories on real-world data. We use the same setting as that in Fig. 2. To train the CNN model, we randomly sample 100 images with odd numbers and 100 figures with even numbers from MNIST as our training data. We set the learning rate of adversarial training as $0.01 \times r$, where r is the perturbation ratio.

B.2 Additional Results

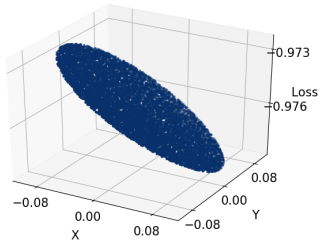
We further analyze the impact of the hidden layer's width on the landscapes in Fig 4 and the landscapes with a different random seed are shown in Fig 5. We try several different random seeds and find that the results are all consistent with our analysis. More details can be found in the code.



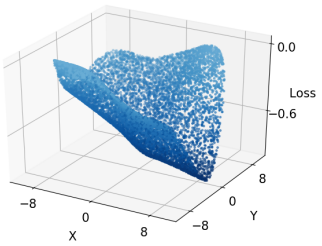
(a) Input scale=0.01, hidden size=16



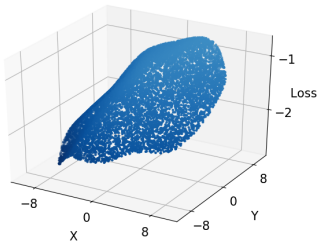
(b) Input scale=0.01, hidden size=128



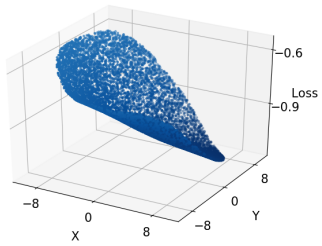
(c) Input scale=0.01, hidden size=1024



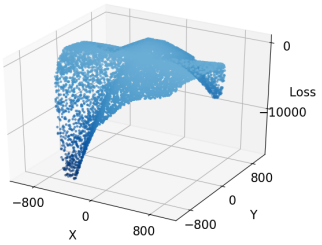
(d) Input scale=1.0, hidden size=16



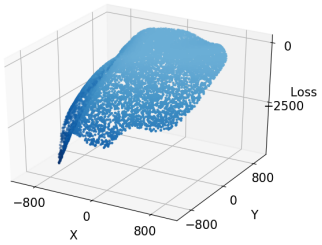
(e) Input scale=1.0, hidden size=128



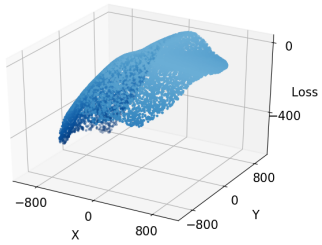
(f) Input scale=1.0, hidden size=1024



(g) Input scale=100, hidden size=16

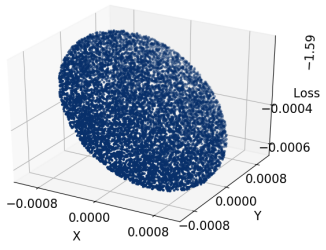


(h) Input scale=100, hidden size=128

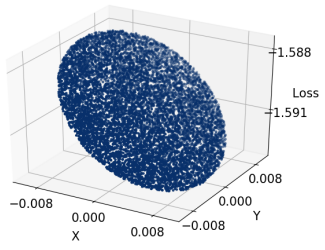


(i) Input scale=100, hidden size=1024

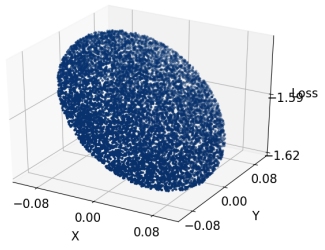
Figure 4: Landscapes of adversarial losses on simulated data with different hidden sizes and different input scales. We here fix the perturbation ratio as 10. We find that wider neural networks lead to more regular landscapes in general.



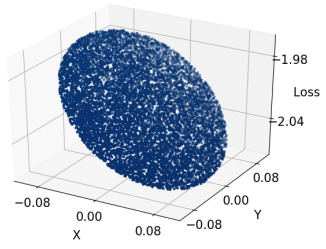
(a) Input scale=0.01, ratio=0.1



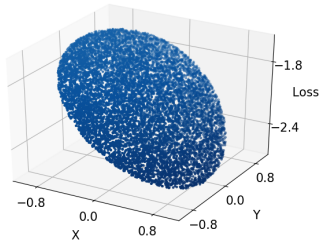
(b) Input scale=0.01, ratio=1.0



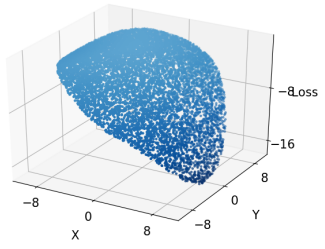
(c) Input scale=0.01, ratio=10



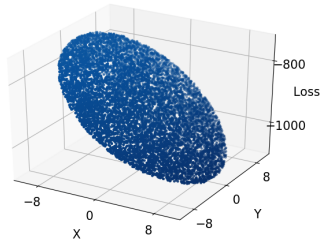
(d) Input scale=1.0, ratio=0.1



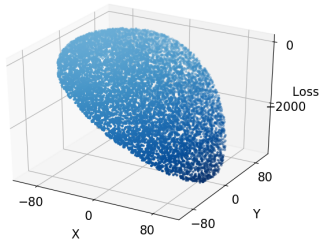
(e) Input scale=1.0, ratio=1.0



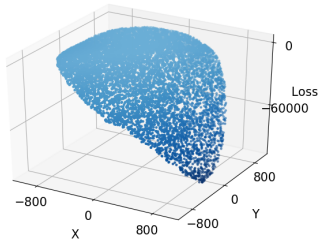
(f) Input scale=1.0, ratio=10



(g) Input scale=100, ratio=0.1



(h) Input scale=100, ratio=1.0



(i) Input scale=100, ratio=10

Figure 5: Landscapes of adversarial losses on simulated data with another random seed.