Deep Networks and the Multiple Manifold Problem

Sam Buchanan

Department of Electrical Engineering, Data Science Institute

Columbia University

Joint with:

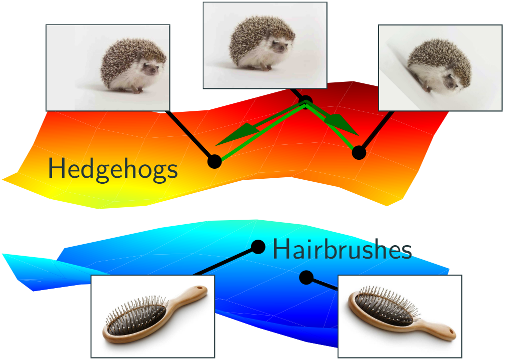
Dar Gilboa

John Wright

Motivation

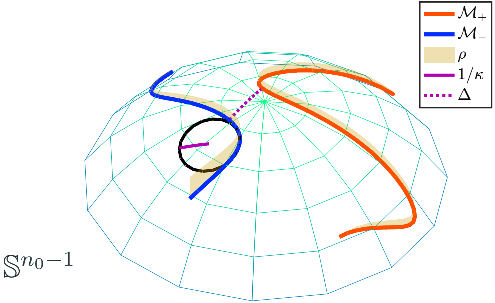
Our focus:

Provable guarantees for training deep networks to classify structured data.



Pope et al. (2021): $\dim(\text{ImageNet}) \approx 43$, $\dim(\text{CIFAR-10}) \approx 26$

Two Manifold Problem (One-Dimensional)

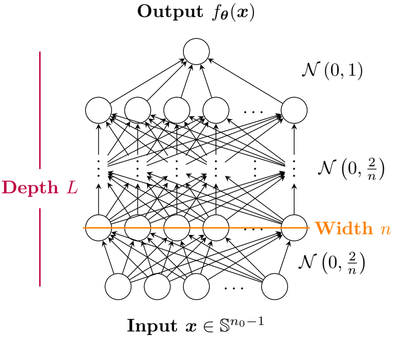


Problem. Given N i.i.d. labeled samples $(\boldsymbol{x}_1, f_\star(\boldsymbol{x}_1))$, ..., $(\boldsymbol{x}_N, f_\star(\boldsymbol{x}_N))$ from $\mathcal{M} = \mathcal{M}_+ \cup \mathcal{M}_-$, use gradient descent to train a deep network $f_{\boldsymbol{\theta}}$ that *perfectly labels the manifolds*:

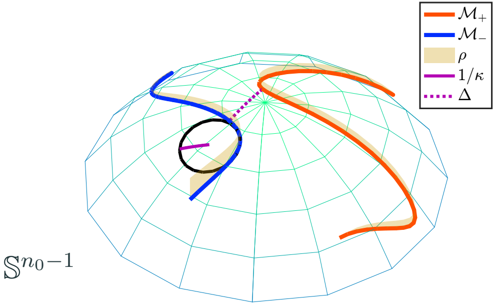
$$\operatorname{sign}(f_{\boldsymbol{\theta}}(\boldsymbol{x})) = f_\star(\boldsymbol{x}) \quad \text{for all} \quad \boldsymbol{x} \in \mathcal{M}.$$

Network Architecture

- Fully connected with ReLUs
- Gaussian initialization $\boldsymbol{\theta}_0$
- Trained with N i.i.d. samples from density ρ by gradient descent on empirical MSE (step size τ)



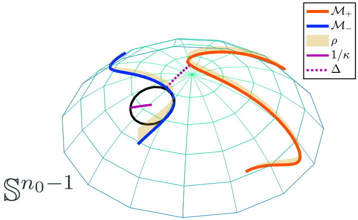
Two Curve Problem: Instance Parameters

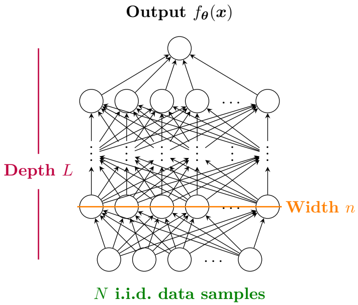


Problem difficulty parameters:

- *Class separation* Δ;
- *Class curvatures* κ;
- *Density properties* $\inf_{\boldsymbol{x}\in\mathcal{M}} \rho(\boldsymbol{x})$, ...

Two Curve Problem: Resource Tradeoffs





Theory question: How should we set our resources (depth L, width n, samples N) relative to the data structure (separation Δ, curvature κ, density ρ) so that *gradient descent succeeds*?

Main Results: Certificates Imply Generalization

Definition. $g : \mathcal{M} \to \mathbb{R}$ is called a *certificate* if for all $\boldsymbol{x} \in \mathcal{M}$

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - f_\star(\boldsymbol{x}) \underset{\text{square}}{\overset{\text{mean}}{\approx}} \int_{\mathcal{M}} \underbrace{\langle \widetilde{\nabla} f_{\boldsymbol{\theta}_0}(\boldsymbol{x}), \widetilde{\nabla} f_{\boldsymbol{\theta}_0}(\boldsymbol{x}') \rangle}_{\text{the ``NTK''}, \Theta(\boldsymbol{x},\boldsymbol{x}')} g(\boldsymbol{x}')\rho(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'$$

and $\int_{\mathcal{M}} (g(\boldsymbol{x}'))^2 \rho(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'$ is small.

Main Results: Certificates Imply Generalization

Definition. $g : \mathcal{M} \to \mathbb{R}$ is called a *certificate* if for all $\boldsymbol{x} \in \mathcal{M}$

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - f_\star(\boldsymbol{x}) \overset{\text{mean}}{\underset{\text{square}}{\approx}} \int_{\mathcal{M}} \underbrace{\langle \widetilde{\nabla} f_{\boldsymbol{\theta}_0}(\boldsymbol{x}), \widetilde{\nabla} f_{\boldsymbol{\theta}_0}(\boldsymbol{x}') \rangle}_{\text{the "NTK"}, \Theta(\boldsymbol{x},\boldsymbol{x}')} g(\boldsymbol{x}') \rho(\boldsymbol{x}') \, \mathrm{d}\boldsymbol{x}'$$

and $\int_{\mathcal{M}} (g(\boldsymbol{x}'))^2 \rho(\boldsymbol{x}') \, \mathrm{d}\boldsymbol{x}'$ is small.

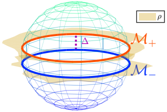
Theorem. If *a certificate exists*, if $\tau \asymp 1/(nL)$, and if

$$L \geq \mathrm{poly}(\kappa, C_\rho, C_{\mathcal{M}}, \log n_0),$$
$$n \asymp \mathrm{poly}(L),$$
$$N \geq \mathrm{poly}(L),$$

then with high probability the manifolds are classified perfectly after no more than L^2 gradient updates.

Main Results: Generalization for a Simple Geometry

Proposition. If additionally $L \gtrsim \Delta^{-1}$, then with high probability a certificate exists for the coaxial circle geometry.



Corollary. *For the two circles geometry*, if $\tau \asymp 1/(nL)$, and

$$
\begin{aligned}
L &\gtrsim \Delta^{-1} + \mathrm{poly}(C_\rho, \log n_0),\\
n &\asymp \mathrm{poly}(L),\\
N &\geq \mathrm{poly}(L),
\end{aligned}
$$

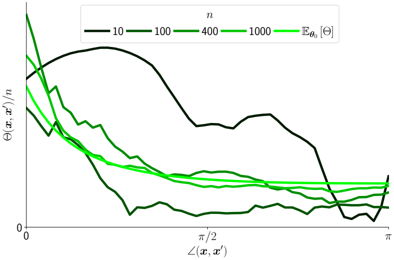
then with high probability the circles are classified perfectly after no more than L^2 gradient updates.

With Tingran Wang: certificates for **general curves**!

Intuitions for the Proof: Width as a Statistical Resource

Key role of width in the analysis:

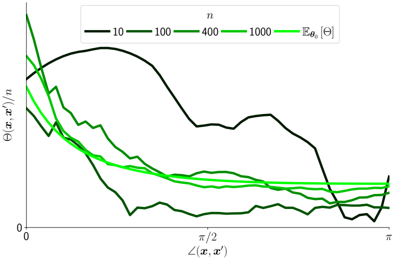
- Ensuring Θ is *uniformly* close to its expectation over $\boldsymbol{\theta}_0$ throughout training.



Intuitions for the Proof: Width as a Statistical Resource

Key role of width in the analysis:

- Ensuring Θ is *uniformly* close to its expectation over $\boldsymbol{\theta}_0$ throughout training.



We prove concentration for manifolds of arbitrary dimension $d_0 \geq 1$.

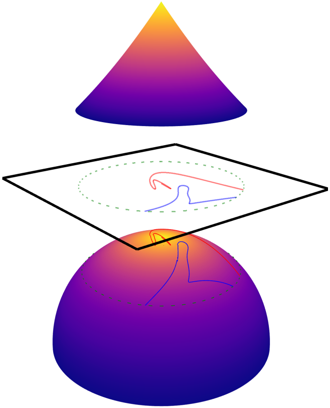
Theorem. If $n \gtrsim L(d_0 \log(n_0 L))^4$, then with high probability, simultaneously for all $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{M}$

$$\frac{\left|\Theta(\boldsymbol{x}, \boldsymbol{x}') - n \lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{x}, \boldsymbol{x}')\right]\right|}{n \lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{x}, \boldsymbol{x}')\right]} \lesssim \sqrt{\frac{L(d_0 \log(n_0 L))^4}{n}}.$$

Intuitions for the Proof: Depth as a Fitting Resource

- $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{x}, \boldsymbol{x}')\right]$ measures gradient descent's ability *to change* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ *without affecting* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x}')$.

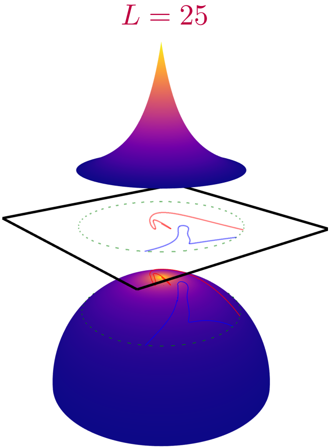
$L = 5$



$\frac{1}{L}\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{e}_1, \boldsymbol{x}')\right]$,
$\boldsymbol{x}' \in \mathbb{S}^2$

Intuitions for the Proof: Depth as a Fitting Resource

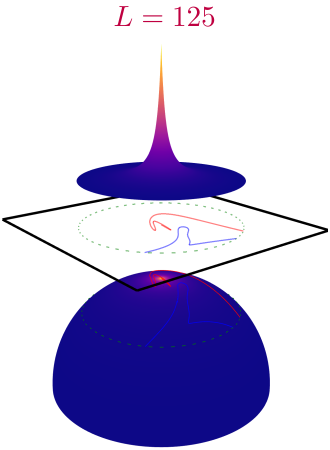
- $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{x}, \boldsymbol{x}')\right]$ measures gradient descent's ability *to change* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ *without affecting* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x}')$.
- Sharpness **increases with depth**.



$\frac{1}{L}\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{e}_1, \boldsymbol{x}')\right]$,
$\boldsymbol{x}' \in \mathbb{S}^2$

Intuitions for the Proof: Depth as a Fitting Resource

- $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{x}, \boldsymbol{x}')\right]$ measures gradient descent's ability *to change* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ *without affecting* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x}')$.
- Sharpness **increases with depth**.

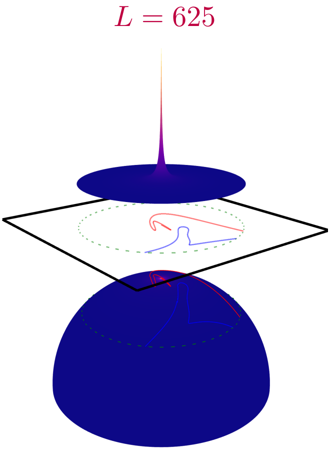


$\frac{1}{L}\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{e}_1, \boldsymbol{x}')\right]$,
$\boldsymbol{x}' \in \mathbb{S}^2$

Intuitions for the Proof: Depth as a Fitting Resource

- $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{x}, \boldsymbol{x}')\right]$ measures gradient descent's ability *to change* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ *without affecting* $f_{\boldsymbol{\theta}_0}(\boldsymbol{x}')$.
- Sharpness **increases with depth**.

$\Longrightarrow$ **set depth based on geometry!**



$$\frac{1}{L}\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\Theta(\boldsymbol{e}_1, \boldsymbol{x}')\right],$$
$$\boldsymbol{x}' \in \mathbb{S}^2$$

For More Details...

1. Technical proof sketch: Section A.4
2. Discussion of open problems: Section 4

Poster Session 11, May 6th
12 p.m.–2 p.m. EDT (UTC−4)

Thanks for listening!