Deep Networks and the Multiple Manifold Problem

John Wright

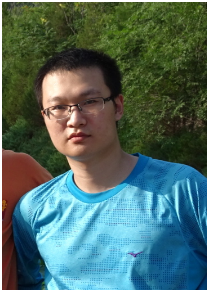
EE / APAM / DSI

Columbia University

Joint with...

Sam Buchanan

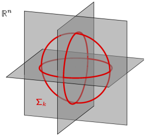
Dar Gilboa

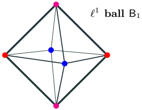


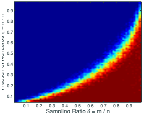
Tingran Wang

Model Problems: Sparse Approximation

Recover **sparse** x_0 from observations $\boldsymbol{y} = \boldsymbol{A}x_0$.



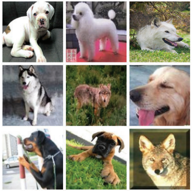


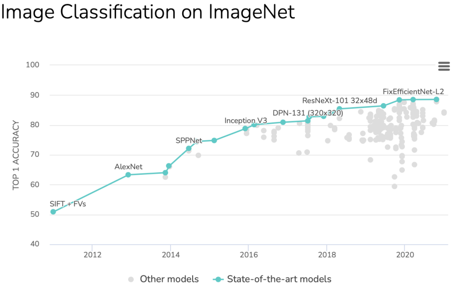


Many insights ... biased subsample:

- Structure
- Isometry
- Certificates of optimality

Model Problems for Deep Learning?





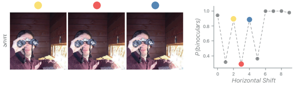
Many insights … biased subsample:

- Depth . . .
- Isometry . . .
- Overparameterization . . .

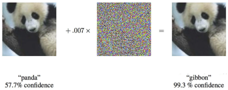
[1] Figure credits: [Deng et. al. '09] (left), `paperswithcode.com` (right)

Mathematical Model Problems for Deep Learning?

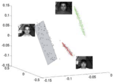
Issues that are hard to address using only datasets:



Uniformity?



Robustness?



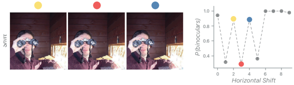
Data Structure?

What are good model problems
for mathematical analysis of deep networks?

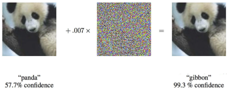
[2] Figure: [Azulay + Weiss]

Mathematical Model Problems for Deep Learning?

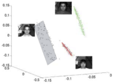
Issues that are hard to address using only datasets:



Uniformity?



Robustness?



Data Structure?

What are good model problems for mathematical analysis of deep networks?

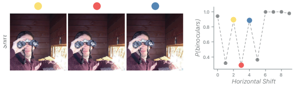
This Talk: one (failed?) attempt to answer this question.

[Plenty of other great existing answers: optimization landscapes, GAN's, implicit regularization, multiple descent, invariance ...]

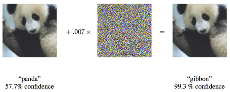
[2] Figure: [Azulay + Weiss]

Mathematical Model Problems for Deep Learning?

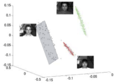
Issues that are hard to address using only datasets:



Uniformity?



Robustness?



Data Structure?

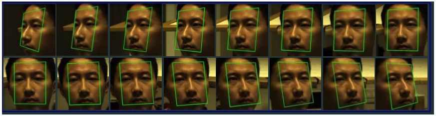
What are good model problems
for mathematical analysis of deep networks?

This Talk: how do deep networks compute with
low-dimensional (manifold) structure?

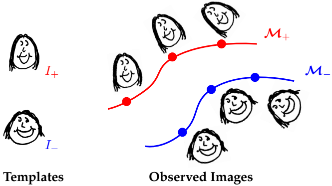
[3] Figure: [Azulay + Weiss]

Manifold Structure: Vision

Statistical and **structural** variabilities in visual data:



Invariant template matching: $\Longrightarrow$ **multiple low-d manifolds**:

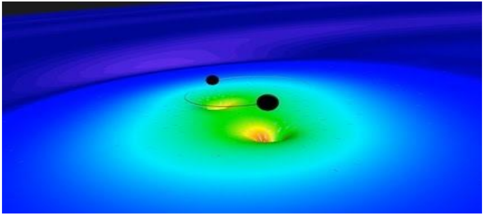


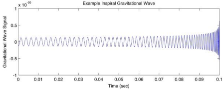
More complicated datasets [Pope et. al.]: CIFAR-10 26-d?, ImageNet 43-d?

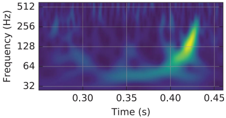
Manifold Structure: Science

Gravitational Wave Astronomy [with Marka, Marka, Yan, Colgan]

One binary black hole merger:



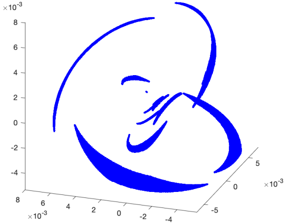




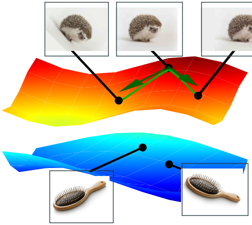
Many mergers

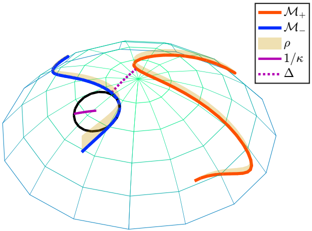
(varying mass M_1, M_2):

$\Longrightarrow$ **low-dim manifold**



The Multiple Manifold Problem

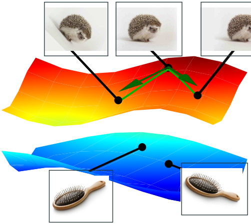


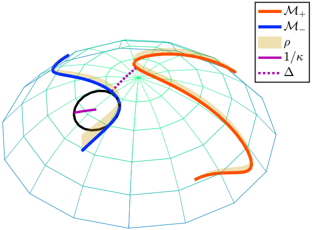


Problem: Given labeled data samples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ lying on manifolds $\mathcal{M}_\pm \subset \mathbb{S}^{n_0-1}$, learn a classifier $f_{\boldsymbol{\theta}}$ that correctly labels *every* point on the two manifolds:

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

Multiple Manifold Problem: Geometric Hypotheses

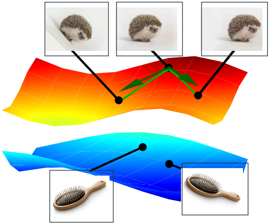


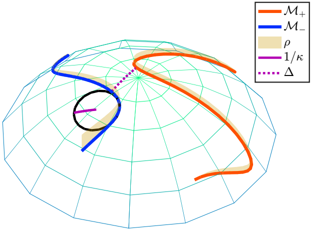


Geometric problem parameters:

- *dimension* d,
- *curvature* κ,
- *separation* Δ,
- *clover number* ✿.

Multiple Manifold Problem: Geometric Hypotheses

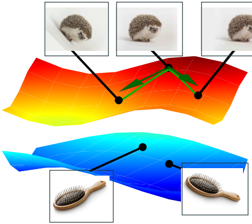


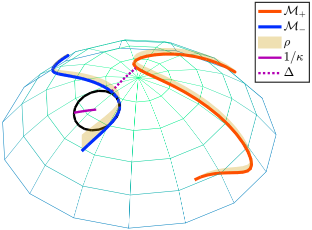


Geometric problem parameters:

- *dimension* d: here, curves – $d = 1$!
- *curvature* κ,
- *separation* Δ,
- *clover number* ✤.

Multiple Manifold Problem: Geometric Hypotheses

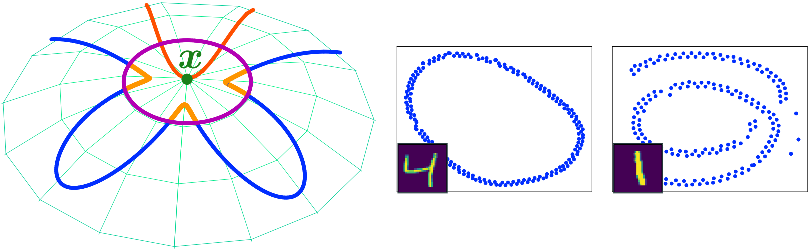




Geometric problem parameters:

- *dimension* d: here, curves – $d = 1$!
- *curvature* κ,
- *separation* Δ,
- *clover number* ✿: next slide...

✤ number: How "loopy" is $\mathcal{M}$?

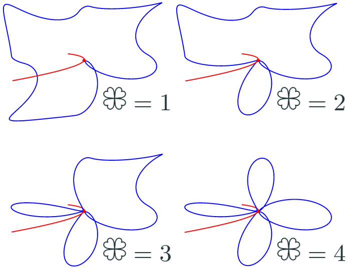


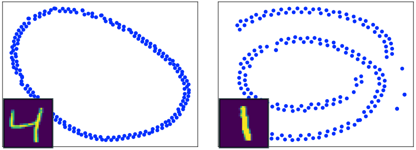
$$✤(\mathcal{M}) = \max_{\boldsymbol{x}\in\mathcal{M}} N_{\mathcal{M}}\left(\left\{\boldsymbol{x}' \;\middle|\; \begin{array}{l} d_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{x}') > \tau_1 \\ \angle(\boldsymbol{x},\boldsymbol{x}') < \tau_2 \end{array}\right\}, \frac{1}{\sqrt{1+\kappa^2}}\right)$$

Here, $N_{\mathcal{M}}(T,\delta)$ is the covering number of $T \subseteq \mathcal{M}$ by δ balls in $d_{\mathcal{M}}$.

Intuition: Number of times that $\mathcal{M}$ loops back on itself.

⌘ number: How “loopy” is $\mathcal{M}$?



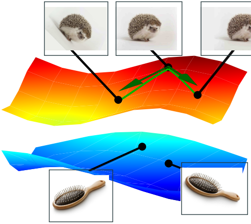


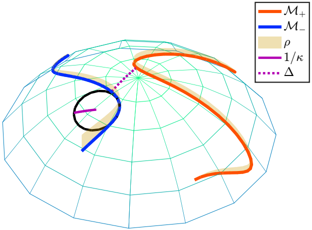
$$\text{⌘}(\mathcal{M}) = \max_{\boldsymbol{x}\in\mathcal{M}} N_{\mathcal{M}}\left(\left\{\boldsymbol{x}' \;\middle|\; \begin{array}{l} d_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{x}') > \tau_1 \\ \angle(\boldsymbol{x},\boldsymbol{x}') < \tau_2 \end{array}\right\}, \frac{1}{\sqrt{1+\kappa^2}}\right)$$

Here, $N_{\mathcal{M}}(T,\delta)$ is the covering number of $T \subseteq \mathcal{M}$ by δ balls in $d_{\mathcal{M}}$.

Intuition: Number of times that $\mathcal{M}$ loops back on itself.

The Multiple Manifold Problem



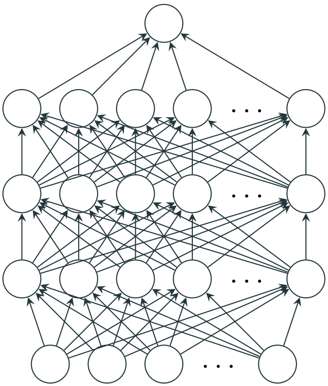


Problem: Given labeled data samples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ lying on manifolds $\mathcal{M}_\pm \subset \mathbb{S}^{n_0-1}$, learn a classifier $f_{\boldsymbol{\theta}}$ that correctly labels *every* point on the two manifolds:

$$\mathrm{sign}(f_{\boldsymbol{\theta}}(\boldsymbol{x})) = \sigma, \text{ for all } \boldsymbol{x} \in \mathcal{M}_\sigma.$$

Network Setup

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



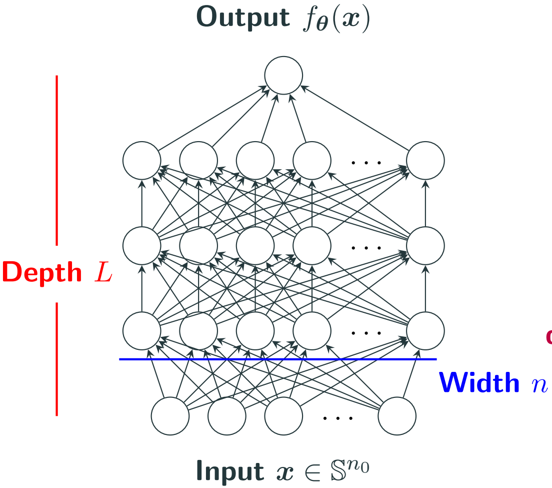
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Fully connected
ReLU network

Weights initialized
iid $\mathcal{N}(0, \frac{2}{n})$

Trained on N iid
data samples $(\boldsymbol{x}_i, y_i)$

Network Setup – Resources

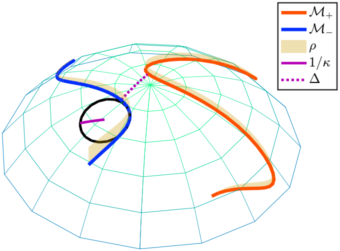


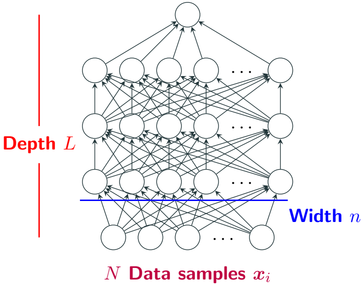
Fully connected
ReLU network

Weights initialized
iid $\mathcal{N}(0, \frac{2}{n})$

Trained on N iid
data samples $(\boldsymbol{x}_i, y_i)$

Multiple Manifold Problem





Theory question: how should **resources** (**depth** L, **width** n, **# samples** N) depend on **geometry** (dimension d, curvature κ, separation Δ, clover number ✿)?

Training?

Objective: Square Loss on Training Data

$$\min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}) \equiv \frac{1}{2} \int_{\boldsymbol{x}} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y(\boldsymbol{x})\right)^2 d\mu_N(\boldsymbol{x}).$$

Does gradient descent correctly label the manifolds?

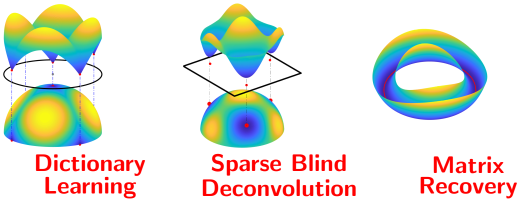
Training?

Objective: Square Loss on Training Data

$$\min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}) \equiv \frac{1}{2} \int_{\boldsymbol{x}} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y(\boldsymbol{x})\right)^2 d\mu_N(\boldsymbol{x}).$$

Does gradient descent correctly label the manifolds?

One Approach: Geometry (from symmetry!) in **parameter space**:



See, e.g., [Sun, Qu, W. '18], [Zhang, Kuo, W. 19], survey [Zhang, Qu, W. 20].

Training?

Objective: Square Loss on Training Data

$$\min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}) \equiv \frac{1}{2} \int_{\boldsymbol{x}} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y(\boldsymbol{x})\right)^2 d\mu_N(\boldsymbol{x}).$$

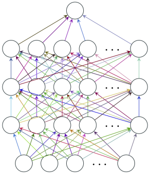
Does gradient descent correctly label the manifolds?

Today's talk: Dynamics in **input-output space**:

Neural Tangent Kernel

$$\Theta(\boldsymbol{x}, \boldsymbol{x}') = \left\langle \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle$$

Measures ease of independently adjusting $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, $f_{\boldsymbol{\theta}}(\boldsymbol{x}')$



Follows [Jacot et. al.], many recent works.

Certificates for Training?

Objective: Square Loss on Training Data

$$\min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}) \equiv \frac{1}{2} \int_{\boldsymbol{x}} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y(\boldsymbol{x})\right)^2 d\mu_N(\boldsymbol{x}).$$

Signed error: $\zeta(\boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y(\boldsymbol{x})$.

Gradient flow: $\dot{\boldsymbol{\theta}}_t = -\nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}_t) = -\int_{\boldsymbol{x}} \frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}(\boldsymbol{x}) \zeta_t(\boldsymbol{x}) d\mu_N(\boldsymbol{x})$.

Certificates for Training?

The error evolves according to the NTK:

$$\begin{aligned}
\dot{\zeta}_t(\boldsymbol{x}) &= \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}}^* \dot{\boldsymbol{\theta}}_t = -\frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}}^* \int_{\boldsymbol{x}'} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}')}{\partial \boldsymbol{\theta}} \zeta_t(\boldsymbol{x}') d\mu_N(\boldsymbol{x}') \\
&= -\int_{\boldsymbol{x}'} \left\langle \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle \zeta(\boldsymbol{x}') d\mu_N(\boldsymbol{x}') \\
&= -\int_{\boldsymbol{x}'} \Theta(\boldsymbol{x}, \boldsymbol{x}') \zeta_t(\boldsymbol{x}') d\mu_N(\boldsymbol{x}') \\
&= -\boldsymbol{\Theta}[\zeta_t](\boldsymbol{x}).
\end{aligned}$$

Fast decay if ζ_t is aligned with lead eigenvectors of $\boldsymbol{\Theta}$.

Certificates for Training?

Objective: Square Loss on Training Data

$$\min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}) \equiv \frac{1}{2} \int_{\boldsymbol{x}} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y(\boldsymbol{x})\right)^2 d\mu_N(\boldsymbol{x}).$$

Signed error: $\zeta(\boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y(\boldsymbol{x})$.

Gradient Method (GD): $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \tau \nabla \varphi(\boldsymbol{\theta}_k)$.

Similar intuition to gradient flow.

We analyze GD with (small) nonzero τ.

Dynamics by Certificates

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

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - f_\star(\boldsymbol{x}) \underset{\text{square}}{\overset{\text{mean}}{\approx}} \int_{\mathcal{M}} \Theta(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') d\mu(\boldsymbol{x}')$$

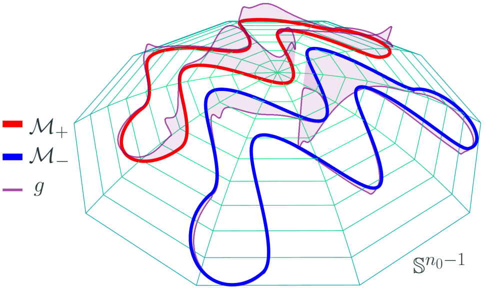
and $\int_{\mathcal{M}} \left(g(\boldsymbol{x}')\right)^2 d\mu(\boldsymbol{x}')$ is small.

Dynamics by Certificates

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

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - f_\star(\boldsymbol{x}) \overset{\text{mean}}{\underset{\text{square}}{\approx}} \int_{\mathcal{M}} \Theta(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') d\mu(\boldsymbol{x}')$$

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

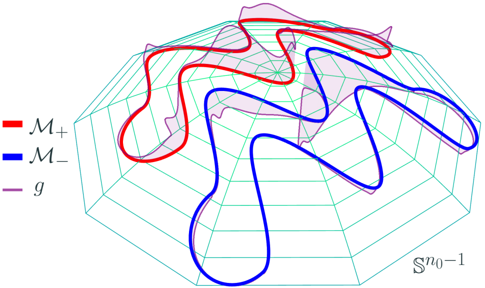


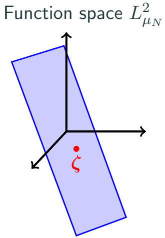
Dynamics by Certificates

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

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - f_\star(\boldsymbol{x}) \overset{\text{mean}}{\underset{\text{square}}{\approx}} \int_{\mathcal{M}} \Theta(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') d\mu(\boldsymbol{x}')$$

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





Error ζ near stable range of *random operator* $\boldsymbol{\Theta}$

Dynamics by Certificates

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

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - f_\star(\boldsymbol{x}) \overset{\text{mean}}{\underset{\text{square}}{\approx}} \int_{\mathcal{M}} \Theta(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') d\mu(\boldsymbol{x}')$$

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

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

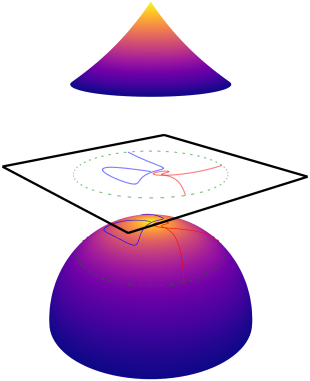
$$L \geq \mathrm{poly}(\kappa, \log n_0, C_\rho, C_{\mathcal{M}}),$$
$$n \geq \mathrm{poly}(L),$$
$$N \geq \mathrm{poly}(L),$$

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

Resource Tradeoffs I: Depth as an Approximation Resource

Increasing depth L sharpens the NTK Θ

$\Rightarrow$ deeper nets fit more complicated geometries

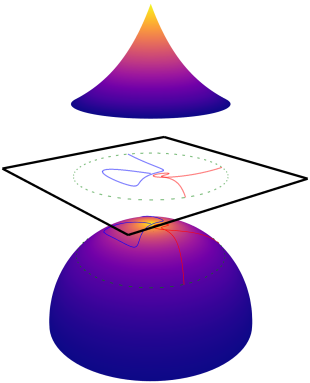


Depth $L = 5$

Resource Tradeoffs I: Depth as an Approximation Resource

Increasing depth L
sharpens the NTK Θ

$\Rightarrow$ deeper nets fit
more complicated geometries

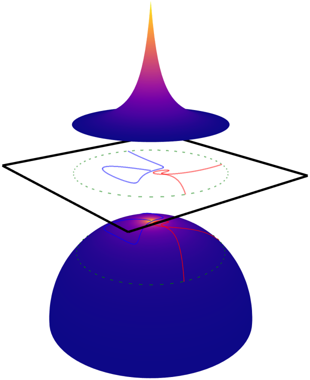


Depth $L = 10$

Resource Tradeoffs I: Depth as an Approximation Resource

Increasing depth L
sharpens the NTK Θ

$\Rightarrow$ deeper nets fit
more complicated geometries

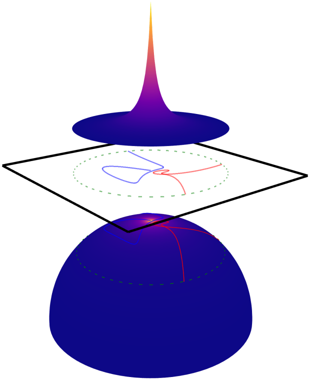


Depth $L = 50$

Resource Tradeoffs I: Depth as an Approximation Resource

Increasing depth L
sharpens the NTK Θ

$\Rightarrow$ deeper nets fit
more complicated geometries

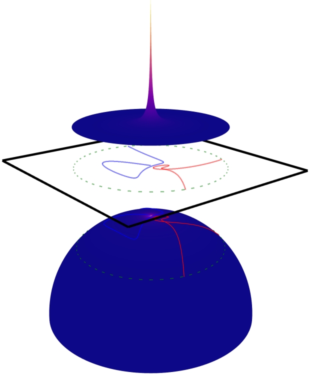


Depth $L = 100$

Resource Tradeoffs I: Depth as an Approximation Resource

Increasing depth L sharpens the NTK Θ

$\Rightarrow$ deeper nets fit more complicated geometries



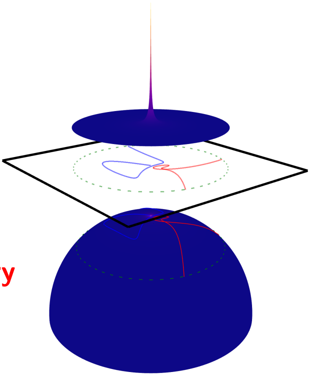
Depth $L = 500$

Resource Tradeoffs I: Depth as an Approximation Resource

Increasing depth L
sharpens the NTK Θ

$\Rightarrow$ deeper nets fit
more complicated geometries

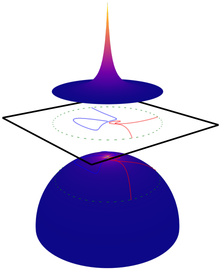
$\Rightarrow$ Set depth L based on geometry



Depth $L = 1,000$

Resource Tradeoffs I: Certificates from Depth

Certificate Problem: $\exists\, g$ small s.t. $\boldsymbol{\Theta} g \approx \zeta$?



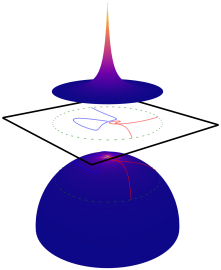
$\boldsymbol{\Theta} \approx \boldsymbol{\Theta}_{\text{near}} + \boldsymbol{\Theta}_{\text{far}} + \boldsymbol{\Theta}_{\text{✿}}$

$\boldsymbol{\Theta}_{\text{near}}$: $\angle$, $d_{\mathcal{M}}$ small wrt κ, Δ
$\approx$ invariant operator M, use Fourier analysis

$\boldsymbol{\Theta}_{\text{far}}$: $\angle$, $d_{\mathcal{M}}$ big, $\boldsymbol{\Theta}_{\text{✿}}$: $\angle$ small, $d_{\mathcal{M}}$ big
Worst-case contributions from these components

Resource Tradeoffs I: Certificates from Depth

Certificate Problem: $\exists\ g$ small s.t. $\boldsymbol{\Theta} g \approx \boldsymbol{\zeta}$?



$\boldsymbol{\Theta} \approx \boldsymbol{\Theta}_{\mathrm{near}} + \boldsymbol{\Theta}_{\mathrm{far}} + \boldsymbol{\Theta}_{✾}$

$\boldsymbol{\Theta}_{\mathrm{near}}$: $\angle$, $d_{\mathcal{M}}$ small wrt κ, Δ

$\approx$ invariant operator M, use Fourier analysis

$\boldsymbol{\Theta}_{\mathrm{far}}$: $\angle$, $d_{\mathcal{M}}$ big, $\boldsymbol{\Theta}_{✾}$: $\angle$ small, $d_{\mathcal{M}}$ big

Worst-case contributions from these components

$$g = \sum_{\ell=0}^{\infty} (-1)^{\ell} \left(\left(\boldsymbol{P}_S \boldsymbol{M} \boldsymbol{P}_S\right)^{-1} \boldsymbol{P}_S \left(\boldsymbol{\Theta} - \boldsymbol{M}\right) \boldsymbol{P}_S \right)^{\ell} \left(\boldsymbol{P}_S \boldsymbol{M} \boldsymbol{P}_S\right)^{-1} \zeta.$$

Resource Tradeoffs I: Certificates from Depth

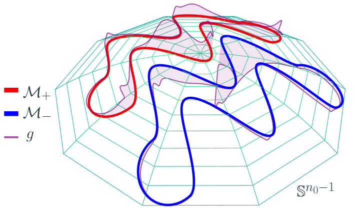
Theorem [Wang, Buchanan, Gilboa, W. '21]. Suppose

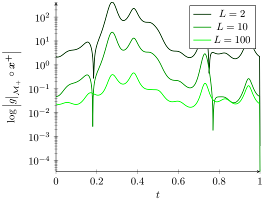
$$L \geq \max\left\{ \left(\frac{1}{\Delta\sqrt{1+\kappa^2}}\right)^{C\text{\ding{96}}(\mathcal{M})}, \right.$$
$$\mathrm{poly}(M_2 \dots M_7, \Delta^{-1}, \rho_{\max}),$$
$$\left. \exp\Big(C'\kappa \mathrm{len}(\mathcal{M})\Big) \right\}.$$

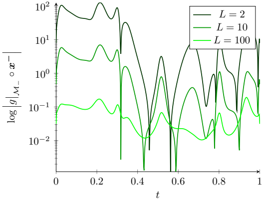
Then there exists a certificate g satisfying

$$\|\mathbf{\Theta}[g] - \zeta\|_{L^2_\mu} \leq \|\zeta\|_{L^\infty} L^{-1},$$
$$\|g\|_{L^2_\mu} \leq \frac{C''\|\zeta\|_{L^2_\mu}}{\rho_{\min} n \log L}.$$

Resource Tradeoffs I: Certificates from Depth







Depth as a fitting resource: Larger L leads to a sharper kernel Θ and a smaller certificate $g \implies$ easier fitting

Dynamics by Certificates

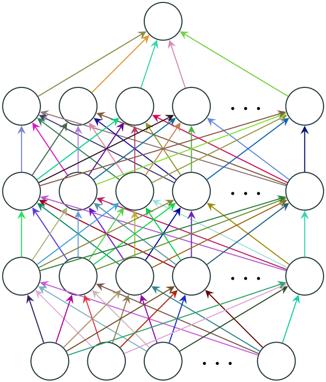
Theorem. If *a certificate exists*, if $\tau \asymp 1/(nL)$, and if

$$
\begin{aligned}
L &\gtrsim \mathrm{poly}(\kappa, \log n_0, C_\rho, C_{\mathcal{M}}), \\
n &\gtrsim \mathrm{poly}(L), \\
N &\geq \mathrm{poly}(L),
\end{aligned}
$$

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

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$

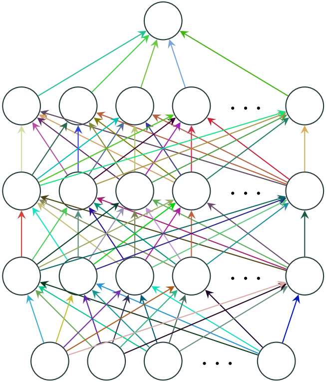


Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

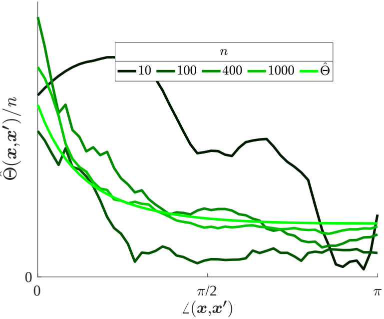
As n increases, $\Theta(\boldsymbol{x}, \boldsymbol{x}')$ concentrates about $\mathbb{E}_{\text{init weights}}[\Theta(\boldsymbol{x}, \boldsymbol{x}')]$

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



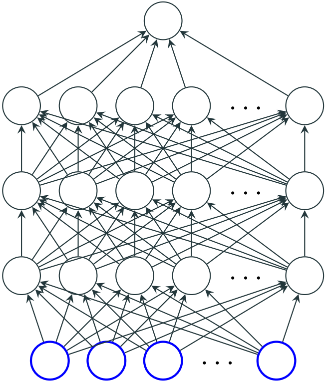
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$



As n increases, $\Theta(\boldsymbol{x}, \boldsymbol{x}')$ concentrates about $\mathbb{E}_{\text{init weights}}[\Theta(\boldsymbol{x}, \boldsymbol{x}')]$

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



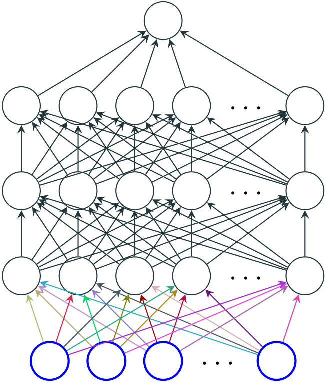
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



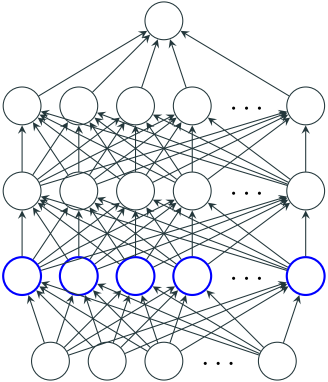
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



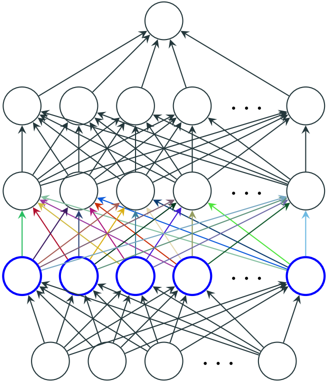
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



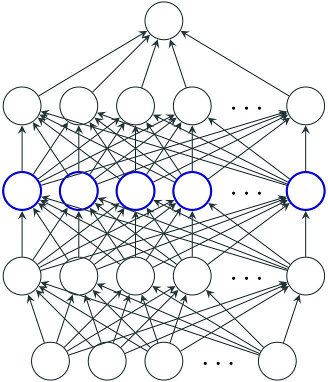
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



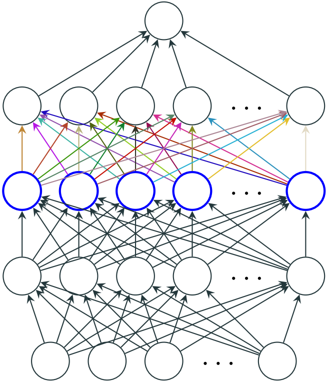
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



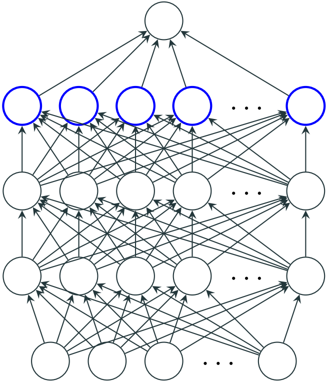
Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$

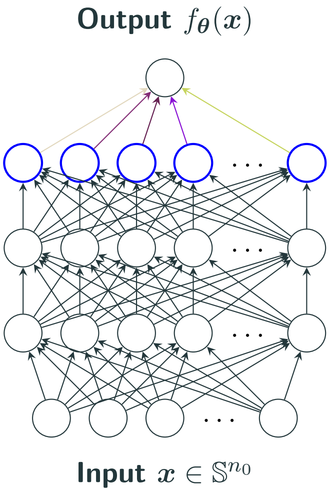


Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

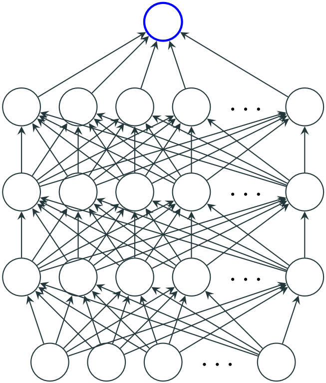


Sequential structure

$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Output $f_{\boldsymbol{\theta}}(\boldsymbol{x})$



Input $\boldsymbol{x} \in \mathbb{S}^{n_0}$

Sequential structure

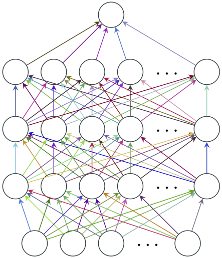
$\Rightarrow$ martingale tools work well

Resource Tradeoffs II: Width as a Statistical Resource

Proposition. Suppose that $n > L\mathrm{polylog}(Ln_0)$. Then

$$\left| \Theta(\boldsymbol{x}, \boldsymbol{x}') - \frac{n}{2} \sum_{\ell} \cos(\varphi^{\ell} \nu) \prod_{\ell'=\ell}^{L-1} \left(1 - \frac{\varphi^{\ell'} \nu}{\pi} \right) \right|$$

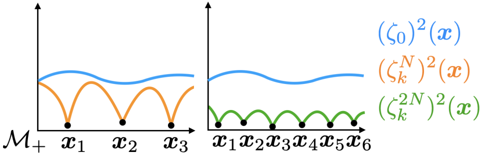
is small (simultaneously) for all $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{M} \times \mathcal{M}$.

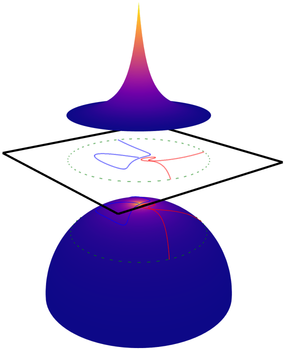


$\Rightarrow$ **set width n based on L**

and implicitly based on κ, Δ

Resource Tradeoffs III: Data as a Statistical Resource





Depth $L = 50$

$\Rightarrow$ Sample complexity N is dictated by kernel "aperature", which depends on geometry (κ, Δ) via L

End-to-End Guarantee

Combining the two results, when ...

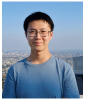
$$
\begin{aligned}
L &\geq \max\left\{ \left(\frac{1}{\Delta\sqrt{1+\kappa^2}}\right)^{C^{\text{✿}}(\mathcal{M})}, \exp\Big(C'\kappa \operatorname{len}(\mathcal{M})\Big) \right. \\
&\qquad\qquad \left. \operatorname{poly}(M_2 \ldots M_7, \Delta^{-1}, \kappa, \rho_{\min}^{-1}, \rho_{\max}) \right\}, \\
n &\geq \operatorname{poly}(L), \\
N &\geq \operatorname{poly}(L),
\end{aligned}
$$

gradient descent correctly classifies every point on $\mathcal{M}$.

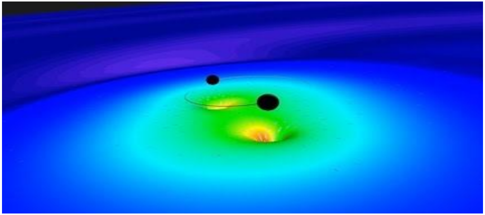
Novelty: end-to-end guarantee of generalization, depending only on the geometry of the data.

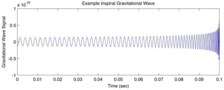
Gravitational Wave Astronomy [with Marka, Marka, Yan, Colgan]

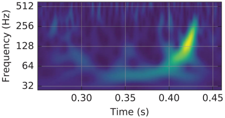




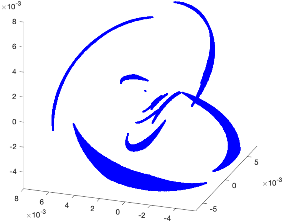

One binary black hole merger:



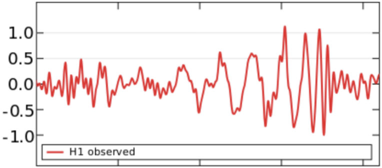


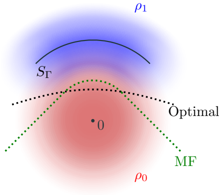


Many mergers
(varying mass M_1, M_2):
$\Longrightarrow$ **low-dim manifold**



Parametric Detection?

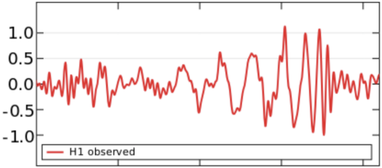


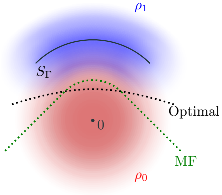


Is observation $x = s_\gamma + z$ or $x = z$?

$\Longrightarrow$ **two (noisy) manifolds!**

Parametric Detection?



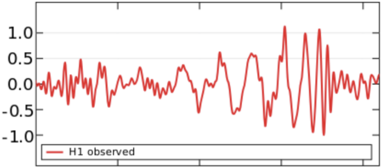


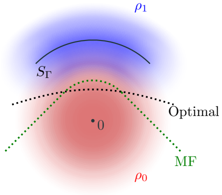
Is observation $\boldsymbol{x} = \boldsymbol{s}_\gamma + \boldsymbol{z}$ or $\boldsymbol{x} = \boldsymbol{z}$?

$\Longrightarrow$ **two (noisy) manifolds!**

Classical approach: template matching $\max_{\boldsymbol{\gamma}} \langle \boldsymbol{a}_{\boldsymbol{\gamma}}, \boldsymbol{x} \rangle > \tau$?

Parametric Detection?





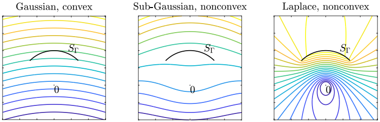
Is observation $\boldsymbol{x} = \boldsymbol{s}_\gamma + \boldsymbol{z}$ or $\boldsymbol{x} = \boldsymbol{z}$?

$\Longrightarrow$ **two (noisy) manifolds!**

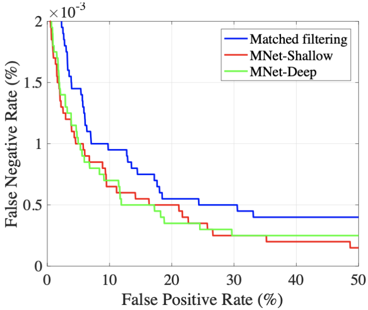
Classical approach: template matching $\max_{\boldsymbol{\gamma}} \langle \boldsymbol{a}_{\boldsymbol{\gamma}}, \boldsymbol{x} \rangle > \tau$?

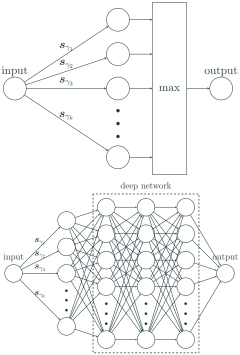
Issues: Optimality? Complexity?

Unknown unknowns? Unknown noise?



Neural Nets – Performance Improvements?





Dedicated constructions of shallow and deep networks, based on equivalence between template matching and (particular) deep models [Yan, Avagyan, Colgan, Veske, Bartos, W., Marka, Marka '21].

Conclusion and Future Directions

End-to-end analysis of learning with data on curves.

More Complicated Geometries

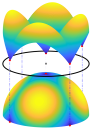
Can still use sharpness of Θ.

Network Structures from Geometry

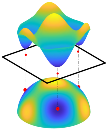
Guaranteed Invariance

Beyond Linearization / Neural Tangent Kernel

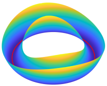
Insights from simpler dictionary / feature learning problems?



Dictionary Learning



Sparse Blind Deconvolution

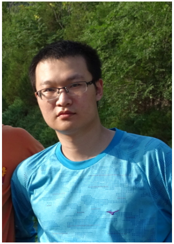


Matrix Recovery

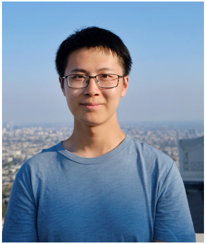
Thanks to ...

Sam Buchanan

Dar Gilboa



Tingran Wang



Jingkai Yan

Deep Networks and the Multiple Manifold Problem, Buchanan, Gilboa, W. '21

Deep Networks Provably Classify Data on Curves

Wang, Buchanan, Gilboa, W. '21

Generalized Approach to Matched Filtering using Neural Networks

Yan, Avagyan, Colgan, Veske, Bartos, W., Marka, Marka, '21