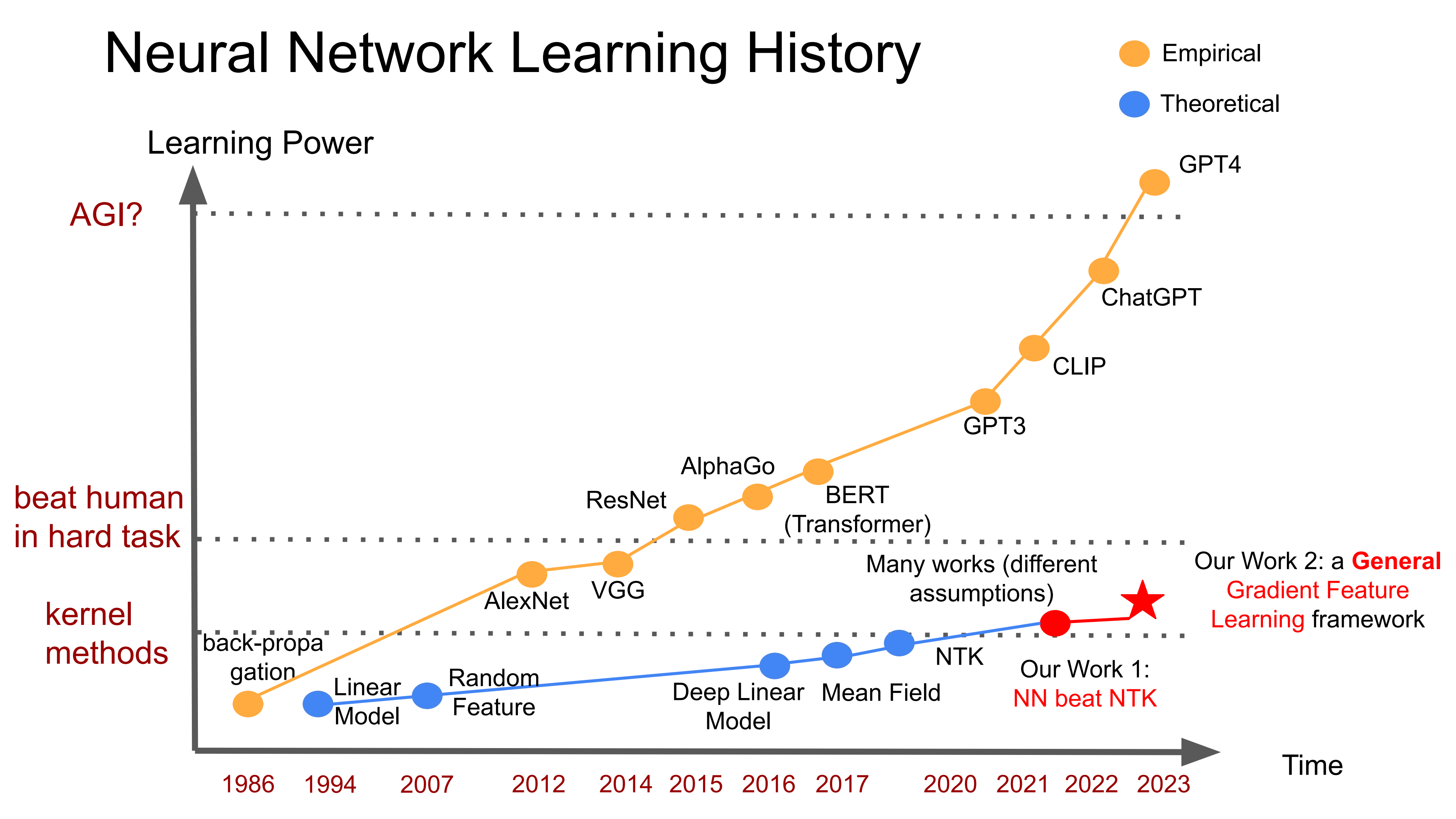
Provable Guarantees for Neural Networks via **Gradient Feature Learning**

Zhenmei Shi*, Junyi Wei*, Yingyu Liang

UW-Madison

NeurIPS 2023



Neural Network Learning History
Empirical
Theoretical
Learning Power
AGI?
beat human in hard task
kernel methods
GPT4
ChatGPT
CLIP
GPT3
AlphaGo
BERT (Transformer)
ResNet
VGG
AlexNet
back-propagation
Linear Model
Random Feature
Deep Linear Model
Mean Field
NTK
Many works (different assumptions)
Our Work 1: NN beat NTK
Our Work 2: a General Gradient Feature Learning framework
Time
1986
1994
2007
2012
2014
2015
2016
2017
2020
2021
2022
2023

Recent Work on Optimization of NNs

Why neural networks success?

- Neural Tangent Kernel (NTK regime or lazy learning regime):
  - Key idea: with heavy overpara, in a close neighborhood of random init, the optimization is almost convex
  - Limitation: impractical overpara; generalization only for simple datasets; approximately a kernel method (fixed feature, no feature learning)
- Feature learning beyond NTK
  - Key idea: on data with specific structure, the training algo exploits the structure to learn specific features as the neuron weights
  - Can handle problems that cannot be handled by fixed feature methods
  - Limitation: specific data (mixture of Gaussians, parity etc)

Recent Work on Optimization of NNs

Why neural networks success?

- Neural Tangent Kernel (NTK regime or lazy learning regime):
  - Key idea: with heavy overpara, in a close neighborhood of random init, the optimization is almost convex
  - Limitation: impractical overpara, generalization only for simple datasets; approximately a kernel method (fixed feature, no feature learning)
- Feature learning beyond NTK
  - Key idea: on data with specific structure, the training algo exploits the structure to learn specific features as the neuron weights
  - Can handle problems that cannot be handled by fixed feature methods
  - Limitation: specific data (mixture of Gaussians, parity etc)

Can we provide a more general analysis framework?

1. for more general data

2. Pin down the principle of feature learning in practical algo

Yes for two-layer networks

Intuition

Gradient captures useful features of data

These features as neuron weights can lead to good performance

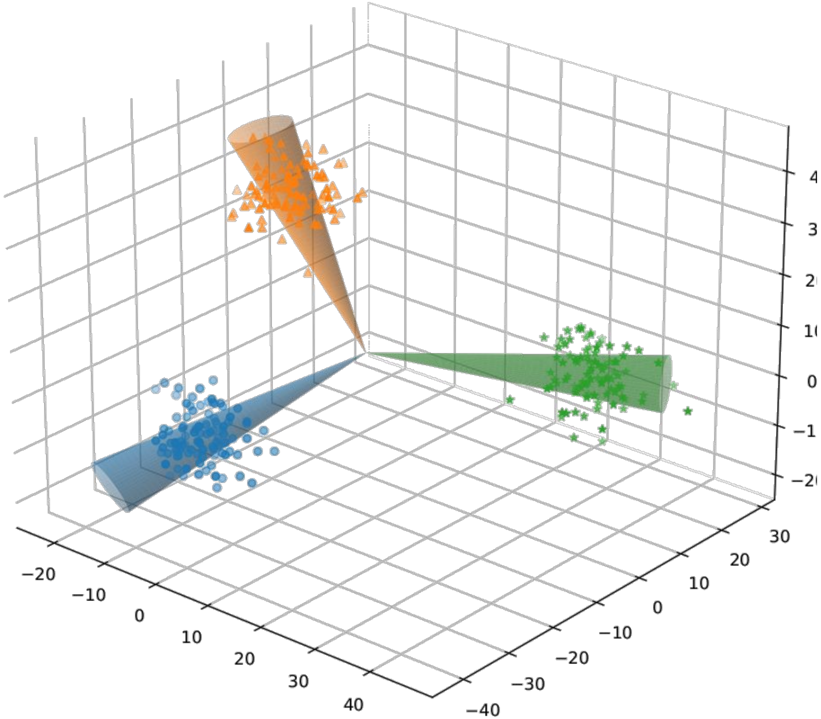
(c) Top Eigenvector of Feature Matrices on CelebA Prediction Tasks

	Lipstick	Eyebrows	5 o'clock shadow	Necktie	Smiling	Rosy Cheeks
Deep Network Feature Matrix: $W_1^T W_1$						
RFM Feature Matrix: $\sum_{i=1}^{n} \nabla f(x_i) \nabla f(x_i)^T$						
Correlation	0.999	0.999	0.999	0.999	0.999	0.999
Deep Network Test Acc.	90.53%	75.71%	85.88%	88.77%	89.83%	87.22%
RFM-T Test Acc.	91.62%	78.11%	88.18%	90.39%	91.24%	88.72%

Visualization of Gradient Feature vs Weight Feature. From: *Feature learning in neural networks and kernel machines that recursively learn features (2022).*

Gradient Features

- Gradient Features as neuron weights can lead to good performance



Gradient Feature Induced Networks

- Gradient Features as neuron weights can lead to good performance
- Induced networks: networks using gradient features as neuron weights
- Optimal approximation using induced networks:

Definition (Optimal Approximation via Gradient Features)

The Optimal Approximation network and loss using gradient feature induced networks $\mathcal{F}_{d,r,B_F,S}$ are defined as:

$$f^* := \text{argmin}_{f \in \mathcal{F}_{d,r,B_F,S}} L_{\mathcal{D}}(f), \qquad \text{OPT}_{d,r,B_F,S} := \min_{f \in \mathcal{F}_{d,r,B_F,S}} L_{\mathcal{D}}(f). \tag{6}$$

Unified Analysis Framework for Two-Layer NNs

Main Theorem (informal)
Two-layer networks can in poly-time w.h.p. achieve test loss close to the optimal approximation loss by Gradient Feature Induced Networks.

General framework that

1. **captures the feature learning from gradients (any data), and**
2. **gives poly error bounds for prototypical problems (beyond NTK)**

Applications of the Framework

- Case studies on prototypical problems:
  - Mixtures of Gaussians
  - Parity functions
  - Linear data
  - Multiple-index data models + polynomial labeling functions

- Explaining some intriguing empirical phenomena
  - Beyond the Kernel Regime
  - Simplicity Bias
  - Lottery Ticket Hypothesis
  - Learning over Different Data Distributions
  - New Perspectives about Roadmaps Forward