Sharp Guarantees for Learning Neural Networks with Gradient Methods

Hossein Taheri, Christos Thrampoulidis, Arya Mazumdar

Motivation

- Deep learning is transforming our lives.

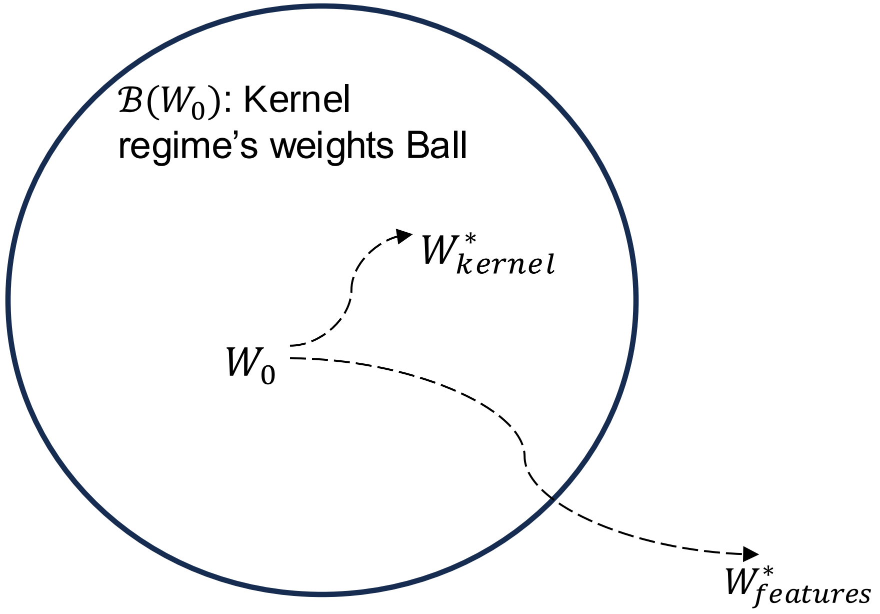
- Large models are ubiquitous in almost all applications.

- Despite their complexity, they also demonstrate good generalization performance.

Motivation: Kernel and Feature Learning

- Neural nets can operate in the kernel regime where the weights stay **close to initialization**.

- By using **large step sizes**, the network weights can move far from initialization, learn underlying features and achieve better test performance.

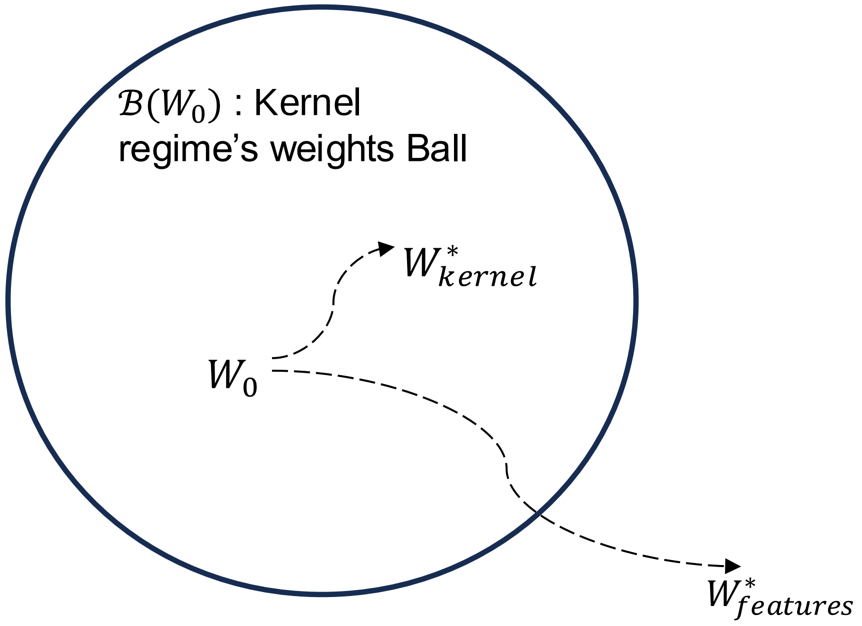


Motivation: Kernel and Feature Learning

- Q1: What is the performance of gradient descent for neural nets in the kernel and feature learning regimes?

- Q2: Can we provably show the benefits of the feature learning regime?

- Q3: How large can the radius of $\mathcal{B}(W_0)$ be?



Kernel regime's results

- If $||W^*_{kernel} - W_0|| \leq m^{O(L^{-1})}$, the network effectively operates in the kernel regime.

- Our analysis leads to better test loss bounds for learning under the kernel regime.

	Width	**Test Loss**
[Chen et al. 2021]	$\Omega(\mathrm{poly}(\frac{\log(n)}{\gamma}))$	$\frac{e^{O(L)}}{\gamma^2}\sqrt{\frac{m}{n}}$
Our result	$\Omega(\mathrm{poly}(\frac{\log(n)}{\gamma}))$	$\frac{e^{O(L)}}{\gamma^2 n}$

Table 1: Comparing our results for learning deep nets under kernel regime to previous results. Here m: width, L: depth, n: number of samples, γ: class margin.

Benefits of Feature learning

- SGD can learn the XOR distribution in both kernel and feature learning regimes.

$$x \in \{\pm 1\}^d, \qquad y = x_1 \cdot x_2$$

- We can study the performance of neural networks in learning the XOR distribution in both regimes.

Benefits of Feature learning

- **Theorem** (informal): A two-layer network of constant width m can achieve zero test error on the XOR problem with $n = \tilde{O}(d)$ samples after $\log(d)$ SGD iterations with step-size $\eta = m$.

	Width	**Iteration**	**Sample**
Kernel regime's result	$\Omega(\mathrm{poly}(d))$	d^2	d^2
Feature learning's results	$\Omega(1)$	$\log(d)$	d

Table 2: Comparison of our results in learning the d-dimensional XOR distribution with large step-size to kernel regime's results.

Experiments on XOR data

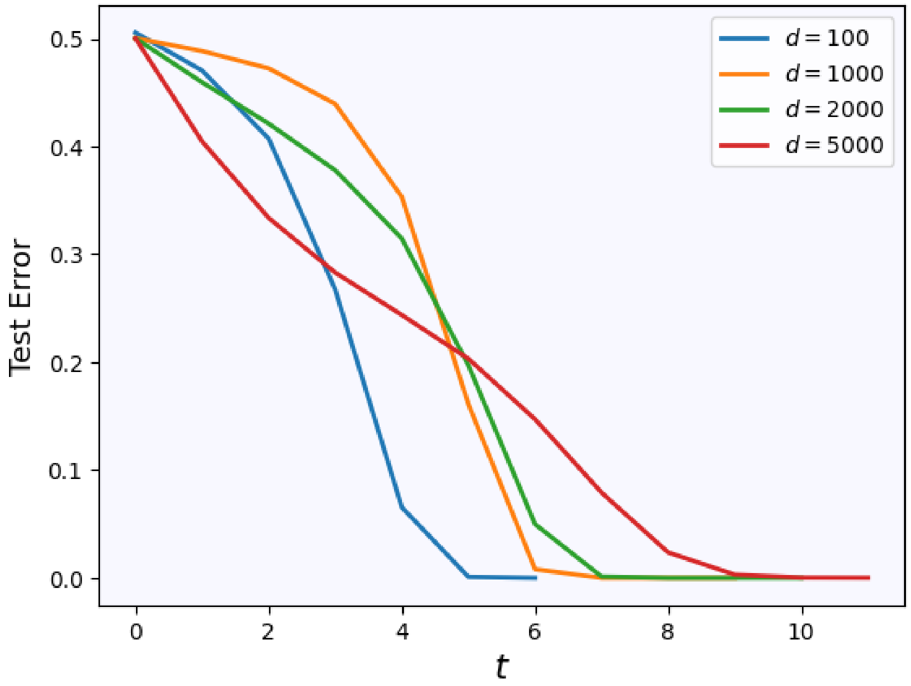


Fig1: Test error vs iteration number for SGD learning of the XOR distribution for different data dimensions. Here $\eta = m = 20$ and $n = 6d$.