# CS7020-Advances in theory of Deep Learning

## Kernel and Rich regimes in Overparameterized models

Shivaram
BE20B032

# Content

# Introduction

- *Kernel Regime*
  - Learned function is a minimum RKHS norm solution.
  - Inherit inductive bias and generalization of RKHS
- *Rich Regime*
  - Other works suggest inductive bias cannot be represented by RKHS.
  - Example: Infinite width ReLU network with infinitesimal weight decay.
- Chizat et al, suggest:
  - Scale of model at initialization controls.
  - Homogenous model with scale at initialization going to infinity.

# Contributions

---

- Implicit bias transition from l2 norm at to $\alpha \to \infty$ l1 in $\alpha \to 0$
- Small initialization good for generalization
- Shape of initialization affects $\alpha \to \infty$ bias and not $\alpha \to 0$
- Depth hastens the transition
- “Width” has an interesting role in controlling transition.

# Setup and Preliminaries

- Consider the models $f : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$, map $\mathbf{w} \in \mathbb{R}^p$ parameters, $\mathbf{x} \in \mathcal{X}$ samples to predictions $f(\mathbf{w}, \mathbf{x}) \in \mathbb{R}$.
- Predictor $F(\mathbf{w}) \in \{f : \mathcal{X} \to \mathbb{R}\}$, linear functional, therefore $f(\mathbf{w}, \mathbf{x}) = \langle \boldsymbol{\beta}_{\mathbf{w}}, \mathbf{x} \rangle$
- D-positive homogeneous models which implies $c \in \mathbb{R}_+, F(c \cdot \mathbf{w}) = c^D F(\mathbf{w})$ Examples: Multi layer ReLU networks, convolutional networks, etc.
- Loss function: $L(\mathbf{w}) = \tilde{L}(F(\mathbf{w})) = \sum_{n=1}^{N} (f(\mathbf{w}, \mathbf{x}_n) - y_n)^2$
- Gradient dynamics: $\dot{\mathbf{w}}(t) = -\nabla L(\mathbf{w}(t))$
- Consider Overparameterized models $N \ll p$

# Kernel Regime

---

- Gradient Descent: $f(\mathbf{w}, \mathbf{x}) = f(\mathbf{w}(t), x) + \langle \mathbf{w} - \mathbf{w}(t), \nabla_{\mathbf{w}} f(\mathbf{w}(t), \mathbf{x}) \rangle + O(\|\mathbf{w} - \mathbf{w}(t)\|^2)$
- Affine model with feature map corresponding to Tangent kernel, $K_{\mathbf{w}(t)}(\mathbf{x}, \mathbf{x}') = \langle \nabla_{\mathbf{w}} f(\mathbf{w}(t), \mathbf{x}), \nabla_{\mathbf{w}} f(\mathbf{w}(t), \mathbf{x}') \rangle$
- Tangent kernel does not change over the course.
- Minimizing the loss using gradient descent, reaches the minimum RKHS norm solution.
- $\alpha \to \infty$ reaches “Kernel Regime”.
- $\alpha \to 0$ leads to rich inductive bias, “Rich Regime”.

# Study of a Depth 2 model

- Consider the following linear functions with squared initialization: $f(\mathbf{w},\mathbf{x}) = \sum_{i=1}^{d} (\mathbf{w}_{+,i}^2 - \mathbf{w}_{-,i}^2)\mathbf{x}_i = \langle \boldsymbol{\beta}_{\mathbf{w}}, \mathbf{x} \rangle$, $\mathbf{w} = \left[\begin{smallmatrix}\mathbf{w}_+\\ \mathbf{w}_-\end{smallmatrix}\right] \in \mathbb{R}^{2d}$, and $\boldsymbol{\beta}_{\mathbf{w}} = \mathbf{w}_+^2 - \mathbf{w}_-^2$
- Diagonal linear neural network(diagonal weight matrices)
- Reason for *"unbiased model"* with 2 weights
  - Ensure the model is truly equivalent to standard linear regression.
  - Allows initialization $F(\alpha \mathbf{w}_0) = 0$
- $\boldsymbol{\beta}_{\alpha,\mathbf{w}_0}^{\infty}$ solution reached with initialization: $\mathbf{w}_+(0) = \mathbf{w}_-(0) = \alpha \mathbf{w}_0$
- Considering the special case $\mathbf{w}_0 = \mathbf{1}$ , tangent kernel at initialization $K_{\mathbf{w}(0)}(\mathbf{x}, \mathbf{x}') = 8\alpha^2 \langle \mathbf{x}, \mathbf{x}' \rangle$
- By Chizat et al, we have minimum l2 solution $\boldsymbol{\beta}_{\ell_2}^* := \arg\min_{X\boldsymbol{\beta}=y} \|\boldsymbol{\beta}\|_2$
- By Gunasekar et al, we have l1 minimization

$$\lim_{\alpha \to 0} \boldsymbol{\beta}_{\alpha,\mathbf{1}}^{\infty} = \boldsymbol{\beta}_{\ell_1}^* := \arg\min_{X\boldsymbol{\beta}=y} \|\boldsymbol{\beta}\|_1$$

# Theorem 1

**Theorem 1** (Special case: $\mathbf{w}_0 = \mathbf{1}$). *For any $0 < \alpha < \infty$, if the gradient flow solution $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}$ for the squared parameterization model in eq. (3) satisfies $X\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}} = \mathbf{y}$, then*

$$\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}} = \arg\min_{\boldsymbol{\beta}} Q_\alpha(\boldsymbol{\beta}) \ \ s.t.\ X\boldsymbol{\beta} = \mathbf{y}, \tag{4}$$

*where $Q_\alpha(\boldsymbol{\beta}) = \alpha^2 \sum_{i=1}^{d} q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2}\right)$ and $q(z) = \int_0^z \operatorname{arcsinh}\left(\frac{u}{2}\right) du = 2 - \sqrt{4+z^2} + z \operatorname{arcsinh}\left(\frac{z}{2}\right)$.*

- Function $Q_\alpha$ ,implicit regularizer biases towards to one of the zero error solution
- $\alpha \to \infty$, we have $\boldsymbol{\beta}_i/\alpha^2 \to 0$ ,thus we have $Q_\alpha(\boldsymbol{\beta}) \propto \sum_i \boldsymbol{\beta}_i^2$
- $\alpha \to 0$, $|\boldsymbol{\beta}_i/\alpha^2| \to \infty$,in this regime $Q_\alpha(\boldsymbol{\beta}) \propto \|\boldsymbol{\beta}\|_1 + O(1/\log(1/\alpha^2))$

# Theorem 2

**Theorem 2.** *For any* $0 < \epsilon < d$*, under the setting of Theorem 1 with* $\mathbf{w}_0 = \mathbf{1}$*,*

$$\alpha \leq \min\left\{ \left(2(1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1\right)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-d/(\epsilon\|\boldsymbol{\beta}^*_{\ell_1}\|_1)\right) \right\} \implies \|\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}\|_1 \leq (1+\epsilon)\,\|\boldsymbol{\beta}^*_{\ell_1}\|_1$$

$$\alpha \geq \sqrt{2(1+\epsilon)(1+2/\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2} \implies \|\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}\|_2^2 \leq (1+\epsilon)\,\|\boldsymbol{\beta}^*_{\ell_2}\|_2^2$$

- Asymmetry in reaching the regimes.
- Polynomially large $\alpha$ suffices to approximate $\beta^*_{\ell_2}$
- Exponentially small $\alpha$ required to approximate $\beta^*_{\ell_1}$
- Conducting experiments in rich regime might be infeasible due to computational reasons.

# Generalization

---

- Consider a sparse regression problem, $\mathbf{x}_1, \ldots, \mathbf{x}_N \sim \mathcal{N}(0, I)$
- In rich limit, $N = \Omega(r^* \log d)$ suffices for l1.
- In kernel limit, l2 requires $N = \Omega(d)$ samples
- Generalization improves with decrease in $\alpha$
- For optimization, w=0, saddle point. Time required to escape the vicinity of zero.
- Work on edge of rich limit, using the largest alpha that allows generalization.

# Shape of w0 and the Implicit Bias

**Theorem 1** (General case). *For any $0 < \alpha < \infty$ and $\mathbf{w}_0$ with no zero entries, if the gradient flow solution $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0}$ satisfies $X\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} = \mathbf{y}$, then*

$$\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} = \arg\min_{\boldsymbol{\beta}} Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) \;\; s.t.\; X\boldsymbol{\beta} = \mathbf{y}, \tag{5}$$

*where* $Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \sum_{i=1}^{d} \alpha^2 \mathbf{w}_{0,i}^2 q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2 \mathbf{w}_{0,i}^2}\right)$ *and* $q(z) = 2 - \sqrt{4+z^2} + z\,\mathrm{arcsinh}\left(\frac{z}{2}\right)$.

- For $\alpha \to \infty$ : $Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \sum_{i=1}^{d} \alpha^2 \mathbf{w}_{0,i}^2\, q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2 \mathbf{w}_{0,i}^2}\right) = \sum_{i=1}^{d} \frac{\boldsymbol{\beta}_i^2}{4\alpha^2 \mathbf{w}_{0,i}^2} + O(\alpha^{-6})$
- For $\alpha \to 0$ : $\frac{1}{\log(1/\alpha^2)} Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \frac{1}{\log(1/\alpha^2)} \sum_{i=1}^{d} \alpha^2 \mathbf{w}_{0,i}^2\, q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2 \mathbf{w}_{0,i}^2}\right) = \sum_{i=1}^{d} |\boldsymbol{\beta}_i| + O\big(1/\log(1/\alpha^2)\big)$
- 
- It affects implicit bias in kernel regime and not in rich regime.

# Explicit Regularization

$$\boldsymbol{\beta}^R_{\alpha,\mathbf{w}_0} := F\left(\arg\min_{\mathbf{w}} \|\mathbf{w} - \alpha\mathbf{w}_0\|_2^2 \text{ s.t. } L(\mathbf{w}) = 0\right) = \arg\min_{\boldsymbol{\beta}} R_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) \text{ s.t. } X\boldsymbol{\beta} = y$$

$$\text{where} \quad R_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \min_{\mathbf{w}} \|\mathbf{w} - \alpha\mathbf{w}_0\|_2^2 \text{ s.t. } F(\mathbf{w}) = \boldsymbol{\beta}.$$

- Implicit bias would be minimizing the Euclidean norm from initialization.
- For special case of w0=1, limiting behavior of the two approaches match.
- We have $R_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) = \sum_i r(\boldsymbol{\beta}_i/\alpha^2)$
- rz is algebraic and qz is transcendental
- Thus $Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \neq R_{\alpha,\mathbf{1}}(\boldsymbol{\beta})$
- Bias of gradient descent and transitions are complex than captured by distances in parameter space

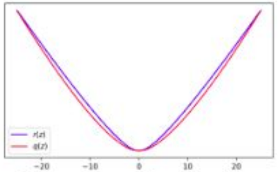


Figure 2: $q(z)$ and $r(z)$.

# Higher Order Models

- Consider: $F_D(\mathbf{w}) = \boldsymbol{\beta}_{\mathbf{w},D} = \mathbf{w}_+^D - \mathbf{w}_-^D$ and $f_D(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}_+^D - \mathbf{w}_-^D, \mathbf{x} \rangle$
- Effect of scale on the implicit bias:

**Theorem 3.** *For any $0 < \alpha < \infty$ and $D \geq 3$, if $X\boldsymbol{\beta}_{\alpha,D}^{\infty} = y$, then*

$$\boldsymbol{\beta}_{\alpha,D}^{\infty} = \arg\min_{\boldsymbol{\beta}} Q_\alpha^D(\boldsymbol{\beta}) \quad s.t. \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

*where $Q_\alpha^D(\boldsymbol{\beta}) = \alpha^D \sum_{i=1}^d q_D(\boldsymbol{\beta}_i/\alpha^D)$ and $q_D = \int h_D^{-1}$ is the antiderivative of the unique inverse of $h_D(z) = (1-z)^{-\frac{D}{D-2}} - (1+z)^{-\frac{D}{D-2}}$ on $[-1, 1]$. Furthermore, $\lim_{\alpha \to 0} \boldsymbol{\beta}_{\alpha,D}^{\infty} = \boldsymbol{\beta}_{\ell_1}^*$ and $\lim_{\alpha \to \infty} \boldsymbol{\beta}_{\alpha,D}^{\infty} = \boldsymbol{\beta}_{\ell_2}^*$.*

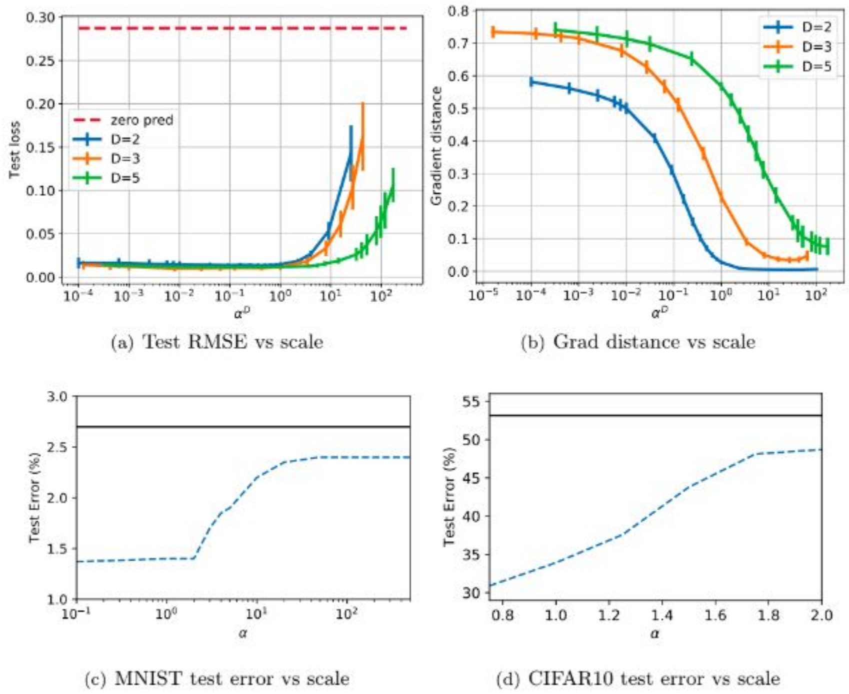
- Two extremes do not change, the intermediate regimes change, sharpness of transition.
- Increasing D hastens the transition to rich limit.

# Effect of Width

---

- Width plays an important role in entering the kernel regime.
- To avoid exploding outputs we used Chizat and Bach’s unbiasing trick.
- Consider, (asymmetric) matrix factorization model
  - $f((\mathbf{U}, \mathbf{V}), \mathbf{X}) = \langle \mathbf{U}\mathbf{V}^\top, \mathbf{X} \rangle$ where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times k}$
  - Linear predictor, $\mathbf{M}_{\mathbf{U},\mathbf{V}} = F(\mathbf{U}, \mathbf{V}) = \mathbf{U}\mathbf{V}^\top$
  - Scale of the model at initialization, $\sigma = \frac{1}{d} \|\mathbf{M}_{\mathbf{U},\mathbf{V}}\|_F$
- Wide factorizations can reach the kernel regime without “unbiasing”

# Experiments




(a) Test RMSE vs scale

(b) Grad distance vs scale

(c) MNIST test error vs scale

(d) CIFAR10 test error vs scale