Kernel and rich regimes in overparameterized models

Jinwon Park

May 20, 2021

Seoul National University

Table of Contents

Implicit bias in Over-Parametrized models

- In many machine learning problems, the model is highly overparametrized
  $\rightarrow$ many possible parameters for which the training loss is zero
- Training algorithm (e.g Gradient descent) can provide "implicit regularization" towards certain solutions over others
- The scale of the initialization α controls the transition between the kernel and rich regimes

The Scale of Initialization

- Model: $f : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$
  - Map parameters $w \in \mathbb{R}^p$ and examples $x \in \mathcal{X}$ to predictions $f(w, x) \in \mathbb{R}$
  - Much of focus will be on models that are linear in x (not in w)
  - D-homogeneous in w: $F(c \cdot w) = c^D F(w)$ for all $c > 0$
- Squared loss: $L(W) = \sum_{n=1}^{N} (f(w, x_n) - y_n)^2$
- Gradient descent/flow: $\dot{w}(t) = -\nabla L(w(t))$
- Scale of initialization: $\alpha \in (0, \infty)$, $w_{\alpha,w_0}(0) = \alpha w_0$

The Kernel Regime

- Gradient flow depends on first-order approximation w.r.t. w

$$f(w, x) = f(w(t), x) + \langle w - w(t), \nabla f(w(t), x) \rangle + O(\|w - w(t)\|^2)$$

- Gradient flow operates on model as if it were an affine model with feature map corresponding to tangent kernel
- Minimizing the loss of affine model reaches the solution nearest to the initialization where distance is measured w.r.t. the RKHS norm
- When does kernel regime happen?
  - "Width" $\to \infty$ leads to kernel regime
  - "Scale of initialization" $\to \infty$ leads to kernel regime

The Rich Regime

- Other studies have shown very different implicit biases
  - Matrix factorization with commutative measurements and $\alpha \to 0$ leads to implicit nuclear norm regularization
  - Deep linear convolutional networks $\to$ implicit $L_{2/depth}$ regularization in frequency domain
  - Infinite-width, depth-2 ReLU networks with infinitesimal weight decay $\to$ minimizes $\int |f``(w,x)| dx$ second order total variations
- These are not Hilbert norms, and cannot be captured by any kernel

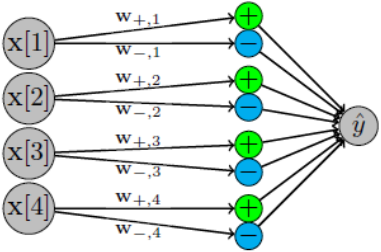
The Transition Regime

- Kernel regime: $\alpha \to \infty$
- Rich regime: $\alpha \to 0$
- Transition regime: finite α (i.e. the regime in which models are actually trained)

Table of Contents

Simple 2-Homogeneous Model

- Diagonal linear neural network
  - linear model with unusual parametrization
  - $f(w,x) = \sum_{i=1}^{d}(w_{+,i}^2 - w_{-,i}^2)x_i = \langle \beta_w, x \rangle$



- Trained with gradient flow to minimize sqare loss
  - $\beta_\alpha^\infty := \lim_{t\to\infty}(w_+^2(t) - w_-^2(t))$ when $w_+(0) = w_-(0) = \alpha w_0$

The Implicit Bias and the Scale of Initialization

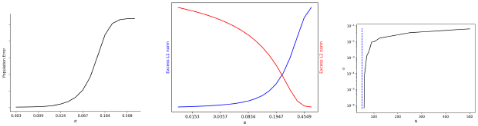
- $\lim_{\alpha \to \infty} \beta_\alpha^\infty = \text{argmin}_\beta \|\beta\|_2 \; s.t. \; L(\beta) = 0$
- $\lim_{\alpha \to 0} \beta_\alpha^\infty = \text{argmin}_\beta \|\beta\|_1 \; s.t. \; L(\beta) = 0$
- Theorem: for any $\alpha \in (0, \infty)$ if the gradient flow solution $\beta_{\alpha, w_0}^\infty$ satisfies $X\beta_{\alpha, w_0}^\infty = y$, then

$$\beta_{\alpha, w_0}^\infty = \underset{\beta}{\text{argmin}} \, Q_{\alpha, w_0}(\beta) \; s.t. \; L(\beta) = 0$$

where $Q_{\alpha, w_0}(\beta) = \sum_{i=1}^d \alpha^2 w_{0,i}^2 q\left(\frac{\beta_i}{\alpha^2 w_{0,i}^2}\right)$
and $q(z) = 2 - \sqrt{4 + z^2} + z \,\text{arcsinh}\left(\frac{z}{2}\right)$

Example: Sparse regression

- $y_n \sim N(\langle \beta^*, x_n \rangle, 0.01)$ for r^*-sparse β^* with non-zero entries
  - $N = \Omega(r^* \log d)$ samples suffice for $\beta^*_{\ell_1}$ to generalize well
  - $N = \Omega(d)$ samples needed for kernel regime solution $\beta^*_{\ell_2}$ to generalize



(a) Generalization (b) Norms of solution (c) Sample complexity

Figure 1: In (a) the population error of the gradient flow solution vs. α in the sparse regression problem described in Section 4. In (b), we plot $\|\beta^{\infty}_{\alpha,1}\|_1 - \|\beta^*_{\ell_1}\|_1$ in blue and $\|\beta^{\infty}_{\alpha,1}\|_2 - \|\beta^*_{\ell_2}\|_2$ in red vs. α. In (c), the largest α such that $\beta^{\infty}_{\alpha,1}$ achieves population error at most 0.025 is shown. The dashed line indicates the number of samples needed by $\beta^*_{\ell_1}$.

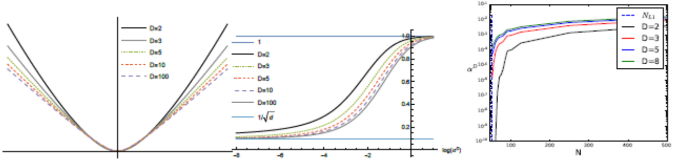
Table of Contents

D-Homogeneous Models

- $F_D(w) = \beta_{w,D} = W_+^D - W_-^D$ and $f_D(w, x) = \langle w_+^D - w_-^D, x\rangle$
- Theorem: For any $\alpha \in (0, \infty)$ and $D \geq 3$, if $X\beta^\infty_{\alpha,D}$, then

$$\beta^\infty_{\alpha,D} = \underset{\beta}{\operatorname{argmin}}\, Q_{\alpha,D}(\beta)\, s.t.\, L(\beta) = 0$$

where $Q^D_\alpha(\beta) = \alpha^D \sum_{i=1}^d q_D\left(\frac{\beta_i}{\alpha^D}\right)$
and $q_D = \int h_D^{-1}$ for $h_D(z) = (1-z)^{-\frac{D}{D-2}} - (1+Z)^{-\frac{D}{D-2}}$ /



(a) Regularizer (b) Approximation ratio (c) Sparse regression simulation

Figure 3: (a) $q_D(z)$ for several values of D. (b) The ratio $\frac{Q^D_\alpha(e_1)}{Q^D_\alpha(\mathbf{1}_d/\|\mathbf{1}_d\|_2)}$ as a function of α, where $e_1 = [1, 0, 0, \ldots, 0]$ is the first standard basis vector and $\mathbf{1}_d = [1, 1, \ldots, 1]$ is the all ones vector in $\mathbb{R}^d$. This captures the transition between approximating the ℓ_2 norm (where the ratio is 1) and the ℓ_1 norm (where the ratio is $1/\sqrt{d}$). (c) A sparse regression simulation as in Figure 1, using different order models. The y-axis is the largest α^D (the scale of $\boldsymbol{\beta}$ at initialization) that leads to recovery of the planted predictor to accuracy 0.025. The vertical dashed line indicates the number of samples needed in order for $\boldsymbol{\beta}^*_{\ell_1}$ to approximate the plant.

Q&A