High dimensional dynamics of neural network generalization error

Paper by Advani & Saxe, 2017

MIT 18.338

December 13, 2023

Introduction: the puzzle of deep learning

- A one-hidden layer neural network is just $y(x) = w^2 \cdot \sigma(w^1 \cdot x)$ for $w^1 \in \mathbb{R}^{N_i}, w^2 \in \mathbb{R}^{N_h}$.
- Puzzles in deep learning
  1. How neural networks can memorize noise as easily as signal
  2. Why they generalize despite this when there is signal to be learned, contradicting classical complexity-based generalization bounds
  3. Weird empirical behavior: double descent, etc

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

The role of random matrices

- We model the data matrix as random Gaussian in $\mathbb{R}^{N \times P}$ for N features and P data points with entries $X_{ij} \sim N(0, 1/N)$.
- This means the data covariance is $\Sigma = XX^T$.
- In what setting could RMT tell us about Σ? $P, N \to \infty$ with $P/N = \alpha \ll \infty$.
- Why are $P, N \to \infty$ and Gaussian justified as a modelling choice?

Roadmap

- Solving out dynamics of a toy linear student-teacher model
- Take-aways from toy model: the role of *eigengaps* and *frozen subspaces* during learning, as well as explaining double descent
- As time allows: the memorization puzzle (solving for train error) and deriving tighter classical Rademacher complexity-based generalization bounds

Toy model setup: baking our cake

- Student $\hat{y} = wX$ learning teacher $y = \bar{w}X + \epsilon$, where $\bar{w} \sim N(0, \sigma_w^2), w \sim N(0, (\sigma_w^0)^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$
- *Goal:* arrive at generalization dynamics

$$E_g(t) = \frac{1}{N} \sum_i \left[\left(\sigma_w^2 + \left(\sigma_w^0 \right)^2 \right) e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i} \left(1 - e^{-\frac{\lambda_i t}{\tau}} \right)^2 \right] + \sigma_\epsilon^2 \quad (1)$$

Step 1: Setting up gradient flow in data eigenbasis

- Begin with MSE

$$E_t(w(t)) = \frac{1}{P}\sum_{\mu=1}^{P} \|y^\mu - \hat{y}^\mu\|_2^2$$

- Write down gradient flow differential equation
- Uncouple w_i by changing variables to $w = zV^T$ for $\Sigma = V\Lambda V^T$

$$\tau \dot{z}(t) = \tilde{s} - z\Lambda \tag{2}$$

for $\tilde{s} = \bar{z}\Lambda + \epsilon\Lambda^{1/2}$.

Step 2: Solve Uncoupled System

- Easy! Notice because of our change of variables, (2) can be treated component-wise as

$$\tau \dot{z}_i = (\bar{z}_i - z_i)\,\lambda_i + \epsilon_i \sqrt{\lambda_i}, \quad i = 1, \cdots, N \tag{3}$$

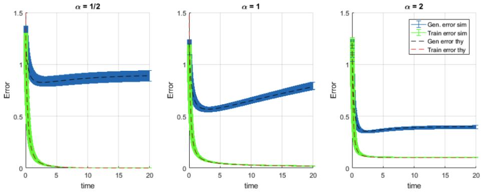
- Which has solution

$$\bar{z}_i - z_i = (\bar{z}_i - z_i(0))\, e^{-\frac{\lambda_i t}{\tau}} - \frac{\tilde{\epsilon}_i}{\sqrt{\lambda_i}} \left(1 - e^{-\frac{\lambda_i t}{\tau}}\right) \tag{4}$$

Step 3: Plug solution into E_g and average over weights

- Observe that

$$E_g(t) = \langle [y(x) - \hat{y}(x)]^2 \rangle_{\bar{w},x,\epsilon} = \left\langle \left[(z - \bar{z})^2 X X^T + \epsilon \right]^2 \right\rangle_{\bar{w},x,\epsilon}$$
$$= \frac{1}{N} \sum_i \left\langle (\bar{z}_i - z_i)^2 \right\rangle + \sigma_\epsilon^2$$
$$= \frac{1}{N} \sum_i \left[\left(\sigma_w^2 + \left(\sigma_w^0\right)^2 \right) e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i} \left(1 - e^{-\frac{\lambda_i t}{\tau}} \right)^2 \right] + \sigma_\epsilon^2 \tag{5}$$



Interpreting (5)

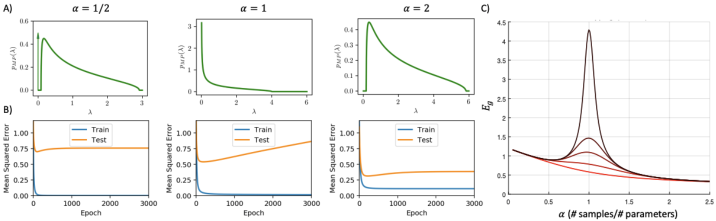
- What does

$$E_g(t) = \frac{1}{N} \sum_i \left[\left(\sigma_w^2 + \left(\sigma_w^0 \right)^2 \right) e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i} \left(1 - e^{-\frac{\lambda_i t}{\tau}} \right)^2 \right] + \sigma_\epsilon^2$$

  tell us?
  - Limits on generalization $E_g(t) \geq \sigma_\epsilon^2$
  - Scale component of error and overfitting component
  - Eigengap λ_{min} determines
    1. Error
    2. Timescales of learning
  - Frozen subspace of weights
  - Can write $\langle E_g(t) \rangle_X$

Double Descent

- Eigengap explains double descent!



- Final error on a *typical* dataset, $\frac{E_g(t)}{\sigma_w^2}$, is then given by

$$\int \rho^{\mathrm{MP}}(\lambda)\left[(1+\mathrm{INR})e^{-\frac{2\lambda t}{\tau}}+\frac{1}{\lambda\cdot\mathrm{SNR}}\left(1-e^{-\frac{\lambda t}{\tau}}\right)^2\right]d\lambda+\frac{1}{\mathrm{SNR}} \quad (6)$$

Reconciling with classical statistics

- Intuition: bigger models are more complex, so they will overfit.
- Formalism: $E_g - E_t \leq f(C(H))$ for some complexity measure $C(H)$ of a class of hypotheses H.
- Known that for Rademacher complexity, $C(H) = R(H)$

$$E_g - E_t \leq 2R(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2P}} \tag{7}$$

- Also known that for a one-hidden layer ReLU network,

$$\mathcal{R}(H) \leq \frac{B_2 B_1 C \sqrt{N_h}}{\sqrt{P}} \tag{8}$$

Naively (classical statistics), this implies bigger models have larger generalization gap. How can we square this with deep learning?

Reconciling with classical statistics

- We will see (8) is very loose, and tighten it using dynamics we derived, to discover that bounds on $R(H)$ for this type of network are actually *decreasing* with model size.
  1. ReLU network is $\hat{y}(x) = \sum_{a=1}^{N_h} \frac{w_a}{\sqrt{N_h}} \phi_a(x)$
  2. Consider modified data matrix $X_h^{a\mu} = \frac{1}{\sqrt{N_h}} \phi_a(W_a^1 \cdot x^\mu)$ so ReLU is "linear model" in these features
  3. Dynamics we derived give

$$z_i^h(t) = \frac{\tilde{z}_i^h}{\lambda_i^h} \left(1 - e^{-\frac{\lambda_i^h t}{\tau}}\right) + z_i^h(0) e^{-\frac{\lambda_i^h t}{\tau}} \tag{9}$$

  4. So that

$$\|w(t)\|^2 = \sum_i z_i^2(t) = \sum_i \left(\frac{\left\|\tilde{z}_i^h\right\|^2}{\left(\lambda_i^h\right)^2} \left(1 - e^{-\frac{\lambda_i^h t}{\tau}}\right)^2 + \left\|z_i^h(0)\right\|^2 e^{-2\frac{\lambda_i^h t}{\tau}} \right) \tag{10}$$

Completing reconciliation: a tighter generalization bound

- See that by *frozen subspace* and *eigengap* properties of $E_g(t)$

$$\frac{\|w\|}{\sqrt{N_h}} \leq \sqrt{\frac{\max_i \left\|\tilde{z}_i^h\right\|^2}{\min_{i,\lambda_i^h>0} \left(\lambda_i^h\right)^2} \frac{\min\left(P, N_h\right)}{N_h}} = B_2 \tag{11}$$

- So that our complexity is instead bounded by

$$\mathcal{R}(H) \leq B_1 C \sqrt{\frac{\max_i \left\|\tilde{z}_i^h\right\|^2}{\min_{i,\lambda_i^h>0} \left(\lambda_i^h\right)^2} \frac{\min\left(P, N_h\right)}{P}} \tag{12}$$

  which is *decreasing* in N_h.
- Decreasing $R(H)$ in model size means generalization gap also decreasing in model size, so gradient-based learning of a simple neural network does not contradict classical bounds.

Wrapping up

Take-aways

- A very simple model of a linear network already gives insight into many "deep learning" puzzles.
- These include a random matrix origin for double descent and tighter generalization bounds.
- The *eigengap* and *frozen subspace* properties are forms of implicit regularization in gradient-based learning.
- Current work is on achieving similar results for deep networks with different nonlinearities, as well as other data distributions. Due to universality, much still relies on RMT techniques!

Thank you for a great semester!