Multiple Descent: Design Your Own Generalization Curve

Lin Chen[1]

Yifei Min[2]

Mikhail Belkin[3]

Amin Karbasi[4]

[1]Simons Institute for the Theory of Computing
[2]Department of Statistics and Data Science, Yale University
[3]Halıcıoğlu Data Science Institute, UCSD
[4]Department of EECS, Yale University

NeurIPS 2021

Outline

Setup

Let $x_1, \cdots, x_n \in \mathbb{R}^D$ be i.i.d. training data of size n, and $x_{\text{test}} \in \mathbb{R}^D$ be the test date point.

- $x_1, \cdots, x_n, x_{\text{test}} \sim \mathcal{D}$

Consider a linear regression problem, where only the first d dimensions of the feature is revealed where $d < D$.

- Denote $\tilde{x}_i = x_i[1 : d] \in \mathbb{R}^d$.

The response is given by $y_i = \tilde{x}_i^\top \beta + \epsilon_i$, $i = 1, \cdots, n$, where the noise $\epsilon_i \sim \mathcal{N}(0, \eta^2)$ i.i.d.

- In this paper, the sample size n is fixed, and the dimension d can vary.

Problem

We want to study the least square estimator $\hat{\beta}$ of β and its excess generalization loss, as d increases.
To find $\hat{\beta}$:

- Denote the design matrix $A = [\tilde{x}_1, \cdots, \tilde{x}_n]^\top \in \mathbb{R}^{n \times d}$.
- We consider the estimator $\hat{\beta} = A^+(A\beta + \epsilon)$, where A^+ denotes the pseudo-inverse of A.
- In the underparametrized regime (i.e. $d < n$), $\hat{\beta}$ defined above is the OLS estimator.
- In the overparametrized regime (i.e. $d > n$), $\hat{\beta}$ is the minimum norm solution that achieves zero training error.

Recap of results in Liang et al. (2020)

- Liang et al. (2020) presented a multiple-descent upper bound on the risk of the minimum-norm interpolation vs. the data dimension.
- **Compared to our work:** A multiple-descent upper bound without a properly matching lower bound does not imply the existence of a multiple-descent generalization curve. We gave an explicit construction and proved the multiple descent of generalization curve itself.

Problem

Excess generalization loss L_d for any $d > 0$ is given by:

$$
\begin{aligned}
L_d &\triangleq \mathbb{E}\left[\left(y - x^\top \hat{\beta}\right)^2 - \left(y - x^\top \beta\right)^2\right] \\
&= \mathbb{E}\left[\left(x^\top (\hat{\beta} - \beta)\right)^2\right] \\
&= \mathbb{E}\left[\left(x^\top \left((A^+ A - I)\beta + A^+ \varepsilon\right)\right)^2\right] \\
&= \mathbb{E}\left[(x^\top (A^+ A - I)\beta)^2\right] + \mathbb{E}\left[(x^\top A^+ \varepsilon)^2\right] \\
&= \underbrace{\mathbb{E}\left[(x^\top (A^+ A - I)\beta)^2\right]}_{bias} + \underbrace{\eta^2 \mathbb{E}\left\|(A^\top)^+ x\right\|^2}_{variance},
\end{aligned}
\tag{1}
$$

where $y = x^\top \beta + \varepsilon_{\text{test}}$ and $\varepsilon_{\text{test}} \sim \mathcal{N}(0, \eta^2)$.

Underparametrized Regime

In the underparametrized regime, if $\mathcal{D}$ is a continuous distribution, the matrix A has independent column almost surely. Then we have

$$L_d = \eta^2 \mathbb{E} \left\| (A^\top)^+ x \right\|^2 .$$

This is because in this case, we have $A^+ A = I$ and therefore the bias $\mathbb{E}\left[(x^\top (A^+ A - I)\beta)^2 \right]$ vanishes.

Theorem (Underparametrized regime)

If $d < n$, we have $L_{d+1} \geq L_d$ irrespective of the data distribution $\mathcal{D}$. Moreover, for any $C > 0$, there exists a distribution $\mathcal{D}$ such that $L_{d+1} - L_d > C$.

Underparametrized Regime

Theorem (Underparametrized regime)

If $d < n$, we have $L_{d+1} \geq L_d$ irrespective of the data distribution $\mathcal{D}$. Moreover, for any $C > 0$, there exists a distribution $\mathcal{D}$ such that $L_{d+1} - L_d > C$.

Remark:

- The theorem says that in the underparametrized regime, the excess generalization loss always increases as d increases.
- The increase can have arbitrary magnitude.
- Our proof shows that $\mathcal{D}$ can be a product distribution (i.e. independence between dimensions), and each single distribution can be as simple as a Gaussian mixture.

Underparametrized Regime

Proof Sketch:

Do a decomposition

$$\begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ = \left[\left(I - \frac{bb^\top}{\|b\|^2} \right) \left(I + \frac{AA^+ bb^\top}{\|b\|^2 - b^\top AA^+ b} \right) (A^+)^\top, \frac{(I - AA^+)b}{\|b\|^2 - b^\top AA^+ b} \right]$$

and we can get

$$\begin{aligned} & L_{d+1} - L_d \\ = & \mathbb{E} \left[\left\| \begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ \begin{bmatrix} x \\ x_1 \end{bmatrix} \right\|^2 - \left\| (A^+)^\top x \right\|^2 \right] \\ \geq & -\underbrace{\mathbb{E}\|(A^+)^\top x\|^2}_{I_1} + \underbrace{\mathbb{E}\left[\frac{x_1^2}{\sum_{i=1}^n b_i^2} \right]}_{I_2}. \end{aligned}$$

Then we show I_1 is finite and I_2 can be made arbitrarily big by some $\mathcal{D}$.

Overparametrized Regime

For the overparametrized regime where $d > n$, we consider two cases:

- $\beta = 0$.
- $\beta \neq 0$.

Overparametrized Regime

$\beta = 0$ case:

- L_d is just the variance $\mathbb{E}\left\|(A^\top)^+ x\right\|^2$.

Theorem (Overparametrized regime, $\beta = 0$)

Let $n < D - 9$. Given any sequence $\Delta_{n+8}, \Delta_{n+9}, \ldots, \Delta_{D-1}$ where $\Delta_d \in \{\uparrow, \downarrow\}$, there exists a distribution $\mathcal{D}$ such that for every $n + 8 \leq d \leq D - 1$, we have

$$L_{d+1} \begin{cases} > L_d, & \text{if } \Delta_d = \uparrow \\ < L_d, & \text{if } \Delta_d = \downarrow. \end{cases}$$

- Remark: The theorem says that we can control the ascent/descent in the overparametrized regime.

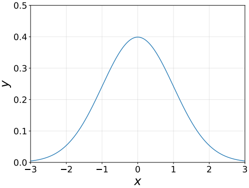
Overparametrized Regime

What is the distribution $\mathcal{D}$?

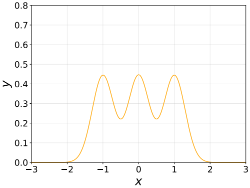
Overparametrized Regime

What is the distribution $\mathcal{D}$?

- It turns out that one can create a descent ($L_{d+1} < L_d$) by adding a Gaussian feature and create an ascent ($L_{d+1} > L_d$) by adding a Gaussian mixture feature.



$\mathcal{N}(0, 1)$ feature



$\mathcal{N}_\sigma^{\mathrm{mix}}$ feature

Overparametrized Regime

Gaussian β setting: we study the setting where each entry of β is i.i.d. $\mathcal{N}(0, \rho^2)$. We show that multiple descent can also be achieved.

Overparametrized Regime

Gaussian β setting: we study the setting where each entry of β is i.i.d. $\mathcal{N}(0, \rho^2)$. We show that multiple descent can also be achieved.
We have bias when $\beta \neq 0$.

$$\mathcal{E}_d \triangleq (x^\top (A^+ A - I)\beta)^2, \quad \mathcal{E}_{d+1} \triangleq \left([x^\top, x_1]([A, b]^+ [A, b] - I) \begin{bmatrix} \beta \\ \beta_1 \end{bmatrix} \right)^2 .$$

Overparametrized Regime

Gaussian β setting: we study the setting where each entry of β is i.i.d. $\mathcal{N}(0, \rho^2)$. We show that multiple descent can also be achieved.
We have bias when $\beta \neq 0$.

$$\mathcal{E}_d \triangleq (x^\top (A^+ A - I)\beta)^2, \quad \mathcal{E}_{d+1} \triangleq \left([x^\top, x_1]([A, b]^+ [A, b] - I) \begin{bmatrix} \beta \\ \beta_1 \end{bmatrix} \right)^2 .$$

Then the expected risks

$$L_d^{\exp} = \mathbb{E}[\mathcal{E}_d] + \eta^2 \mathbb{E} \left\| (A^\top)^+ x \right\|^2,$$

$$L_{d+1}^{\exp} = \mathbb{E}[\mathcal{E}_{d+1}] + \eta^2 \mathbb{E} \left\| \begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ \begin{bmatrix} x \\ x_1 \end{bmatrix} \right\|^2 ,$$

where $\beta \sim \mathcal{N}(0, \rho^2 I_d)$ and $\beta_1 \sim \mathcal{N}(0, \rho^2)$.

Overparametrized Regime

Theorem (informal)

Under mild conditions, the following holds:

1. *If* $x_1, b_1, \ldots, b_n \overset{iid}{\sim} \mathcal{N}^{\text{mix}}_{\sigma,\mu}$*, for any* $C > 0$*, there exist* μ, σ *such that* $L^{\text{exp}}_{d+1} - L^{\text{exp}}_{d} > C$*.*
2. *If* $x_1, b_1, \ldots, b_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$*, there exists* $\sigma > 0$ *such that for all*

$$\rho \leq \eta \sqrt{\frac{\mathbb{E}[\|(A^\top A)^+ x\|^2]}{\mathbb{E}\|A^{+\top} x\|^2 + 1}},$$

we have $L^{\text{exp}}_{d+1} < L^{\text{exp}}_{d}$*.*

Summary

- Our work proves that the expected risk of linear regression can manifest multiple descents when the number of features increases and sample size is fixed.
- This is done by designing the distribution of each feature.
- Specifically, the procedure enables us to control local maxima in the underparametrized regime and control ascents/descents freely in the overparametrized regime.

Thank you!

Chen, Min, Belkin, Karbasi, “Multiple Descent: Design Your Own Generalization Curve,”

LIANG, T., RAKHLIN, A. and ZHAI, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*. PMLR.