Triple Descent phenomenon

Ghassan Najjar, Ruben Illouz

Ecole Polytechnique

Mar 08, 2022

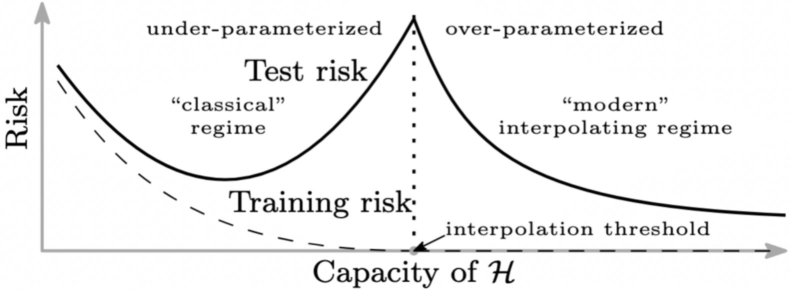
Plan

1. First observations

2. Random features model

3. Results

Double descente



Double descent curve, *medium.com*

Details

Notations :

- P : the number of parameters of the model
- N : the number of observations of the dataset
- D : the dimension of each observation

Two types of peaks

Now, it is no longer the number of parameters that is varied but the number of elements in the training dataset.

1. Linear peak : $N = D$
2. Non-linear peak : $N = P$

Questions: two different phenomena? Can they coexist?

Linear spike $N = D$

Linear regression example

In the case of a linear regression with N samples of dimension D, we obtain an interpolation peak for $N = D = P$

in the case $N = D$, we have N equations with N unknowns. same peak is observed in the case of a neural network with linear activation functions.

Non-linear peak

In what cases?

In the case of neural networks with non-linear activation functions (in a sense that we will specify), we obtain a peak that no longer depends on the dimension D and that appears for $N = P$ (non-linear peak).

then what happens with intermediate activation functions (between linear and non-linear) the peak $N = D$, $N = P$, both remain?

Experimentation: Random feature model

- Dataset : $X \in \mathbb{R}^{N \times D}$, N lines drawn in iid ways according to $\mathcal{N}(0,1)$.
- Label generator : f^*, we obtain the labels $y = f^\star(x) + \epsilon$ where ϵ follows $\mathcal{N}(0,1)$.

Definition of f^*

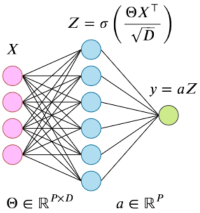
Let β be a vector of dimension D with coordinates randomly drawn according to $\mathcal{N}(0,1)$.
Then we define $f^\star(\boldsymbol{x}) = \langle \boldsymbol{x} \rangle / \sqrt{D}$

Random Feature Model

Definition

Let $Theta be$N in $mathbbR^{P \times D}$ containing the P vectors of random features.We can see the random feature model as a double layer network whose first layer is the Θ matrix.

fact, Θ will send $x \in R^D$ on $z \in R^P f(\boldsymbol{x}) = \sum_{i=1}^{P} \boldsymbol{a}_i(\boldsymbol{x}) = sum_{i=1}^{P} \boldsymbol{a}_i = f(\boldsymbol{x})$

- 

source : Article d'étude

Training the Random feature model

In the case of a linear regression on the modified features, we seek to find the vector $\boldsymbol{a}$ in the framework of a Ridge regression.

Training

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}\in\mathbb{R}^P}{\arg\min}\left[\frac{1}{N}\left(\boldsymbol{y} - \boldsymbol{a}\boldsymbol{Z}^\top\right)^2 + \frac{P\gamma}{D}\|\boldsymbol{a}\|_2^2\right] = \frac{1}{N}\boldsymbol{y}^\top\boldsymbol{Z}\left(\boldsymbol{\Sigma} + \frac{P\gamma}{D}\mathbb{I}_P\right)^{-1}$$

$$\boldsymbol{Z}_i^\mu = \sigma\left(\frac{\langle\boldsymbol{\Theta}_i, \boldsymbol{X}_\mu\rangle}{\sqrt{D}}\right) \in \mathbb{R}^{N\times P}, \quad \boldsymbol{\Sigma} = \frac{1}{N}\boldsymbol{Z}^\top\boldsymbol{Z} \in \mathbb{R}^{P\times P}$$

Test loss space phase

Test loss

The test loss is calculated by drawing $\boldsymbol{x} \sim \mathcal{N}(0,1) : \mathcal{L}_g = \mathbb{E}_{\boldsymbol{x}} \left[(f(\boldsymbol{x}) - f^\star(\boldsymbol{x}))^2 \right]$.

Special case of the random features model

In our particular case, we know explicitly the loss test in the frame

$$N, D, P \to \infty, \quad \frac{D}{P} = \psi = \mathcal{O}(1), \quad \frac{D}{N} = \phi = \mathcal{O}(1)$$

Gaussian equivalence theorem

Important variables

$$\eta = \int \frac{e^{-z^2/2}}{sqrt 2\pi} \sigma^2(z) dz, \quad \zeta = \left[\int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \sigma^{prime}(z) \right]^2 \quad \text{and} \quad r = frac \zeta \eta \in [0, 1]$$

r is the degree of linearity of the activation function σ.

this sense, the absolute value function has a zero degree of linearity.

Gaussian equivalence theorem

$$\boldsymbol{Z} = \sigma\left(\frac{\boldsymbol{X}\Theta^\top}{\sqrt{D}}\right) \to \sqrt{\zeta}\frac{\boldsymbol{X}\Theta^\top}{\sqrt{D}} + \sqrt{\eta - \zeta}\boldsymbol{W}, \quad \boldsymbol{W} \sim \mathcal{N}(0, 1)$$

Spectral analysis

The link between eigenvalues and peak

The peak (double descent) of an unregularised linear regression on iid data is related to the presence of small (but non-zero) eigenvalues of the covariance matrix of the model data.

Note: the *random feature model* is nothing more than a linear regression on the modified coordinates $\mathbf{Z} \in \mathbb{R}^{N \times P}$ this case, we study the spectrum of the matrix $\boldsymbol{\Sigma} = \frac{1}{N} \boldsymbol{Z}^\top \boldsymbol{Z}$.

Appearance on the spectral density

Some results from the theory of random matrices.

Calculation of the spectral density

We can calculate the spectral density $\rho(\lambda)$ by solving the implicit equation below:

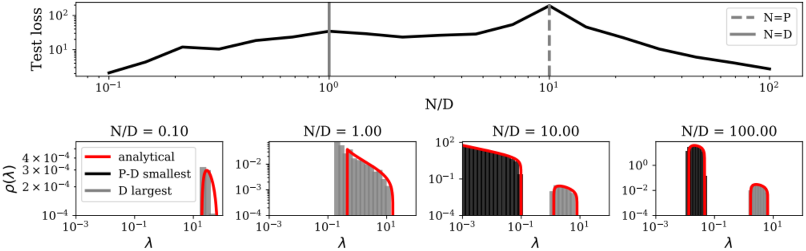
$$\rho(\lambda) = \frac{1}{\pi} \lim_{epsilon \to 0^+} \operatorname{Im} G(\lambda - i\epsilon), \quad G(z) = \frac{1}{z} A\left(\frac{1}{z\psi} z\right) + \frac{1-\psi}{z} \quad A(t) = 1 + (\eta - \zeta) t A_\phi(t) A$$

where $A_\phi(t) = 1 + (A(t) - 1)$ and $A_\psi(t) = 1 + (A(t) - 1)$

- We obtain a theoretical spectral density towards which the empirical spectral density converges when $N, D, P \to \infty, \quad \frac{D}{P} = \psi = \mathcal{O}(1), \quad \frac{D}{N} = phi = \mathcal{O}(1)$

Spectral density plots

- La résolution de l'équation implicite nous donne la densité spectrale dessinée en rouge, on observe une séparation pour $N > D$
- L'histogramme est l'histogramme empirique des valeurs propres de $\boldsymbol{\Sigma} = \frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{Z} \in \mathbb{R}^{P\times P}$.
- la formule explicite du *test loss* nous donne le tracé noir.



source : Article d'étude

Results for spectral density

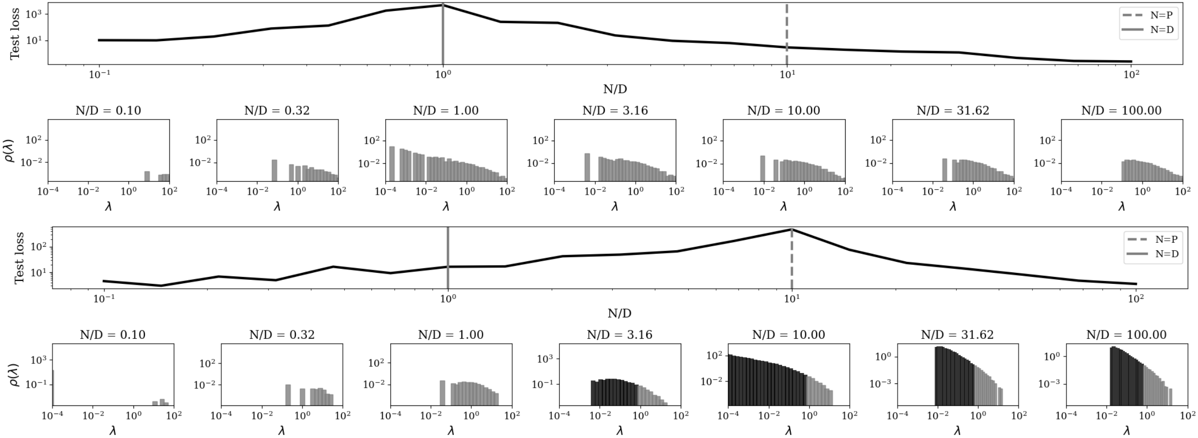


Figure: Spectral density for MNIST $\gamma = 10^{-5}$, $P/D = 10$, $SNR = 0.2$ in the case σ linear then $\sigma = abs$

Results for spectral density

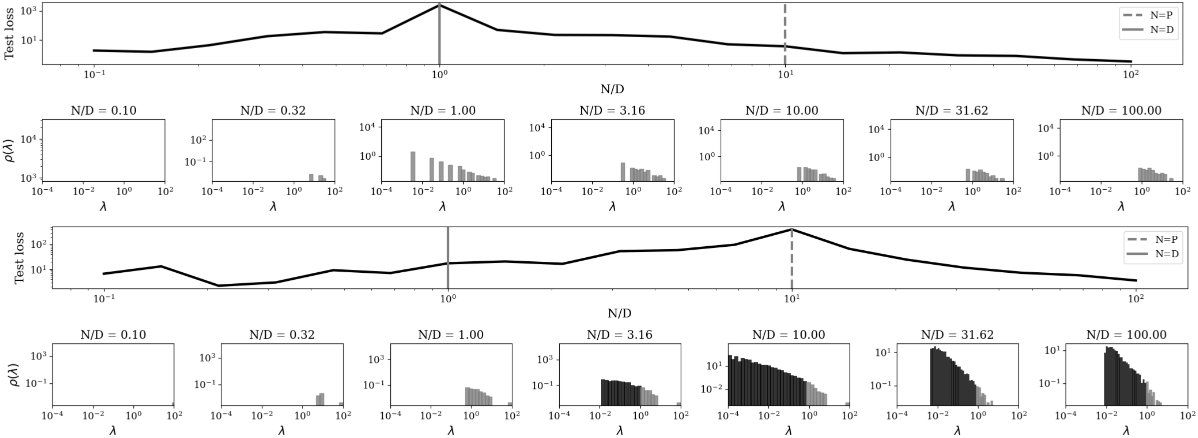
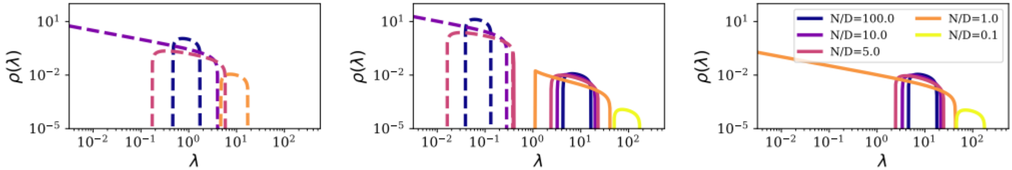


Figure: Spectral density for KMNIST $\gamma = 10^{-5}, P/D = 10, SNR = 0.2$ where σ is linear and $\sigma = abs$

Influence of the linearity degree r r



(a) Absolute value ($r = 0$) (b) Tanh ($r \simeq 0.92$) (c) Linear ($r = 1$)

source : Article

lines : represent the non-linear eigenvalues. lines: represent linear eigenvalues.

Influence of the non-linearity

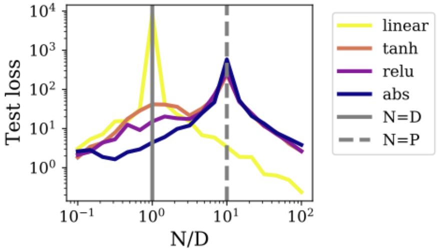


Figure: Effect of the activation function, *Article*

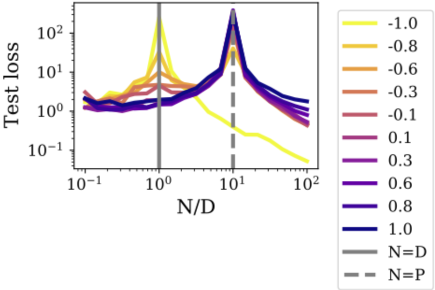


Figure: Effet of the linearity, *Article*

Résultats for other datasets

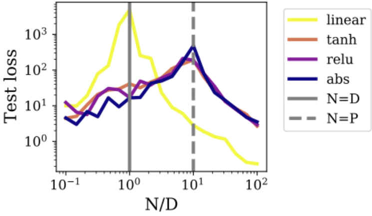


Figure: MNIST

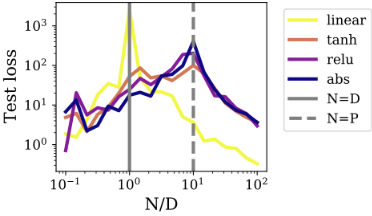


Figure: KMNIST

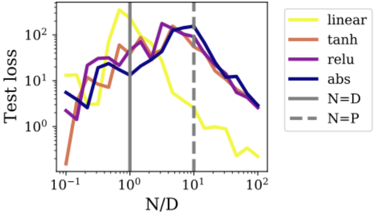
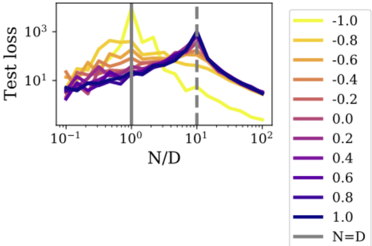
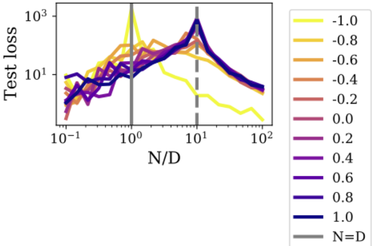
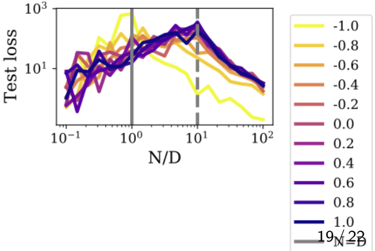


Figure: Fashion MNIST







Impact of the regularisation and of the ensembling

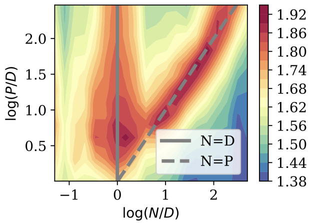


Figure: With Standard (with $\sigma = Tanh$, $\gamma = 0$,K=1), *Article*

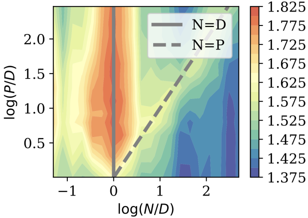


Figure: With Ensembling (with $\sigma = Tanh$, K=10), *Article*

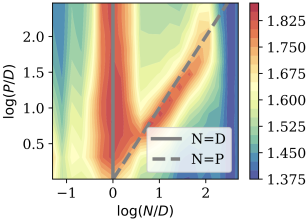


Figure: With regularisation ($\sigma = Tanh$, $\gamma = 0.05$), *Article*

- Regularisation and ensembling act mainly on the non-linear peak

Appearance of the non-linear peak

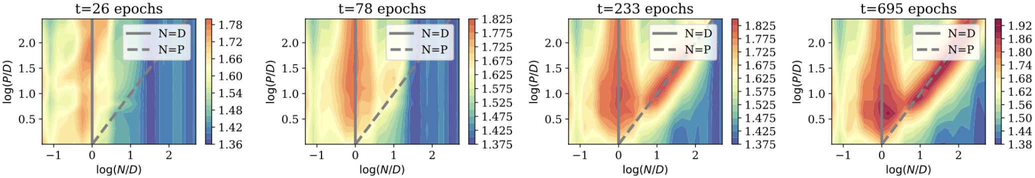


Figure: Effect of the time of training on the NN model ($\sigma = Tanh$)*Article*

Conclusion

- Choice of N, P must be judicious
- The two peaks are quite distinct but can appear at the same time
- The relative size of the peaks is related to linearity