A Note on Double Descent

University College London Centre for AI Research Note

David Barber

November 24, 2023

Abstract

Double Descent is the phenomenon that the test error in a learning system displays non-monotonic behaviour as the number of train datapoints increases. Double Descent was well known in the 1990s and this brief note adds some references and details around how to calculate the test error and suggests an explanation for the phenomenon.

Whilst the phenomenon of Double Descent may have been recently generally observed, see for example [1], the phenomenon also occurs in simple linear systems in which it has been previously well studied mathematically - see for example [2] for some historical links. We will add some more references and details on the calculation, and give an explanation for the phenomenon.

1 Linear Regression

Consider a simple learning system with data (x_i, y_i), $i = 1, \ldots, P$, with vector inputs $x_i \in \mathbb{R}^N$. The model we will fit to this data is

$$y = \frac{1}{\sqrt{N}} w^\mathsf{T} x \tag{1}$$

where $w \in \mathbb{R}^N$ is the weight vector.

We assume that the data is generated by a model of the same form, but with unknown parameter w_0 and additive Gaussian noise $\epsilon_i \in \mathbb{R}$, drawn i.i.d from a zero mean Gaussian with variance σ^2. That is, each observation is generated from

$$y_i = \frac{1}{\sqrt{N}} w_0^\mathsf{T} x_i + \epsilon_i \tag{2}$$

We further assume that each $x_i \in \mathbb{R}$ is i.i.d drawn from a zero mean unit covariance Gaussian.

Then for a test point x the error is

$$(y - y_0)^2 = \left(\frac{1}{\sqrt{N}} w^\mathsf{T} x - \frac{1}{\sqrt{N}} w_0^\mathsf{T} x + \epsilon \right)^2 = \left(\frac{1}{\sqrt{N}} (w - w_0)^\mathsf{T} x + \epsilon \right)^2 \tag{3}$$

Averaging this over the test noise ϵ we obtain

$$\left\langle (y - y_0)^2 \right\rangle_\epsilon = \frac{1}{N} (w - w_0)^\mathsf{T} x x^\mathsf{T} (w - w_0) + \sigma^2 \tag{4}$$

We assume that the train and test data are drawn i.i.d from a zero mean unit covariance Gaussian. Averaging the test error over the test input x we obtain

$$\left\langle (y - y_0)^2 \right\rangle_{\epsilon,x} = \frac{1}{N} (w - w_0)^2 + \sigma^2 \tag{5}$$

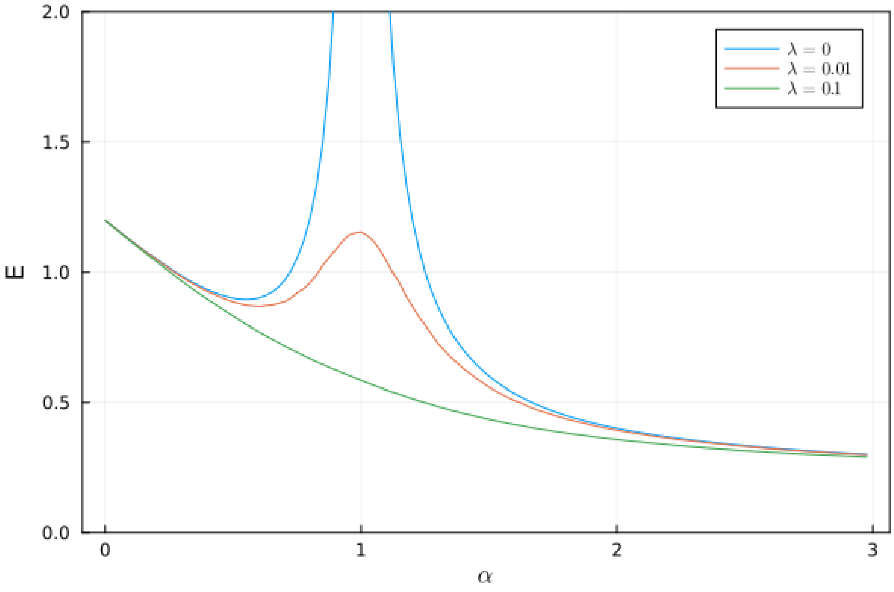

Figure 1: The generalisation error against $\alpha = P/N$ the number of datapoints ratio, for a Linear Perception with additive Gaussian noise $\sigma^2 = 0.2$, $N = 1000$ and different regularisation values λ. The graph shows the characteristic "double descent" in which for underregularised systems, the generalisation error increases as the number of datapoints nears the dimension of the model, before decreasing with increasing number of datapoints.

1.1 Ordinary Least Squares

Based on the training objective

$$E_{train}(w) \equiv \sum_{i=1}^{P} (y_i - \frac{1}{\sqrt{N}} w^\mathsf{T} x_i)^2 + \lambda w^2 \tag{6}$$

with regularisation parameter λ, the solution that minimises E_{train} is

$$\begin{aligned} w &= \left(\frac{1}{N} \sum_i x_i x_i^\mathsf{T} + \lambda I \right)^{-1} \frac{1}{\sqrt{N}} \sum_i y_i x_i & (7) \\ &= \left(\frac{1}{N} \sum_i x_i x_i^\mathsf{T} + \lambda I \right)^{-1} \frac{1}{\sqrt{N}} \sum_i \left(\frac{1}{\sqrt{N}} w_0^\mathsf{T} x_i + \epsilon_i \right) x_i & (8) \end{aligned}$$

Let $A = \frac{1}{N} \sum_{i=1}^{P} x_i x_i^\mathsf{T}$ and define $M = A + \lambda I$

$$w - w_0 = M^{-1} \frac{1}{N} \sum_i \left(w_0^\mathsf{T} x_i + \sqrt{N} \epsilon_i \right) x_i - w_0 \tag{9}$$

Averaging over the train noise ϵ_i we have

$$\begin{aligned} \langle (w - w_0)^2 \rangle_{\epsilon, x, \epsilon_{1:P}} &= \left(M^{-1} A w_0 - w_0 \right)^2 + \frac{\sigma^2}{N} \text{trace} \left(M^{-1} \sum_i x_i x_i^\mathsf{T} M^{-1} \right) & (10) \\ &= \left(M^{-1} A w_0 - w_0 \right)^2 + \sigma^2 \text{trace} \left(M^{-1} A M^{-1} \right) & (11) \end{aligned}$$

Assuming that w_0 is drawn from a zero mean unit covariance Gaussian, the first term averages over w_0 to

$$\left\langle \left(M^{-1} A w_0 - w_0 \right)^2 \right\rangle_{w_0} = \text{trace} \left(M^{-1} A - I \right)^2 = \lambda^2 \text{trace} \left(M^{-2} \right) \tag{12}$$

This gives the final expression for the test error $E(X) = \langle (y - y_0)^2 \rangle_{\epsilon, x, \epsilon_{1:P}, w_0}$

$$E(X) = \sigma^2 + \sigma^2 \frac{1}{N} \text{trace} \left(M^{-1} \right) + \lambda \left(\lambda - \sigma^2 \right) \frac{1}{N} \text{trace} \left(M^{-2} \right) \tag{13}$$

Ideally, for mathematical elegance, we would also average the test error over the train data to define

$$E = \langle E(X) \rangle_X \tag{14}$$

where X are the train inputs. In general there is no known closed form expression for $\langle E(X) \rangle_X$. However, there are exact results for $N \to \infty$ and separately the Pseudo-Inverse limit $\lambda \to 0$ for finite N. We will outline the Pseudo-Inverse case below. For further details and the $N \to \infty$ case see for example [3, 4, 5, 6, 7, 8, 9].

2 Pseudo Inverse

In the limit $\lambda \to 0$, the OLS weight vector w tends to the Pseudo-Inverse solution,

$$w = X^{\mathsf{T}} \left(X^{\mathsf{T}} X \right)^{-1} Y \tag{15}$$

where we define the matrix $X^{\mathsf{T}} = (x_1, \ldots, x_P)$ and vector $Y^{\mathsf{T}} = (y_1, \ldots, y_P)$. In this case, there is an exact expression for the generalisation error (see [9],[6, p. 40, 41]) which is based on the fact that A is Wishart distributed, and gives

$$E = \begin{cases} 1 - \frac{P}{N} + \sigma^2 \frac{N-1}{N-P-1} & \text{if } P < N-1 \\ \sigma^2 + \frac{N\sigma^2}{P-N-1} & \text{if } P > N+1 \end{cases} \tag{16}$$

A brief derivation of this result is given below. These details are given for example in [9] and [6].

2.1 $P > N + 1$

In this case we have simply

$$E(X) = \sigma^2 + \sigma^2 \text{trace}\left(A^{-1}\right) \tag{17}$$

Using the fact that A^{-1} is inverse Wishart distributed, we have the result

$$\left\langle \text{trace}\left(A^{-1}\right) \right\rangle = \frac{N}{P-N-1} \tag{18}$$

and

$$\langle E(X) \rangle_X = \sigma^2 + \sigma^2 \frac{N}{P-N-1} \tag{19}$$

2.2 $P < N - 1$

We first evaluate the terms of equation(13) for finite λ. For the situation $P < N - 1$, we note that the matrix $M = A + \lambda I$ will have $N - P$ eigenvalues of value λ (since A has an $N - P$ dimensional null space). The remaining eigenvalues of $M = X^{\mathsf{T}} X / N + \lambda I$ for eigenvalue γ and eigenvector e satisfy

$$\left(X^{\mathsf{T}} X / N + \lambda I \right) e = \gamma e \tag{20}$$

Equivalently, we can write

$$X X^{\mathsf{T}} X / N e + \lambda X e = \gamma X e \tag{21}$$

that is

$$\left(X X^{\mathsf{T}} / N + \lambda I \right) X e = \gamma X e \tag{22}$$

Hence the eigenvalues are also the eigenvalues of the $P \times P$ matrix $\tilde{M} \equiv X X^{\mathsf{T}} / N + \lambda I$. Hence

$$E(X) = \sigma^2 + \frac{\sigma^2}{N} \left(\frac{N-P}{\lambda} + \text{trace}\left(\tilde{M}^{-1} \right) \right) + \frac{\lambda \left(\lambda - \sigma^2 \right)}{N} \left(\frac{N-P}{\lambda^2} + \text{trace}\left(\tilde{M}^{-2} \right) \right) \tag{23}$$

Defining $B = XX^\mathsf{T}/N$, for the matrix $\tilde{M} = B + \lambda I$, the eigenvalues/vectors satisfy

$$(B + \lambda I)^{-1} e = \gamma e \tag{24}$$

then

$$\gamma^{-1} e = (B + \lambda I) e = (\beta + \lambda)\, e \tag{25}$$

where β is an is an eigenvalue of B. Hence the eigenvalues of $(B + \lambda I)^{-1}$ are given by $\gamma = 1/(\beta + \lambda)$ and $\mathrm{trace}\left(\tilde{M}^{-2}\right) = \sum_{i=1}^{P} 1/(\beta_i + \lambda)^2$. Simplifying $E(X)$ and taking the limit $\lambda \to 0$, the term $\mathrm{trace}\left(\tilde{M}^{-2}\right)$ does not contribute and we arrive at

$$E(X) = 1 - \frac{P}{N} + \sigma^2 \left(1 + \frac{1}{N} \mathrm{trace}\left(B^{-1}\right)\right) \tag{26}$$

Since B is also a correlation matrix, it is also Wishart distributed, and we can reuse the result equation(18) by interchanging P and N to derive $\left\langle \mathrm{trace}\left(B^{-1}\right)\right\rangle_X$, giving equation(16).

In figure(1) we plot the test error for the linear regression problem against $\alpha = P/N$ and a system dimension $N = 1000$. For the case $\lambda = 0$ the curve is exact and represents the full average, including over the train inputs. For $\lambda > 0$ the plots are simply the result of a single train set, rather than averaging over train sets. As we see, for an under-regularised system, the Double Descent phenomenon can occur in which, despite the amount of train data increasing, the test error can increase, before subsequently decreasing.

3 An Explanation

The Double Descent phenomenon involves a delicate relationship on the regularisation amount λ, the amount of noise and the ratio of train data to model dimension.

It is probably easiest to explain why this happens for the limit of no regularisation. In this case ($\lambda = 0$) the weight vector w is determined by the data subspace and the noise on the data labels (see equation(26)) increases the test error. Consider the general setting of selecting P points at random in a P-dimensional space. For large P, there is a high probability that at least one direction will not be well spanned by the selected points. More precisely, for a $P \times P$ matrix with random Gaussian[1] entries, there is a high probability that at least one of the singular values will be very small. Since the eigenvalues of B are the square of the singular values, $\mathrm{trace}\left(B^{-1}\right)$ will be the sum of the squared inverse singular values. If any of them is small, this will create a large value for $\mathrm{trace}\left(B^{-1}\right)$. This is almost guaranteed to happen for randomly chosen datapoints in a high dimensional space. This phenomenon is well studied in random matrix theory – see for example [8].

In this sense, one explanation for Double Descent is that in the presence of noise, as the amount of train data increases (up to the model dimension), so does the chance that the data effectively lie on a lower dimensional subspace. The noise dominates in those directions of the space that are not well covered, causing a spike in the test error. In more general non-linear settings, the explanation for Double Descent is likely the same, namely that not all directions are covered well by the data (and the probability of this happening increases with the training data up to the effective dimension of the model), meaning that the noise in those underspecified directions has a dominant and undesirable effect on the determination of the model.

References

[1] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc Natl Acad Sci, USA*, 116(32):15849–15854, 2019.

[1] The phenomenon of small singular values is not restricted to Gaussian distributed x data. For example, random binary attributes would also have the same property.

[2] M. Loog, T. Viering, A. Mey, Krijthe J.H., and D.M.J. Tax. A brief prehistory of double descent. *Proc Natl Acad Sci*, 117(20):10625–10626, May 2020.

[3] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11), 1990.

[4] A. Krogh and J. Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25:1135, 1991.

[5] P. Sollich. Learning in large linear perceptrons and why the thermodynamic limit is relevant to the real world. *Advances in Neural Information Processing Systems*, 7:207–214, 1994.

[6] D. Barber. *Finite Size Effects in Neural Network Algorithms*. PhD thesis, University of Edinburgh, 1996.

[7] D. Barber, D. Saad, and P. Sollich. Test Error Fluctuations in Finite Linear Perceptrons. *Neural Computation*, 7(4):809–821, 7 1995.

[8] A. Krogh. Learning with noise in a linear percepton. *J. Phys. A: Math. Gen.*, 25:1119, 1992.

[9] L. K. Hansen. Stochastic linear learning: exact test and training error averages. *Neural Networks*, 6:393–396, 1993.