An attempt to explain double descent in modern machine learning

Jochen Merker, Gregor Schuldt
HTWK Leipzig, PF 30 11 66, D-04251 Leipzig

This article aims to explain mathematically, why the so called double descent observed by Belkin et al., Reconciling modern machine-learning practice and the classical bias-variance trade-off, PNAS 116(32) (2019), p. 15849-15854, occurs on the way from the classical approximation regime of machine learning to the modern interpolation regime. We argue that this phenomenon may be explained by a decomposition of mean squared error plus complexity into bias, variance and an unavoidable irreducible error inherent to the problem. Further, in case of normally distributed output errors, we apply this decomposition to explain, why LASSO provides reliable predictors avoiding overfitting.

1. Introduction

While standard statistical machine learning theory [1] predicts that bigger models should be more prone to overfitting, modern machine learning practitioners claim that bigger models are always better. Within the approximation regime of machine learning, the first point of view is supported by the classical property of bias-variance trade-off well-known in literature since the pioneering work [2]. However, in the modern interpolation regime of deep learning, where there are so many parameters that data points do not have to be approximated but may be interpolated, a decay of the bias towards an asymptote was observed by [3] in neural networks as the network width increases. Belkin et al. [4] provided strong evidence for the existence and ubiquity of such a double descent (see Figure 1) for a wide spectrum of data models including decision trees and simple neural networks.

A concise explanation of this double descent phenomenon still seems to be missing in literature, although there are recent attempts [5] using a fine-grained bias-variance decomposition of the mean-squared error (MSE). In this article, we follow a similar strategy and aim to provide an explanation for the occurence of double descent in machine learning using complexity [6]. This correponds to the fact that

- an underfitted model has a high MSE on training data, high bias, low variance, low complexity,
- an overfitted model has a low MSE on training data, low bias, high variance, high complexity,
- a complexity reduced interpolation model has zero MSE on training data, low bias, reduced variance, reduced complexity.

Further, we will offer an explanation why LASSO [7], i.e. L^1-regularized approximation, occurs naturally in statistical learning and provides reliable predictors avoiding overfitting.

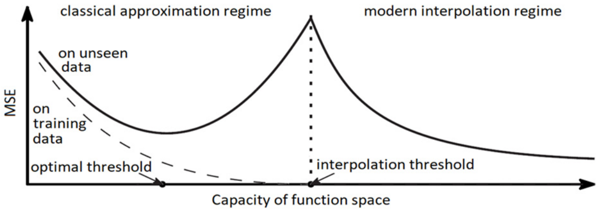

Figure 1: MSE on unseen data is a double descent curve [4]

2. Standard Bias-Variance Decomposition

Machine learning methods estimate an unknown deterministic input-output map from given inputs $u_i \in \mathcal{U}$ and corresponding outputs $y_i \in \mathcal{Y}$, $i = 1, \dots, N$, as training data. In the mathematical framework of statistical learning theory, it is assumed that

- the inputs u_i, $i = 1, \dots, N$, are drawn from a measurable space $\mathcal{U}$ independently according to a fixed but unknown distribution of a random variable U,
- for every input u_i an output y_i is drawn from a measurable space $\mathcal{Y}$ according to a fixed but unknown conditional distribution of a random variable Y given $U = u_i$.

For simplicity, let us further assume that $\mathcal{U}$ and $\mathcal{Y}$ are finite-dimensional real Hilbert spaces, and that the distribution of the input resp. the conditional distribution of the output to an input are continuous, i.e. have densities $f_U(u)$ resp. $f_{Y|U}(y|u)$ w.r.t. Lebesgue measure. Then, the density of the joint distribution of (U, Y) is given by $f_{(U,Y)}(u, y) = f_{Y|U}(y|u) f_U(u)$, and the unknown deterministic input-output map $F: \mathcal{U} \longrightarrow \mathcal{Y}$ is the expected output given an input, i.e.

$$Fu \;=\; E(Y|U = u) = \int_{\mathcal{Y}} y\, f_{Y|U}(y|u)\, dy\,.$$

Because a machine learning method selects on the basis of training data $(u_{i,}, y_i)$, $i = 1, \dots, N$, a map $F_{(u_1,y_1),\dots,(u_N,y_N)}$ from a D-dimensional space $\mathcal{F}$ of functions mapping $\mathcal{U}$ to $\mathcal{Y}$ as estimation of the unknown deterministic input-output map $F = E(Y|U)$, in statistical learning theory the predictor itself has to be considered as random variable $F_T = F_{(U_1,Y_1),\dots,(U_N,Y_N)}$ induced by the machine learning method and by the random variables (U_i, Y_i), $i = 1, \dots, N$, which are i.i.d. as (U, Y).

Within this theory, the failure of overfitted models – low MSE on training data, but high variation w.r.t. training data and therefore bad generalization to unseen data – can be mathematically explained by the decomposition

$$E_{(U,Y),T}\left(|F_T U - Y|_{\mathcal{Y}}^2\right) = E_{U,T}\left(|F_T U - E_T(F_T)U|_{\mathcal{Y}}^2\right) + E_U\left(|E_T(F_T)U - E(Y|U)|_{\mathcal{Y}}^2\right) + E_{(U,Y)}\left(|Y - E(Y|U)|_{\mathcal{Y}}^2\right) \tag{1}$$

of the expected squared error of F_T w.r.t. training data T into three parts: The variance of F_T w.r.t. training data, the squared bias of the predictor $E_T(F_T)$ expected from training data, and the variance of Y as an unavoidable irreducible error inherent to the problem. Now, if you enhance the capacity of the function space $\mathcal{F}$ by increasing the dimension D beyond an optimal threshold, then the squared bias of the predictor $E_T(F_T)$ often decreases, because $E_T(F_T)U$ can better resemble the output $FU = E(Y|U)$ expected at U. However, variance $E_{U,T}\left(|F_T U - E_T(F_T)U|_{\mathcal{Y}}^2\right)$ w.r.t. training data T often increases more strongly, leading in total to a larger expected squared error of F than at the optimal threshold. This bias-variance trade-off explains the failure of overfitted models in the classical approximation regime of machine learning. Yet, there is no necessity that a machine learning method shows this behaviour, and in fact, a simultaneous decrease of variance and bias can be observed in the modern interpolation regime of machine learning [3,4].

3. Bias-Variance-Complexity Decomposition

In this section, we aim to explain double descent and simultaneous decrease of variance and bias in the modern interpolation regime of machine learning by a bias-variance decomposition of mean squared error plus complexity. Let us consider the mean squared error with random training data, i.e. the random variable

$$\frac{1}{N}\sum_{i=1}^{N} |F_T U_i - Y_i|_{\mathcal{Y}}^2$$

If we try to calculate the expected value of this random variable w.r.t. training data and to obtain a similar decomposition as (1), it makes sense to notationally separate the i-th variable from the other variables by setting $F_{(u_i,y_i)} \coloneqq E_{T_i}\left(F_{(u_i,y_i),T_i}\right)$, where T_i does not contain (U_i, Y_i). As a first partial result, we then obtain the decomposition

$$\begin{aligned} E_T\left(|F_T U_i - Y_i|_{\mathcal{Y}}^2\right) &= E_T\left(|F_T U_i - F_{(U_i,Y_i)} U_i|_{\mathcal{Y}}^2\right) \\ &\quad + E_{(U_i,Y_i)}\left(|F_{(U_i,Y_i)} U_i - Y_i|_{\mathcal{Y}}^2\right) \end{aligned} \tag{2}$$

of the expected squared error at the i-th input into variance of F_T w.r.t. training data T and expected squared error of the predictor $F_{(U_i,Y_i)} U_i$ with variable i-th input for training, because $E_T = E_{(U_i,Y_i)} E_{T_i}$ and the expectation of the middle term in

$$|F_T U_i - Y_i|_{\mathcal{Y}}^2 = |F_T U_i - F_{(U_i,Y_i)} U_i|_{\mathcal{Y}}^2$$

$$- 2\,(F_T U_i - F_{(U_i,Y_i)} U_i)(F_{(U_i,Y_i)} U_i - Y_i) + |F_{(U_i,Y_i)} U_i - Y_i|_{\mathcal{Y}}^2$$

vanishes due to $E_T\left(F_T U_i - F_{(U_i,Y_i)} U_i\right) = 0$. However, in the further decomposition of $E_{(U_i,Y_i)}\left(|F_{(U_i,Y_i)} U_i - Y_i|_{\mathcal{Y}}^2\right)$ the middle term does not vanish: The middle term on the right of

$$\begin{aligned} &E_{(U_i,Y_i)}\left(|F_{(U_i,Y_i)} U_i - Y_i|_{\mathcal{Y}}^2\right) \\ &= E_{(U_i,Y_i)}\left(F_{(U_i,Y_i)} U_i - E(Y_i|U_i)|_{\mathcal{Y}}^2\right) \\ &\quad - 2\,E_{(U_i,Y_i)}\left(\left(F_{(U_i,Y_i)} U_i - E(Y_i|U_i)\right)\left(Y_i \right.\right. \\ &\qquad\qquad \left.\left. - E(Y_i|U_i)\right)\right) \\ &\quad + E_{(U_i,Y_i)}\left(|Y_i - E(Y_i|U_i)|_{\mathcal{Y}}^2\right) \end{aligned} \tag{3}$$

is given by

$$\begin{aligned} &E_{(U_i,Y_i)}\left(\left(F_{(U_i,Y_i)} U_i - E(Y_i|U_i)\right)\left(Y_i - E(Y_i|U_i)\right)\right) \\ &= \int_{\mathcal{U}}\left(\int_{\mathcal{Y}} \left(F_{(u,y)} u - Fu\right)(y - Fu)\, f_{Y|U}(y|u) dy\right) f_U(u) du\,, \end{aligned}$$

and this term could be named expected complexity of F_T. Note that the scalar product inside the integral is positive, if the deviation of the predicted output $F_{(u,y)} u$ from the deterministic output $Fu = E(Y|U = u)$ at u points in the same direction as the deviation of the stochastic output y, i.e. the complexity is positive if $F_{(U_i,Y_i)} U_i$ captures the behaviour of the noise in the output Y_i. Of course, this is an undesired property, thus a machine learning method should be so that complexity of the predictor is low. While complexity has some similarities with covariance and correlation, see e.g. [8], the term is not a covariance or correlation, as $E(Y_i|U_i)$ is not the expectation value of $F_{(U_i,Y_i)} U_i$.

If moreover – like for most machine learning methods – the order of the training data is not important for the estimation of the input-output map, then from (2) and (3) we obtain the following Bias-Variance-Complexity decomposition.

3.1 Main Theorem

Under the assumptions of statistical learning theory (and assuming existence of the expectation values), with the above definitions the decomposition

$$\begin{aligned} &E_T\left(\frac{1}{N}\sum_{i=1}^{N} |F_T U_i - Y_i|_{\mathcal{Y}}^2\right) \\ &+ 2\,E_{(U,Y)}\left(\left(F_{(U,Y)} U - E(Y|U)\right)\left(Y - E(Y|U)\right)\right) \\ &= E_T\left(\frac{1}{N}\sum_{i=1}^{N} |F_T U_i - F_{(U_i,Y_i)} U_i|_{\mathcal{Y}}^2\right) \\ &\quad + E_{(U,Y)}\left(F_{(U,Y)} U - E(Y|U)|_{\mathcal{Y}}^2\right) \\ &\quad + E_{(U,Y)}\left(|Y - E(Y|U)|_{\mathcal{Y}}^2\right) \end{aligned} \tag{4}$$

of the expected mean squared error w.r.t. training data T plus twice the complexity into variance of F_T w.r.t. training data T, squared bias of the predictor $F_{(U,Y)} U$ with one input variable for training, and the variance of Y as an unavoidable irreducible error inherent to the problem holds.

Hence, if the machine learning method is such that on an enhancement of the capacity of the function space $\mathcal{F}$ the mean squared error plus complexity decreases, then variance plus bias decreases simultaneously, too. This seems to be the case for many machine learning methods in the modern interpolation regime and may explain that after the interpolation threshold there occurs a second descent of the error on unseen data, which is related to a decrease of complexity $E_{(U,Y)}\left(\left(F_{(U,Y)}U - E(Y|U)\right)\left(Y - E(Y|U)\right)\right)$.

4. Complexity and LASSO

Complexity is tightly related to LASSO, i.e. L^1-regularized approximation. In fact, if $\mathcal{Y} = \mathbb{R}^k$ and $Y - E(Y|U = u) \sim N(0, \sigma_u^2\, Id)$ is normally distributed, then the inner integral in the definition of complexity allows a partial integration

$$\begin{aligned}
&\frac{1}{\sigma_u (2\pi)^{k/2}} \int_{\mathbb{R}^k} \left(F_{(u,y)}u - Fu\right)(y - Fu)\, \mathrm{e}^{\frac{(y-Fu)^2}{2\sigma_u^2}}\, dy \\
&= \frac{-\sigma_u}{(2\pi)^{k/2}} \int_{\mathbb{R}^k} \left(F_{(u,y)}u - Fu\right)\nabla_y \left(\mathrm{e}^{\frac{(y-Fu)^2}{2\sigma_u^2}}\right) dy \\
&= \frac{\sigma_u}{(2\pi)^{k/2}} \int_{\mathbb{R}^k} \mathrm{div}_y\left(F_{(u,y)}u\right)\mathrm{e}^{\frac{(y-Fu)^2}{2\sigma_u^2}}\, dy\ .
\end{aligned}$$

Thus, complexity is given in this case by

$$\begin{aligned}
&E_{(U,Y)}\left(\left(F_{(U,Y)}U - E(Y|U)\right)\left(Y - E(Y|U)\right)\right) \\
&= E_{(U,Y)}\left(\sigma_U^2\, \mathrm{div}_Y\left(F_{(U,Y)}U\right)\right)
\end{aligned}$$

and the decomposition (4) reads as

$$\begin{aligned}
&E_T\left(\frac{1}{N}\sum_{i=1}^{N} |F_T U_i - Y_i|_{\mathcal{Y}}^2\right) \\
&\qquad\qquad + \ 2\, E_{(U,Y)}\left(\sigma_U^2\, \mathrm{div}_Y\left(F_{(U,Y)}U\right)\right) \\
&= E_T\left(\frac{1}{N}\sum_{i=1}^{N} |F_T U_i - F_{(U_i,Y_i)}U_i|_{\mathcal{Y}}^2\right) \\
&\quad + E_{(U,Y)}\left(F_{(U,Y)}U - E(Y|U)|_{\mathcal{Y}}^2\right) \\
&\quad + E_{(U,Y)}\left(|Y - E(Y|U)|_{\mathcal{Y}}^2\right)
\end{aligned} \tag{5}$$

Hence, under the assumption that the true input-output map $F\ =\ E(Y|U)$ has a derivative $\frac{\partial}{\partial u}Fu =: \gamma_u$ w.r.t. u and the estimated input-output map $F_{(\tilde{u},y)}u$ has approximately the same derivative w.r.t. u at $\tilde{u} = u$, $y = Fu$, i.e. if there are $\sigma_i \approx \sigma_{u_i}^2 \gamma_{u_i}$ such that

$$\sigma_{u_i}^2\, \mathrm{div}_{y_i}\left(F_{(u_i,y_i)}u_i\right) =\ \sigma_i \nabla F_{(\tilde{u},y)}u_i|_{\tilde{u}=u_i, y=y_i}\,,$$

a machine learning method could estimate for training data $t =\ (u_{1,}, y_1), \dots, (u_{N,}, y_N)$ the input-output map by solving

$$\frac{1}{N}\sum_{i=1}^{N} |Fu_i - y_i|_{\mathcal{Y}}^2 + \frac{1}{N}\sum_{i=1}^{N} \sigma_i : \nabla F u_i = \min_{F \in \mathcal{F}}! \tag{6}$$

to simultaneously decrease variance and bias when enhancing the capacity of the function space $\mathcal{F}$. Now assume that σ_i is bounded by a constant C, then the second term is dominated by

$$\frac{C}{N}\sum_{i=1}^{N} |\nabla F u_i|\ .$$

This is exactly the L^1-regularizer in LASSO regression, and if this regularizer is used instead of the second term in (6), then the minimal value dominates the left and right hand side of (5). Hence, under the above assumptions LASSO regression provides reliable predictors in machine learning and leads to a simultaneously decrease of variance and bias when enhancing the capacity of the function space $\mathcal{F}$.

Further, even if no LASSO regression problem is solved, often machine learning methods – like e.g. stochastic gradient descent – generate complexity reduced interpolation models, and thus automatically lead to a descent of the MSE on unseen data after the interpolation threshold. Let us mention in 4.1 one such case, and in 4.2 a case where LASSO is explicitly used.

4.1 Autoencoder neural networks

In the example of an autoencoder neural network in [9] trained by data $y_i = u_i$, where gradient descent is used to minimize mean squared error for a deep neural network in the modern interpolation regime of machine learning, the weights do not converge to the identity but to the matrix W of lowest rank satisfying $Wu_i = u_i$. Hence, MSE is minimal and complexity is reduced, so that variance and bias are simultaneously low and the predictor is nearly the identity on training data.

4.2 Sparse Support Vector Machines

To generate a sparse soft margin support vector machine [10], the constrained optimization problem

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{N} |\xi_i|^2 + C\sum_{j=1}^{D} |\omega_j| = \min_{(\omega,,\xi,\beta)}! \\
&\mathrm{s.t.}\,(\omega \cdot u_i + \beta)y_i\ \geq 1 - \xi_i
\end{aligned}$$

may be solved for data points u_i with given classification $y_i = \pm 1$. The predictor then is the linear classifier $Fu = \omega \cdot u + \beta$, and $\nabla F = \omega$ holds so that above really LASSO is used, i.e. $\sum_{j=1}^{D} |\omega_j|$ is the L^1-regularization term. Therefore, variance plus bias decreases simultaneously when enhancing the capacity of the function space $\mathcal{F}$ by increasing the dimension D.

Acknowledgements

The first author gratefully acknowledges the financial support of the German Federal Ministry for Economic Affairs and Energy within the ZIM project 16KN072420. The second author gratefully acknowledges the financial support of the Saxonian Ministry of Science and Arts as well as the European Union within the research project SmartKMU (Project No. 100 38 21 59).

Bibliography

[1] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[2] S. Geman, E. Bienenstock, R. Doursat, Neural computation 4 (1992), 1-58.

[3] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, M. Wyart, Journal of Physics A: Mathematical and Theoretical 52 (2019), 474001.

[4] M. Belkin, D. Hsu, S. Ma, S. Mandal, Proceedings of the National Academy of Sciences 116 (2019), 15849-15854.

[5] B. Adlam, J. Pennington, arXiv:2011.03321 (2020).

[6] B. Ghojogh, M. Crowley, arXiv:1905.12787 (2019).

[7] T. Hastie, R. Tibshirani, M. Wainwright, Statistical learning with sparsity: the lasso and generalizations, CRC press, 2015.

[8] J. Merker, G. Schuldt, Proceedings of ICoMS 2020, ACM (2020), DOI: 10.1145/3409915.3409920

[9] A. Radhakrishnan, M. Belkin, C. Uhler, Proceedings of the National Academy of Sciences 117 (2020), 27162-27170.

[10] J. Merker, Journal of Advances in Applied Mathematics 2 (2017), 109-114.