Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 6 Issue 1 (2026)

Classic bias-variance trade-off in modern statistical learning context: a position paper and theoretical review

Saurabh Rawal
Decision Sciences Area, Indian Institute of Management Lucknow, India Email: phd21006@iiml.ac.in

Abstract:

Classical statistical learning theory suggests that models learn data intricacies by fitting a parsimonious model, which can lead to irreducible error components. Overfitting occurs when a model fits the data so well that it becomes too good to be true, and when assessed for unseen data, it performs poorly. The bias-variance trade-off deals with balancing complexity and generalization error. Increasing the number of parameters in models increases the chances of poorly sampling in specific directions, leading to higher variance. This phenomenon is known as benign overfitting. This paper presents a position paper and brief theoretical review that synthesizes key analytical results from the growing literature on benign overfitting. It focuses on overparameterized linear regression analysis, which has two major benefits: decreasing the likelihood of overfitting and uncovering hidden trends in data. However, the generalization bounds for overparameterized models do not explain important empirical observations, and the case of dataset shift remains unexplored in this regime. Rather than proposing a new algorithm or empirical method, the paper aims to clarify conceptual mechanisms underlying benign overfitting and to highlight limitations of current theoretical explanations, particularly under dataset shift. The paper concludes by identifying open theoretical questions relevant to the foundational understanding of modern machine learning systems.

Keywords: benign overfitting; interpolation; generalization error; double descent; statistical learning; conceptual review

Introduction:

Memorization of data has not been suggested in classical statistical learning theory. Instead, the effort is made to fit a parsimonious model (in accordance with Occam's razor). There is a certain amount of inherent noise that leads to an irreducible error component, which is always present in the data. The irreducible error component can be present due to several possible reasons, such as unmeasured latent variables, unmeasurable variation, or imprecision in data recording. Therefore, it is suggested in the extant literature that a model that learns the data intricacies to the also be learning noise ilearnss, and hence this practice must be avoided (James, Witten, Hastie & Tibshirani, 2017, The term overfitting is used to describe this situation where the model fits the data so perfectly that it becomes too good to be true, and when such a model is assessed for unseen data, it performs poorly. Moreover, learning is an ill-posed problem, and it is not possible to do without inductive bias (due to the assumptions made in the model). Thus, the fundamental aim of dividing the original dataset, fitting the model on its major portion (training set), and assessing its performance on the remaining part of the data (validation or test set to measure the quality of the inductive bias) is lost. The model fails to generalize well in the unseen data. Balancing the amount of complexity and generalization error in the model is known as the bias-variance tradeoff (or the approximation-generalization trade-off). The situation is sometimes referred to as the triple trade-off (between the class complexity, the amount of training data, and the generalization error) (Alpaydin., 2016).

The purpose of this paper is not to propose a new algorithm or estimator. Instead, this article is intended as a position paper and theoretical review that synthesizes and organizes existing analytical results on benign overfitting in overparameterized linear regression models. The objective is to clarify the conceptual mechanisms underlying this phenomenon and to reconcile different analytical perspectives presented in the literature.

From an informatics perspective, understanding benign overfitting is essential for interpreting how modern learning systems process information in high-dimensional settings. Clarifying the theoretical foundations of generalization in overparameterized models contributes to a deeper conceptual understanding of machine learning systems which are increasingly utilized in educational, scientific, and decision-support contexts.

Literature Review:

Different notions of measuring model complexity exist. Some of the frequently used metrics are as follows:

1. Vapnik–Chervonenkis complexity
2. Condition number of the regression matrix
3. Number of parameters
4. The number of neighbours averaged in the nearest neighbour estimator
5. The scale of estimate in a Reproducing Kernel Hilbert Space (RKHS)

In addition to the above measures, the *predicted R-squared* is used to check the generalizability of the linear regression model by detecting overfitting (Montgomery, Peck, and Vining, 2014). The statistic is calculated using the cross-validation approach. The value of the predicted R-squared is compared with that of the R-squared, and if the former is significantly less than the latter, it indicates overfitting. The predicted R-squared is calculated as follows:

$$\text{Predicted R-squared}=\left[1\text{-}\left(\frac{\text{PRESS}}{\text{total sum of squares}}\right)\right]\times 100$$

Decomposition of the generalization error

The generalization error, measured as the mean squared error (or MSE) can be decomposed into three components. The test error is averaged over all possible test sets. The components are:

1. Variance
2. Squared bias
3. Noise

Let y be the observed response and $g(x)$ be the predicted response.

MSE = $E[(y - g(x))^2]$

$= E(g(x)^2) - 2E(g(x))\ E(y) + E(y^2)$

$= Var(g(x)) + E(g(x))^2 - 2E(g(x))f(x) + Var(y) + f(x)^2$

$= Var(g(x)) + [E(g(x)) - f(x)]^2 + \sigma^2$

Here, in the above equation, the MSE has been decomposed into three terms: variance, squared-bias, and noise, respectively.

Eigenvalue distribution

To understand the phenomenon of double-descent, we need to explore the expected eigenvalue distribution (*Marchenko-Pastur distribution*) of the covariance matrix ∑. The covariance matrix in overparameterized regime is dependent only on its eigenvalues.

The underparameterized regime is characterized by a finite gap in eigenvalue spectrum with no small eigenvalue at the interpolation threshold, the minimum eigenvalue becomes zero. The overparameterized regime is characterized by finite gap in the eigenvalue spectrum with many zero eigenvalues (indicating unimportant directions).

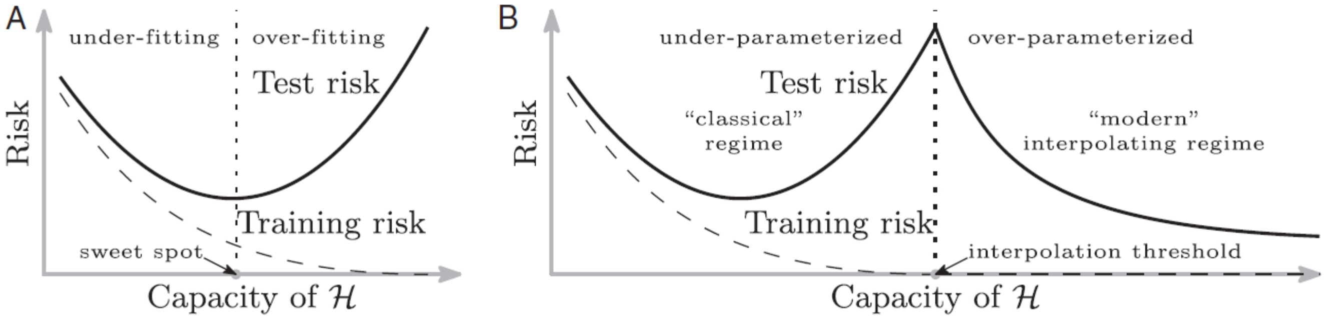
In the infinite-dimensional space, there is a very narrow range of decay rates of eigenvalues of the covariance matrix, corresponding to benign overfitting. In contrast, in the case of a finite (but high) dimensional space, this range of decay rates widens.

Interpolation

Interpolation refers to the situation where the training error becomes zero. The specific point at which it occurs is referred to as the interpolation threshold. It is located at a point where the number of principal components equals the number of data points. In the underparameterized regime, interpolation is the same as overfitting.

Increasing the number of parameters leads to sampling in a greater number of directions. However, it increases the chances of poorly sampling in any specific direction, leading to higher variance. It is the point of interpolation threshold where the generalization error begins diverging due to an increase in variance. Interpolation in overparameterized models is also known as *benign overfitting* (the term highlighting the contrast to the ill-effects of overfitting). Divergence in generalization error occurs due to small eigenvalues (corresponding to the poorly sampled directions in feature space) of the Hessian matrix.

On the contrary, when we move beyond the interpolation threshold, as more parameters are added to the model, it reduces the effects of overfitting by reducing variance, as it allows for richer sampling in the directions captured by the training data.



Source of figure: Belkin, M., Hsu, D., Ma, S. and Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32), pp.15849-15854

In the above figure A, training and test risks have been plotted against the capacity of H (which represents the complexity of the model). By increasing the class capacity, we aim to expand the hypotheses that the model can learn. The training risk curve (represented by a dotted line) indicates that training risk decreases monotonically and ultimately approaches zero. On the other hand, the test risk curve decreases (in the under-fitting regime), but beyond a specific point, it starts increasing (in the over-fitting regime), thereby forming its U-shape. The specific point at which the test risk reaches its minimum is the sweet spot that we aim to achieve in classical statistical learning models. In a way, it summarizes the process of balancing bias and variance in the test error.

In the adjacent figure B, we can see that it is the extension of figure A, where we are increasing the capacity of H further. Eventually, we reach a specific point, called the interpolation threshold, where the training error is zero. Increasing the model capacity even further brings the model into the "modern interpolating regime", where we observe another descent in the test risk curve. Beyond the interpolation threshold, we fit the overparameterized regime (where the number of parameters exceeds the number of data points).

Bias

Bias in MSE is due to wrong assumptions which reduces the ability of the model to learn patterns. There are two sources of bias in the overparameterized regime: the model is incapable of fully capturing the full data distribution, and incomplete sampling of the data's feature space. In the underparameterized regime, the bias remains constant as we approach the interpolation threshold. The conventional definition of bias (pertaining to incorrect assumptions) is not suitable for the overparameterized regime. Contrary to what the existing literature suggests, Rocks J. & Mehta P. (2022) establish that the bias never diverges; it remains finite and decreases monotonically, regardless of the regularization in the model. Non-zero bias may exist if training data spans only a part of the feature space. In this case, the variance component will remain non-zero in the interpolating regime as well.

In the overparameterized regime, bias occurs due to two reasons: the model is unable to fully capture the data distribution, and there is incomplete sampling of the feature space.

Variance

Variance in the model represents the oversensitivity of the model to the training data such that it learns the quirks of the data which are actually noise. In linear regression, the eigenvalue of the empirical covariance $X^T X/M$ represents the empirical variance along each associated principal direction. In classical statistical learning, the overfitting is chiefly attributed to increase in variance. When we enter the overparametrized regime, in the presence of noise, it is not only the variance but also the bias that may lead to overfitting. In other words, in the case of non-zero bias, the model may treat part of the training data as noise, as it is unable to express the underlying data distribution fully. In the overparameterized regime, increasing the number of parameters leads to solutions with smaller norms (which act as the inductive bias). This results in decreasing the variance beyond the point of interpolation threshold.

Studies have attempted (Muthukumar, Vodrahalli & Sahai, 2019) explaining the benign overfitting in the presence of noise in the data (known as *favourable noise-fitting*, which is a novel concept). The lower bound on generalization error in the overparameterized regime (number of parameters (d) >> number of cases(n)) is

$$\varepsilon^*_{test}=\mathbf{W}^T_{train}(B_{train}B^T_{train})^{-1}\mathbf{W}_{train}+\sigma^2$$

where **W** = Gaussian noise (N(0, σ^2)), B_{train} = $A_{train}\Sigma^{-1}$, A_{train} = $[a(X_1)^T \quad a(X_2)^T \quad \dots \ a(X_n)^T]^T$.

Favourable noise-fitting represents the scenario in which fitting the noise in an overparameterized regime gives an interpolative solution such that the effect of fitting noise on generalization error decays to zero as the number of features in the model approaches infinity. Few studies have used the minimum norm interpolating method to investigate the, otherwise underdetermined model, in context of benign overfitting (Bartlett, Long, Lugosi & Tsigler, 2020). Additionally, the bounds for the excess prediction error (or test error or generalization error) hold for arbitrary finite dimensional spaces and finite sample sizes. The unique minimum norm interpolating solution can be obtained by calculating the pseudoinverse of the matrix.

The minimum norm interpolating solution can be obtained by solving the following optimization problem

$$min_{\theta\in\mathbb{H}} \|\theta\|^2 \ such\ that \|X\theta-y\|^2=min_{\beta}\|X\beta-y\|^2$$

Estimated parameter $(\hat{\theta})=X^T(XX^T)^{\dagger}\boldsymbol{y}$

(Bartlett, Long, Lugosi & Tsigler, 2020) Two notions of *effective ranks* have been computed to put bounds on the generalization error. If $\sum_{i=1}^{\infty}\lambda_i<\infty$ and $\lambda_{k+1}>0$, for $k\geq0$

$$r_k(\Sigma)=\frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}},\ R_k(\Sigma)=\frac{(\sum_{i>k}\lambda_i)^2}{\sum_{i>k}\lambda_i^2}$$

where λ_i denotes the eigenvalues of the covariance matrix Σ such that $\lambda_1\geq\lambda_2\geq\lambda_i\geq\ldots$

Here, $r_k(\sum)$ denotes the effective rank of the subspace with the highest k eigenvalues removed. If all the eigenvalues are identical, then both notions of effective ranks are equal to the rank of ∑. Generally speaking, the large eigenvalues of Σ denote the directions which are critical for prediction, and the error made in these directions is particularly detrimental for the overall generalization error. For benign overfitting, $r_0(\Sigma)$ should be smaller than the sample size(n) and $r_{k*}(\Sigma)$ *and* $R_{k*}(\Sigma)$ should be larger than n. Therefore, the number of non-zero eigenvalues should be larger than the sample size n. In other words, the number of unimportant directions must be much, much greater than the sample size (so that the label noise gets distributed among these directions) to observe the double descent in the generalization risk curve.

We can write the upper bound on the generalization error in terms of these effective ranks:

With $\lambda_n>0$, if k^* = min $\{k\geq0{:}r_k(\Sigma)\geq bn\}$, then with high probability:

$$R(\hat{\theta})\leq c\left(\|\theta^*\|^2\sqrt{\frac{tr(\Sigma)}{n}}+\sigma^2\left(\frac{k^*}{n}\right)+\frac{n}{R_{k^*}(\Sigma)}\right)$$

σ_{min}^2 is the minimum non-zero eigenvalue of the Hessian. At the interpolation threshold, it approaches zero. The underparameterized regime is characterized by finite gap in the eigenvalue spectrum with no small eigenvalues. On the other hand, the overparameterized regime is characterized by a finite gap in the eigenvalue spectrum with many zero eigenvalues (indicating the unimportant directions).

In the case of a minimum norm interpolating solution, the generalization error can be decomposed into noise and noiseless cases as follows:

$$\varepsilon_{test}(\hat{\alpha})-\sigma^2\leq\varepsilon_{test}(\hat{\alpha};\ noiseless)+\varepsilon_{test}(\hat{\alpha};noise)$$

where $\varepsilon_{test}(\hat{\alpha};noiseless)=\left\|\Sigma^{1/2}(A_{train}^{\dagger}A_{train}-I)\alpha^*\right\|_2^2$ and $\varepsilon_{test}(\hat{\alpha};noise)=W_{train}^T(A_{train}^{\dagger})^T\Sigma A_{train}^{\dagger}W_{train}$

The benign overfitting is also looked into the domain of separable Hilbert space (which is an infinite-dimensional vector space with an inner product operator defined in it). The covariance structure ∑ and optimal parameters β^* play a crucial role in linear regression.

It has been highlighted that in the infinite-dimensional space, there is a very narrow range of decay rates of eigenvalues of the covariance matrix, which corresponds to benign overfitting. In contrast, in the case of a finite (but high) dimensional space, this range of decay rates widens. (Hastie, Montanari, Rosset & Tibshirani, 2022) While in low dimensions, a positive regularization is necessary to achieve good interpolation, in certain high-dimensional settings, interpolation can be nearly optimal. To select one solution from the overparameterized model, two approaches have been considered: min-norm (or minimum l_2 norm) regression and ridge regression. Since X matrix is of full rank in the overparameterized model, the min-norm estimator is an interpolator. Considered cases of isotropic features (where the covariance matrix is the identity matrix) and latent space features model (where the covariance matrix = WW^T + I, where W $\in \mathbb{R}^{p \times d}$, $d << p$ and β lies in the span of columns of W).

By making a suitable choice of covariance matrix ∑ and coefficients β, we can achieve a minimum generalization error in the overparameterized regime, and this result will be robust to the distribution of (y_i, x_i).

In the case of isotropic features, asymptotic generalization error depends only on the norm $\|\beta^2\|_2^2$ or signal-to-noise ratio ($\|\beta^2\|_2^2/\sigma^2$). The global minimum of the generalization error is achieved in the underparameterized (overparameterized) regime for a well-specified model (a misspecified model). On the other hand, for the anisotropic features, the risk depends on the alignment of β with the eigenvectors of ∑.

In the overparameterized regime, the min-norm estimate of β is constrained to lie on the row space of X, which is a subspace of dimension d. With an increase in dimensions, the bias continues to increase since this row space accounts for a lesser part of the feature space. Whereas the variance decreases with an increase in γ since the min-norm solution will have decreasing l_2 norm, thereby leading to the second descent in the generalization error curve.

While these studies have significantly advanced theoretical understanding, the literature remains fragmented across different analytical perspectives. In particular, conceptual connections between bias–variance behavior, eigenvalue decay, and generalization under interpolation are often treated separately. This paper addresses this gap by providing an integrated conceptual synthesis.

Methodology:

This paper adopts a conceptual and analytical review methodology. This methodological approach is appropriate given the paper's objective of conceptual clarification and theoretical synthesis. By integrating results across multiple analytical frameworks, the paper aims to provide a coherent understanding of benign overfitting without introducing new empirical claims.

Discussion and Conclusion:

This paper has presented a position paper and a theoretical review of benign overfitting in overparameterized linear regression. By synthesizing existing analytical results, the paper provides conceptual clarity on how interpolation, eigenvalue structure, and implicit regularization interact to enable generalization.

References:

1. Alpaydin. (2016). *Introduction to Machine Learning*. [S.l.]: PHI learning.
2. Bartlett, P., Long, P., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings Of The National Academy Of Sciences*, *117*(48), 30063-30070. doi: 10.1073/pnas.1907378117
3. Belkin, M., Hsu, D., Ma, S. and Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), pp.15849-15854.
4. Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, *50*(2), 949-986. doi: 10.1214/21-aos2133
5. James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2017). *An Introduction to Statistical Learning: With Applications in R*. Springer.

6. Montgomery, D., Peck, E., & Vining, G. (2014). *Introduction to Linear Regression Analysis* (3rd ed., pp. 152-154). Wiley India Pvt. Ltd.
7. Muthukumar, V., Vodrahalli, K., & Sahai, A. (2019). Harmless interpolation of noisy data in regression. *2019 IEEE International Symposium On Information Theory (ISIT)*. doi: 10.1109/isit.2019.8849614
8. Rocks, J., & Mehta, P. (2022). Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, *4*(1). doi: 10.1103/physrevresearch.4.013201