Double Descent Experiments

November 2019

1 Overview

Broadly, I have conducted three experiments which are summarised as follows:

1. Reproducing the double descent experiment for misspecified model assuming Gaussian distribution from the paper 'Two models of double descent for weak features'
2. Bias-Variance analysis for the above model
3. Empirically observing the variation of signal bleed with the spectrum of the covariance matrix

2 Double Descent in misspecified model

The model is assumed to be $Y = \sum_{i=1}^{D} \beta_i x_i$, where each x_i is distributed according to $N(0, 1)$. The number pf data points are kept at a constant 200 and the number of features included in the model are increased starting from 1 till D. The test error for minimum-l_2-norm interpolator is plotted against the number of features included for various choices of β_i.

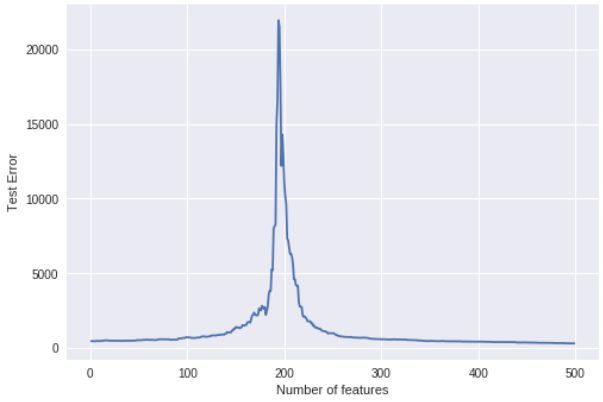

Figure 1: β sampled from $N(0, 1)$

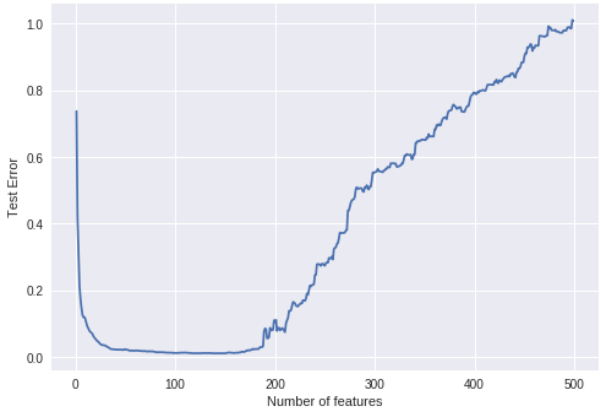

Figure 2: $\beta_t = \frac{1}{t}$

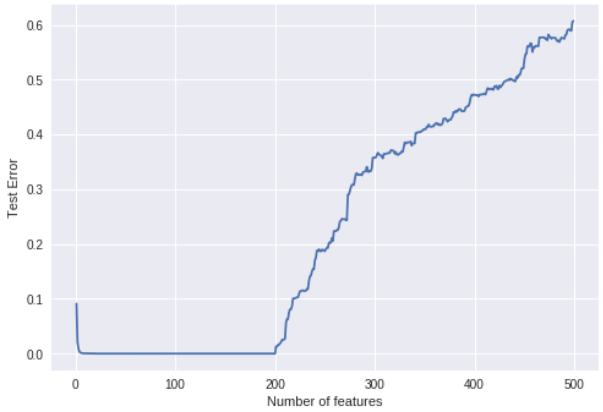

Figure 3: $\beta_t = \frac{1}{t^2}$

When I conducted above experiments, I always thought that bias must be going down as we include more features and the phenomenon of double descent must be dependent on variance. For verifying the same, I conducted the experiment below:

3 Bias-Variance analysis

Following the same setup as above, in this case, 20 different draws of training samples are used to estimate the bias and variance in the case above. The plots obtained are surprising:

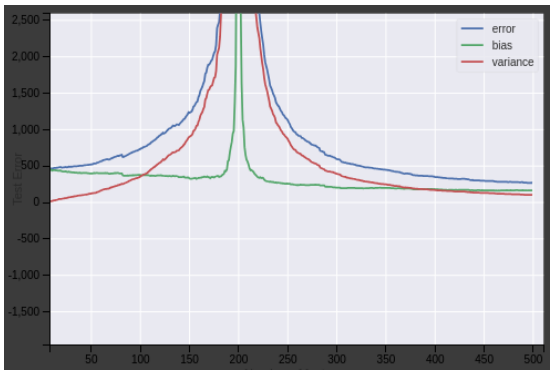

Figure 4: β sampled from $N(0, 1)$

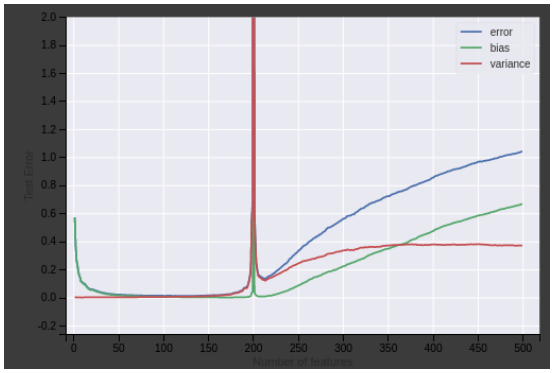

Figure 5: $\beta_t = \frac{1}{t}$

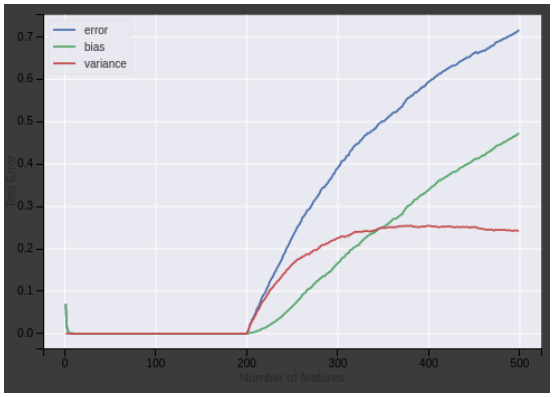

Figure 6: $\beta_t = \frac{1}{t^2}$

The test error in case of $\beta_t = \frac{1}{t}$ and $\beta_t = \frac{1}{t^2}$ goes up due to increase in bias. In the paper 'Surprises in high dimensional linear regression', I came across an intuitive reason for this. Basically, the minimum-l_2-norm solution is constrained to lie the row space of X_T. Thus, it always lies in a subspace of dimension n against increasing p. This is why the bias goes up in prescient cases. However, in $N(0,1)$ case as the new features added are as important as the already existing ones, the increase in bias due to the increase in p is compensated by addition of more features.

4 Variation of signal bleed with spectrum of the covariance matrix

The paper 'Harmless interpolation of noisy data in regression' stated that eigenvectors with smaller eigenvalues are impacted by much worse signal bleed in overparameterized regime. To verify this, I generated a random training data matrix with 500 features having the spectrum of covariance matrix equally spaced between 1 and 100. The generative model for Y is assumed to be $Y = \langle x, \alpha^* \rangle + W$, where W is sampled from $N(0, 0.1)$. α^* is set to the smallest eigenvector, 100th eigenvector and so on till the top eigenvector. For each of these cases, the $\hat{\alpha}$ obtained using minimum-l_2-norm interpolation is projected along the different eigenvector directions. Ideally, we expect this distribution to concentrate on the eigenvector it came from. But, the results are shown below:

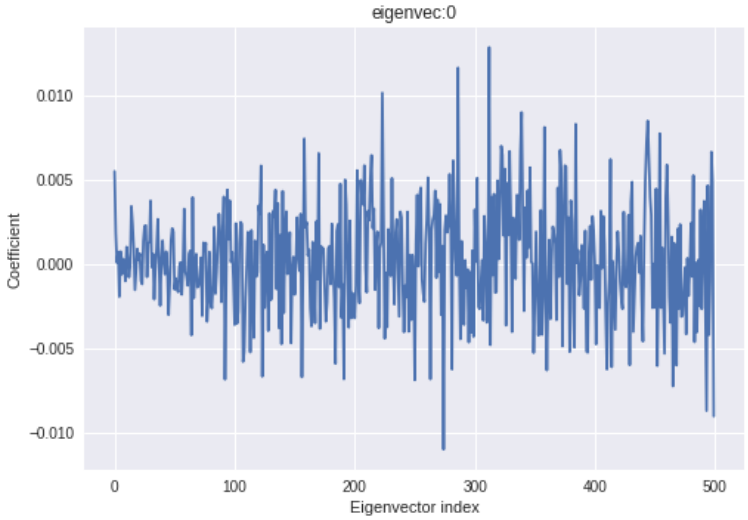

Figure 7: Eigenvector with eigenvalue 1

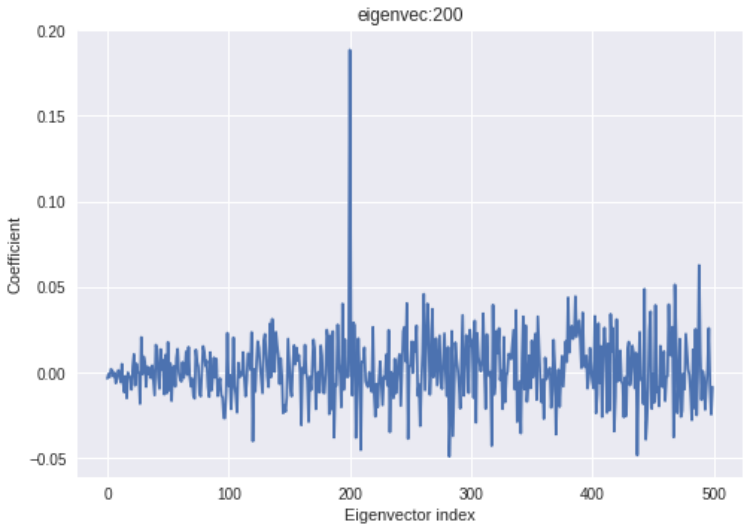

Figure 8: Eigenvector with eigenvalue 40

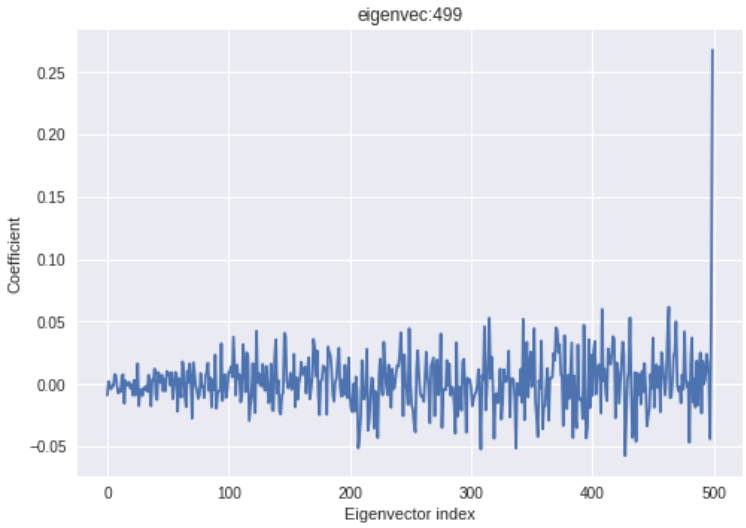

Figure 9: Eigenvector with eigenvalue 100

As expected, the coefficient bleed decreases as the eigenvalue of the vector along which α^* is concentrated goes up. However, this doesn't necessarily imply that the prediction error will go down as well, because

$$E[x^T(\hat{\alpha} - \alpha^*)]^2 = (\hat{\alpha} - \alpha^*)^T \Sigma (\hat{\alpha} - \alpha^*)$$

Thus, any error along the higher eigenvalue direction contributes more to the total error, as compared to a lower eigenvalue direction. Thus, if the signal is concentrated along a higher eigenvalue direction, even though it will bleed less, but even a minimal bleed will contribute higher to the prediction error. To study the variation of test error, same experiment was repeated with varying number of data points, however another modification was necessary.
Consider $Y = XQ\gamma$, where $\Sigma = Q\Lambda Q^T$ represents the eigenvalue decomposition of population covariance matrix.

$$||Y||^2 = \gamma^T Q^T X^T X Q \gamma$$

Now, Although $X^T X$ may not be exactly equal to the population covariance matrix, but it can be approximated as $Q\Lambda Q^T$. Thus, $||Y||^2$ is on the order of $\gamma^T \Lambda \gamma$. Now, to keep the scale of Y to be similar for different γ, they are scaled using $\frac{1}{\sqrt{\Lambda}}$. This is the factor of $\sqrt{\frac{5}{j+5}}$ in the code. the result obtained is shown below:

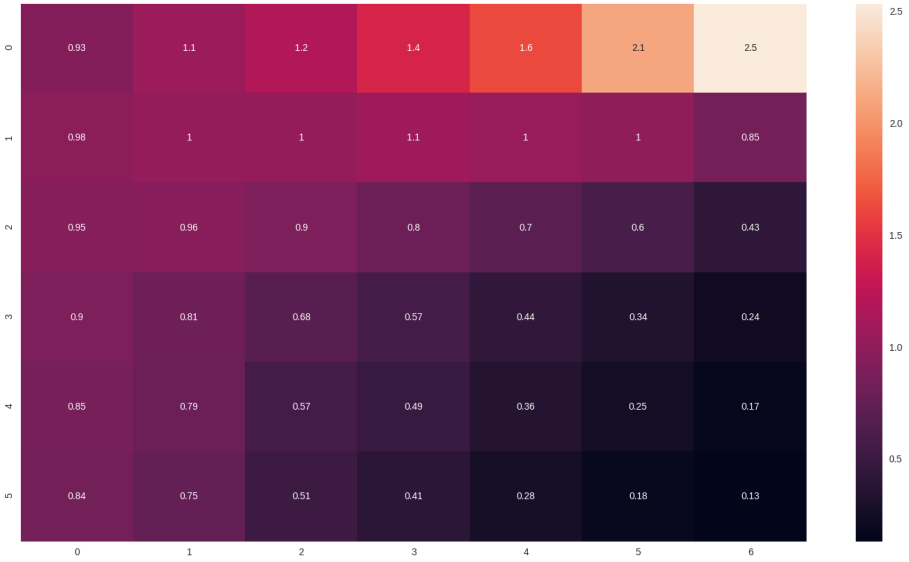

Figure 10: Variation of test accuracy with datapoints and eigenvalue (Eigenvalues increase towards the bottom and datapoints increase towards the right)

Thus, we can see that the error goes down with increasing data points as well as increasing eigenvalue.
Although initially I thought this would be a trivial statement to prove, turns out its not as trivial. Basically, we know that minimum-l_2-norm solution always lies in row space of X. Thus, we need to show that the row space of X is likely to comprise of the larger eigenvectors of X, which should involve random matrix theory.