Double Descent and Over-Parameterization

Alex Wang, Victor Wei, Gregory Sheyn

Abstract

Despite being unwanted in classical machine learning models, over-parameterization in deep learning is powerful. Models with a lot of parameters not only fit the training data perfectly, but also generalize well, contrary to our intuition from traditional bias-variance tradeoff; people named it “Double Descent”. In this project, we reproduce results of a paper on the phenomenon of Double Descent in simple Linear Regression models and the effects of applying PCA (principal component analysis) in this context. Furthermore, we extend authors’ work by investigating the effects of different covariance structures that were not considered.

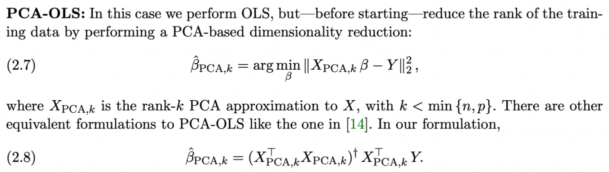
Introduction

Overparameterization has always been an unwanted state of model in classical machine learning, where various tools and a lot of effort were put into fixing the issue of over-fitting. For example, in this course, we have spent a lot of time studying common regularization techniques such as Ridge and Lasso.

Things change when we step into the world of deep learning, where we rely on very large and deep neural network models to perform useful tasks that could even surpass human performance. At the same time, to many people’s surprise, one observed phenomenon was that overparameterization in deep neural networks could sometimes improve the overall performance, contrary to the traditional Bias-Variance Tradeoff concept from classical machine learning [1]. To better understand this phenomenon, in which people named “Double Descent”, researchers started to reproduce the Double Descent in even simpler models, such as a two-layer simple neural network [2]; or even simpler, linear regression model [3].

Despite Double Descent occurring in the modern deep learning models, the same phenomenon existing in the old-school simple classical models like linear regression was not expected by intuition. Hastie et al. at Stanford did a very detailed analysis on Double Descent in linear regression models, and argued that the asymptotic bounds they derived have intimate connection with the modern sophisticated neural network models [3]. The paper we choose to reproduce in this project is a continuation of the same effort, namely investigating Double Descent in simpler models and hoping to find connection with nowadays deep learning models. In particular, Huang et al. (authors of our chosen paper) investigated how PCA (principal component analysis) is able to alter the Double Descent curve and possibly find a solution without overparameterization. In this paper, we reproduce their results first, then we investigate the effects of different covariance structure and SNR (signal-noise-ratio).

Model Description and Datasets



PCA-OLS: In this case we perform OLS, but—before starting—reduce the rank of the training data by performing a PCA-based dimensionality reduction:

$$\hat{\beta}_{\mathrm{PCA},k} = \arg\min_{\beta} \|X_{\mathrm{PCA},k}\,\beta - Y\|_2^2\,, \tag{2.7}$$

where $X_{\mathrm{PCA},k}$ is the rank-k PCA approximation to X, with $k < \min\{n,p\}$. There are other equivalent formulations to PCA-OLS like the one in [14]. In our formulation,

$$\hat{\beta}_{\mathrm{PCA},k} = (X_{\mathrm{PCA},k}^{\top} X_{\mathrm{PCA},k})^{\dagger}\, X_{\mathrm{PCA},k}^{\top}\, Y. \tag{2.8}$$

Figure 1: Mathematical formulation of PCA-OLS, the focus of the paper.

In the chosen paper, the authors generated datasets consisting of n data points (x,y) from a joint Gaussian distribution, with or without noise. Various experiments were then carried out using different models, on datasets with or without noise. The noise is also a Gaussian distribution added to the y data points, specified by the variance and can be manipulated using SNR defined to be the norm of the model coefficient divided by the variance of the Gaussian noise.

For the various models, OLS (ordinary least square) refers to the vanilla version of linear regression from course lectures, note that for p>n (number of parameters > number of samples), we take the pseudo-inverse to ensure a unique solution. Ridge or ridge regression is also identical to the one from course lectures, where we add a L2 penalty term.

PCA-OLS also known as PCR is where we perform a PCA-based dimensionality reduction procedure before the actual OLS fitting. This is also known as PCR (principal component regression) and is different from its variant oracle-PCA, where the true population covariance is known and the projection is onto the k principal components of the population covariance. See Figure 1 for the mathematical formulation.

Another model that will appear in the authors' comparison plot is PLS (partial least square), where similar to PCA-OLS, it is a dimensionality reduction technique not only maximizes the variance of the projected features, but also the covariance of the projected response and projected features.

Finally, the authors also introduced the Generative and latent-variable models, where there is an extra layer/matrix of latent variables that can map/project the input data to higher or lower dimensions, before the actual model fitting step. We classify the projection matrices into data-dependent and data-independent. For example, PCA-OLS involves a data-dependent projection into lower dimensions.

Results

(Reproduce Results)

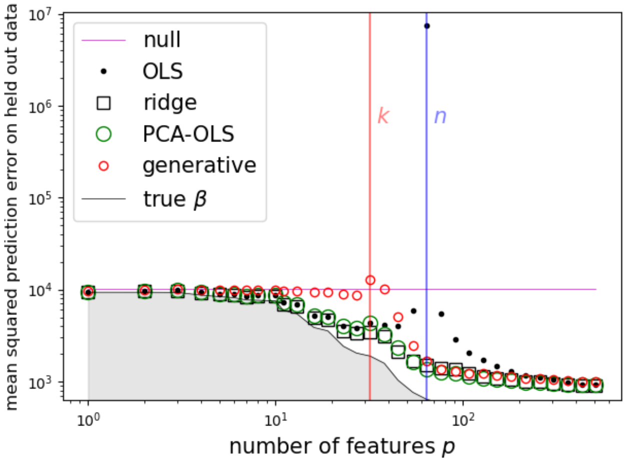


Figure 2: MSE of prediction of y on datasets without additional Gaussian noise. Notice that the number of features are varying through a different projection matrix. Null estimator refers to the case of beta = 0, and PCA-OLS performed on a dataset reduced to dimension k = 32.

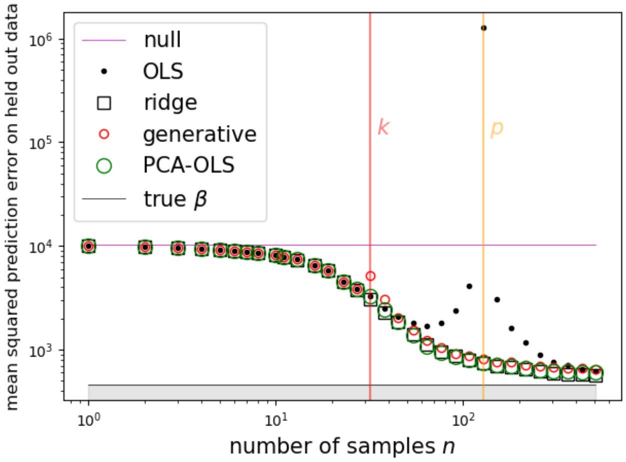


Figure 3: MSE of prediction of y on datasets without additional Gaussian noise. Notice that the number of samples are varying through a different projection matrix. Null estimator refers to the case of beta = 0, and PCA-OLS performed on a dataset reduced to dimension k = 32.

As shown in Figure 2 and 3, we indeed observe the expected Double Descent phenomenon in unregularized models OLS and the generative model. In Figure 2, we varied p, the number of model parameters, as a way to vary model complexity, and indeed observed a peak in generalization error around n=p, or k=p for the generative model. Similarly, in Figure 3 we varied the number of samples, and again observed such peaks in generalization error in around the same region.

On the other hand, the regularized models, including ridge and PCA-OLS, do not exhibit Double Descent peaks and have a monotonically decreasing generalization error. While this phenomenon with ridge regression has already been carefully studied by Hastie et al. [3], and could be explained in terms of the condition number of the inverted matrix, the same effect achieved by PCA-OLS is brand new. This sheds light onto the potential benefits of PCA in more sophisticated deep learning models, where Double Descent peaks could be avoided.

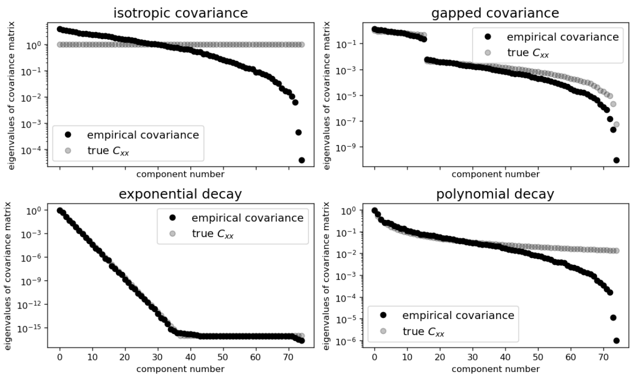


Figure 4: Different covariance structures in the experiment with Gaussian noise added. Both true and empirical covariance structures are plotted.

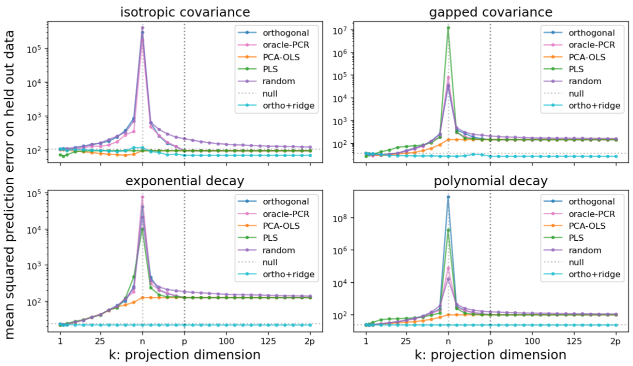


Figure 5: Double Descent plots with high SNR = 16 for various models, with Gaussian noise added.

As shown in Figure 4 and 5, Double Descent curves were plotted for various models with different covariance structure and SNR values (for more SNR values, see the notebooks). Here the number of samples n=50 and number of parameters p=75 are fixed, and we vary the projection dimension instead to increase model complexity. In Figure 4, isotropic covariance structure refers to Cxx being the identity matrix; gapped covariance corresponds to a planted eigenvalue gap at component #16; exponential and polynomial decay correspond to the two decay patterns of the eigenvalues. All largest eigenvalues was chosen to be 1. Note that in the plots of Figure 4, we see a sudden drop of eigenvalues after component 50, around the number of samples. This was expected, as the number of samples give an upper limit to the number of "directions" we can have in high-dimensional space.

Figure 5 shows the familiar Double Descent curves for unregularized models, and flatter curves for PCA-OLS and ortho-ridge (variant of ridge). The performances were evaluated by averaging over 10 trials with 256 test samples. Ridge regularization yielded the best performance overall, followed by PCA-OLS; but keep in mind that PCA can offer some unique advantages for specific problems of interest, such as feature selection issues that require dimensionality reduction. We also see that with or without regularization, the solution in the overparameterization regime may not be the best one, as one can easily observe in the cases of gapped and decaying covariance structure.

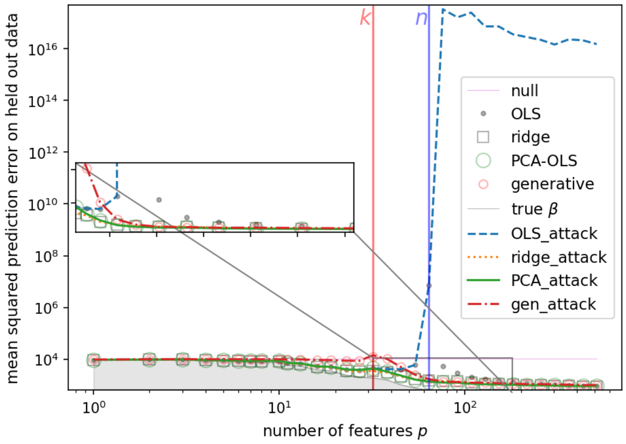


Figure 6: Double Descent plots with datasets under data-poisoning attacks of magnitude epsilon=1, with fixed number of samples n=64 and varying p.

The last piece of significant results from the authors is resilience of PCA-OLS from data-poisoning. As shown in Figure 6, while the unregularized OLS suffers from the poisoned data in the overparameterized regime, we see that PCA-OLS performs extremely well in any regions, especially the over-parameterized regime, as good as the ridge-regularized curve.

(New Results)

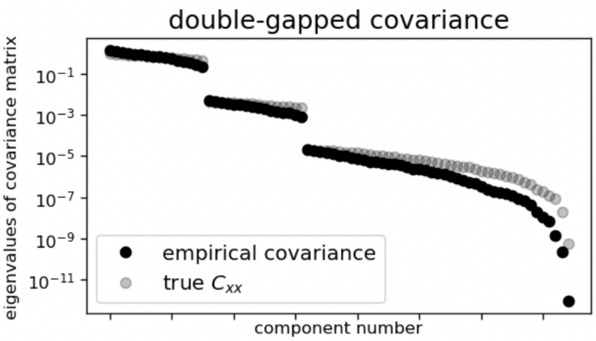


Figure 7: Double-gapped covariance structure. Note that here the empirical covariance structure does not necessarily have n=50 samples, so the tail looks

non-flat. This figure is shown here for a visualization of double-gapped covariance structure.

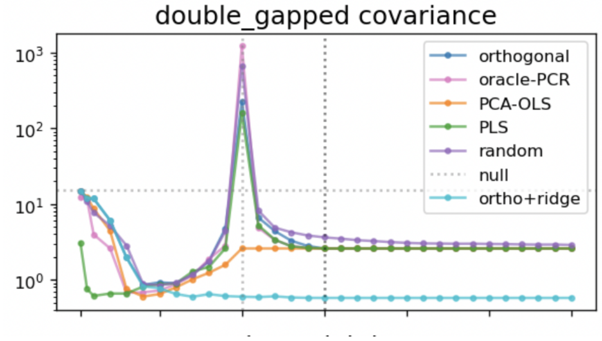


Figure 8: Double Descent plots with high SNR = 16 for double-gapped covariance structure, with Gaussian noise added.

As shown in Figure 7, we add a new double-gapped covariance structure and re-run all the experiments. In Figure 8, we plotted the Double Descent curve with SNR=16, so we can directly compare it with the other four covariance structure plots in Figure 5. We see that adding an extra gap to the covariance structure has pretty much no effects on the Double Descent curve, after comparing Figure 8 with the single-gapped version in Figure 5. This is expected, as the second half of the covariance structure plot plays minimal roles, so reducing their weights on a log-scale should not affect the overall trend.

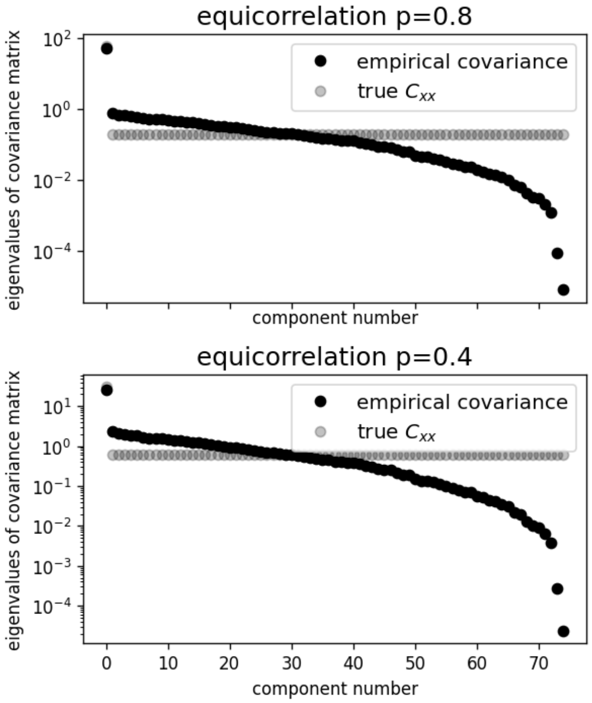


Figure 8: True and empirical covariance plots under high SNR = 16 for equicorrelation covariance structure, with Gaussian noise added.

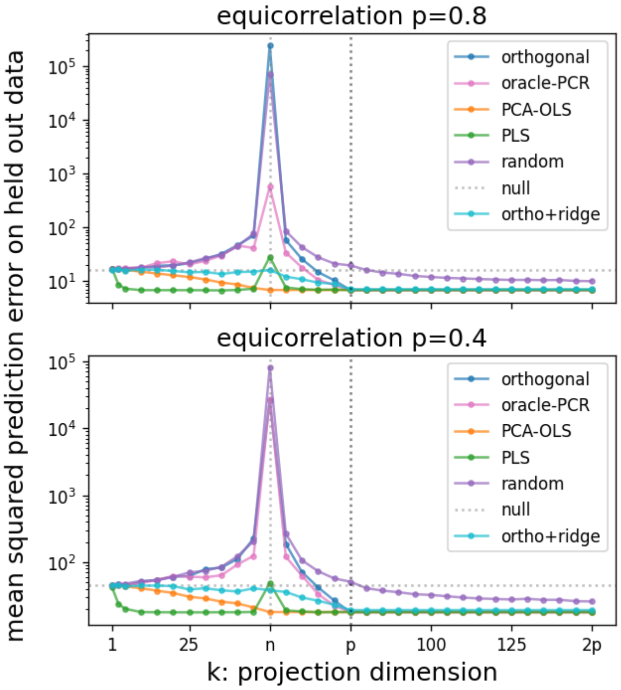


Figure 9: Double Descent plots with high SNR = 16 for equicorrelation covariance structure, with Gaussian noise added.

We also tried equicorrelation matrices, which are Toeplitz matrices where all diagonal entries are 1 and all non-diagonal entries are ρ. An equicorrelation matrix is observed from data with features that are equally correlated with each other, with ρ as the correlation coefficient. The data must also be standardized with variance 1 in all features.

In our experiment, we used equicorrelation matrices with $\rho = 0.8$ and $\rho = 0.4$ and observed some interesting results relative to those of the authors. The covariance plots in Figure 8 are almost identical to that of the isotropic structure by the authors. These plots are so similar because there are only two differences between isotropic and equicorrelation structures: equicorrelation assumes that each feature is standardized to have a variance of 1 and isotropic assumes all features are mutually independent.

More interesting are the double descent plots in Figure 9, where we observe that PCA-OLS outperformed ortho+ridge at all values of k. In contrast, ortho+ridge outperformed PCA-OLS at all values of k in all except one of the author's trials. Overall, the algorithms performed very similarly on equicorrelation compared to isotropic, except that oracle-PCR's peak MSE was much lower for p = 0.8.

Let us discuss the eigendecomposition of an equicorrelation matrix, especially within the context of PCA. The eigenvalues of a $N \times N$ equicorrelation matrix with correlation coefficient ρ are

$$\boldsymbol{\lambda_1 = 1 + (n-1)\rho} \textbf{ with multiplicity 1}$$

$$\boldsymbol{\lambda_2 = 1 - \rho} \textbf{ with multiplicity } \boldsymbol{n-1}$$

The matrix has n eigenvectors, where the j-th entry of the k-th eigenvector is

$$(v_k)_j = \begin{cases} 1 & j = 1 \\ -1 & j = k \\ 0 & \text{otherwise} \end{cases}$$

Since the matrix only has two distinct eigenvalues, all principal components explain roughly the same proportion of variance in the data. The eigenvectors contain only two non-zero entries and therefore correspond to one pair of opposing features at a time.

Overall, PCA helps OLS to avoid overfitting a training dataset with equicorrelation because the principal components are very simplistic and equally weighted. This factor contributes to the excellent performance of PCA-OLS on unseen data following an equicorrelation structure in Figure 9.

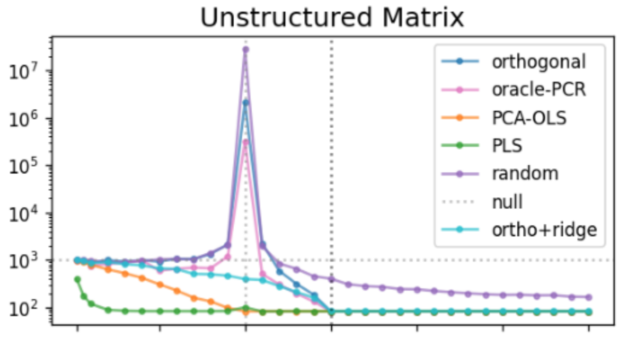


Figure 10:Double Descent plots with high SNR = 16 for Unstructured covariance structure.

The last covariance matrix structure tested was the Unstructured matrix. Within an Unstructured matrix, there is no pattern within the data, meaning each covariance and variance calculated is entirely different and has no relation to one another.

In our experiment, we tested the performance of this covariance structure. The results for the Double Descent plots in Figure 10 follow a similar method structure to the isotropic results reported by the authors. But there is a rather significant difference in the peak-MSE reported for all the descent methods compared to the isotropic structure, with larger max and minimum MSE values for all methods in the unstructured model.

These shifts reflect the worse performance of all algorithms on Unstructured structures, potentially due to the nature of the structure possessing no significant patterns. All methods, including the best performing PCA-OLS method, are likely to overfit training data with an unstructured matrix. This is because the principal components for an unstructured matrix are not equally weighted and have no clear discernible relationships leading to potential overfitting of random effects. This factor contributes to the poor performance of all methods on unseen data following an unstructured matrix structure as seen in Figure 10.

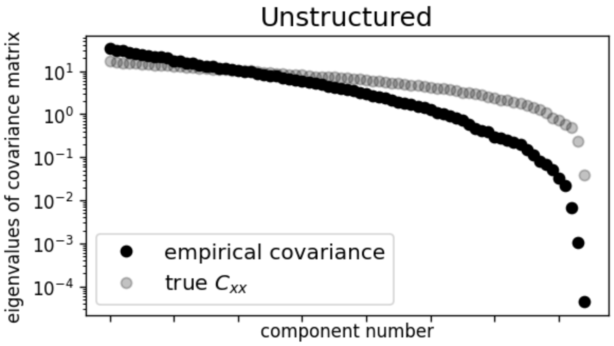


Figure 11: True and empirical covariance plots under high SNR=16 for unstructured covariance structure

Lastly , the covariance plots for our unstructured covariance structure in Figure 11 achieve higher covariance values compared to the structures reported by the authors. These higher values can be attributed to the unstructured model posing no constraints on its values , while other structures do.

Discussion and Conclusion

Modern-day deep learning models such as deep neural networks are still black-boxses for us, as compared to the well-understood classical machine learning models like SVM (support-vector-machine). While we understand the working mechanism of each node, little formalism was established in explaining why the model as a whole works so well, so we still have to rely on doing experiments to investigate the behaviors of these “black-boxes”, for example, when we observe Double Descent in over-parameterized neural network.

Since we know the classical machine learning models very well and little is known for the deep learning ones, it would be very useful if we can map these behaviors to the simpler models, and start from there. This paper is a perfect example for this type of research, where we first observe the same phenomenon in simpler models, and do simple experiments to make general conclusions by leveraging the highly explainable model structures in use. As a possible future direction, we should investigate Double Descent in other simple well-studied classical machine learning models, so that experimenting with them could help us understand deep learning better.

References

[1] Huang, N.T., Hogg, D.W., & Villar, S. (2022). Dimensionality reduction, regularization, and generalization in overparameterized regressions. *ArXiv, abs/2011.11477*.

[2] J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang, Generalization of two-layer neural networks: An asymptotic viewpoint, in International Conference on Learning Representations, 2020.

[3] Hastie, T., Montanari, A., Rosset, S. & Tibshirani, R. J. (2019). Surprises in High-Dimensional Ridgeless Least Squares Interpolation

Statement of Contributions

All authors contributed equally to this project.