UNDERSTANDING DOUBLE DESCENT THROUGH THE LENS OF PRINCIPAL COMPONENT REGRESSION

A PREPRINT

Christine H. Lind
Department of Electrical and Computer Engineering,
University of California San Diego,
La Jolla, CA, USA
clind@eng.ucsd.edu

Angela J. Yu
Department of Cognitive Science & HalIcioglu Data Science Institute,
University of California San Diego,
La Jolla, CA, US
ajyu@.ucsd.edu

April 27, 2021

ABSTRACT

Several recent papers have studied the double descent phenomenon: a classic U-shaped empirical risk curve when the number of parameters is smaller or equal to the number of data points, followed by a decrease in empirical risk (referred to as "second descent") as the number of features is increased past the interpolation threshold (the minimum number of parameters needed to have 0 training error). In a similar vein as several recent papers on double descent, we concentrate here on the special case of over-parameterized linear regression, one of the simplest model classes that exhibit double descent, with the aim of better understanding the nature of the solution in the second descent and how it relates to solutions in the first descent. In this paper, we show that the final second-descent model (obtained using all features) is equivalent to the model estimated using principal component (PC) regression when all PCs of training data are included. It follows that many properties of double descent can be understood through the relatively simple and well-characterized lens of PC regression. In particular, we will identify a set of conditions that will guarantee final second-descent performance to be better than the best first-descent performance: it is the scenario in which PC regression using all features does not suffer from over-fitting and can be guaranteed to outperform any other first-descent model (any linear regression model using no more features than training data points). We will also discuss how this work relates to transfer learning, semi-supervised learning, few-shot learning, as well as theoretical concepts in neuroscience.

***K*eywords** Double Descent · Principal Component Regression

1 Introduction

Several recent papers (Belkin et al., 2019, 2020; Bartlett et al., 2020) have studied the double descent phenomenon: a classical U shaped empirical risk curve when the number of parameters is smaller or equal to the number of data points, followed by a decrease in empirical (referred to as "second descent") as the number of features is increased past the interpolation threshold (the minimum number of parameters needed to have training error equal to 0).

In the vein of several the other recent papers on double descent (Belkin et al., 2019; Bartlett et al., 2020), we concentrate here on the special case of over-parameterized linear regression. This is one of the simplest model known to exhibit the double-descent phenomenon, while its simplicity allows analytical examination. While other papers have tried to

identical conditions under which the final second-descent model (including all features) can be expected to perform well compared to first-descent variants (Bartlett et al., 2020; Belkin et al., 2020), we concentrate here on the exact nature of the final second-descent solution. One interesting question is whether second descent finds a completely novel kind of solution to the regression problem that generalizes especially well, or whether it is equivalent to a particular first-descent solution.

As we will show in this paper, in the case of linear regression, the final second-descent model is exactly equivalent to a first-descent regression algorithm that uses principal components (PCs) as predictor variables instead of the original features, when all PCs of training data are included. A transformed regression model that uses PC features are predictor variables is also known as PC regression or PCR (Hotelling, 1957; Kendall, 1957; Park, 1981). Due to PCR being relatively simple and well understood, viewing double descent via PCR yields novel insight into the nature and performance of linear regression in the second descent. In particular, we will identify a set of conditions that will guarantee final second-descent performance to be better than first-descent performance using any combination or linear transformation of the original features (including PC transformation).

Finally, we will discuss how this work relates to semi-supervised learning, transfer learning, few-shot learning, as well as theoretical concepts in neuroscience.

2 Definitions and Notations

2.1 Linear Regression

We consider an ordinary linear regression problem with the standard linear-Gaussian setup: $y = \mathbf{x} \cdot \boldsymbol{\beta} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is normally distributed i.i.d. noise. Given a data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, containing m data points of n features, and a corresponding vector of observations $\mathbf{y} \in \mathbb{R}^{\mathbf{m}}$, we wish to estimate the coefficients $\hat{\boldsymbol{\beta}} \in \mathbb{R}^n$ as a function of $\mathbf{X}$ and $\mathbf{y}$. Without loss of generality, we assume $\mathbf{y}$ is centered and $\mathbf{X}$ has full row rank, i.e. rank($\mathbf{X}$)=m.

When $m < n$ (first descent), there is generally no way to reduce training error to 0, i.e. finding an exact solution for $\hat{\beta}$ such that $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. In this under-parameterized regime, the least-squares estimate, $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^{\mathbf{2}}$, minimizes the expected empirical risk (measured as squared error) on test data. This least-squares estimate is also the maximum-likelihood estimate (MLE) under the linear-Gaussian generative assumptions.

When $m = n$, there is a unique solution to $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and, as such, a unique solution for $\hat{\boldsymbol{\beta}}$. $m = n$ is known as the interpolation threshold, being the smallest number of parameters necessary to reduce training error to 0 (i.e. the estimated regression plane goes exactly through all the training data).

When $m > n$ (second descent), there is an infinite number of solutions for $\hat{\boldsymbol{\beta}}$ that satisfy $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$ (0 training error). In this over-parameterized regime, one standard way of finding a unique solution is to minimize the L2 norm of the parameter vector, $||\hat{\boldsymbol{\beta}}||^2$, which previously has been shown to exhibit the double-descent phenomenon (Belkin et al., 2019, 2020; Bartlett et al., 2020).

In all cases, $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{y}$, where $\mathbf{X}^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{X}$. The pseudoinverse can be obtained from a singular value decomposition (SVD) of $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$, as $\mathbf{X}^\dagger = \mathbf{V}\boldsymbol{\Sigma}^\dagger\mathbf{U}^{\mathbf{T}}$, where $\boldsymbol{\Sigma}$ is a rectangular diagonal matrix containing the singular values along the diagonal, and $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices whose columns consist of the left and right singular vectors, respectively.

2.2 Principal Component Regression

Principal component regression (PCR) is a regression technique based on principal component analysis (PCA), whereby the projection of the data matrix $\mathbf{X}$ along the PCs are used as the predictor variables in the regression analysis. PCR assumes the first k PC dimensions are retained for regression, where $k \leq m$. If $\mathbf{V}_k \in \mathbb{R}^{n \times k}$ contains the first k right singular vectors, then $\mathbf{Z}_k = \mathbf{X}\mathbf{V}_k$ is the representation of $\mathbf{X}$ projected onto the first k PCs. The linear regression problem is then transformed into finding $\hat{\boldsymbol{\beta}}$ that minimizes $||\mathbf{y} - \mathbf{Z_k}\hat{\boldsymbol{\beta}}||^{\mathbf{2}}$ in the first descent, or minimizing $||\hat{\boldsymbol{\beta}}||^2$ in the second descent assuming $||\mathbf{y} - \mathbf{Z_k}\hat{\boldsymbol{\beta}}||^{\mathbf{2}} = \mathbf{0}$.

3 Results

3.1 Final Second Descent Equivalent to PC Regression with All PC Features

We first note that test data predictions are invariant under orthogonal transformation of the (potentially overcomplete) basis. An important implications then follows: "final" second-descent predictions (using all features) on test data are exactly equivalent to PCR using all PCs of training data $\mathbf{X}$ at the interpolation threshold, or $\mathbf{Z} = \mathbf{XV}$.

We start by showing that model predictions are invariant under orthogonal transformations of the (potentially overcomplete) basis.

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be the training data matrix, $\mathbf{y} \in \mathbb{R}^{\mathbf{m}}$ be the corresponding responses, $\mathbf{W} \in \mathbb{R}^{n \times n}$ be an orthogonal matrix, and $\tilde{\mathbf{X}} = \mathbf{XW}$ be the data matrix in the transformed representation. We will show that the predictions made using the two representations are identical.

Let $\mathbf{X}'$ denote the test data, and $\tilde{\mathbf{X}}' = \mathbf{X}'\mathbf{W}$ denote the test data in the alternative representation. Then $\hat{\tilde{\mathbf{y}}}' = \tilde{\mathbf{X}}'\hat{\boldsymbol{\beta}} = \tilde{\mathbf{X}}'\tilde{\mathbf{X}}^{\dagger}\mathbf{y} = \mathbf{X}'\mathbf{W}(\mathbf{XW})^{\dagger}\mathbf{y} = \mathbf{X}'\mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{X}^{\dagger}\mathbf{y} = \mathbf{X}'\mathbf{X}^{\dagger}\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}'$. Note that because $\mathbf{W}$ is an orthogonal matrix, $\mathbf{W}^{\dagger} = \mathbf{W}^{T} = \mathbf{W}^{-1}$.

The above shows that the predictions for $\mathbf{y}'$ are invariant with respect to orthogonal transformations of the original feature representation $\mathbf{X}$. Since $\mathbf{V}$, containing the right singular vectors of $\mathbf{X}$, is orthogonal, PCR using all training PCs as predictor variables, or $\mathbf{Z} = \mathbf{XV}$, yields identical test predictions as the original feature representation. Since the output equivalence holds for arbitrary test data $\mathbf{X}'$, we can see that over-parameterized linear regression that minimizes the L2-norm of the regression coefficient vector is exactly identical to PCR at the interpolation threshold using all PCs.

In other words, at least in the case of overparameterized *linear regression*, the final second-descent solution, independent of the feature representation, is not a novel kind of regression solution, but exactly equivalent to the first-descent PC regression that does not discard any PCs. An important implication is that we can analyze how final second-descent performance can be expected to behave by appealing to what is known about PCR.

3.2 Insight into Second Descent via PCR

In this section, we appeal to known properties of PCR to identify conditions under which the final second-descent solution can be expected to outperform the best first-descent solution (i.e. the best under-parameterized regression model) on test data.

Let σ_i be the i-th largest singular value of the data matrix $\mathbf{X}$, i.e. the sample standard deviation of the i-th PC component, and σ_y be the (unknown) standard deviation of the Gaussian noise ϵ on y. Suppose we wanted to find the best first-descent linear regression model for $(\mathbf{X}, \mathbf{y})$ in the sense of minimizing expected MSE on test data, $E[(y - \hat{y}^2]$, by allowing a linear transformation of $\mathbf{X}$ and restricting to $k \leq m$ features. Park (1981) showed that if only k features are allowed in the linear regression, then the best such linear transformation is PC projection using the k largest PCs, i.e. $\Lambda X = \mathbf{XV}_k$, where $\mathbf{V}_k$ consists of the first k right singular vectors. Furthermore, if one wants to optimize for k in terms of minimizing MSE on test data, Park (1981) showed that the i-th largest PC should be dropped if and only if:

$$\sigma_i^2 < \frac{\sigma_y^2}{||\boldsymbol{\beta}||^2/m} \,, \tag{1}$$

where $\boldsymbol{\beta}$ is the true (unknown) parameter vector, σ is the (unknown) noise standard deviation, and m is the number of PCs. In other words, if the spread of the training data along the i-th PC direction is below some threshold, then this PC feature, along with all subsequent PC features (as well as all subsequent PCs, whose projected variances are even smaller), should not be included in the regression problem. According to the right side of the inequality, this threshold is larger (one should be more willing to drop PCs) if the noise is y is larger; conversely, this threshold is smaller (one should be less willing to drop PCs) if the average expected magnitude of the beta coefficients are larger. Both dependencies make intuitive sense: the magnitude of the quantity to be estimated (true magnitude of beta's) increases the signal-to-noise ratio, while noise in observations decrease it.

Interestingly, this also sets up the condition under which *none* of the PCs should be dropped in PCR, i.e. the smallest PC should be kept if and only if:

$$\sigma_m^2 \geq \frac{\sigma_y^2}{||\boldsymbol{\beta}||^2} \,. \tag{2}$$

In other words, when the smallest singular value is large enough, PCR including all the training PCs can be expected to outperform PCR with any proper subset of PCs, as well as any other possible linearly transformed representation of

$\mathbf{X}$. In particular, when the above inequality is satisfied, it follows that the test RMSE curve will not blow up like most first-descent models do (i.e. it does not over-fit or suffer from a classical bias-variance trade-off) as the interpolation threshold is reached, but instead monotonically decreases in a graceful manner.

Given the equivalence between PCR with all PCs included, and final second-descent performance, the above inequality being satisfied also implies that final-second descent performance can be expected to outperform first-descent models *regardless of the data representation* (up to a linear transformation of the original features).

The above relationship requires knowledge of the true noise variance σ_y^2 and the true parameter vector $\boldsymbol{\beta}$, which is unrealistic. Park (1981) recommends that they be respectively replaced by the MSE on training data, $\hat{\sigma}_y$, and the estimated regression coefficient vector with a correction for a tendency to inflate the magnitude of estimated β, especially when the spread of x is small or the noise in y is large, i.e.

$$\hat{\gamma}^2 = \hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\beta}} - \hat{\sigma}_y^2 \sum_{i=1}^{m} \frac{1}{\sigma_i^2} \tag{3}$$

3.3 The Nested Nature of PCR Models

The previous section showed that the final second-descent regression solution, independent of the initial feature representation for the data $\mathbf{X}$, is equivalent to PCR using all training data PCs. But we can leverage the orthogonality of PCs to gain further insight into the nature of PCR including all PCs, and thus also the nature of the final second-descent solution. In particular, we will show that the PCR models are nested, in the sense that the first i estimated regression coefficients remain the same, for all PCR models that utilize $k \geq i$ PC features.

Given the training data and its SVD decomposition $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, the representation of $\mathbf{X}$ in the PC representation is $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma}$, and the estimated parameter vector is $\hat{\boldsymbol{\beta}} = \mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^\dagger\mathbf{U}^T\mathbf{y} = \boldsymbol{\Sigma}^\dagger\mathbf{U}^\mathbf{T}\mathbf{y}$. For any $k \leq n$, let $\mathbf{X}_k = \mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k^T$ be the SVD decomposition of $\mathbf{X}$ using only the first $k \leq n$ features, where $\mathbf{U}_k$ and $\mathbf{V}_k$ contain the first k left and right singular vectors, respectively, and $\boldsymbol{\Sigma}_k$ is a diagonal matrix consisting of the first k singular values on the diagonal. Then, the first k regression coefficients obtained by using only the first k PC features remain identical if all n PC features are included in the model, $\hat{\boldsymbol{\beta}}_k^k = \boldsymbol{\Sigma}_k^\dagger\mathbf{U}_k^T\mathbf{y} = (\boldsymbol{\Sigma}^\dagger\mathbf{U}^\mathbf{T})_\mathbf{k} = \hat{\boldsymbol{\beta}}_\mathbf{n}^\mathbf{k}$. In fact, it is obvious that $\hat{\boldsymbol{\beta}}_k^k = \hat{\boldsymbol{\beta}}_h^k$ for any $h \geq k$, i.e. PCR models are nested.

4 Discussion

In this work, we identified the equivalence between overparameterized linear regression with PC regression using all PC features. In particular, when the smallest singular value of the data matrix exceeds a well-characterized threshold, PCR is guaranteed to have the smallest MSE by including all PC features, and thus the final second-descent performance can be expected to be better than or no worse than any first-descent model that has no more features than training data points (e.g. either PCR using some subset of PCs, or any other linear transformation of the predictor variables). In this regime, PCR tells us that all features should be included because the signal-to-noise ratio is high enough for even the last PC (smallest projected standard deviation), such that its regression coefficient can be estimated well enough to make a positive contribution over all. Interestingly, when PCR's empirical risk curve is monotonically non-increasing, it implies that PCR using all PC features does not blow up at the interpolation threshold as most models do. This marks out PCR as a nice exception to the general finding that MSE peaks at the interpolation threshold (Belkin et al., 2020).

Although obviously restricted to the linear regression domain, this work has interesting relevance to theoretical neuroscience and broader implications for several topical areas within machine learning. While modern machine learning has achieved incredible performance in some areas (He et al., 2015; Silver et al., 2016), relative to the human brain, it still falls short on several important respects. For example, while deep neural networks perform impressively on many perceptual classification tasks (He et al., 2015, 2016), it is extremely labeled-data intensive and computation-intensive (Krizhevsky et al., 2012; Thompson et al., 2020). In contrast, the developing brain receives a great deal of unlabeled data but only a very small amount of labeled data, i.e. putting it more in the regime of semi-supervised learning than supervised learning. It also completes a large variety of visual perception tasks with high accuracy and computational efficiency (Lake et al., 2017, 2019). Related to this, deep neural networks typically generalizes poorly in terms of transfer learning (Recht et al., 2018, 2019) and few-shot learning (Oreshkin et al., 2018). But the human brain appears capable of effortlessly and accurately generalizing to completely novel perceptual tasks. The current work suggests that one way to tackle this problem may be to have a massive over-parameterized feature representation that is largely learned in an unsupervised manner, and then train another decoding layer *as needed* for a novel supervised learning task (e.g. regression). The results (final second-descent) can be guaranteed to be as good as or better than any specialized model that does feature selection for each supervised learning task, based on

the results we described here. Interestingly, it has been shown that across species and sensory modalities, the brain tends to massively expand the feature representation from the initial sensory receptor level (Olshausen and Field, 1997; Dasgupta et al., 2018). Previously, it was suggested that such feature expansion may serve some appealing goals within unsupervised learning, such as reducing energetic needs in the brain (sparsification) or preserving semantic similarity among sensory input. However, our result suggests that this feature expansion in the brain may instead serve as a powerful *universal* representation that can readily support future unspecified supervised learning needs with high accuracy and sample-efficiency.

At a technical level, what we have shown are some sufficient conditions for guaranteeing the final second-descent performance to exceed that of any first-descent model (up to linear transformation of the predictor variables), which is a particular lower bound on the smallest singular value of the predictor data matrix. However, if the smallest singular value falls short of the inclusion threshold, it does not necessarily mean that final second descent performance would not be better than best first-descent performance. It only implies that final second-descent performance would be worse than eliminating one or more of the smallest PCs in PCR. If on-line PCA of training data is not possible for some reason, then final second-descent performance might still out-perform the best first-descent model. How the two compare depends both on how small the smallest singular values are (those that fall below the inclusion threshold), and how *far* the feature representation is from training data PC representation. An arbitrary linear transformation (away from the PC representation) can lead to arbitrarily small smallest singular values, which can in turn lead to large empirical risk on test data. For example, in the case of the brain, if most of the unsupervised learning in the visual cortex is done beforehand (e.g. during a critical period during development (Reh et al., 2020)), and very limited additional neural plasticity is available later on when encountering novel supervised learning tasks, then it is unreasonable to assume that the visual cortex can reorganize itself to obtain a PC representation of new training data.

This work can be developed in multiple fruitful future directions. For example, it is worthwhile to examine how the results presented here generalize to other supervised learning tasks (e.g. classification), or relate to more complex multi-level architecture (instead of only single-level linear regression). Another interesting direction is to identify more precisely the general feature representational settings under which final second-descent can be expected to out-perform best first-descent. Yet another direction is to examine how things generalize with a different norm for minimization in the second descent. For example, in theoretical neuroscience, L1 norm (sparsification) has received more attention than L2 norm minimization in overcomplete representations (Olshausen and Field, 1997); an interesting question is which first-descetn solution L1 norm minimization might be equivalent to, and, more generally, how one should choose the norm to minimize in the second descent.

References

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL `https://www.pnas.org/content/116/32/15849`.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Sanjoy Dasgupta, Timothy C Sheehan, Charles F Stevens, and Saket Navlakha. A neural data structure for novelty detection. *Proceedings of the National Academy of Sciences*, 115(51):13093–13098, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Harold Hotelling. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79, 1957.

Maurice George Kendall. A course in multivariate analysis. Technical report, 1957.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018.

Sung H. Park. Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics*, 23(3):289–295, 1981. ISSN 00401706. URL http://www.jstor.org/stable/1267793.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

Rebecca K Reh, Brian G Dias, Charles A Nelson, Daniela Kaufer, Janet F Werker, Bryan Kolb, Joel D Levine, and Takao K Hensch. Critical period regulation across multiple timescales. *Proceedings of the National Academy of Sciences*, 117(38):23242–23251, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 2020.