M.S. THESIS

Review on a double descent in the high dimensional regime

고차원에서의 이중하강현상에 대한 고찰

BY

Woonyoung Chang

August 2021

DEPARTMENT OF STATISTICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Review on a double descent in the high dimensional regime

고차원에서의 이중하강현상에 대한 고찰

지도교수 정 성 규

이 논문을 이학석사 학위논문으로 제출함

2021년 4월

서울대학교 대학원

통계학과

장 우 녕

장우녕의 이학석사 학위논문을 인준함

2021년 7월

위 원 장 _____ 이 상 열 _____

부위원장 _____ 정 성 규 _____

위 원 _____ PARK JUN YONG _____

Abstract

Modern statistical methods, such as deep neural network, give rise to many thought-provoking phenomena. We deal with one of them, *double descent*. Double descent occurs when a model with a sufficiently large number of parameters has good generalization. This seems to conflict with the classical notion of bias-variance tradeoff. Recent studies have tried to explain this interesting phenomenon across a variety of statistical methodologies. Here, we review and compare proposed interpretations in linear regression, linear discrimination, and deep neural network.

keywords: Modern statistical methods, Double descent, Linear regression, Linear discrimination, Deep neural network

student number: 2019-22321

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

A model selection has been an ongoing issue in the fields of statistical modeling. One of the main goals of the model selecting procedure is to guarantee the generalization power, that is, to minimize the test error. The traditional criterion, such as Akaike Information Criteria [20] and Bayesian Information Criteria [21], attempts to balance the bias-variance trade-off, suggesting that the overfitted model tends to have poor generalization.

However, current experiments in modern statistic modeling lead us to design sufficiently complex training procedures. This is because deep models have been shown outstanding test performance even with huge model complexity, and these models often almost fully interpolate the training data, that is, achieve vanishing training error. In addition, recent papers present a new perspective to understand these phenomena. [3] suggests that the concept of interpolation and that of generalization are separate rather than contradictory, and [11] shows that the two concepts can coexist in ridgeless kernel estimation.

This paper is a survey on the new concepts of bias-variance trade-off and one of its derivatives, a double descent. A double descent is an unconventional trend in risk curves in the context of modern high-complexity learners. With an increasing model complexity p, the risk initially decreases, attains local minimum, and increases

until p reaches a certain threshold (interpolation threshold). Beyond the interpolation threshold, the risk decreases again, resulting in a twofold descent-shaped risk curve. See Figure 1.1.

Although recent machine learning algorithms shed a light on the double descent phenomenon, it is difficult to say that this phenomenon has been overlooked so far. [6, 7, 12, 18] gave empirical evidence contradicting the conventional model selection, and their work rooted the analysis on the double descent. To cover the double descent phenomenon, recent research have been conducted on various learning algorithms such as linear regressions [4, 17, 15, 8], classifications [5, 9, 14], random feature regressions [13], and deep neural networks [16, 13, 19]. Some of them discover non-intuitive and interesting phenomena: more training data may hurt the test performance [16] and negatively tuned regularization helps the test performance [10]. Also, neural network algorithms often achieve nearly optimal prediction accuracy even when they are trained on contaminated training data[23].

In Chapter 2, we present a simple example of a double descent and illustrate basic frameworks. The current results on a double descent in various statistical models, including linear regression, linear classification, and neural networks, are provided in Chapter3. We conclude with a discussion in Chapter 4.

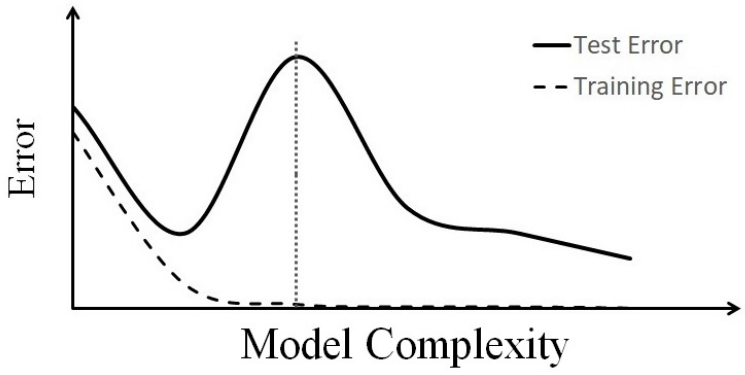

Figure 1.1: Example 1.

Chapter 2

Preliminary

2.1 Simple example

We are going to bring out the concept of a double descent with a simple, but intuitive, example. Consider a binary classification problem. From each p-dimensional multivariate normal distribution, $\mathcal{N}(0, I_p)$ and $\mathcal{N}(\sqrt{p}e_1, I_p)$, 100 independent samples are generated. Here, I_p denotes a $p \times p$ identity matrix and e_1 is the first canonical basis of $\mathbb{R}^p$. To classify new sample, we make a classification rule as follows.

1. If $p < 100$, apply Fisher's linear discriminant analysis (LDA).
2. Otherwise, perform a linear discrimination with the maximal data piling (MDP) direction.

Maximal piling direction is a natural extension of Fisher's linear LDA direction to high dimension, and detailed descriptions are made in Section 1.

Figure 1 depicts the estimated error rate, by 1000 times sample regeneration, for each dimension p. Here, two conspicuous features can be seen in the test error curve. First, if the number of samples $n = 200$ is larger than the dimension p, the shape of the error curve is consistent with the traditional concept of bias-variance trade-off. With an increasing p, or increasing the number of covariates in terms of linear regression,

the bias decreases while the variance increases. The error rate, or the risk, reflecting both trends forms of U-shaped. Second, surprisingly, it can be seen that the error rate decreases again after dimension p acrosses n, which is consistent with recent findings.

We can also check in Figure 2.1 that the training error monotonously decreases as p increases. In particular, beyond $p \approx n$, we have nearly 0 training error. The MDP direction perfectly separates the training groups as shown in Section 2.3.3. In this example, $p \approx n$ serves as a threshold that discriminates two regimes, under-parametrized regime and over-parametrized regime.

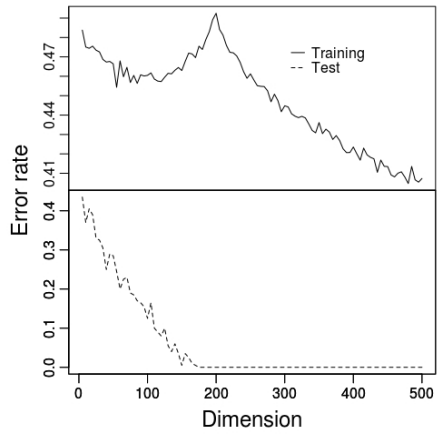

Figure 2.1: Test and training error with increasing dimension. Errors are estimated thorugh 1000 times of sample regeneration.

2.2 Classical bias-variance trade-off

Given a sample of training data $\{(x_i, y_i)\}_{i=1}^n$ from a probability distribution $\mathcal{P}$ on $\mathbb{R}^p \times \mathbb{R}$, a learning procedure develops a predictor $f_n : \mathbb{R}^p \to \mathbb{R}$ to predict the unseen y of a new data point x. The prediction algorithms vary, but a lot of them fall into the scope of empirical risk minimization (ERM). The goal of ERM is to find the f_n that minimizes empirical risk $\sum_{i=1}^n l(y_i, f(x_i))/n$ among $f \in \mathcal{F}$. Here, l is a loss

function, such as a square loss $l(y, a) = (y - a)^2$ in the case of regression or 0-1 loss $l(y, a) = I(y \neq a)$ in the case of classification. We estimate true f minimizing true risk $\mathbb{E}_P l(y, h(x))$ with empirical risk minimizer f_n. However, there exists a gap between them since we generally do not know the true probability distribution P from which new observation is drawn. Therefore, we have been forced to determine how much *information* from the training data should we use. In other words, we should balance the degree of under-fitting and that of over-fitting for good generalization power as suppressed in the classical U-shaped risk curve on an under-parametrized regime in Figure 1.1 and 2.1.

One of the direct approaches is to control the *size* of the function class $\mathcal{F}$.

1. If $\mathcal{F}$ is too small, all predictors in $\mathcal{F}$ may not enough catch the noticeable features in the data (under-fit).

2. If $\mathcal{F}$ is too large, the empirical risk minimizer may over-concentrate on superfluous features in the training data (over-fit) and thus poorly predict on new data.

In particular, classical regularization methods that constrain the size or sparsity of models can also be considered. These approaches focus on finding the sweet spot between two extreme situations.

2.3 Interpolation threshold and interpolaters

Modern statistical models, such as deep neural networks and random forests, involve a large number of parameters. In particular, most models are developed to perfectly interpolate the data. The *interpolation threshold* is the first moment when the model fits the training sample with significantly small training error. If the model does not reach this criterion, we call the model under-parametrized. On the other hand, if this criterion is exceeded, the model is said to be over-parametrized. Finally, the model is called critically parametrized if it is near this criterion.

Estimators used in well-known analyses for the under-parametrized models are often undefined in the over-parametrized models, e.g., the least square estimator (LSE) and Fisher's linear discriminant analysis (LDA). Hence, we here introduce notable estimators, which is called *interpolators* in linear regression and linear classification. Interpolators have a close connection with its under-parametrized-model version, and a lot of research on the over-parametrized model is conducted through these estimators.

2.3.1 Min-norm least square estimator

Assume training data $\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}$ from a model

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \dots, n$$

where x_i and ϵ_i are independently drawn from $(0, \Sigma)$ and $(0, \sigma^2)$ for each $i = 1, \dots, n$. Least square estimator $\hat{\beta}^{\mathrm{LSE}}$ is the ERM with the square loss when $p \leq n$,

$$\hat{\beta}^{\mathrm{LSE}} = (X^T X)^{-1} X^T y = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

where X is an $n \times p$ data matrix and $y \in \mathbb{R}^p$ is a vector consisting independent variables. When $p > n$, there exists a number of *interpolators* $\tilde{\beta}$ which literally interpolates the training data, i.e., $y_i = x_i^T \tilde{\beta}$ for $i = 1, \dots, n$. Among those we choose $\hat{\beta}$ with minimum l_2-norm,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p : y = X\beta}{\operatorname{argmin}} \|\beta\|_2. \tag{2.1}$$

The solution of (2.1) is given by $\hat{\beta} = (X^T X)^{\dagger} X^T y$ where $A^{\dagger}$ denotes the Moore-Penrose inverse of A. We sometimes call $\hat{\beta}$ the ridgeless least square estimator since

$$\hat{\beta} = \lim_{\lambda \to 0+} (X^T X + \lambda I)^{-1} X^T y.$$

Hence, we can consider $\hat{\beta}$ as an empirical risk minimizer with square loss and vanishing l_2-penalty. Also, it is known that the gradient descent (GD) on the empirical risk with the square loss converges to least square estimator $\hat{\beta}^{\mathrm{LSE}}$ for $p \leq n$ and to

min-norm least square estimator $\hat{\beta}$ for $p > n$ [8, Proposition 1]. Therefore, theoretical results undergo using $\hat{\beta}^{\mathrm{LSE}}$ and $\hat{\beta}$ rather than directly focusing on the empirical risk itself.

2.3.2 Maximum margin classifier

In this subsection, we consider a binary classification problem under a logistic model. For $x \in \mathbb{R}^p$ and $y \in \{-1, 1\}$,

$$\mathbb{P}(y = 1|x) = f(x^T\theta_0) = 1 - \mathbb{P}(y = -1|x) \tag{2.2}$$

for some unkwnown regressor θ_0 and a sigmoid function f. We obtain the estimate $\tilde{\theta}$ via ERM principle for a logistic loss l. Since the emprical risk in this case does not have a closed-form minimizer, we construct gradient descent iteration, that is, for $k \in \mathbb{N}$,

$$\tilde{\theta}^{(k+1)} = \tilde{\theta}^{(k)} - s_k \nabla R_{\mathrm{emp}}(\tilde{\theta}^{(k)}) \tag{2.3}$$

where $\nabla R_{\mathrm{emp}}(\tilde{\theta}^{(k)})$ is a gradient of the emprirical risk at $\tilde{\theta}^{(k)}$ and s_k is the kth-step size.

The limit of gradient descent procedure depends on the linear separability of data. We say the training data is linearly separable if and only if there exists a linear classifier which perfectly separates two training groups, i.e., $\exists \beta \in \mathbb{R}^p$ satisfying $y_i x_i^T \beta \geq 1$ for $i = 1, \ldots, n$. When data is linearly separable, then the hard-margin support vector machine (SVM) uniquely exists,

$$\hat{\beta}^{\mathrm{SVM}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{\|\beta\| : y_i x_i^T \beta \geq 1 \text{ for } i = 1, \ldots, n\}. \tag{2.4}$$

In addition, [5] shoed that the gradient descent iteration converges to the hard margin SVM,

$$\tilde{\theta}^{(k)}/\|\tilde{\theta}^{(k)}\| \to \hat{\beta}^{\mathrm{SVM}}/\|\hat{\beta}^{\mathrm{SVM}}\|$$

as $k \to \infty$. For non-separable data, it is known that the GD converges to the maximum likelihhod estimator, $\hat{\beta}^{\mathrm{ML}}$.

2.3.3 Maximum data piling

Maximum data piling direction is a binary classifier which generally exists when $p > n$. Suppose that we have two classes of samples in $\mathbb{R}^p$. For $i = 1, 2$, let the ith class consists of the p-dimensional sample $x_{i,j}$ ($j = 1, \ldots, n_i$) which is drawn from a continuous ditribution. Denote the within-class covariance matrix as $S_w = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{i,\cdot})(x_{i,j} - \bar{x}_{i,\cdot})^T/(n_1 + n_2)$ and between-class covariance as $S_b = n_1 n_2 (\bar{x}_{1,\cdot} - \bar{x}_{2,\cdot})(\bar{x}_{1,\cdot} - \bar{x}_{2,\cdot})^T/(n_1 + n_2)^2$ where $\bar{x}_{i,\cdot}$ denotes the sample mean of the ith class for $i = 1, 2$. Then, the total sample covariance matrix $S_t = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{\cdot,\cdot})(x_{i,j} - \bar{x}_{\cdot,\cdot})^T/(n_1 + n_2)$ with the total sample mean $\bar{x}_{\cdot,\cdot}$ satisfies $S_t = S_w + S_b$.

The Fisher's linear discriminant analysis (LDA) finds a classfier w which maximizes the ratio between between scatter and within scatter, i.e.,

$$w_{\mathrm{LDA}} = \underset{w \in \mathbb{R}^p : \|w\| = 1}{\operatorname{argmax}} \frac{w^T S_b w}{w^T S_w w}. \tag{2.5}$$

The solution of (2.5) is given by $w \propto S_t^{-1} d$ for $p \leq n_1 + n_2 - 2$. However, when $p > n_1 + n_2 - 2$, there exists a number of w's which make the within-class scatter 0. If each data is projected on such w, all points within the same class are projected onto a single point. In other words, they perfectly separate two training classes. The maximal data piling (MDP) direction proposed by [1] maximizes the between scatter while degenrating the within scatter,

$$w_{\mathrm{MDP}} = \underset{w \in \mathbb{R}^p : \|w\| = 1}{\operatorname{argmax}} \; w^T S_b w \quad \text{subject to} \quad w^T S_w w = 0. \tag{2.6}$$

The MDP direction is *maximal* in the sense that it induces the longest distance between projections of each class.

Chapter 3

Modern bias-variance trade-off in statistical models

In this chapter, we introduce some recent findings on bias-variance trade-off in various statistical models including linear regression, linear classification, and deep neural networks.

3.1 Linear regression

Many of theoretical results consider an asymptotic setup where n and p increase at the same rate, i.e., $p/n \to \gamma \in (0, \infty)$. We focus on an asymptotic risk $R(\gamma)$ induced by the min-norm least square estimator $\hat{\beta}$ in (2.1),

$$R(\gamma) = \lim_{n,p\to\infty} \mathbb{E} l(y, x^T \hat{\beta}). \tag{3.1}$$

The limit in (3.1) shows a sharp difference between the models with $\gamma < 1$ and those with $\gamma > 1$. Table 3.1 is a precise characterization of $R(\gamma)$ for isotropic x and the square loss $l(y, a) = (y - a)^2$ [8, Theorem 1&2]. The results in Table 3.1 suggests that $\gamma = 1$ serves as an interpolation threshold. This is quite intutive recalling a shallow linear regression model. We make some discussion in the perspective of bias-varaince trade-off.

Table 3.1: The asymptotic risk for the isotropic feature under the square loss. Here, $r^2 = \|\beta\|_2^2$.

	Bias	Variance	$R(\gamma)$
Under-parametrized ($\gamma < 1$)	0	$\sigma^2 \frac{\gamma}{1-\gamma}$	$\sigma^2 \frac{\gamma}{1-\gamma}$
Over-parametrized ($\gamma > 1$)	$r^2 \frac{\gamma-1}{\gamma}$	$\sigma^2 \frac{\gamma}{1-\gamma}$	$r^2 \frac{\gamma-1}{\gamma} + \sigma^2 \frac{1}{\gamma-1}$

Under-parametrized regime. There is no bias term and the variance term $\sigma^2 \frac{\gamma}{1-\gamma}$ dominates the risk. The variance increases with γ and diverges as $\gamma \to 1-$. This result fits with our conventional wisdom.

Over-parametrized regime. The bias term $r^2 \frac{\gamma-1}{\gamma}$ increases in γ. It is perceptive since $\hat{\beta}$ lies on an n-dimensional subspace, which is the rowspace of data matrix X, so that it accounts for less amount of p-dimensional feature space as p, n grows. On the other hand, the variance term $\sigma^2 \frac{1}{\gamma-1}$ decreases with γ. [8] gives a nice intuition about this. Considering the simple linear system $y = Xb$, $\hat{\beta}$ is a candidate for b with smallest l_2-norm. Hence, the elements of $\hat{\beta}$ are "well-distributed", considering the concept of Cauchy Schwartz inequality. As p grows, distributing them over more columns results in decreasing l_2-norm, considering the concept of Hölder's inequality. This explanation ahs a close connection with the variation of the ratio between largest and smallest singular values of the covariance matrix which is described in [19]. Combining the bias and the variance, the shape of the risk curve in the over-parameterized model critically depends on the signal to noise ratio (SNR). In particular, the risk monotonely decreases for sufficiently small SNR while the risk has local minimum for sufficiently large SNR.

3.2 Linear classification

We denote the feature vector $x \in \mathbb{R}^p$ and the binary class label $y \in \{\pm 1\}$. Supervised binary classification problems are studied under popular models including a generative logistic model and a discriminative Gaussian mixture model[9, 22]. Here we assume a generative logistic model (2.2).

In classification, a misclassification rate serves as a measure for a performance of a classifier. We introduce [5] who focused on the asymptotic error of the max-margin classifier (2.4). They made the asymptotic setup which is the same as the setup presented in the previous subsection. First, they obtained the asymptotic interpolation threshold $\gamma^* \in (0, 1/2)$ in the sense that the data are (asymptotically) linearly separable if and only if $\gamma > \gamma^*$. Next, they made a full characterization of asymptotic test error and showed that the error has a sharp distinction between the under-parametrized model, $\gamma < \gamma^*$ and the over-parametrized model, $\gamma > \gamma^*$.

The double descent behavior of the test error can be observed in the results of [5]. In all cases they considered, the test errors have two local minima corresponding to two regimes of learning. However, the global minima have a close connection with the SNR. For example, the risk tends to have its global minimum on the overparameterized regime for large SNR.

Also, see [5] for the characterization of the error of the ML estimator and [9] for the similar work under the square loss.

3.3 Neural networks

In this subsection, we present recent evidence of a double descent in deep neural networks. As the training procedure of neural networks tends to be complex, it is difficult to rigorously identify the interpolation threshold. Therefore, [16] suggest an effective model complexity (EMC). The EMC is a function of training algorithm and depends

on a sample distribution and a sufficiently small cut-off ϵ:

$$\mathrm{EMC}_{\epsilon,\mathcal{P}}(\mathcal{T}) = \max\{\mathrm{n} : \mathbb{E}_{\mathrm{S}\sim\mathcal{P}^{\mathrm{n}}}[\mathrm{err}\{\mathcal{T}(\mathrm{S})\}] < \epsilon\}. \tag{3.2}$$

Here, $\mathcal{S} = \{(x_i, y_i) : i = 1, \ldots, n\}$ denotes a training sample. The EMC can be understood as the maximum number of samples achieving sufficiently small test error. Hence, we say that We introduce an informal hypothesis in [16] suggesting that the double descent in a neural network is a function of the EMC, so that is a function of the training procedure $\mathcal{T}$. Suppose that we have n training samples.

Under-parametrized regime. When $\mathrm{EMC}(\mathcal{T}) \ll n$, the variation on $\mathcal{T}$ increasing the EMC tends to decrease the test error.

Over-parametrized regime. When $\mathrm{EMC}(\mathcal{T}) \gg n$, the variation on $\mathcal{T}$ increasing the EMC tends to decrease the test error.

Critiacally parametrized regime. When $\mathrm{EMC}(\mathcal{T}) \approx n$, the variation on $\mathcal{T}$ increasing the EMC may either decrease or increase the test error.

In the following subsections, we describe various forms of double descent.

3.3.1 Model-wise double descent

It has been shown that model-wise double descent occurs in various deep models, such as VGGnet and Resnets [16, 15, 23, 2]. In the training procedure, we often noise label or augment the data. This modification tends to intensify the double descent feature; however, [16] stated that it may not be the actual cause of the phenomenon. For example, when Resnet18 is trained on CIFAR-100, we can see that a model-wise double descent occurs with or without label noise.

3.3.2 Epoch-wise double descent

Double descent also occurs over the course of training for large architecture. For small models, the test error monotonously decreases with respect to the number of iterations. In contrast, non-monotonicity occurs for the bigger models. Precisely, the test error strikingly decreases at beginning of the training procedure, then increases, and gradually decreases beyond the EMC. Interestingly, the error hardly re-decreases for the medium-size model. It may indicate that the EMC of a medium-size model can be larger than that of a large-size model. Also, the epoch-wise double descent has turned out to be robust to optimization (SGD and Adam) and degree of wight decay or learning rate.

3.3.3 Sample-wise double descent

The test error is also affected by the number of training samples. It has been observed that the double descent phenomenon can be alleviated for a large number of samples while it still occurs. Although an increase in sample size reduces the test error, it also shifts the peak of the error. For example, the performance of VGGnet, trained on CIFAR-100, improves with an increasing number of samples. At the same time, the peak at which the error increases due to excessive learning also moves to the right. In summary, more data ultimately helps the performance; however, the combination of peak-move and sample-wise double descent may result in the situation, more data hurts.

Chapter 4

Conclusion

We dealt with the double descent phenomenon found in linear regression analysis, linear discriminant analysis, and deep neural network models. The double descent phenomenon in linear regression analysis and linear discriminant analysis is not perfect, but it has some theoretical support. The reality is that deep neural network models generally lack theoretical and mathematical support. [16] did enough research to give intuition through a vast amount of experimentation, but there was little mention of the theoretical part that would supplement him. This is expected not only to explain the double descent phenomenon, but also to be a task for further deep neural network research.

Bibliography

[1] Jeongyoun Ahn and J. S. Marron. The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259, January 2010.

[2] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[3] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

[4] Prasad Cheema and Mahito Sugiyama. A geometric look at double descent risk: Volumes, singularities, and distinguishabilities. *arXiv preprint arXiv:2006.04366*, 2020.

[5] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification, 2020.

[6] Robert PW Duin. Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964. PROCEEDINGS PUBLISHED BY VARIOUS PUBLISHERS, 1995.

[7] Robert PW Duin. Classifiers in almost empty spaces. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pages 1–7. IEEE, 2000.

[8] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[9] Ganesh Kini and Christos Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. *arXiv preprint arXiv:2001.11572*, 2020.

[10] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

[11] Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

[12] Marco Loog and Robert PW Duin. The dipping phenomenon. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 310–317. Springer, 2012.

[13] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[14] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[15] Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent, 2019.

[16] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and

Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

[17] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent, 2020.

[18] Manfred Opper. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pages 922–925, 1995.

[19] Tomaso Poggio, Gil Kur, and Andrzej Banburski. Double descent in the condition number. *arXiv preprint arXiv:1912.06190*, 2019.

[20] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81, 1986.

[21] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[22] Matthias Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106, 2004.

[23] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

초 록

심층 신경망과 같은 현대 통계적 방법론들은 다양한 생각을 불러일으키는 현상을 일으키고 있다. 우리는 그 중 하나인 이중하강현상을 다룬다. 이중하강현상은 충분히 많은 수의 매개변수가 있는 모델이 좋은 일반화를 가질 때 발생한다. 이것은 고전적인 편향-분산 절충의 개념과 충돌하는 것으로 보인다. 최근 연구들은 다양한 통계 방법론에 걸쳐 이 흥미로운 현상을 설명하려 노력하였다. 여기에서 우리는 선형 회귀분석, 선형 판별분석 및 심층 신경망에서 제안된 해석을 검토하고 비교한다.

주요어: 현대 통계적 방법론, 이중하강현상, 선형 회귀분석, 선형 판별분석, 심층 신경망

학번: 2019-22321