DOUBLE DESCENT AND STATISTICAL PHYSICS

Bui Gia Khanh *
Department of Physics Hanoi University of Science, Vietnam National University
Hanoi, Vietnam
{fujimiyaamane}@outlook.com

ABSTRACT

The analysis of learning action, machine learning and related practices in the field theoretically has been made by utilizing Computational Learning Theory (CoLT) Valiant (1984) and Statistical Learning Theory (SLT) Vapnik (1999). These two theories, while overlapped, provides a general framework in analysing and justifying learning actions and learning model constructions, aiding in the formation of modern practices. One of the famous insight using such framework is the *bias-variance tradeoff* Geman et al. (1992), which states that model complexity and generality inversely affect each other, thus guarantee the need for a safe bracket between them. However, recent literatures, Belkin et al. (2019) has indicated the fallout of such dilemma by the new phenomenon called *double descent*, where there exists an interpolation threshold that renders the current statistical justification for bias-variance inconsequential to a given region. Various 'anomaly' has also been detected similarly, with varying degree of sophistication and potential. In this paper, we investigate the phenomenon under new insight, and with renewed modelling architectures aiding in the process of experimental analysis. In turn, it would also indict on the topic of model selection principle.

1 INTRODUCTION

The theory of machine learning and its modern practice has been developed and researched, of a substantial portion by empirical and heuristic approach, either by advancing new practices, architectures or by try-and-test modification. From its early onset of a regression estimator $\theta(x, y)$ on the linear regression problem, machine learning has developed substantially. Certain model concept with great successes includes regression-classification model, Bayesian modelling, generative learning model, support vector machine, Gaussian processes, and more. Of all such, the more formal and complex model architecture created, is the concept of a *neural network*. With increasingly sophisticated architecture, heuristic approach become popular, the fast-paced advancement of the field comes with new method, new results, new observations, and its far-reaching application which led to even bigger and larger scale deployment, there has been questions about the formation and status of a theoretical ground, a rigorous matter on the side of **theoretical machine learning**.

While rigorous and well-formulated in a sense, classical and theoretical machine learning was dwarfed by the modern advancement of machine learning as a whole, leading to several anecdotal problems regarding the interpretation of phenomena, the re-evaluation of the theory to fit the more updated analysis, and explanation to more sophisticatedly designed system. This and many more, plus as present, many of such advancements and improvements are heuristic, and the general theory and conceptual understanding remain limited, led to the choice to often opt for analogies and empirical workaround. Ultimately, this resulted in multiple 'failures' in explaining, interpreting, and predicting behaviours of new phenomena appeared in the modern landscape of machine learning. Thereby, we suggest some particular insights and analysis, and a fix within interpretation of the phenomena.

Our main focus is to review on the topic of the umbrella term classical learning theory, and the recent phenomena observed named 'double descent'. Statistical learning theory (SLT) and computational learning theory (CLT) Vapnik (1999); Mohri et al. (2012); Shalev-Shwartz & Ben-David (2014); Hajek & Raginsky (2021); Bousquet et al. (2020) has been prominent in constructing a well-rounded formal theory surrounding learning problems, models, and machine learners analysis. Valuable insights have been dissected from treatment of statistical theory and mathematical modelling on

*Undergraduate student, affiliated

models, including the *bias-variance tradeoff* Geman et al. (1992); Domingos (2000), which serves as a bound for efficient learning and model configuration (or complexity). However, recently there has been observations of *double descent* Belkin et al. (2019); Schaeffer et al. (2023); Nakkiran et al. (2019); Lafon & Thomas (2024) which refute the famous tradeoff assumption, and hence brings question to the establishment of the theory, as well as several assumptions and insight in the framework. Further events and phenomena observed also includes grokking and triple, to n-descent Davies et al. (2023); d' Ascoli et al. (2020). Furthermore, many problems of defining and formalizing notions used in designing and implementing machine learning models are inconclusive, such as, for example, *model complexity* and others.

On itself, the phenomena *double descent* has been investigated somewhat comprehensively, firstly introduced by Belkin et al. (2019), which gives it distinctive saddle escape illustration. Further analysis was made by several literatures, particularly attributed the existence of double descent to the concept of model complexity and inductive bias, and potential explanations of such. Nakkiran et al. (2019) expanded the phenomena into deep neural network models. Their conclusion is reached by considering the perturbation of a learning procedure $\mathcal{T}$ on the effective model complexity $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ defined in the paper, separating the eventual phenomena into regions of observations, thereby in one way or another, predicting the tendency of double descent. This is also the first, arguably, "concise" definition of double descent, even though its nature is an empirical definition, including the notion of model complexity. Recently, there is also the Neural Tangent Kernel (Jacot et al. (2018)) which can be used to explain double descent, however further researches are still required. Preliminary works are done by Lafon & Thomas (2024), Schaeffer et al. (2023), and Liu & Flanigan (2023) on the role of optimization in double descent. Davies et al. (2023) attempted to unifying grokking with double descent, theorized a possibility of similarity during the generalization phase transition of the inference period, and Olmin & Lindsten (2024) attempted to explain epoch-wise double descent within the model of two-layer linear network. However, it is mentioned that it can also be expanded into deep nonlinear networks.

2 Problem statements

Before continuing further, we must review the problem statement that led to the problem of double descent. We will have a look at the theory, the original conflict, and the illustrative phenomenon in action.

2.1 Bias-variance tradeoff

In classical sources, James et al. (2013); Goodfellow et al. (2016); Hajek & Raginsky (2021); Mohri et al. (2012); Shalev-Shwartz & Ben-David (2014), bias-variance tradeoff is a classical principle used in the categorization of technique called *model selection*. Specifically, this refers to the progress of choosing the best predictor h^* that minimize the generalization error that is the main goal of a particular machine learning model, during the process of training and inference. Originally, the theory that bias and variance often come in to conjunction is **Estimation theory**, of which there exists an estimator $\hat{\theta}$ that use a set of observations, to estimate the concept, or the underlying mechanics of certain system that outputted the observation set, by the **probabilistic perspective** - that is, it is governed by appearance by an independently and identically distributed process of parameterized probability $\theta \in \Theta$. Here, parameterized means being expressed by a set, often finite, of parameters, by E. L. Lehmann (1998); Paninski (2005). It is only when Geman et al. (1992) introduced it to machine learning that we have an equivalent structure that is similar to bias-variance in machine learning.

2.1.1 Precursor (Geman et al., 1992)

The original definition of bias-variance tradeoff by Geman et al. (1992) is first constructed using the means-square error, which is regarded as a normal measure in the real encoding space. Their approach is to justify bias-variance via decomposition of the loss function ℓ, for such to find an alternative reasonable form of such loss landscape. Suppose of a regression problem to construct a hypothesis function $f(x)$ from $(x_1, y_1, \ldots, x_N, y_N)$ for the purpose of generalization - that is, predicting unseen variational values for different pair $(x_j, ?)$ such that $? = y_j + \epsilon$ for a conceivable implicit error. To be explicit about the relation of this problem, or f on the given data $\mathcal{D} = \{(x_i, y_i) \mid i \leq N\}$, denote $f(x; \mathcal{D})$ instead of f, the natural mean-square measure as a predictor is:

$$\mathcal{M}(f, y) = \mathbb{E}\left[((y - f(x; \mathcal{D})))^2 \mid x, \mathcal{D}\right] \tag{1}$$

for $\mathbb{E}[\cdot]$ the expectation wrt to a distribution P. Decomposing the right-hand side, we have:

$$\mathcal{M}(f,y) = \mathbb{E}\left[((y - f(x;\mathcal{D})))^2 \mid x, \mathcal{D}\right] = \mathbb{E}\left[(y - \mathbb{E}[y \mid x])^2 \mid x, \mathcal{D}\right] + (f(x;\mathcal{D}) - \mathbb{E}[y \mid x])^2 \quad (2)$$

Here, $\mathbb{E}\left[(y - \mathbb{E}[y \mid x])^2 \mid x, \mathcal{D}\right]$ does not depend on $\mathcal{D}$, but simply the statistical variance of y given x. The term $(f(x;\mathcal{D}) - \mathbb{E}[y \mid x])^2$ is considered a natural measure of effectiveness on $\mathbb{R}^n$ as a singular predictor of y. Now, for $\mathbb{E}_{\mathcal{D}}\left[(f(x;\mathcal{D}) - \mathbb{E}[y \mid x])^2\right]$ which depends on the training set $\mathcal{D}$ in its computation, is decomposed into the form of *bias-variance decomposition* terms, by derivation:

$$\mathbb{E}_{\mathcal{D}}\left[(f(x;\mathcal{D}) - \mathbb{E}[y \mid x])^2\right] = \underbrace{\{\mathbb{E}_{\mathcal{D}}[f(x;\mathcal{D})] - \mathbb{E}[y \mid x]\}^2}_{\text{bias term}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left\{(f(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(x;\mathcal{D})])^2\right\}}_{\text{variance term}} \quad (3)$$

We summarize this in the following statement.

Theorem 2.1 (Bias-variance decomposition)**.** *Suppose the model $f(x;\mathcal{D})$ for the data $\mathcal{D} = (x_i, y_i)$ and its parameter x is defined. For y_i of the target concept's responses y, and consider a regression problem with the loss measure $\mathcal{M}(f,y)$ of mean squared risk, the following statement is true:*

$$\mathbb{E}\big[\mathcal{M}(f,y)\big] = \mathcal{B}(f,y) + \mathcal{V}(f,y) + \mathbb{E}\Big[\mathbb{E}\left[((y - f(x;\mathcal{D})))^2 \mid x, \mathcal{D}\right]\Big] \quad (4)$$

for $\mathbb{E}[\,\cdot \mid x, \mathcal{D}]$ any expression with dependencies on x and $\mathcal{D}$. The bias and variance term is subsequently expressed by

$$\mathcal{B}(f,y) = \underbrace{\{\mathbb{E}_{\mathcal{D}}[f(x;\mathcal{D})] - \mathbb{E}[y \mid x]\}^2}_{\textit{bias}}, \quad \mathcal{V}(f,y) = \underbrace{\mathbb{E}_{\mathcal{D}}\left\{(f(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(x;\mathcal{D})])^2\right\}}_{\textit{variance}} \quad (5)$$

The above decomposition principle is often expressed into a form where there exists the intrinsic noise Brown & Ali (2024):

$$\begin{aligned}\mathbb{E}_D\left[\mathbb{E}_{xy}\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}_x\left[\left(y^* - \mathbb{E}_D[\hat{f}(x)]\right)^2\right] + \mathbb{E}_x\left[\mathbb{E}_D\left(\hat{f}(x) - \mathbb{E}_D[\hat{f}(x)]\right)^2\right] \\ + \mathbb{E}_{xy}\left[(y - y^*)^2\right]\end{aligned} \quad (6)$$

For simplicity of notation, we adopt the similar form in the standard case of Adlam & Pennington (2020):

$$\mathbb{E}\left[\hat{y}(\mathbf{x}) - y(\mathbf{x})\right]^2 = \left(\mathbb{E}\hat{y}(\mathbf{x}) - \mathbb{E}y(\mathbf{x})\right)^2 + \mathbb{V}\left[\hat{y}(\mathbf{x})\right] + \mathbb{V}[y(\mathbf{x})] \quad (7)$$

A main common theme of criticism toward bias-variance tradeoff is the fact that the decomposition is much more general, and intrinsic for the class of *mean squared loss*. However, when considering the naturalness of an error measure and then its direct applicant, mean square error on real space of sufficient support and measure comes of as a very natural choice of an error measure, at least in consideration of the regression setting; for that, we then can also define classification as another top-layer above a regression's similar real continuous space, i.e. a decision layer. Furthermore, it can also be shown Brown & Ali (2024); Pfau (2013) that it also holds for the class of Bregman divergence measure.

What does the decomposition say about model selection principle in general? In general, bias-variance is typically presented of its decomposition only, and the asymptotic prediction of its behaviours. In fact, one of the reason that it became the rule-of-thumb for ML practitioner, as well as generally statistical learning (Lafon & Thomas (2024) provides a quite rigorous treatment of bias-variance tradeoff in the section on statistical learning theory) solidify the trade-off as a particular model selection principle. Generally, this tradeoff can be summarized as followed:

Theorem 2.2 (Bias-variance tradeoff)**.** *For the expected loss of any given hypothesis h, the bias $\mathcal{B}(f,y)$ and variance $\mathcal{V}(f,y)$ is inversely proportional, that is, $\mathcal{B}(f,y) \propto \lambda^{-1}\mathcal{V}(f,y)$ for some proportionality λ that may or may not be constant. In the most general case possible, $\lambda = -1$ on the entire error range.*

The tradeoff is then of inverse proportionality. Indeed, statistically, we have such tradeoff on a statistical framework in a more concrete sense. For the bias to increase, variance will increase, of which the criterion is inverse - we would like to have more bias but lower variance, and vice versa, according to such theory.

2.2 DOUBLE DESCENT

Nevertheless, of bias-variance and the analogous statistical learning theory concept, the target is the same. It is the dilemma of which is presented in **Occam's razor**, for choosing the sufficient model of good complexity, or bias, for tradeoff of its generalization ability, or variance. Then there must exist a sweet spot between the axis of bias and variance, since they are as exhibited above inversely proportional to each other. However, double descent seemingly broke the status quo, and insists on an interesting phenomenon - under the same setting, if we 'crank' the complexity high enough, we will then reach a point then called the **interpolation threshold**, such that the trend reverse and the error rate, instead of being theorized to go up, goes down to a certain line of lower bound.

The first identification of the double descent phenomena dated back to the paper of Belkin - Belkin et al. (2019), in which the title is literally "reconciling" modern machine learning practice and the bias-variance tradeoff. In modern machine learning practice, or state-of-the-art developments, models are now bigger than ever. If to notice, we will see that currently models are inherently large, for example, a normal large language model will have from 900 millions (900M) to a few billions, for example 10 billions (10B) parameters. That is not taking into account the overall dynamics and structure of the model, which dictates the operating range and efficiency of the model itself. These model, based on the neural network architecture are somewhat trained to exactly fit (or interpolate) the data, almost certainly so that it turn from a prediction setting to an estimation setting. By statistical learning theory, this would be considered overfitting, and yet, they often obtain very high accuracy on test data.

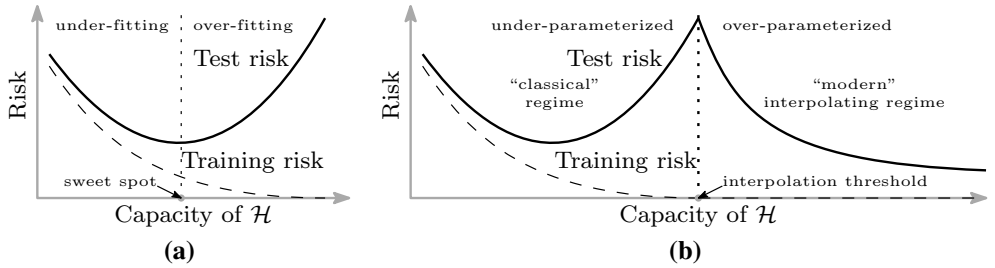

Figure 1: **Curves for training risk (dashed line) and test risk (solid line).** (**a**) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (**b**) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behaviour from using high capacity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk. Reproduced from Belkin et al. (2019).

The main finding that Belkin found is a pattern for how the apparent performance on unseen data depends on model capacity and the mechanism underlying the emergence of double descent. When function class capacity is below the "interpolation threshold", learned predictors exhibit the classical U-shaped curve from Figure 1. The 'modern' interpolating regime marks the opposite trend to the right, where the risk starts to decrease up to a lower bound, which then can be called the *optimal descent bound*.

> The bottom of the U is achieved at the sweet spot which balances the fit to the training data and the susceptibility to over-fitting: to the left of the sweet spot, predictors are under-fit, and immediately to the right, predictors are over-fit. When we increase the function class capacity high enough (e.g., by increasing the number of features or the size of the neural network architecture), the learned predictors achieve (near) perfect fits to the training data—i.e., interpolation. Although the learned predictors obtained at the interpolation threshold typically have high risk, we show that increasing the function class capacity beyond this point leads to decreasing risk, typically going below the risk achieved at the sweet spot in the "classical" regime. (Belkin et al. (2019))

Another prominent result to look at is Nakkiran et al. (2019), on the double descent of deep learning models. This is the first step toward identifying double descent to be perhaps, universal.

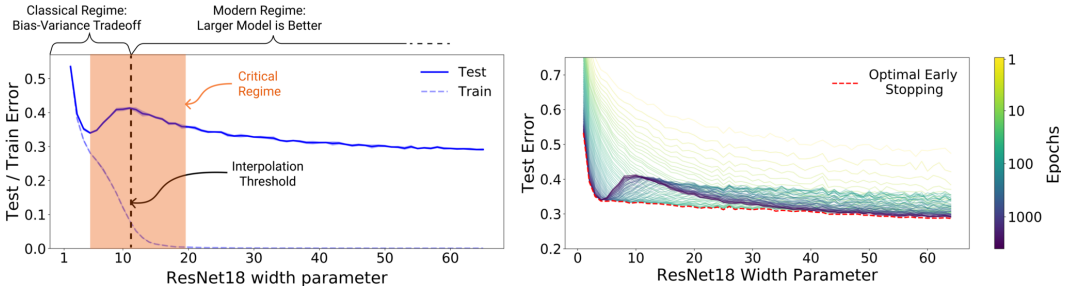

Figure 2: **Left:** Train and test error as a function of model size, for ResNet18s of varying width on CIFAR-10 with 15% label noise. **Right:** Test error, shown for varying train epochs. All models trained using Adam for 4K epochs. The largest model (width 64) corresponds to standard ResNet18. Resued from Nakkiran et al. (2019).

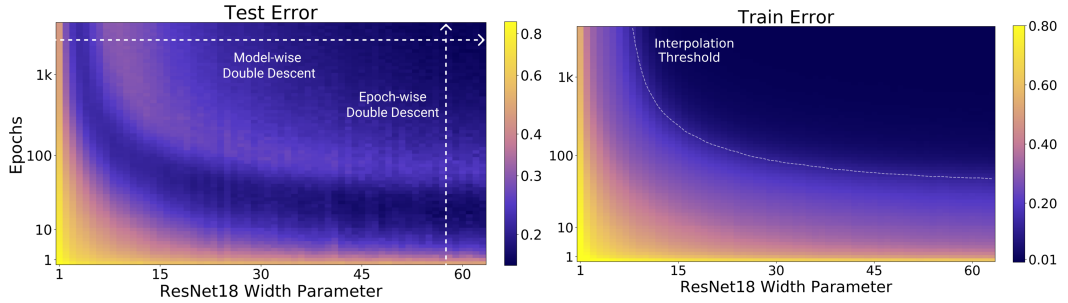

Figure 3: **Left:** Test error as a function of model size and train epochs. The horizontal line corresponds to model-wise double descent–varying model size while training for as long as possible. The vertical line corresponds to epoch-wise double descent, with test error undergoing double-descent as train time increases. **Right** Train error of the corresponding models. All models are Resnet18s trained on CIFAR-10 with 15% label noise, data-augmentation, and Adam for up to 4K epochs. Reused from Nakkiran et al. (2019)

They define *effective model complexity* of $\mathcal{T}$ (w.r.t. distribution $\mathcal{D}$) to be the maximum number of samples n on which $\mathcal{T}$ achieves on average ≈ 0 *training error*. This is an entirely empirical definition, similarly per definition as VC-dimension.

Definition 2.1 (Effective Model Complexity)**.** *The* Effective Model Complexity *(EMC) of a training procedure $\mathcal{T}$, with respect to distribution $\mathcal{D}$ and parameter $\epsilon > 0$, is defined as:*

$$\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) := \max\{n \mid \mathbb{E}_{S\sim\mathcal{D}^n}[\mathrm{Error}_S(\mathcal{T}(S))] \leq \epsilon\}$$

where $\mathrm{Error}_S(M)$ is the mean error of model M on train samples S.

Using this definition, their main hypothesis can then be stated as the following three-fold regions:

Hypothesis 2.1 (Generalized Double Descent hypothesis, informal)**.** *For any natural data distribution $\mathcal{D}$, neural-network-based training procedure $\mathcal{T}$, and small $\epsilon > 0$, if we consider the task of predicting labels based on n samples from $\mathcal{D}$ then:*

Under-parameterized regime. *If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ is sufficiently smaller than n, any perturbation of $\mathcal{T}$ that increases its effective complexity will decrease the test error.*

Over-parameterized regime. *If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ is sufficiently larger than n, any perturbation of $\mathcal{T}$ that increases its effective complexity will decrease the test error.*

Critically parameterized regime. *If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) \approx n$, then a perturbation of $\mathcal{T}$ that increases its effective complexity might decrease* **or increase** *the test error.*

It is to notice that this definition is also only observational. By that, it only outlines the specific region of interest where the paradigm shift is identified through experimental results. It has no effective

predictive power, or rather descriptive power more than setting up hypothesis with respect to the effective model complexity, and some arbitrary perturbation. Nakkiran et al. (2019) also noticed this difficulty in providing a theoretical definition and theorem regarding such hypothesis, as said in their manuscript. The existence of the critical regime is also not defined.

The behaviour itself has been particularly investigated, in certain settings. For example, before Belkin et al. (2019), Advani & Saxe (2017) investigated the generalization error in neural networks of high-dimensional measure. Of such, various behaviours that bear similarity to double descent can be observed. Belkin et al. (2018) expanded on his previous research on the particular subset of kernel learning models, providing a theoretical analysis of specifically the Laplacian kernel on standard neural network. Mei & Montanari (2020) investigated random feature regression in such regard, and also found similar results.

The fact that double descent is not well-defined itself is a problem on its own. Up to the author's knowledge and of myself, there has been no effective definition or given description that outlines specifically how double descent can be structured. Instability in reproducing experiments and the effective range of the phenomena remains a question on its own.

REFERENCES

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition, 2020. URL https://arxiv.org/abs/2011.03321.

Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks, 2017. URL https://arxiv.org/abs/1710.03667.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning, 2018. URL https://arxiv.org/abs/1802.01396.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.*, 116(32):15849–15854, August 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1903070116. URL http://arxiv.org/abs/1812.11118. arXiv:1812.11118 [cs, stat].

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning, 2020. URL https://arxiv.org/abs/2011.04483.

Gavin Brown and Riccardo Ali. Bias/variance is not the same as approximation/estimation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=4TnFbv16hK.

Stéphane d' Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where & why do they appear? In *Advances in Neural Information Processing Systems*, volume 33, pp. 3058–3069. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1fd09c5f59a8ff35d499c0ee25a1d47e-Abstract.html.

Xander Davies, Lauro Langosco, and David Krueger. Unifying Grokking and Double Descent, March 2023. URL http://arxiv.org/abs/2303.06173. arXiv:2303.06173 [cs].

Pedro M. Domingos. A unified bias-variance decomposition for zero-one and squared loss. In *AAAI/IAAI*, 2000. URL https://api.semanticscholar.org/CorpusID:2063488.

George Casella E. L. Lehmann. *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, New York, 1998. ISBN 978-0-387-98502-2. doi: 10.1007/b98854. URL http://link.springer.com/10.1007/b98854.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992. doi: 10.1162/neco.1992.4.1.1.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Bruce Hajek and Maxim Raginsky. *Statistical Learning Theory*, volume 1. 2021. URL https://maxim.ece.illinois.edu/teaching/SLT/.

Arthur Jacot, Fran¸cois Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. Kernel view of wide-network behavior.

Gareth James, Trevor Hastie, Robert Tibshirani, and Daniela Witten. *An introduction to statistical learning : with applications in R*. New York : Springer, [2013] ©2013, 2013. URL https://search.library.wisc.edu/catalog/9910207152902121.

Marc Lafon and Alexandre Thomas. Understanding the Double Descent Phenomenon in Deep Learning, March 2024. URL http://arxiv.org/abs/2403.10459. arXiv:2403.10459 [cs, stat].

Chris Yuhao Liu and Jeffrey Flanigan. Understanding the role of optimization in double descent, 2023. URL https://arxiv.org/abs/2312.03951.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve, 2020. URL https://arxiv.org/abs/1908.05355.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 026201825X.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt, December 2019. URL http://arxiv.org/abs/1912.02292. arXiv:1912.02292 [cs, stat].

Amanda Olmin and Fredrik Lindsten. Towards understanding epoch-wise double descent in two-layer linear neural networks, 2024. URL https://arxiv.org/abs/2407.09845.

Liam Paninski. Statistics 4107: Intro to Math Stat (fall 2005), 2005. URL https://sites.stat.columbia.edu/liam/teaching/4107-fall05/.

David Pfau. A generalized bias-variance decomposition for bregman divergences. Technical report, 2013.

Rylan Schaeffer, Mikail Khona, Zachary Robertson, Akhilan Boopathy, Kateryna Pistunova, Jason W. Rocks, Ila Rani Fiete, and Oluwasanmi Koyejo. Double Descent Demystified: Identifying, Interpreting & Ablating the Sources of a Deep Learning Puzzle, March 2023. URL http://arxiv.org/abs/2303.14151. arXiv:2303.14151 [cs, stat].

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer: New York, 1999.