Mila Deep Learning Theory Group

“The double descent phenomenon”

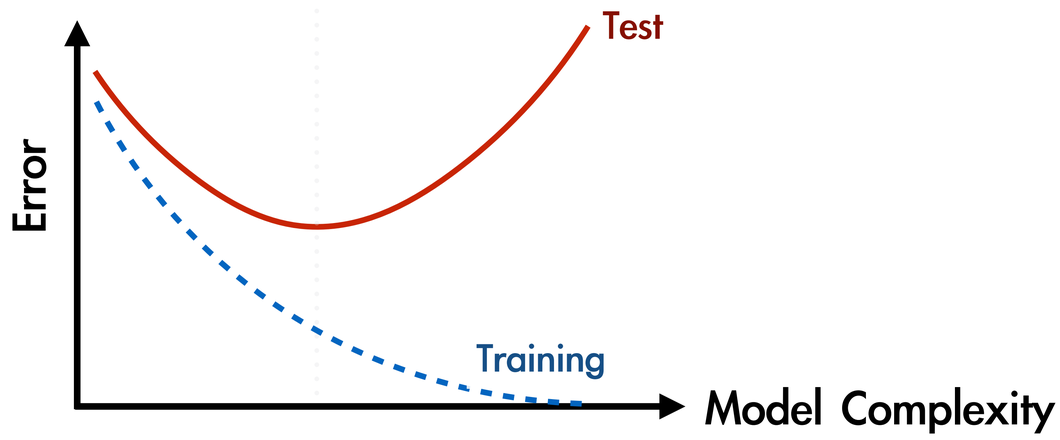
By Mohammad Pezeshki
July 9th, 2020

- A ~20 minutes presentation.
- Followed by ~30 minutes free discussion.
- Finally ~10 minutes conclusion.

1. What is Double Descent?
2. The history of Double Descent.
3. How people justify it?

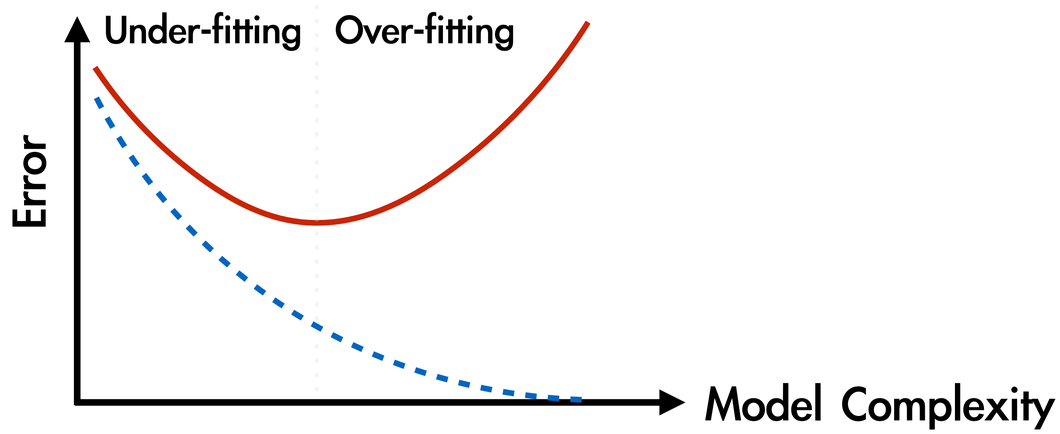
1. What is Double Descent?

Conventional wisdom:



1. What is Double Descent?

Conventional wisdom:

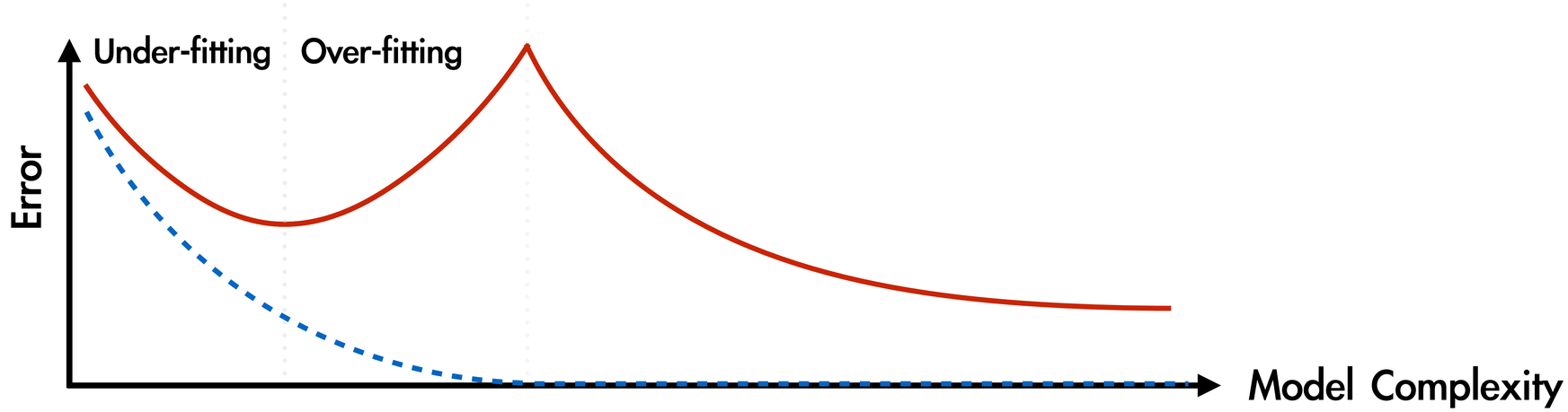


1. What is Double Descent?

Belkin et al (2018):

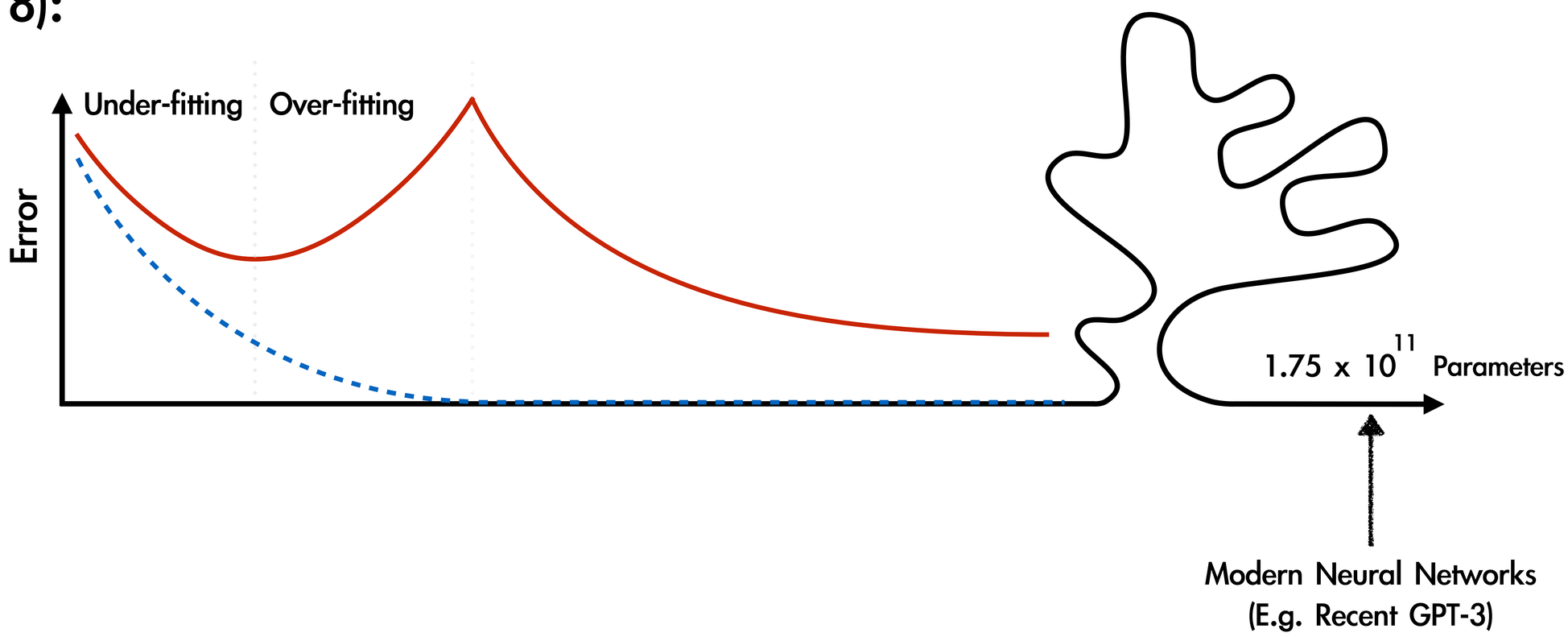
1. What is Double Descent?

Belkin et al (2018):



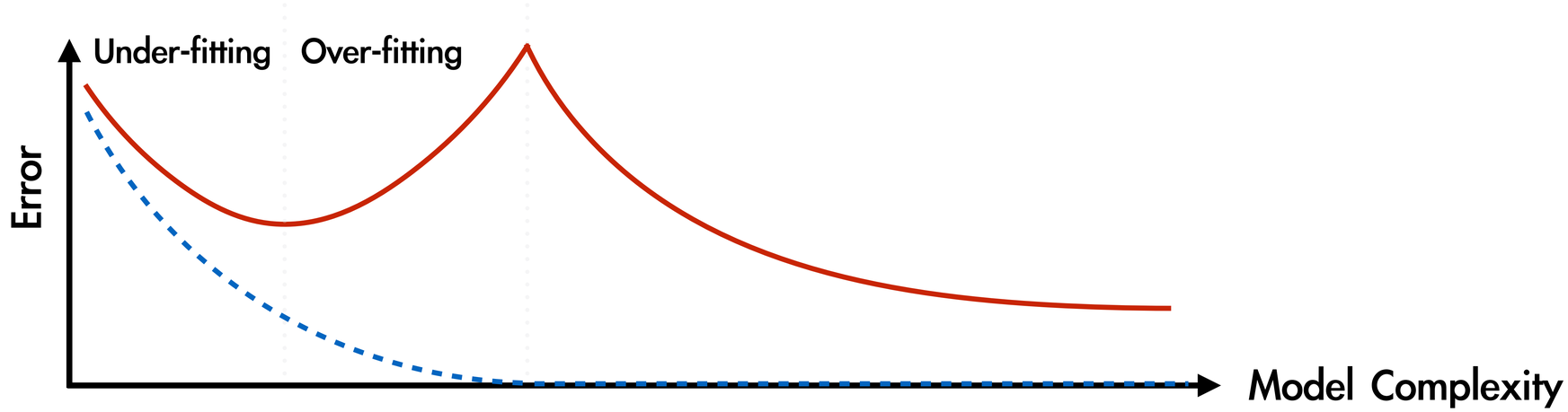
1. What is Double Descent?

Belkin et al (2018):

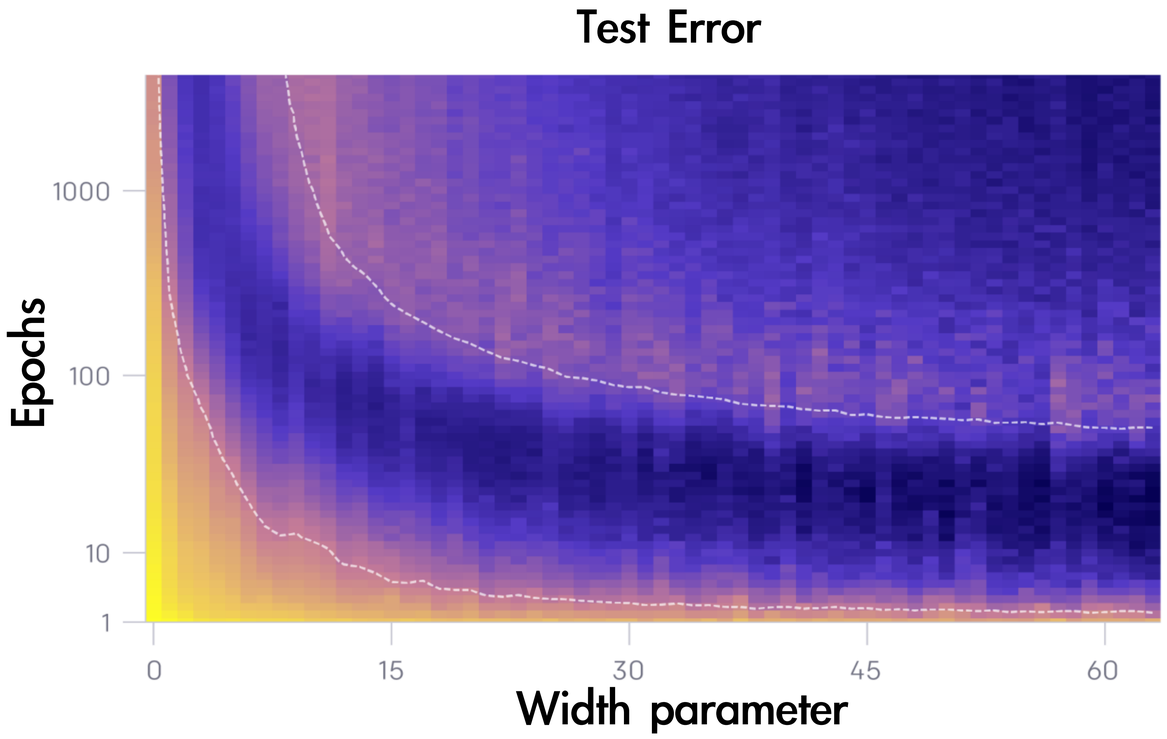


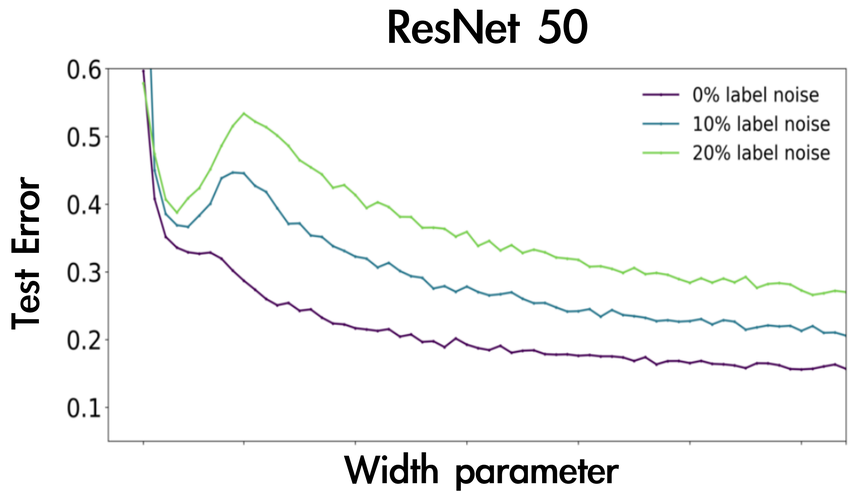
1. What is Double Descent?

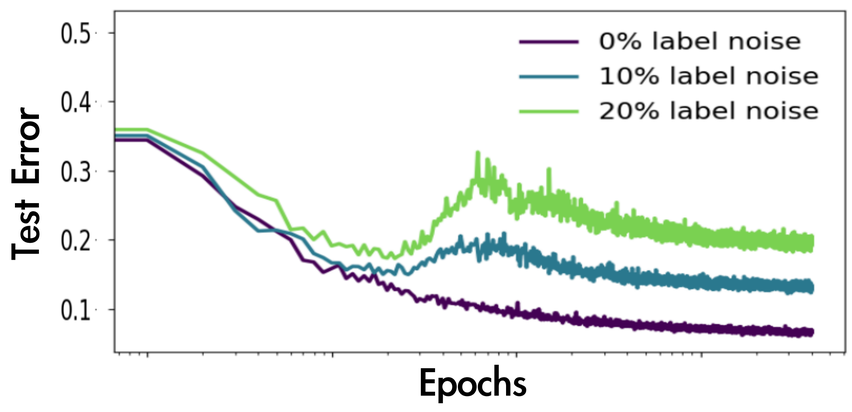
Belkin et al (2018):



Nakkiran et al (2019):

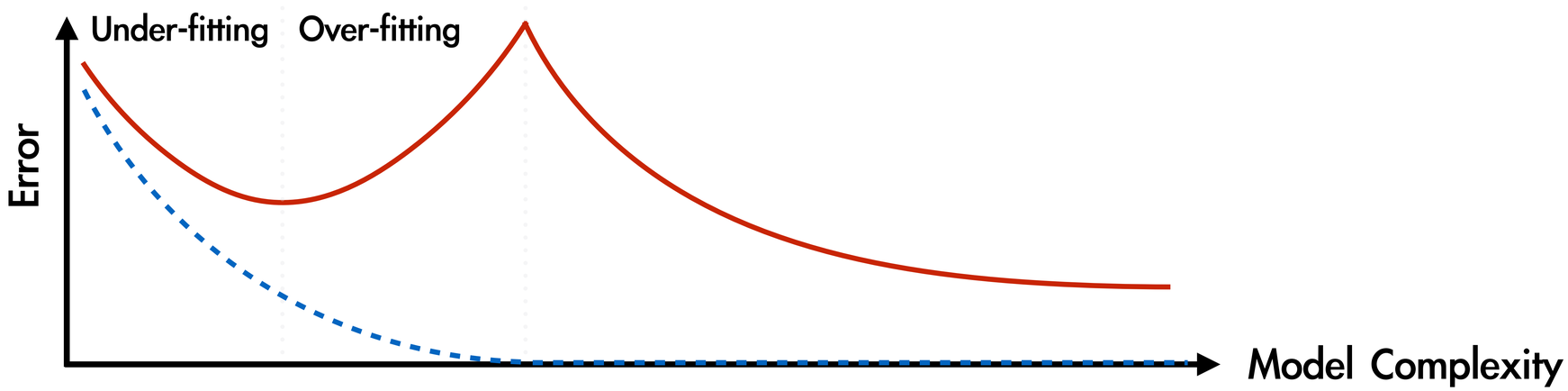




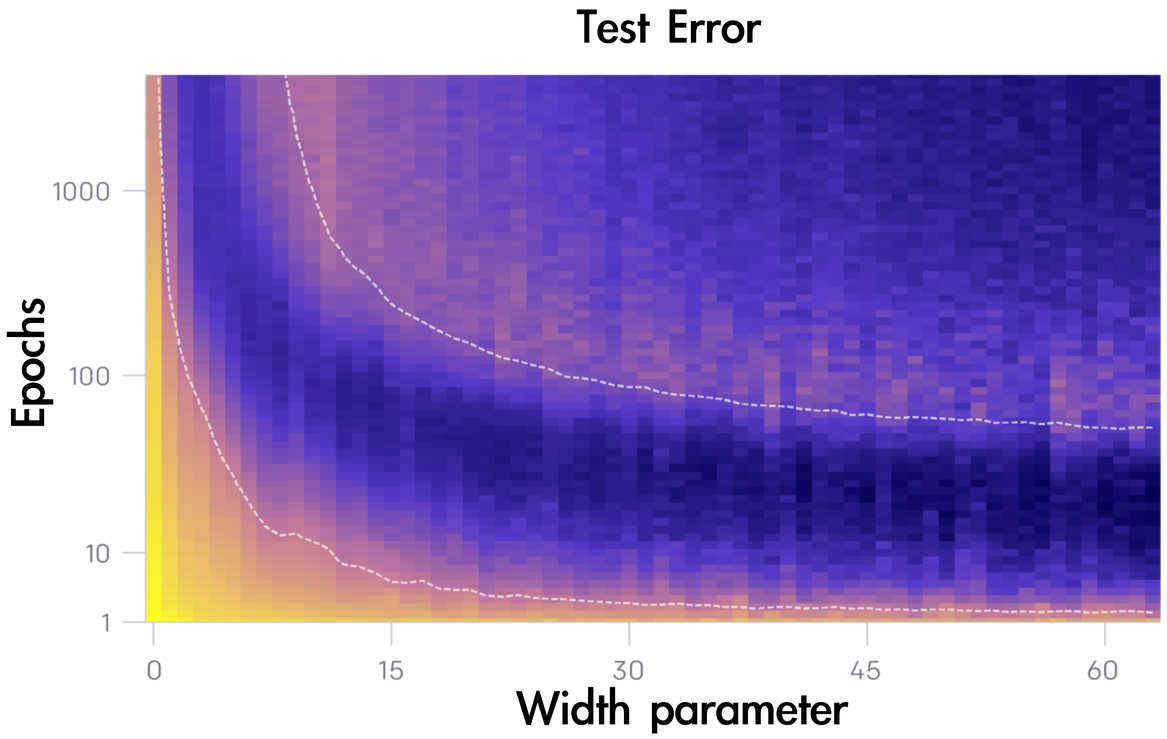


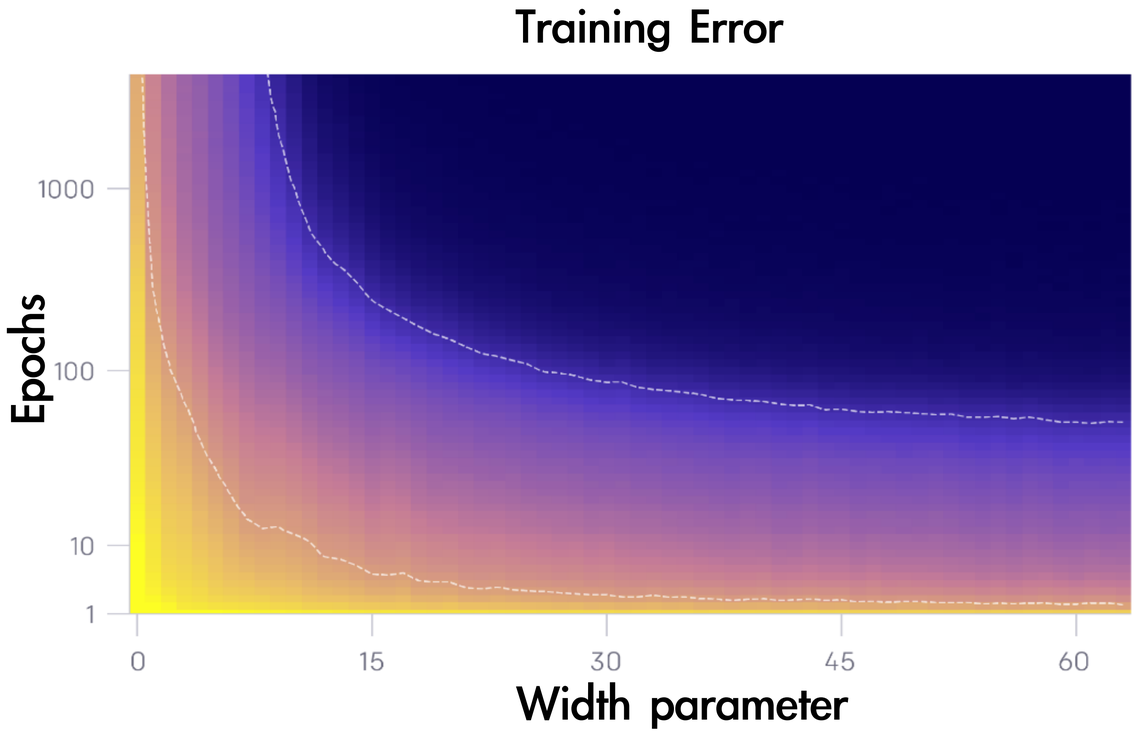
1. What is Double Descent?

Belkin et al (2018):



Nakkiran et al (2019):





2. The history of Double Descent

Belkin et al (Dec. 2018):

Reconciling modern machine learning practice and the bias-variance trade-off

Mikhail Belkin[a], Daniel Hsu[b], Siyuan Ma[a], and Soumik Mandal[a]

[a]*The Ohio State University, Columbus, OH*
[b]*Columbia University, New York, NY*

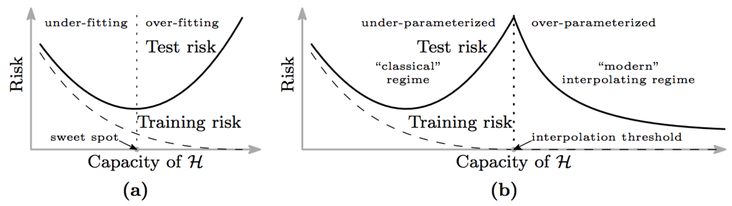
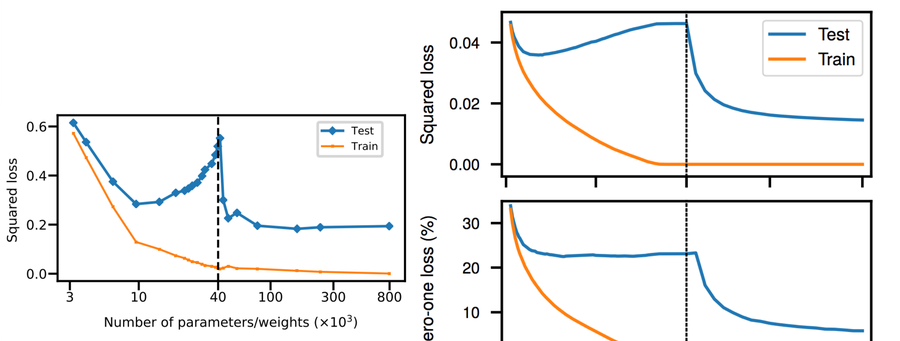


Figure 1: **Curves for training risk (dashed line) and test risk (solid line).** (**a**) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (**b**) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior 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.



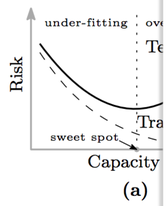
2. The history of Double Descent

Belkin et al (Dec. 2018):

Reconciling
and

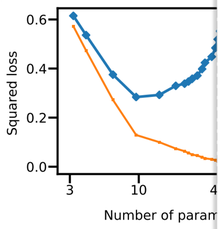
Mikhail Belkin[a],

[a]*The*
[b]*C*



(a)

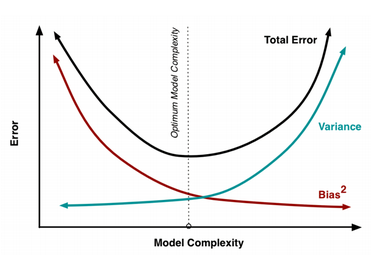
Figure 1: **Curves for t**
U-shaped risk curve ari
which incorporates the
behavior from using hig
arated by the interpola
have zero training risk.



Neal et al (Oct. 2018)

A Modern Take on the Bias-Variance Tradeoff in Neural Networks

Brady Neal **Sarthak Mittal** **Aristide Baratin** **Vinayak Tantia** **Matthew Scicluna**
Simon Lacoste-Julien[†,‡] **Ioannis Mitliagkas**[†]
Mila, Université de Montréal
[†]Canada CIFAR AI Chair [‡]CIFAR Fellow



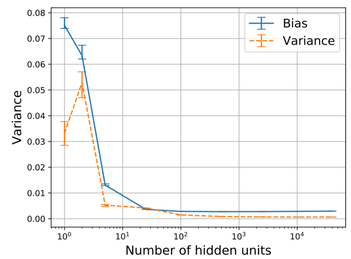
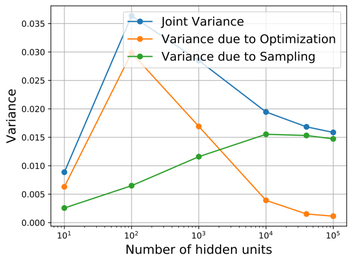


Figure 1: On the left is an illustration of the common intuition for the bias-variance tradeoff (Fortmann-Roe, 2012). We find that *both* bias and variance decrease when we increase network width on MNIST (right) and other datasets (Section 4). These results seem to contradict the traditional intuition of a strict tradeoff.



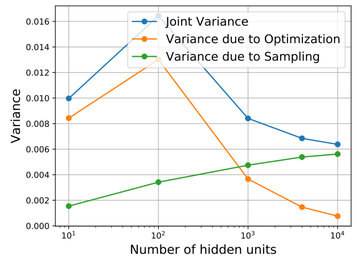


Figure 2: Trends of variance due to sampling and variance due to optimization with width on CIFAR10 (left) and on SVHN (right). Variance due to optimization decreases with width, once in the over-parameterized setting. Variance due to sampling plateaus and remains constant. This is in contrast with what the bias-variance tradeoff would suggest.

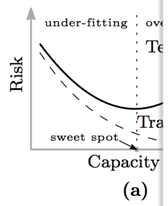
2. The history of Double Descent

Belkin et al (Dec. 2018):

Reconciling
and

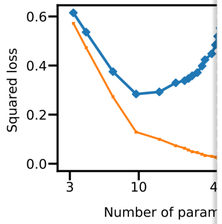
Mikhail Belkin[a],

[a]*The*
[b]*C*



(a)

Figure 1: **Curves for t**
U-shaped risk curve ari
which incorporates the
behavior from using hig
arated by the interpola
have zero training risk.



Neal et al (Oct. 2018)

A M

Brady Neal Sa

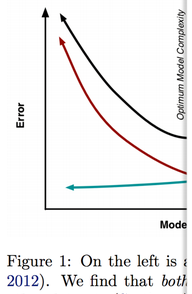


Figure 1: On the left is
2012). We find that *both*
other datasets (Section 4)

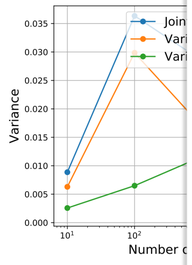


Figure 2: Trends of varian
on SVHN (right). Varian
Variance due to sampling
would suggest.

Advani & Saxe (2017)

High-dimensional dynamics of generalization error in neural networks

Madhu S. Advani*
Center for Brain Science
Harvard University
Cambridge, MA 02138

Andrew M. Saxe*
Center for Brain Science
Harvard University
Cambridge, MA 02138

MADVANI@FAS.HARVARD.EDU

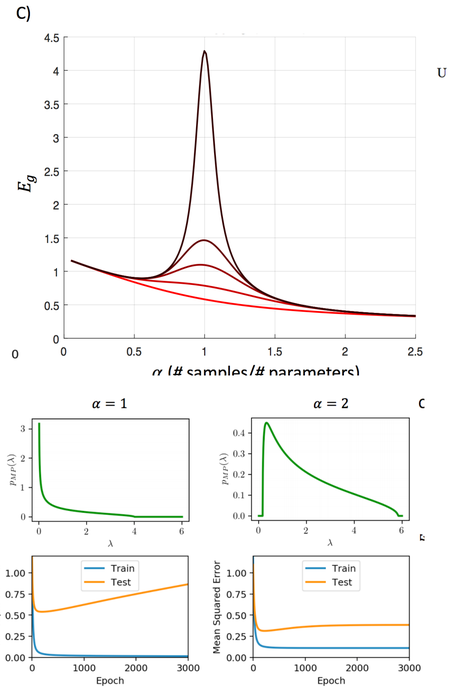


Figure 2: The Marchenko-Pasteur distribution and high-dimensional learning dynamics. A) Different ratios of number training samples (P) to network parameters (N) ($\alpha = \frac{P}{N}$) yield different eigenvalue densities in the input correlation matrix. For large N, this density is described by the MP distribution (14), which consists of a 'bulk' lying between $[\lambda_-, \lambda_+]$, and, when $\alpha < 1$, an additional delta function spike at zero. When there are fewer samples than parameters ($\alpha < 1$, left column), some fraction of eigenvalues are exactly zero (delta-function arrow at origin), and the rest are appreciably greater than zero. When the number of samples is on the order of the parameters ($\alpha = 1$, center column), the distribution diverges near the origin and there are many nonzero but arbitrarily small eigenvalues. When there are more samples than parameters ($\alpha > 1$, right column), the smallest eigenvalues are appreciably greater than zero. B) Dynamics of learning. From (13), the generalization error is harmed most by small eigenvalues; and these are the slowest to be learned. Hence for $\alpha = 1/2$ and $\alpha = 2$, the gap in the spectrum near

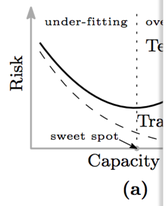
2. The history of Double Descent

Belkin et al (Dec. 2018):

Reconciling
and

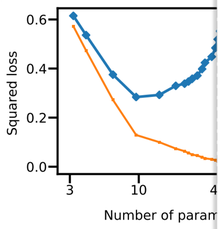
Mikhail Belkin[a],

[a]The
[b]C



(a)

Figure 1: **Curves for t**
U-shaped risk curve ari
which incorporates the
behavior from using hig
arated by the interpola
have zero training risk.



Neal et al (Oct. 2018)

A M

Brady Neal Sa

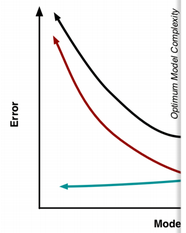


Figure 1: On the left is
2012). We find that *both*
other datasets (Section 4)

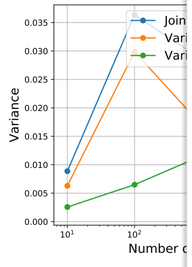


Figure 2: Trends of varian
on SVHN (right). Varian
Variance due to sampling
would suggest.

Advani & Saxe (2017)

High-d

Madhu S. Adva
Center for Brain Sc
Harvard University
Cambridge, MA 021

Andrew M. Sax
Center for Brain Sc
Harvard University
Cambridge, MA 021

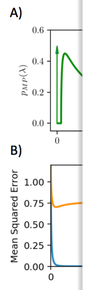


Figure 2: The March
Different ra
$\frac{P}{N}$) yield di
N, this den
lying betwe
zero. Wher
fraction of
rest are ap
order of the
origin and
are more sa
are appreci
generalizat
to be learn

Yann Le Cun et al. (1991)

VOLUME 66, NUMBER 18 PHYSICAL REVIEW LETTERS 6 MAY 1991

Eigenvalues of Covariance Matrices: Application to Neural-Network Learning

Yann Le Cun,[(1)] Ido Kanter,[(2)] and Sara A. Solla[(1)]
[(1)]*AT&T Bell Laboratories, Holmdel, New Jersey 07733*
[(2)]*Department of Physics, Bar Ilan University, Ramat Gan, 52100, Israel*

(Received 2 January 1991)

The learning time of a simple neural-network model is obtained through an analytic computation of the eigenvalue spectrum for the Hessian matrix, which describes the second-order properties of the objective function in the space of coupling coefficients. The results are generic for symmetric matrices obtained by summing outer products of random vectors. The form of the eigenvalue distribution suggests new techniques for accelerating the learning process, and provides a theoretical justification for the choice of centered versus biased state variables.

$$\rho(\lambda) = -\frac{2}{N\pi}\mathrm{Im}\frac{\partial}{\partial\lambda}\lim_{n\to 0}\frac{1}{n}\{[\det{}^{-1/2}(\mathbf{1}\lambda-\mathbf{R})]^n-1\}\,, \tag{10}$$

the Fresnel representation yields

$$\rho(\lambda) = -\frac{2}{N\pi}\mathrm{Im}\frac{\partial}{\partial\lambda}\lim_{n\to 0}\frac{1}{n}\left\{\left[\frac{e^{i\pi/4}}{\sqrt{\pi}}\right]^{Nn}\int_{-\infty}^{\infty}\prod_k^N dy_k\exp\left[-i\sum_{k,l}^{N}\sum_{\gamma}^{n}y_k^\gamma(\lambda\delta_{kl}-R_{kl})y_l^\gamma\right]-1\right\}. \tag{11}$$

The expression in curly brackets in Eq. (11) can be written as

$$\{\cdots\} = \int\prod_{\gamma\beta}^{n}\frac{dq_{\gamma\beta}\,d\varphi_{\gamma\beta}}{2\pi}\int\prod_{\gamma}^{n}\frac{dM_\gamma}{2\pi}\exp\left[N\left\{i\sum_{\gamma\beta}\varphi_{\gamma\beta}q_{\gamma\beta}\ln\left[\int\prod_{\gamma}^{n}dy_\gamma\exp\left[-i\lambda\sum_{\gamma}y_\gamma^2-i\sum_{\gamma\beta}\varphi_{\gamma\beta}y_\gamma y_\beta\right]\right]\right.\right.$$
$$\left.\left.+\alpha\ln\left[\int\prod_{\gamma}^{n}\frac{dt_\gamma}{2\pi}\exp\left[-\tfrac{1}{2}\sum_{\gamma}t_\gamma^2+\sqrt{2i}m\sum_{\gamma}t_\gamma M_\gamma+iv\sum_{\gamma\beta}t_\gamma t_\beta q_{\gamma\beta}\right]\right]\right\}\right], \tag{12}$$

and can be evaluated in the $N\to\infty$ thermodynamic limit using a saddle-point method.

E through gradient descent. If the prescribed error $\tilde{E}$ is

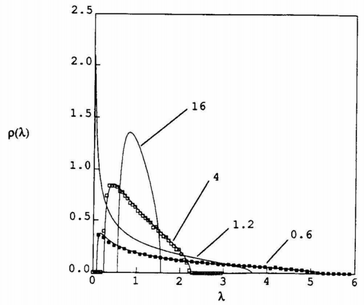


FIG. 1. Spectral density $\rho(\lambda)$ predicted by Eq. (13) for $m=0$, $v=1$, and $\alpha=0.6$, 1.2, 4, and 16. Experimental histograms for $\alpha=0.6$ (solid squares) and $\alpha=4$ (open squares) are averages over 100 trials with $N=200$ and $x_i^\mu=\pm 1$ with probability $\frac{1}{2}$ each.

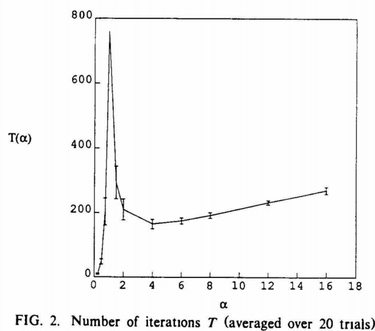


FIG. 2. Number of iterations T (averaged over 20 trials) needed to train a linear neuron with $N=100$ inputs. The x_i^μ are uniformly distributed between -1 and $+1$. Initial and target couplings $\mathbf{W}$ are chosen randomly from a uniform distribution within the $[-1,+1]^N$ hypercube. Gradient descent, with $\eta=1/2\lambda_{\max}$, is considered complete when the error reaches the prescribed value $\tilde{E}=0.001$ above the $E_0=0$ minimum value.

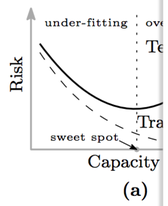
2. The history of Double Descent

Belkin et al (Dec. 2018):

Reconciling
and

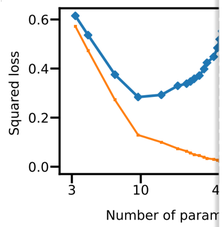
Mikhail Belkin[a],

[a]The
[b]C



(a)

Figure 1: **Curves for t**
U-shaped risk curve ari
which incorporates the
behavior from using hig
arated by the interpola
have zero training risk.



Neal et al (Oct. 2018)

A M

Brady Neal Sar

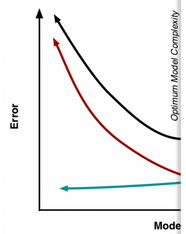


Figure 1: On the left is
2012). We find that *both*
other datasets (Section 4)

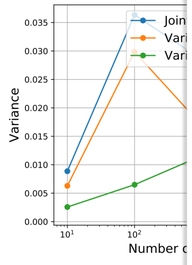


Figure 2: Trends of varian
on SVHN (right). Varian
Variance due to sampling
would suggest.

Advani & Saxe (2017)

High-d

Madhu S. Adva
Center for Brain Sc
Harvard University
Cambridge, MA 021

Andrew M. Sax
Center for Brain Sc
Harvard University
Cambridge, MA 021

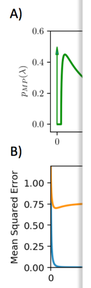


Figure 2: The March
Different ra
$\frac{P}{N}$) yield di
N, this den
lying betwe
zero. Wher
fraction of
rest are ap
order of the
origin and
are more sa
are appreci
generalizati
to be learn

Yann Le Cun et al. (1991)

VOLUME 66, NUMBER 18

Eigenvalues of C

[2]*Dep*

The learning time
the eigenvalue spectr
jective function in th
tained by summing o
new techniques for
choice of centered ver

$$\rho(\lambda) = -\frac{2}{N\pi}\mathrm{Im}\frac{\delta}{\delta}$$

the Fresnel representati

$$\rho(\lambda) = -\frac{2}{N\pi}\mathrm{Im}\frac{\delta}{\delta}$$

The expression in curly

$$\{\cdots\} = \int \prod_{\gamma\beta}^{n} \frac{dq_{\gamma\beta}\, d\varphi_{\gamma\beta}}{2\pi}$$

and can be evaluated in

E through gradient descent

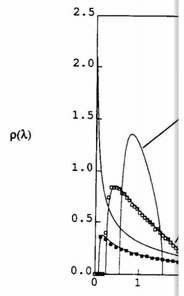


FIG. 1. Spectral density
$m=0$, $v=1$, and $\alpha=0.6$, 1.2,
grams for $\alpha=0.6$ (solid squar
averages over 100 trials with N
bility $\frac{1}{2}$ each.

2398

Krogh and Hertz (1991)

Generalization in a linear perceptron in the presence of noise

Anders Krogh†‡ and John A Hertz§
† The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen, Denmark
§ Nordita, Blegdamsvej 17, DK-2100 Copenhagen, Denmark

Received 2 April 1991, in final form 4 November 1991

Abstract. We study the evolution of the generalization ability of a simple linear perceptron with N inputs which learns to imitate a 'teacher perceptron'. The system is trained on $p = \alpha N$ example inputs drawn from some distribution and the generalization ability is measured by the average agreement with the teacher on test examples drawn from the same distribution. The dynamics may be solved analytically and exhibits a phase transition from imperfect to perfect generalization at $\alpha = 1$, when there are no errors (static noise) in the training examples. If the examples are produced by an erroneous teacher, overfitting is observed, i.e. the generalization error starts to increase after a finite time of training. It is shown that a weight decay of the same size as the variance of the noise (errors) on the teacher improves on the generalization and suppresses the overfitting. The generalization error as a function of time is calculated numerically for various values of the parameters. Finally dynamic noise in the training is considered. White noise on the input corresponds on average to a weight decay, and can thus improve generalization, whereas white noise on the weights or the output degrades generalization. Generalization is particularly sensitive to noise on the weights (for $\alpha < 1$) where it makes the error constantly increase with time, but this effect is also shown to be damped by a weight decay. Weight noise and output noise acts similarly above the transition at $\alpha = 1$.

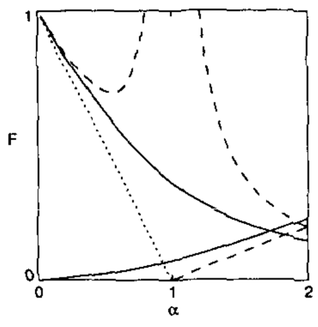


Figure 2. Generalization and training error with static noise of size $\sigma^2 = 0.2$ on the input or the output of the teacher. Curves starting in the upper left corner represent the generalization error F while the lower ones represent the training error. The full curve is for $\lambda = \sigma^2$, and the broken curve $\lambda = 0$. The dotted line is the generalization error with no noise and $\lambda = 0$ for reference.

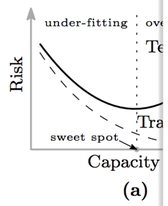
2. The history of Double Descent

Belkin et al (Dec. 2018):

Reconciling
and

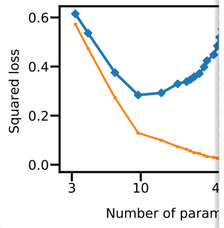
Mikhail Belkin[a],

[a]The
[b]C



(a)

Figure 1: **Curves for t**
U-shaped risk curve ari
which incorporates the
behavior from using hig
arated by the interpola
have zero training risk.



Neal et al (Oct. 2018)

A M

Brady Neal Sa

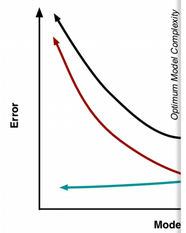


Figure 1: On the left is
2012). We find that *both*
other datasets (Section 4)

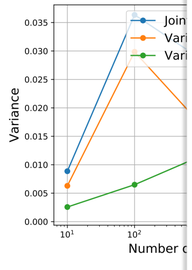


Figure 2: Trends of varian
on SVHN (right). Varian
Variance due to sampling
would suggest.

Advani & Saxe (2017)

High-d

Madhu S. Adva
Center for Brain Sc
Harvard University
Cambridge, MA 021

Andrew M. Sax
Center for Brain Sc
Harvard University
Cambridge, MA 021

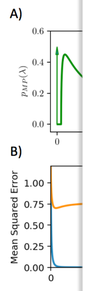


Figure 2: The March
Different ra
$\frac{P}{N}$) yield di
N, this den
lying betwe
zero. Wher
fraction of
rest are ap
order of the
origin and
are more sa
are appreci
generalizati
to be learn

Yann Le Cun et al. (1991)

VOLUME 66, NUMBER 18

Eigenvalues of

[2]*Dep*

The learning time
the eigenvalue spectr
jective function in th
tained by summing o
new techniques for
choice of centered ver

$$\rho(\lambda) = -\frac{2}{N\pi}\mathrm{Im}\frac{\delta}{\delta}$$

the Fresnel representati

$$\rho(\lambda) = -\frac{2}{N\pi}\mathrm{Im}\frac{\delta}{\delta}$$

The expression in curly

$$\{\cdots\} = \int \prod_{\gamma\beta}^{n} \frac{dq_{\gamma\beta}\, d\varphi_{\gamma\beta}}{2\pi}$$

and can be evaluated in

E through gradient descent

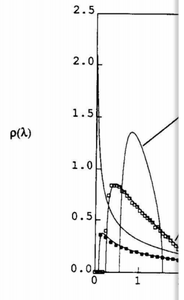


FIG. 1. Spectral density
$m=0$, $v=1$, and $\alpha=0.6$, 1.2,
grams for $\alpha=0.6$ (solid squar
averages over 100 trials with N
bility $\frac{1}{2}$ each.

2398

Krogh and Hertz (1991)

Generalization

Anders
† The Ni
§ Nordita

Received

Abstract.
with N i
$p = \alpha N$
is measur
the same
transition
(static no
teacher,
finite tim
of the no
overfitting
various va
White noi
generaliza
Generaliz
the error
a weight
$\alpha = 1$.

Figure 2. Gene
input or the out
the generalizatio
curve is for $\lambda =$
error with no n

Opper et al. (1989)

On the ability of the optimal perceptron to generalise

M Opper, W Kinzel, J Kleinz and R Nehl

Institut für Theoretische Physik, Justus-Liebig-Universität Giessen, D-6300 Giessen, Federal Republic of Germany

Abstract. A linearly separable Boolean function is derived from a set of examples by a perceptron with optimal stability. The probability to reconstruct a pattern which is not learnt is calculated analytically using the replica method.

To see this take a random input $\boldsymbol{S}$ and consider the variables

$$x = \sum_j J_j S_j \qquad \text{and} \qquad y = \sum_j B_j S_j. \tag{9}$$

For different inputs $\boldsymbol{S}$, x and y are correlated Gaussian variables with

$$\langle x\rangle = \langle y\rangle = 0 \qquad \langle x^2\rangle = J^2 \qquad \langle y^2\rangle = 1 \qquad \langle xy\rangle = R \tag{10}$$

where $\langle \ldots \rangle$ means an average over the random inputs $\boldsymbol{S}$. Hence the distribution $P(x, y)$ of x and y is given by

$$P(x,y) = \frac{1}{2\pi}\frac{1}{J\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{J^2} - 2\rho\frac{xy}{J} + y^2\right)\right]. \tag{11}$$

By definition $G(\alpha)$ is the probability that $xy > 0$, hence one has

$$G(\alpha) = 2\int_0^\infty dx \int_0^\infty dy\, P(x,y). \tag{12}$$

A straightforward calculation gives

$$G(\alpha) = 1 - \frac{1}{\pi}\cos^{-1}\rho. \tag{13}$$

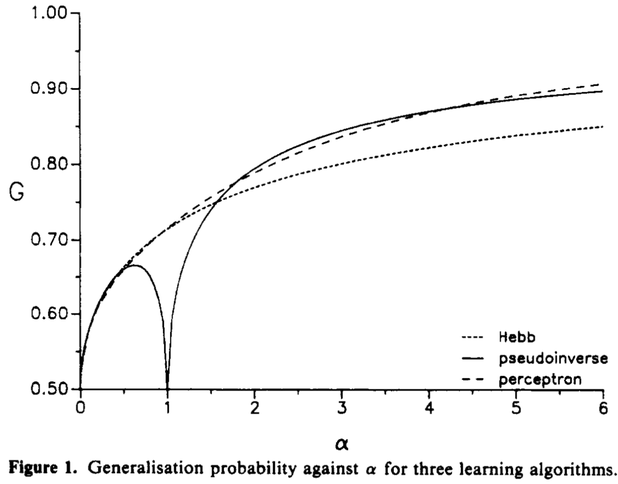


Figure 1. Generalisation probability against α for three learning algorithms.

2. The history of Double Descent

Belkin et al (Dec. 2018):

Neal et al (Oct. 2018)

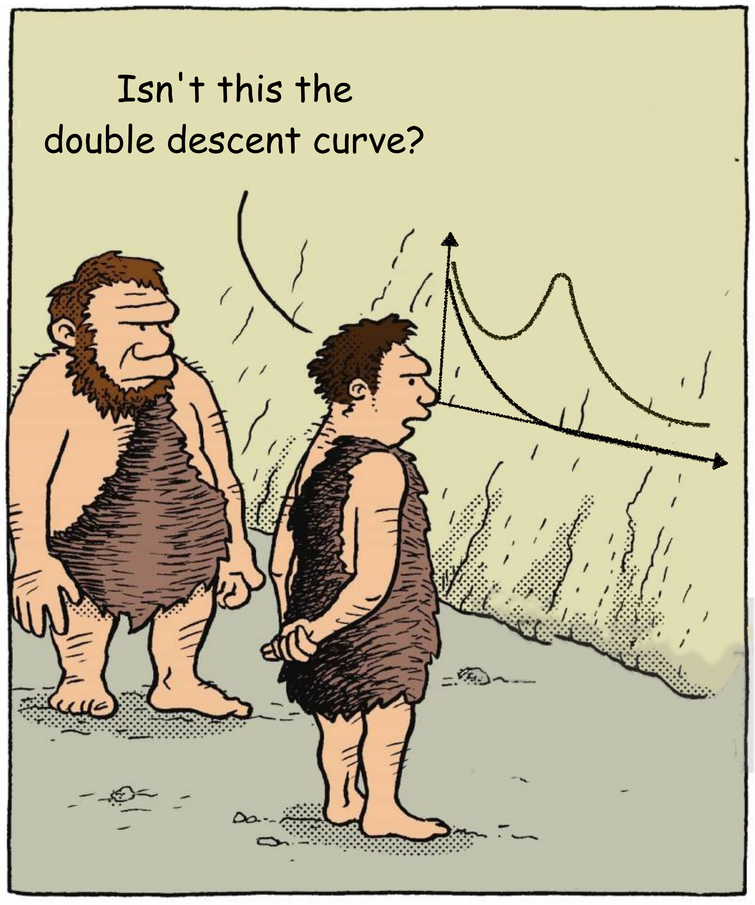
Advani & Saxe (2017)

Yann Le Cun et al. (1991)

Krogh and Hertz (1991)

Opper et al. (1989)

Homosapien (3000 BC)



3. How people justify it?

Point of view of:

Random Matrix Theory

Krogh and Hertz (1991)

Yann Le Cun et al. (1991)

Advani & Saxe (2017)

High-dimensional dynamics of generalization error in neural networks

Madhu S. Advani*
Center for Brain Science
Harvard University
Cambridge, MA 02138

Andrew M. Saxe*
Center for Brain Science
Harvard University
Cambridge, MA 02138

MADVANI@FAS.HARVARD.EDU

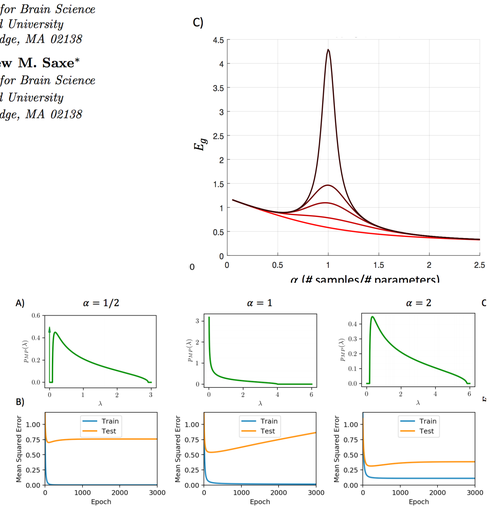


Figure 2: The Marchenko-Pasteur distribution and high-dimensional learning dynamics. A) Different ratios of number training samples (P) to network parameters (N) ($\alpha = \frac{P}{N}$) yield different eigenvalue densities in the input correlation matrix. For large N, this density is described by the MP distribution (14), which consists of a 'bulk' lying between $[\lambda_-, \lambda_+]$, and, when $\alpha < 1$, an additional delta function spike at zero. When there are fewer samples than parameters ($\alpha < 1$, left column), some fraction of eigenvalues are exactly zero (delta-function arrow at origin), and the rest are appreciably greater than zero. When the number of samples is on the order of the parameters ($\alpha = 1$, center column), the distribution diverges near the origin and there are many nonzero but arbitrarily small eigenvalues. When there are more samples than parameters ($\alpha > 1$, right column), the smallest eigenvalues are appreciably greater than zero. B) Dynamics of learning. From (13), the generalization error is harmed most by small eigenvalues; and these are the slowest to be learned. Hence for $\alpha = 1/2$ and $\alpha = 2$, the gap in the spectrum near

3. How people justify it?

Point of view of:

Random Matrix Theory

Krogh and Hertz (1991)

Yann Le Cun et al. (1991)

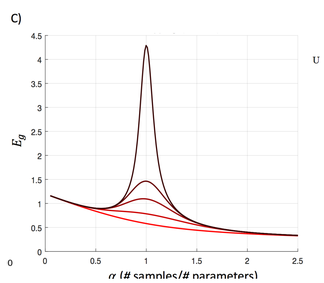
Advani & Saxe (2017)

High-dimensional dynamics of generalization error in neural networks

Madhu S. Advani*
Center for Brain Science
Harvard University
Cambridge, MA 02138

Andrew M. Saxe*
Center for Brain Science
Harvard University
Cambridge, MA 02138

MADVANI@FAS.HARVARD.EDU



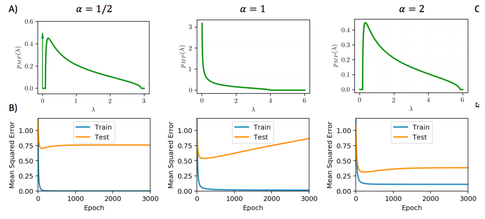


Figure 2: The Marchenko-Pasteur distribution and high-dimensional learning dynamics. A) Different ratios of number training samples (P) to network parameters (N) ($\alpha = \frac{P}{N}$) yield different eigenvalue densities in the input correlation matrix. For large N, this density is described by the MP distribution (14), which consists of a 'bulk' lying between $[\lambda_-, \lambda_+]$, and, when $\alpha < 1$, an additional delta function spike at zero. When there are fewer samples than parameters ($\alpha < 1$, left column), some fraction of eigenvalues are exactly zero (delta-function arrow at origin), and the rest are appreciably greater than zero. When the number of samples is on the order of the parameters ($\alpha = 1$, center column), the distribution diverges near the origin and there are many nonzero but arbitrarily small eigenvalues. When there are more samples than parameters ($\alpha > 1$, right column), the smallest eigenvalues are appreciably greater than zero. B) Dynamics of learning. From (13), the generalization error is harmed most by small eigenvalues; and these are the slowest to be learned. Hence for $\alpha = 1/2$ and $\alpha = 2$, the gap in the spectrum near

Bias/Variance trade-off

Belkin et al (Dec. 2018)

Neal et al (Oct. 2018)

Yang et al (Mar. 2020)

Rethinking Bias-Variance Trade-off for Generalization of Neural Networks

Zitong Yang [*1] **Yaodong Yu** [*1] **Chong You** [1] **Jacob Steinhardt** [1 2] **Yi Ma** [1]

Abstract

The classical bias-variance trade-off predicts that bias decreases and variance increases with model complexity, leading to a U-shaped risk curve. Recent work calls this into question for neural networks and other over-parameterized models, for which it is often observed that larger models generalize better. We provide a simple explanation for this by measuring the bias and variance of neural networks: while the bias is *monotonically decreasing* as in the classical theory, the variance is *unimodal* or *bell-shaped*: it increases then decreases with the width of the network. We vary the network architecture, loss function, and choice of dataset and confirm that variance unimodality occurs robustly for all models we considered. The risk curve is the sum of the bias and variance curves and displays different qualitative shapes depending on the relative scale of bias and variance, with the double descent curve observed in recent literature as a special case. We corroborate these empirical results with a theoretical analysis of two-layer linear networks with random first layer. Finally, evaluation on out-of-distribution data shows that most of the drop in accuracy comes from increased bias while variance increases by a relatively small amount. Moreover, we find that deeper models decrease bias and increase variance for both in-distribution and out-of-distribution data.

increasing as a function of the complexity of the model. The *variance* measures sensitivity to fluctuations in the training set and is often attributed to a large number of model parameters. Classical wisdom predicts that model variance increases and bias decreases *monotonically* with model complexity (Geman et al., 1992). Under this perspective, we should seek a model that has neither too little nor too much capacity and achieves the best trade-off between bias and variance.

In contrast, modern practice for neural networks repeatedly demonstrates the benefit of increasing the number of neurons (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Zhang et al., 2017), even up to the point of saturating available memory. The inconsistency between classical theory and modern practices suggests that some arguments in the classical theory can not be applied to modern neural networks.

Geman et al. (1992) first studied the bias and variance of the neural networks and give experimental evidence that the variance is indeed increasing as the width of the neural network increases. Since Geman et al. (1992), Neal et al. (2019) first experimentally measured the variance of modern neural network architectures and shown that the variance can actually be decreasing as the width increases to a highly overparameterized regime. Recently, Belkin et al. (2019a; 2018; 2019b) directly studied the risk of modern machine learning models and proposed a *double descent* risk curve, which has also been analytically characterized for certain regression models (Mei & Montanari, 2019; Hastie et al., 2019; Spigler et al., 2019; Deng et al., 2019; Advani & Saxe, 2017). However, there exists two mysteries around

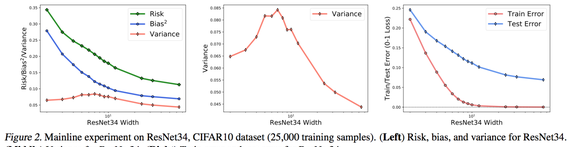
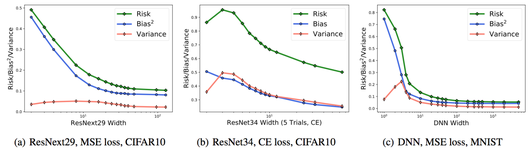


Figure 2. Mainline experiment on ResNet34, CIFAR10 dataset (25,000 training samples). (**Left**) Risk, bias, and variance for ResNet34. (**Middle**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.



(a) ResNext29, MSE loss, CIFAR10 (b) ResNet34, CE loss, CIFAR10 (c) DNN, MSE loss, MNIST

3. How people justify it?

Point of view of:

Random Matrix Theory

Krogh and Hertz (1991)

Yann Le Cun et al. (1991)

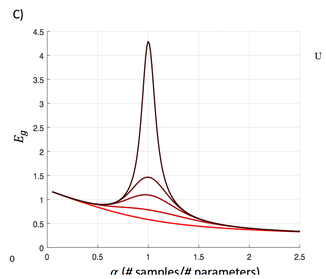
Advani & Saxe (2017)

High-dimensional dynamics of generalization error in neural networks

Madhu S. Advani*
Center for Brain Science
Harvard University
Cambridge, MA 02138

Andrew M. Saxe*
Center for Brain Science
Harvard University
Cambridge, MA 02138

MADVANI@FAS.HARVARD.EDU



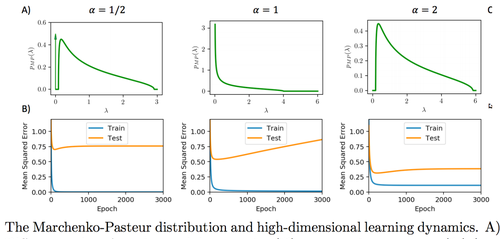


Figure 2: The Marchenko-Pasteur distribution and high-dimensional learning dynamics. A) Different ratios of number training samples (P) to network parameters (N) ($\alpha = \frac{P}{N}$) yield different eigenvalue densities in the input correlation matrix. For large N, this density is described by the MP distribution (14), which consists of a 'bulk' lying between $[\lambda_-, \lambda_+]$, and, when $\alpha < 1$, an additional delta function spike at zero. When there are fewer samples than parameters ($\alpha < 1$, left column), some fraction of eigenvalues are exactly zero (delta-function arrow at origin), and the rest are appreciably greater than zero. When the number of samples is on the order of the parameters ($\alpha = 1$, center column), the distribution diverges near the origin and there are many nonzero but arbitrarily small eigenvalues. When there are more samples than parameters ($\alpha > 1$, right column), the smallest eigenvalues are appreciably greater than zero. B) Dynamics of learning. From (13), the generalization error is harmed most by small eigenvalues; and these are the slowest to be learned. Hence for $\alpha = 1/2$ and $\alpha = 2$, the gap in the spectrum near

Bias/Variance trade-off

Belkin et al (Dec. 2018)

Neal et al (Oct. 2018)

Yang et al (Mar. 2020)

Rethinking Bias-Variance Trade-off for Generalization of Neural Networks

Zitong Yang [*1] **Yaodong Yu** [*1] **Chong You** [1] **Jacob Steinhardt** [1 2] **Yi Ma** [1]

Abstract

The classical bias-variance trade-off predicts that bias decreases and variance increases with model complexity, leading to a U-shaped risk curve. Recent work calls this into question for neural networks and other over-parameterized models, for which it is often observed that larger models generalize better. We provide a simple explanation for this by measuring the bias and variance of neural networks: while the bias is *monotonically decreasing* as in the classical theory, the variance is *unimodal* or *bell-shaped*: it increases then decreases with the width of the network. We vary the network architecture, loss function, and choice of dataset and confirm that variance unimodality occurs robustly for all models we considered. The risk curve is the sum of the bias and variance curves and displays different qualitative shapes depending on the relative scale of bias and variance, with the double descent curve observed in recent literature as a special case. We corroborate these empirical results with a theoretical analysis of two-layer linear networks with random first layer. Finally, evaluation on out-of-distribution data shows that most of the drop in accuracy comes from increased bias while variance increases by a relatively small amount. Moreover, we find that deeper models decrease bias and increase variance for both in-distribution and out-of-distribution data.

increasing as a function of the complexity of the model. The *variance* measures sensitivity to fluctuations in the training set and is often attributed to a large number of model parameters. Classical wisdom predicts that model variance increases and bias decreases *monotonically* with model complexity (Geman et al., 1992). Under this perspective, we should seek a model that has neither too little nor too much capacity and achieves the best trade-off between bias and variance.

In contrast, modern practice for neural networks repeatedly demonstrates the benefit of increasing the number of neurons (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Zhang et al., 2017), even up to the point of saturating available memory. The inconsistency between classical theory and modern practices suggests that some arguments in the classical theory can not be applied to modern neural networks.

Geman et al. (1992) first studied the bias and variance of the neural networks and give experimental evidence that the variance is indeed increasing as the width of the neural network increases. Since Geman et al. (1992), Neal et al. (2019) first experimentally measured the variance of modern neural network architectures and shown that the variance can actually be decreasing as the width increases to a highly overparameterized regime. Recently, Belkin et al. (2019a; 2018; 2019b) directly studied the risk of modern machine learning models and proposed a *double descent* risk curve, which has also been analytically characterized for certain regression models (Mei & Montanari, 2019; Hastie et al., 2019; Spigler et al., 2019; Deng et al., 2019; Advani & Saxe, 2017). However, there exists two mysteries around

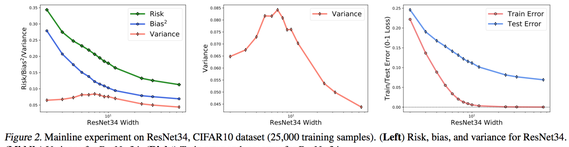
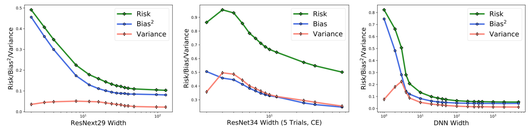


Figure 2. Mainline experiment on ResNet34, CIFAR10 dataset (25,000 training samples). (**Left**) Risk, bias, and variance for ResNet34. (**Middle**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.



(a) ResNext29, MSE loss, CIFAR10 (b) ResNet34, CE loss, CIFAR10 (c) DNN, MSE loss, MNIST

Statistical Mechanics

Engel and Van den Broeck (2005)

Opper et al. (1989)

Gerace et al. (Feb. 2020)

Generalisation error in learning with random features and the hidden manifold model

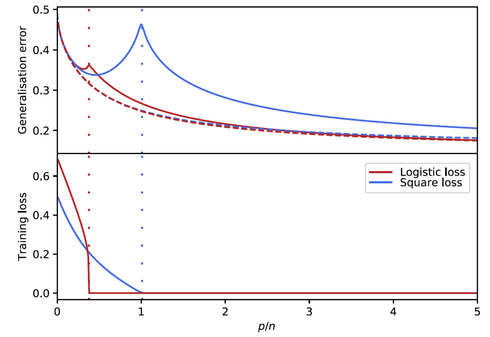
Federica Gerace[†], Bruno Loureiro[†],

Florent Krzakala[⋆], Marc Mézard [⋆], Lenka Zdeborová[†]

† *Institut de Physique Théorique*
CNRS & CEA & Université Paris-Saclay, Saclay, France
⋆ *LPENS, CNRS & Sorbonnes Universites,*
Ecole Normale Superieure, PSL University, Paris, France

Abstract

We study generalised linear regression and classification for a synthetically generated dataset encompassing different problems of interest, such as learning with random features, neural networks in the lazy training regime, and the hidden manifold model. We consider the high-dimensional regime and using the replica method from statistical physics, we provide a closed-form expression for the asymptotic generalisation performance in these problems, valid in both the under- and over-parametrised regimes and for a broad choice of generalised linear model loss functions. In particular, we show how to obtain analytically the so-called double descent behaviour for logistic regression with a peak at the interpolation threshold, we illustrate the superiority of orthogonal against random Gaussian projections in learning with random features, and discuss the role played by correlations in the data generated by the hidden manifold model. Beyond the interest in these particular problems, the theoretical formalism introduced in this manuscript provides a path to further extensions to more complex tasks.



3. How people justify it?

Random Matrix Theory:

Consider a simple Linear Regression task:

Dataset

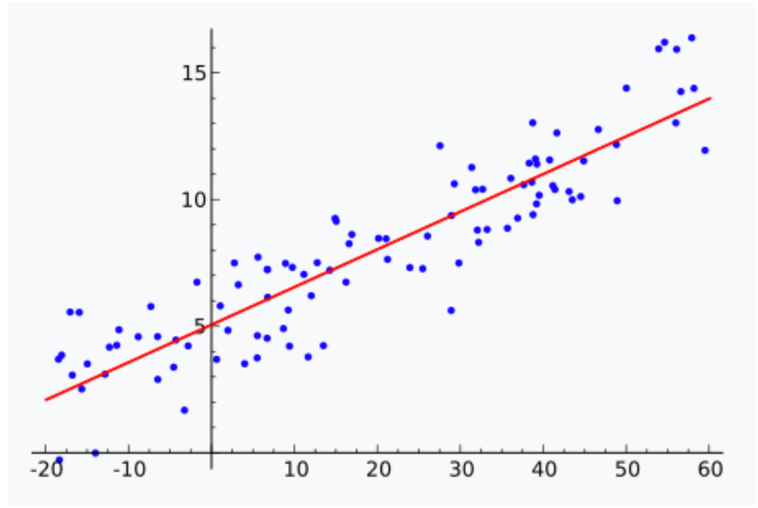
$$\mathcal{D} = \{X, y\}$$

$$\hat{y} = XW$$

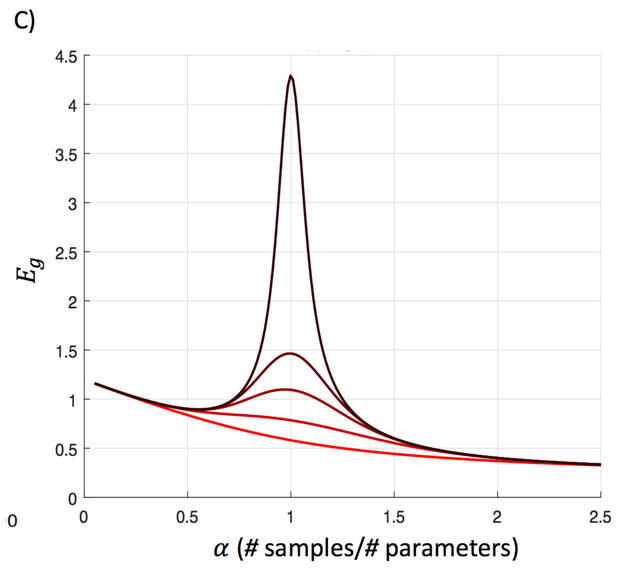
N x dim

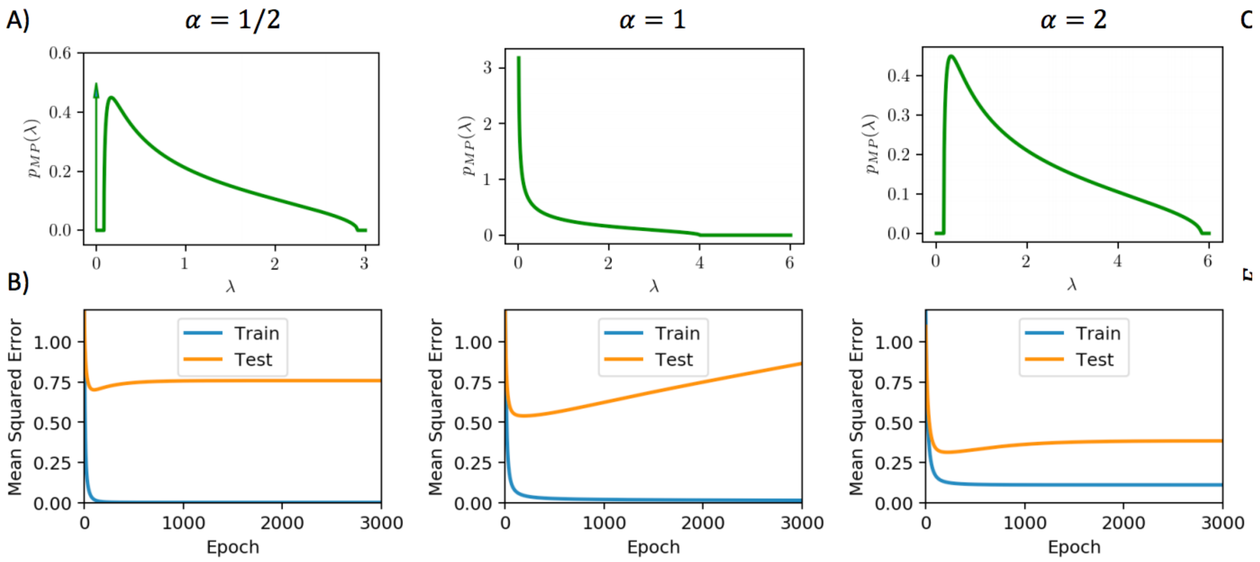
alpha = dim / N

A measure of over-parameterization



$$\mathcal{L} = ||y - \hat{y}||_2^2 \longrightarrow \nabla_W \mathcal{L} = 0 \longrightarrow \hat{\boldsymbol{\theta}}_P := \begin{cases} (\boldsymbol{X}_P^\top \boldsymbol{X}_P)^{-1} \boldsymbol{X}_P^\top \boldsymbol{y} & \text{if dim < N} \\ \boldsymbol{X}_P^\top (\boldsymbol{X}_P \boldsymbol{X}_P^\top)^{-1} \boldsymbol{y} & \text{if dim > N} \end{cases}$$





3. How people justify it?

Bias/Variance trade-off

$$\mathbb{E}_{\epsilon_i}\left[\|(\boldsymbol{y} - f(\boldsymbol{x}, \mathcal{T}_{\epsilon_i})\|_2^2\right] =$$
$$\underbrace{\left[\|(\boldsymbol{y} - \bar{f}(\boldsymbol{x})\|_2^2\right]}_{\textbf{Bias}^2} + \underbrace{\mathbb{E}_{\epsilon_i}\left[\|(f(\boldsymbol{x}, \mathcal{T}_{\epsilon_i}) - \bar{f}(\boldsymbol{x})\|_2^2\right]}_{\textbf{Variance}},$$

3. How people justify it?

Statistical Mechanics

In large random microscopic heterogeneity, striking levels of almost deterministic macroscopic order can arise in ways that do not depend on the details of the heterogeneity.

The idea here is to rewrite the high-dimensional learning dynamics, in terms of low-dimensional but high level macroscopic variables.

$$y = X\bar{W}$$

$$\hat{y} = XW$$

$$P(x, y) = \frac{1}{2\pi} \frac{1}{J\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{J^2} - 2\rho \frac{xy}{J} + y^2 \right) \right]. \tag{11}$$

By definition $G(\alpha)$ is the probability that $xy > 0$, hence one has

$$G(\alpha) = 2 \int_0^\infty dx \int_0^\infty dy\, P(x, y). \tag{12}$$

A straightforward calculation gives

$$G(\alpha) = 1 - \frac{1}{\pi} \cos^{-1} \rho. \tag{13}$$

$$R = \sqrt{\frac{2}{\pi}}\, \alpha \qquad J^2 = \frac{\alpha - R^2}{1-\alpha} \qquad (\alpha < 1)$$

$$R = \sqrt{\frac{2}{\pi}} \qquad J^2 = \frac{1 + (2/\pi)(\alpha - 2)}{\alpha - 1} \qquad (\alpha > 1)$$

which using (13) yields

$$G(\alpha) = 1 - \frac{1}{\pi} \cos^{-1} \left(\frac{2\alpha(1-\alpha)}{\pi - 2\alpha} \right)^{1/2} \qquad (\alpha < 1)$$

$$G(\alpha) = 1 - \frac{1}{\pi} \cos^{-1} \left(\frac{2(\alpha - 1)}{\pi + 2\alpha - 4} \right)^{1/2} \qquad (\alpha > 1).$$

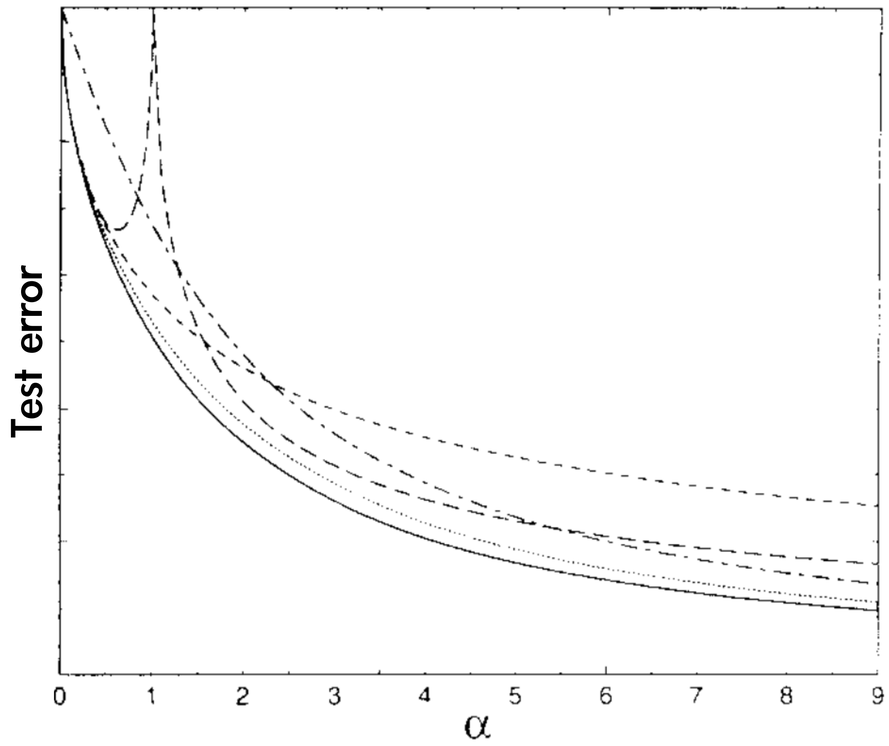
What is missing?

A model of modern model-wise Double Descent and for classification setup

A theory on epoch-wise Double Descent

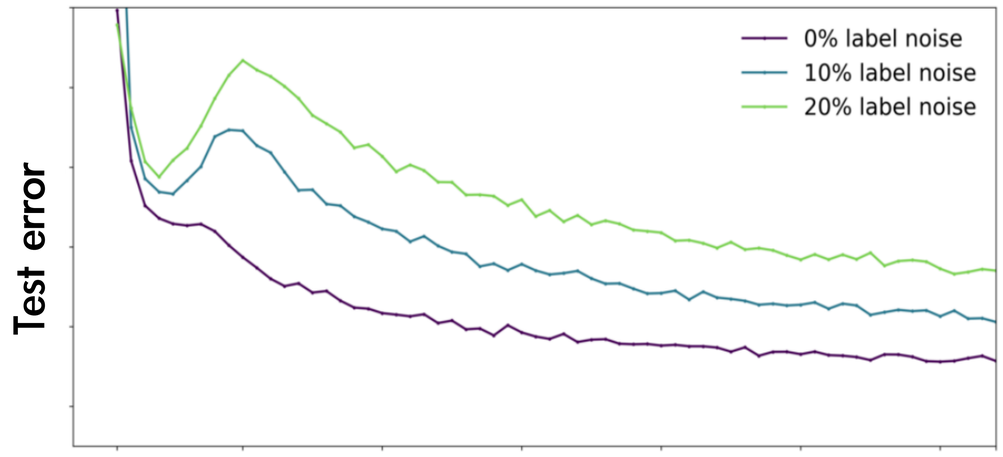
An approach to mitigate Double Descent

Classical Double Descent



(#samples / #parameters)

Modern Double Descent



(#parameters / #samples)

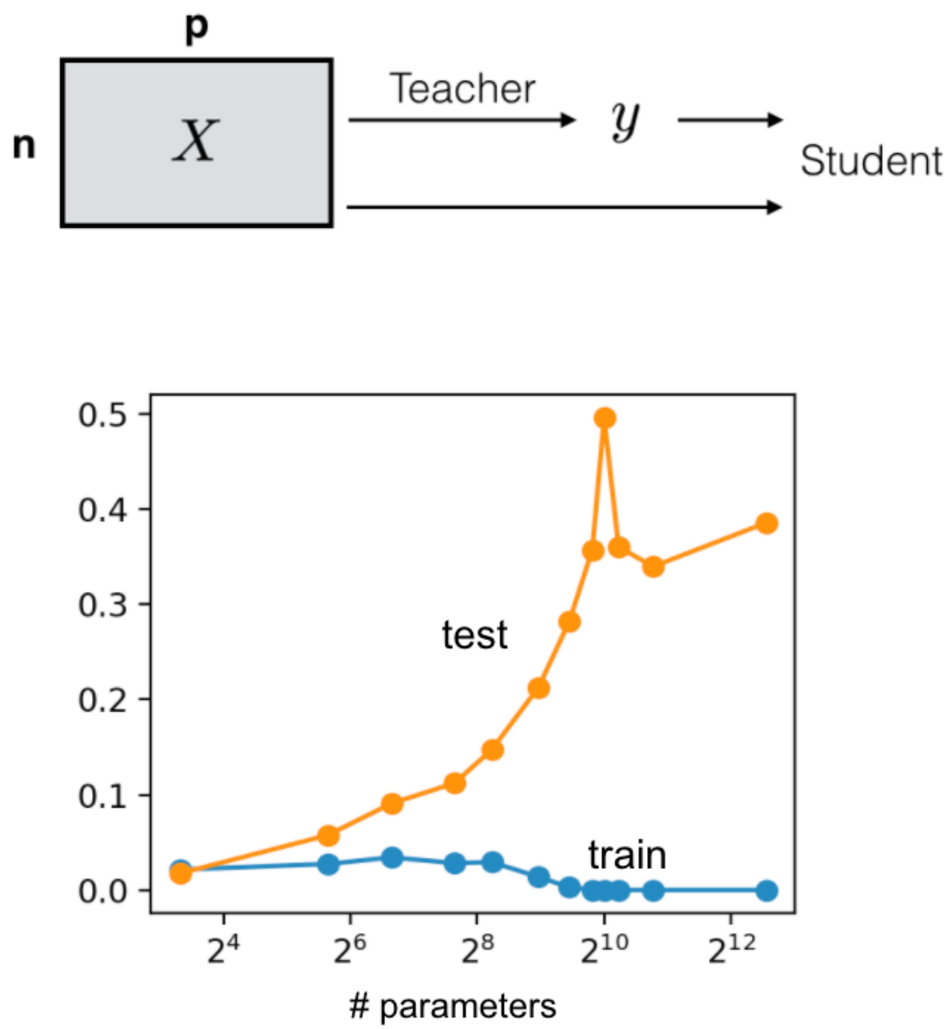
What is missing?

A model of modern model-wise Double Descent and for classification setup

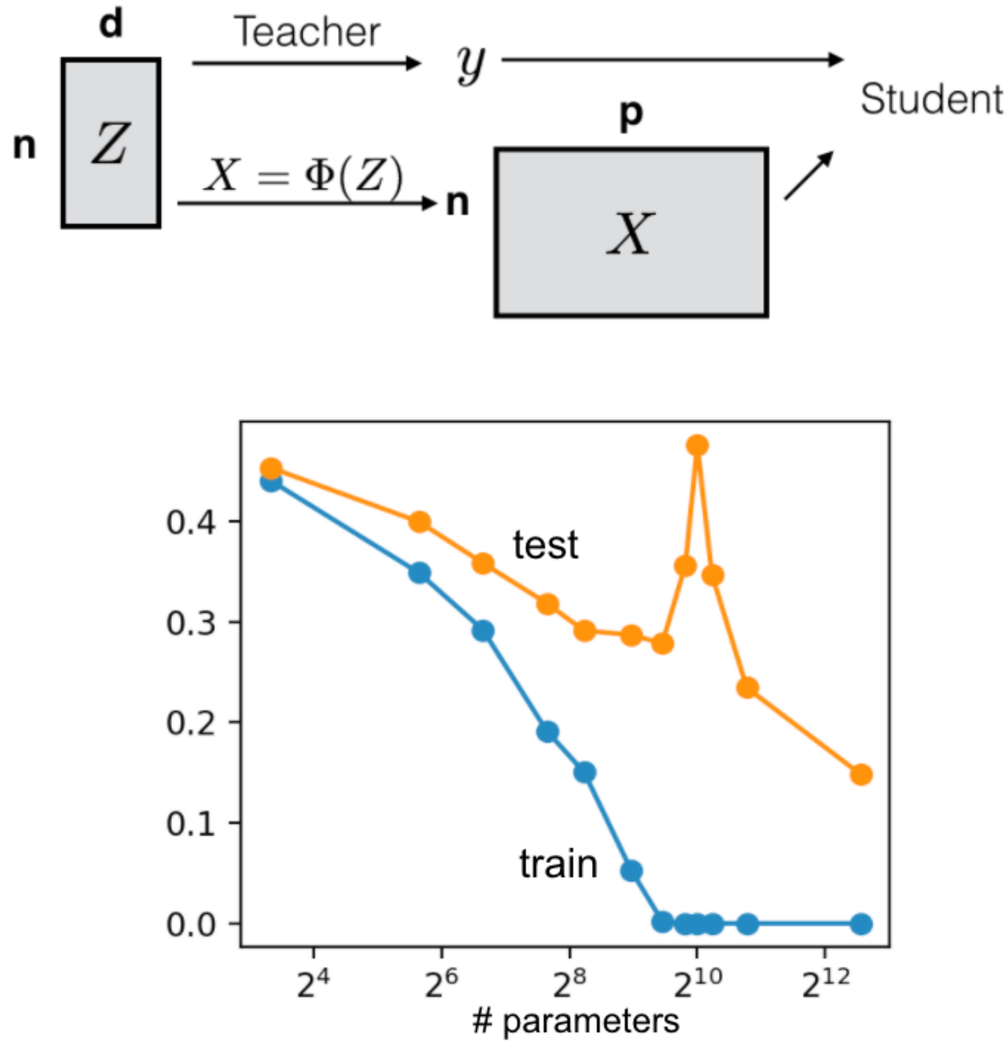
A theory on epoch-wise Double Descent

An approach to mitigate Double Descent

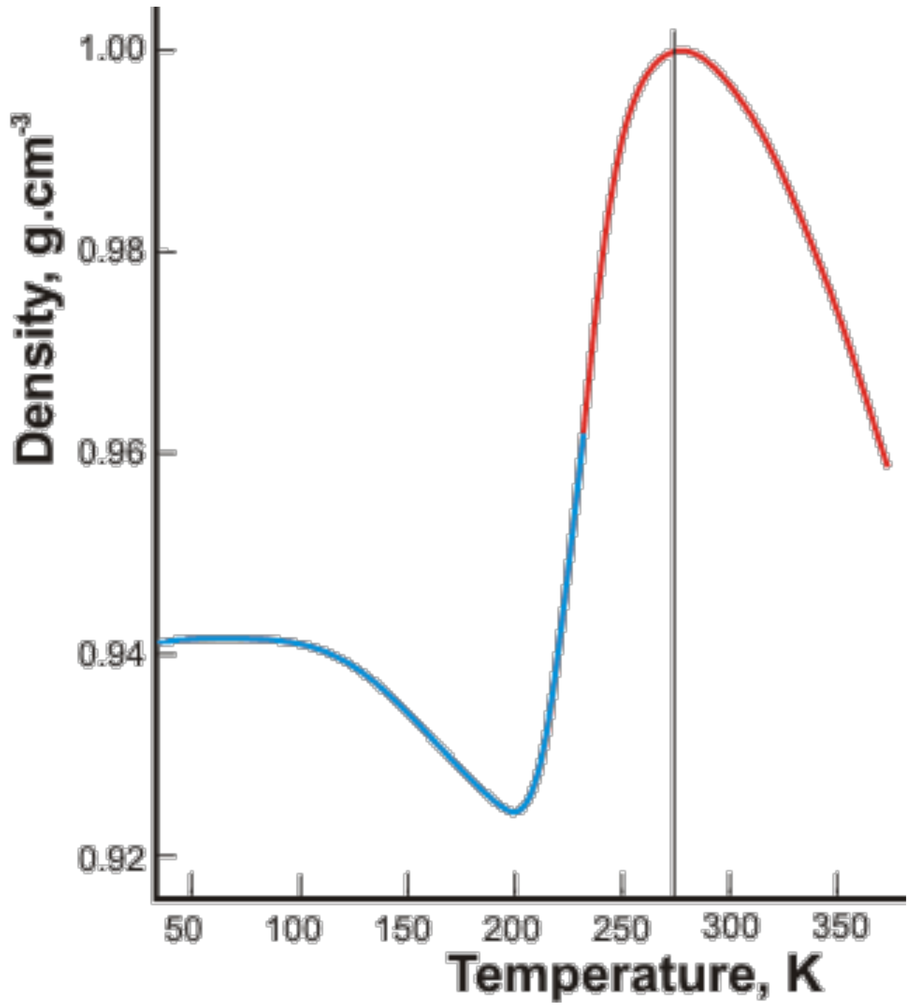
Classical Double Descent



Modern Double Descent

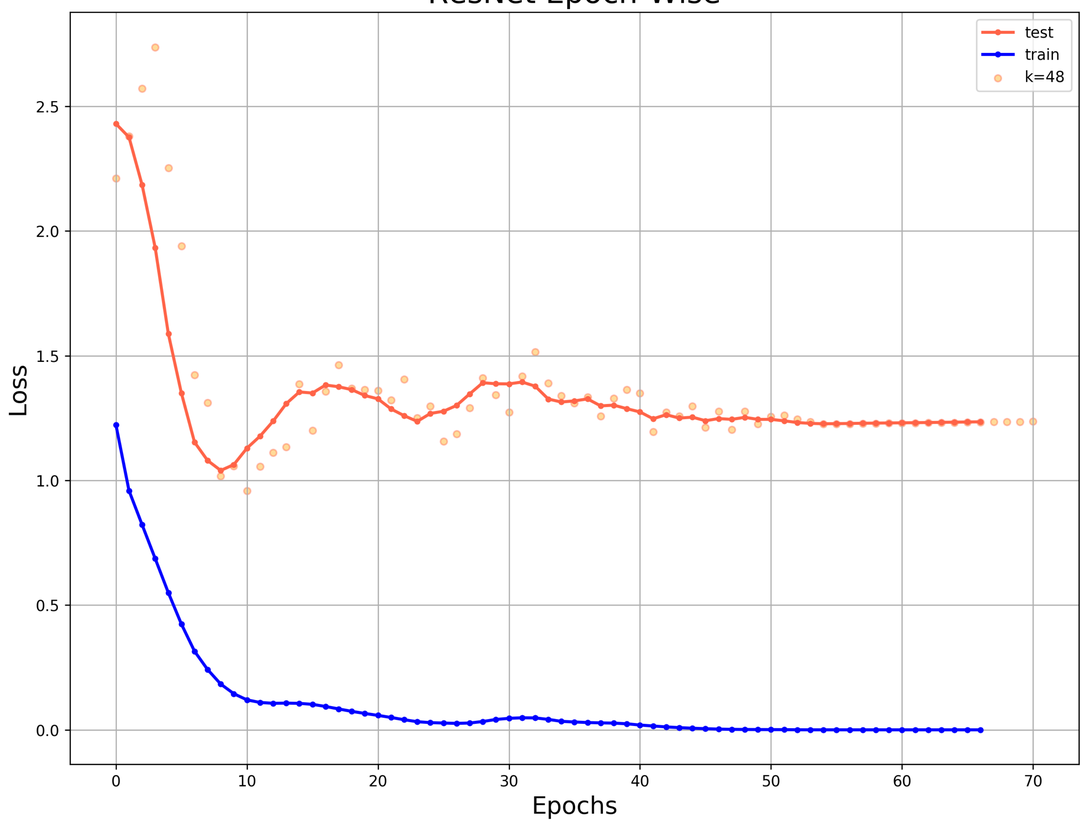


What is missing?



What is missing?

Without batch-norm



With batch-norm