Review : Generalization of Two-Layer Neural Networks : An Asymptotic Viewpoint(ICLR, 2020)

DongYoon Yang

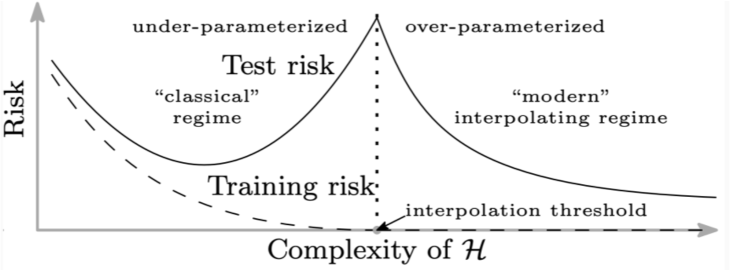
Seoul National University

January 17, 2022

Table of Contents

Table of Contents

Introduction : the Double Descent Phenomenon



Double descent phenomenon : Second decrease in population risk beyond the interpolation threshold.

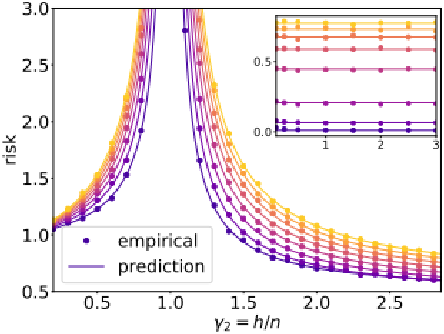
Problem Setup and Assumptions

- $y_i = F(x_i) + \varepsilon_i$, $\mathbb{E}(x_i) = 0$, $Cov(x_i) = \Sigma$, $var(\varepsilon_i) = \sigma^2$
- $R(f) = \mathbb{E}_{\boldsymbol{x}}(f(\boldsymbol{x}) - F(\boldsymbol{x}))^2$: population risk
- Trainable Network $f(\boldsymbol{x}) = \sum_{i=1}^{h} a_i \phi(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle)$
- (A1) Data : $x_i \sim N(0, I_d)$
- (A2) True model : $F(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$, $\|\boldsymbol{\beta}\| = r$
- Objective : $\ell(f) = \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2$
- $\frac{d}{n} \to \gamma_1 \in (0, \infty)$, $\frac{h}{n} \to \gamma_2 \in (0, \infty)$ as $n, d, h \to \infty$
- Optimization : Gradient flow on either the fist or second layer.
- Overparametrization corresponds to increasing $\gamma_2 = h/n$

Table of Contents

Linear Network

- $\phi(x) = x$



- Brighter color indicate larger γ_1.
- Main figure : second layer coefficients are optimized.
- Subfigure : first layer coefficients are optimized.

Table of Contents

Optimizing The Second Layer: Nonlinear

- $w_i \overset{i.i.d.}{\sim} N(0, d^{-1} I_d)$
- $\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{h} \hat{a}_i \phi(\langle \boldsymbol{x}, w_i \rangle)$

$$
\begin{aligned}
R(\hat{f}) &= \mathbb{E}_{\boldsymbol{x} \sim P_x} \left[\left\| \phi\left(\boldsymbol{x}^\top W\right) \hat{\boldsymbol{a}} - F(\boldsymbol{x}) \right\|_2^2 \mid \boldsymbol{X}, W \right] \\
&= \underbrace{\mathbb{E}_{\boldsymbol{x}} \left[\left\| \mathbb{E}\left[\phi\left(\boldsymbol{x}^\top W\right) \hat{\boldsymbol{a}} \mid \boldsymbol{X}, W \right] - F(\boldsymbol{x}) \right\|_2^2 \right]}_{B = \text{ bias}} \\
&\quad + \underbrace{\mathbb{E}_{\boldsymbol{x}} \left[\left\| \phi\left(\boldsymbol{x}^\top W\right) \hat{\boldsymbol{a}} - \mathbb{E}\left[\phi\left(\boldsymbol{x}^\top W\right) \hat{\boldsymbol{a}} \right] \right\|_2^2 \mid \boldsymbol{X}, W \right]}_{V = \text{ variance}}
\end{aligned}
$$

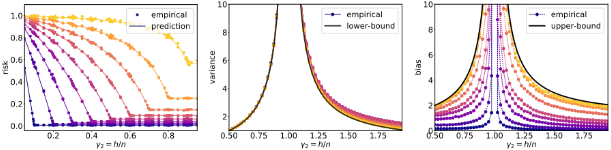
Optimizing The Second Layer: Nonlinear

Theorem

If we let $\gamma_1 \to \infty$, the variance is equal to
$V_{(\gamma_1 \to \infty)} = \sigma^2 \min\{\gamma_2, 1\} / |1 - \gamma_2|$

Theorem

Given(A1, A2) and $w_i \sim N(0, I_d)$, then $B \to \infty$ as $\gamma_2 \to 1$. And, B is finite when $\gamma_2 > 1$.



(a) risk ($\gamma_1 < 1$). (b) variance (ReLU). (c) bias (ReLU).

Figure 2: Population risk of two-layer neural networks with optimized second layer under (A1)(A2). Brighter color indicates larger γ_1. (a) risk of linear network with $r^2/\sigma^2 = 16$ and $\gamma_1 < 1$. ($\gamma_1 > 1$ is shown in Figure 1) (b) variance of network with ReLU activation. Black line corresponds to $\gamma_1 \to \infty$ predicted by Corollary 5. (c) bias of network with ReLU activation. Black line corresponds to $\gamma_1 \to \infty$ for linear network, which is empirically observed as an upper-bound. Note that as $\gamma_2 \to 1$ both bias and variance becomes unbounded.

Table of Contents

- First layer parameters $a_i \overset{i.i.d.}{\sim} \text{Unif}\left\{-1/\sqrt{h}, 1/\sqrt{h}\right\}$
- (A3) : ϕ is smooth, Lipchitz and monotone with $\phi'(0) = 0, |\phi'(\pm x) - \phi'(\infty)| = O(\exp^{-x})$
- (A4) : $\forall i \in [1, h], \exists! j \in [1, h]$ s.t. $a_i w_i^{init} = -a_j w_j^{\text{init}}$
- Vanishing initialization : $w_i(0) \sim N(0, I_d/dh^{1+\varepsilon})$

$$\|W(t) - W(0)\|_F \gg \|W(0)\|_F$$

- Non-vanishing initialization : $w_i(0) \sim N(0, I_d/d^{-\varepsilon})$

$$\|W(t) - W(0)\|_F \ll \|W(0)\|_F$$

Vanishing Initialization

Theorem

Given $(A1-3)$. *Let* $T = O(\log\log h)$ *and* $\hat{f}(\cdot) = \hat{f}^{van}(\cdot, W(T))$, *then as* $n, d, h \to \infty$, *the gradient flow reaches a* $o(1)$ *first-order stationary point at time* T, *at which point the population risk is given as*

$$R(\hat{f}) \to \max\left\{0, \frac{\gamma_1 - 1}{\gamma_1}\right\} r^2 + \frac{\min\{\gamma_1, 1\}}{|1-\gamma_1|}\sigma^2 \tag{1}$$

- Independent to γ_2. Double descent does not appear about overparametrization.
- Double descent appears about γ_1.
- It is same as the risk of linear neural network.

Non-Vanishing Initialization

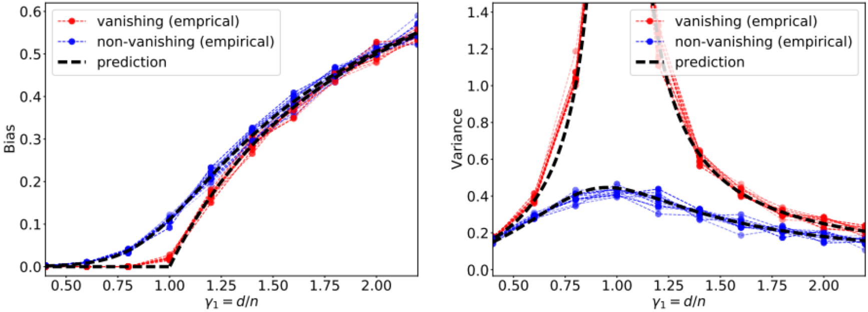
Theorem

Given $(A1-4)$ *and let* $n, d, h \to \infty$ *the stationary solution* $\hat{f}(\cdot) = \hat{f}^{\text{non-van}}(\cdot, W(T))$ *has the following risk*

$$R(\hat{f}) \to \left(\frac{\gamma_1 - 1}{2\gamma_1} + \frac{\gamma_1(\gamma_1 + \gamma_1 m + m - 2) + 1}{2\gamma_1\sqrt{\gamma_1(\gamma_1 + m(\gamma_1(m+2)+2) - 2) + 1}} \right) r^2 + \left(\frac{\gamma_1 + \gamma_1 m + 1}{4\sqrt{\gamma_1(\gamma_1 + m(\gamma_1(m+2)+2) - 2) + 1}} - \frac{1}{4} \right) \sigma^2, \tag{2}$$

where $m = b_1^2/b_0^2$, $b_0^2 = \mathbb{E}\left[\phi'(G)\right]^2$, *and* $b_1^2 = \mathbb{E}\left[\phi'(G)^2\right] - b_0^2$, $G \sim \mathcal{N}(0,1)$.

- Independent to γ_2
- Double descent does not appear.



(a) bias. (b) variance.

Figure 3: Bias and variance of two-layer sigmoid network with optimized first layer under (A1)(A2). Individual dotted lines correspond to different γ_2 (from 0.2 to 2) which is independent to the risk. The bias and variance for both initializations is well-aligned with Theorem 7 and Theorem 8.

Table of Contents

Summary

- The double descent phenomenon of the bias-variance decomposition may be observed when the second layer weights are optimized assuming that the first layer weights are constant.
- When the first layer is optimized, the population risk is independent to overparametrization.
- But, two different initialization show different patterns.
  1. For vanishing initialization : the gradient flow solution is asymptotically close to a two layered linear network.
  2. For non-vanishing initialization : the gradient flow solution is well approximated by a kernel model.

Summary

	Opt 2nd layer	Opt 1st layer Vanishing	Opt 1st layer Non-Vanishing
Bias	γ_1 : No, γ_2 : Yes	γ_1 : No, γ_2 : No	γ_1 : No, γ_2 : No
Variance	γ_1 : No, γ_2 : Yes	γ_1 : Yes, γ_2 : No	γ_1 : No, γ_2 : No

Table: The presence/ absence of double descent