Double Descent In Machine Learning

数学院 周括

2020/ 04 /21

1统计学习基础定理：泛化误差界：

1）损失函数：

$$L(Y, f(X))$$

2）期望风险：

$$R(f) = E_p[L(Y, f(X))]$$

3）经验风险：

$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))]$$

泛化误差的上界由以下概率不等式来表示：

$$\varepsilon(d,N,\delta)=\sqrt{\frac{1}{2N}\left(\log d+\log\frac{1}{\delta}\right)}$$

$$P(R(f)<\hat{R}(f)+\varepsilon)\geq 1-\delta$$

由上述不等式可得到的机器学习传统观念：

（1）要尽量降低训练误差

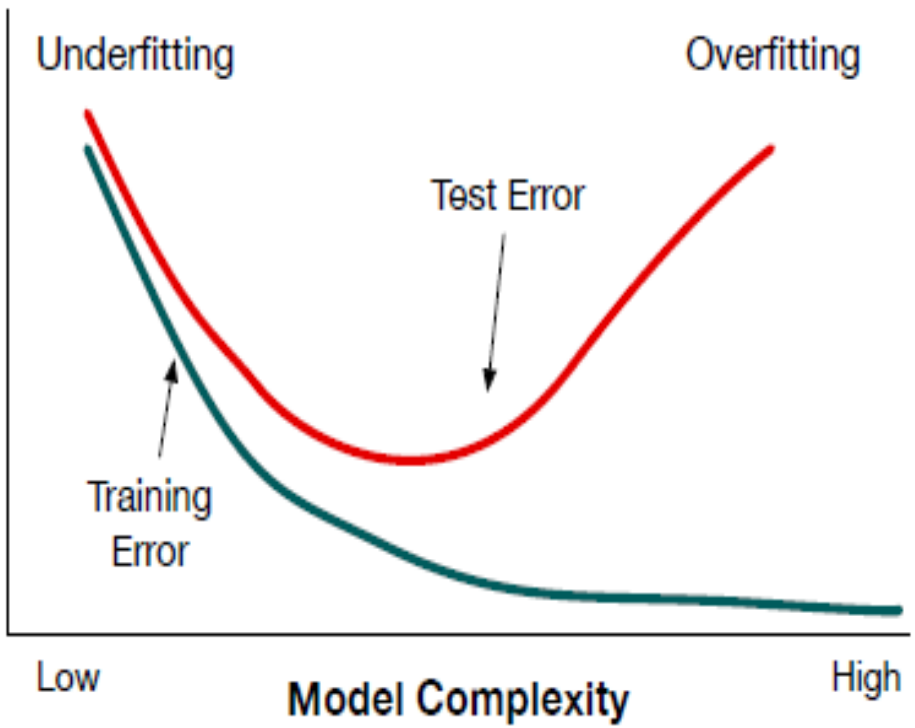
（2）不能让模型太复杂，导致其容量过大

（3）利用更多的数据可以学习到更好的分类器

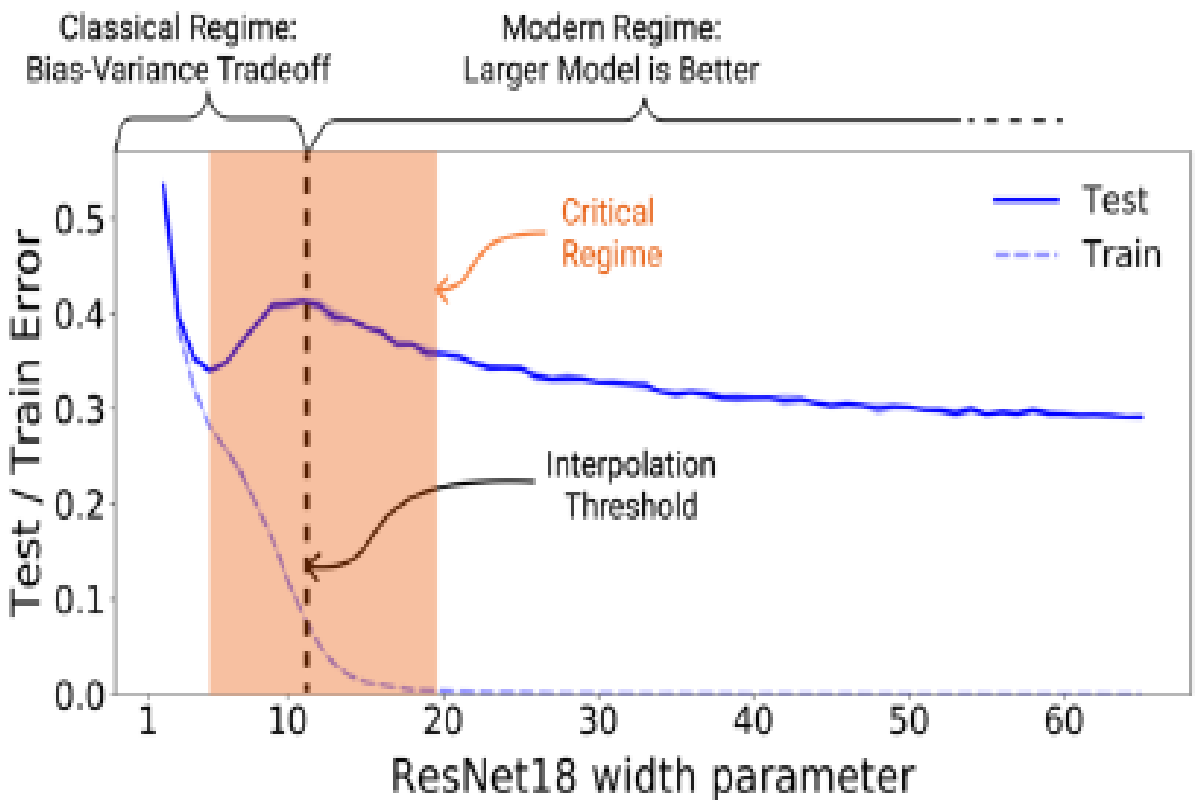
$$\varepsilon(d,N,\delta)=\sqrt{\frac{1}{2N}\left(\log d+\log\frac{1}{\delta}\right)}$$

$$P(R(f)<\hat{R}(f)+\varepsilon)\geq 1-\delta$$

模型复杂度与经验风险和泛化误差的关系



2 Double descent 现象



DEEP DOUBLE DESCENT: WHERE BIGGER MODELS AND MORE DATA HURT

- 这篇文章介绍了在深度学习中出现的双下降现象，作者指出：
- （1）传统的基于模型容量的理论分析是不够的，不但要考虑模型，还要考虑训练的步数，人为引入的噪声，训练方法，数据分布等。
- （2）用更多的数据去训练以降低测试集上的误差，有时可能会适得其反。

EMC（Effective Model Complexity）

- The Effective Model Complexity (EMC) of a training procedure T , with respect to distribution D and parameter > 0, is defined as:

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

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

对EMC的理解：

- 这是衡量所用的学习程序（即一个输入数据集，输出一个条件分布的程序）

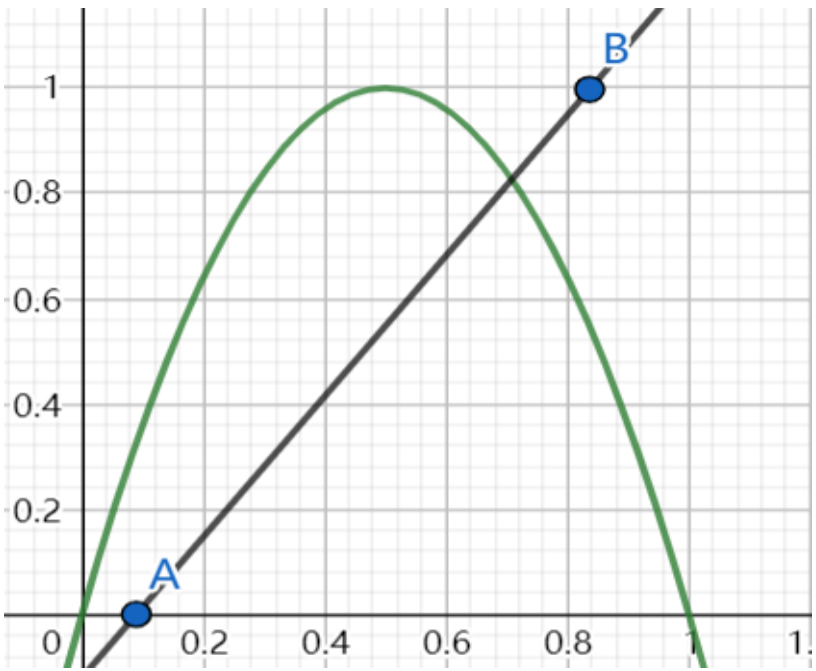
对于某个分布的拟合能力的度量。

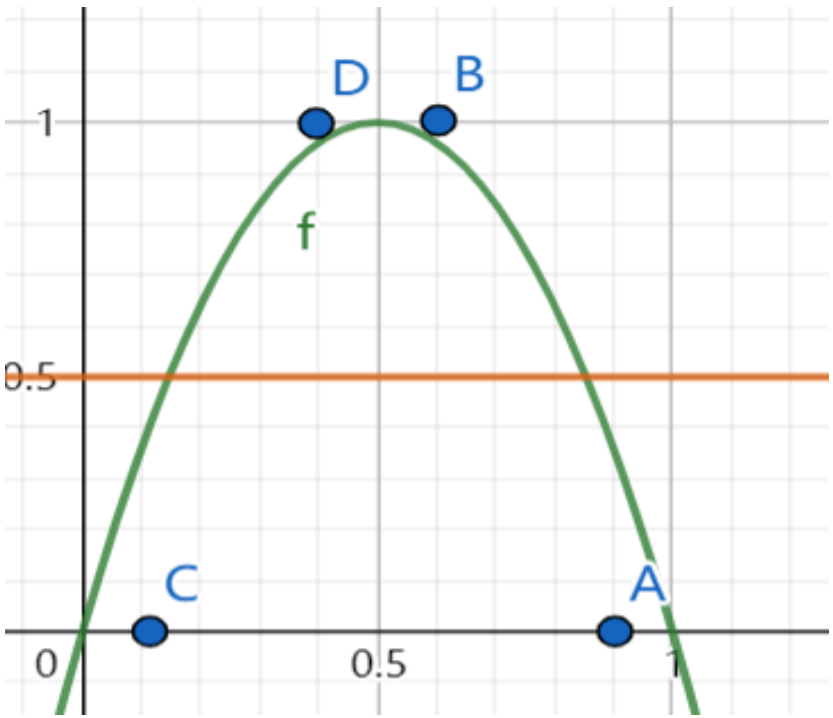
例：设随机变量X取值于[0,1]的均匀分布，随机变量Y取值于{0，1}，并且满足

$E[Y=1 \mid X=t] = 4* t *(1- t)$

假设我们采用线性模型来拟合条件分布：则可以看出其EMC应该为2

当训练集只有两个点时，无论如何都可以满足，但若训练集有3，4个点时，就不太可能在训练集上表现好了。





- 我们可以看出：对于某个分布D，某个学习程序T，其EMC值是比较大是T程序能生成较好的判别器的： **必要但未必充分条件**
- 但EMC是定义式，而非决定式，EMC的真正影响因素有
  （1）模型
- （2）训练步数
- （3）训练算法
- （4）人工噪声 等等因素

通过一系列实验得到的三个基本规律：

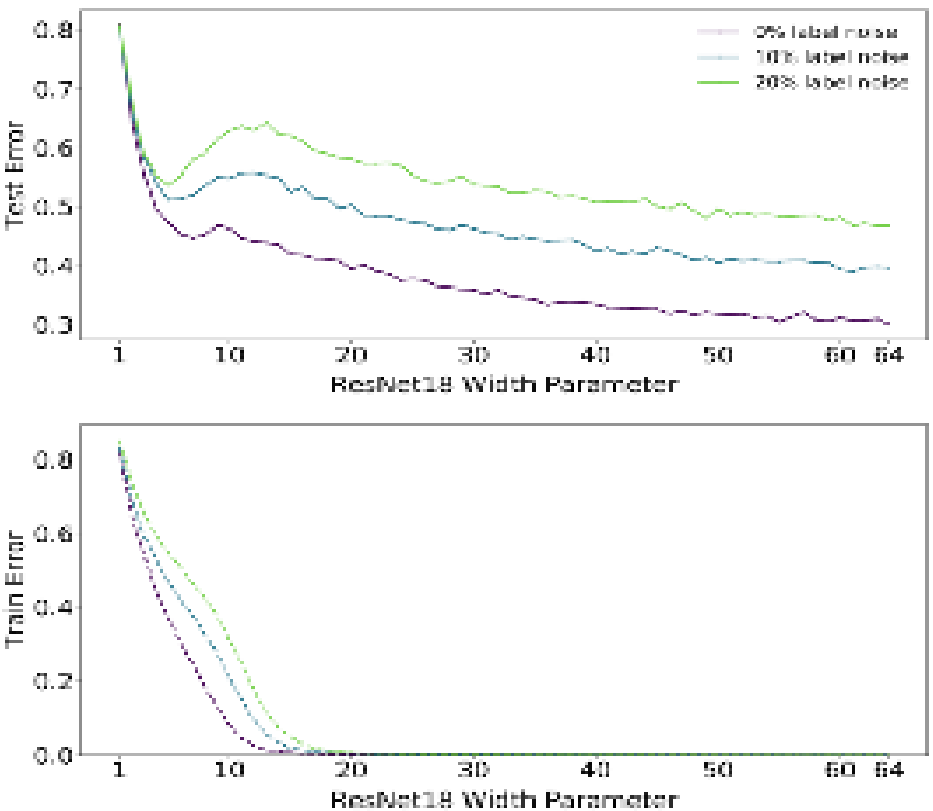
Hypothesis 1 (Generalized Double Descent hypothesis, informal) *For any natural data distribution $\mathcal{D}$, neural-network-based training procedure $\mathcal{T}$, and small $\epsilon > 0$, if we consider the task of predicting labels based on n samples from $\mathcal{D}$ then:*

Under-paremeterized 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.*

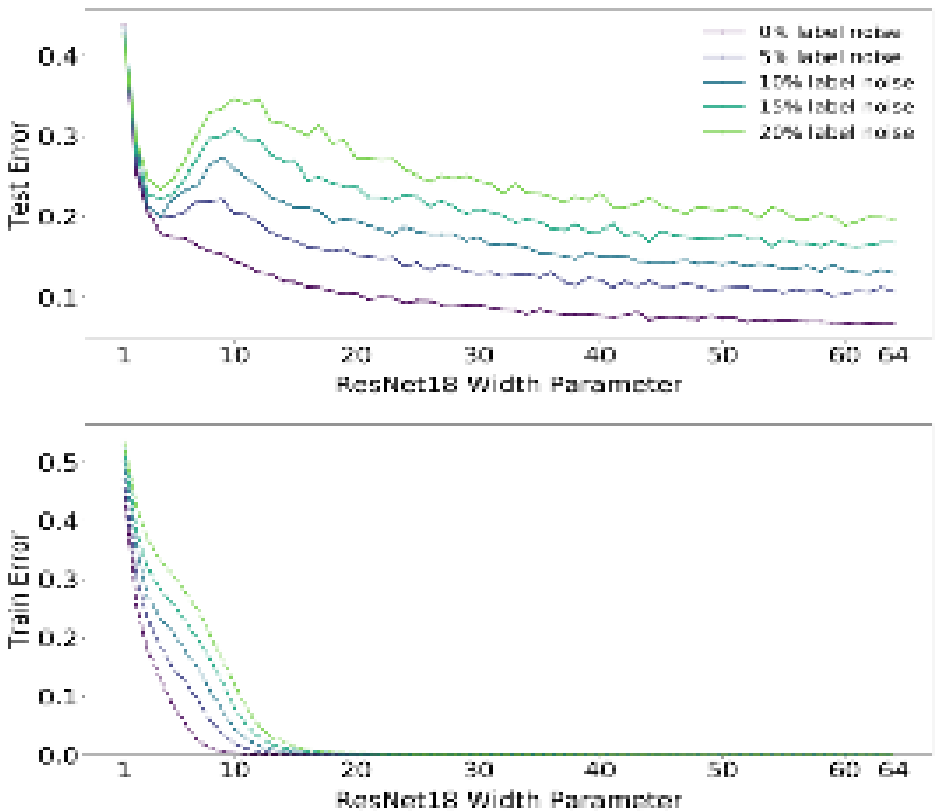
Over-parameterized regime. *If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ is sufficiently larger than n, any perturbation of $\mathcal{T}$ that increases its effective complexity will decrease the test error.*

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.*

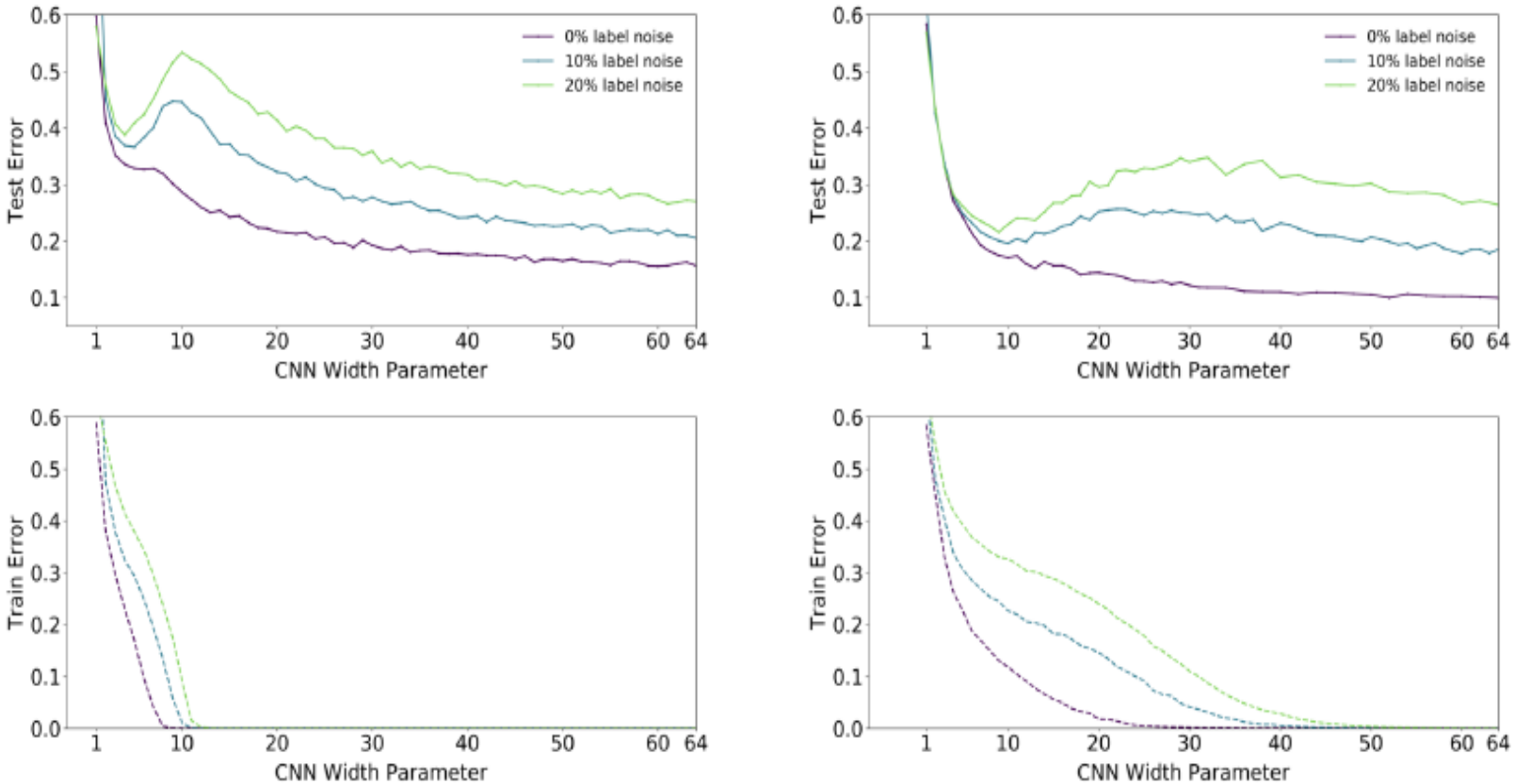
实验一 Model wise



(a) **CIFAR-100.** There is a peak in test error even with no label noise.



(b) **CIFAR-10.** There is a "plateau" in test error around the interpolation point with no label noise, which develops into a peak for added label noise.



(a) Without data augmentation.

(b) With data augmentation.

Figure 5: **Effect of Data Augmentation.** 5-layer CNNs on CIFAR10, with and without data-augmentation. Data-augmentation shifts the interpolation threshold to the right, shifting the test error peak accordingly. Optimized using SGD for 500K steps. See Figure 27 for larger models.

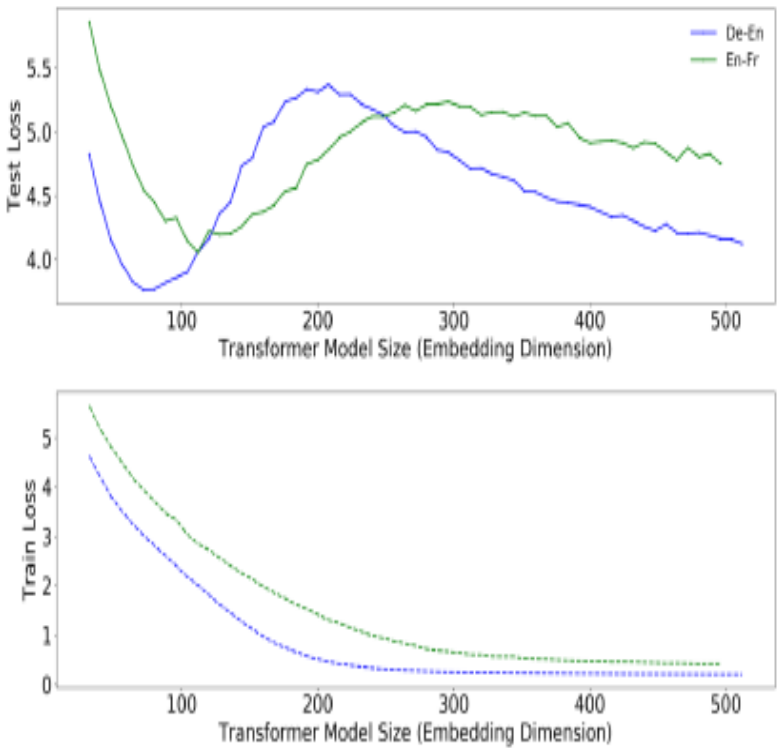
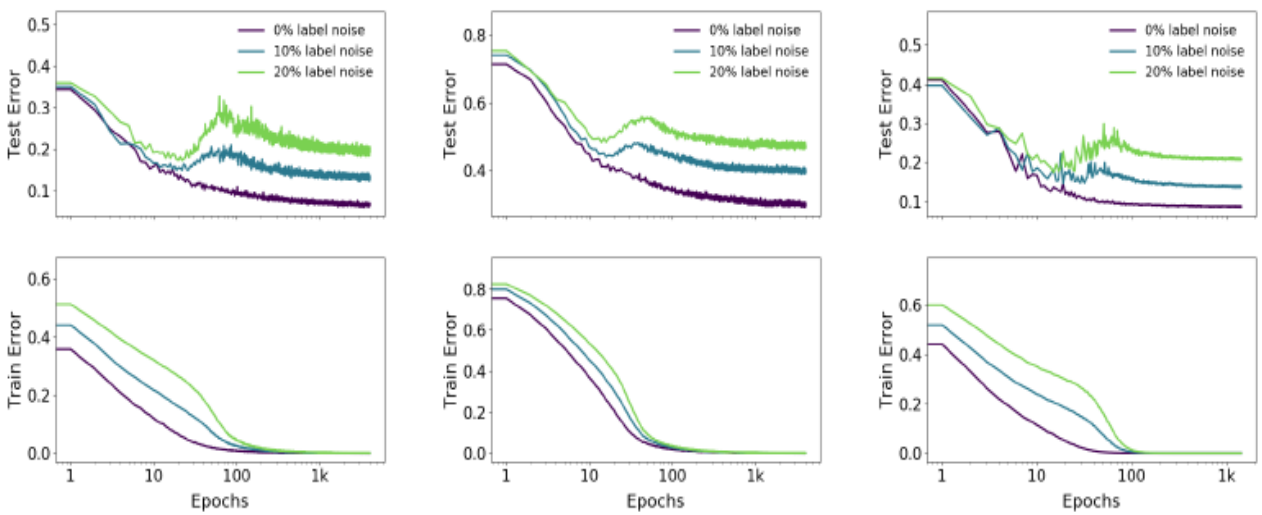


Figure 8: **Transformers on language translation tasks:** Multi-head-attention encoder-decoder Transformer model trained for 80k gradient steps with labeled smoothed cross-entropy loss on IWSLT'14 German-to-English (160K sentences) and WMT'14 English-to-French (subsampled to 200K sentences) dataset. Test loss is measured as per-token perplexity.

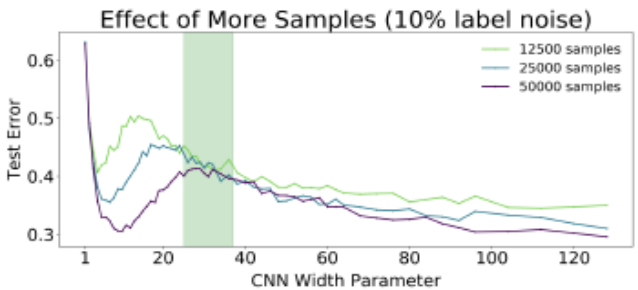
实验二 训练步数



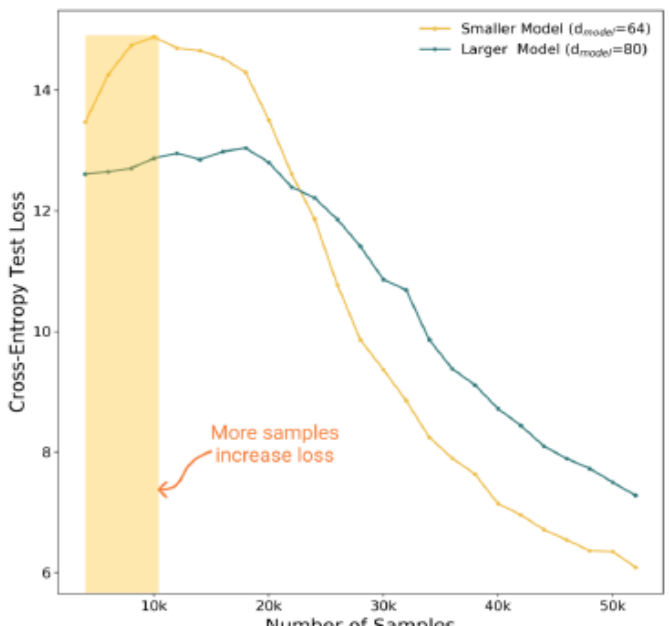
(a) ResNet18 on CIFAR10. (b) ResNet18 on CIFAR100. (c) 5-layer CNN on CIFAR 10.

Figure 10: **Epoch-wise double descent** for ResNet18 and CNN (width=128). ResNets trained using Adam with learning rate 0.0001, and CNNs trained with SGD with inverse-squareroot learning rate.

实验三： 样本数量



(a) Model-wise double descent for 5-layer CNNs on CIFAR-10, for varying dataset sizes. **Top:** There is a range of model sizes (shaded green) where training on 2× more samples does not improve test error. **Bottom:** There is a range of model sizes (shaded red) where training on 4× more samples does not improve test error.



(b) **Sample-wise non-monotonicity.** Test loss (per-word perplexity) as a function of number of train samples, for two transformer models trained to completion on IWSLT'14. For both model sizes, there is a regime where more samples hurt performance. Compare to Figure 3, of model-wise double-descent in the identical setting.

Figure 11: Sample-wise non-monotonicity.

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

- 本篇文章通过传统的偏差-方差分解来为双下降现象寻找一个解释。

偏差方差分解：

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

其中：

通过一些方法来估计模型的偏差和方差

- 1 将数据集分割为一些子集的无交并，通常每个子集包含两个元素：

$$\mathcal{T} = \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_N$$

- 2 均值 $\bar{f}(x) = \mathbb{E}_{\mathcal{T}} f(x, \mathcal{T})$.的估计：

$$\sum_{i=1}^{N} \frac{1}{N} f(x, \mathcal{T}_i)$$

- 3方差的估计：

$$\widehat{\mathrm{var}}(x, \mathcal{T}) \quad \frac{1}{N-1} \sum_{i=1}^{N} \left\| f(x, \mathcal{T}_i) \quad \sum_{i=1}^{N} \frac{1}{N} f(x, \mathcal{T}_i) \right\|_2^2$$

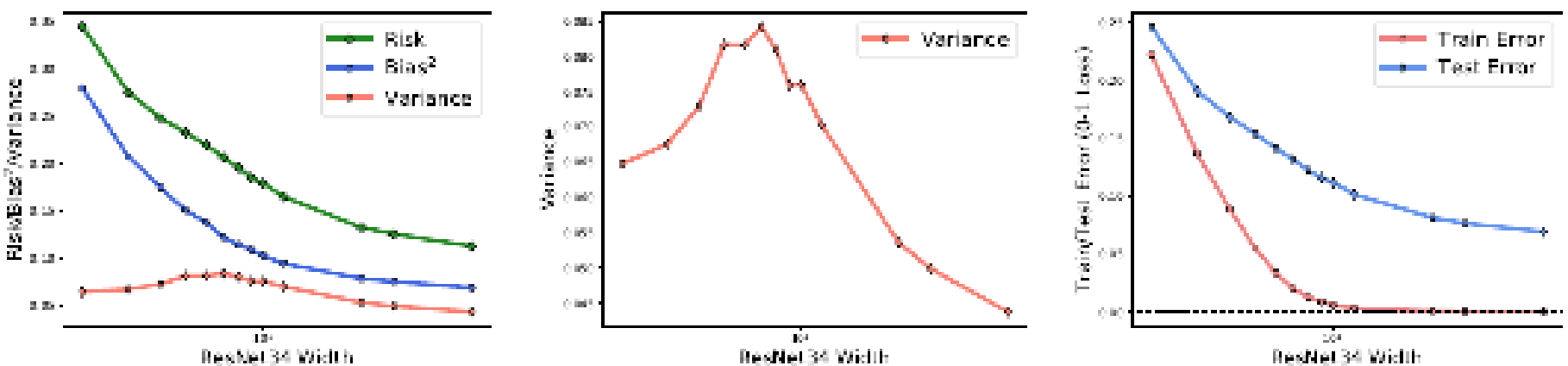


Figure 2. Mainline experiment on ResNet34, CIFAR10 dataset (25,000 training samples). **(Left)** Risk, bias, and variance for ResNet34. **(Middle)** Variance for ResNet34. **(Right)** Train error and test error for ResNet34.

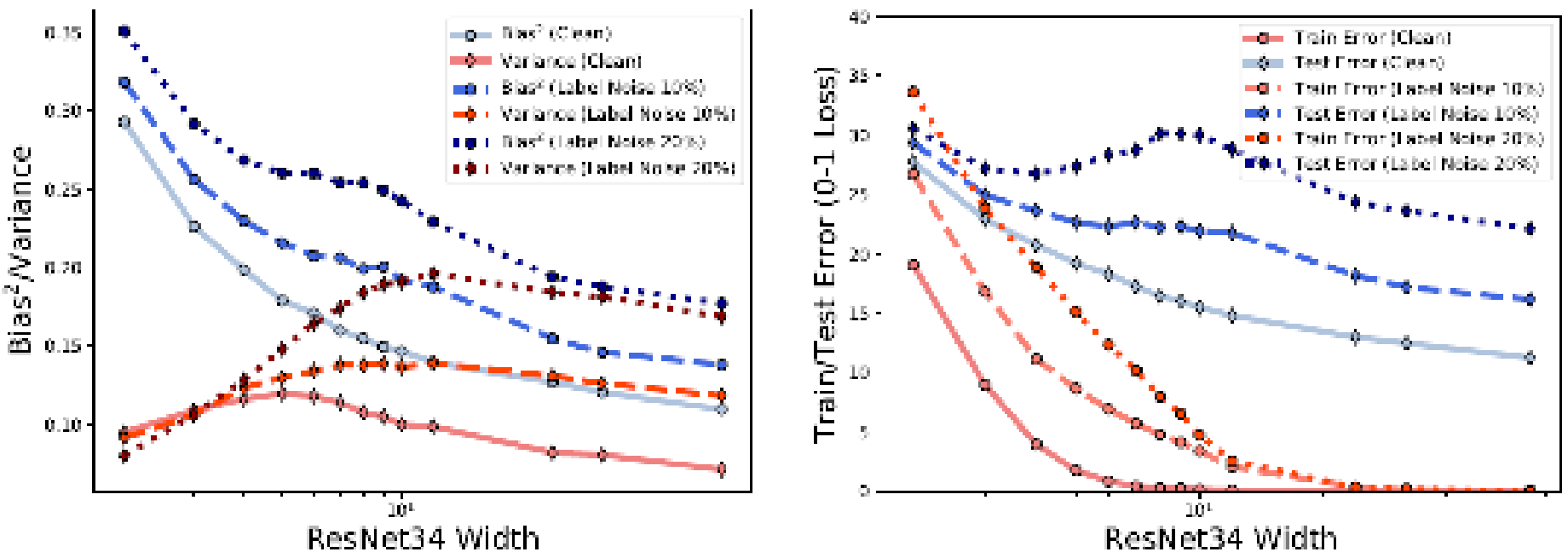
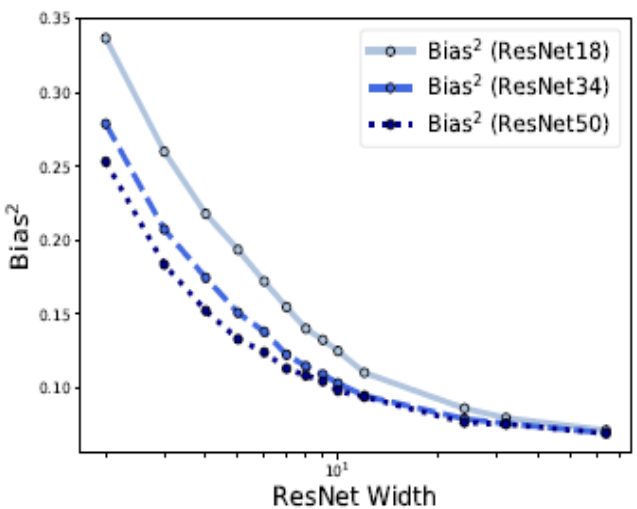
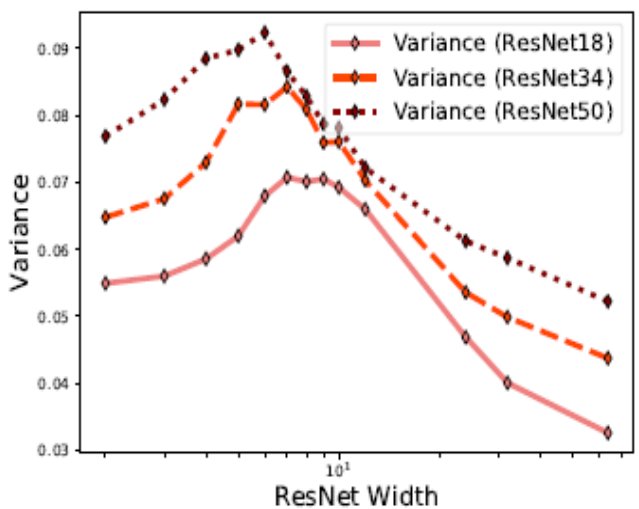


Figure 4. Increasing label noise leads to double-descent. **(Left)** Bias and variance under different label noise percentage. **(Right)** Training error and test error under different label noise percentage.

模型深度的影响：



(b) Bias of model with different depth



(c) Variance of model with different depth

本篇文章遗留的问题：

- 1 虽然偏差和方差随模型复杂度的变化规律解释了双下降现象，但是，方差的变化规律依然没有得到证明的，本文相当于把问题转化了，但并未从根本上解决问题。
- 2 主要是研究宽度变化的影响，对深度的影响研究不足。

更复杂的网络，意味着包含“更好”的函数

- 以二层神经网络为例，研究其在手写字体识别上表现
- 得到的结论是随着网络变宽，学习到的函数的范数会更小，意味着函数更加平滑。
- 观点：函数范数的双下降导致了测试误差的双下降

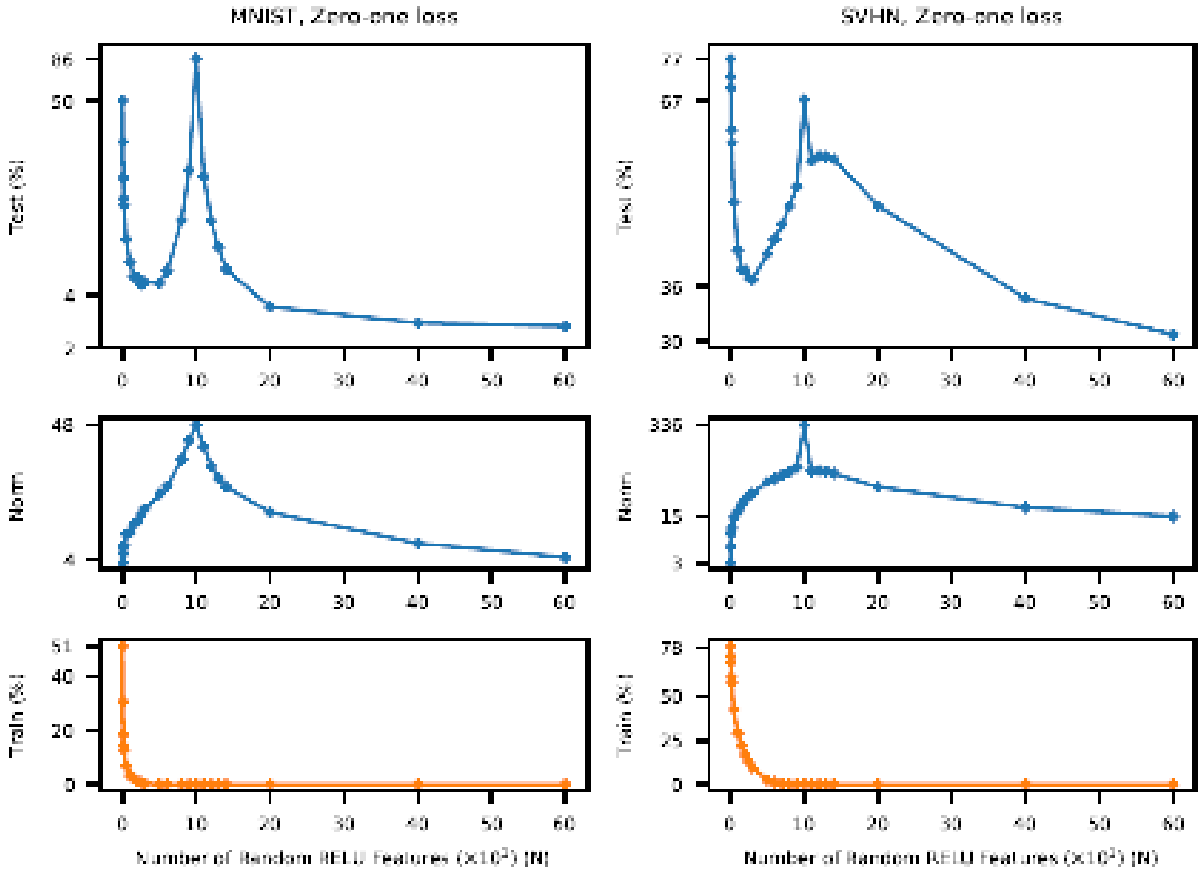


Figure 8: **Double descent risk curve for Random ReLU model.** Test risks (log scale), coefficient ℓ_2 norms (log scale), and training risks of the Random ReLU Features model predictors $h_{n,N}$ learned on subsets of MNIST and SVHN data ($n = 10^4$). The interpolation threshold is achieved at $N = 10^4$. Regularization of $4 \cdot 10^{-6}$ is added for SVHN to ensure numerical stability near interpolation threshold.

线性模型的双下降理论分析

- 由于复杂网络的理论证明难度较大，这里用线性模型来证明双下降。

We consider a regression problem where the response y is equal to a linear function $\beta = (\beta_1, \ldots, \beta_D) \in \mathbb{R}^D$ of D real-valued variables $x = (x_1, \ldots, x_D)$ plus noise $\sigma\epsilon$:

$$y = x^\top \beta + \sigma\epsilon = \sum_{j=1}^{D} x_j \beta_j + \sigma\epsilon.$$

The learner observes n iid copies $((x^{(i)}, y^{(i)}))_{i=1}^n$ of (x, y), but fits a linear model to the data only using a subset $T \subseteq [D] := \{1, \ldots, D\}$ of $p := |T|$ variables.

Let $X := [x^{(1)}|\cdots|x^{(n)}]^\intercal$ be the $n \times D$ design matrix, and let $y := (y^{(1)}, \ldots, y^{(n)})$ be the vector of responses. For a subset $A \subset [D]$ and a D-dimensional vector v, we use $v_A := (v_j : j \in A)$ to denote its $|A|$-dimensional subvector of entries from A; we also use $X_A := [x_A^{(1)}|\cdots|x_A^{(n)}]^\intercal$ to denote the $n \times |A|$ design matrix with variables from A. For $A \subset [D]$, we denote its complement by $A^c := [D] \setminus A$. Finally, $\|\cdot\|$ denotes the Euclidean norm.

The learner fits regression coefficients $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_D)$ with

$$\hat{\beta}_T := X_T^\dagger y, \quad \hat{\beta}_{T^c} := 0.$$

Theorem 1. *Assume the distribution of x is the standard normal in $\mathbb{R}^D$ and $y = x^T\beta$ for some $\beta \in \mathbb{R}^D$. Pick any $p \in \{0, \ldots, D\}$ and $T \subseteq [D]$ of cardinality p. The risk of $\hat{\beta}$, where $\hat{\beta}_T = X_T^{\dagger} y$ and $\hat{\beta}_{T^c} = 0$, is*

$$\mathbb{E}[(y - x^T\hat{\beta})^2] = \begin{cases} \|\beta_{T^c}\|^2 \cdot \left(1 + \frac{p}{n} \frac{p}{p} 1\right) & \text{if } p \le n-2; \\ +\infty & \text{if } n-1 \le p \le n+1 \text{ and } \beta_{T^c} \neq 0; \\ \|\beta_T\|^2 \cdot \left(1 \quad \frac{n}{p}\right) + \|\beta_{T^c}\|^2 \cdot \left(1 + \frac{n}{p-n-1}\right) & \text{if } p > n+2; \\ \|\beta_T\|^2 \cdot \max\left\{1 - \frac{n}{p}, 0\right\} & \text{if } \beta_{T^c} = 0. \end{cases}$$

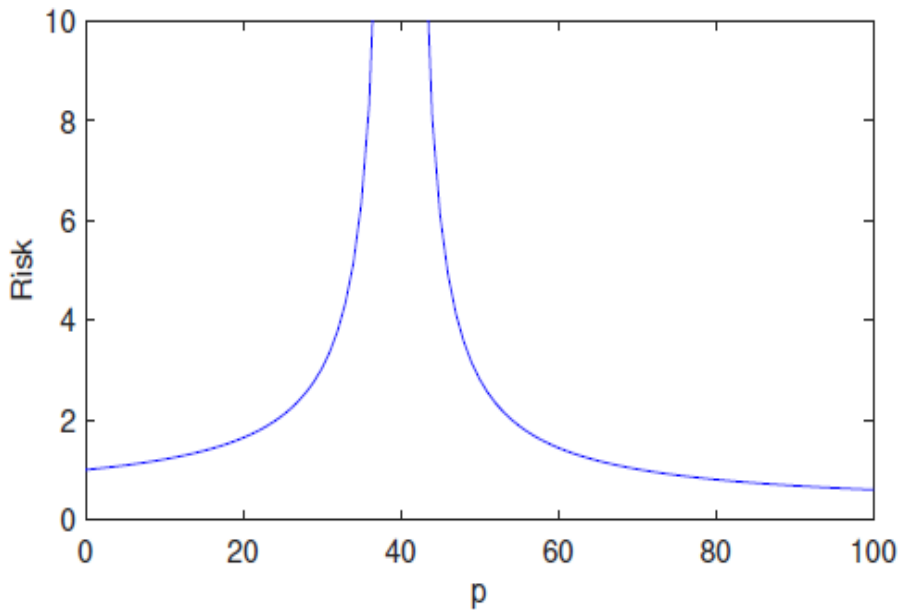


Figure 1: Plot of risk $\mathbb{E}[(y - x^\intercal \hat{\beta})^2]$ as a function of p, under the random selection model of T. Here, $\|\beta\|^2 = 1$, $\sigma^2 = 0$, $D = 100$, and $n = 40$.

模型的容量是非常关键的

- 只有当模型的容量足够大时，才能选到比较“好”的函数。
- 而深度学习的泛化能力强，可能正是因为深度神经网络所能表达的函数更多。
- 这一点在16年左右已经被发现了，证明了相同宽度下三层网络的容量是远远高于两层网络的。

Assumption 1 *Given the activation function σ, there is a constant $c_\sigma \geq 1$ (depending only on σ) such that the following holds: For any L-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ which is constant outside a bounded interval $[-R, R]$, and for any δ, there exist scalars $a, \{\alpha_i, \beta_i, \gamma_i\}_{i=1}^{w}$, where $w \leq c_\sigma \frac{RL}{\delta}$, such that the function*

$$h(x) = a + \sum_{i=1}^{w} \alpha_i \cdot \sigma(\beta_i x - \gamma_i)$$

satisfies

$$\sup_{x \in \mathbb{R}} |f(x) - h(x)| \leq \delta.$$

Assumption 2 *The activation function σ is (Lebesgue) measurable and satisfies*

$$|\sigma(x)| \leq C(1 + |x|^\alpha)$$

for all $x \in \mathbb{R}$ and for some constants $C, \alpha \geq 0$.

对于三层网络所能拟合的函数，二层网络需要指数级宽度

Theorem 1 *Suppose the activation function $\sigma(\cdot)$ satisfies assumption 1 with constant c_σ, as well as assumption 2. Then there exist universal constants $c, C > 0$ such that the following holds: For every dimension $d > C$, there is a probability measure μ on $\mathbb{R}^d$ and a function $g : \mathbb{R}^d \to \mathbb{R}$ with the following properties:*

1. *g is bounded in $[-2, +2]$, supported on $\{\mathbf{x} : \|\mathbf{x}\| \leq C\sqrt{d}\}$, and expressible by a 3-layer network of width $Cc_\sigma d^{19/4}$.*

2. *Every function f, expressed by a 2-layer network of width at most ce^{cd}, satisfies*

$$\mathbb{E}_{\mathbf{x}\sim\mu} \left(f(\mathbf{x}) - g(\mathbf{x})\right)^2 \geq c.$$

总结：

- 目前解释双下降的角度主要有两个
- 1偏差-方差分解
- 2模型的容量

- 但是，还是有很多疑问有待解决，比如学习算法是如何自动选择比较好的函数。

Thanks for your listening!

Q & A