人工智能与深度学习

3-1 偏差-方差分析

王中雷

经济学院、王亚南经济研究院、邹至庄经济研究院

厦门大学

（第一版）

内容摘要

1. 模型超参数

2. 偏差与方差

3. 双下降现象

回顾

1．我们已经讨论了全连接神经网络模型，其包含两类参数：

- 模型参数：$\{(\boldsymbol{b}^{[l]}, \boldsymbol{W}^{[l]}) : l = 1, \ldots, L\}$
  - ▷ 它们可以通过梯度下降法进行学习
- 模型超参数，通常不可通过梯度下降法进行估计

2．我们的目标变了

- 不再关心如何通过减小训练误差，来估计模型参数
- 聚焦泛化误差来选择模型超参数

模型超参数

1. α：学习率
2. L：神经网络模型的层数
3. $\{d^{[l]} : l = 1, \ldots, L-1\}$：第$l$ 层神经元个数
4. m：小批量梯度下降法对应的样本量
5. 梯度下降法
6. 梯度下降法的参数更新所需要的迭代次数
7. $\cdots$

符号

1. $\boldsymbol{x}$：特征向量
2. y：观测标签
3. $S = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$：训练集（具有随机性）
4. y_t：对应于特征 $\boldsymbol{x}$ 的真实标签，产生于真实模型（例如，$y_t = E(y \mid \boldsymbol{x})$）
5. $\hat{y}$：基于训练集S，并利用（工作）模型，对真实标签y_t 的估计
6. y_t 和 $\hat{y}$ 均是特征 $\boldsymbol{x}$ 的函数

基于平均平方损失的视角

1．给定一个特征 $\boldsymbol{x}$

$$
\begin{aligned}
E_S\{(\hat{y}-y_t)^2 \mid \boldsymbol{x}\} &= E_S[\{\hat{y}-E_S(\hat{y} \mid \boldsymbol{x})+E_S(\hat{y} \mid \boldsymbol{x})-y_t\}^2 \mid \boldsymbol{x}] \\
&= E_S[\{E_S(\hat{y} \mid \boldsymbol{x})-y_t\}^2+2\cdot\{E_S(\hat{y} \mid \boldsymbol{x})-y_t\}\cdot\{\hat{y}-E_S(\hat{y} \mid \boldsymbol{x})\}+\{\hat{y}-E_S(\hat{y} \mid \boldsymbol{x})\}^2 \mid \boldsymbol{x}] \\
&= E_S[\{E_S(\hat{y} \mid \boldsymbol{x})-y_t\}^2 \mid \boldsymbol{x}]+2\cdot\{E_S(\hat{y} \mid \boldsymbol{x})-y_t\}\cdot E_S[\{\hat{y}-E_S(\hat{y} \mid \boldsymbol{x})\} \mid \boldsymbol{x}] \\
&\quad + E_S[\{\hat{y}-E_S(\hat{y} \mid \boldsymbol{x})\}^2 \mid \boldsymbol{x}] \\
&= \{E_S(\hat{y} \mid \boldsymbol{x})-y_t\}^2+E_S[\{\hat{y}-E_S(\hat{y} \mid \boldsymbol{x})\}^2 \mid \boldsymbol{x}] \\
&= \text{（偏差）}^2+\text{方差}
\end{aligned}
$$

- $E_S(\cdot)$：关于训练集S 随机性的条件期望
- $E_S(\hat{y} \mid \boldsymbol{x})-y_t$：没有随机性，衡量用一个预设的模型近似真实标签 y_t 的平均偏差
- $\hat{y}-E_S(\hat{y} \mid \boldsymbol{x})$：随机变量，衡量模型估计标签 $\hat{y}$，对训练集 S 依赖的敏感性

偏差和方差

1．偏差（Bias）

$$\text{Bias}(\hat{y}) = E_S(\hat{y} \mid \boldsymbol{x}) - y_t$$

- $E_S(\cdot)$：针对训练集S 的条件期望

2．方差（Variance）

$$\text{Variance}(\hat{y}) = E_S[\{\hat{y} - E_S(\hat{y} \mid \boldsymbol{x})\}^2 \mid \boldsymbol{x}]$$

3．偏差和方差，都是针对一个给定的特征$\boldsymbol{x}$ 计算的

例子--均值估计

1. 考虑如下问题

$$y \mid \boldsymbol{x} = \mu + \epsilon$$

- $E(\epsilon \mid \boldsymbol{x}) = 0, \quad \text{Variance}(\epsilon \mid \boldsymbol{x}) = \sigma^2$
- 真实回归模型是关于特征$\boldsymbol{x}$ 的常值函数
- 真实标签：$y_t = E(y \mid \boldsymbol{x}) = \mu$

2. 对于一个新的特征 $\boldsymbol{x}$，其对应的标签估计值为

$$\hat{y} = \hat{\mu}$$

- $\hat{\mu} = n^{-1} \sum_{i=1}^{n} y_i$
- (工作) 模型也是常值函数$f(\boldsymbol{x}) = c$，其中，模型参数为c

例子--均值估计

1．偏差（Bias）

$$\begin{aligned}\text{Bias}(\hat{y}) &= E_S(\hat{y} \mid \boldsymbol{x}) - y_t \\ &= E_S(\hat{\mu}) - \mu = 0\end{aligned}$$

2．方差（Variance）

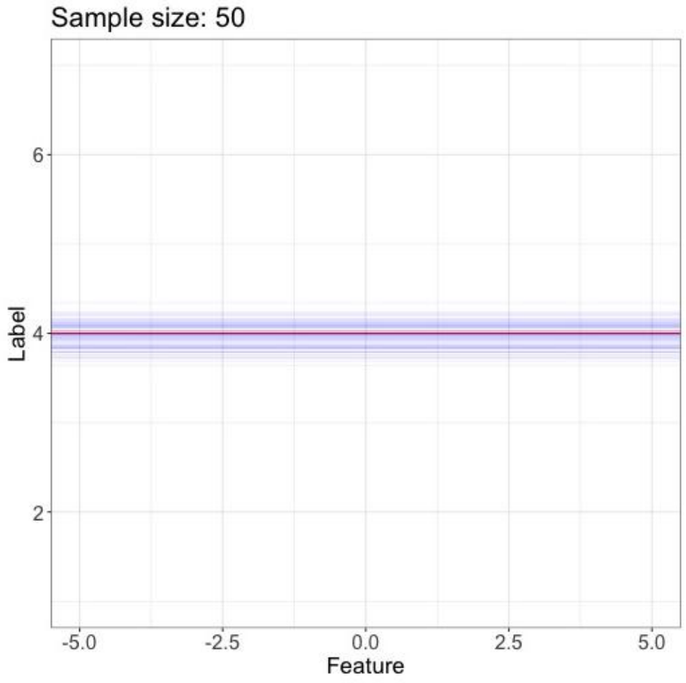
$$\begin{aligned}\text{Variance}(\hat{y}) &= E_S[\{\hat{y} - E_S(\hat{y})\}^2 \mid \boldsymbol{x}] \\ &= \text{Variance}(\hat{\mu}) \\ &= n^{-1}\sigma^2\end{aligned}$$

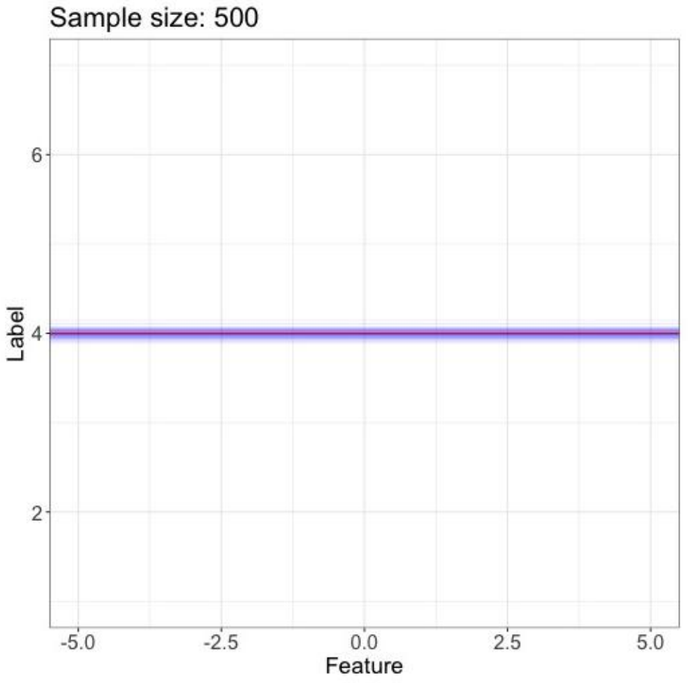
3．我们已经在数理统计课程中，讨论过以上内容

例子--均值估计

1. 一个小例子

- $x_i \sim \text{Uniform}[-5, 5], \quad y_i \mid x_i = 6 + \epsilon_i, \quad \epsilon_i \sim N(0, 1^2) \quad (i = 1, \ldots, n)$





例子--线性回归

1．考虑下面的模型设定

$$y \mid \boldsymbol{x} = b_0 + \boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{w}_0 + \epsilon$$

- 真实模型是关于特征$\boldsymbol{x}$ 的线性函数，模型参数是 b_0 和 $\boldsymbol{w}_0$
- 真实标签为：$y_t = E(y \mid \boldsymbol{x}) = b_0 + \boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{w}_0$
- $E(\epsilon \mid \boldsymbol{x}) = 0, \quad \mathrm{Variance}(\epsilon \mid \boldsymbol{x}) = \sigma^2$

例子--线性回归

1．对于一个新的特征 $\boldsymbol{x}$，对应的估计标签为

$$\hat{y} = \hat{b} + \boldsymbol{x}^{\mathrm{T}} \cdot \hat{\boldsymbol{w}}$$

- (工作) 模型是$f(\boldsymbol{x}; \boldsymbol{\theta}) = b + \boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{w}$，模型参数为$\boldsymbol{\theta} = (b, \boldsymbol{w}^{\mathrm{T}})^{\mathrm{T}}$
- $\hat{b}, \hat{\boldsymbol{w}}$： 最小化如下代价函数

$$n^{-1} \sum_{i=1}^{n} (y_i - b - \boldsymbol{x}_i^{\mathrm{T}} \cdot \boldsymbol{w})^2$$

- 关于计算细节，请参见第一章

例子--线性回归

1．我们有如下结论

$$E_S(\hat{b}) = b_0 \quad E_S(\hat{\boldsymbol{w}}) = \boldsymbol{w}_0$$

- 模型参数估计是无偏的

2．估计标签的偏差（Bias）

$$\begin{aligned}\text{Bias}(\hat{y}) &= E_S(\hat{y} \mid \boldsymbol{x}) - y_t \\ &= E_S(\hat{b} + \boldsymbol{x}^{\mathrm{T}} \cdot \hat{\boldsymbol{w}} \mid \boldsymbol{x}) - (b_0 + \boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{w}_0) = 0\end{aligned}$$

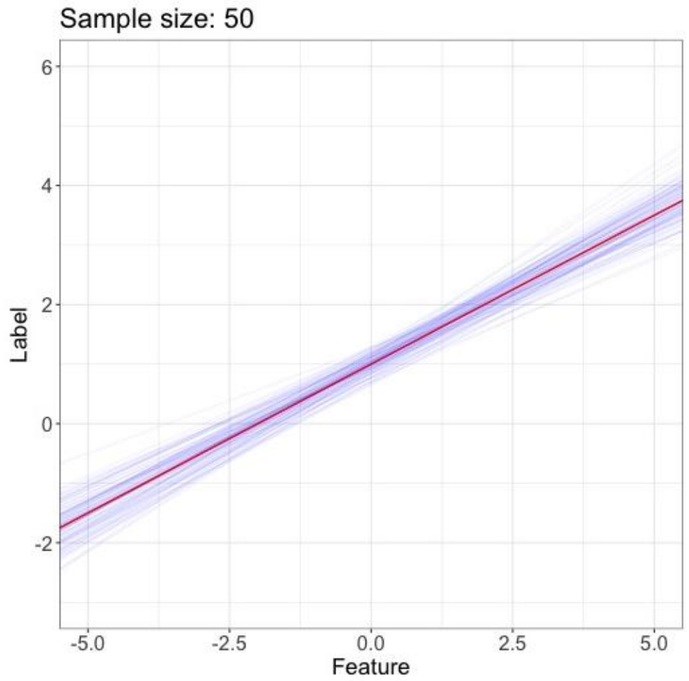
3．估计标签的方差（Variance）

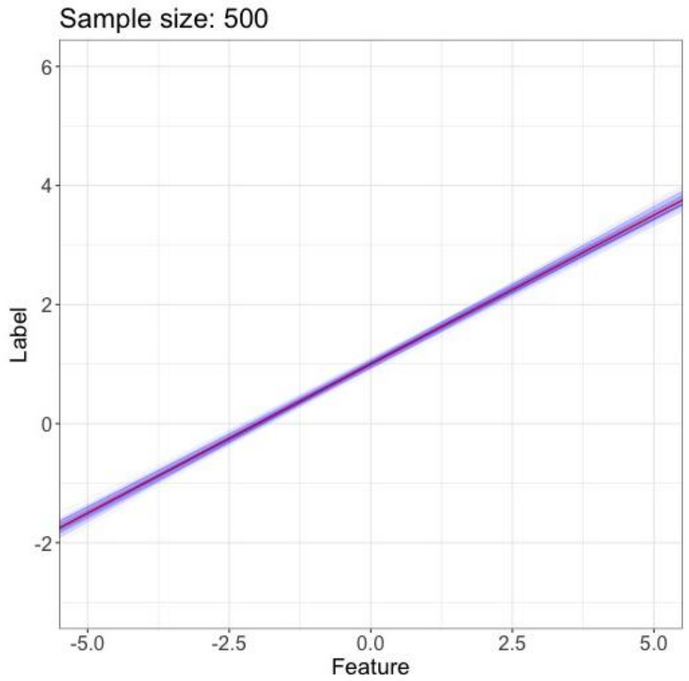
$$\begin{aligned}\text{Variance}(\hat{y}) &= E_S[\{\hat{y} - E_S(\hat{y})\}^2 \mid \boldsymbol{x}] \\ &= \text{Variance}(\hat{b} + \boldsymbol{x}^{\mathrm{T}} \cdot \hat{\boldsymbol{w}} \mid \boldsymbol{x}) = \text{（参见对应参考书）}\end{aligned}$$

例子--线性回归

1. 一个小例子

- $x_i \sim \text{Uniform}[-5, 5], \quad y_i \mid x_i = 1 + 0.5x_i + \epsilon_i, \quad \epsilon_i \sim N(0, 1^2) \quad (i = 1, \ldots, n)$





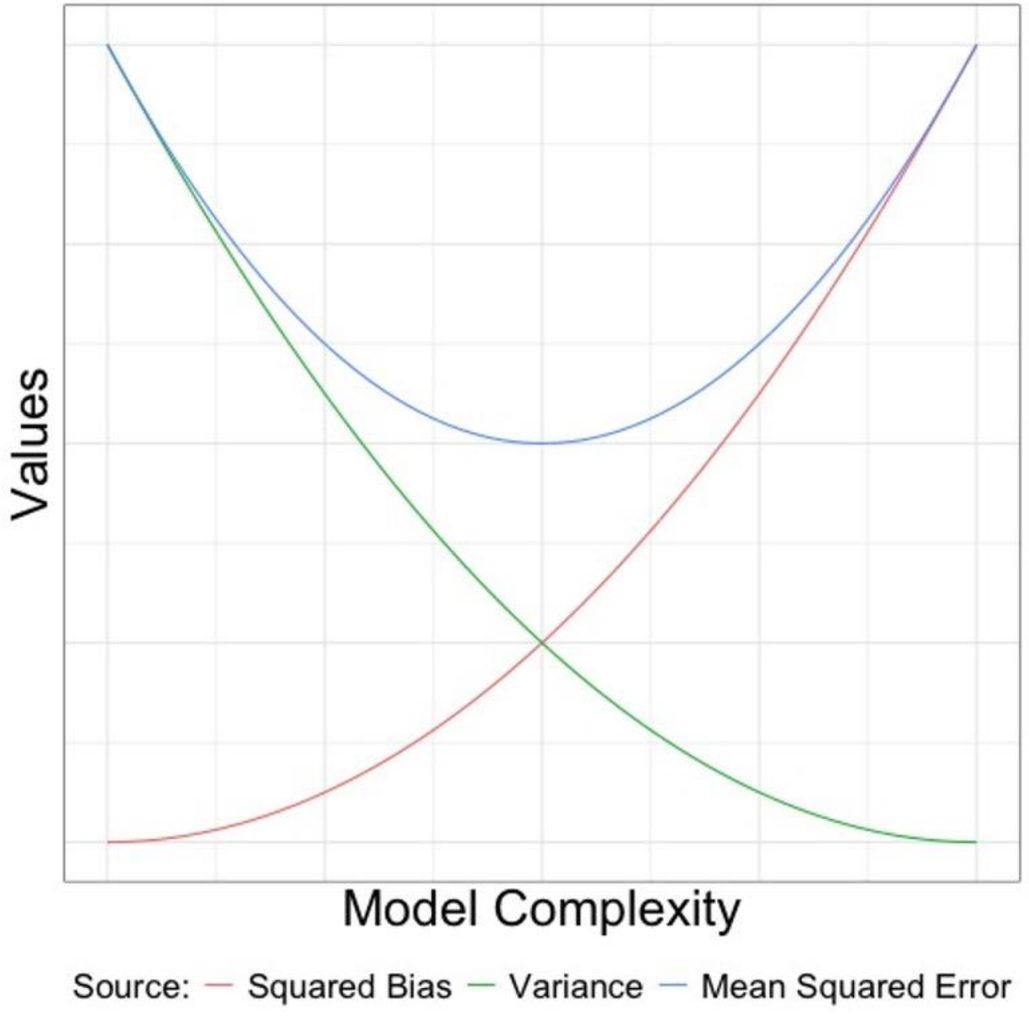
例子--岭回归

1. 我们仍然考虑线性回归模型
2. 模型参数，通过最小化如下代价函数获得

$$\sum_{i=1}^{n}(y_i - b - \boldsymbol{x}_i^{\mathrm{T}} \cdot \boldsymbol{w})^2 + \lambda \sum_{j=1}^{d} w_j^2$$

- $\boldsymbol{w} = (w_1, \ldots, w_d)^{\mathrm{T}}$
- 超参数 $\lambda > 0$ 用来控制估计模型的复杂程度
- 当 $\lambda > 0$ 时，岭回归的参数估计量通常是有偏的（请参见统计学相关教材）

经典偏差-方差分析


Values
Model Complexity
Source: Squared Bias Variance Mean Squared Error

调整经典模型的超参数

1．交叉验证（Cross validation），常被用于传统统计学习模型的超参数选择

2．通常来说，其并不用于深度神经网络模型超参数的选择

3．对于神经网络模型，我们常通过一个验证集，选择模型的超参数

- 训练集：训练神经网络模型
- 验证集：评价不同超参数下，模型的泛化性能，并选择一组最优的超参数
  - 不同超参数，对应着不同的神经网络模型
  - 选择一组好的超参数，等价于选择一个好的模型
- 测试集：在模型结构和超参数全部确定后，仅用于最终泛化性能评价

调整超参数

1．原则：

- 首先，降低模型的估计偏差
  - 考虑更加复杂的模型结构、改进特征、减小正则化、充分优化
- 当模型偏差得到控制时，我们再考虑降低方差
  - 增加训练集规模（较为昂贵）、适当降低模型复杂度、加强正则化

双下降（Double descent）现象

1．对于传统统计模型而言，

- 简单的模型，往往具有较大偏差但较小方差
- 复杂的模型，往往具有较小偏差但较大方差

2．基于经典理论，随着模型复杂度的增加

- 偏差降低
- 方差升高
- 因此，“复杂的模型往往不是最优的!”

双下降（Double descent）现象

1．深度学习的兴起，从根本上挑战了经典的偏差-方差分析理论

- 随着模型复杂度的增加，测试误差可能先下降、后上升，并在越过插值阈值后再次下降
- 双下降现象出现于模型规模的进一步增加，也会随着训练周期的增加而出现
- 关于双下降现象，请参见论文 (Belkin et al., 2019; Nakkiran et al., 2019)

双下降（Double descent）现象

1．下面的图片来自于 Nakkiran et al. (2019) 的图 1

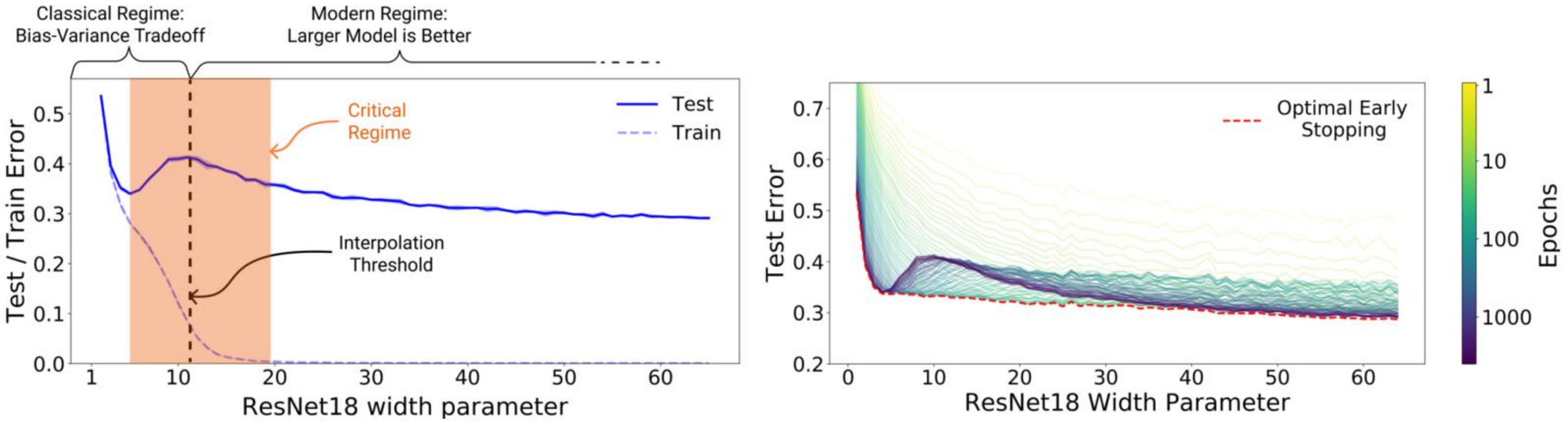


Figure 1: **Left:** Train and test error as a function of model size, for ResNet18s of varying width on CIFAR-10 with 15% label noise. **Right:** Test error, shown for varying train epochs. All models trained using Adam for 4K epochs. The largest model (width 64) corresponds to standard ResNet18.

双下降（Double descent）现象

1. 泛化误差的三个区间
   - 传统区间：模型参数规模远小于训练样本个数
   - 临界区间：模型参数规模与训练样本个数具有相同数量级
   - 现代区间：模型参数规模远大于训练样本个数
2. 更大规模的训练集，反而可能暂时降低模型的泛化能力（不一定发生）
   - 更多的训练样本，可能将模型从现代区间拉到临界区间
   - 增加数据后应重新评估模型复杂度和训练设置

双下降（Double descent）现象

1．下面的图片来自于 Nakkiran et al. (2019) 的图 2

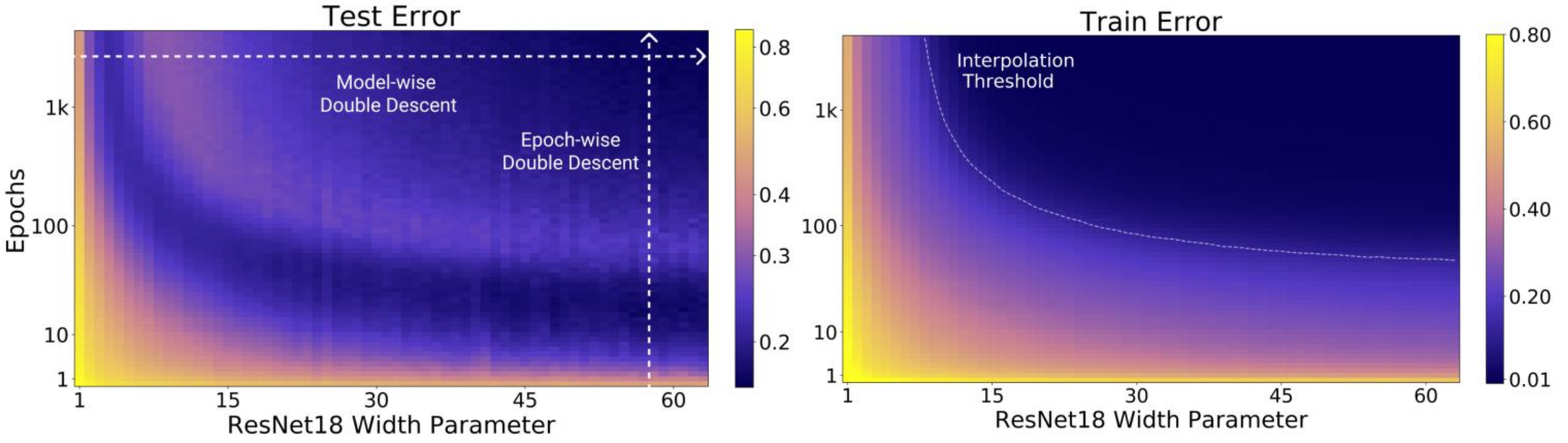


Figure 2: **Left:** Test error as a function of model size and train epochs. The horizontal line corresponds to model-wise double descent–varying model size while training for as long as possible. The vertical line corresponds to epoch-wise double descent, with test error undergoing double-descent as 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.

课程总结

1．超参数是决定模型复杂度和泛化性能的重要因素

2．偏差-方差分析，可帮助我们选择超参数

3．经典 U 型偏差-方差曲线并非唯一可能的泛化误差形态

4．双下降现象告诉我们，一个更复杂的模型，其泛化性能并不一定会变得更差