3-1 偏差-方差分析

王中雷

厦门大学王亚南经济研究院和经济学院, 2025

内容摘要

1. 模型超参

2. 模型分析

回顾

1．我们已经讨论了权利爱你接神经网络，其包含两类参数：

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

模型超参

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次数
7. $\cdots$

符号

1. $\boldsymbol{x}$：特征向量
2. y：标签
3. $S = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$：训练集
4. y_t：对应于特征向量 $\boldsymbol{x}_t$ 的标签（例如，$y_t = E(y \mid \boldsymbol{x}_t)$）
5. $\hat{y}$：基于S，并利用（工作）模型对真实标签y_t 的估计

偏差和方差

1．偏差（Bias）

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

- $E_S(\cdot)$：给定样本S 时的条件概率
- 偏差和方差都是针对一个给定的特征$\boldsymbol{x}$ 计算的

2．方差（Variance）

$$\text{Variance}(\hat{y}) = E_S\{\hat{y} - E_S(\hat{y})\}^2$$

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向量$\boldsymbol{x}$ 的常值函数
- 真实标签：$y_t = E(y \mid \boldsymbol{x}_t) = \mu$

2. 对于一个新的特征向量 $\boldsymbol{x}$，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）

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

2．方差（Variance）

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

3．我们已经在数理统计课程中讨论过以上内容

例子--线性回归

1．考虑下面的模型设定

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

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

例子--线性回归

1．对于一个新的特征向量 $\boldsymbol{x}$，其对应指标的估计量为

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

- (工作) 模型是$f(\boldsymbol{x};\boldsymbol{\theta}) = b + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$，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

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

- 计算细节参见第一章

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）

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

3．估计标签的方差（Variance）

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

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

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

- $\boldsymbol{w} = (w_1, \ldots, w_d)^{\mathrm{T}}$
- 超参 λ 用来控制估计模型的复杂程度
- 估计量不再是无偏的，请参见统计学相关教材

双下降现象（Double descent）

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

2．一般而言，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!”

双下降现象（Double descent）

1．对于深度神经网络，随着模型复杂程度的升高，偏差和方差出现了神奇的

双下降现象

- 随着模型规模的增加，模型性能先下降再升高
- 双下降现象出现于模型规模增加，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 (Nakkiran et al., 2019)

双下降现象（Double descent）

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

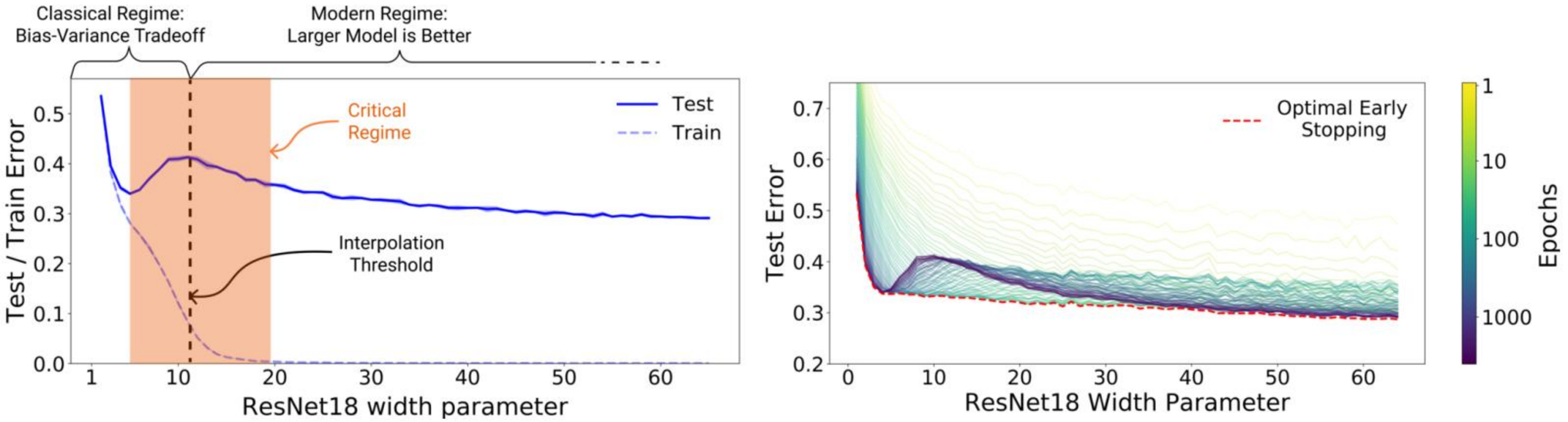

Figure 1: **Left:** Train and test error as a function of model size, for ResNet18s of varying width on CIFAR-10 with 15% label noise. **Right:** Test error, shown for varying train epochs. All models trained using Adam for 4K epochs. The largest model (width 64) corresponds to standard ResNet18.

双下降现象（Double descent）

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

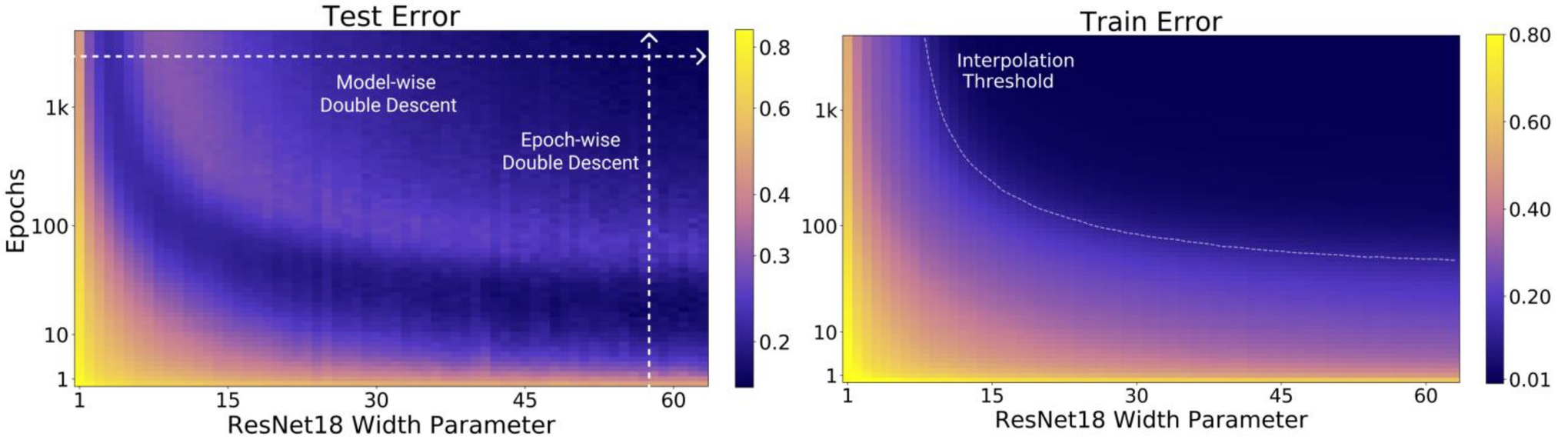

Figure 2: **Left:** Test error as a function of model size and train epochs. The horizontal line corresponds to model-wise double descent–varying model size while training for as long as possible. The vertical line corresponds to epoch-wise double descent, with test error undergoing double-descent as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

调整超参

1. 交叉验证（Cross validation）常被用于传统统计模型的超参选择
2. 并不适用于深度神经网络模型
3. 对于神经网络模型，我们常通过一个验证集调整超参
   - 训练集：训练神经网络
   - 验证集：评价不同超参下不同模型的好坏，并选择一个最优模型
     - 不同超参对应着不同的神经网络模型
     - 选择一组好的超参等价于选择一个 好的模型
   - 测试集（可有可无）：在真实应用场景中，测试已经选择好的模型

调整超参

1．原则：

- 首先，降低模型的估计偏差
  - 增加训练集规模（较为昂贵）
  - 考虑更加复杂的模型结构
- 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）
  - 考虑正则化方法