机器学习A
3.性能评估

王翔

中国科学技术大学
数据科学实验室LDS

Quantifying Mistakes
量化错误

- Model → Algorithm → Estimated Params 模型 → 算法 → 估计的参数
- Predictions → Decisions → Outcomes 预测 → 决策 → 结果
- How do we quantify "happiness" with the outcomes?
  我们如何量化对结果的"满意度"?
  - Predicted too low 预测过低
  - Predicted too high 预测过高

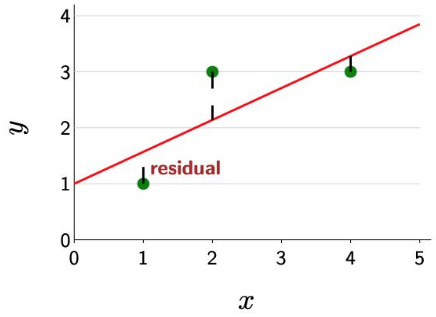
Loss/Cost Function
损失/成本函数

- Loss Function 损失函数
  - Squared Loss 平方损失
  - Absolute Error 绝对误差
  - ……

$$f_{\mathbf{w}}(x) = \mathbf{w} \cdot \phi(x)$$
$$\mathbf{w} = [1, 0.57]$$
$$\phi(x) = [1, x]$$

training data $\mathcal{D}_{\text{train}}$

x	y
1	1
2	3
4	3



$$\text{Loss}(x, y, \mathbf{w}) = (f_{\mathbf{w}}(x) - y)^2$$ **squared loss**

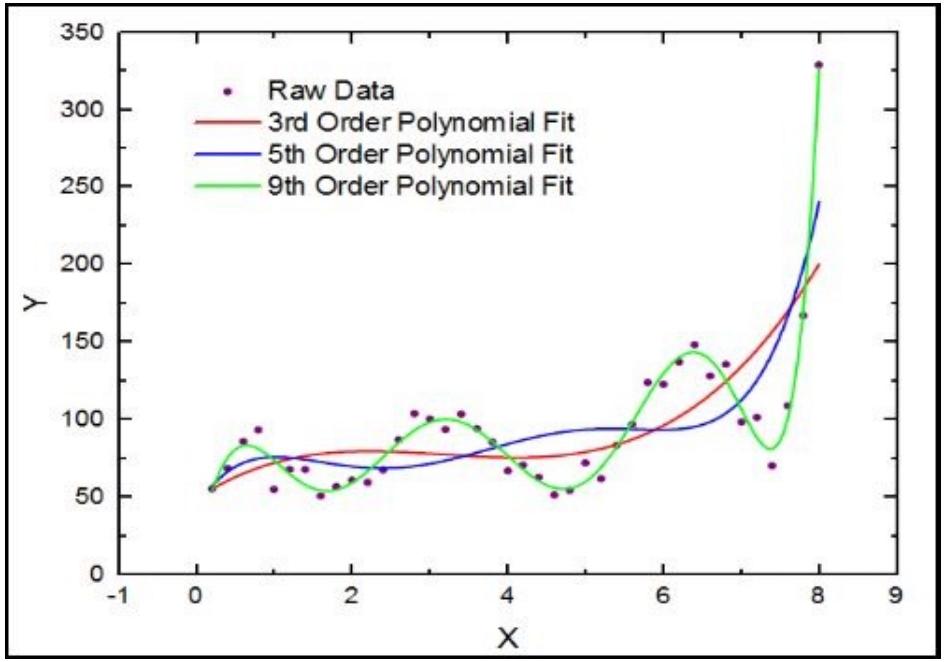
$$\text{Loss}(1, 1, [1, 0.57]) = ([1, 0.57] \cdot [1, 1] - 1)^2 = 0.32$$
$$\text{Loss}(2, 3, [1, 0.57]) = ([1, 0.57] \cdot [1, 2] - 3)^2 = 0.74$$
$$\text{Loss}(4, 3, [1, 0.57]) = ([1, 0.57] \cdot [1, 4] - 3)^2 = 0.08$$

$$\text{TrainLoss}(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y) \in \mathcal{D}_{\text{train}}} \text{Loss}(x, y, \mathbf{w})$$

$$\text{TrainLoss}([1, 0.57]) = 0.38$$

Smallest Loss
最小损失

- Perfect Predictions → Loss = 0 完美预测 → 损失为0
- What “fit” (in the plot on the right) has the lowest loss? 哪种拟合（见右图）具有最低的损失？
- **Do you like it?**
  你喜欢吗？



Roadmap
课程安排

Training Loss → **Generalization Error** → **Test Error** → **Bias-Variance**

Training Loss
训练损失

- Thus far, we have been talking about loss on the training data
  目前为止，我们讨论的是训练数据集上的损失
  - Perfect Predictions (on training data) → Training Loss = 0
    在训练数据上，完美预测 → 训练损失为0

- Determining Training Loss/Error 决定训练损失、误差:
  - Define a loss function, e.g. squared loss, absolute error, …
    定义一个损失函数，例如平方损失，绝对误差

$$L(\mathrm{y}, f_{\hat{w}}(\mathrm{x}))$$

  - Training Error 训练误差:
    - Average loss on training data 平均训练数据上的损失

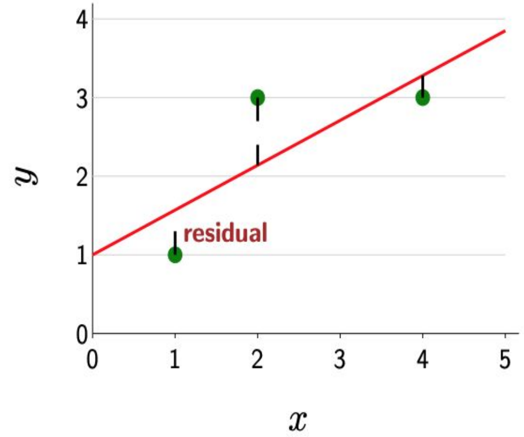
$$\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} L(y, f_{\hat{w}}(x))$$

Training Loss Example
训练损失样例

$$f_{\mathbf{w}}(x) = \mathbf{w} \cdot \phi(x)$$
$$\mathbf{w} = [1, 0.57]$$
$$\phi(x) = [1, x]$$

training data $\mathcal{D}_{\text{train}}$

x	y
1	1
2	3
4	3



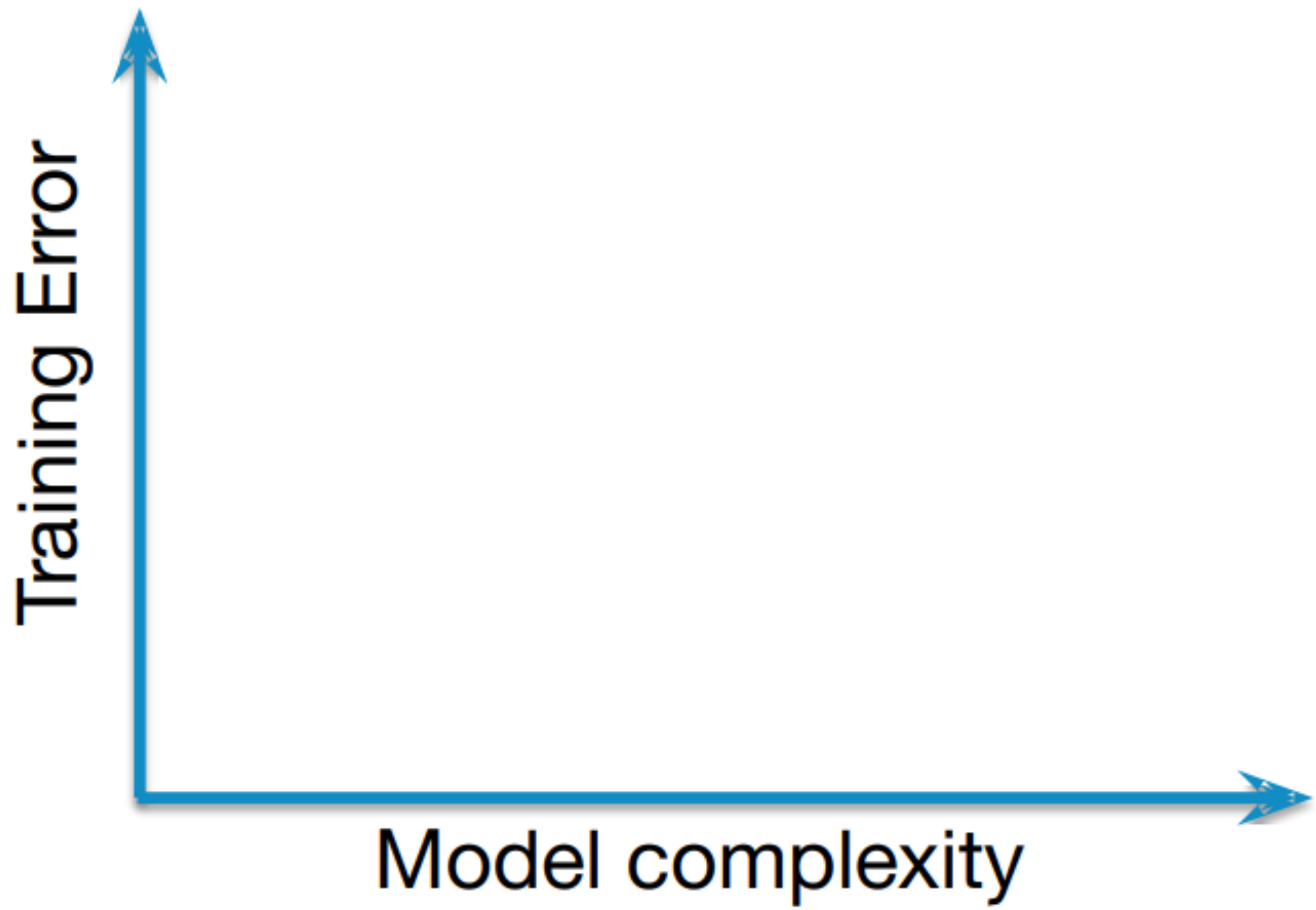
$$\text{Loss}(x, y, \mathbf{w}) = (f_{\mathbf{w}}(x) - y)^2$$ **squared loss**

$$\text{Loss}(1, 1, [1, 0.57]) = ([1, 0.57] \cdot [1, 1] - 1)^2 = 0.32$$
$$\text{Loss}(2, 3, [1, 0.57]) = ([1, 0.57] \cdot [1, 2] - 3)^2 = 0.74$$
$$\text{Loss}(4, 3, [1, 0.57]) = ([1, 0.57] \cdot [1, 4] - 3)^2 = 0.08$$

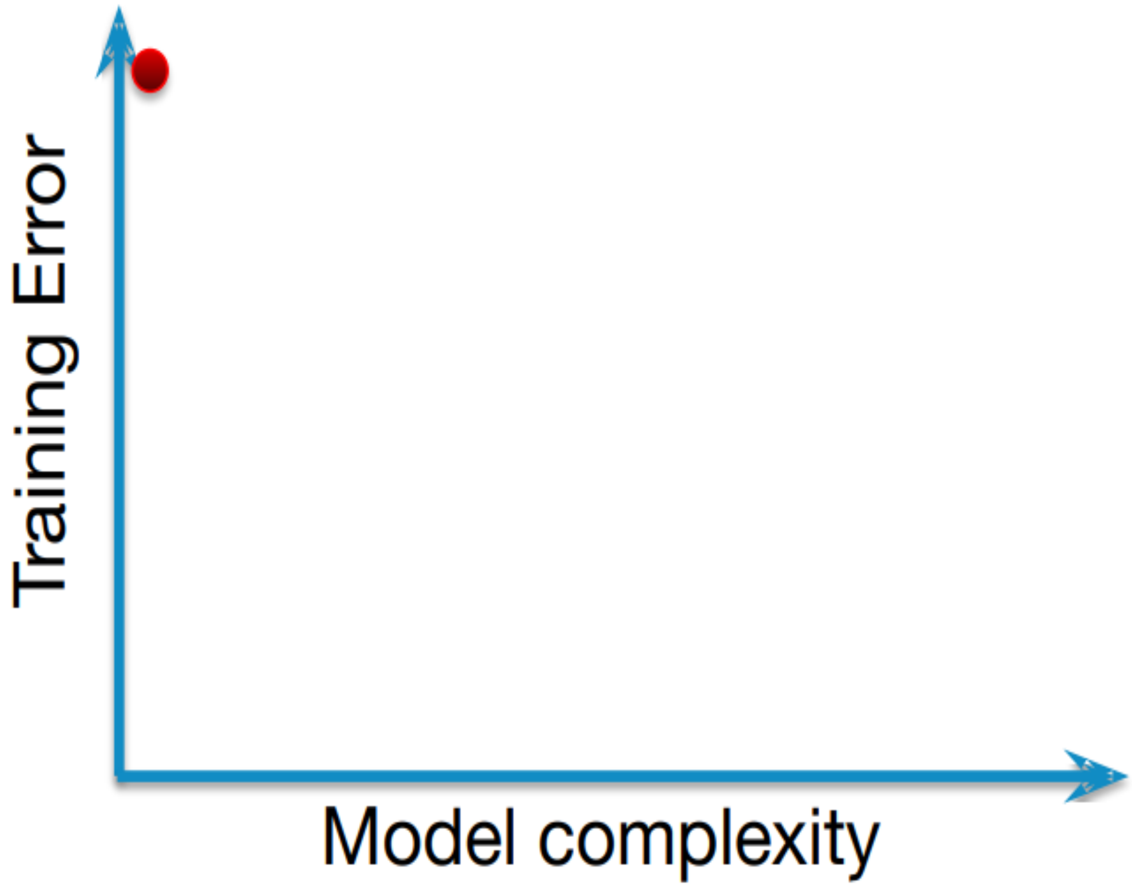
$$\text{TrainLoss}(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y) \in \mathcal{D}_{\text{train}}} \text{Loss}(x, y, \mathbf{w})$$

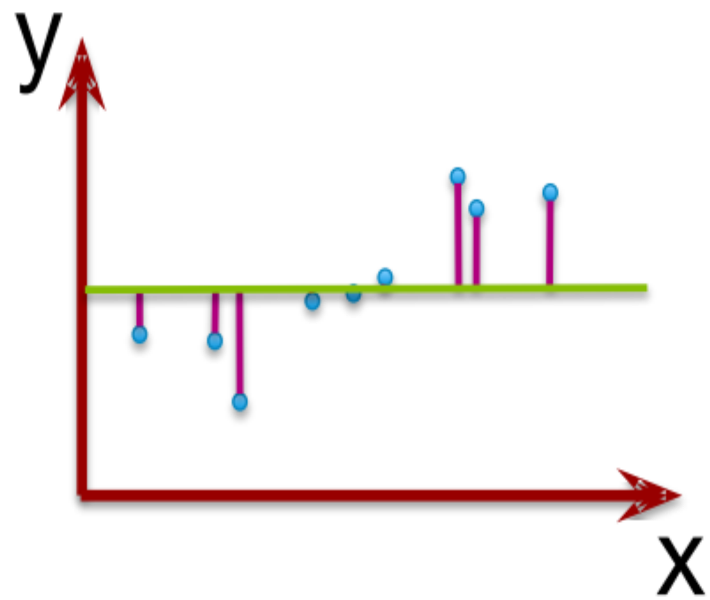
$$\text{TrainLoss}([1, 0.57]) = 0.38$$

Training Loss vs Model Complexity
训练损失 vs 模型复杂度

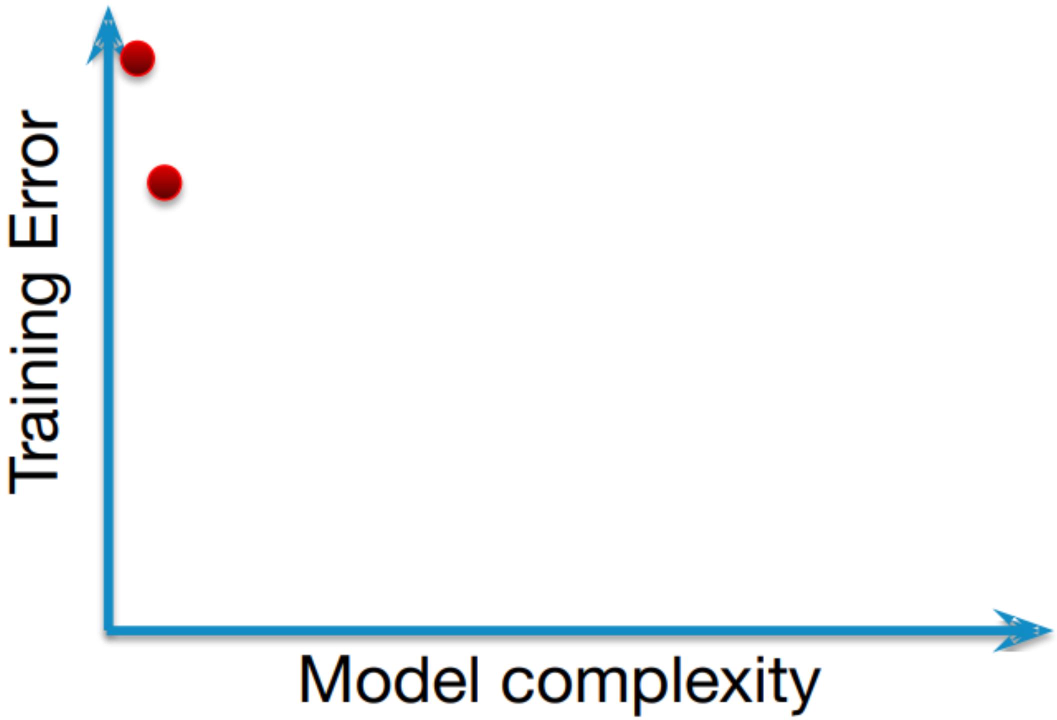


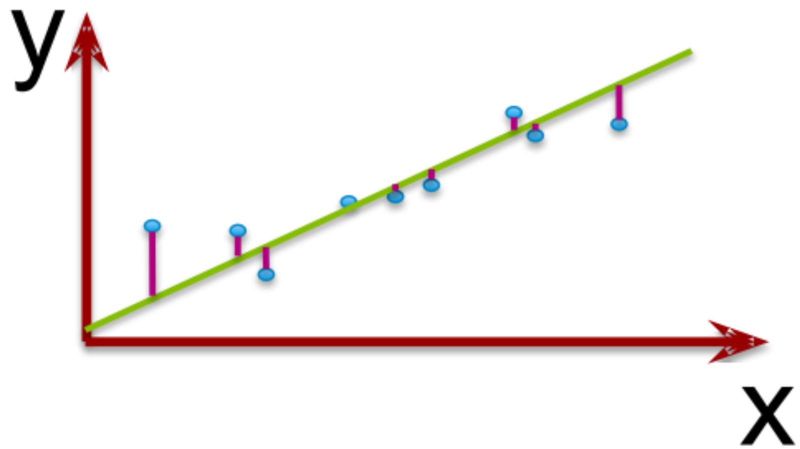
Training Loss vs Model Complexity
训练损失 vs 模型复杂度



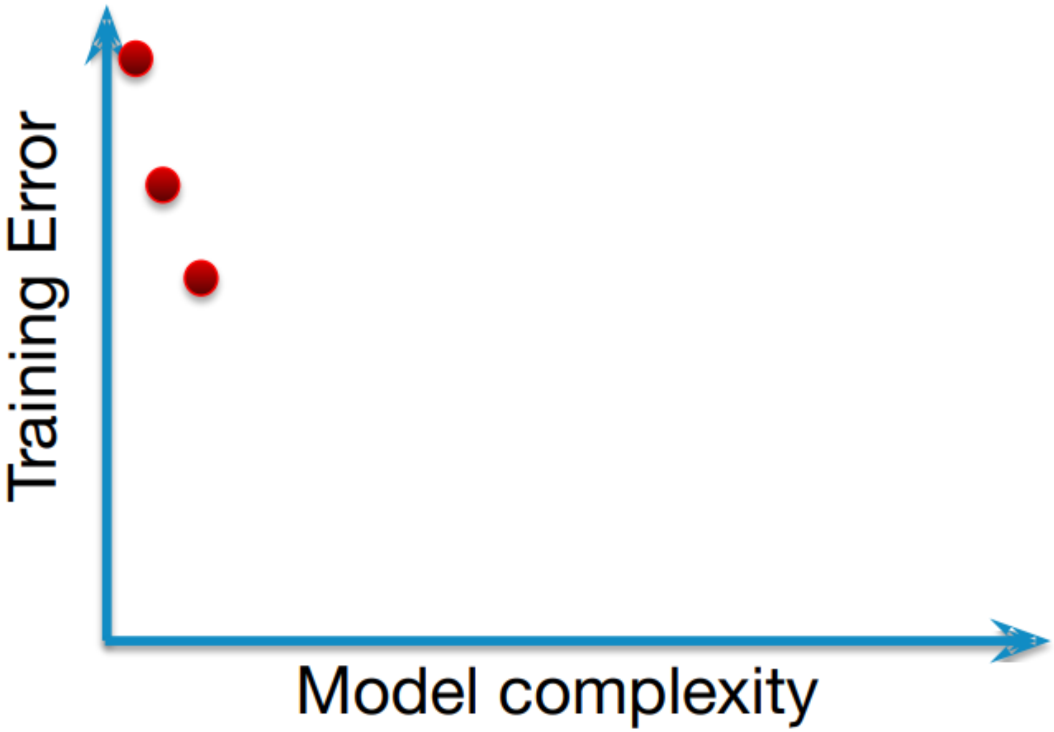


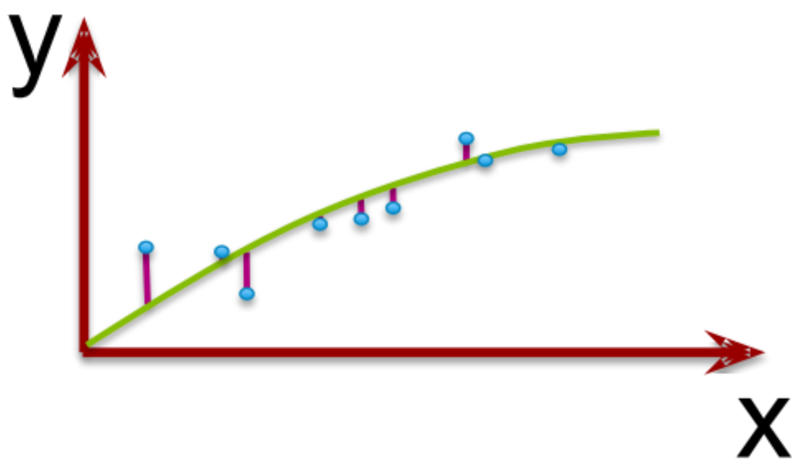
Training Loss vs Model Complexity
训练损失 vs 模型复杂度



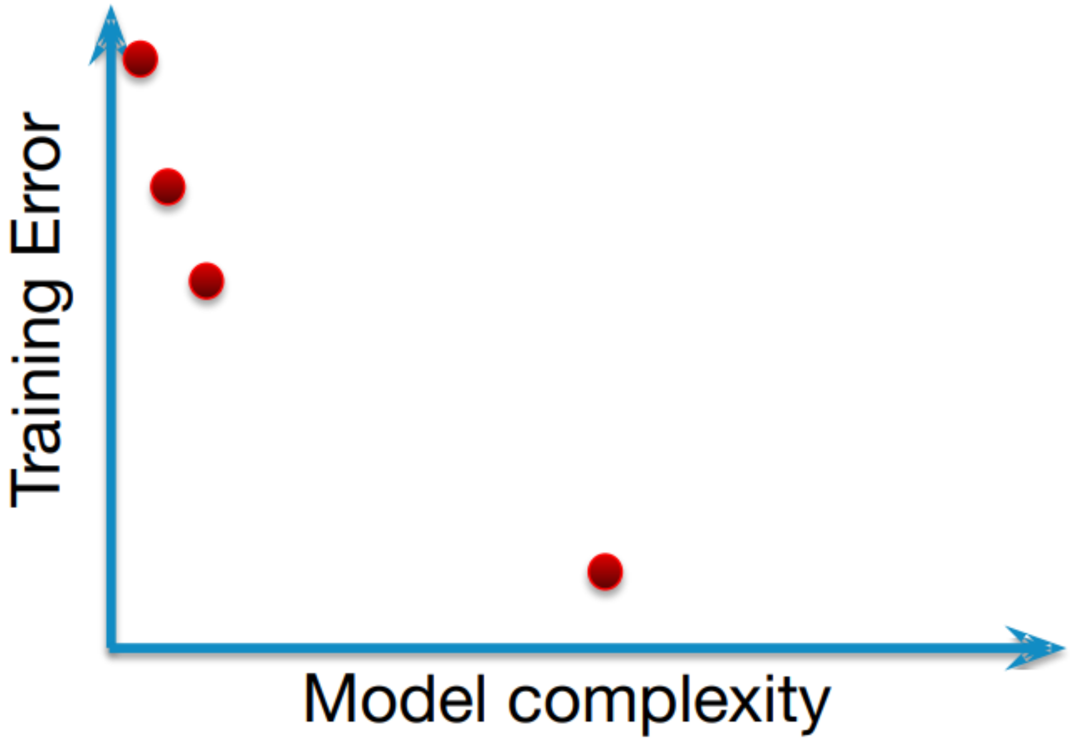


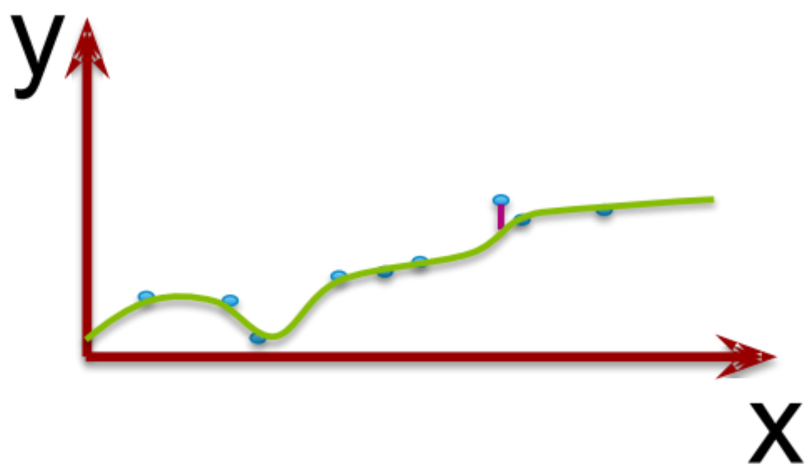
Training Loss vs Model Complexity
训练损失 vs 模型复杂度



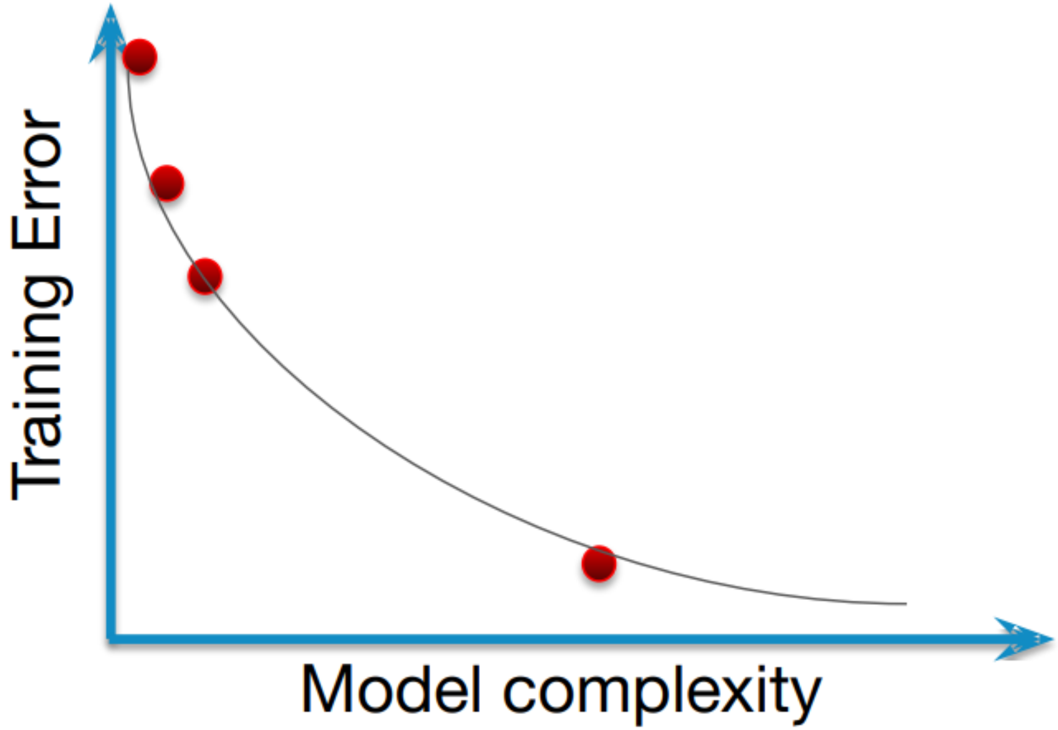


Training Loss vs Model Complexity
训练损失 vs 模型复杂度



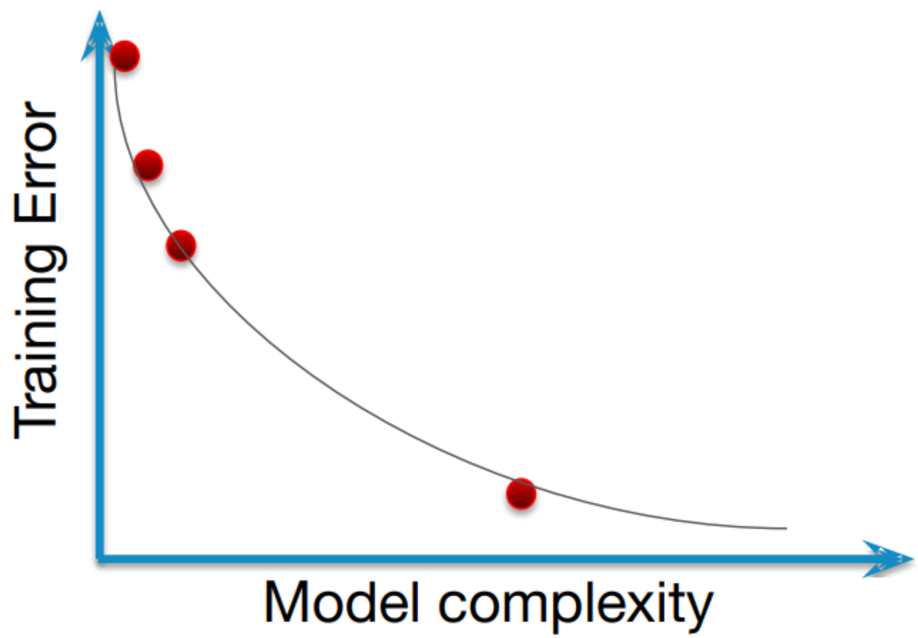


Training Loss vs Model Complexity
训练损失 vs 模型复杂度



Training Loss vs Model Complexity
训练损失 vs 模型复杂度

- Increase Model Complexity → Lower Training Error
  增加模型复杂度→降低训练误差
- **Is this a desirable thing**? 这值得吗？



Training Error → Predictive Performance?
训练误差 → 预测性能?

- Complex model → Low Training Error 复杂模型 → 低训练误差
- Select a new/unseen data point 选择一个新的/未见的数据点
- Make prediction 预测
- **Are you happy with the prediction?** 该预测合理吗?

- Training error is overly optimistic
  训练误差过于乐观
- Params (w) were fit to training data
  参数（w）是针对训练数据拟合的

- **Small training error ≠> Good predictions***
  小的训练误差 **≠>** 好的预测

*unless training data includes everything you will ever see
*除非训练数据包含了能看到的一切

Roadmap
课程安排

Training Loss → **Generalization Error** → Test Error → Bias-Variance

Generalization Error
泛化误差

- Ideally, want to estimate Loss over all possible data (x, y)
  理想情况下，想要估计所有可能的数据 (x, y) 上的损失
  - This would be the *true* error
    这应该是真实误差

- But not all (x, y) are equally likely
  每个数据点(x,y)出现的可能性是不同的

- Ideally, weigh the Loss at (x_i, y_i) by how likely (x_i, y_i) really are
  - Likelihood of (x_i, y_i): p(x_i, y_i). This is the joint distribution, p(x, y)
    联合概率分布

  - Alternatively, and more naturally, we look for p(x) and p(y | x)
    - How likely is x_i, i.e. p(x_i)? 贝叶斯定理
    - Given x_i, how likely is y_i, i.e. p(y_i | x_i)?

Generalization Error Definition
泛化误差定义

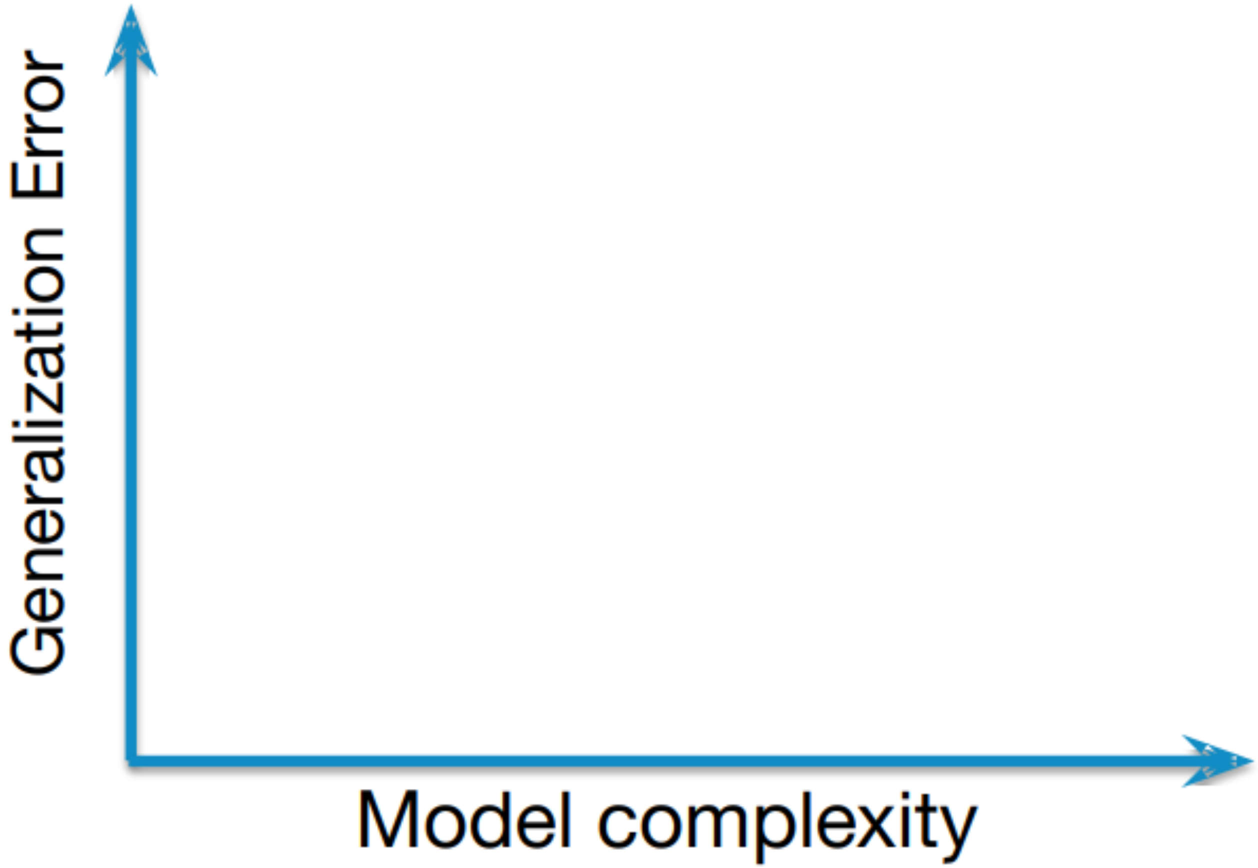
$$\text{Generalization Error} = \sum_{\forall (x_i, y_i) \in \mathcal{D}} L(y_i, f_{\hat{w}}(x_i)) p(x, y)$$
$$= \sum_{\forall (x_i, y_i) \in \mathcal{D}} L(y_i, f_{\hat{w}}(x_i)) p(x) p(y|x)$$

$$\text{Generalization Error} = \int_{(x,y) \in \mathcal{D}} L(y, f_{\hat{w}}(x)) p(x, y) dx dy$$
$$= \int_{(x,y) \in \mathcal{D}} L(y, f_{\hat{w}}(x)) p(x) p(y|x) dx dy$$

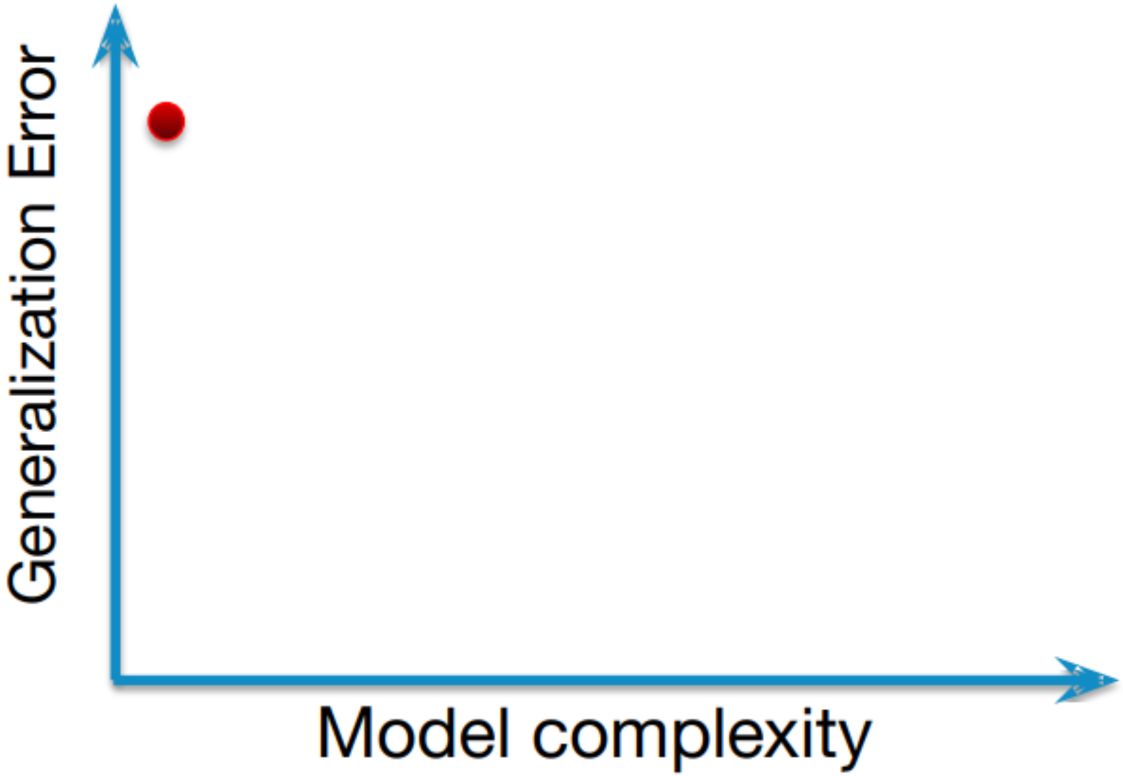
Generalization Error Definition
泛化误差定义

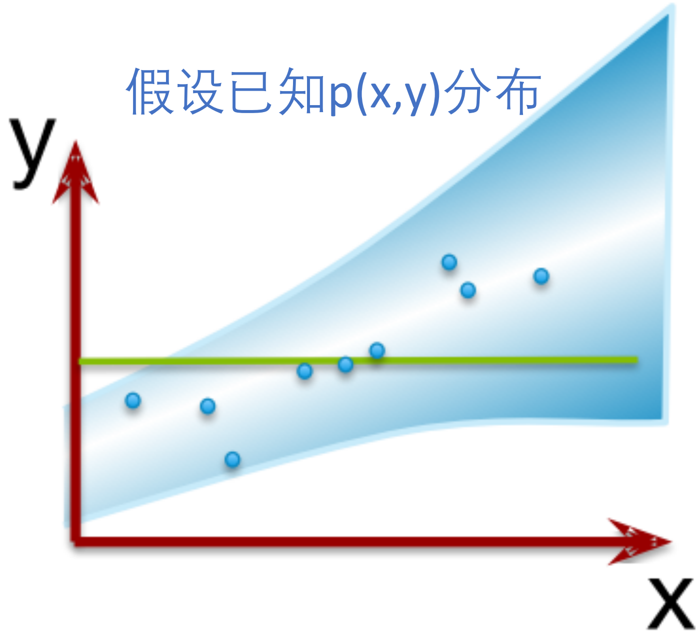
$$\text{Generalization Error} = \sum_{\forall (x_i, y_i) \in \mathcal{D}} L(y_i, f_{\hat{w}}(x_i))p(x, y)$$
$$= \mathbb{E}_{\mathrm{x},\mathrm{y}}\left[L(\mathrm{y}, f_{\hat{w}}(\mathrm{x}))\right]$$

Generalization Error vs Model Complexity
泛化误差 vs 模型复杂度

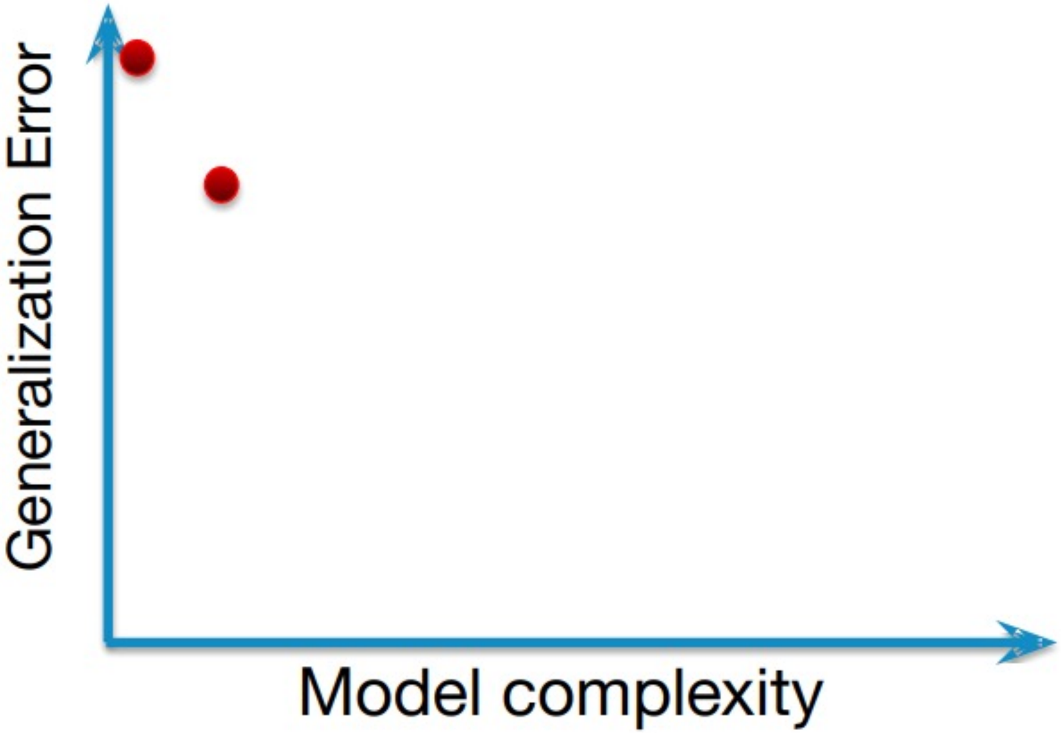


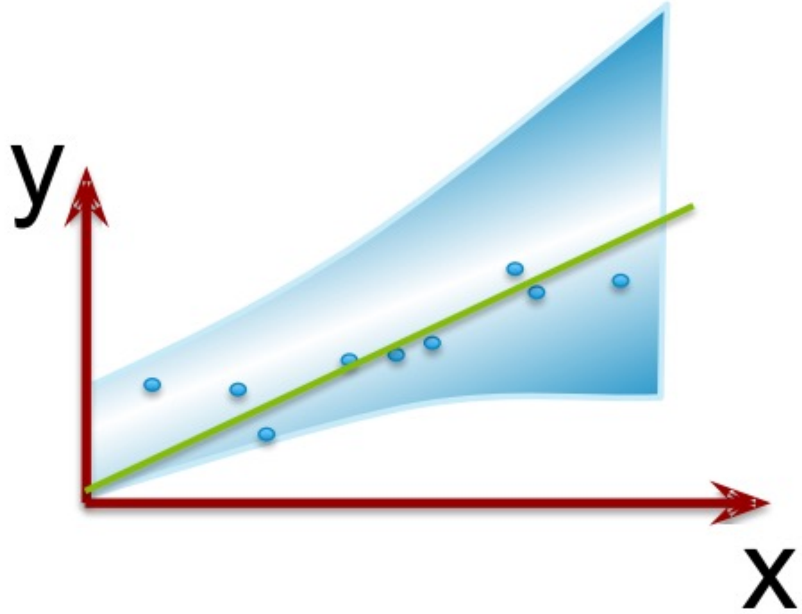
Generalization Error vs Model Complexity
泛化误差 vs 模型复杂度



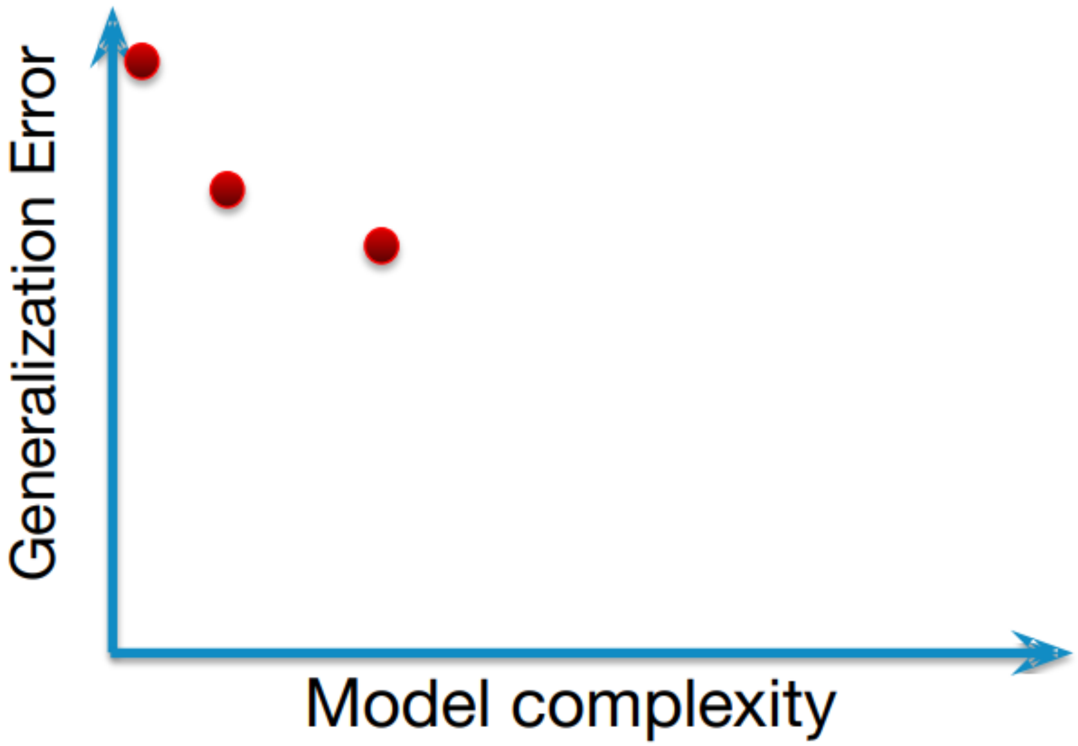


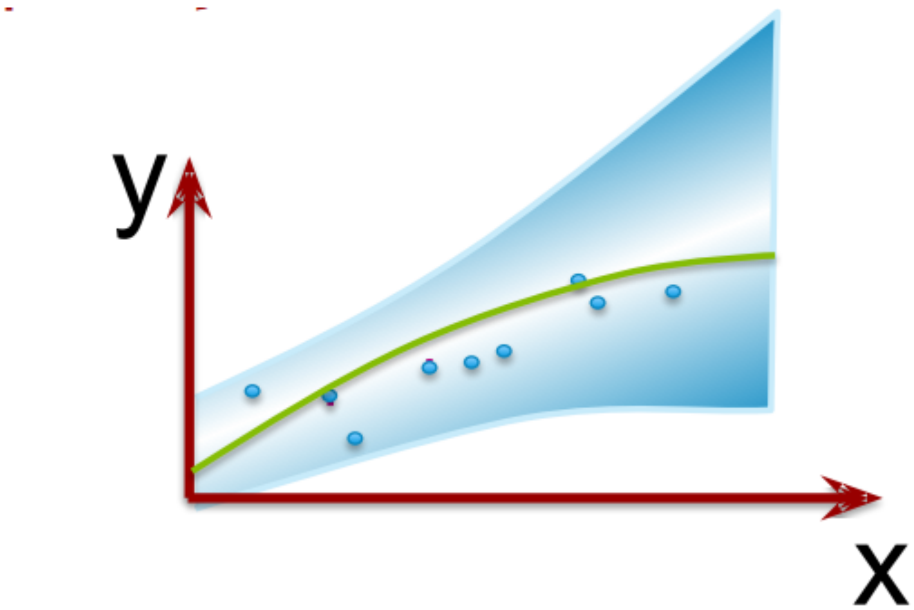
Generalization Error vs Model Complexity
泛化误差 vs 模型复杂度



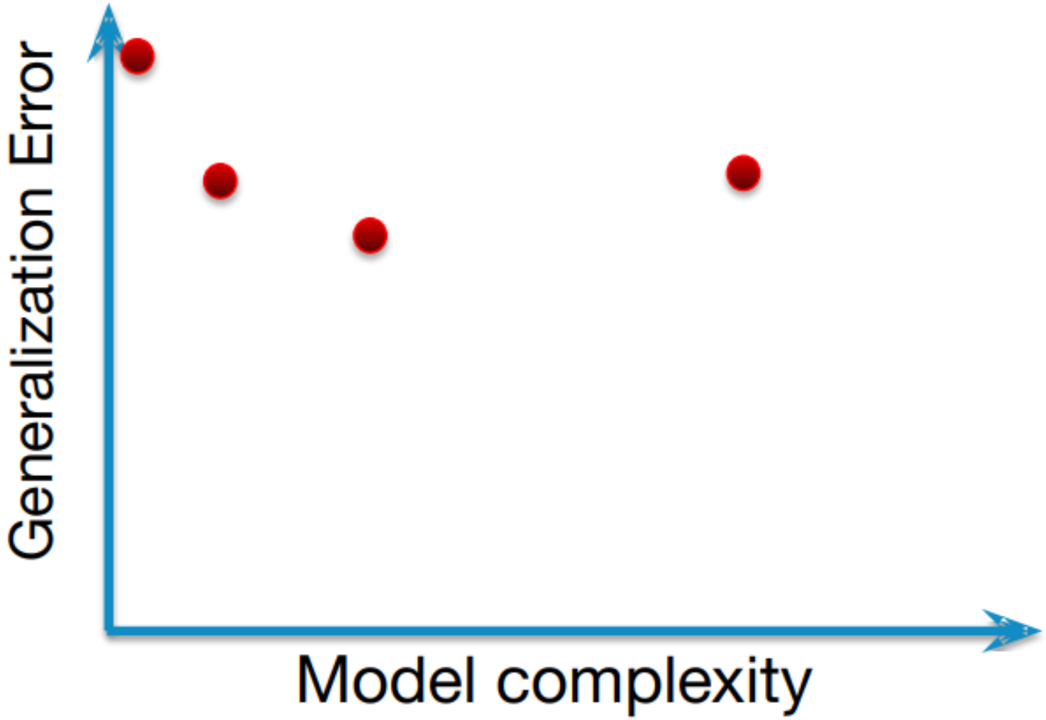


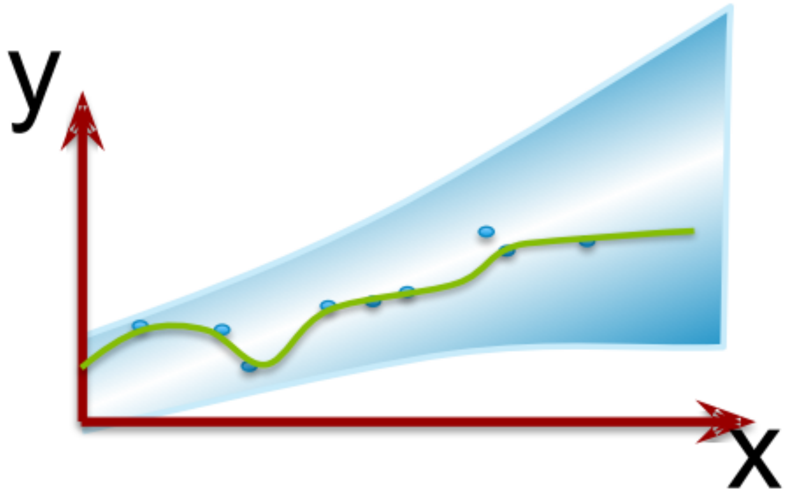
Generalization Error vs Model Complexity
泛化误差 vs 模型复杂度



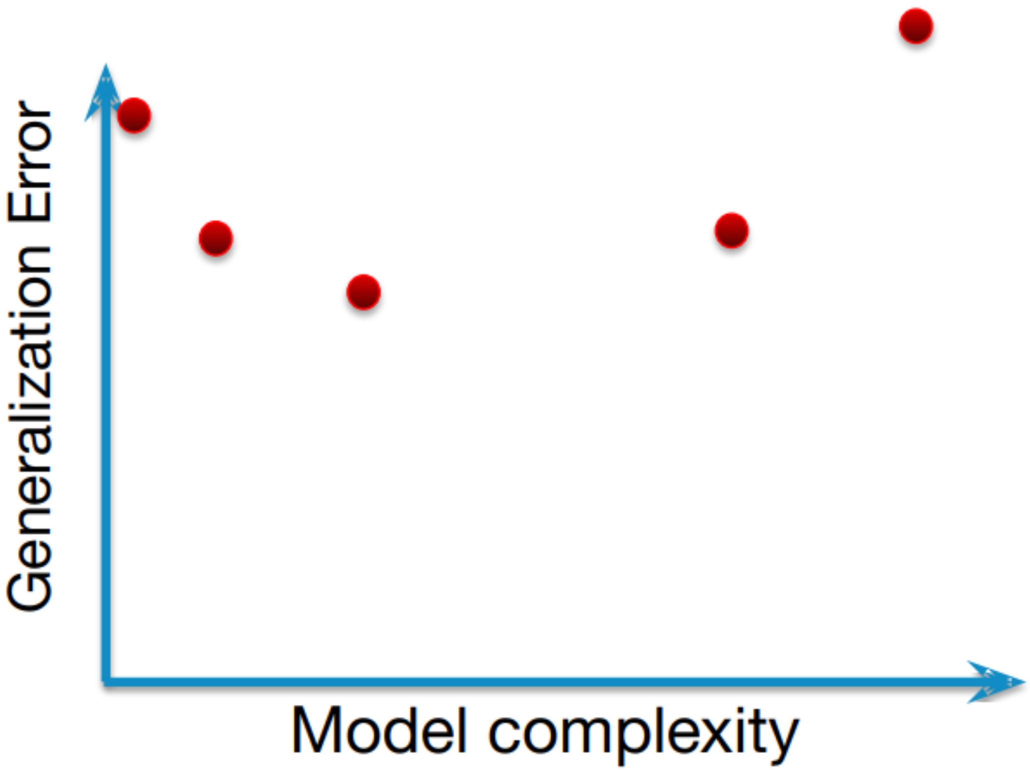


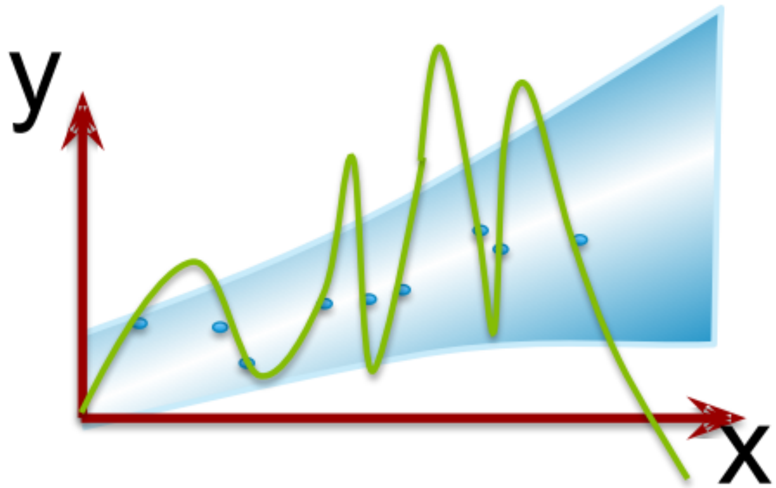
Generalization Error vs Model Complexity
泛化误差 vs 模型复杂度





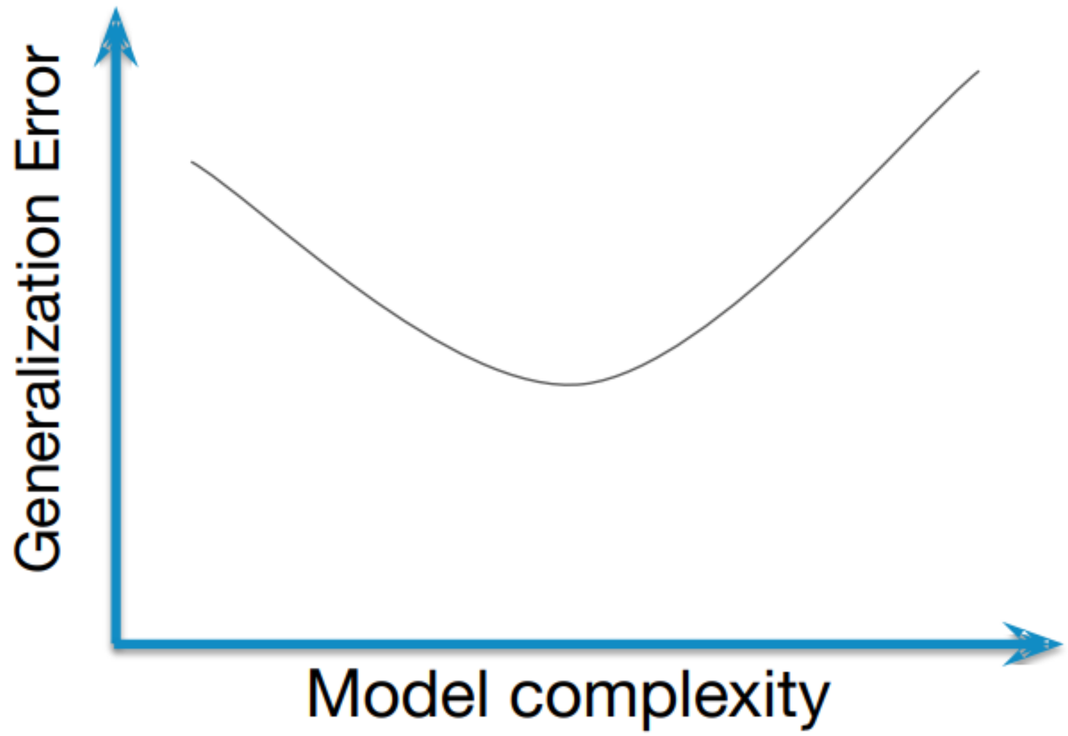
Generalization Error vs Model Complexity
泛化误差 vs 模型复杂度





Generalization Error vs Model Complexity
泛化误差 vs 模型复杂度

- Can't compute! 无法计算!
- We don't have access to the entire data and/or the distribution of data
  我们不太可能访问到所有数据和/或数据分布



Approximating Generalization Error
近似泛化误差

- Not possible to look at all (x, y) to compute generalization error
  不可能查看所有 (x, y) 来计算泛化误差

- Instead, approximate generalization error by looking at data which is not in the training set
  代替的方法是通过未包含在训练集中的数据近似估计泛化误差

- Split the available data into Train/Test
  将现有数据分为训练集/测试集
  - **Never, ever, train your models on test data!**
    永远不要在测试集上训练!
- Test Set becomes a Proxy for “everything one might see”
  测试集成为“你可能看到的一切”的代理

Roadmap
课程安排

Training Loss → Generalization Error → Test Error → Bias-Variance

Test Error Definition
测试误差定义

- Test Error = Avg Loss on the Test Set
  测试误差 = 测试集上的平均损失

Estimated on the Training Dat

$$\text{Test Error} = \frac{1}{|\mathcal{D}_{\text{Test}}|} \sum_{\forall (x_i, y_i) \in \mathcal{D}_{\text{Test}}} L(y_i, f_{\hat{w}}(x_i))$$

Approximate

$$\boxed{\text{Generalization Error} = \int_{(x,y)\in\mathcal{D}} L(y, f_{\hat{w}}(x)) p(x,y) dx dy}$$

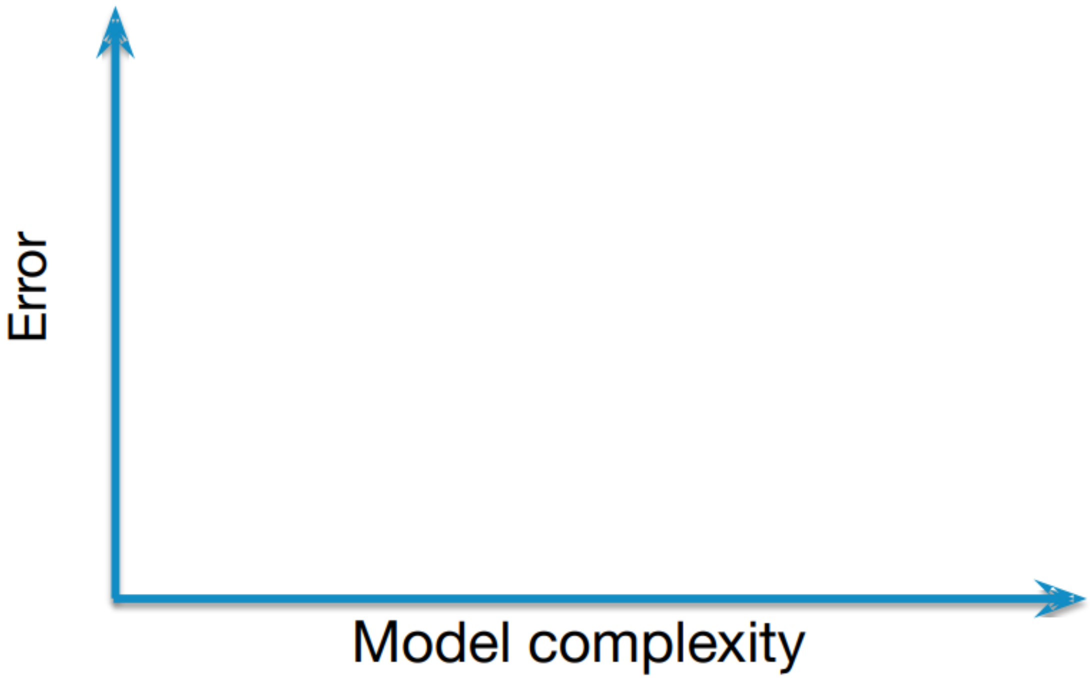
Training → Test
训练 → 测试

- Split data to hold out “test” data 将数据分割，保留“测试”数据
  - In practice, Train/Test split: 70%/30%, 80%/20%
    实践中，训练集/测试集的比例：70%/30%，80%/20%
  - What if 20%/80%?
    若是20%/80%呢？

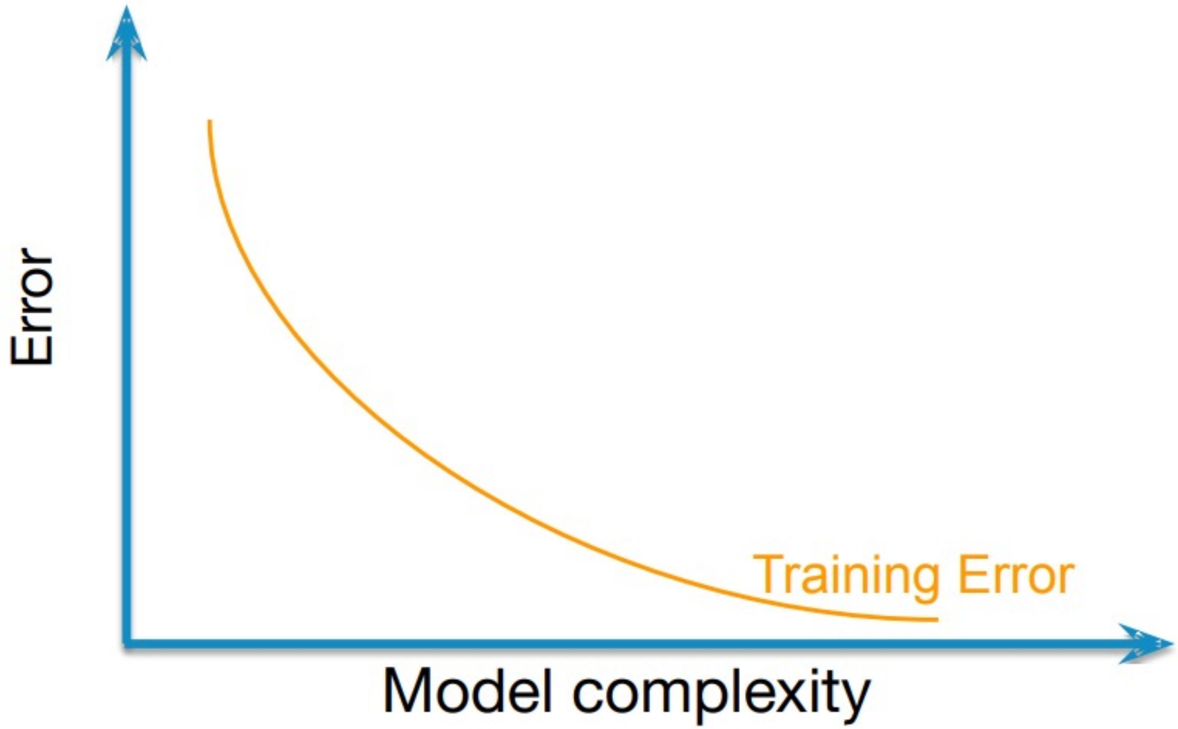
- Train the model (estimate the params) using only the training data
  仅用训练集训练数据（估计参数）
- Split the available data into Train/Test
  将现有数据分为训练集/测试集
  - **Never, ever, train your models on test data!**
    永远不要在测试集上训练！
  - Done by minimizing the Training Loss
    通过最小化训练损失完成

- Test Set becomes a Proxy for “everything one might see”
  通过计算测试数据上的测试损失测试“性能”

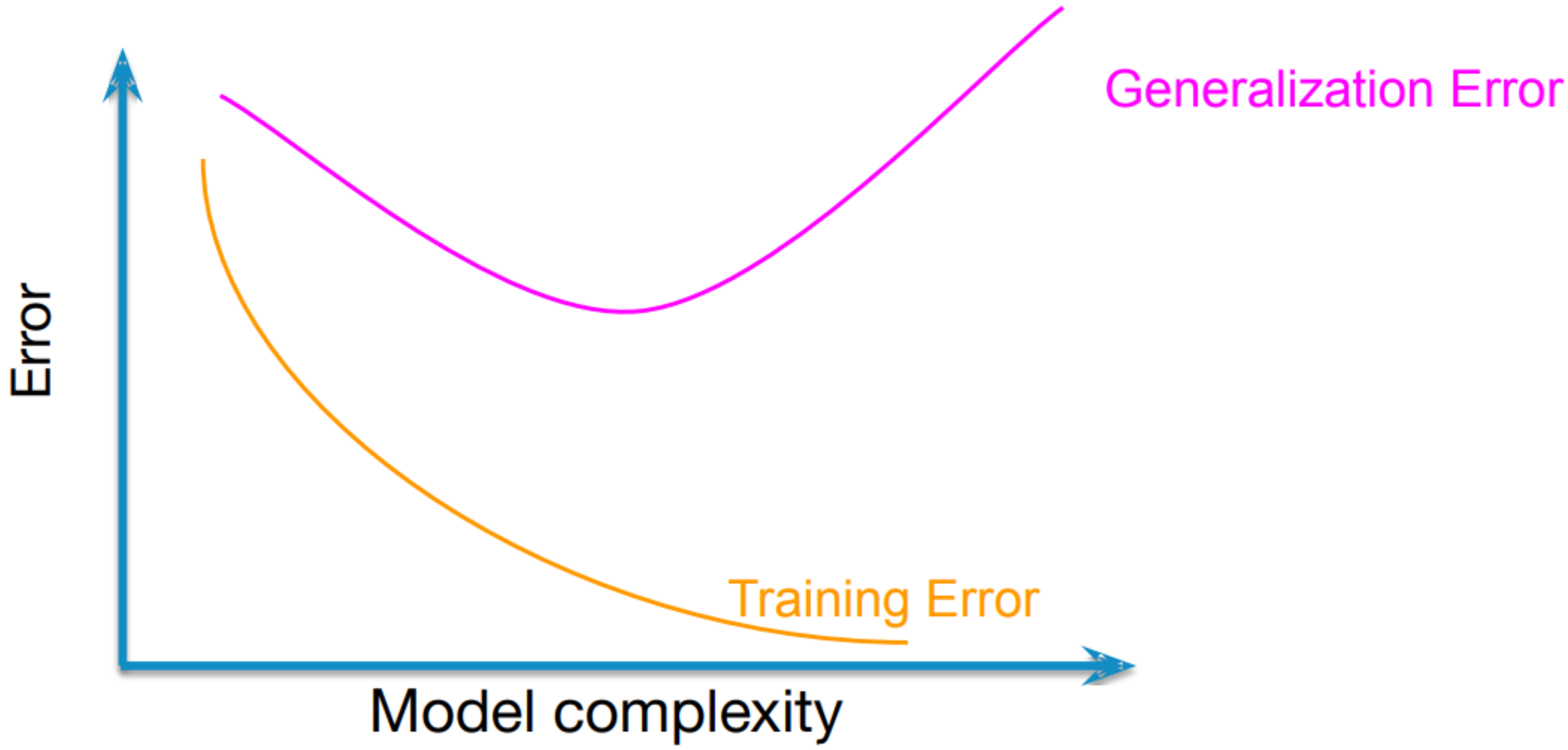
Train, Generalization, Test Error vs Model Complexity
训练、泛化、测试误差与模型复杂度的关系



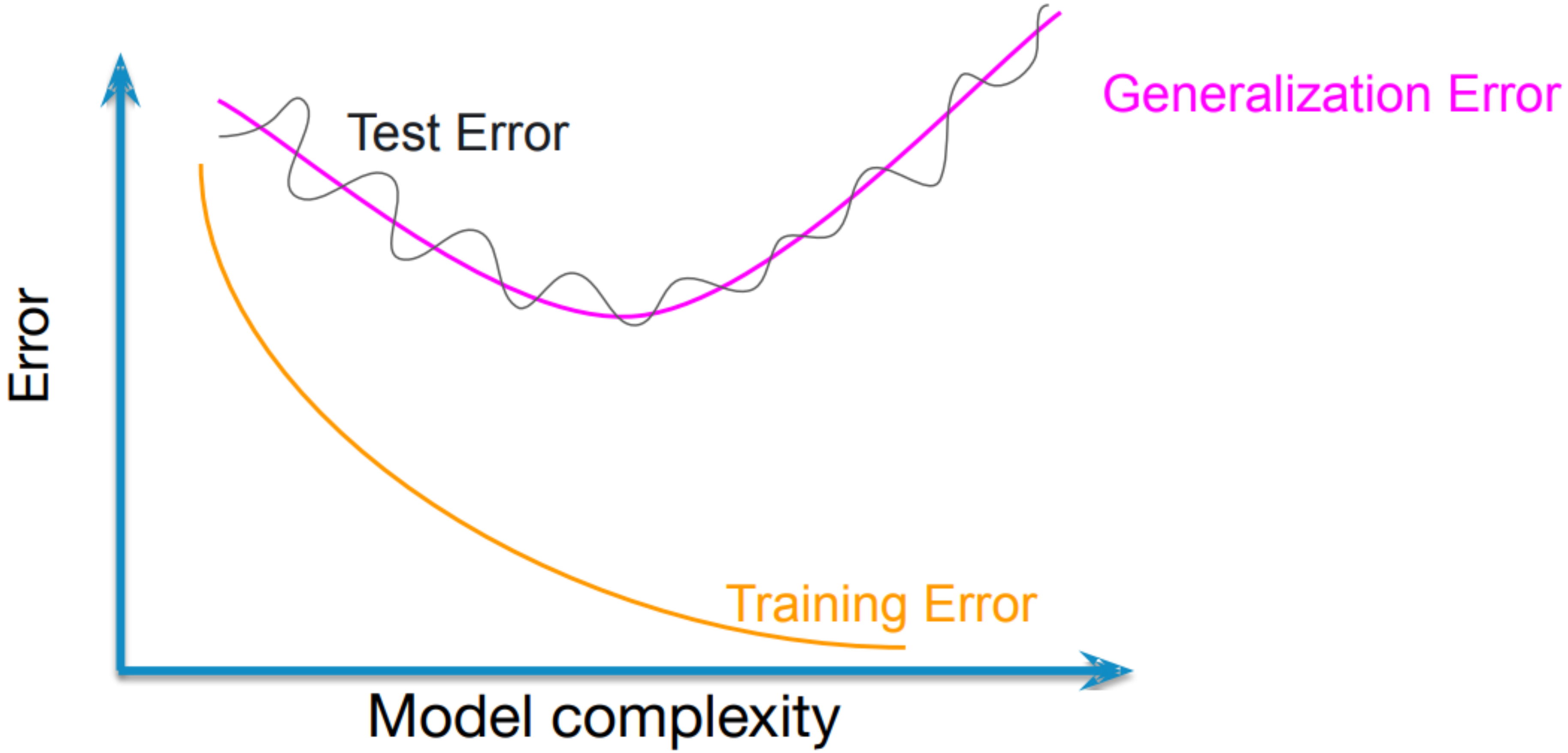
Train, Generalization, Test Error vs Model Complexity
训练、泛化、测试误差与模型复杂度的关系



Train, Generalization, Test Error vs Model Complexity
训练、泛化、测试误差与模型复杂度的关系



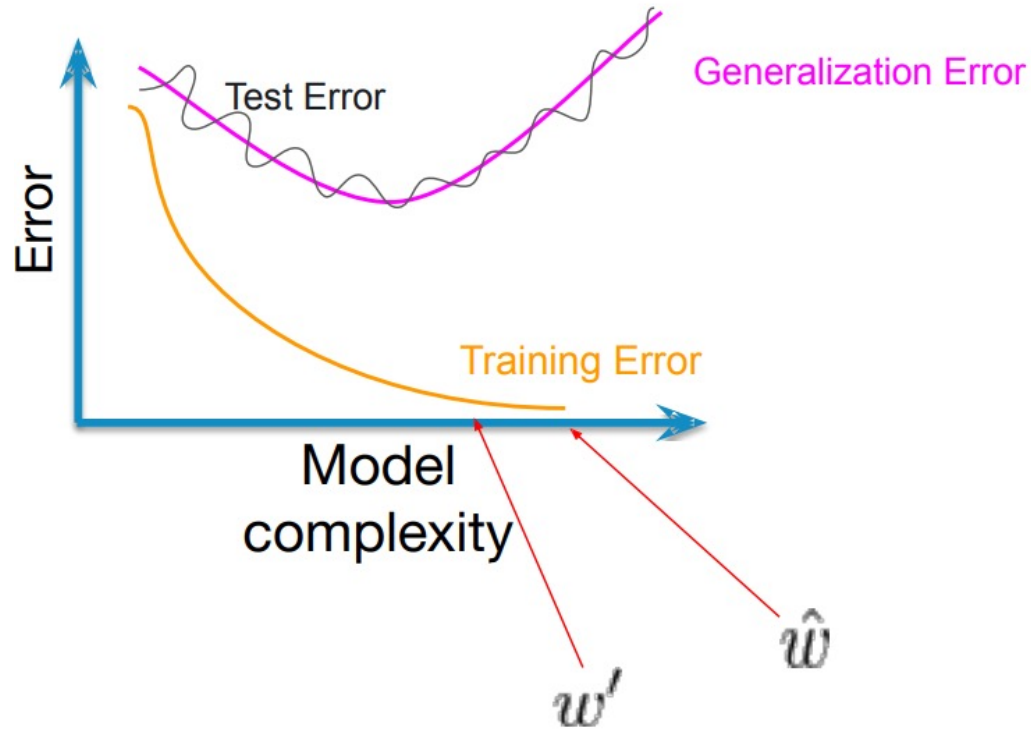
Train, Generalization, Test Error vs Model Complexity
训练、泛化、测试误差与模型复杂度的关系



Overfitting
过拟合

- Using $\widehat{\omega}$
  使用 $\widehat{\omega}$

- if there exists a model with estimated params, ω' such that:
  如果存在一个模型，其估计的参数 ω' 使得:



$$1.\text{Training Error}(\hat{w}) < \text{Training Error}(w')$$

$$2.\text{Generalization Error}(\hat{w}) > \text{Generalization Error}(w')$$

Roadmap
课程安排

Training Loss → **Generalization Error** → **Test Error** → **Bias-Variance**

Sources of Error
误差的来源

- Noise 噪声
- Bias 偏差
- Variance 方差

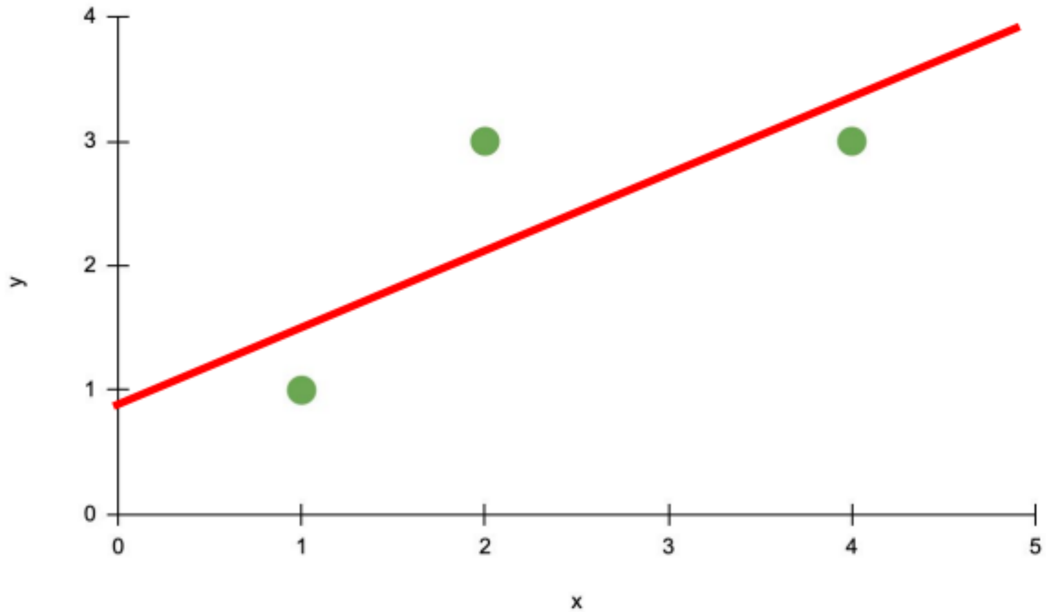
$$\text{Expected Prediction Error}(x) = \text{Noise} + \text{Bias}^2 + \text{Variance}$$

$$\text{MSE}(x) = \sigma^2 + \text{Bias}^2 + \text{Variance}$$

Noise
噪声

- Data is inherently noisy 数据本质上是有噪声的
- Irreducible error 无法减少的误差
- Regression Model 回归模型:

$$y_i = f(x_i) + \epsilon_i$$
$$\mathbb{E}[\epsilon_i] = 0$$



- Even if we estimate f exactly, there will still be some error (noise).
  即使我们精确地估计了 f，仍然会有一些误差（噪声）

Setup
设置

- We create N different training sets by sampling
  我们通过抽样创建 N 个不同的训练集

- Each training set produces estimated model params
  每个训练集会产生估计的模型参数
- Use the “average” predictions of all the N estimated models, denoted by: $f_{\bar{w}}$
  使用所有 N 个估计模型的“平均”预测，记为： $f_{\bar{w}}$
- The “average” fit is akin to an expected fit
  “平均”拟合类似于预期拟合

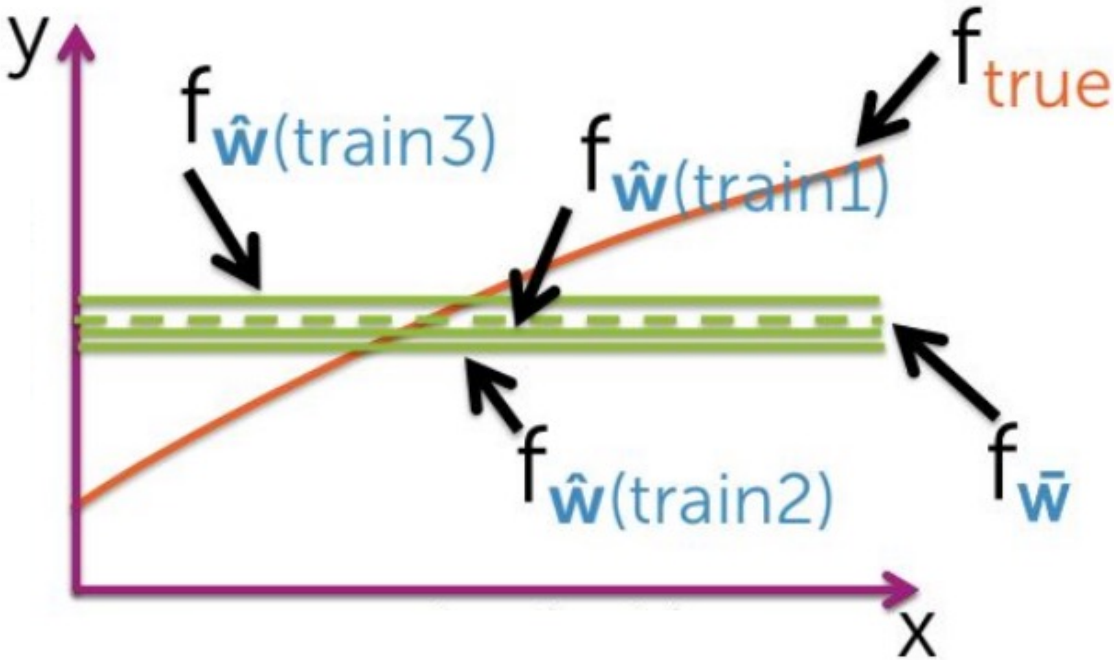
Bias
偏差

$$\mathrm{Bias}(x) = f_{\mathrm{true}}(x) - f_{\bar{\mathrm{w}}}(x)$$

- Is our approach flexible enough to capture $f_{true}(x)$
  我们的方法是否足够灵活以捕捉$f_{true}(x)$
- Bias is high when the hypothesis class is unable to capture $f_{true}(x)$
  当假设类无法捕捉 $f_{true}(x)$时，偏差很高

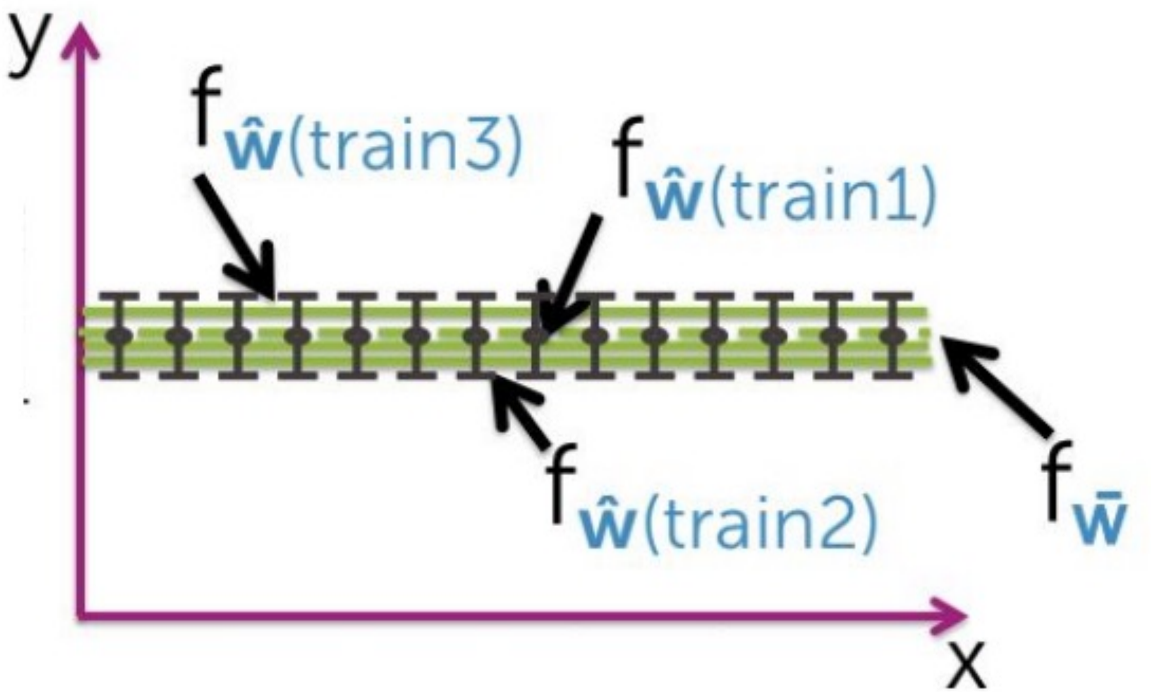
Low Complexity → High Bias
低复杂度 → 高偏差

- Constant Function has Low Model Complexity
  常量函数具有低模型复杂度
- Usually not capable of capturing the true relationship
  通常无法捕捉真实关系
- Low complexity models lead to High Bias
  低复杂度模型导致高偏差



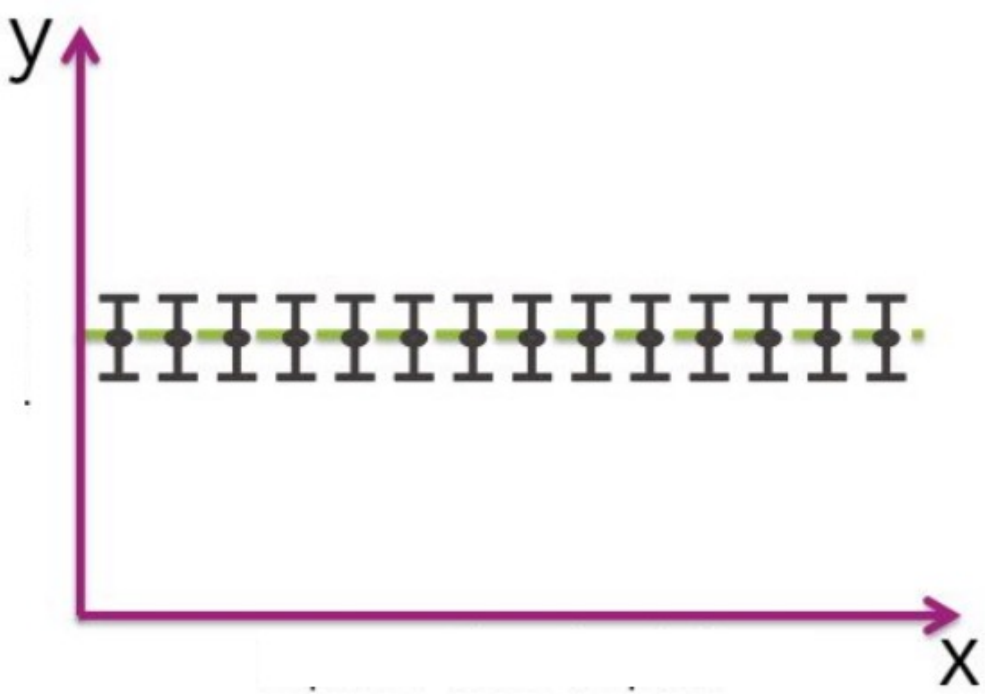
Variance
方差

- How much do specific fits (from among the N fits) vary from the “average” fit?
  在 N 个拟合中，具体拟合与“平均”拟合之间的差异有多大？



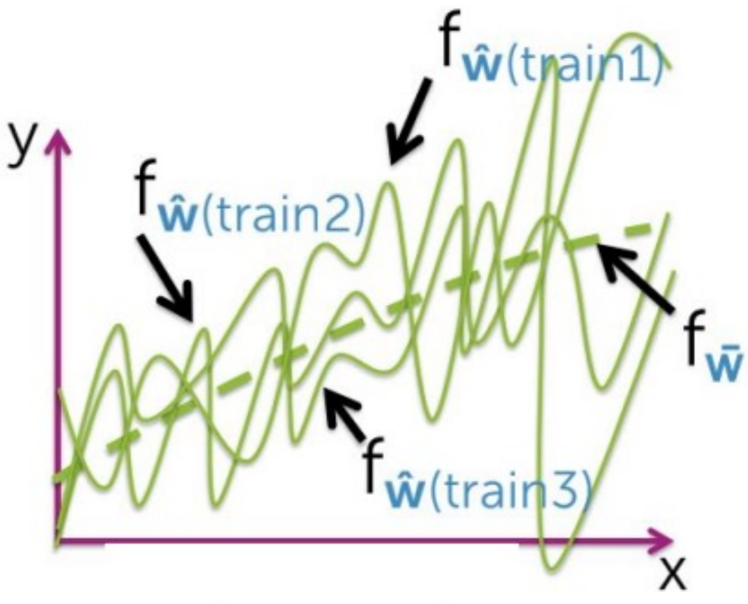
Low Complexity → Low Variance
低复杂度 → 低方差

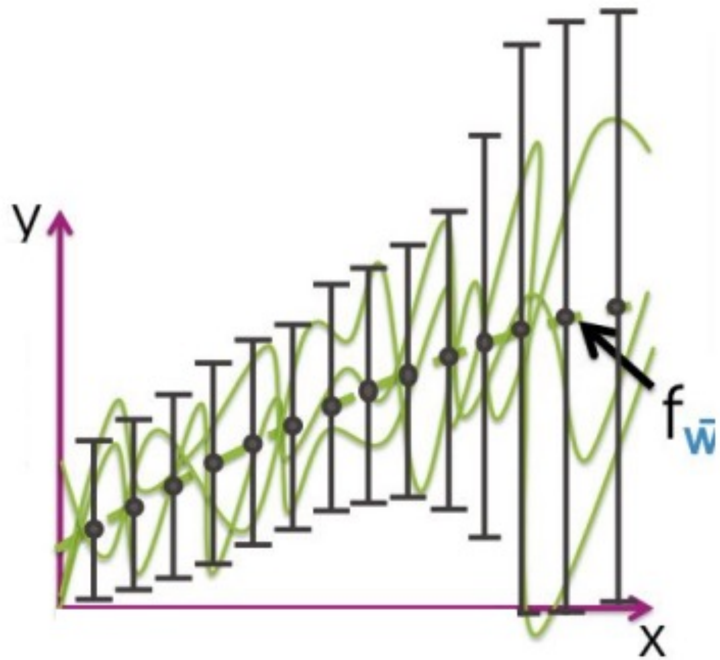
- In a low complexity setting, specific fits do not vary widely 在低复杂度设置中，具体拟合的变化不大
- This means, Low Complexity models have Low Variance 这意味着，低复杂度模型具有低方差



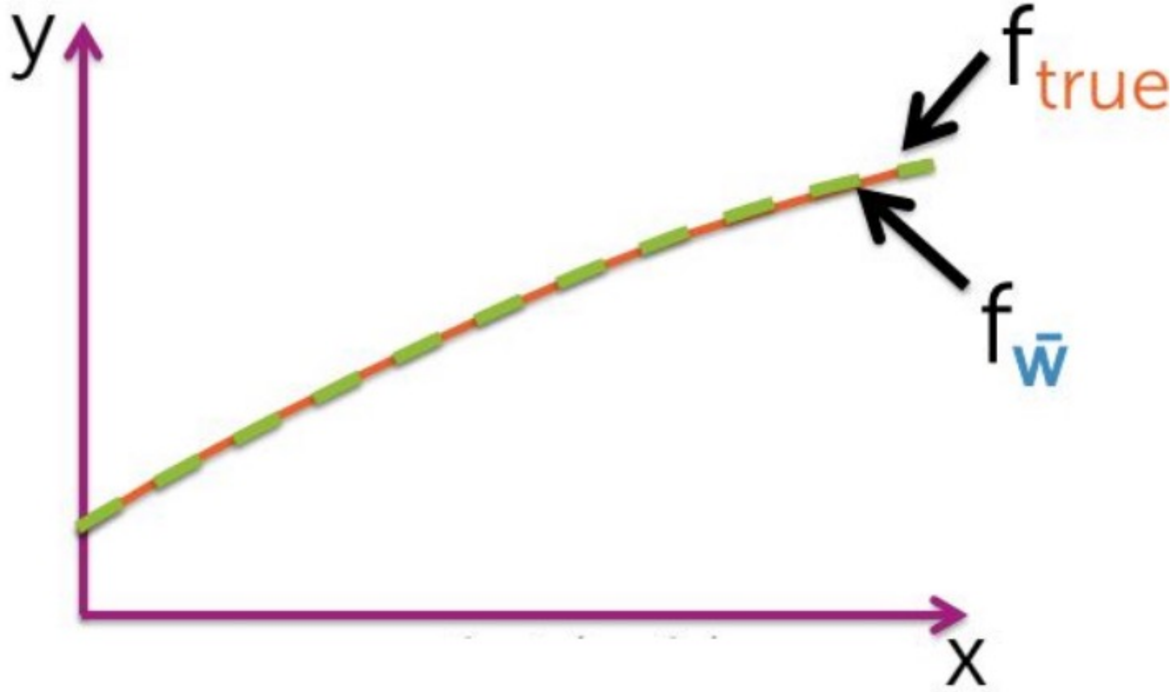
High Complexity → High Variance
高复杂度 → 高方差

- In a high complexity setting, specific fits do vary widely
  在高复杂度设置中，具体拟合的变化很大
- This means, High Complexity models have High Variance 这意味着，高复杂度模型具有高方差

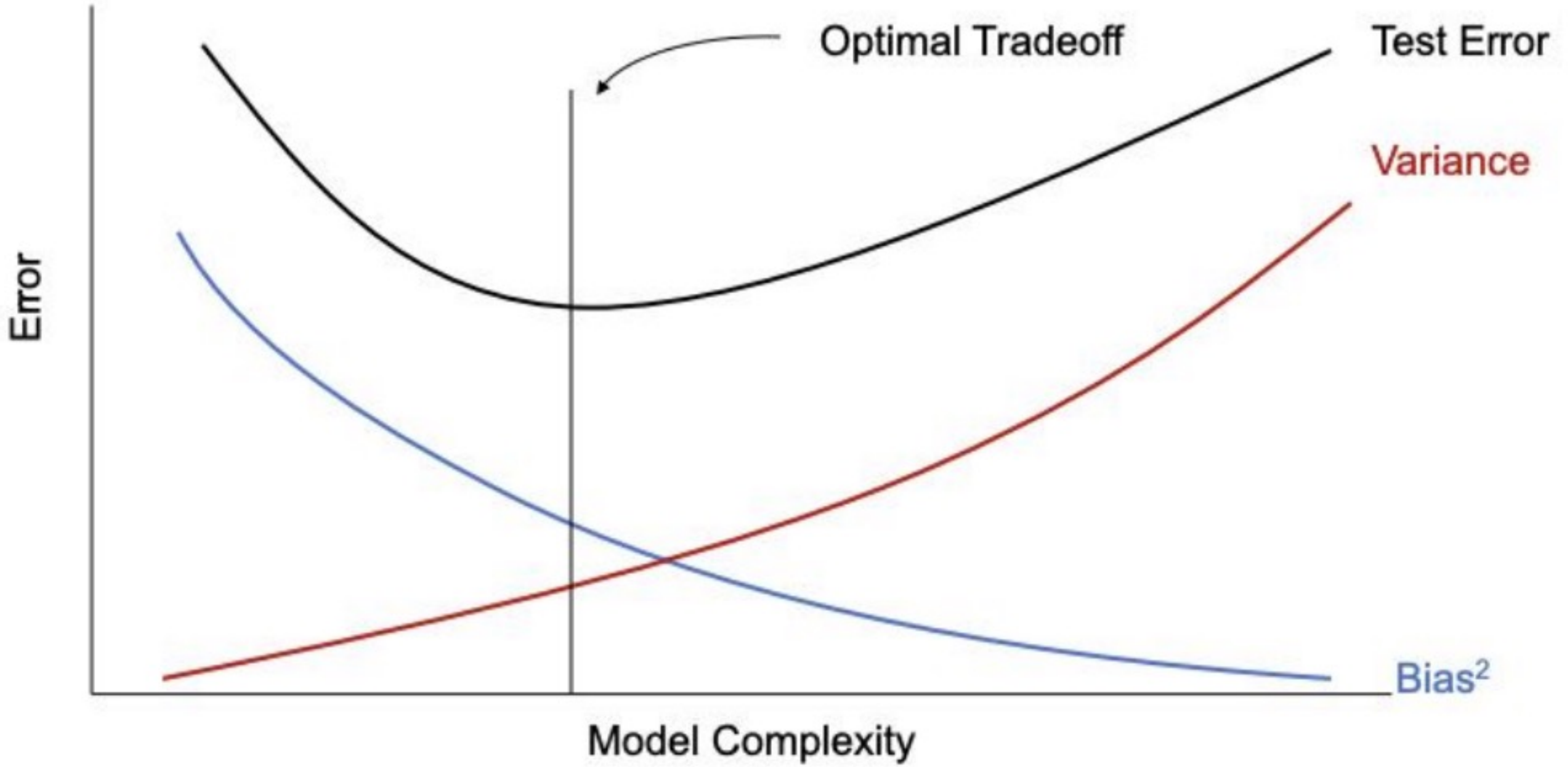




High Complexity → Low Bias
高复杂度 → 低偏差



Bias-Variance Trade-off
偏差-方差权衡



Why 3 sources of error?
为什么有三种来源?

Expected prediction error

$= E_{train}$ [generalization error of $\hat{\mathbf{w}}$(train)]

$= E_{train}\left[E_{\mathbf{x},y}\left[L(y, f_{\hat{\mathbf{w}}(train)}(\mathbf{x}))\right]\right]$

1. Look at specific $\mathbf{x}$
2. Consider $L(y, f_{\hat{\mathbf{w}}}(\mathbf{x})) = (y - f_{\hat{\mathbf{w}}}(\mathbf{x}))^2$

Expected prediction error at $\mathbf{x}$

$= E_{train}\left[E_{y|\mathbf{x}}\left[(y - f_{\hat{\mathbf{w}}(train)}(\mathbf{x}))^2\right]\right]$

Why 3 sources of error?
为什么有三种来源?

Expected prediction error at **x**

$$= E_{train}\left[E_{y|\mathbf{x}}\left[(y - f_{\hat{\mathbf{w}}(train)}(\mathbf{x}))^2\right]\right]$$

$$= E_{train}\left[E_{y|\mathbf{x}}\left[(\overbrace{(y - f_{true}(\mathbf{x}))}^{a} + \overbrace{(f_{true}(\mathbf{x}) - f_{\hat{\mathbf{w}}(train)}(\mathbf{x}))}^{b})^2\right]\right]$$

$$= \cancel{E_{train}}\left[\underbrace{E_{y|x}[(\underbrace{y-f}_{\epsilon^2})^2]}_{=\sigma^2 \text{ by defn}}\right] + 2\,E_{train}\left[\underbrace{E_{y|x}[(\underbrace{y-f}_{\epsilon})(f-\hat{f})]}_{E[\epsilon]E[f-\hat{f}]=0}\right]$$

$$+ \underbrace{E_{train}[\cancel{E_{y|x}}[(f-\hat{f})^2]]}_{\triangleq MSE(\hat{f}) \text{ mean square error}}$$

$$= \sigma^2 + MSE(\hat{f})$$

Shorthand:
$f_{true} \rightarrow f$
$f_{\hat{w}(train)} \rightarrow \hat{f}$

$E[(a+b)^2]$
$= E[a^2] +$
$2E[ab]$
$+ E[b^2]$

$E[ab] = E[a]E[b]$
if a, b
uncorr
(or, ind.)

Why 3 sources of error?
为什么有三种来源?

$$\mathrm{MSE}(\mathbf{x})$$

$$= E_{\text{train}}\Big[\big(\overbrace{f_{\text{true}}(\mathbf{x})}^{f} - \overbrace{f_{\hat{\mathbf{w}}(\text{train})}(\mathbf{x})}^{\hat{f}}\big)^2\Big]$$

$$= E_{\text{train}}\Big[\big((f_{\text{true}}(\mathbf{x}) - f_{\bar{\mathbf{w}}}(\mathbf{x})) + (\overbrace{f_{\bar{\mathbf{w}}}(\mathbf{x})}^{\bar{f}} - f_{\hat{\mathbf{w}}(\text{train})}(\mathbf{x}))\big)^2\Big]$$

$$= \underbrace{\cancel{E_{\text{train}}}[(f-\bar{f})^2]}_{=\text{bias}^2(\hat{f}) \text{ by defn}} + \underbrace{2E_{\text{train}}[\underbrace{(f-\bar{f})}_{\text{not fcn of "train"}}(\bar{f}-\hat{f})]}_{2(f-\bar{f})\underbrace{E_{\text{train}}[\bar{f}-\hat{f}]}_{=\bar{f}-E[\hat{f}]\,=0 \text{ by defn}}} + \underbrace{E_{\text{train}}[(\bar{f}-\hat{f})^2]}_{\text{var}(\hat{f})}$$

$\bar{f} = E_{\text{train}}[\hat{f}]$

its mean

random var.

$$= \text{bias}^2(\hat{f}) + \text{var}(\hat{f})$$

Why 3 sources of error?
为什么有三种来源?

Expected prediction error at $\mathbf{x}$

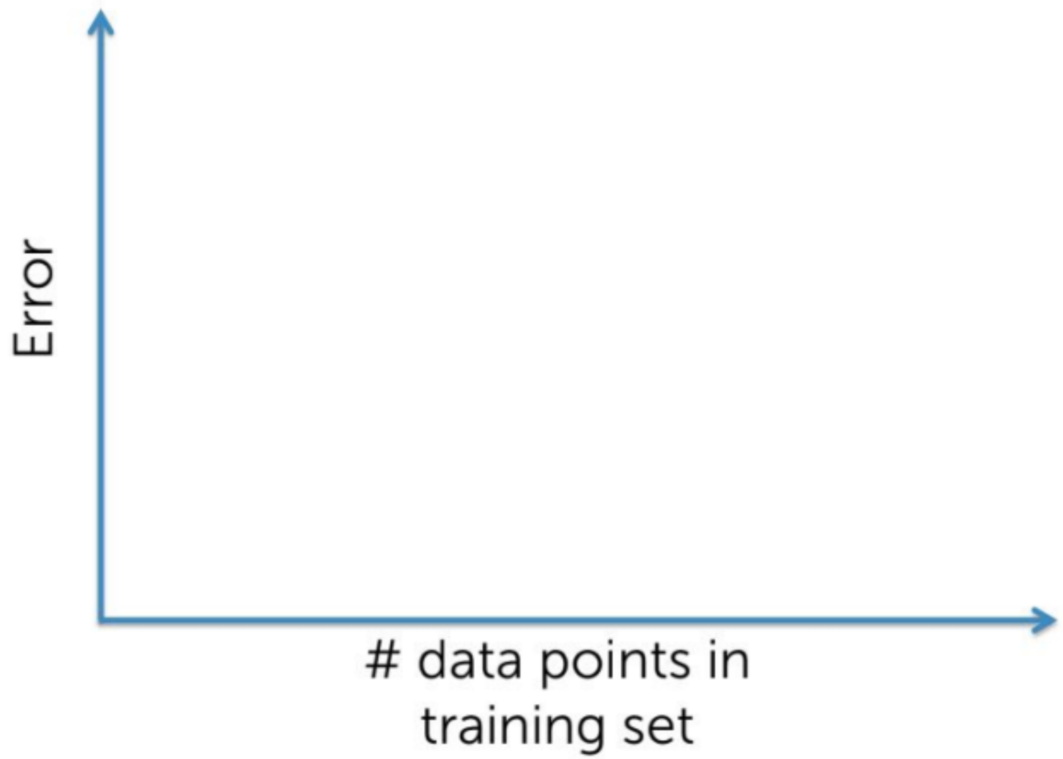
$$= \sigma^2 + \text{MSE}(f_{\hat{\mathbf{w}}}(\mathbf{x}))$$

$$= \sigma^2 + [\text{bias}(f_{\hat{\mathbf{w}}}(\mathbf{x}))]^2 + \text{var}(f_{\hat{\mathbf{w}}}(\mathbf{x}))$$

3 sources of error

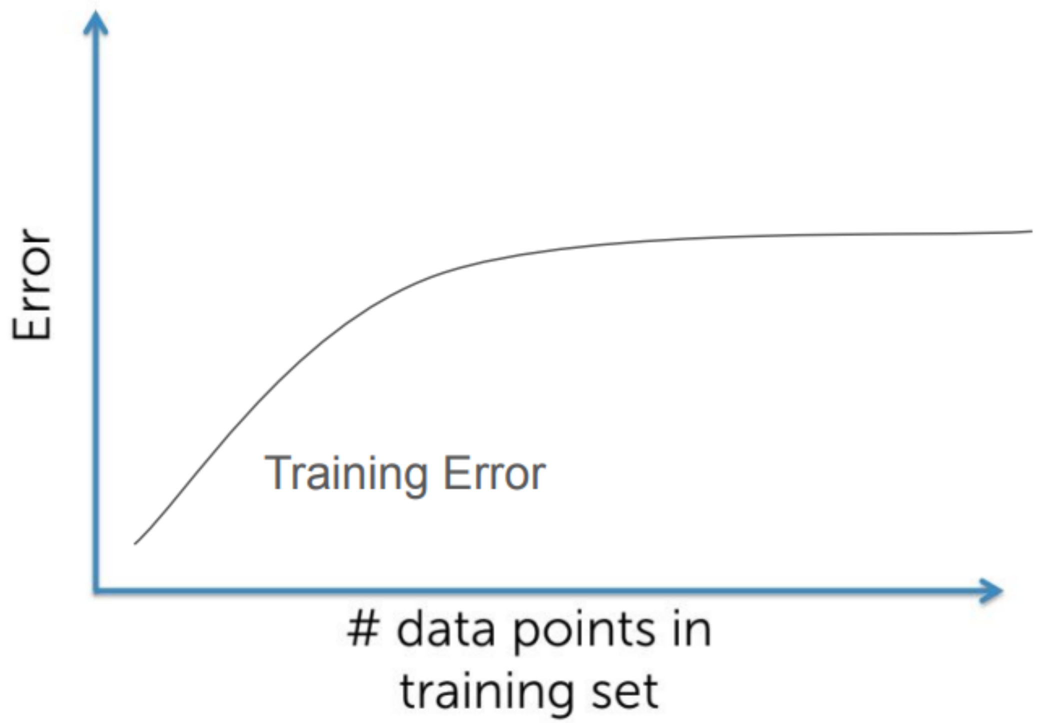
Error vs Amount of (Training) Data
误差 vs（训练）数据量的关系

Fixed Model Complexity



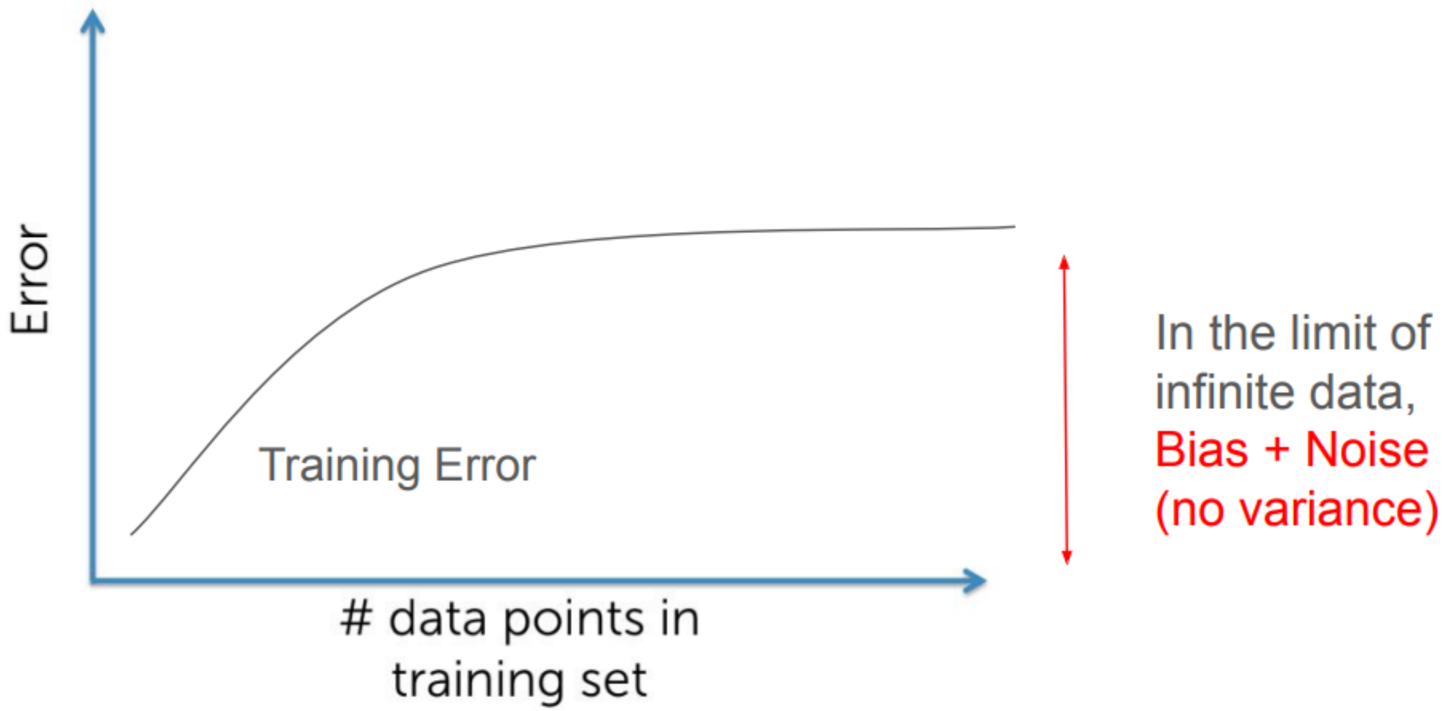
Error vs Amount of (Training) Data
误差 vs（训练）数据量的关系

Fixed Model Complexity

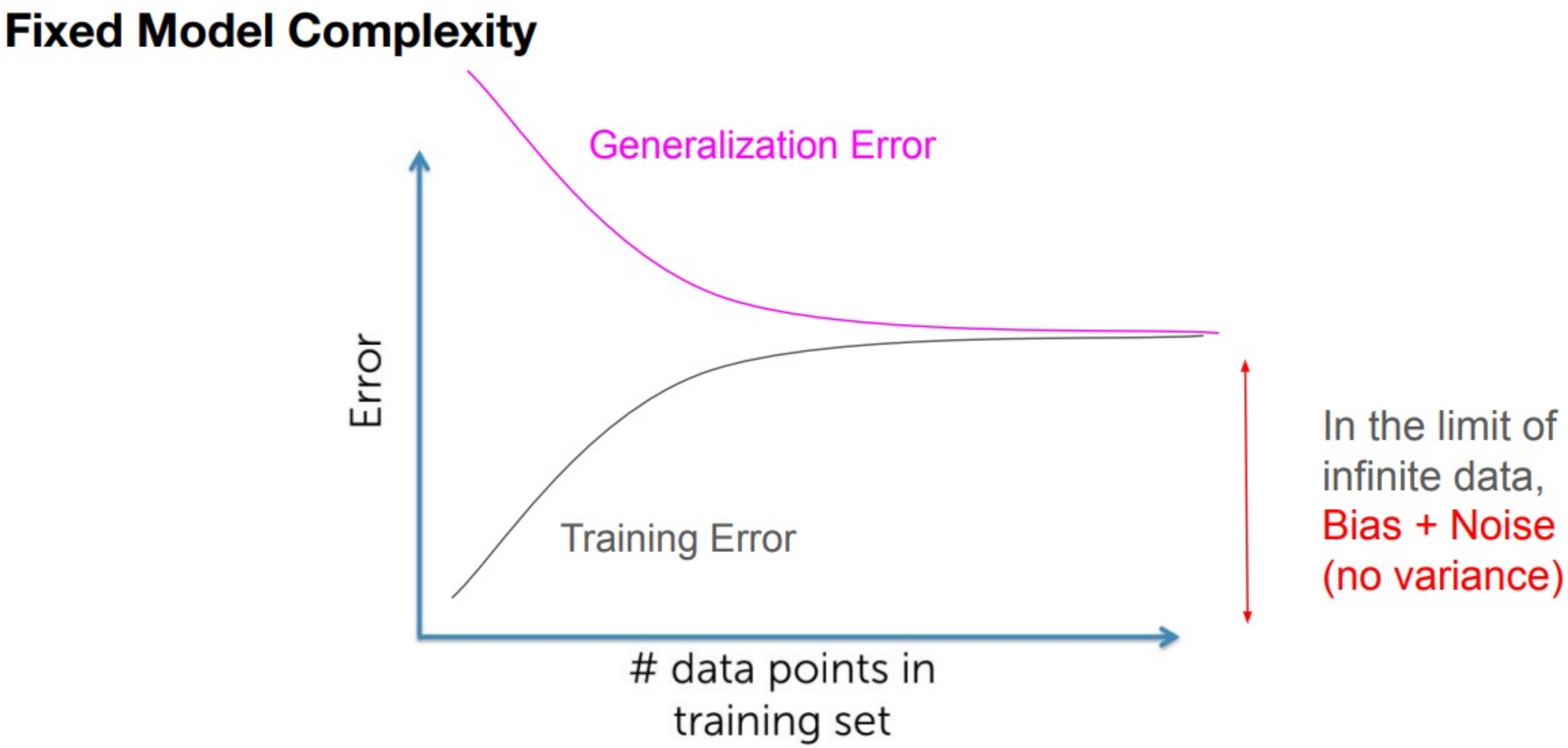


Error vs Amount of (Training) Data
误差 vs（训练）数据量的关系

Fixed Model Complexity



Error vs Amount of (Training) Data
误差 vs（训练）数据量的关系



Error vs Amount of (Training) Data
误差 vs（训练）数据量的关系

- What happens if we increase model complexity? 如果我们增加模型复杂度会发生什么？
- Noise? 噪声？
- Variance? 方差？
- Bias? 偏差？