# Machine Learning

**Supervised Machine Learning – Regression**

**Part 2: Model Flexibility, Overfitting, Bias-Variance Trade-Off**

# Parametric Modeling – Part 2

# Parametric Modeling

Consider that $f$ can take the parametric form

$$f(\boldsymbol{x}) \approx m(\boldsymbol{x}; \boldsymbol{a}) \;\; \forall \boldsymbol{x} \in I, \boldsymbol{a} \in H$$

<u>Training:</u>

1. Identify parametric model $m$
2. Identify model parameters $\boldsymbol{a}$

# Training the Parameters (recap)

Assume the simplified case that true $m$ is known.

Then, to train $\hat{f}$ is to train the parameters of the model, $\boldsymbol{a}$, which can take values in set $H$.

Given training data $S = \{(y_n, \boldsymbol{x}_n)\}_{n=1}^{N}$ and loss function $L$, we train

$$\hat{f}(\boldsymbol{x}) = m(\boldsymbol{x}; \widehat{\boldsymbol{a}})$$

where

$$\widehat{\boldsymbol{a}} = \operatorname{argmin}_{\boldsymbol{b} \in H} \; L(\boldsymbol{b};\; S)$$

We call this "<u>fitting</u> the model parameters to the training data."

For <u>MSE training</u>, $L(\boldsymbol{b}, S) = \frac{1}{N} \sum_{n=1}^{N} (y_n - m(\boldsymbol{x}_n; \boldsymbol{b}))^2$

Other loss functions?

Better parameter training:
- Fewer parameters
- More training data $N$
- Lower noise variance, $\sigma_\epsilon^2$

# Identifying Model

- True knowledge of $m$ in $f$ is rarely the case. Needs domain expertise and simple $f$ .

- Need to select hypothetical model $\hat{m}$ (call it hypothesis) to approximate the true $m$ in $f$

# Model Flexibility

- Affine model:

$$m(\boldsymbol{x}, \boldsymbol{a}) = a_0 + a_1 x_1 + a_2 x_2$$

- Quadratic model:

$$m'(\boldsymbol{x}, \boldsymbol{a}) = a_o + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2$$

- It holds that:

$$m(\boldsymbol{x}, [a_0, a_1, a_2]^T) = m'(\boldsymbol{x}, [a_0, a_1, a_2, a_3 = 0, a_4 = 0, a_5 = 0]^T)$$

- $m'$ is more “flexible” than $m$
    - $m'$ has more parameters than $m$
    - $m'$ can *become* $m$ for certain parameter configuration

# Select Hypothesis and Train Parameters – Considerations

We want $\hat{f}(\boldsymbol{x}) = \widehat{m}(\boldsymbol{x}; \widehat{\boldsymbol{a}}) \approx f(\boldsymbol{x}) = m(\boldsymbol{x}; \boldsymbol{a})$.

What if $m$ is more flexible than $\widehat{m}$?

What if $\widehat{m}$ is more flexible than $m$?

What is more important, to select $\widehat{m}$ correctly or to train $\widehat{\boldsymbol{a}}$ well?

Place in order of preference:

- ❑ $\widehat{m} = m$, poor parameter training
- ❑ $\widehat{m}$ more flexible than $m$, excellent parameter training
- ❑ $\widehat{m}$ less flexible than $m$, poor parameter training
- ❑ $\widehat{m} = m$, excellent parameter training
- ❑ $\widehat{m}$ less flexible than $m$, excellent parameter training

# Select Hypothesis and Train Parameters – Considerations

We want $\hat{f}(\boldsymbol{x}) = \widehat{m}(\boldsymbol{x}; \widehat{\boldsymbol{a}}) \approx f(\boldsymbol{x}) = m(\boldsymbol{x}; \boldsymbol{a})$.

Given ample low-noise training data…

- ❑ Overestimate flexibility of $m$?
- ❑ Underestimate flexibility of $m$?

Given limited and or highly-noisy training data…

- ❑ Overestimate flexibility of $m$?
- ❑ Underestimate flexibility of $m$?

# Hypothesis vs Training Data Quality – Example

$d = 1$; $m(x; \boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x) = m(x; a) + \epsilon$; $E\{\epsilon^2\} = \sigma_\epsilon^2$

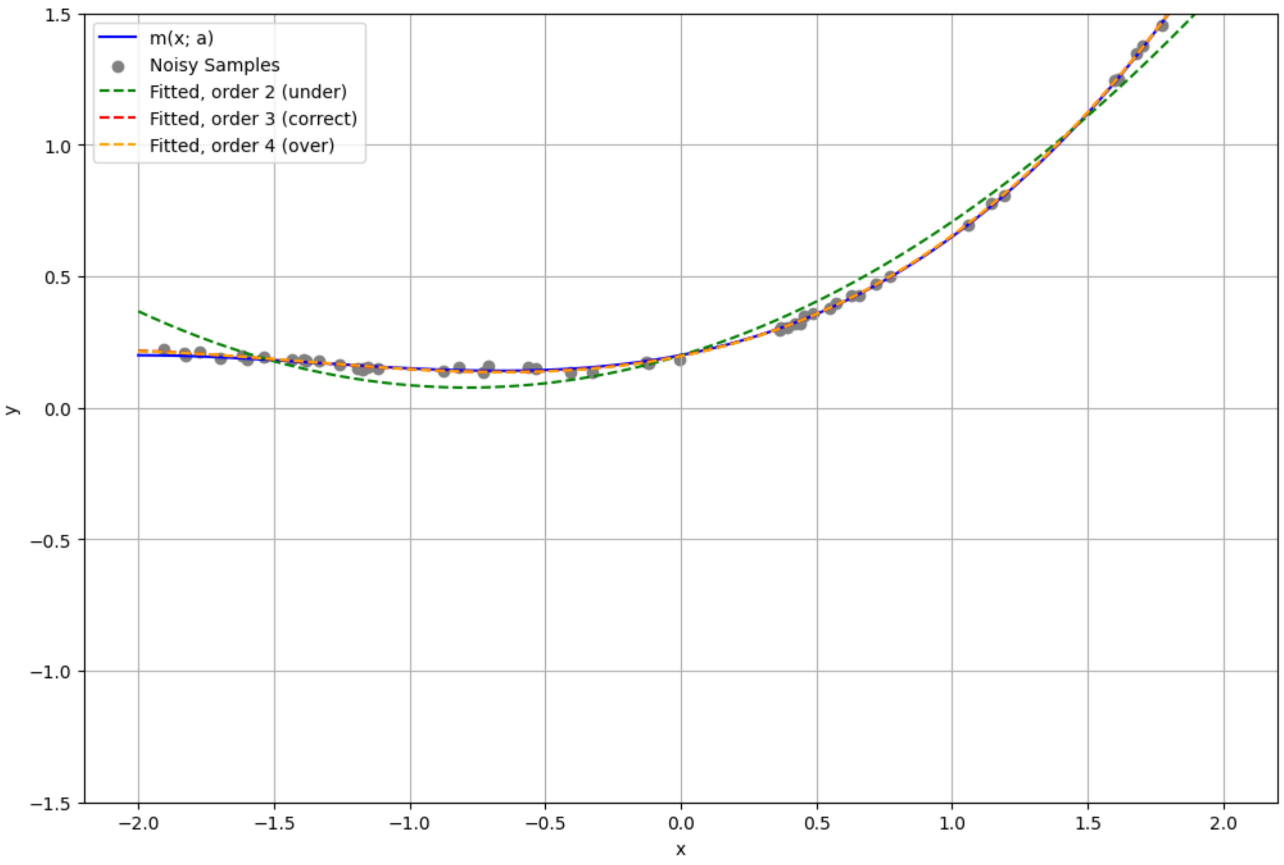
Model Estimation:

$\hat{m}$ estimated as:

- 2nd deg. poly. (3 param.)
- 3rd deg. poly. (4 param.)
- 4th deg. poly. (5 param.)

Mod. Est. Quality:

2nd deg. too low
3rd deg. is correct
4th deg. can be




Param. Training:

$\hat{m}$ fitted with:

- N=50 samples
- $\sigma_\epsilon = 0.01$

Param. Tr. Quality:

Enough low-noise data. All models train well.

Correct and high-flexibility trained models perform best.

# Hypothesis vs Training Data Quality – Example

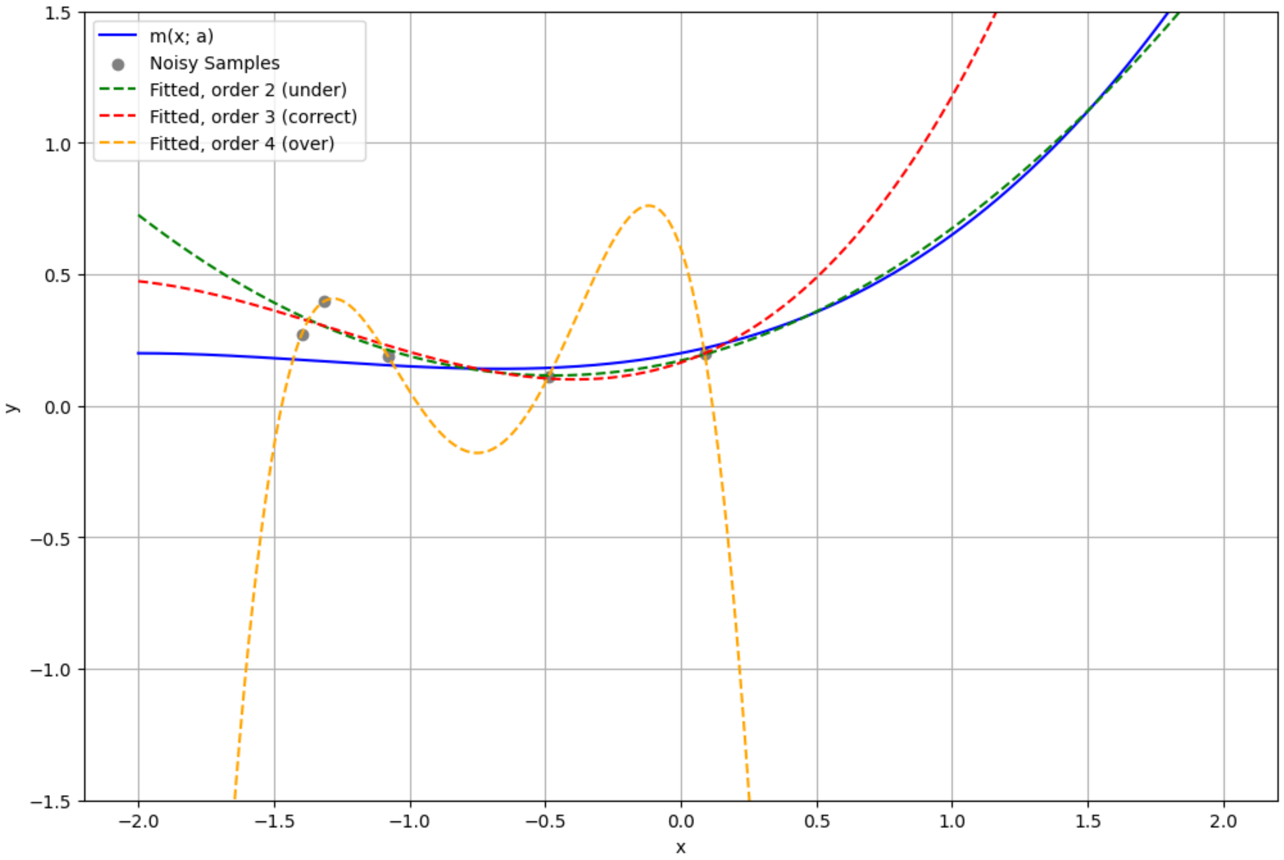
$d=1$; $m(x;\boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x)=m(x;a)+\epsilon$; $E\{\epsilon^2\}=\sigma_\epsilon^2$

<u>Model Estimation:</u>

$\hat{m}$ estimated as:
- $2^{nd}$ deg. poly. (3 param.)
- $3^{rd}$ deg. poly. (4 param.)
- $4^{th}$ deg. poly. (5 param.)

<u>Mod. Est. Quality:</u>

$2^{nd}$ deg. too low
$3^{rd}$ deg. is correct
$4^{th}$ deg. can be




<u>Param. Training:</u>

$\hat{m}$ fitted with:
- N=5 samples
- $\sigma_\epsilon = 0.1$

<u>Param. Tr. Quality:</u>

Limited noisy data. Flexible models do not train well.

Moderately-trained low-flexibility model performs better (!) than poorly-trained correct-flexibility and high-flexibility models.

# Hypothesis vs Training Data Quality – Example

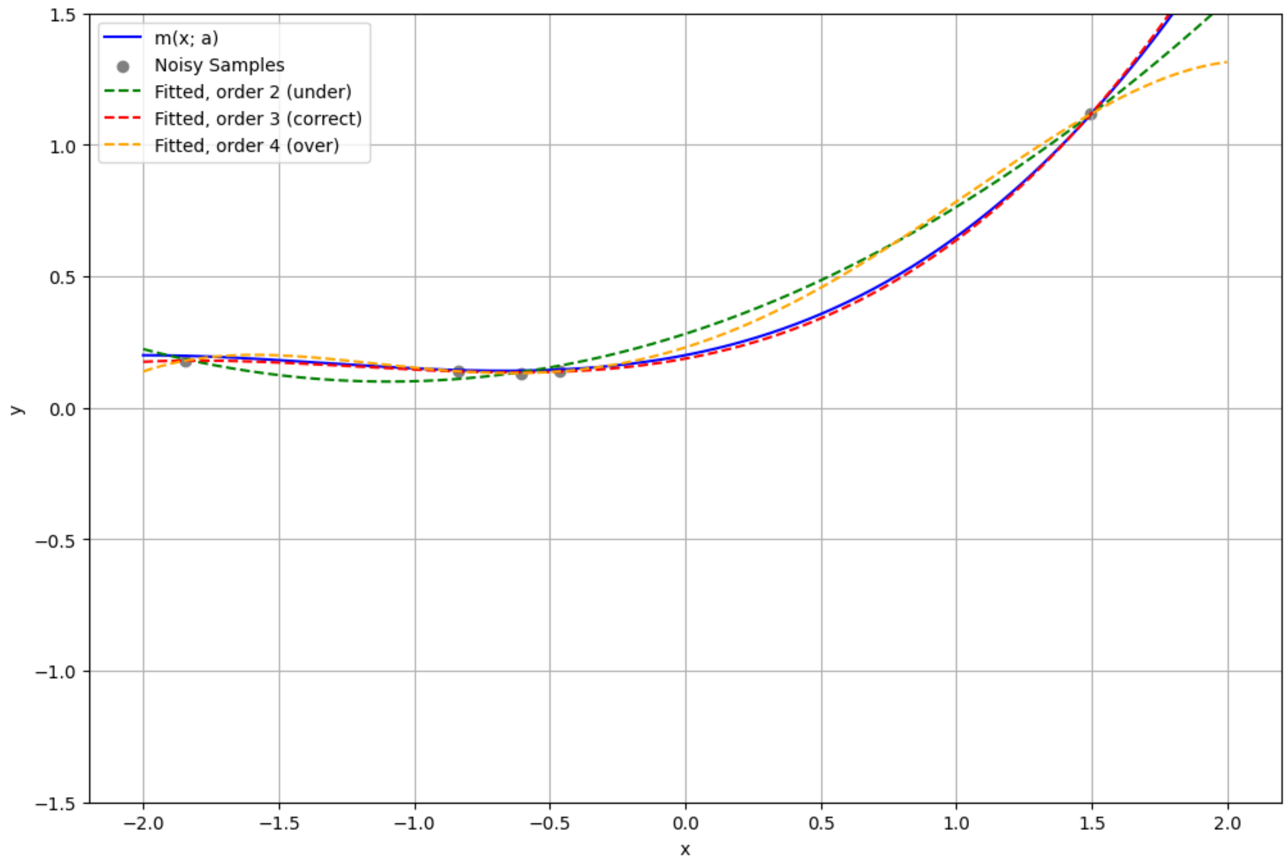
$d = 1$; $m(x; \boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x) = m(x; a) + \epsilon$; $E\{\epsilon^2\} = \sigma_\epsilon^2$

<u>Model Estimation:</u>

$\hat{m}$ estimated as:
- $2^{nd}$ deg. poly. (3 param.)
- $3^{rd}$ deg. poly. (4 param.)
- $4^{th}$ deg. poly. (5 param.)

<u>Mod. Est. Quality:</u>

$2^{nd}$ deg. too low
$3^{rd}$ deg. is correct
$4^{th}$ deg. can be




<u>Param. Training:</u>

$\hat{m}$ fitted with:
- N=5 samples
- $\sigma_\epsilon = 0.01$

<u>Param. Tr. Quality:</u>

Limited low-noise data. Still, high-flexibility model does not train very well.

Correct-flexibility trained model performs best.

# Hypothesis vs Training Data Quality – Example

$d = 1$; $m(x; \boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x) = m(x; a) + \epsilon$; $E\{\epsilon^2\} = \sigma_\epsilon^2$

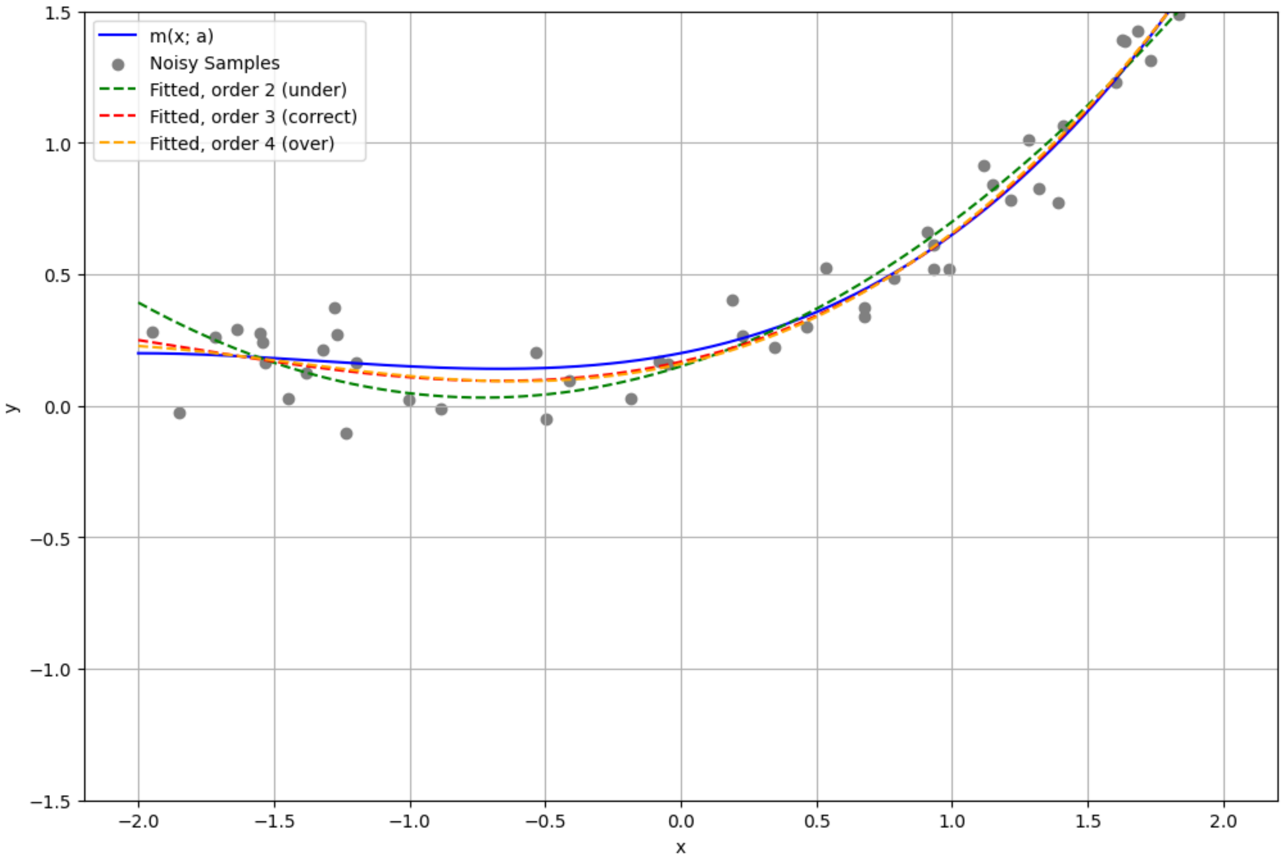
<u>Model Estimation:</u>

$\hat{m}$ estimated as:

- 2nd deg. poly. (3 param.)
- 3rd deg. poly. (4 param.)
- 4th deg. poly. (5 param.)

<u>Mod. Est. Quality:</u>

2nd deg. too low
3rd deg. is correct
4th deg. can be




<u>Param. Training:</u>

$\hat{m}$ fitted with:

- N=50 samples
- $\sigma_\epsilon = 0.1$

<u>Param. Tr. Quality:</u>

Enough noisy data. High-flexibility model does not train very well.

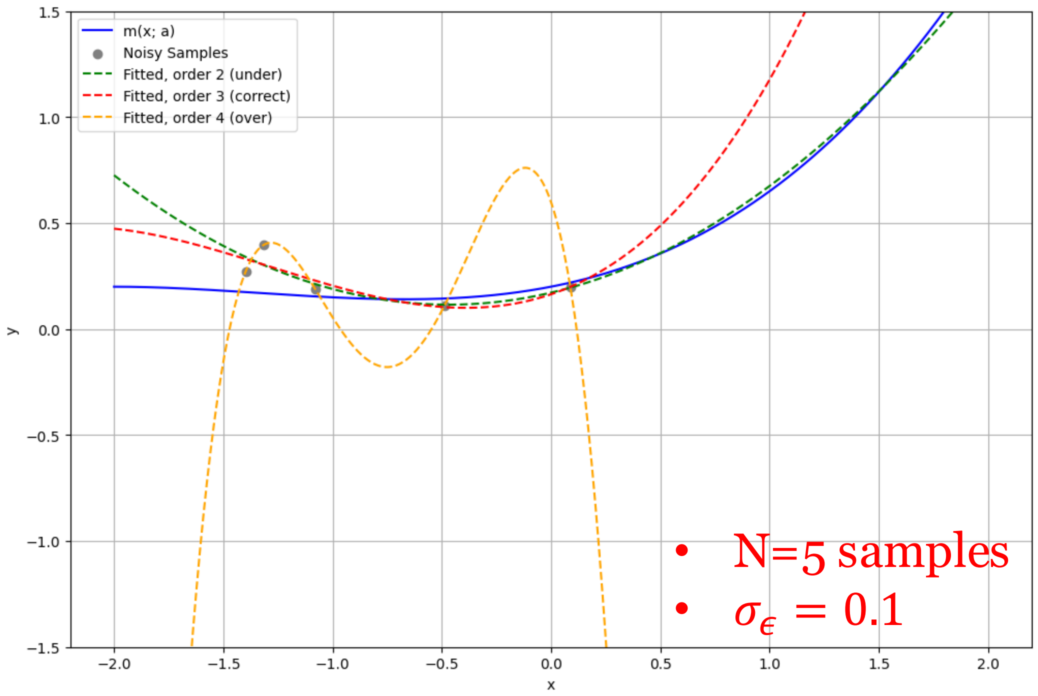
Correct-flexibility trained model performs best.

# Overfitting

- For any $\hat{m}$, $\hat{\boldsymbol{a}}$ is optimized so that $\hat{f}(\boldsymbol{x}) = \hat{m}(\boldsymbol{x}; \hat{\boldsymbol{a}})$ fits the training data <u>as closely as allowed by the</u> $\hat{m}$.
- If $\hat{m}$ is flexible, $\hat{f}(\boldsymbol{x})$ will fit well the training data.

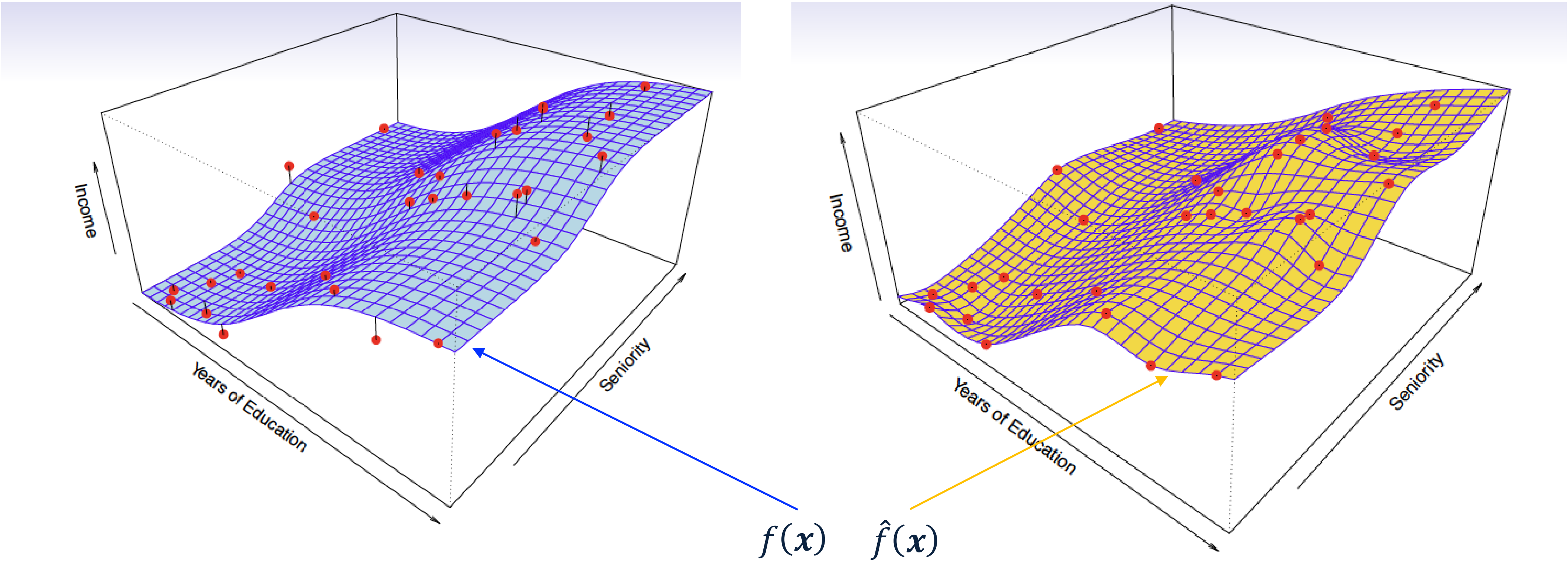
- This is positive, for many and/or low-noise data.
- This is negative, for limited and/or noisy data.

- **Overfitting**: The trained model $\hat{f}$ overfits data and captures the noise within it. For high noise intensity, overfitted model fails to represent $m$ and express other "unseen" data ("generalize").




- N=5 samples
- $\sigma_\epsilon = 0.1$

# Overfitting (cont'd)



# Measuring Model Accuracy

- Consider $\hat{f}(\boldsymbol{x}) = \widehat{m}(\boldsymbol{x}; \boldsymbol{a})$ trained on data $S_{\mathrm{tr}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$.

- We could examine MSE on $S_{tr}$.

$$\mathrm{MSE}_{tr} = \frac{1}{|S_{\mathrm{tr}}|} \sum_{(y,\boldsymbol{x}) \in S_{\mathrm{tr}}} \left| y - \hat{f}(\boldsymbol{x}) \right|^2$$

- This will be low. This is exactly what parameter training/fitting optimized.

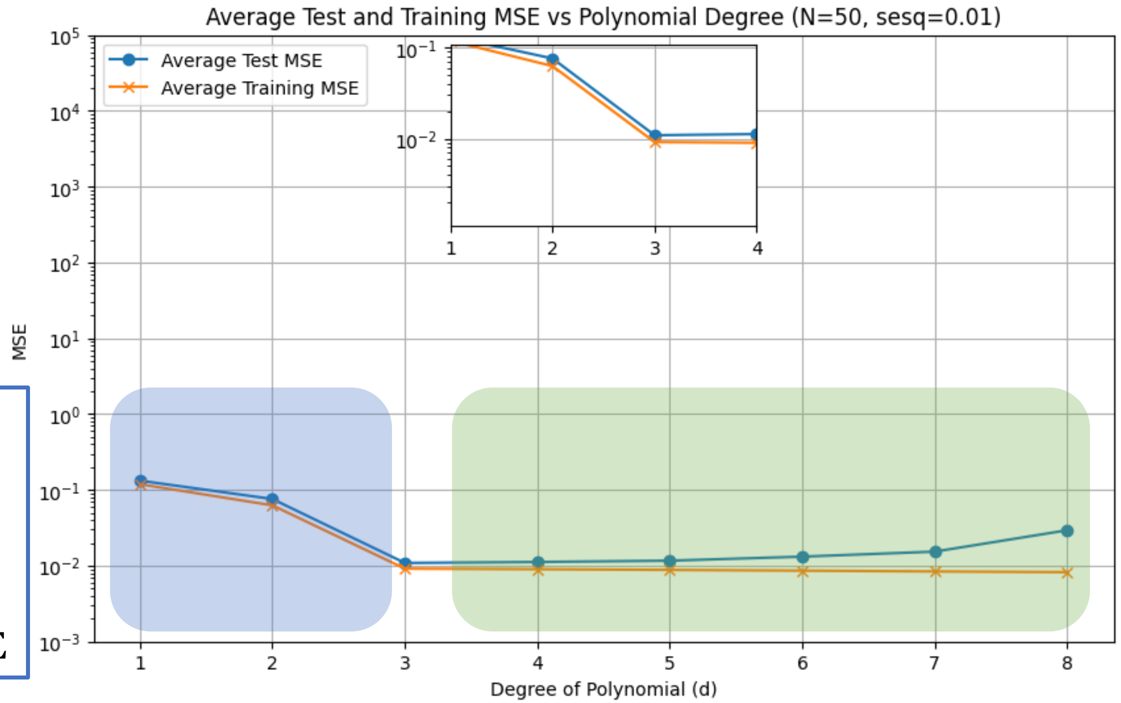
- Instead, we should examine MSE on fresh (unseen) test data $S_{\mathrm{te}} = \{(\boldsymbol{x}_i, y_i)\}_{\mathrm{i}=1}^{M}$:

$$\mathrm{MSE}_{\mathrm{te}} = \frac{1}{|S_{\mathrm{te}}|} \sum_{(y,\boldsymbol{x}) \in S_{\mathrm{te}}} \left| y - \hat{f}(\boldsymbol{x}) \right|^2$$

# Train-MSE and Test-MSE vs Flexibility for Various Data

$d = 1$; $m(x; \boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x) = m(x; a) + \epsilon$; $E\{\epsilon^2\} = \sigma_\epsilon^2$

- Moderate fitting to training data (underfitting).
- $\hat{m}$ not flexible enough to express $m$.
- Increased train-MSE AND test-MSE

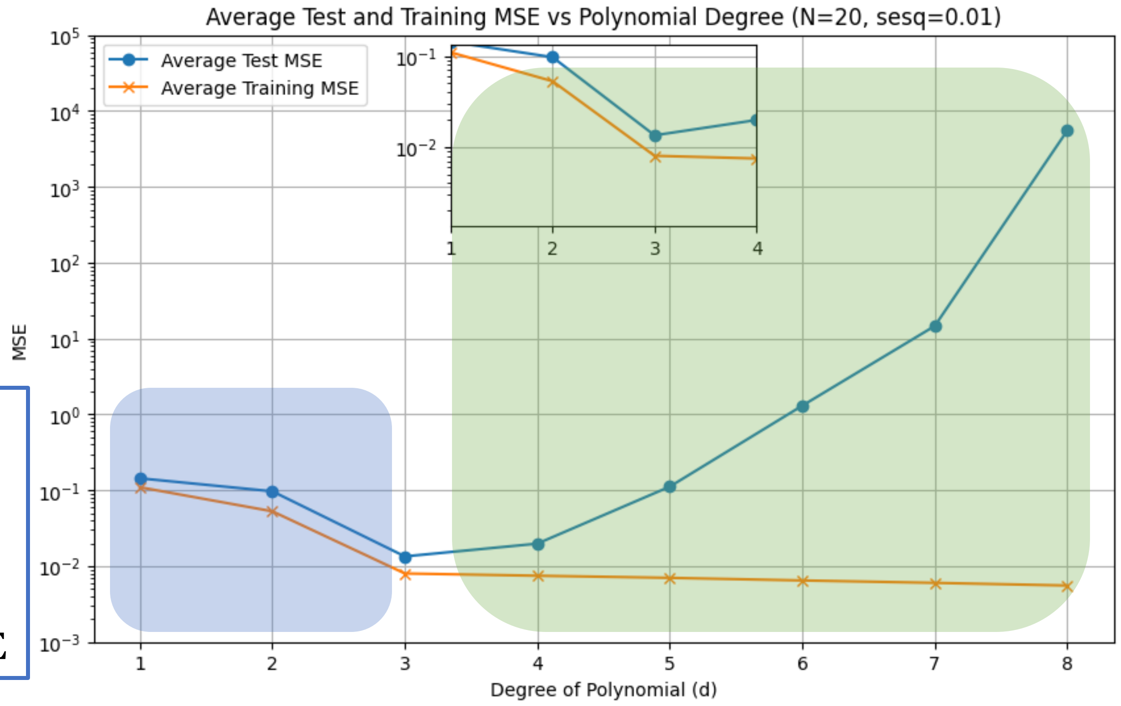



- Fitting well the training data.
- Slightly overfitting as $d$ increases because of enough training data and low noise intensity. Limited effect on test-MSE.

Param. Training: N=50 samples; $\sigma_\epsilon^2 = 0.01$

# Train-MSE and Test-MSE vs Flexibility for Various Data

$d = 1$; $m(x; \boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x) = m(x; a) + \epsilon$; $E\{\epsilon^2\} = \sigma_\epsilon^2$




- Pronounced overfitting as $d$ increases due to limited data. test-MSE increases fast.

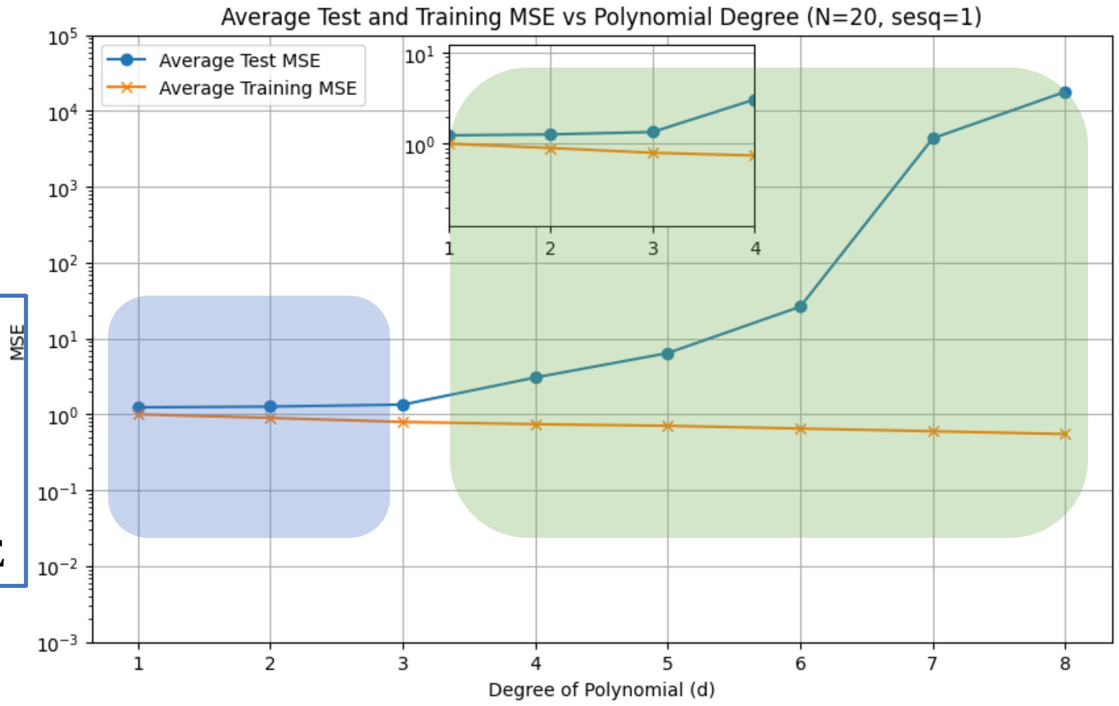
- Moderate fitting to training data (underfitting).
- $\hat{m}$ not flexible enough to express $m$.
- Increased train-MSE AND test-MSE

Param. Training: N=20 samples; $\sigma_\epsilon^2 = 0.01$

# Train-MSE and Test-MSE vs Flexibility for Various Data

$d = 1$; $m(x; \boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x) = m(x; a) + \epsilon$; $E\{\epsilon^2\} = \sigma_\epsilon^2$

- Moderate fitting to data (underfitting).
- $\hat{m}$ not flexible enough to express $m$.
- Increased train-MSE AND test-MSE




Param. Training: N=20 samples; $\sigma_\epsilon^2 = 1$

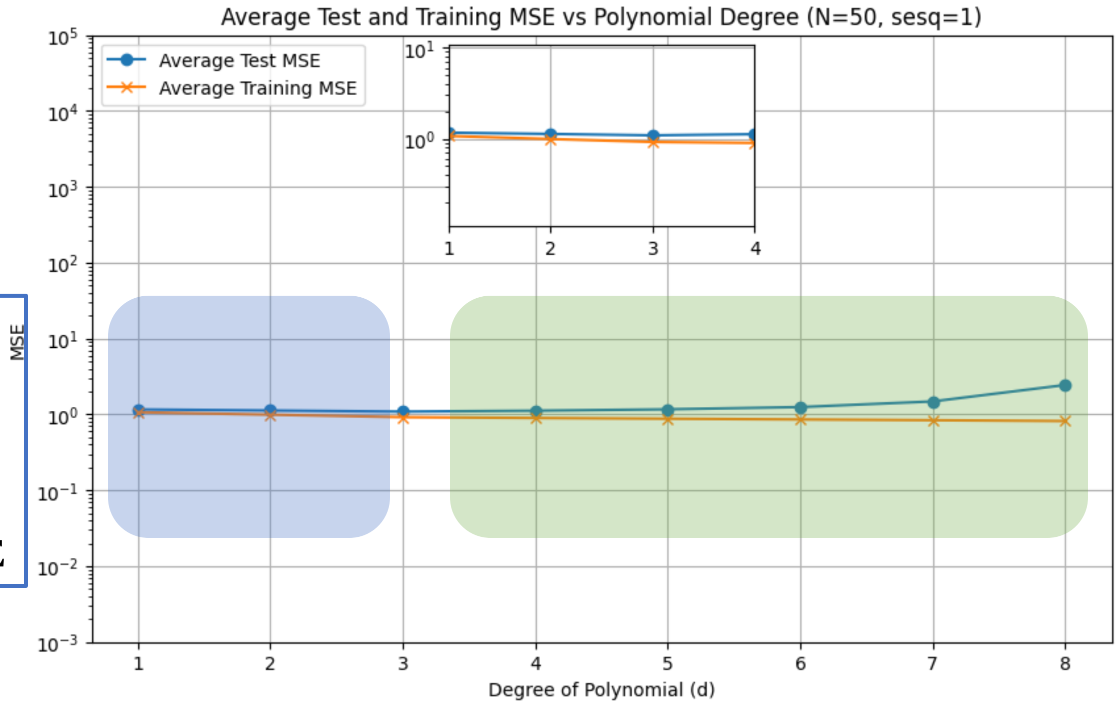
- Moderate fitting to data (higher train-MSE floor), even for high $d$, due to high $\sigma_\epsilon^2$.
- Yet, higher test-MSE than before due to capturing part of the high-intensity noise.

- Previous figure:
  Overfitting (low train-MSE) at high $d$, capturing most of the low-intensity noise. High test-MSE.
- This figure:
  Moderate fitting (higher train-MSE) at high $d$, capturing part of the high-intensity noise. Higher test-MSE.

# Train-MSE and Test-MSE vs Flexibility for Various Data

$d = 1$; $m(x; \boldsymbol{a})$ is 3rd degree polynomial (4 parameters); $f(x) = m(x; a) + \epsilon$; $E\{\epsilon^2\} = \sigma_\epsilon^2$

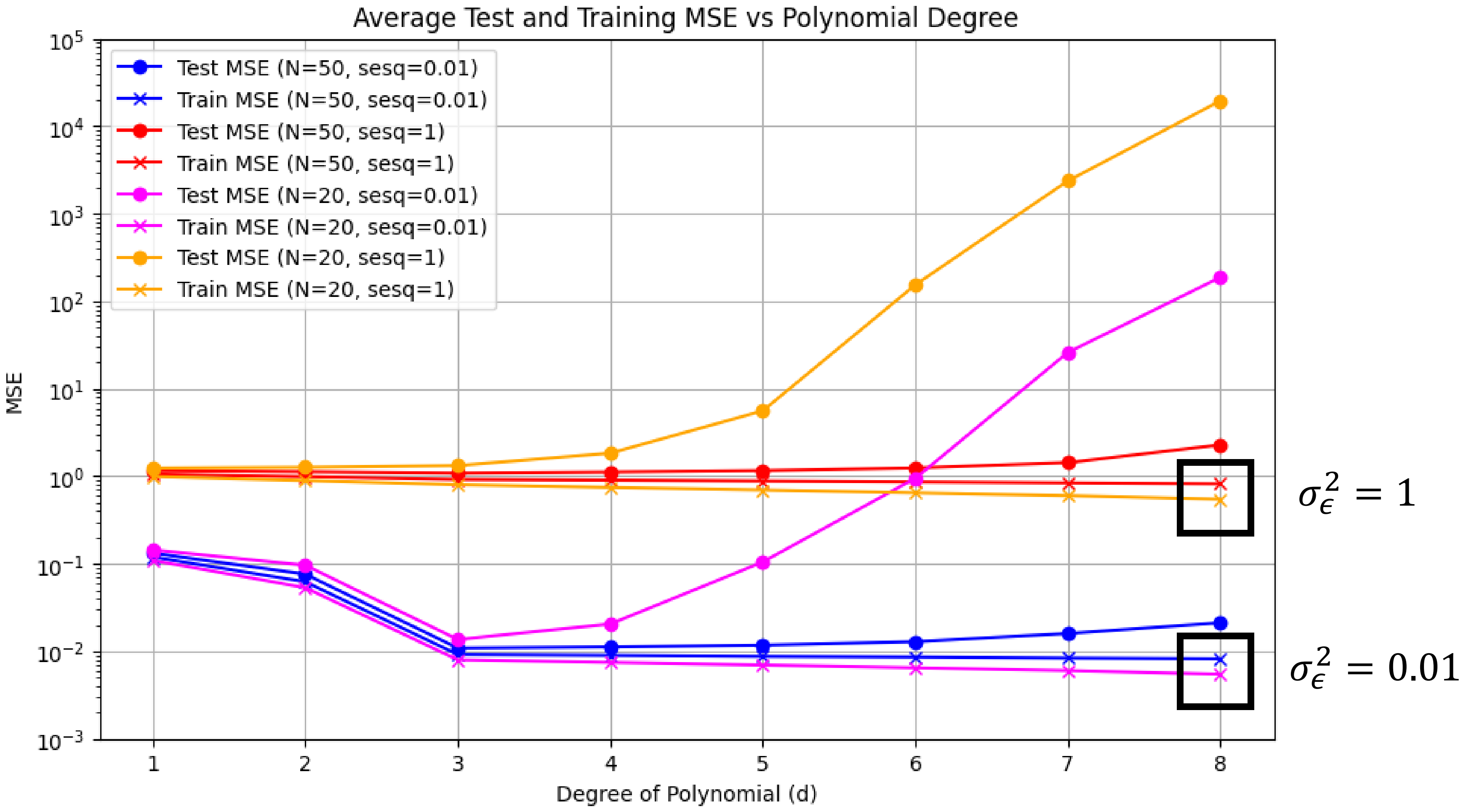
- Moderate fitting to training data (underfitting).
- $\hat{m}$ not flexible enough to express $m$.
- Increased train-MSE AND test-MSE



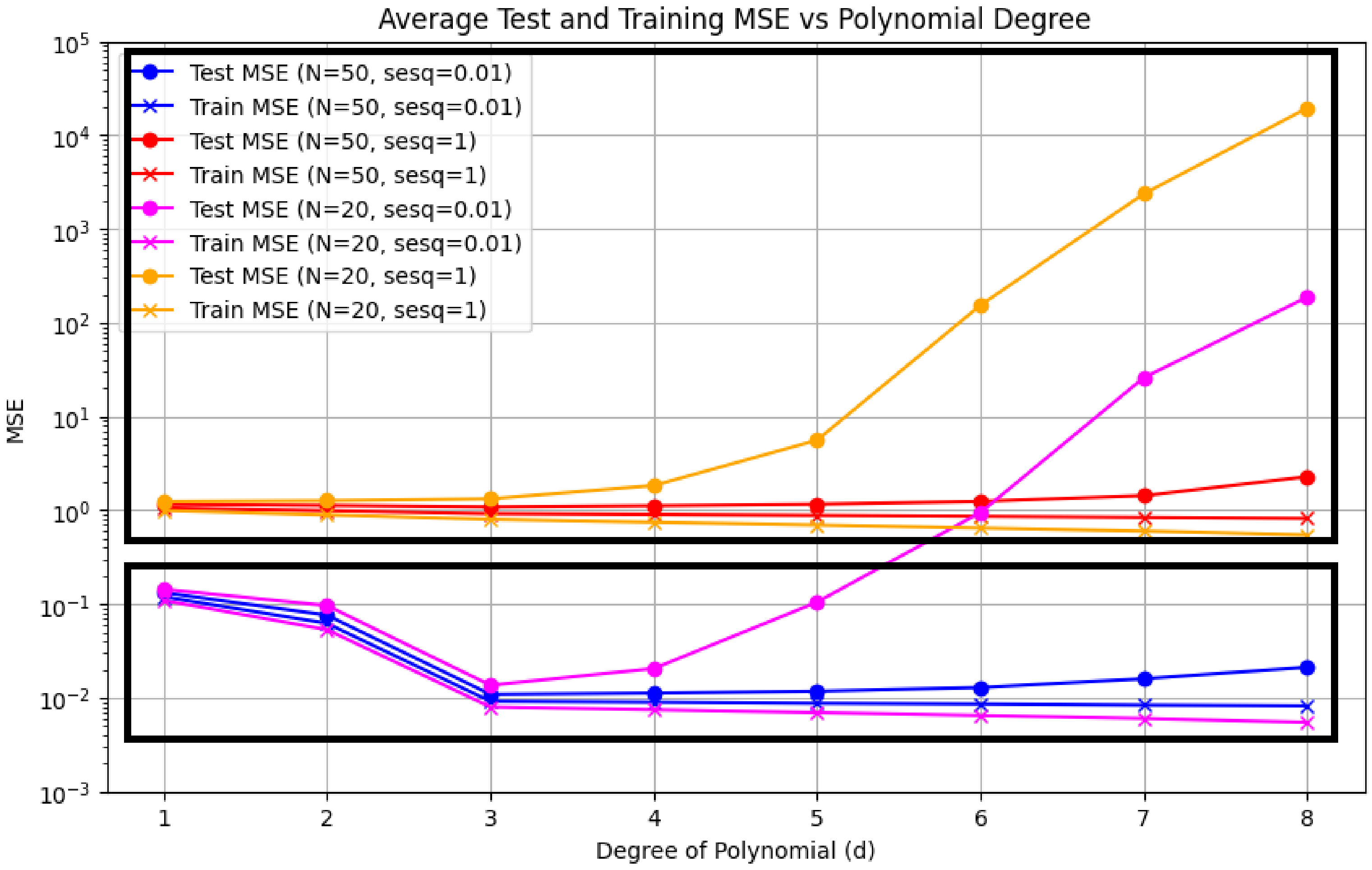

- Low fitting to data (flat train-MSE floor), even for high $d$, due to both low $N$ and high $\sigma_\epsilon^2$.
- Captures little of the high-intensity noise, which is averaged out due to high $N$. Low test-MSE.

<u>Param. Training:</u> N=50 samples; $\sigma_\epsilon^2 = 1$

# Train-MSE and Test-MSE vs Flexibility for Various Data



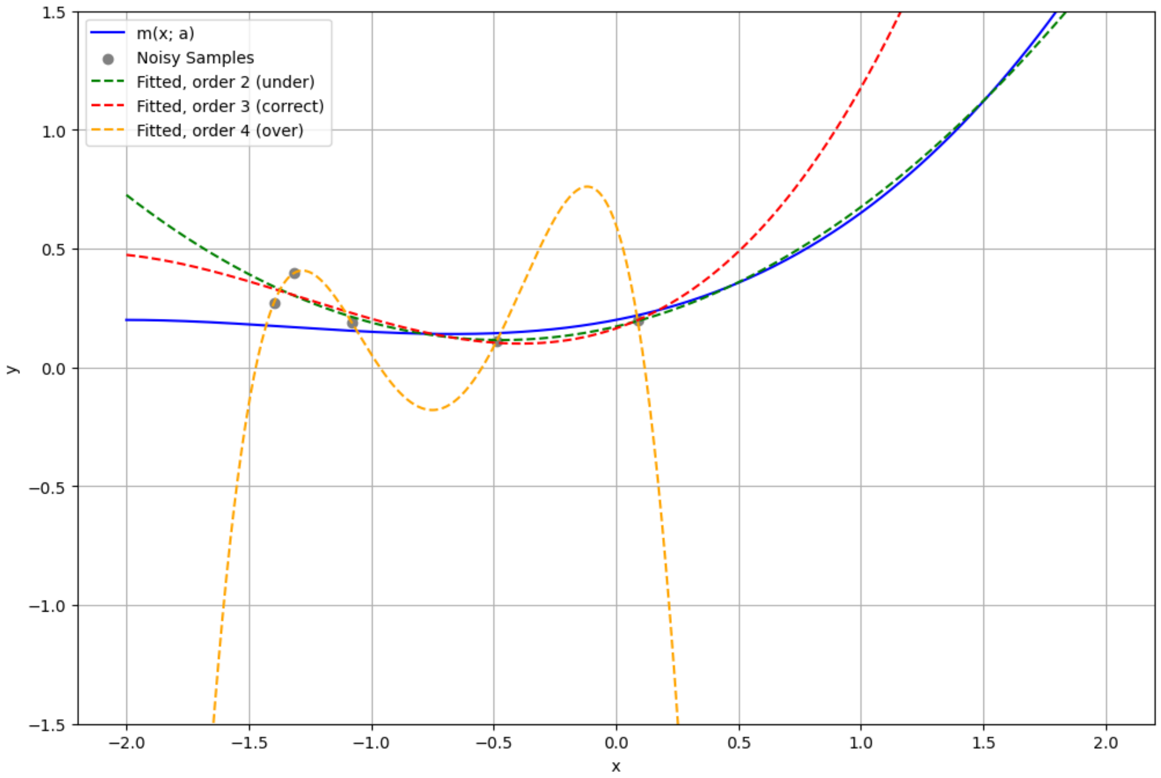
# Measuring Model Accuracy




High noise vs low noise

# How to Combat Overfitting?

- **Buy better training data:**
  - More examples
  - Less noise
- **Fixed training data:**
  - Lower hypothesis flexibility
  - Suboptimal parameter fitting



# Bias and Variance

- Assume that you have chosen $\hat{m}$ and you train the model over **random** dataset $S$ to obtain $\hat{f}$.
- For any unseen input $\boldsymbol{x}$ and corresponding output $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$, the model exhibits:

  - Bias:

    Mean over $S$

    $$\text{Bias}_S\left(\hat{f}(\boldsymbol{x})\right) = f(\boldsymbol{x}) - E_S\left[\hat{f}(\boldsymbol{x})\right]$$

    Bias: Error of $\hat{f}(\boldsymbol{x})$ to express $f(\boldsymbol{x})$, in the mean over $S$.

  - Variance:

    $$\text{Var}_S\left(\hat{f}(\boldsymbol{x})\right) = E_S\left[\left(\hat{f}(\boldsymbol{x}) - E_S[\hat{f}(\boldsymbol{x})]\right)^2\right]$$

    Variance of trained $\hat{f}(\boldsymbol{x})$, over $S$.

  - MSE:

    $$\text{MSE}_{S,\epsilon}\left(\hat{f}_S(\boldsymbol{x})\right) = E_{S,\epsilon}\left[\left(y(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\right)^2\right]$$

    SE attained by trained $\hat{f}(\boldsymbol{x})$ on unseen $\boldsymbol{x}$, in the mean over $S$ and error in $y(\boldsymbol{x})$.

# Bias and Variance (cont'd)

Simplify notation: $Bias(\hat{f}) = f - E\left[\hat{f}\right]$, $Var(\hat{f}) = E\left[\left(\hat{f} - E[\hat{f}]\right)^2\right]$, and $MSE(\hat{f}) = E\left[\left(y - \hat{f}\right)^2\right]$

Then we find:

$$\mathrm{MSE}_{\mathrm{S},\epsilon} = E\left[\left(y - \hat{f}\right)^2\right] = E\left[\left(f + \epsilon - \hat{f}\right)^2\right] = E\left[\left(f - \hat{f}\right)^2 + \epsilon^2 + 2\epsilon\left(f - \hat{f}\right)\right]$$

$$= E\left[\left(f - \hat{f}\right)^2\right] + E[\epsilon^2] + 2E\left[\left(f - \hat{f}\right)\epsilon\right] = E\left[\left(f - \hat{f}\right)^2\right] + \sigma_\epsilon^2$$

In turn we find:

$$E\left[\left(f - \hat{f}\right)^2\right] = E\left[f^2 + \hat{f}^2 - 2f\hat{f}\right] = f^2 + E\left[\hat{f}^2\right] - 2fE\left[\hat{f}\right]$$

$$= f^2 + E\left[\hat{f}^2\right] - 2fE\left[\hat{f}\right] + E\left[\hat{f}\right]^2 - E\left[\hat{f}\right]^2 - 2E\left[\hat{f}\right]^2 + 2E\left[\hat{f}\right]^2$$

$$= f^2 + E\left[\hat{f}^2 + E\left[\hat{f}\right]^2 - 2E\left[\hat{f}\right]^2\right] - 2fE\left[\hat{f}\right] + E\left[\hat{f}\right]^2$$

$$= \left(f - E\left[\hat{f}\right]\right)^2 + E\left[\hat{f}^2 + E\left[\hat{f}\right]^2 - 2E\left[\hat{f}\right]\hat{f}\right] = \left(f - E\left[\hat{f}\right]\right)^2 + E\left[\left(\hat{f} - E\left[\hat{f}\right]\right)^2\right] = Bias^2(\hat{f}) + Var(\hat{f})$$

# Bias and Variance (cont'd)

We proved that, for any given $\boldsymbol{x}$, the MSE is:

$$\mathrm{MSE}_{S,\epsilon}\left(\hat{f}(\boldsymbol{x})\right) = \left(\mathrm{Bias}_S\left(\hat{f}(\boldsymbol{x})\right)\right)^2 + \mathrm{Var}_S\left(\hat{f}(\boldsymbol{x})\right) + \sigma_\epsilon^2$$

The mean MSE over all (random) unseen data is:

$$\mathrm{MSE} = E_{\boldsymbol{x}}\left[\mathrm{MSE}_{S,\epsilon}\left(\hat{f}(\boldsymbol{x})\right)\right] = E_{\boldsymbol{x}}\left[\left(\mathrm{Bias}_S\left(\hat{f}(\boldsymbol{x})\right)\right)^2 + \mathrm{Var}_S\left(\hat{f}(\boldsymbol{x})\right)\right] + \sigma_\epsilon^2$$

# Bias and Variance (cont'd)

- We proved that the MSE is:

$$\text{MSE} = E_{\boldsymbol{x}}\left[\left(\text{Bias}_S\left(\hat{f}(\boldsymbol{x})\right)\right)^2 + \text{Var}_S\left(\hat{f}(\boldsymbol{x})\right)\right] + \sigma_\epsilon^2$$

$$\text{MSE} = \underbrace{E_{\boldsymbol{x}}\left[\left(f(\boldsymbol{x}) - E_S\left[\hat{f}(\boldsymbol{x})\right]\right)^2\right]}_{B} + \underbrace{E_{\boldsymbol{x}}\left[E_S\left[\left(\hat{f}(\boldsymbol{x}) - E_S[\hat{f}(\boldsymbol{x})]\right)^2\right]\right]}_{V} + \sigma_\epsilon^2$$

- How to reduce MSE?
  - Consider $\sigma_\epsilon^2$ given.
  - Reduce Bias and/or Variance.
  - Two things to tune: hypothesis and number of training data.

# High Bias

$$\text{MSE} = \underbrace{E_{\boldsymbol{x}}\left[\left(f(\boldsymbol{x}) - E_S\left[\hat{f}(\boldsymbol{x})\right]\right)^2\right]}_{B} + \underbrace{E_{\boldsymbol{x}}\left[E_S\left[\left(\hat{f}(\boldsymbol{x}) - E_S[\hat{f}(\boldsymbol{x})]\right)^2\right]\right]}_{V} + \sigma_\epsilon^2$$

High $B$:

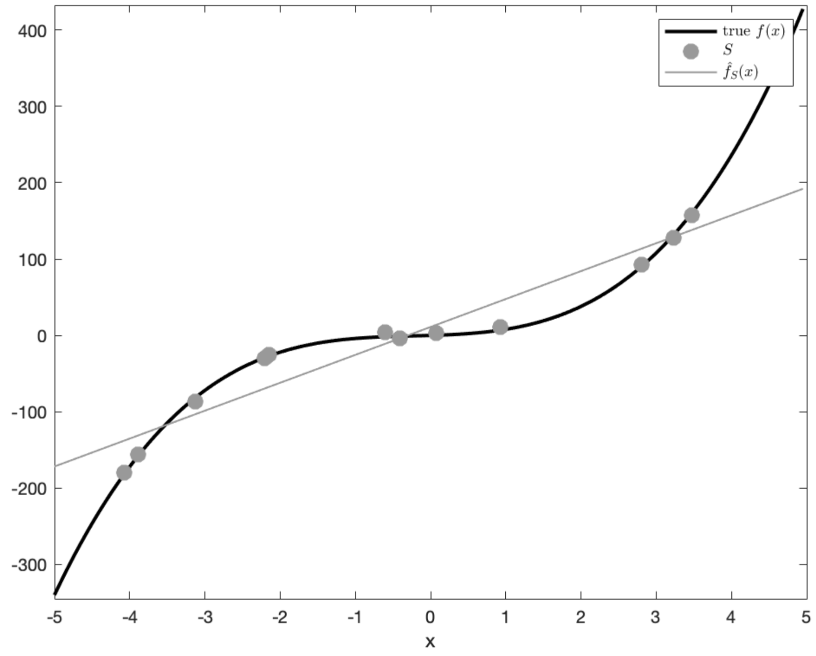
- Over the unseen points (in the mean), over the possible training datasets of size $N$ (in the mean), your model is far from true $f$.
- This is because hypothesis $\hat{m}$ is too rigid (not flexible enough).
- Cannot fit true $f$ and generalize to unseen data.

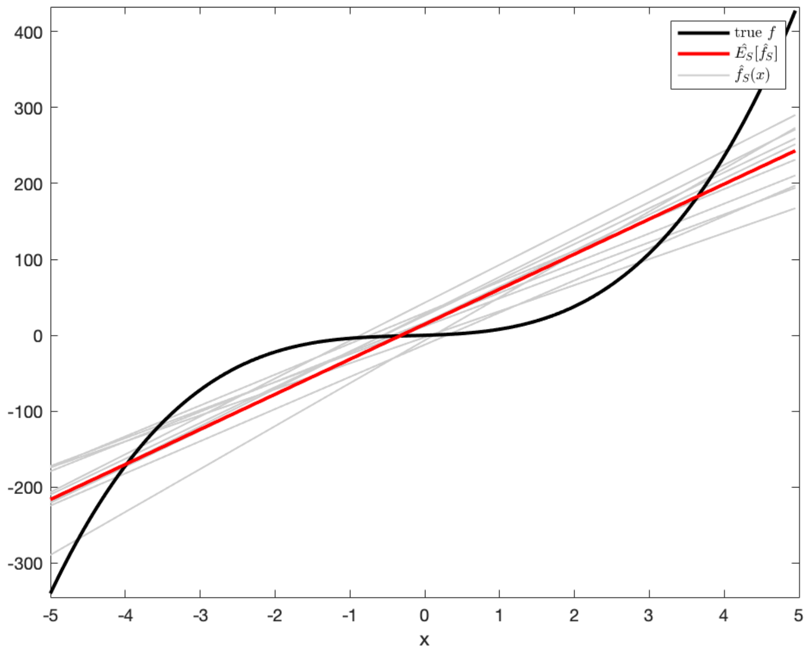
Remedy:

- For fixed $N$, increase the flexibility (e.g., #parameters) of $\hat{m}$.

# Example 4 – High Bias

True: $f(x) = 3x^3 + 2x^2 + 3x$. Hypothesis: line (simpler). $N = 12, \sigma_\epsilon = 5$.



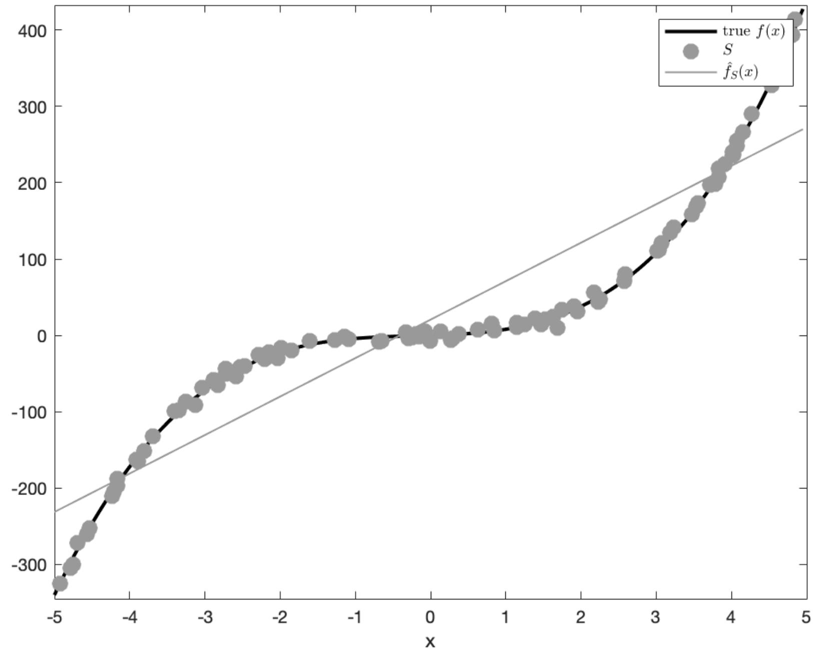

Model cannot fit the training data (underfitting).

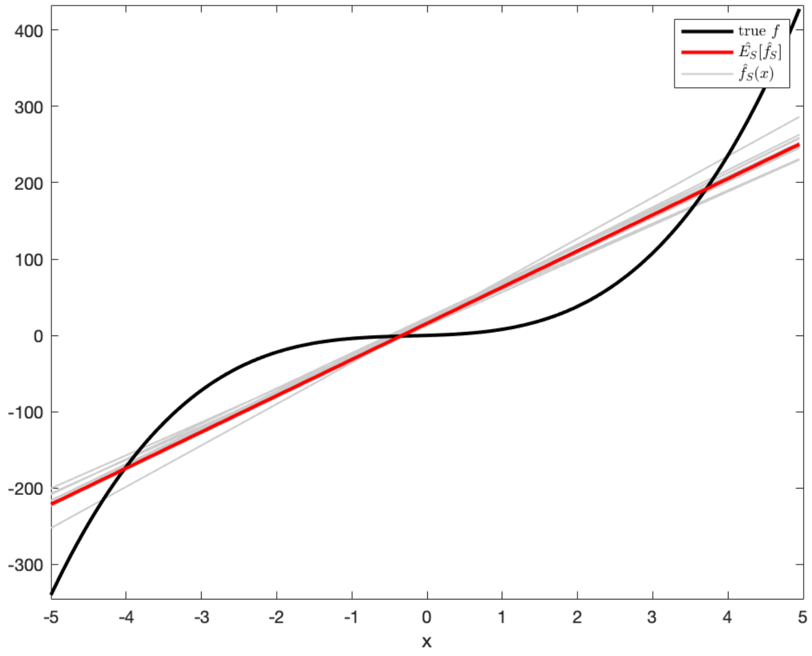
$$B = E_{\boldsymbol{x}}\left[\left(f(\boldsymbol{x}) - E_S\left[\hat{f}(\boldsymbol{x})\right]\right)^2\right]$$

- Moderate deviation of **model instances** is far from the mean model (over all training datasets of size $N$). Both model instances and mean model are far from the true $f$.

# Example 4 - Increase Data

Same simple hypothesis. Increase $N = 64$.



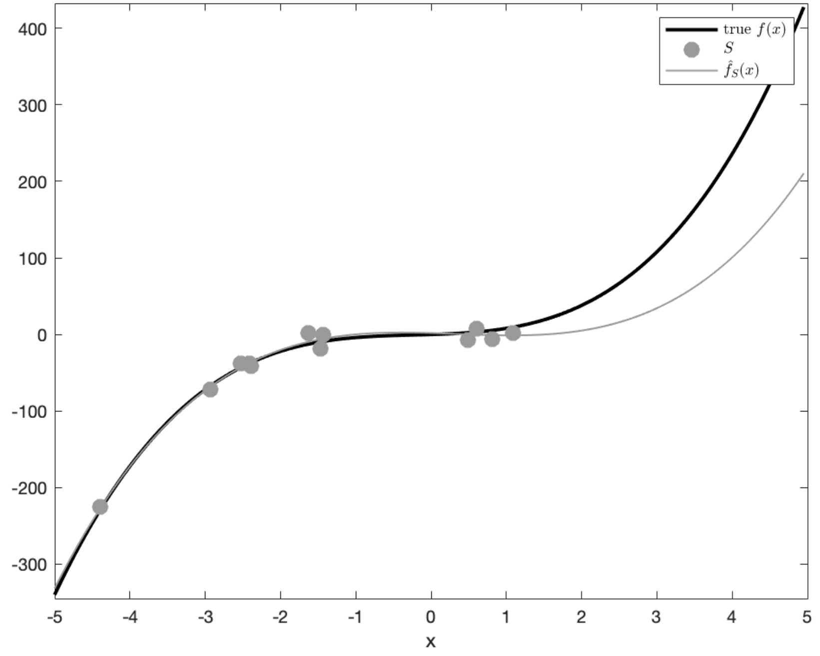



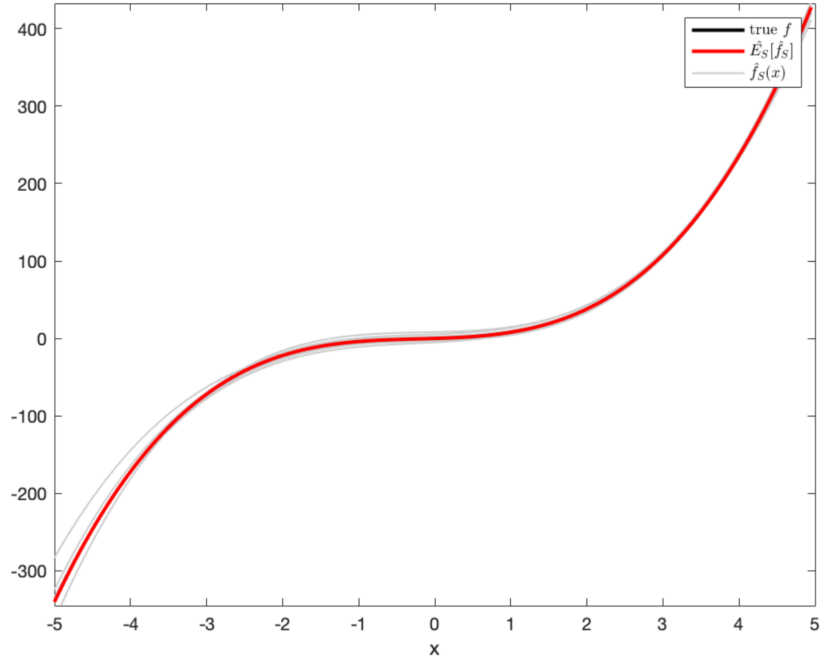

Still underfitting.

- Bias remains. Variance (and, thus, MSE) somewhat decreased.
- Increasing the amount of data typically drops variance.

# Example 4 - Increase Flexibility

Keep $N = 12$ data. Increase flexibility to deg-3 polynomial.




- Fits data better, but not all of them and not perfectly.
- Since $\sigma > 0$ this is good; the model can generalize.

- Bias eliminated.
- Variance remains.

# High Variance

$$\text{MSE} = \underbrace{E_x\left[\left(f(x) - E_S\left[\hat{f}(x)\right]\right)^2\right]}_{B} + \underbrace{E_{\boldsymbol{x}}\left[E_S\left[\left(\hat{f}(\boldsymbol{x}) - E_S\left[\hat{f}(\boldsymbol{x})\right]\right)^2\right]\right]}_{V} + \sigma_\epsilon^2$$

High $V$:

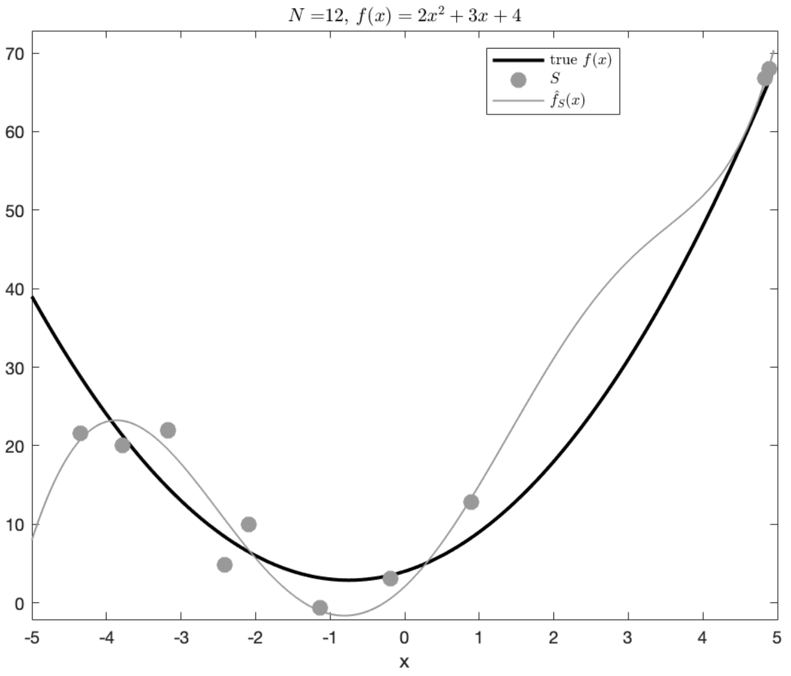
- Over the unseen points (in the mean), model instances across different training datasets of size $N$ deviate a lot from the mean model (in the mean).
- This is because our hypothesis is too flexible and overfits to each specific training dataset.
- Each training dataset contains error and deviates from true $f$.
- Thus, overfitted model is far from the true $f$ and cannot express unseen data (generalize).

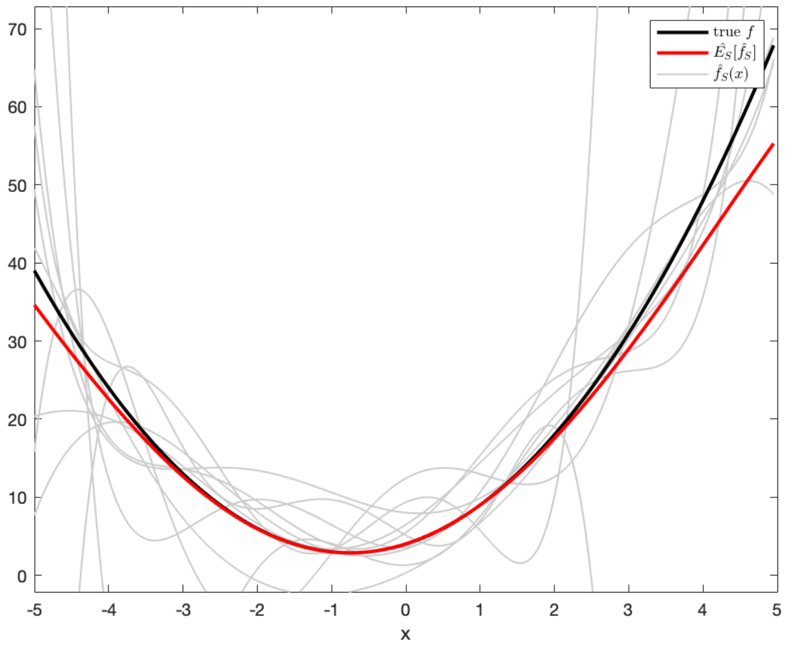
Remedy:

- For same number of data, reduce flexibility to reduce variance.
- For same hypothesis, increase the number of data to reduce variance.

# Example 5 - High Variance

True: $f(x) = 2x^2 + 3x + 4$. Hypothesis: deg-6 polynomial (more complex). $N = 12, \sigma = 5$.



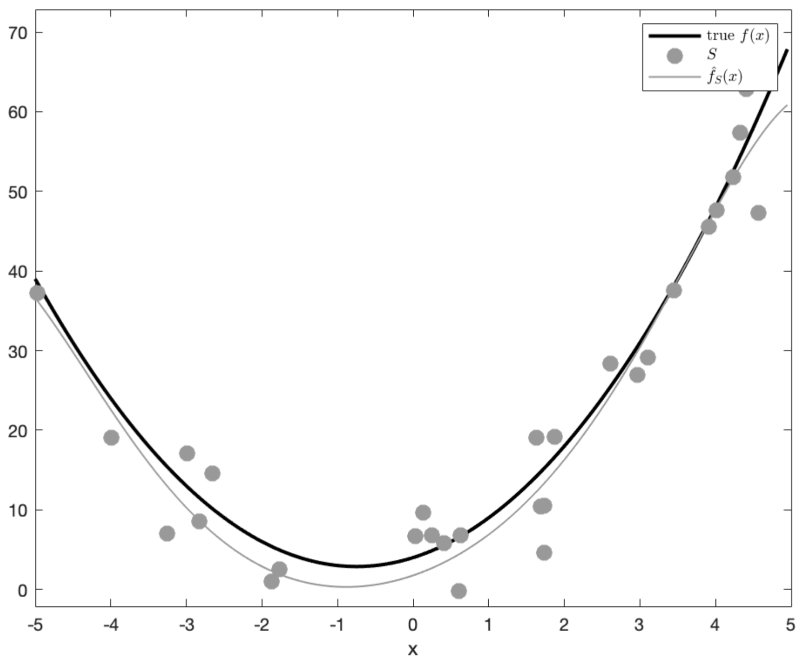



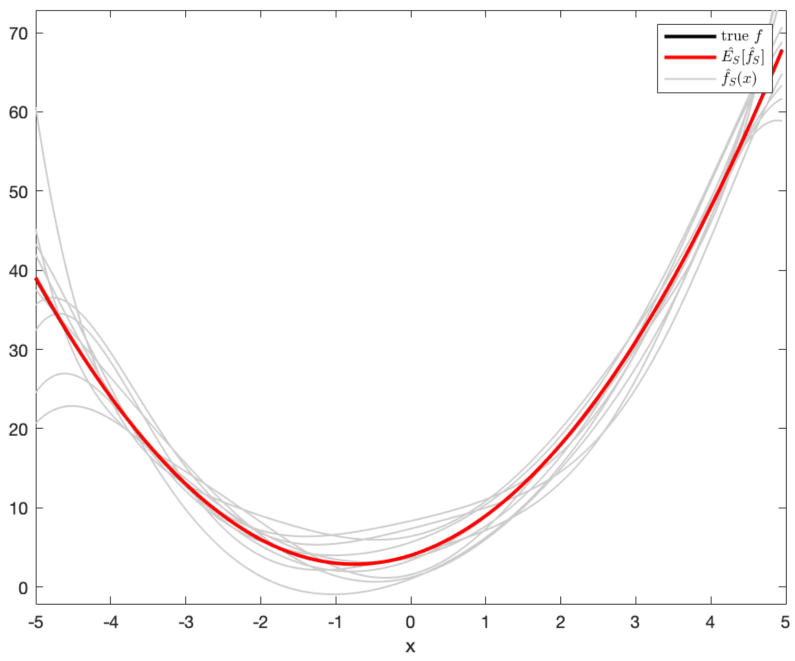

Model fits more to the training data than true $f$. **Overfitting.**

- The mean model (over all datasets of size $N$) matches the true $f$. **Minimal Bias.**
- But each **model instance** is far from the mean model and the true $f$. **High Variance.**

# Example 5 - Increase Data

Same flexible hypothesis. **Increase $N = 32$.**



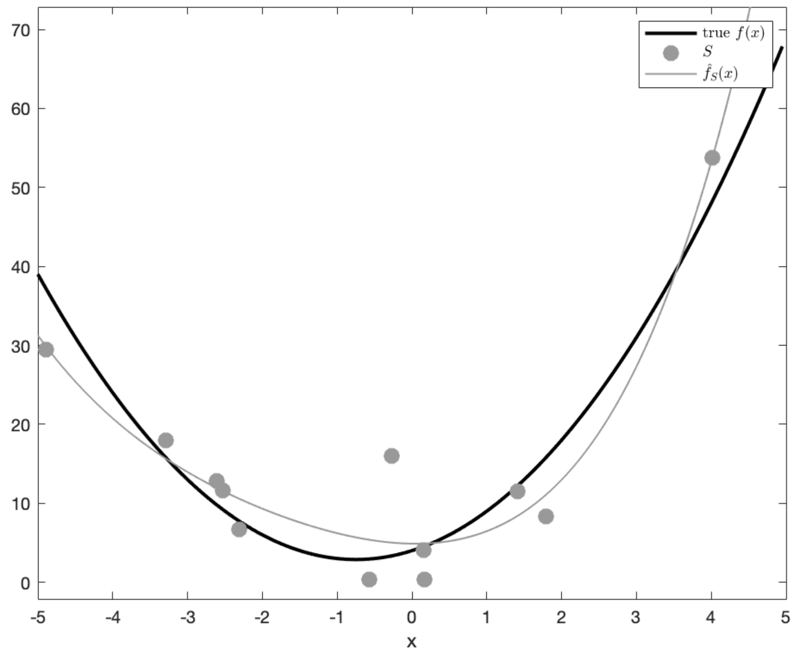



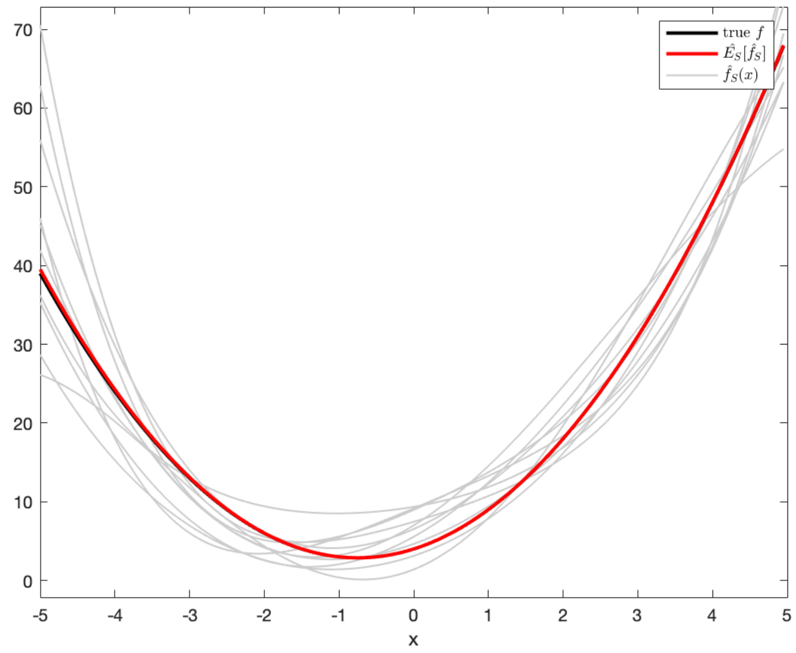

Model cannot fit to training data. Not complex enough for increased $N$. Model balances among training data, staying closer to the true $f$.

- The mean model matches the true $f$. **No Bias.**
- Now each **model instance** is closer to the mean model and the true $f$. **Variance reduced.**

# Example 5 - Reduce Flexibility

Keep same $N$. **Reduce flexibility to deg-4 polynomial.**



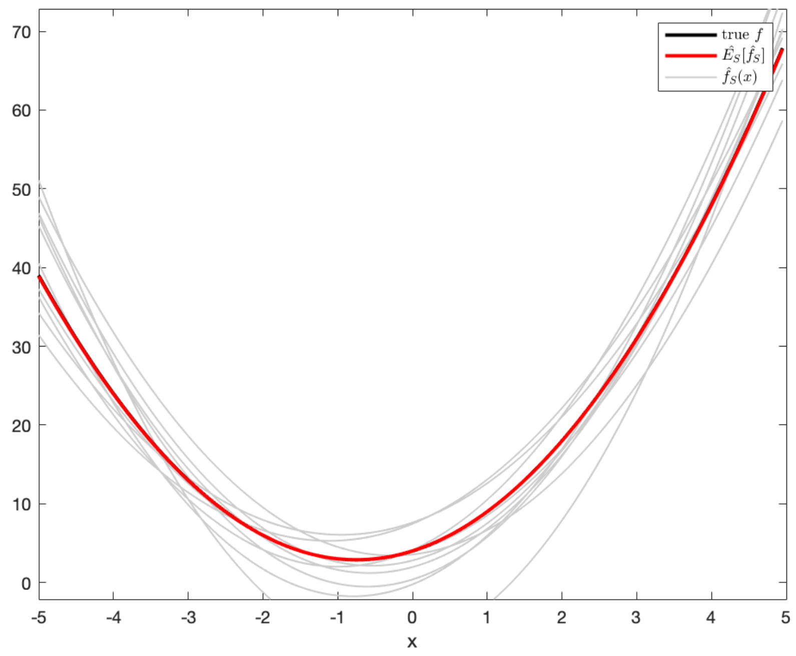




Model cannot fit to all training data. Not complex enough. Model balances among training data, staying closer to the true $f$.

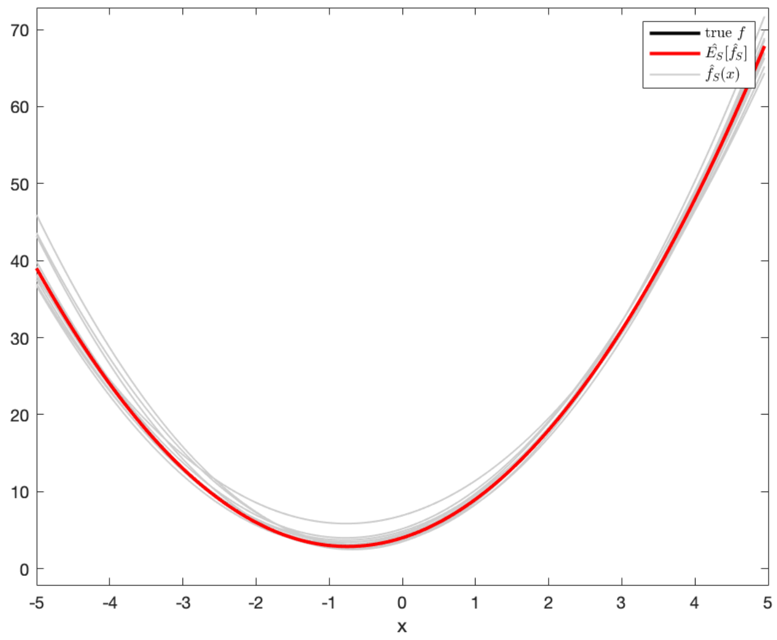
- The mean model almost the true $f$. **Low bias.**
- Each **model instance** is closer to the mean model and the true $f$. **Variance reduced.**

# Example 5 - Correct Hypothesis

For correct hypothesis, even $N = 6$ suffices. More training data will be better of course.

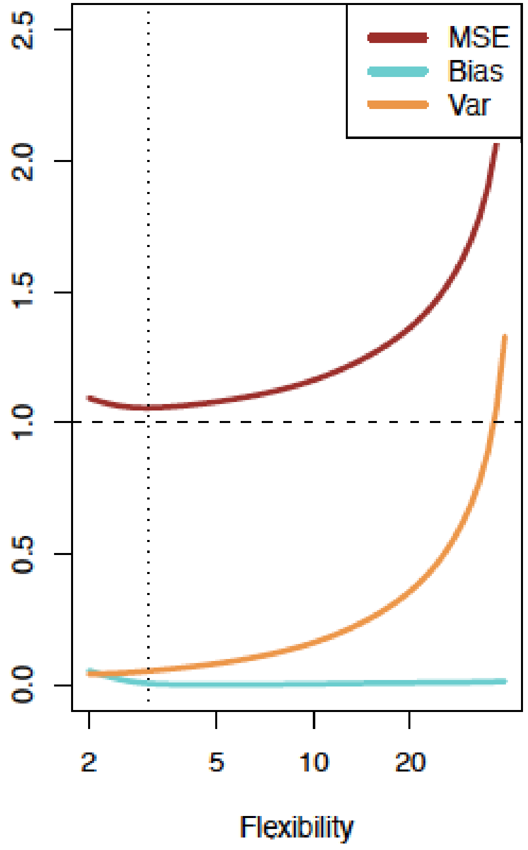



$N = 6$




$N = 32$

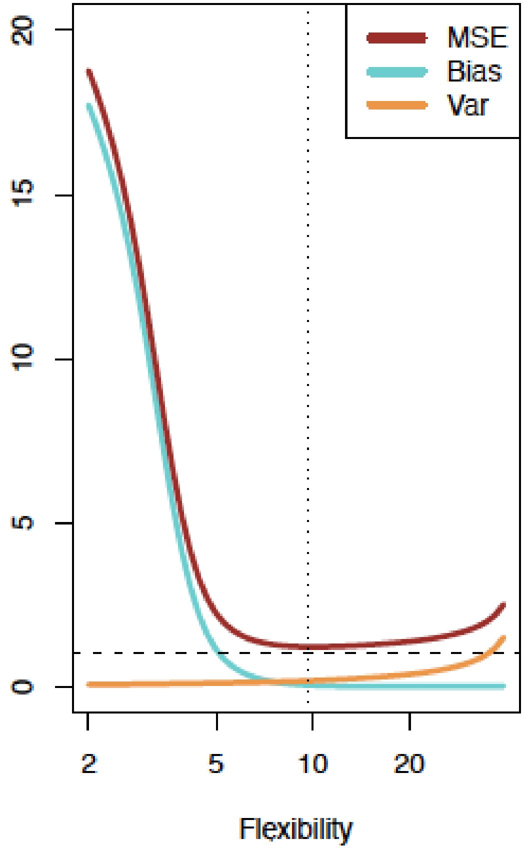
# Bias-Variance Trade-Off




Simple $f$.

- For lower model flexibility,
  - **Some very little bias exists.** Simple model underfits.
  - Assuming enough data, variance is low.
  - Therefore, MSE is low.
- As flexibility increases to the best spot,
  - **Bias further reduces.**
  - Variance starts increasing but remains low.
  - MSE remains low.
- As flexibility increases excessively,
  - Variance increases. Flexible model overfits to training data.
  - Bias remains low.
  - Following the variance, MSE increases.

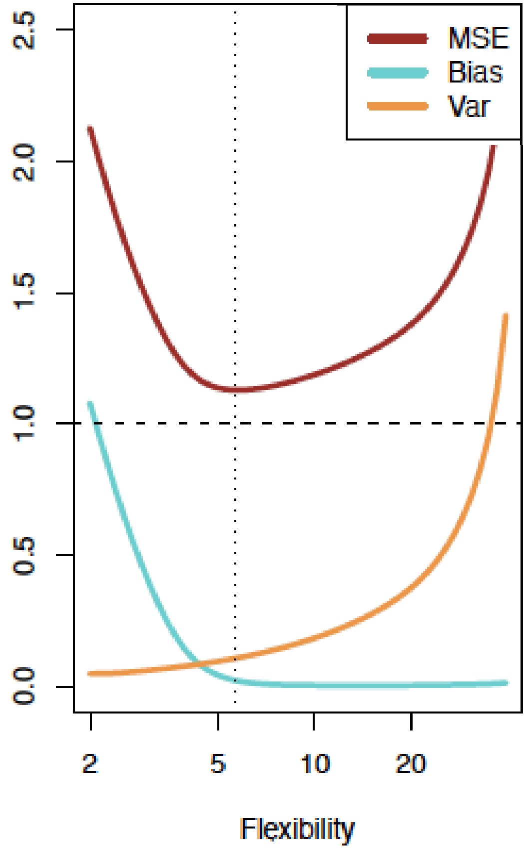
# Bias-Variance Trade-Off (cont'd)




Complex $f$.

- For lower model flexibility,
  - **Bias is very high. Sever underfitting.**
  - Assuming enough data, variance is low.
  - Following the bias, MSE is very high.
- As flexibility increases toward the best spot,
  - Bias rapidly drops.
  - Variance remains low.
  - MSE rapidly drops, following the bias.
- As flexibility starts increasing excessively,
  - Bias remains low.
  - Variance starts increasing. As #parameters increase, while $N$ remains fixed, mild overfitting starts appearing.
  - Following the variance, MSE starts increasing.

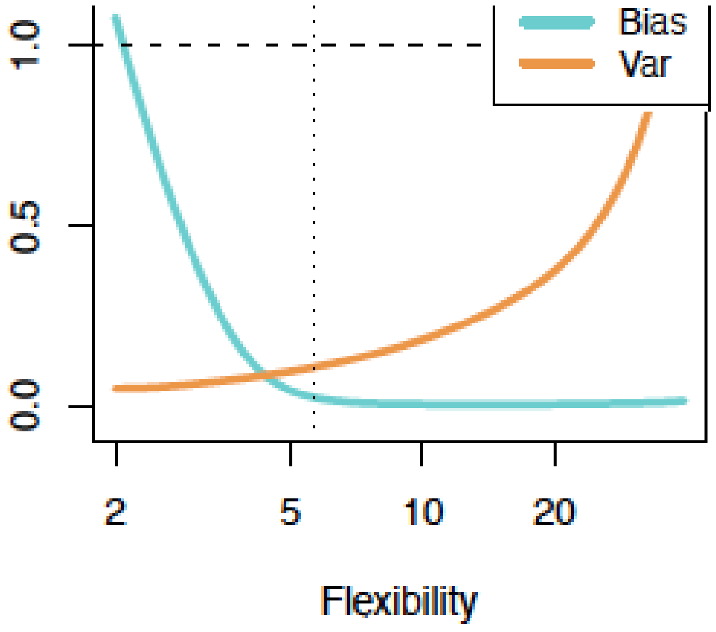
# Bias-Variance Trade-Off (cont'd)




Moderately complex $f$.

- ❑ For low flexibility,
  - ❑ Bias is very high, leading to high MSE. Underfitting.
  - ❑ Assuming enough data, variance is low.
- ❑ As flexibility increases to the best spot (complexity of true *f*),
  - ❑ Bias drops drastically.
  - ❑ Variance starts increasing but remains relatively low.
  - ❑ Following the bias, MSE drops drastically.
- ❑ As flexibility increases excessively,
  - ❑ Given data are not enough any more to sustain low variance. Model too flexible for the given amount of data. Variance increases. Overfitting.
  - ❑ Bias remains low.
  - ❑ Following the variance, MSE increases.

# Bias-Variance Trade-Off (cont'd)




- Figure expresses what is known as Bias-Variance Trade-Off.
- We want to be about where Bias and Variance meet.
- But we do not have the Bias/Variance curves when we design/train the model...
  - This is why we wish we have a good guess of what the true *f* complexity is.