

# Lecture 5:

## Bias-Variance decomposition

## Train-Validation-Test framework

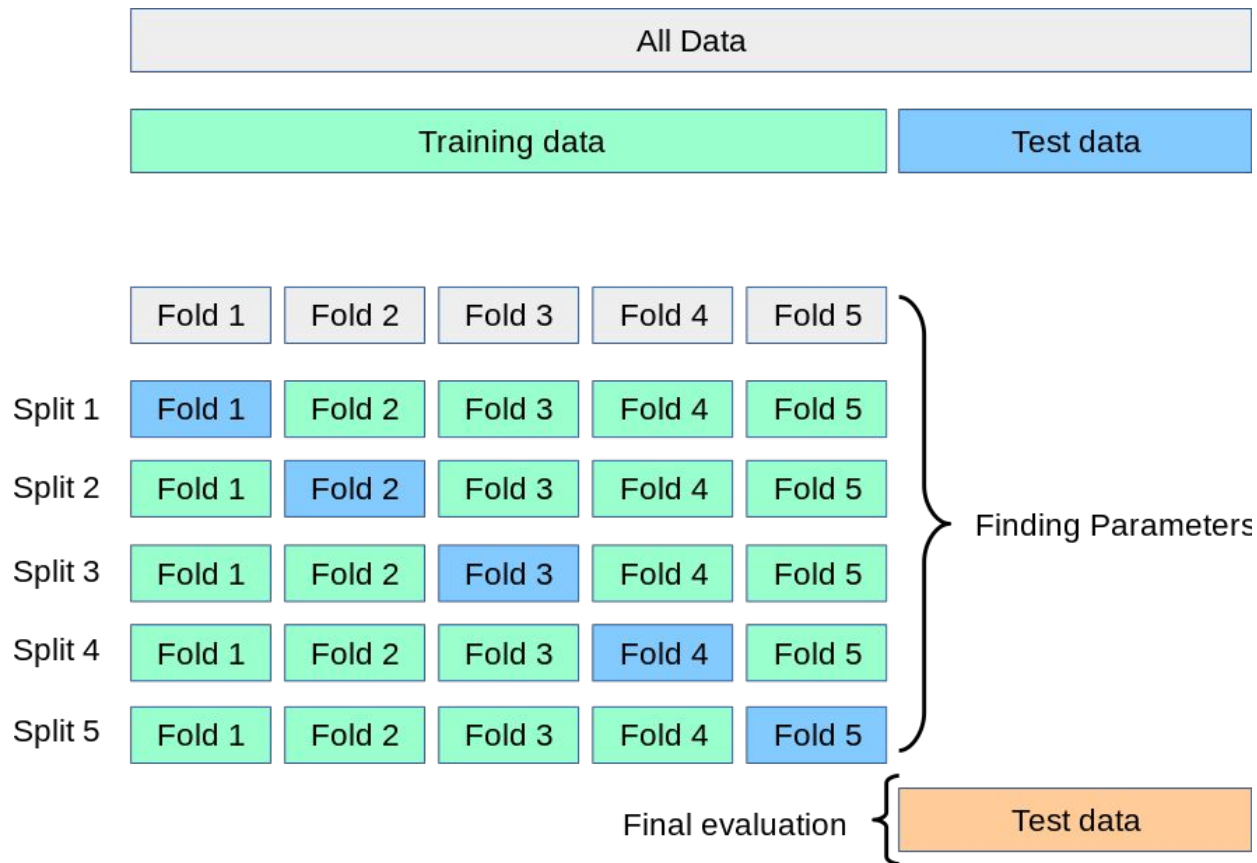
Iurii Efimov

MIPT, 2020

1. Validation strategies
2. Bias-variance decomposition intuition
3. Optional: bias-variance derivation for MSE

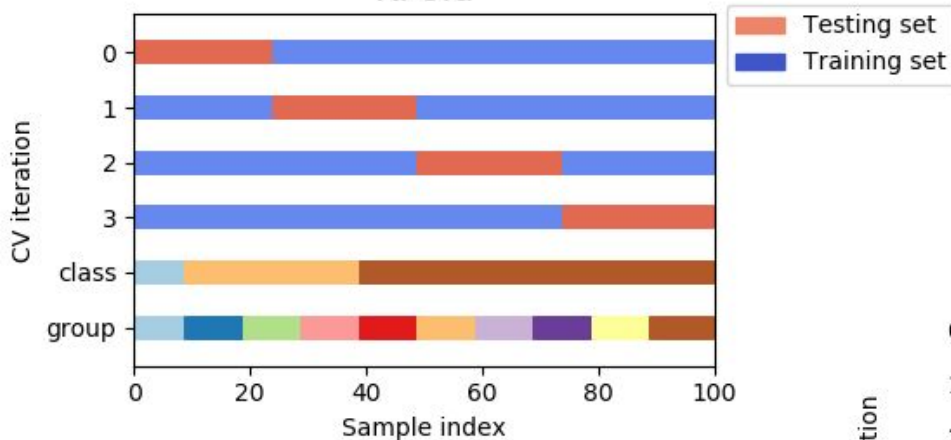
# Validation strategies

# Validation strategies

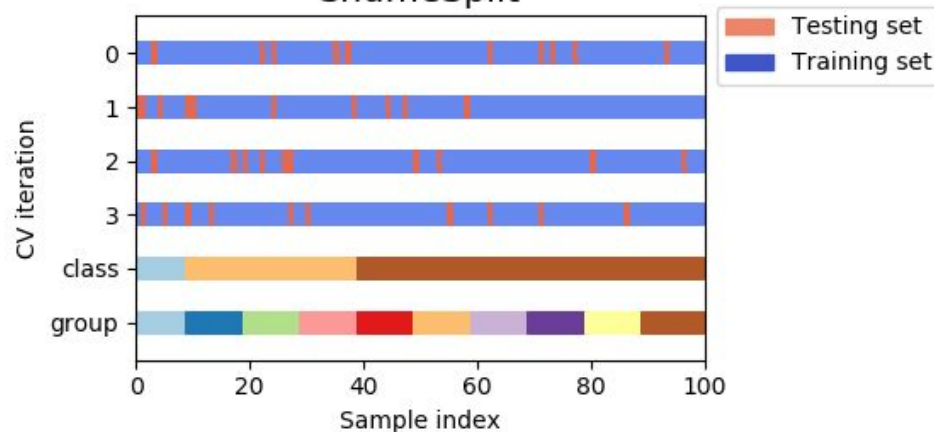


# Validation strategies

KFold

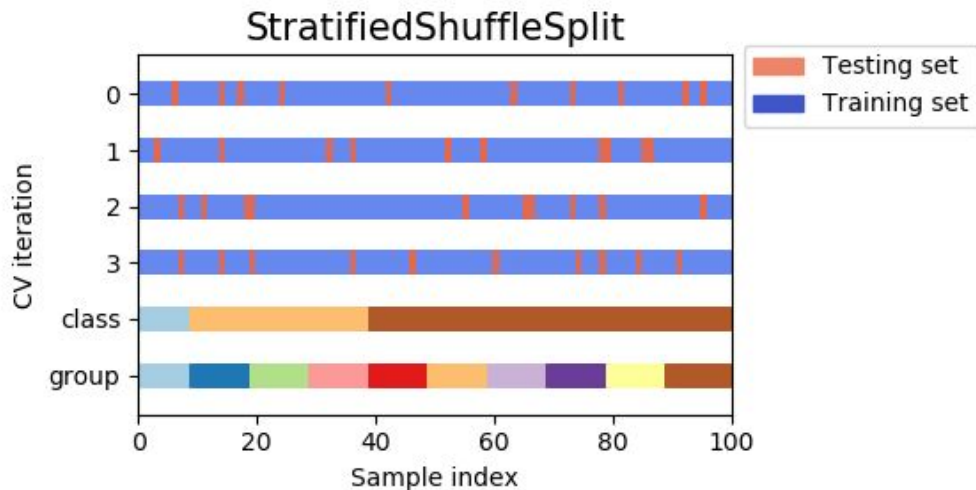
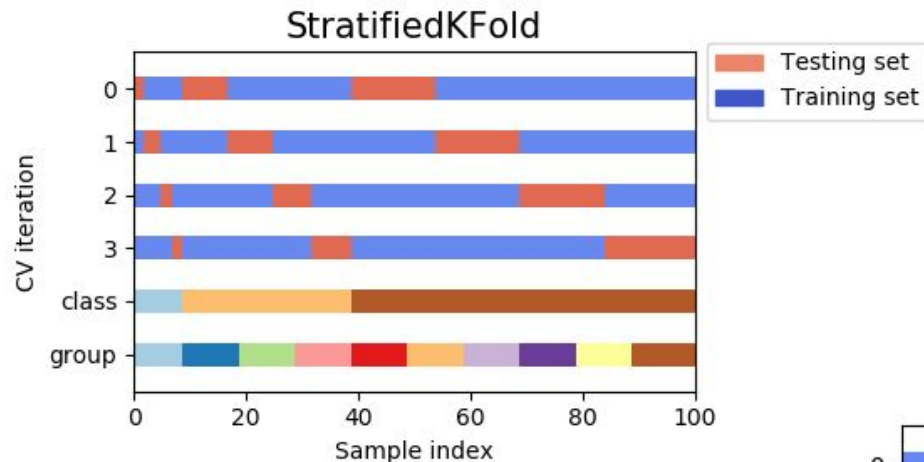


ShuffleSplit

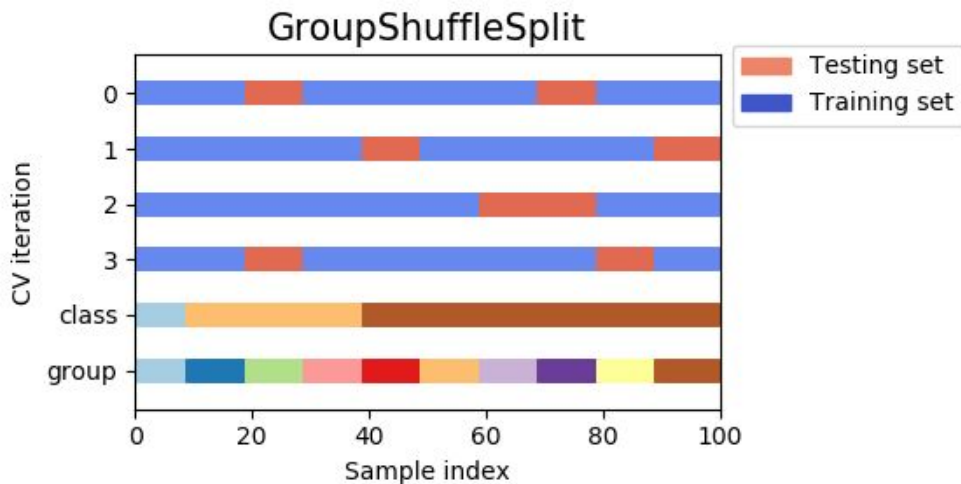
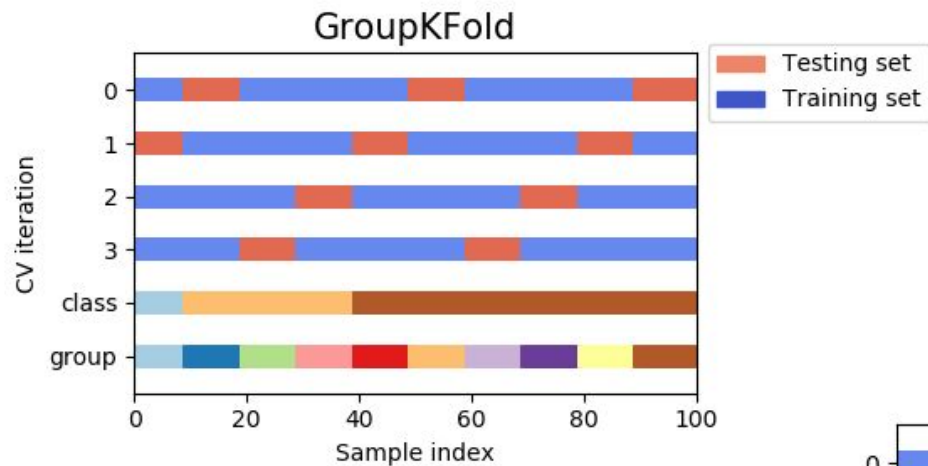


Special case: Leave One Out (LOO) - good for tiny datasets

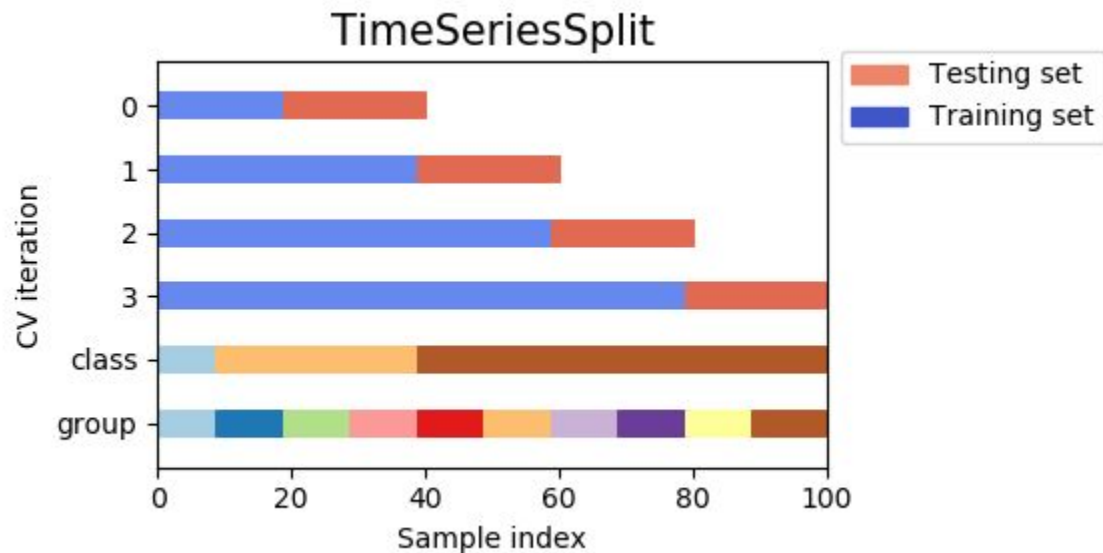
# Validation strategies



# Validation strategies



# Special case: time series

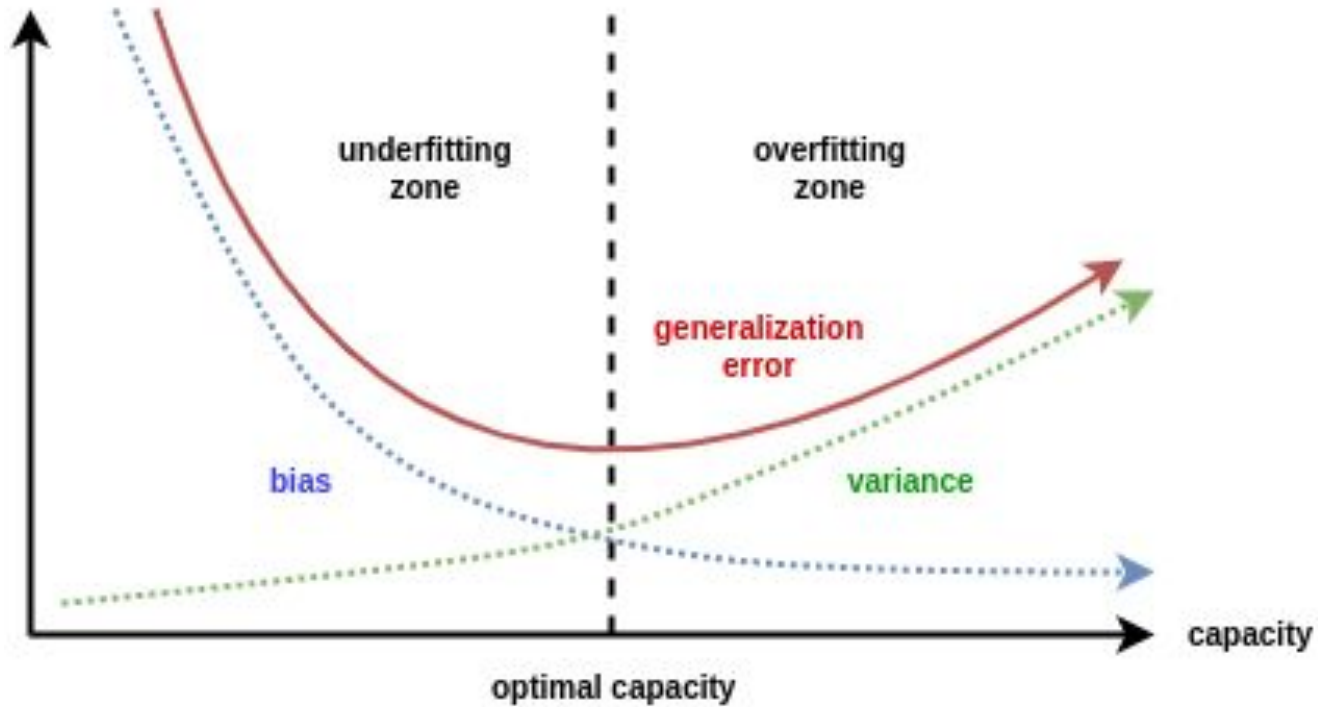


Never use `train_test_split` in this case!



Optional:  
Bias-variance decomposition intuitions

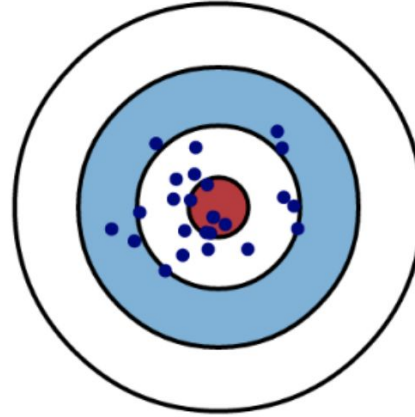
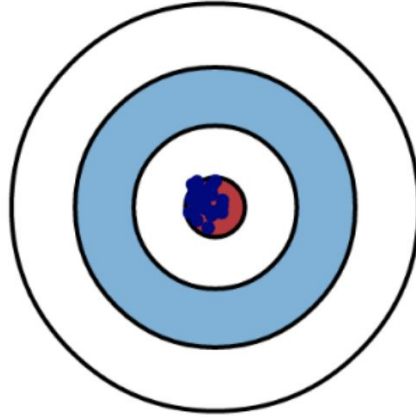
# Bias-variance tradeoff



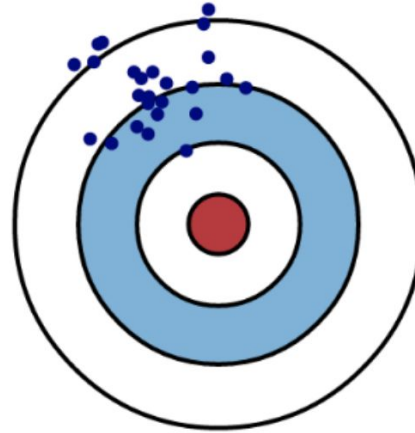
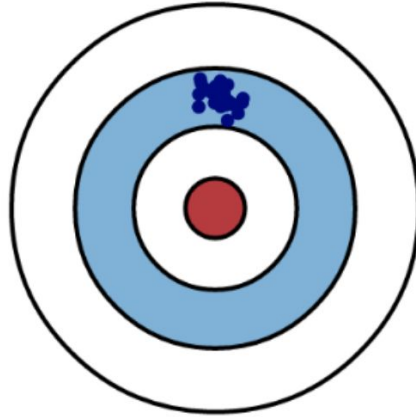
Low Variance

High Variance

Low Bias



High Bias



Optional:  
Bias-variance decomposition derivation

# Bias-variance decomposition

The dataset  $X = (x_i, y_i)_{i=1}^{\ell}$  with  $y_i \in \mathbb{R}$  for regression problem.

# Bias-variance decomposition

The dataset  $X = (x_i, y_i)_{i=1}^{\ell}$  with  $y_i \in \mathbb{R}$  for regression problem.

Denote loss function  $L(y, a) = (y - a(x))^2$ .

# Bias-variance decomposition

The dataset  $X = (x_i, y_i)_{i=1}^{\ell}$  with  $y_i \in \mathbb{R}$  for regression problem.

Denote loss function  $L(y, a) = (y - a(x))^2$ .

The corresponding risk estimation is

$$R(a) = \mathbb{E}_{x,y} \left[ (y - a(x))^2 \right] = \int_{\mathbb{X}} \int_{\mathbb{Y}} p(x, y) (y - a(x))^2 dx dy.$$

# Bias-variance decomposition

Let's show that  $a_*(x) = \mathbb{E}[y | x] = \int_{\mathbb{Y}} yp(y | x)dy$



# Bias-variance decomposition

Let's show that  $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} yp(y \mid x)dy = \arg \min_a R(a).$

# Bias-variance decomposition

Let's show that  $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} yp(y \mid x)dy = \arg \min_a R(a).$

$$L(y, a(x)) = (y - a(x))^2$$

# Bias-variance decomposition

Let's show that  $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} yp(y \mid x)dy = \arg \min_a R(a).$

$$L(y, a(x)) = (y - a(x))^2 = (y - \mathbb{E}(y \mid x) + \mathbb{E}(y \mid x) - a(x))^2 =$$

# Bias-variance decomposition

Let's show that  $a_*(x) = \mathbb{E}[y | x] = \int_{\mathbb{Y}} yp(y | x)dy = \arg \min_a R(a).$

$$\begin{aligned} L(y, a(x)) &= (y - a(x))^2 = (y - \mathbb{E}(y | x) + \mathbb{E}(y | x) - a(x))^2 = \\ &= (y - \mathbb{E}(y | x))^2 + 2(y - \mathbb{E}(y | x))(\mathbb{E}(y | x) - a(x)) + (\mathbb{E}(y | x) - a(x))^2. \end{aligned}$$

Let's return to the risk estimation:

# Bias-variance decomposition

Let's show that  $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} yp(y \mid x)dy = \arg \min_a R(a).$

$$\begin{aligned} L(y, a(x)) &= (y - a(x))^2 = (y - \mathbb{E}(y \mid x) + \mathbb{E}(y \mid x) - a(x))^2 = \\ &= (y - \mathbb{E}(y \mid x))^2 + 2(y - \mathbb{E}(y \mid x))(\mathbb{E}(y \mid x) - a(x)) + (\mathbb{E}(y \mid x) - a(x))^2. \end{aligned}$$

Let's return to the risk estimation:

$$R(a) = \mathbb{E}_{x,y} L(y, a(x))$$

# Bias-variance decomposition

Let's show that  $a_*(x) = \mathbb{E}[y | x] = \int_{\mathbb{Y}} yp(y | x)dy = \arg \min_a R(a).$

$$\begin{aligned} L(y, a(x)) &= (y - a(x))^2 = (y - \mathbb{E}(y | x) + \mathbb{E}(y | x) - a(x))^2 = \\ &= (y - \mathbb{E}(y | x))^2 + 2(y - \mathbb{E}(y | x))(\mathbb{E}(y | x) - a(x)) + (\mathbb{E}(y | x) - a(x))^2. \end{aligned}$$

Let's return to the risk estimation:

$$\begin{aligned} R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\ &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\ &+ 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x))(\mathbb{E}(y | x) - a(x)). \end{aligned}$$

$$\begin{aligned}
 R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
 &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\
 &\quad + 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)).
 \end{aligned}$$

Focus on the last term:

$$\begin{aligned}
 R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
 &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\
 &\quad + 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)).
 \end{aligned}$$

Focus on the last term:


$$\mathbb{E}_x \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)) \mid x \right] =$$



$$\begin{aligned}
 R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
 &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\
 &\quad + 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)).
 \end{aligned}$$

Focus on the last term:

Does not depend on y




$$\mathbb{E}_x \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)) \mid x \right] =$$

$$\begin{aligned}
 R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
 &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\
 &\quad + 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)).
 \end{aligned}$$

Focus on the last term:

Does not depend on y



$$\begin{aligned}
 \mathbb{E}_x \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)) \mid x \right] &= \\
 = \mathbb{E}_x \left( (\mathbb{E}(y | x) - a(x)) \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) \mid x \right] \right) &=
 \end{aligned}$$

$$\begin{aligned}
 R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
 &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\
 &\quad + 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)).
 \end{aligned}$$

Focus on the last term:

$$\begin{aligned}
 \mathbb{E}_x \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)) \mid x \right] &= \\
 = \mathbb{E}_x \left( (\mathbb{E}(y | x) - a(x)) \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) \mid x \right] \right) &= \\
 = \mathbb{E}_x \left( (\mathbb{E}(y | x) - a(x)) (\mathbb{E}(y | x) - \mathbb{E}(y | x)) \right) &=
 \end{aligned}$$

$$\begin{aligned}
 R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
 &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\
 &\quad + 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)).
 \end{aligned}$$

Focus on the last term:

$$\begin{aligned}
 &\mathbb{E}_x \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)) \mid x \right] = \\
 &= \mathbb{E}_x \left( (\mathbb{E}(y | x) - a(x)) \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) \mid x \right] \right) = \\
 &= \mathbb{E}_x \left( (\mathbb{E}(y | x) - a(x)) (\mathbb{E}(y | x) - \mathbb{E}(y | x)) \right) = \\
 &= 0
 \end{aligned}$$

$$\begin{aligned}
 R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
 &= \mathbb{E}_{x,y} (y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y} (\mathbb{E}(y | x) - a(x))^2 + \\
 &\quad + 2\mathbb{E}_{x,y} (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)).
 \end{aligned}$$

Focus on the last term:

0

$$\begin{aligned}
 \mathbb{E}_x \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) (\mathbb{E}(y | x) - a(x)) \mid x \right] &= \\
 = \mathbb{E}_x \left( (\mathbb{E}(y | x) - a(x)) \mathbb{E}_y \left[ (y - \mathbb{E}(y | x)) \mid x \right] \right) &= \\
 = \mathbb{E}_x \left( (\mathbb{E}(y | x) - a(x)) (\mathbb{E}(y | x) - \mathbb{E}(y | x)) \right) &= \\
 = 0
 \end{aligned}$$

Does not depend on  $a(x)$

So the risk takes form:

$$R(a) = \mathbb{E}_{x,y}(y - \mathbb{E}(y | x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y | x) - a(x))^2.$$

The minimum is reached when  $a(x) = \mathbb{E}(y | x)$ .

So the optimal regression model with square loss is

$$a_*(x) = \mathbb{E}(y | x) = \int_{\mathbb{Y}} yp(y | x)dy.$$

Denote  $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \rightarrow \mathcal{A}$ , where  $\mathcal{A}$  is some family of algorithms.

Denote  $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \rightarrow \mathcal{A}$ , where  $\mathcal{A}$  is some family of algorithms.

So  $L(\mu) = \mathbb{E}_X \left[ \mathbb{E}_{x,y} \left[ (y - \mu(X)(x))^2 \right] \right]$ , where  $X$  dataset.



Denote  $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \rightarrow \mathcal{A}$ , where  $\mathcal{A}$  is some family of algorithms.

So  $L(\mu) = \mathbb{E}_X \left[ \mathbb{E}_{x,y} \left[ (y - \mu(X)(x))^2 \right] \right]$ , where  $X$  dataset.

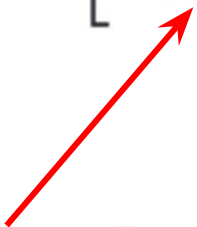
If  $X$  is fixed, then

$$\mathbb{E}_{x,y} \left[ (y - \mu(X))^2 \right] = \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right].$$

Denote  $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \rightarrow \mathcal{A}$ , where  $\mathcal{A}$  is some family of algorithms.

So  $L(\mu) = \mathbb{E}_X \left[ \mathbb{E}_{x,y} \left[ (y - \mu(X)(x))^2 \right] \right]$ , where  $X$  dataset.

If  $X$  is fixed, then


$$\mathbb{E}_{x,y} \left[ (y - \mu(X)(x))^2 \right] = \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X)(x))^2 \right].$$

Let's combine the latter equations:

Denote  $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \rightarrow \mathcal{A}$ , where  $\mathcal{A}$  is some family of algorithms.

So  $L(\mu) = \mathbb{E}_X \left[ \mathbb{E}_{x,y} \left[ (y - \mu(X)(x))^2 \right] \right]$ , where  $X$  dataset.

If  $X$  is fixed, then

$$\mathbb{E}_{x,y} \left[ (y - \mu(X))^2 \right] = \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right].$$

Let's combine the latter equations:

$$L(\mu) = \mathbb{E}_X \left[ \underbrace{\mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right]}_{\text{Does not depend on } X} + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right]$$

$$L(\mu) = \mathbb{E}_X \left[ \underbrace{\mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right]}_{\text{Does not depend on X}} + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] =$$

Does not depend on X

$$L(\mu) = \mathbb{E}_X \left[ \underbrace{\mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right]}_{\text{Does not depend on X}} + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] =$$

Does not depend on X

$$= \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right].$$

$$\begin{aligned} L(\mu) &= \mathbb{E}_X \left[ \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] = \\ &= \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right]. \end{aligned}$$

Focus on the second term:

$$\begin{aligned}
 L(\mu) &= \mathbb{E}_X \left[ \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] = \\
 &= \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right].
 \end{aligned}$$

Focus on the second term:

$$\mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] =$$

$$\begin{aligned}
 L(\mu) &= \mathbb{E}_X \left[ \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] = \\
 &= \mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right].
 \end{aligned}$$

Focus on the second term:

$$\begin{aligned}
 \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] &= \\
 &= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)] + \mathbb{E}_X [\mu(X)] - \mu(X))^2 \right] \right]
 \end{aligned}$$



$$\begin{aligned}\mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mu(X) \right)^2 \right] \right] &= \\ &= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)] + \mathbb{E}_X [\mu(X)] - \mu(X) \right)^2 \right] \right] =\end{aligned}$$

$$\begin{aligned}
& \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mu(X) \right)^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)] + \mathbb{E}_X [\mu(X)] - \mu(X) \right)^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \underbrace{\mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)] \right)^2 \right]} + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}_X [\mu(X)] - \mu(X) \right)^2 \right] \right] + \right. \\
&\quad \left. + 2 \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)] \right) \left( \mathbb{E}_X [\mu(X)] - \mu(X) \right) \right] \right].
\end{aligned}$$

$$\begin{aligned}
& \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)] + \mathbb{E}_X [\mu(X)] - \mu(X))^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \underbrace{\mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)])^2 \right]}_{\text{Does not depend on X}} \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}_X [\mu(X)] - \mu(X))^2 \right] \right] + \\
&\quad + 2\mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)]) (\mathbb{E}_X [\mu(X)] - \mu(X)) \right] \right].
\end{aligned}$$

$$\begin{aligned}
& \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y | x] - \mu(X) \right)^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)] + \mathbb{E}_X [\mu(X)] - \mu(X) \right)^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \underbrace{\mathbb{E}_X \left[ \left( \mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)] \right)^2 \right]}_{\text{Does not depend on } X} \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}_X [\mu(X)] - \mu(X) \right)^2 \right] \right] + \\
&\quad + 2 \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)] \right) \left( \mathbb{E}_X [\mu(X)] - \mu(X) \right) \right] \right].
\end{aligned}$$

Just a bit further, we are almost there

$$\begin{aligned}
& \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)] + \mathbb{E}_X [\mu(X)] - \mu(X))^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)])^2 \right] \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}_X [\mu(X)] - \mu(X))^2 \right] \right] + \\
&\quad + 2\mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X [\mu(X)]) (\mathbb{E}_X [\mu(X)] - \mu(X)) \right] \right].
\end{aligned}$$

Focus on this term

$$\mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)] \right) \left( \mathbb{E}_X [\mu(X)] - \mu(X) \right) \right] =$$

$$\begin{aligned}\mathbb{E}_X \left[ \left( \mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)] \right) \left( \mathbb{E}_X [\mu(X)] - \mu(X) \right) \right] &= \\ &= \left( \mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)] \right) \mathbb{E}_X \left[ \mathbb{E}_X [\mu(X)] - \mu(X) \right] =\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_X \left[ (\mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)]) (\mathbb{E}_X [\mu(X)] - \mu(X)) \right] &= \\
&= (\mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)]) \mathbb{E}_X [\mathbb{E}_X [\mu(X)] - \mu(X)] = \\
&= (\mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)]) [\mathbb{E}_X [\mu(X)] - \mathbb{E}_X [\mu(X)]] =
\end{aligned}$$



$$\begin{aligned}
\mathbb{E}_X \left[ (\mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)]) (\mathbb{E}_X [\mu(X)] - \mu(X)) \right] &= \\
&= (\mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)]) \mathbb{E}_X [\mathbb{E}_X [\mu(X)] - \mu(X)] = \\
&= (\mathbb{E}[y \mid x] - \mathbb{E}_X [\mu(X)]) [\mathbb{E}_X [\mu(X)] - \mathbb{E}_X [\mu(X)]] = \\
&= 0.
\end{aligned}$$

$$\begin{aligned}
& \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mu(X))^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X[\mu(X)] + \mathbb{E}_X[\mu(X)] - \mu(X))^2 \right] \right] = \\
&= \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X[\mu(X)])^2 \right] \right] + \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}_X[\mu(X)] - \mu(X))^2 \right] \right] + \\
&\quad + 2\mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ (\mathbb{E}[y | x] - \mathbb{E}_X[\mu(X)])(\mathbb{E}_X[\mu(X)] - \mu(X)) \right] \right].
\end{aligned}$$

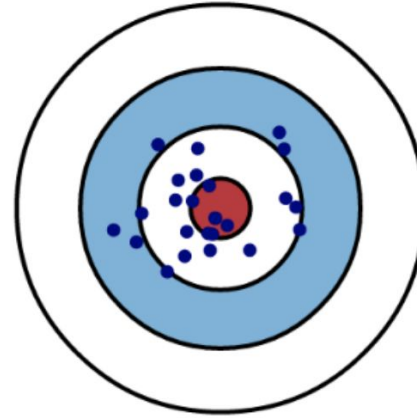
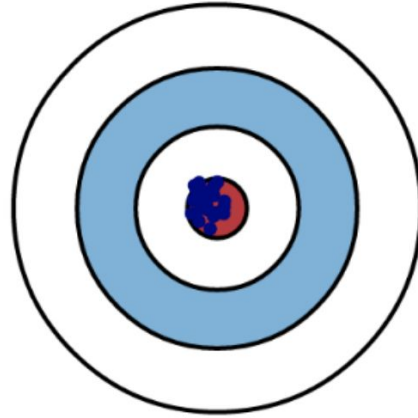
0

$$\begin{aligned}
L(\mu) = & \underbrace{\mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right]}_{\text{noise}} + \\
& + \underbrace{\mathbb{E}_x \left[ (\mathbb{E}_X [\mu(X)] - \mathbb{E}[y | x])^2 \right]}_{\text{bias}} + \underbrace{\mathbb{E}_x \left[ \mathbb{E}_X \left[ (\mu(X) - \mathbb{E}_X [\mu(X)])^2 \right] \right]}_{\text{variance}}.
\end{aligned}$$

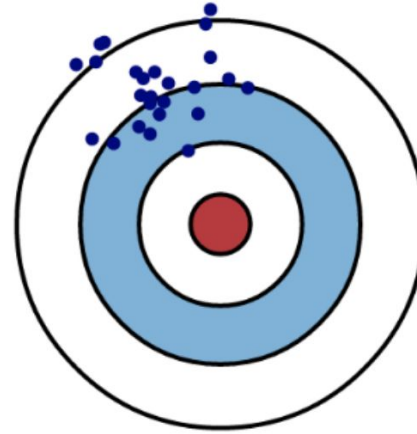
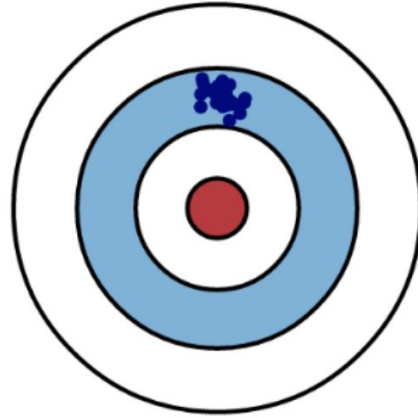
Low Variance

High Variance

Low Bias



High Bias



$$\begin{aligned}
L(\mu) = & \underbrace{\mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right]}_{\text{noise}} + \\
& + \underbrace{\mathbb{E}_x \left[ (\mathbb{E}_X [\mu(X)] - \mathbb{E}[y | x])^2 \right]}_{\text{bias}} + \underbrace{\mathbb{E}_x \left[ \mathbb{E}_X \left[ (\mu(X) - \mathbb{E}_X [\mu(X)])^2 \right] \right]}_{\text{variance}}.
\end{aligned}$$

This exact form of bias-variance decomposition is correct for square loss in regression.

However, it is much more general. See extra materials for more exotic cases.

# Bagging = Bootstrap aggregating

Denote dataset  $\tilde{X}$  bootstrapped from  $X$ .

Denote  $\mu: \tilde{\mu}(X) = \mu(\tilde{X})$ . Let  $b_n(x)$  be basic algorithm.

Denote the ensemble:

$$a_N(x) = \frac{1}{N} \sum_{n=1}^N b_n(x) = \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x).$$

$$\begin{aligned}
L(\mu) = & \underbrace{\mathbb{E}_{x,y} \left[ (y - \mathbb{E}[y | x])^2 \right]}_{\text{noise}} + \\
& + \underbrace{\mathbb{E}_x \left[ (\mathbb{E}_X [\mu(X)] - \mathbb{E}[y | x])^2 \right]}_{\text{bias}} + \underbrace{\mathbb{E}_x \left[ \mathbb{E}_X \left[ (\mu(X) - \mathbb{E}_X [\mu(X)])^2 \right] \right]}_{\text{variance}}.
\end{aligned}$$

The **bias** term takes the following form:

$$\mathbb{E}_{x,y} \left[ \left( \mathbb{E}_X \left[ \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) \right] - \mathbb{E}[y | x] \right)^2 \right] =$$



The **bias** term takes the following form:

$$\begin{aligned}\mathbb{E}_{x,y} \left[ \left( \mathbb{E}_X \left[ \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) \right] - \mathbb{E}[y | x] \right)^2 \right] &= \\ &= \mathbb{E}_{x,y} \left[ \left( \frac{1}{N} \sum_{n=1}^N \mathbb{E}_X [\tilde{\mu}(X)(x)] - \mathbb{E}[y | x] \right)^2 \right] =\end{aligned}$$

The **bias** term takes the following form:

$$\begin{aligned}\mathbb{E}_{x,y} \left[ \left( \mathbb{E}_X \left[ \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) \right] - \mathbb{E}[y | x] \right)^2 \right] &= \\ &= \mathbb{E}_{x,y} \left[ \left( \frac{1}{N} \sum_{n=1}^N \mathbb{E}_X [\tilde{\mu}(X)(x)] - \mathbb{E}[y | x] \right)^2 \right] = \\ &= \mathbb{E}_{x,y} \left[ \left( \mathbb{E}_X [\tilde{\mu}(X)(x)] - \mathbb{E}[y | x] \right)^2 \right].\end{aligned}$$

The **bias** term takes the following form:

$$\begin{aligned} \mathbb{E}_{x,y} \left[ \left( \mathbb{E}_X \left[ \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) \right] - \mathbb{E}[y | x] \right)^2 \right] &= \\ &= \mathbb{E}_{x,y} \left[ \left( \frac{1}{N} \sum_{n=1}^N \mathbb{E}_X [\tilde{\mu}(X)(x)] - \mathbb{E}[y | x] \right)^2 \right] = \\ &= \mathbb{E}_{x,y} \left[ \left( \mathbb{E}_X [\tilde{\mu}(X)(x)] - \mathbb{E}[y | x] \right)^2 \right]. \end{aligned}$$

One algorithm bias

The **variance**:  $\mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) - \mathbb{E}_X \left[ \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) \right] \right)^2 \right] \right].$

$$\begin{aligned} \left( \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) - \mathbb{E}_X \left[ \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x) \right] \right)^2 &= \\ &= \frac{1}{N^2} \left( \sum_{n=1}^N \left[ \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right] \right)^2 = \\ &= \frac{1}{N^2} \sum_{n=1}^N \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right)^2 + \\ &\quad + \frac{1}{N^2} \sum_{n_1 \neq n_2} \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \end{aligned}$$

The **variance**:

$$\begin{aligned}
 & \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \frac{1}{N^2} \sum_{n=1}^N \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right)^2 + \right. \right. \\
 & \quad \left. \left. + \frac{1}{N^2} \sum_{n_1 \neq n_2} \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \right] \right] = \\
 & = \frac{1}{N^2} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \sum_{n=1}^N \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right)^2 \right] \right] + \\
 & \quad + \frac{1}{N^2} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \sum_{n_1 \neq n_2} \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \times \right. \right. \\
 & \quad \quad \left. \left. \times \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \right] \right] = \text{One algorithm} \\
 & = \frac{1}{N} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right)^2 \right] \right] + \text{variance} * 1/N \\
 & \quad + \frac{N(N-1)}{N^2} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \times \right. \right. \\
 & \quad \quad \left. \left. \times \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \right] \right]
 \end{aligned}$$

The **variance**:

$$\begin{aligned}
 & \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \frac{1}{N^2} \sum_{n=1}^N \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right)^2 + \right. \right. \\
 & \quad \left. \left. + \frac{1}{N^2} \sum_{n_1 \neq n_2} \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \right] \right] = \\
 & = \frac{1}{N^2} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \sum_{n=1}^N \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right)^2 \right] \right] + \\
 & \quad + \frac{1}{N^2} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \sum_{n_1 \neq n_2} \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \times \right. \right. \\
 & \quad \quad \left. \left. \times \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \right] \right] = \text{One algorithm} \\
 & = \frac{1}{N} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right)^2 \right] \right] + \text{variance} * 1/N \\
 & \quad + \frac{N(N-1)}{N^2} \mathbb{E}_{x,y} \left[ \mathbb{E}_X \left[ \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \times \right. \right. \\
 & \quad \quad \left. \left. \times \left( \tilde{\mu}(X)(x) - \mathbb{E}_X [\tilde{\mu}(X)(x)] \right) \right] \right] \quad \text{Basic algorithms covariance}
 \end{aligned}$$

