Linear regression

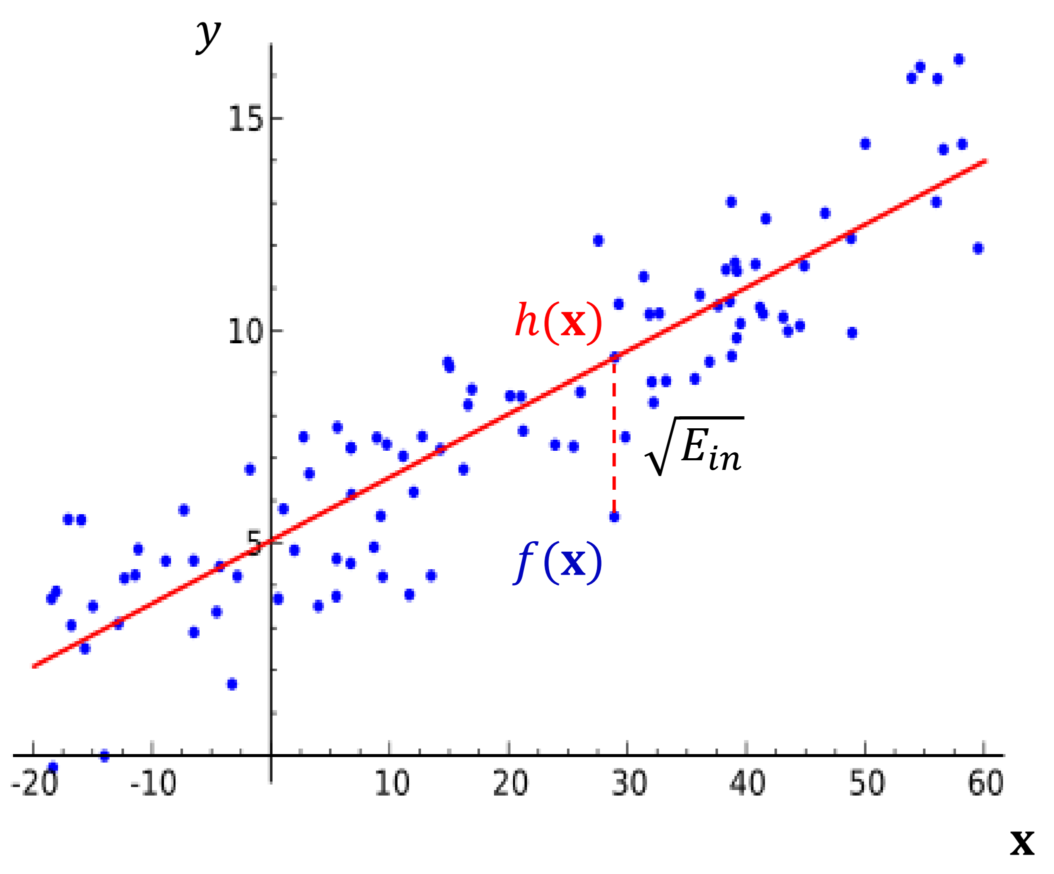
Linear regression

$$E_{out}(h, \mathbf{x}) = E\big(h(\mathbf{x}) - f(\mathbf{x})\big)^2$$

$$E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i\right)^2 = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

$$L_{linear}(\mathbf{w}) = \sum_{i=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i\right)^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}_1^{\mathrm{T}} & - \\ - & \mathbf{x}_2^{\mathrm{T}} & - \\ & \vdots & \\ - & \mathbf{x}_N^{\mathrm{T}} & - \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$



Linear regression

$$L_{linear}(\mathbf{w}) = \left|\left|\mathbf{X}\mathbf{w} - \mathbf{y}\right|\right|_2^2$$

$$\nabla L_{linear}(\mathbf{w}) = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) = 0$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \mathbf{w} = \mathbf{X}^{\dagger}\mathbf{y}$$

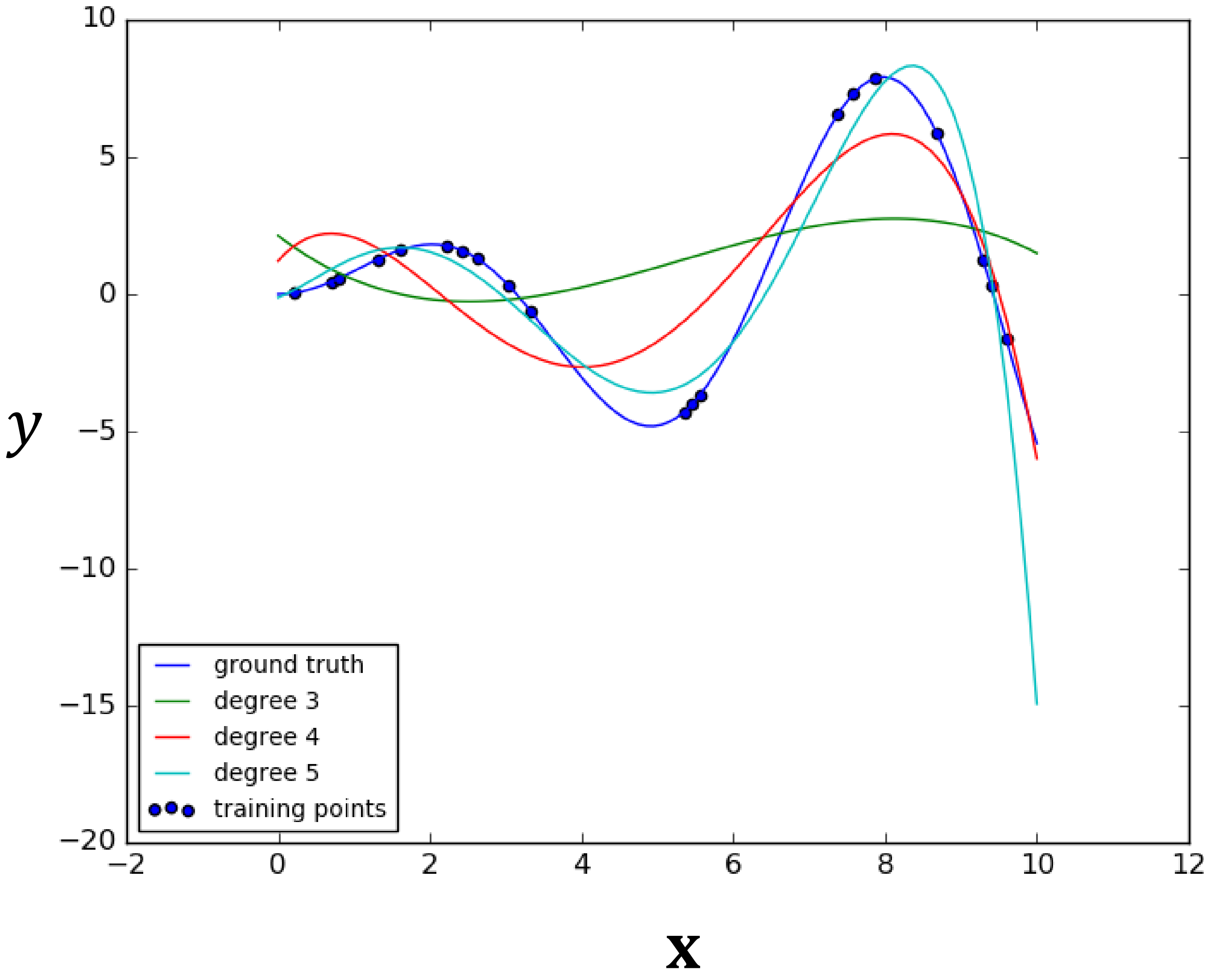
Polynomial regression

$$X \rightarrow Z$$

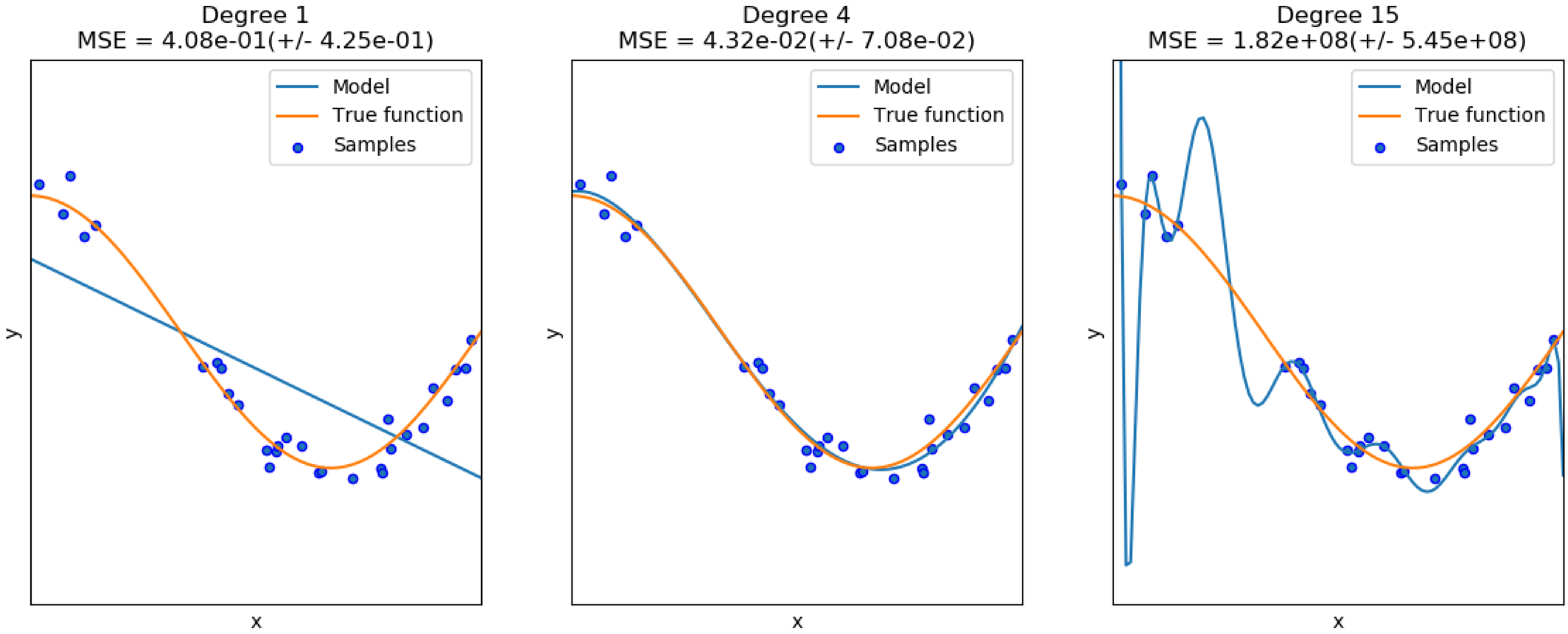
$$x \rightarrow [1, x, x^2]$$

$$[x_1, x_2] \rightarrow [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]$$

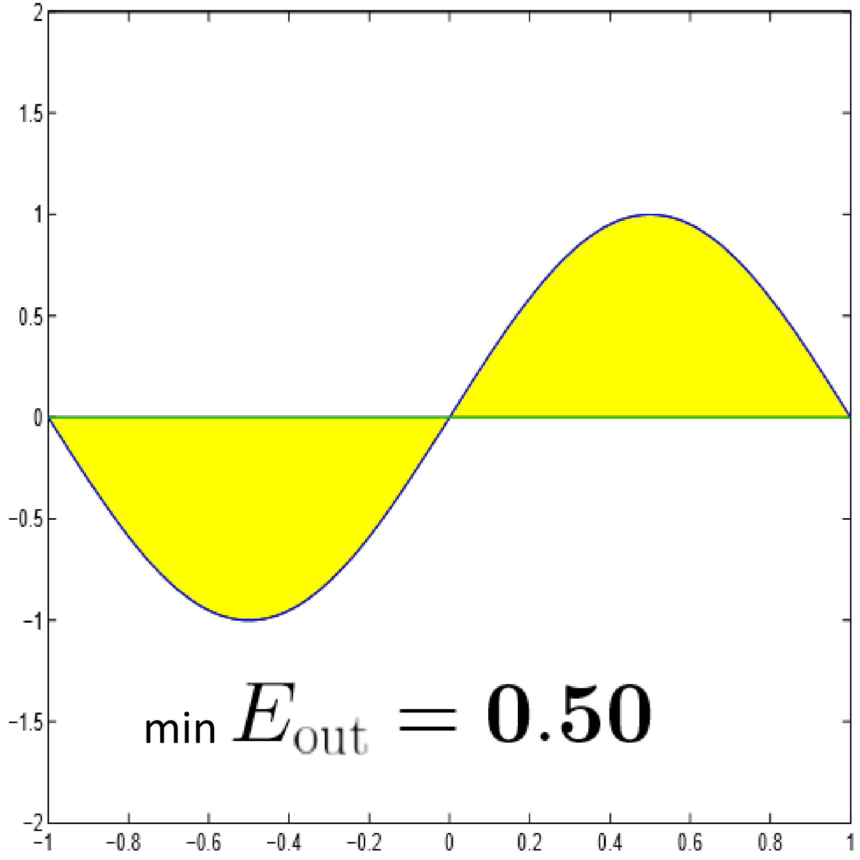
etc…

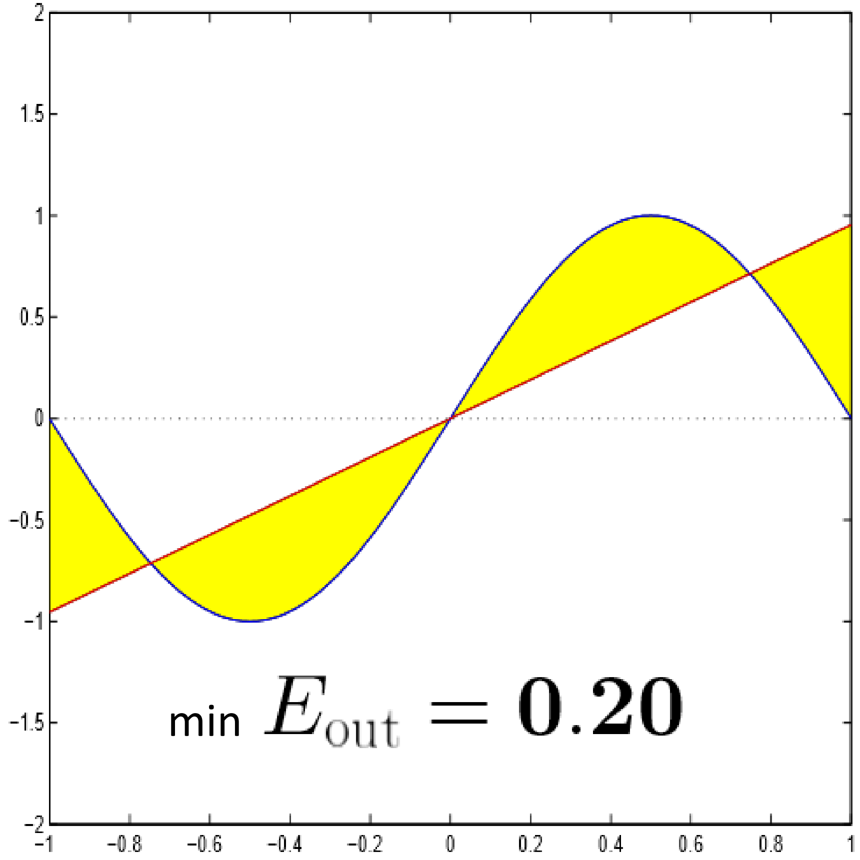


Polynomial regression

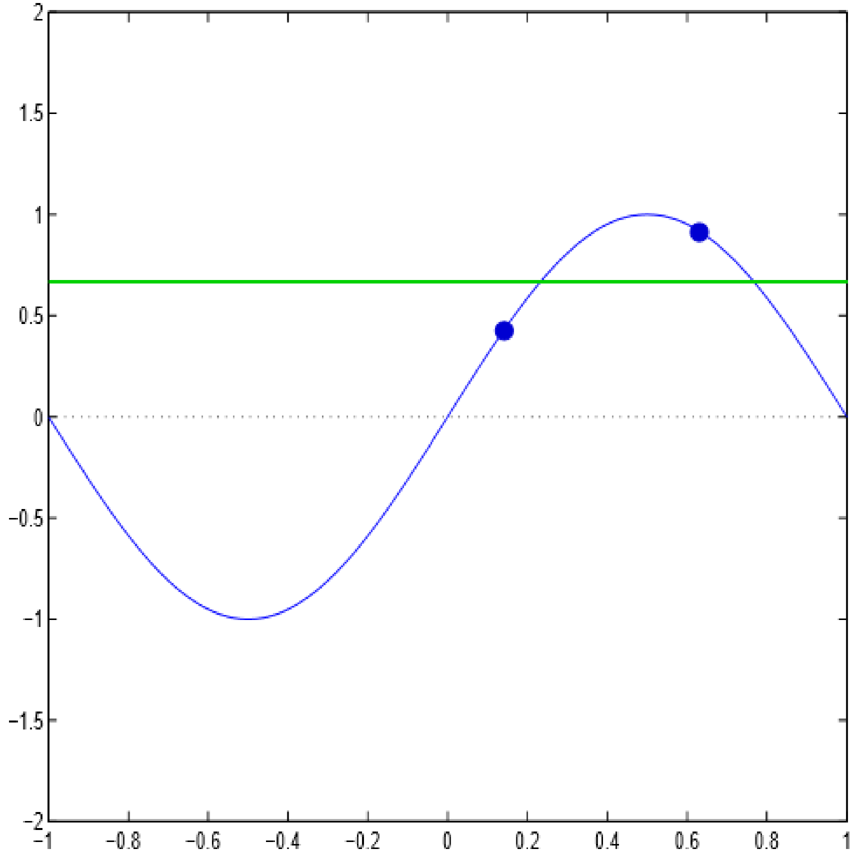


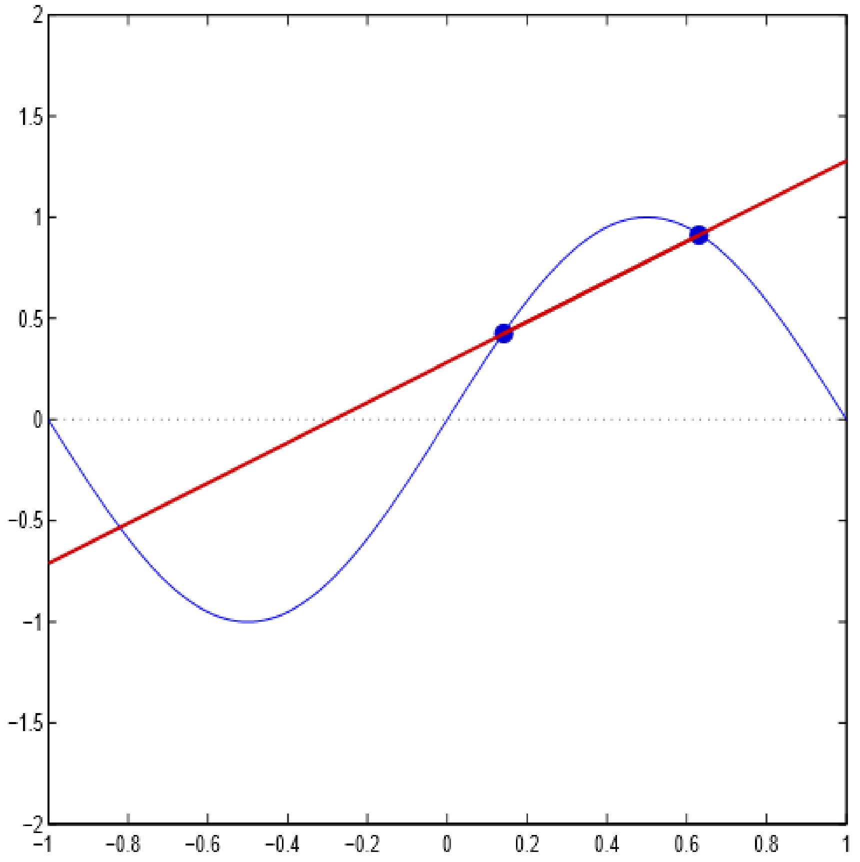
Sine target function



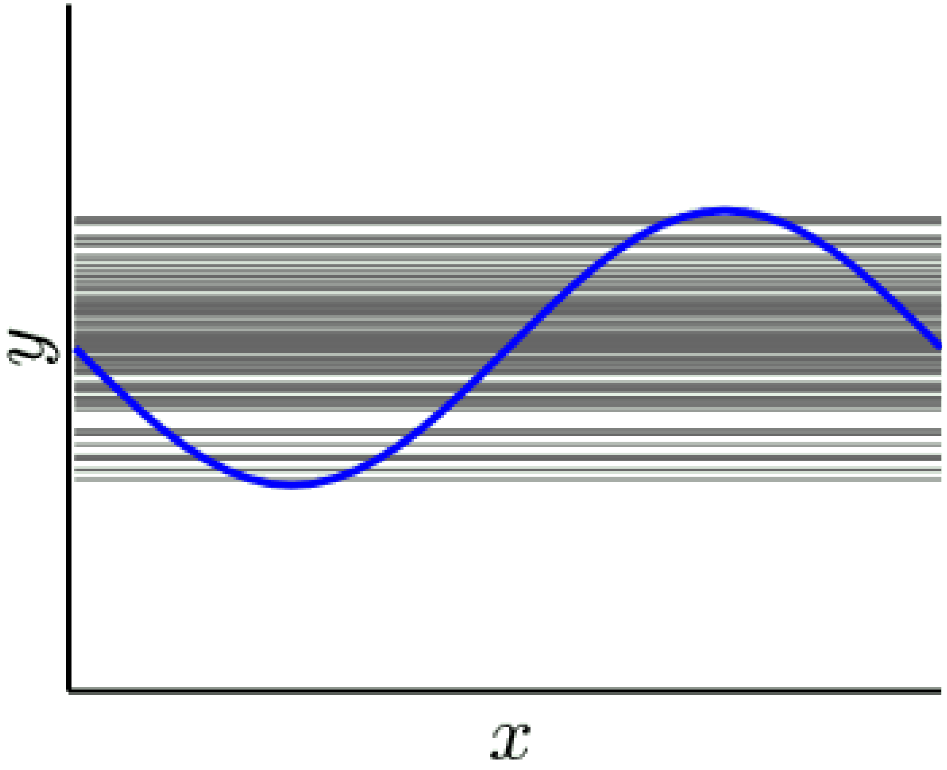


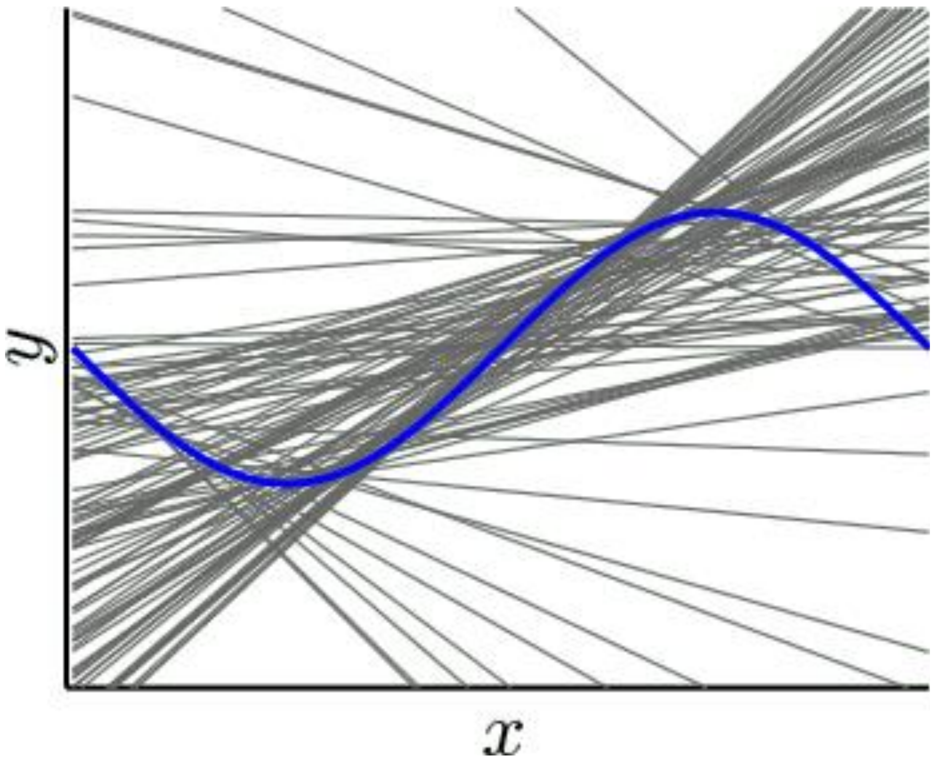
Sine target function





Sine target function





Bias and Variance

$$E_{out}(h^D) = \mathbb{E}_{\mathbf{x}}\left[\left(h^D(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \quad \color{blue}{D - data}$$

$$\mathbb{E}_D[E_{out}(h^D)] = \mathbb{E}_D\left[\mathbb{E}_{\mathbf{x}}\left[\left(h^D(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_D\left[\left(h^D(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$$

$$\bar{h}(\mathbf{x}) = \mathbb{E}_D[h^D(\mathbf{x})]$$ mean hypothesis

$$\mathbb{E}_D\left[\left(h^D(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \mathbb{E}_D\left[\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x}) + \bar{h}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] =$$

$$= \mathbb{E}_D\left[\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)^2 + \left(\bar{h}(\mathbf{x}) - f(\mathbf{x})\right)^2 + 2\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x}) - f(\mathbf{x})\right)\right] =$$

$$= \mathbb{E}_D\left[\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)^2\right] + \left(\bar{h}(\mathbf{x}) - f(\mathbf{x})\right)^2$$

Bias and Variance

$$\mathbb{E}_D\left[\left(h^D(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \mathbb{E}_D\left[\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)^2\right] + \left(\bar{h}(\mathbf{x}) - f(\mathbf{x})\right)^2$$

$$\mathbb{E}_D[E_{out}(h^D)] = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[\left(h^D(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)^2\right] + \left(\bar{h}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] =$$

$$= \mathbb{E}_{\mathbf{X}}[variance(\mathbf{x}) + bias(\mathbf{x})] = bias + variance$$

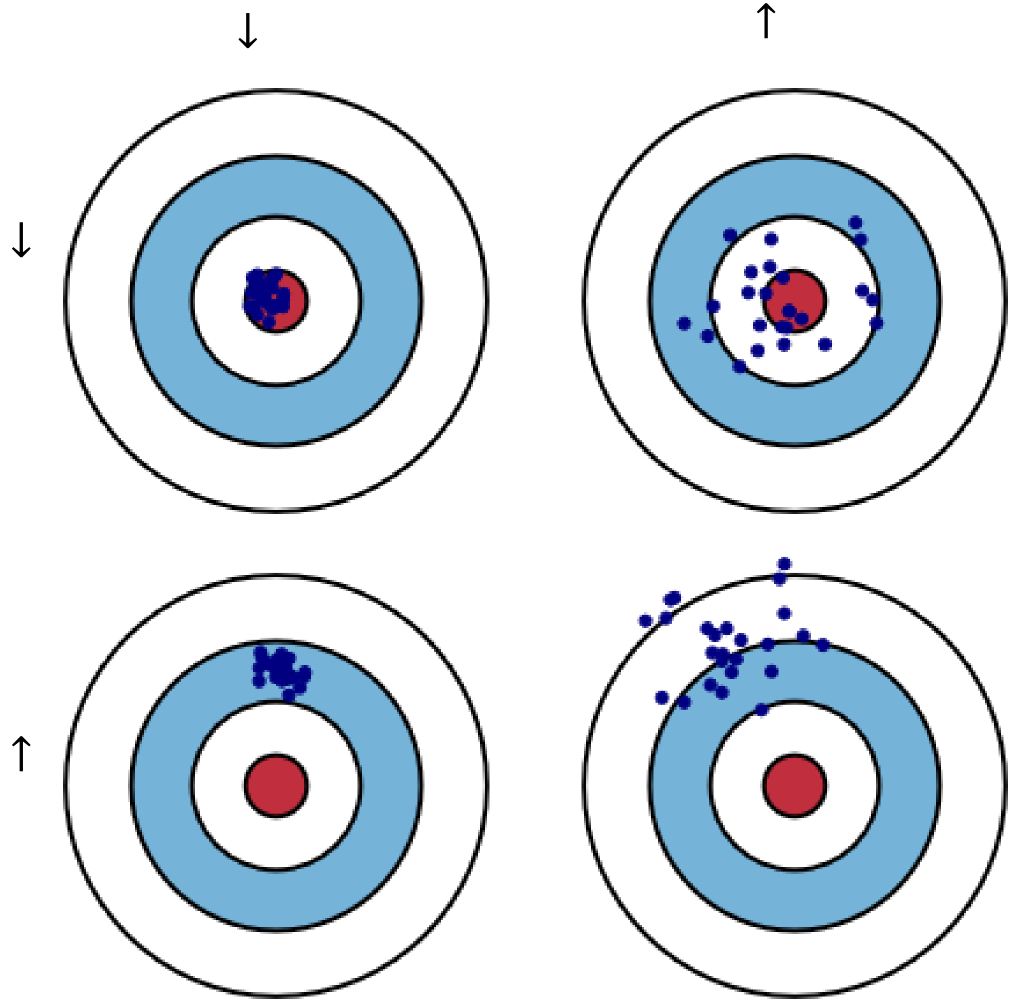
$$bias = \mathbb{E}_{\mathbf{X}}\left[\left(\bar{h}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

$$variance = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)^2\right]\right]$$

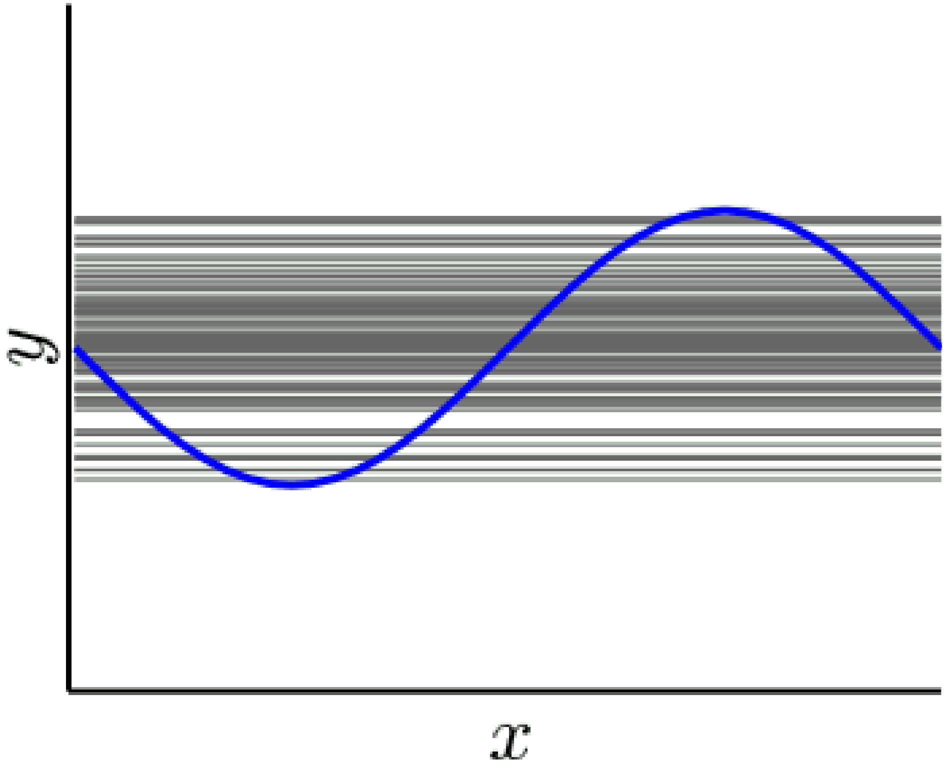
Bias and Variance

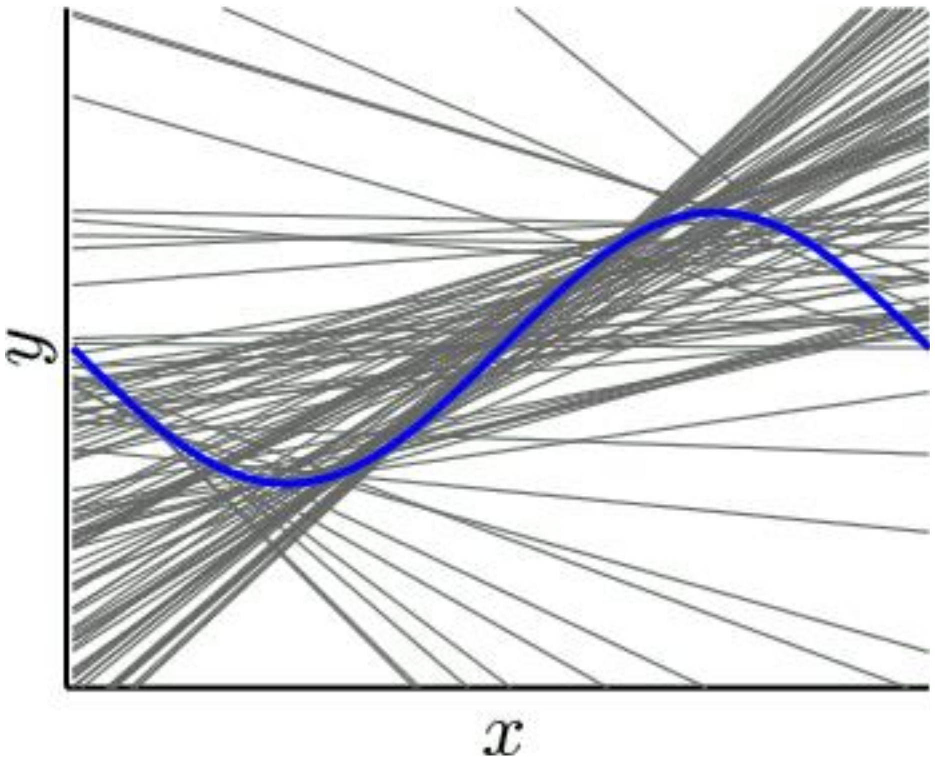
$$variance = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_D\left[\left(h^D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)^2\right]\right]$$

$$bias = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

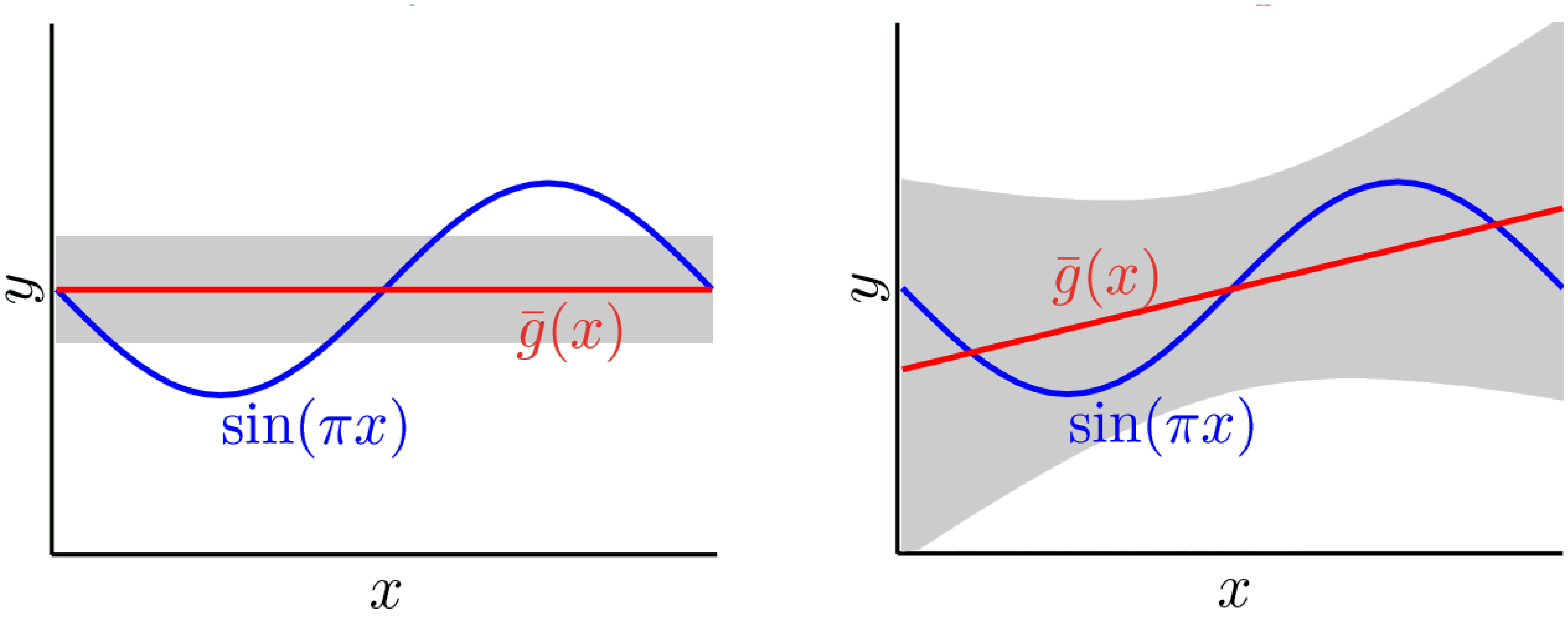


Sine target function





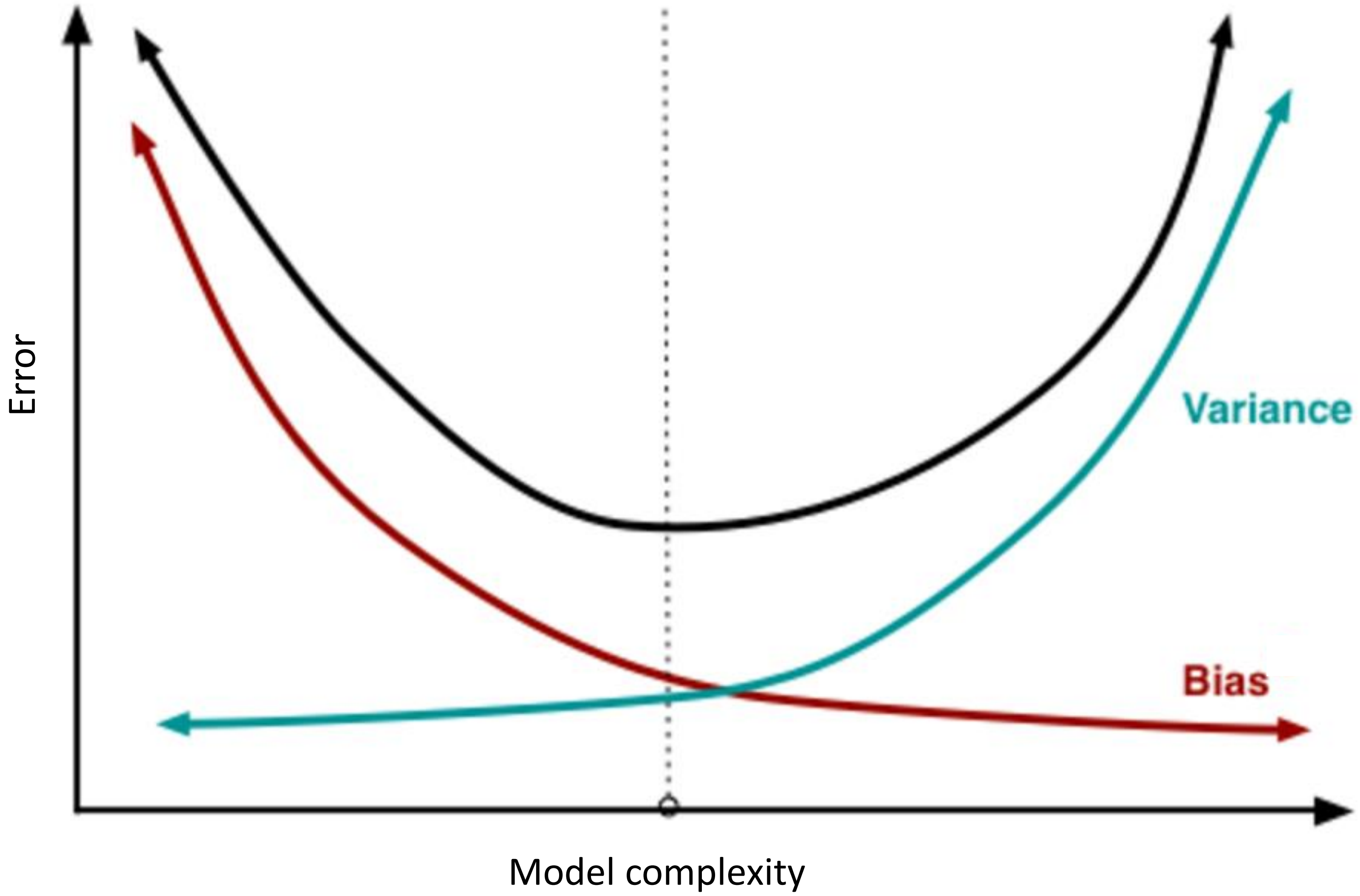
Sine target function



bias $=$ **0.50** var $=$ **0.25**

bias $=$ **0.21** var $=$ **1.69**

Bias and Variance



Regularization

L2 regularization (Ridge regression)

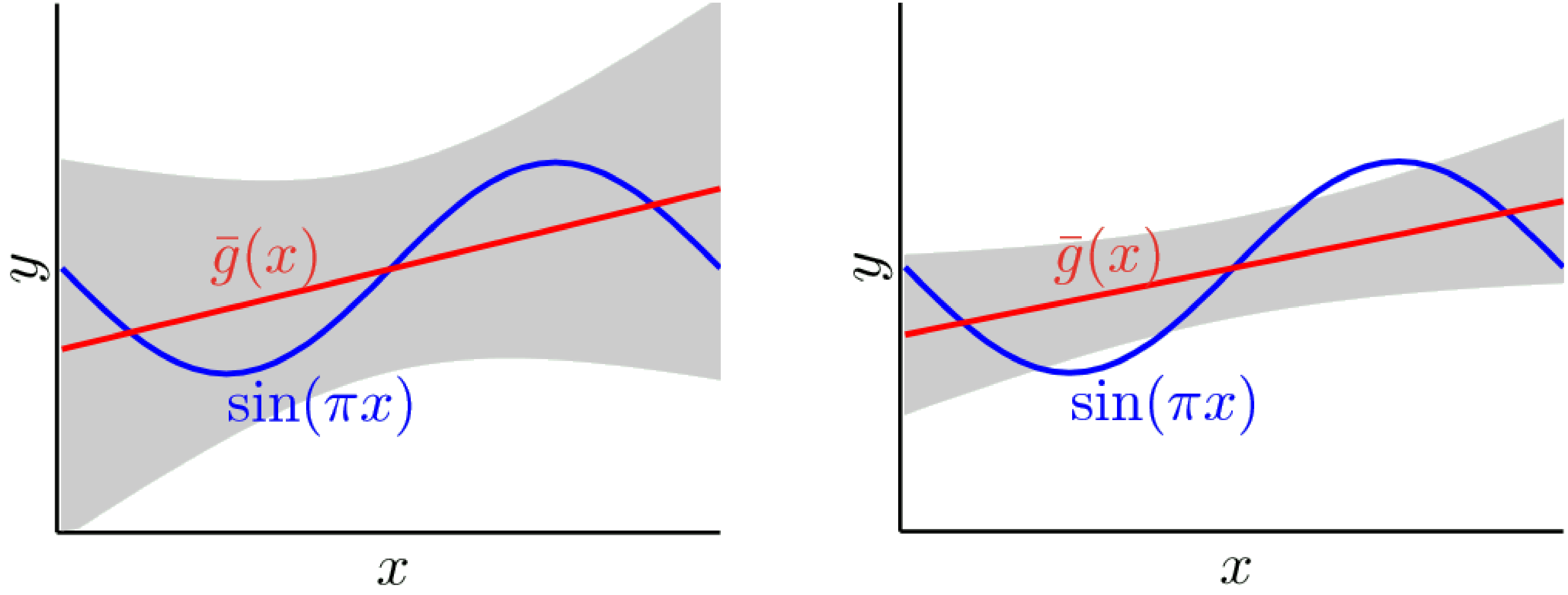
$$L_{linear}(\mathbf{w}) = \left|\left|\mathbf{Xw} - \mathbf{y}\right|\right|_2^2$$

$$L_{ridge} = L_{linear}(\mathbf{w}) + \boldsymbol{\alpha}\mathbf{w}^{\mathrm{T}}\mathbf{w} = \left((\mathbf{Xw} - \mathbf{y})^{\mathrm{T}}(\mathbf{Xw} - \mathbf{y}) + \alpha\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$$

$$\nabla L_{ridge}(\mathbf{w}) = 2\left(\mathbf{X}^{\mathrm{T}}(\mathbf{Xw} - \mathbf{y}) + \alpha\mathbf{w}\right) = 0$$

$$\mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

L2 regularization example

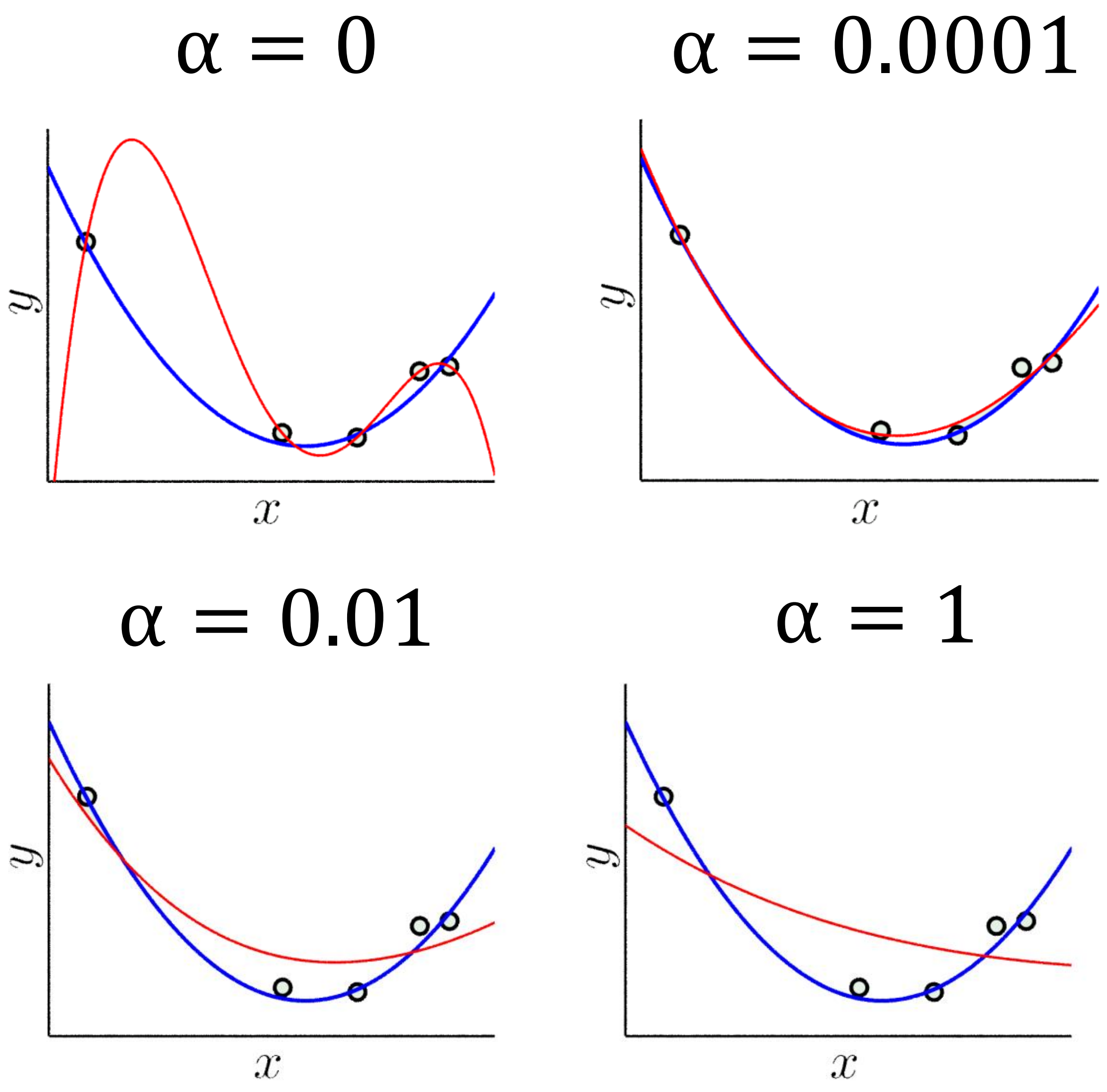


bias $= \mathbf{0.21}$ var $= \mathbf{1.69}$

bias $= \mathbf{0.23}$ var $= \mathbf{0.33}$

Overfitting and underfitting

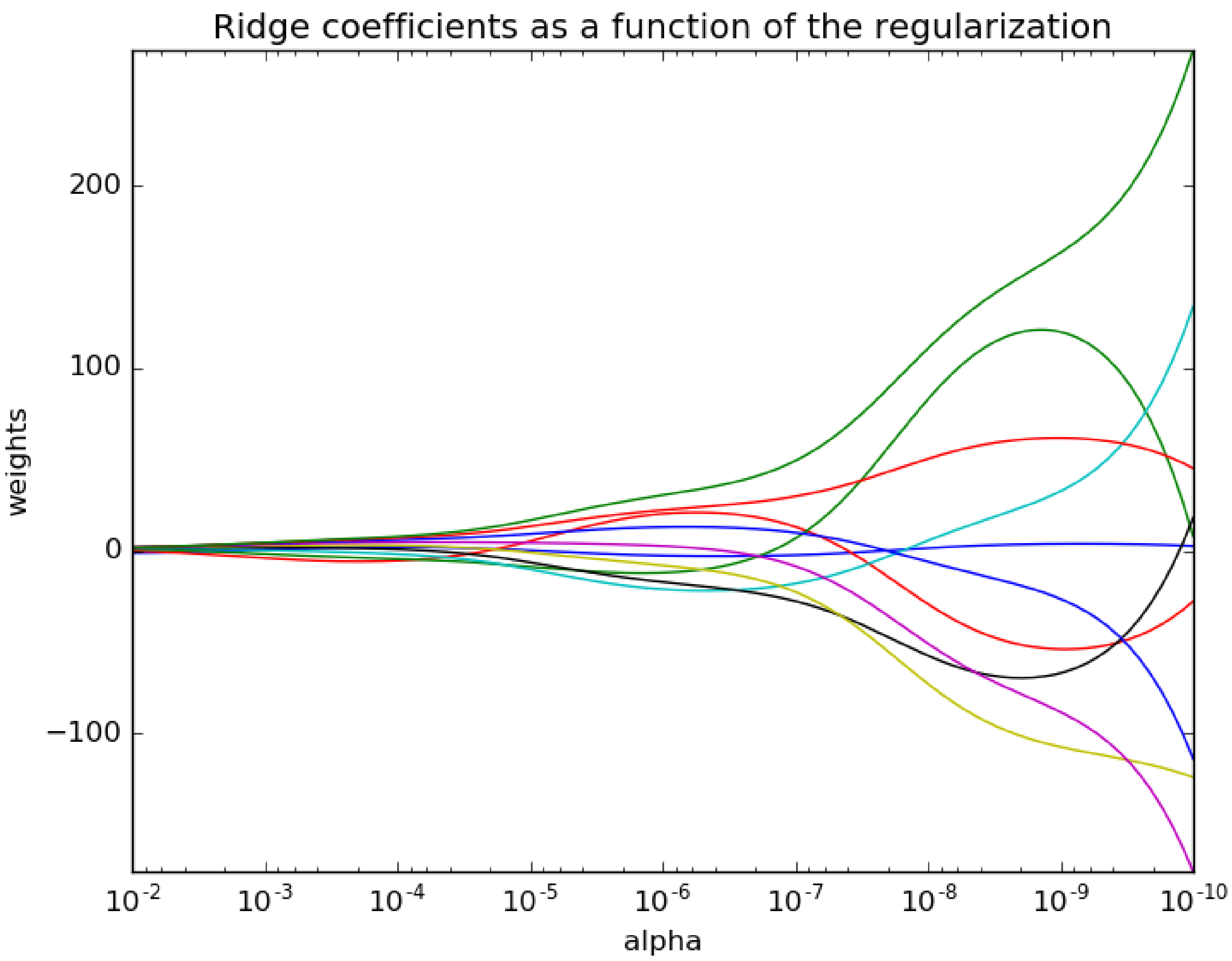
$$L_{ridge} = L_{linear}(\mathbf{w}) + \boldsymbol{\alpha}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$



Ridge regression

$$L_{ridge} = ||Xw - Y||_2^2 + \alpha||w||_2^2$$

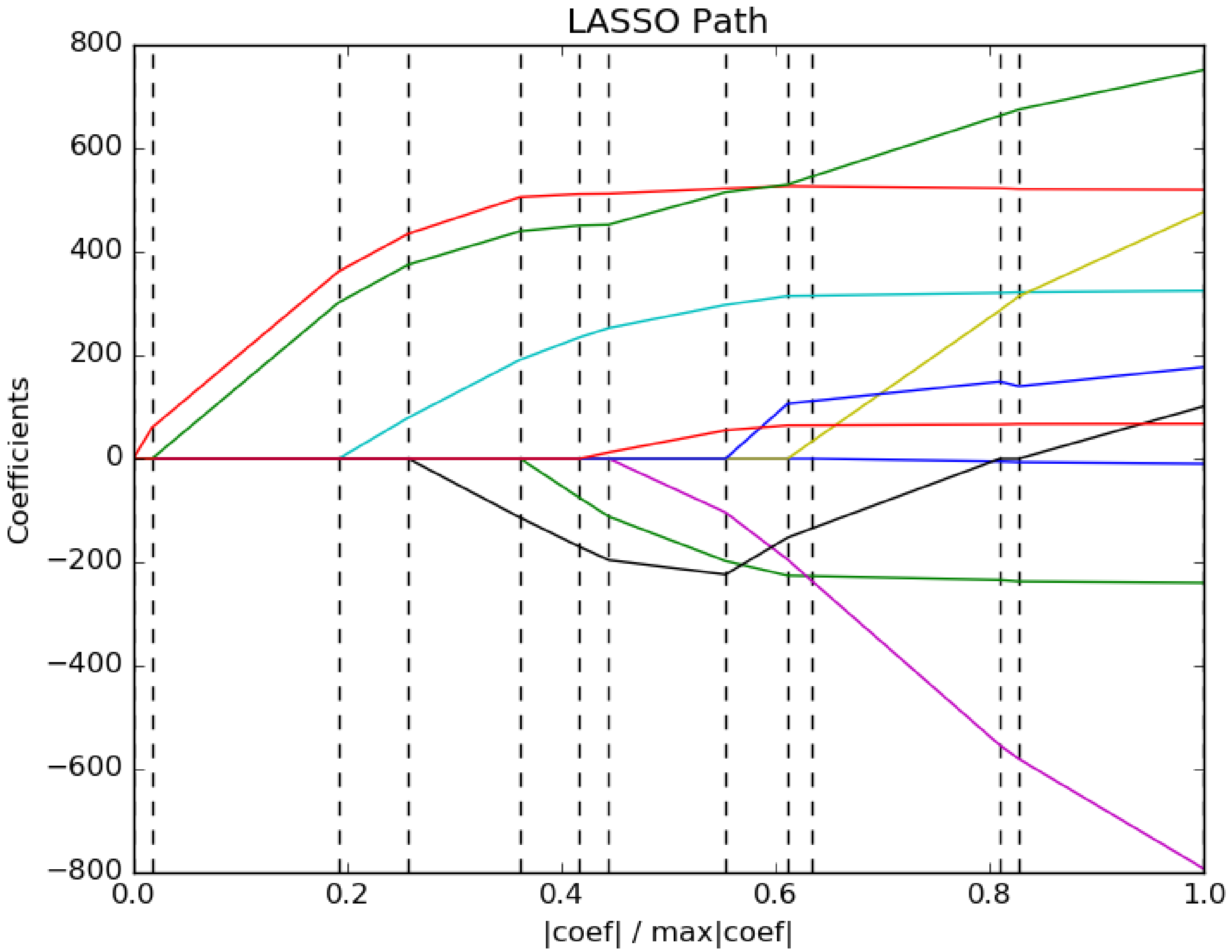
$$w = (X^T X + \alpha I)^{-1} X^T Y$$



L1 regularization and LASSO
(Least Absolute Shrinkage and Selection Operator)

$$L_{lasso} = \left|\left|Xw - Y\right|\right|_2^2 + \alpha\left|\left|w\right|\right|_1$$

Solve with coordinate descent or LARS.



LARS (Least Angle Regression)

1. Take feature x_i that has the highest absolute correlation with y.

2. Introduce coefficient β_1 as a multiplier for x_i and increase it (or decrease, in the case of negative correlation) while correlation of x_i with residual $r = y - \hat{y}$ is the maximum.

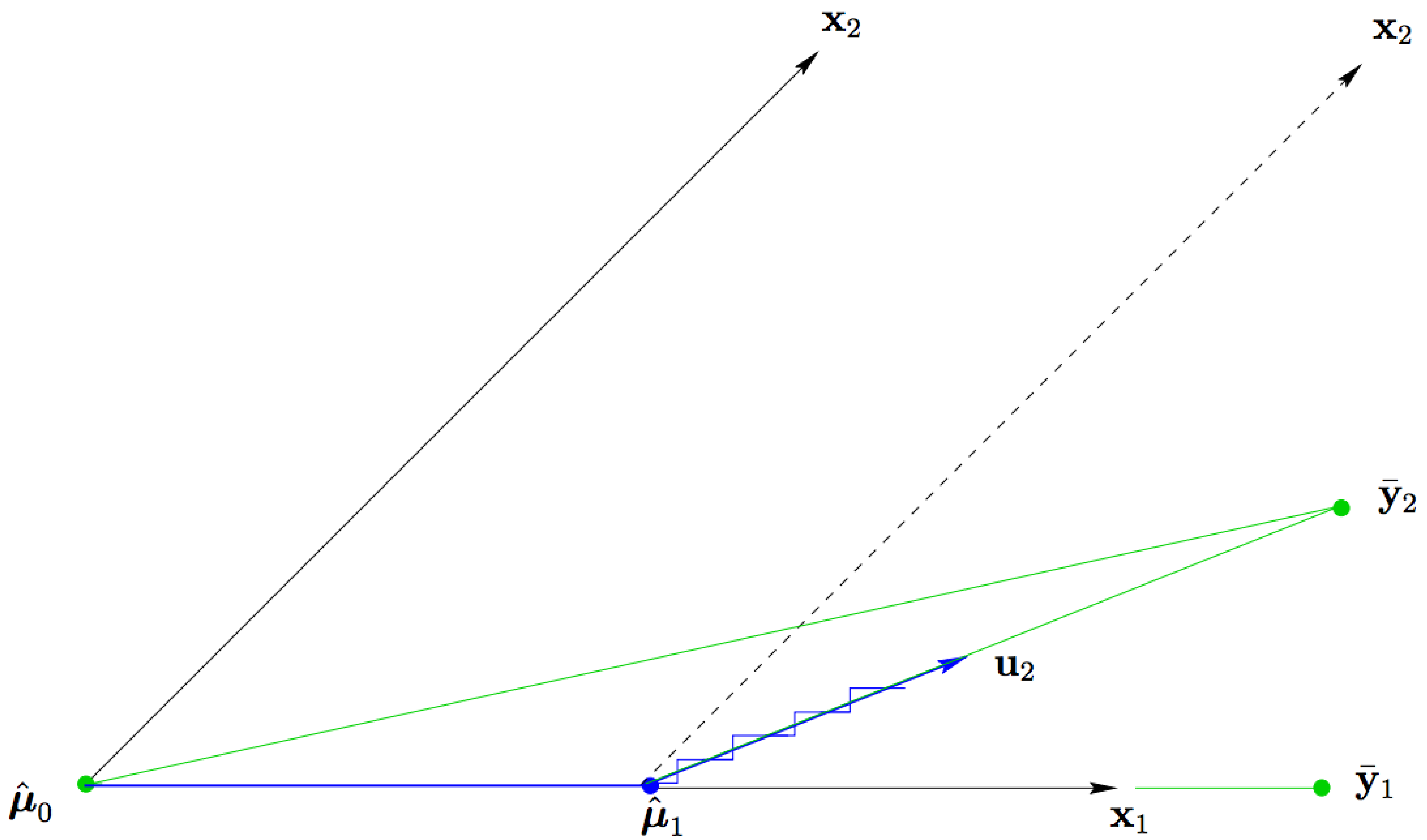
3. At the point where the condition from 2 breaks we have a new feature x_j with the same correlation.

4. Introduce β_2 as a multiplier for $\left(x_i \pm x_j\right)^*$.

5. $\rightarrow 2$

6. We stop, when the increase of the sum of the coefficients (multiplied by α) is less that the decrease in error.

LARS

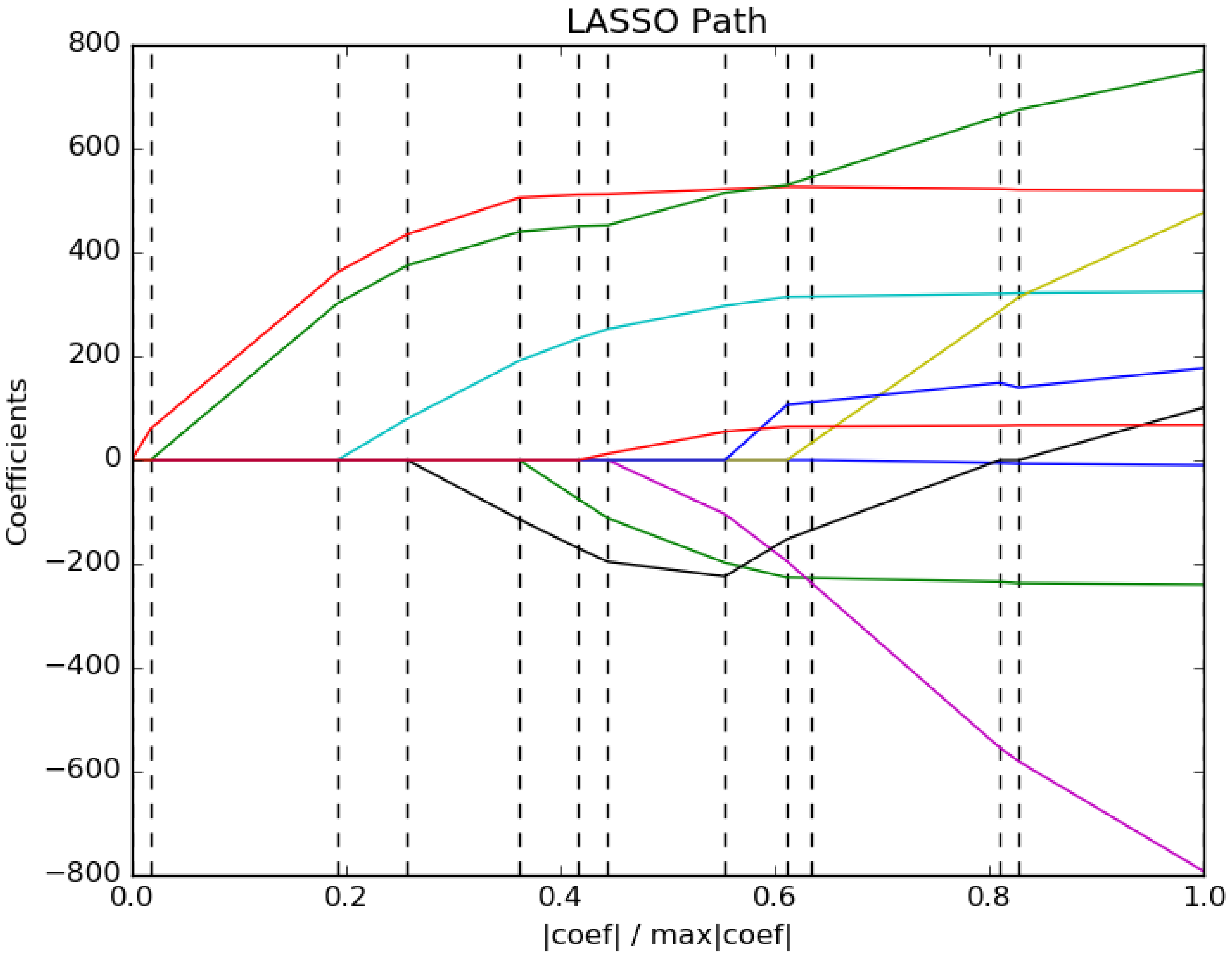


$\mathbf{x}_2$
$\mathbf{x}_2$
$\bar{\mathbf{y}}_2$
$\mathbf{u}_2$
$\hat{\boldsymbol{\mu}}_0$
$\hat{\boldsymbol{\mu}}_1$
$\mathbf{x}_1$
$\bar{\mathbf{y}}_1$

LASSO
(Least Absolute Shrinkage and Selection Operator)

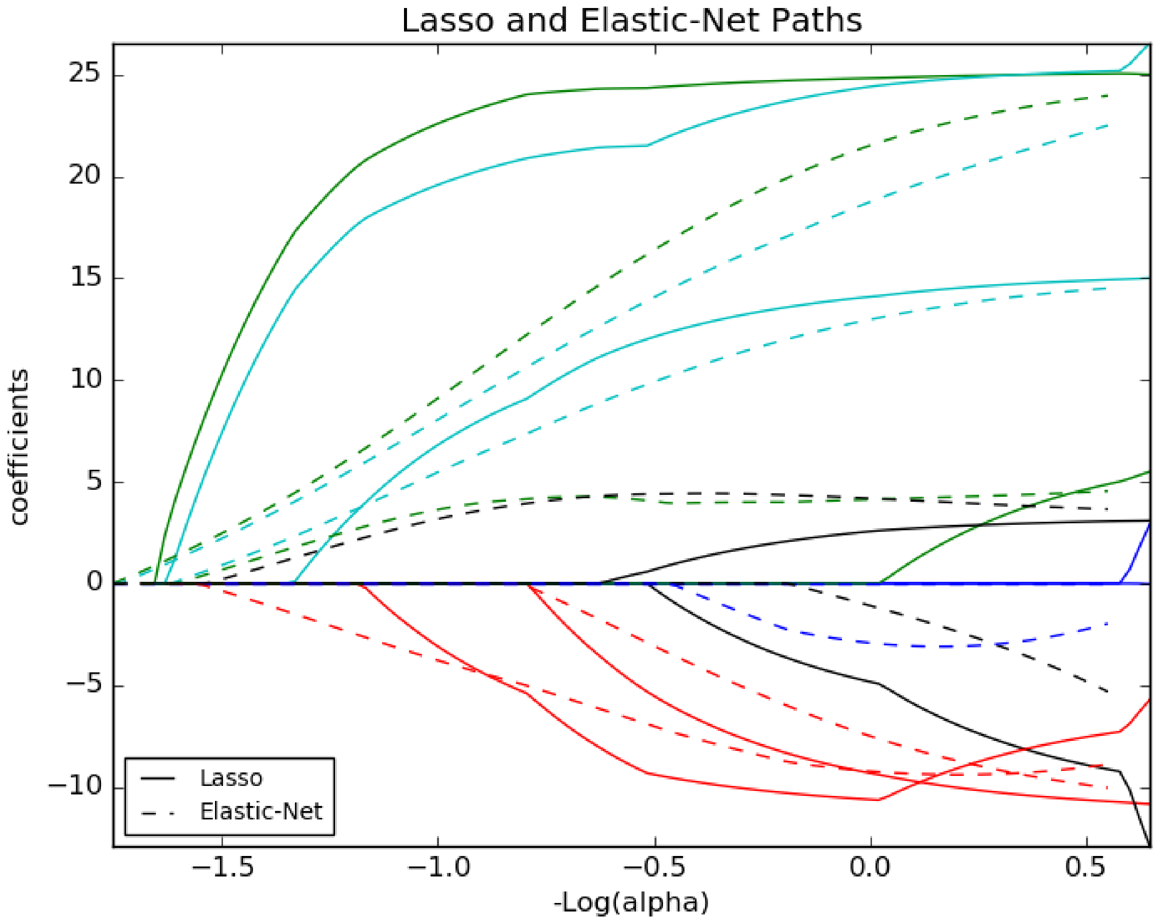
$$L_{lasso} = \left|\left|\mathrm{Xw} - \mathrm{Y}\right|\right|_2^2 + \alpha\left|\left|\mathrm{w}\right|\right|_1$$

Solve with coordinate descent or LARS.



Elastic Net

$$L_{elastic} = \left\|Xw - Y\right\|_2^2 + \alpha(1 - l1_{ratio})\left\|w\right\|_2^2 + \alpha(l1_{ratio})\left\|w\right\|_1$$



R^2-score

A useful metric for regression:

$$R^2 = 1 - \frac{u}{v}$$

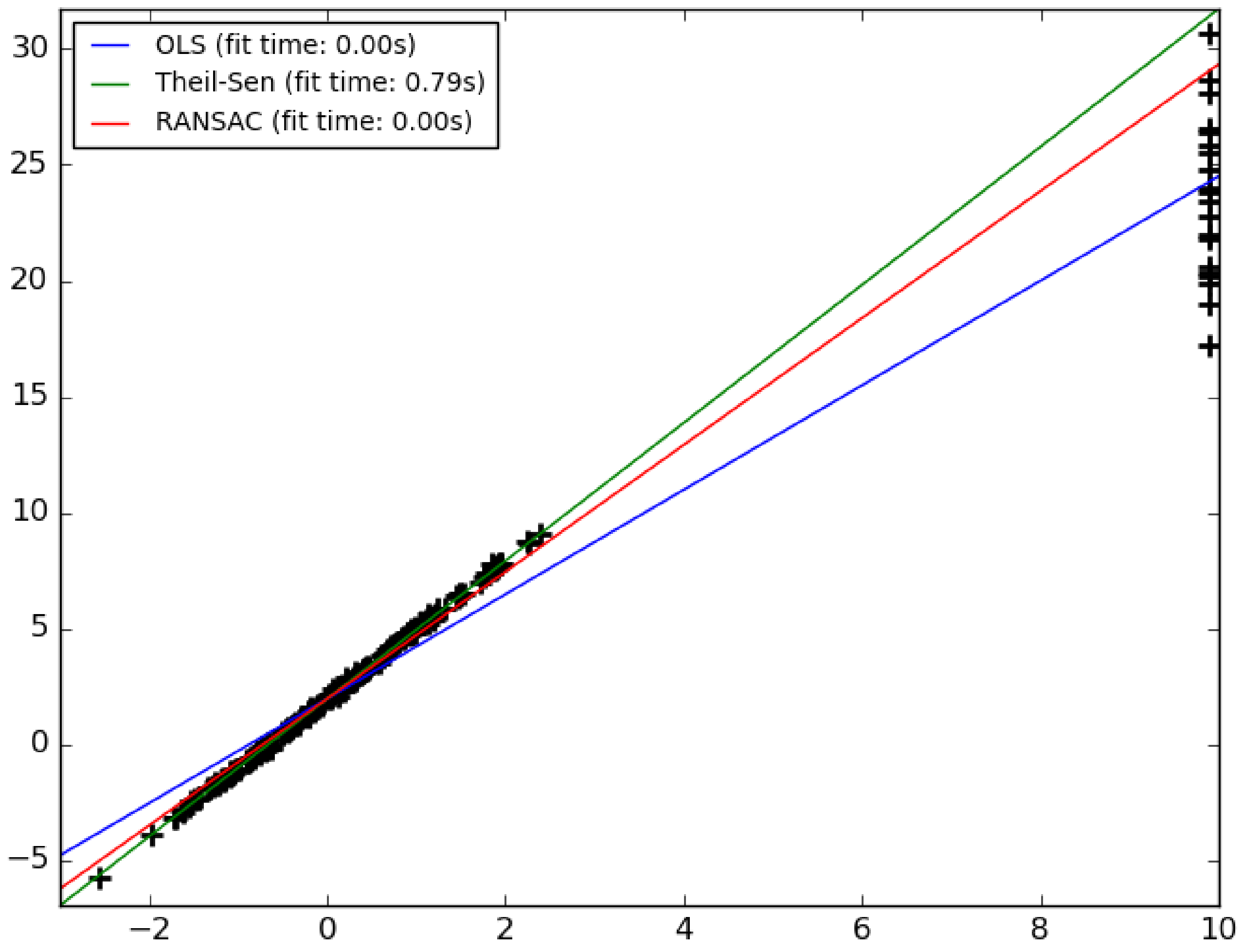
$$u = \sum(h(\mathrm{x}_i) - y_i)^2 \qquad v = \sum(\bar{y} - y_i)^2 \qquad \bar{y} = \frac{1}{N}\sum y_i$$

Fighting outliers

Theil-Sen Regressor

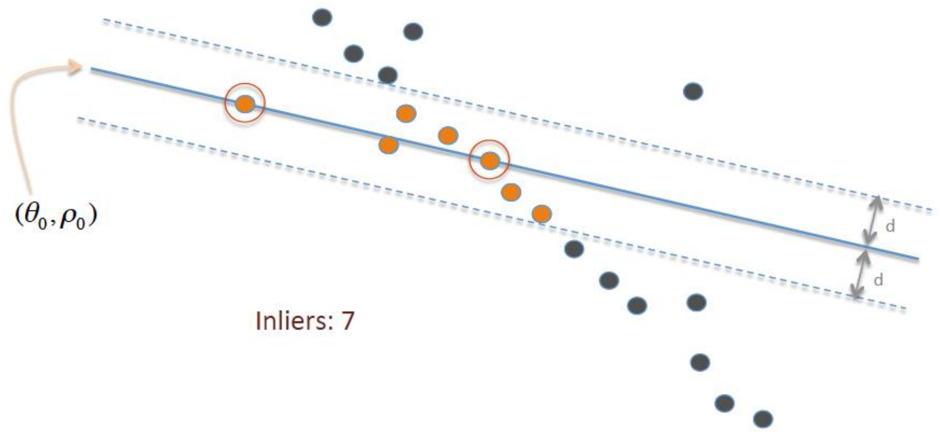
Train models on subsets of X.

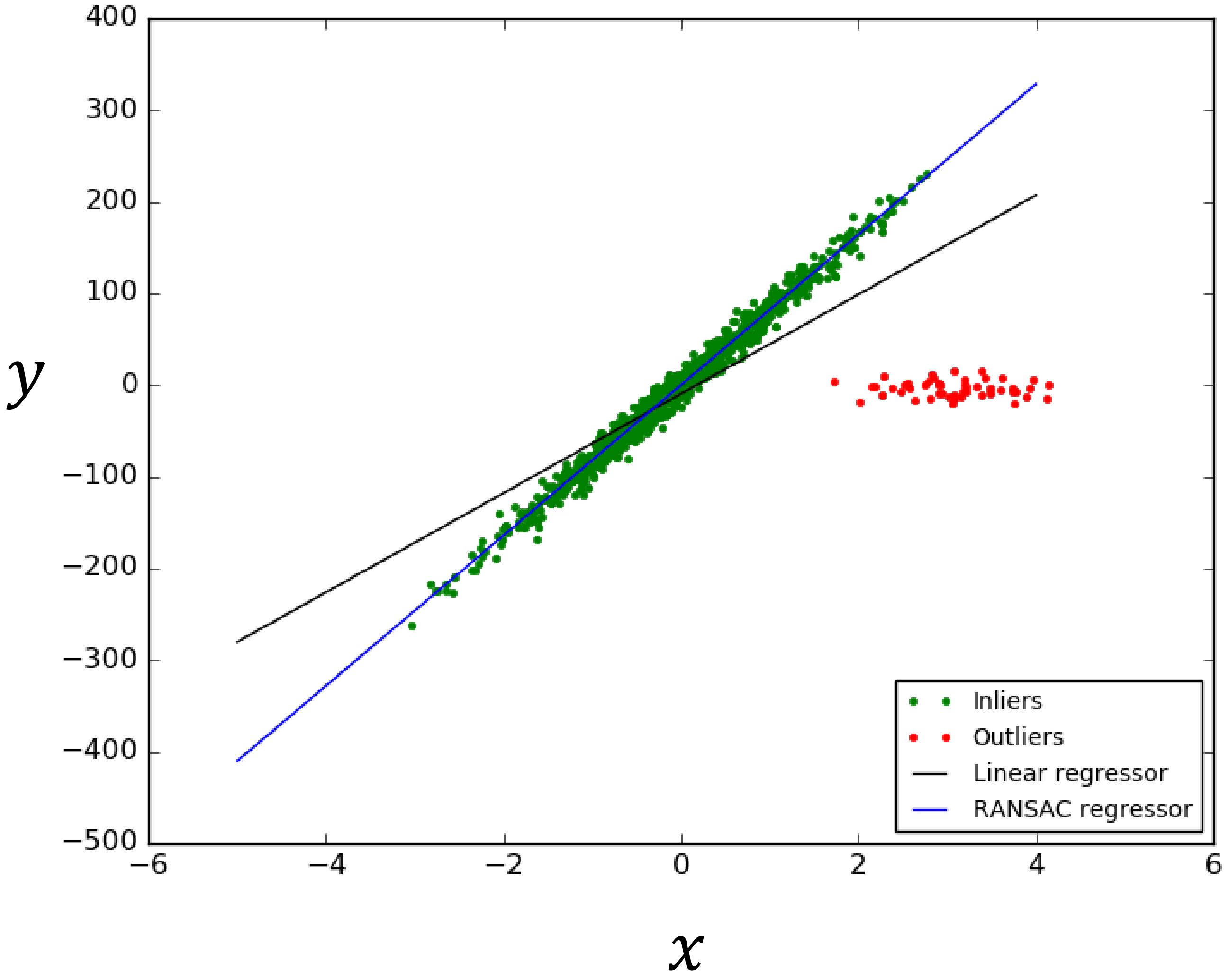
The result is the marginal median of trained models.



RANSAC: RANdom SAmple Consensus

1. Build models on subsets of X.
2. Pick the best one in terms of the number of inliers and train a new one on all these inliers.





Huber Regressor

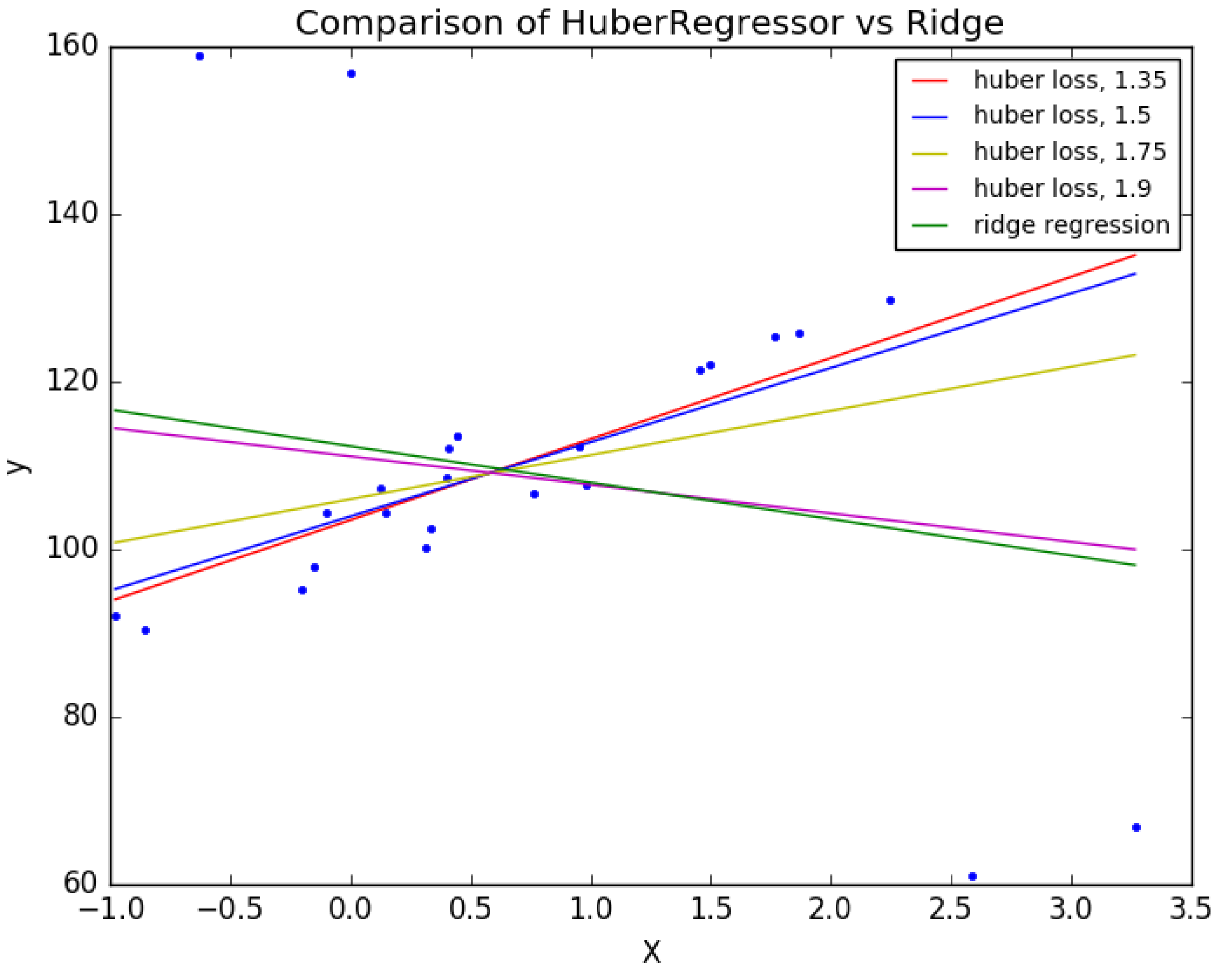
Square error for inliers, linear for outliers.

$$\min_{w,\sigma} \sum_{i=1}^{N} \left(\sigma + H_\epsilon\left(\frac{\mathrm{x}_i \mathrm{w} - y_i}{\sigma}\right)\sigma \right) + \alpha ||\mathrm{w}||_2^2$$

$$H_\epsilon(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon \\ 2\epsilon|z| - \epsilon^2, & \text{if } |z| \geq \epsilon \end{cases}$$

σ – scaling constant.

It is advised to set the parameter ϵ to 1.35 to achieve 95% statistical efficiency.



RANSAC vs. Theil-Sen vs. Huber