Overfitting is the outcome of noise creeping into the signal

difficult to avoid with noisy data

Regularization is a procedure to control overfitting

consider fitting a linear hypothesis: $\widehat{y} = \mathbf{x}^T \mathbf{w}$

In **regularized regression**, we define a cost function

$$\underbrace{E = \frac{1}{n} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) + \overset{\text{penalty term}}{\lambda \mathbf{w}^T \mathbf{w}}}_{\text{Ridge regression (Tikonov regularization)}}$$

λ : penalty parameter

$\lambda \to 0$: classical least square regression

$\lambda \to \infty$: $\widehat{y} \to 0$

minimization of E requires

$$\nabla E(\mathbf{w}) = \mathbf{0} \Rightarrow \left(\mathbf{X}^T \mathbf{X} + n\lambda \mathbf{I}\right) \mathbf{w} = \mathbf{X}^T \mathbf{y}$$

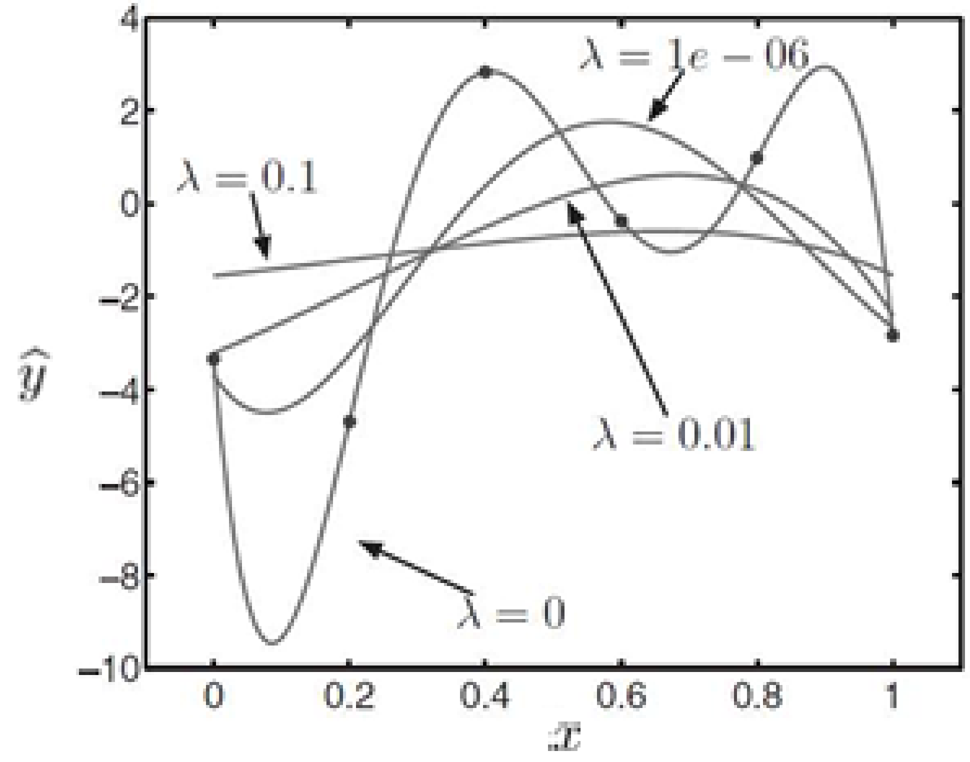
Thus regularization tends to reduce the model complexity by reducing $\mathbf{w}$

λ is decided based on cross-validation consider fitting the linear hypothesis:

$$\widehat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$$

$$E = \frac{1}{n}\left(\mathbf{Xw} - \mathbf{y}\right)^T\left(\mathbf{Xw} - \mathbf{y}\right) + \lambda \mathbf{w}^T\mathbf{w}$$

$$\Rightarrow \left(\mathbf{X}^T\mathbf{X} + n\lambda\mathbf{I}\right)\mathbf{w} = \mathbf{X}^T\mathbf{y}$$



increasing λ reduces fluctuations

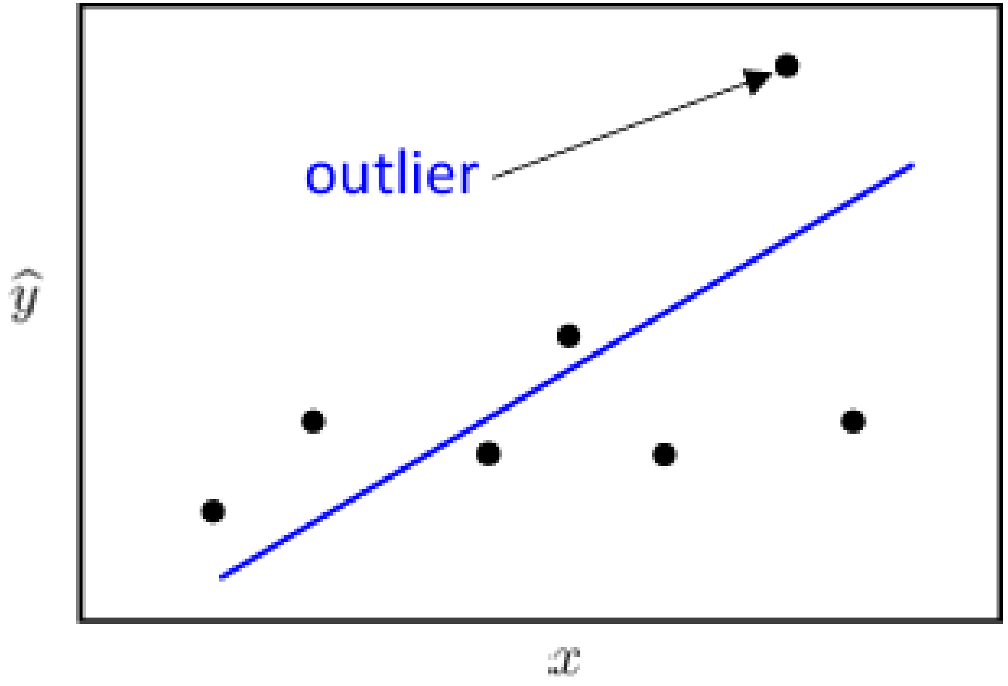
Other forms of regularization

Lasso regression:

$$E = \frac{1}{n}\left(\mathbf{Xw} - \mathbf{y}\right)^T\left(\mathbf{Xw} - \mathbf{y}\right) + \lambda\|\mathbf{w}\|_1 \qquad \|\mathbf{w}\|_1 = \sum |w|$$

Elastic net regression: $E = \frac{1}{n}\left(\mathbf{Xw} - \mathbf{y}\right)^T\left(\mathbf{Xw} - \mathbf{y}\right) + \lambda\left[\|\alpha\mathbf{w}\|_1 + (1-\alpha)\,\mathbf{w}^T\mathbf{w}\right]$

Susceptibility to **Outlier**



least square fit, due to squaring of residual, is heavily influenced by outliers

Least absolute deviation fit is often used to reduce the dependence on outlier

$$E = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_1$$

Least absolute deviation has zero double derivative, precludes use of some optimization algorithms

In most cases, Least absolute deviation reduces cost function to a lower value than that of the least square regression

Outlier may also be removed based on appropriate criterion of loss function

Nonlinear regression: normal equations are nonlinear

Linear regression: $\widehat{y} = \mathbf{x}^T\mathbf{w}$ $\widehat{y}_i = \mathbf{x}_i^T\mathbf{w}$ $i = 1, 2, \cdots, n$

$\mathbf{x}^T = \begin{bmatrix}1 & x_1 & x_2 & \dots & x_k\end{bmatrix}$

$\mathbf{x}_i^T = \begin{bmatrix}1 & x_{i1} & x_{i2} & \dots & x_{ik}\end{bmatrix}$

$\mathbf{w}^T = \begin{bmatrix}w_0 & w_1 & w_2 & \dots & w_k\end{bmatrix}$

In general, $\widehat{y} = f(\mathbf{x}, \mathbf{w})$ $\widehat{y}_i = f(\mathbf{x}_i, \mathbf{w})$

Cost function $E(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left[f(\mathbf{x}_i, \mathbf{w}) - y_i\right]^2$

We wish to find

$$\arg\min_{\mathbf{w}} E(\mathbf{w})$$

$$\frac{\partial E}{\partial w_j} = 0 \Rightarrow \sum_{i=1}^{n}\left[f(\mathbf{x}_i, \mathbf{w}) - y_i\right]\frac{\partial f(\mathbf{x}_i, \mathbf{w})}{\partial w_j} = 0 \quad j = 0, 1, 2, \cdots, k$$

$$\Rightarrow \sum_{i=1}^{n}\left[f(\mathbf{x}_i, \mathbf{w}) - y_i\right]\nabla f(\mathbf{x}_i, \mathbf{w}) = 0$$

Normal equations; solves $\mathbf{w}$

In nonlinear regression, normal equations are **nonlinear**

- direct minimization of $E(\mathbf{w})$
- solving nonlinear normal equations using suitable numerical methods
- linearization

Example: Given a set of discrete data points $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{n}$

We wish to fit $\widehat{y} = w_0 x^{w_1}$

Cost function: $E(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w_0 x_i^{w_1}\right)^2$

$$\frac{\partial E}{\partial w_0} = 0 = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - w_0 x_i^{w_1}\right) x_i^{w_1} \qquad \frac{\partial E}{\partial w_1} = 0 = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - w_0 x_i^{w_1}\right) w_0 \ln(x_i)\, x_i^{w_1}$$

We can calculate w_0, w_1 by solving the normal equations

$$\boxed{\begin{aligned} &\sum_{i=1}^{n}\left(y_i - w_0 x_i^{w_1}\right) x_i^{w_1} = 0 \\ &\sum_{i=1}^{n}\left(y_i - w_0 x_i^{w_1}\right) w_0 \ln(x_i)\, x_i^{w_1} = 0 \end{aligned}}$$

the normal equations are nonlinear

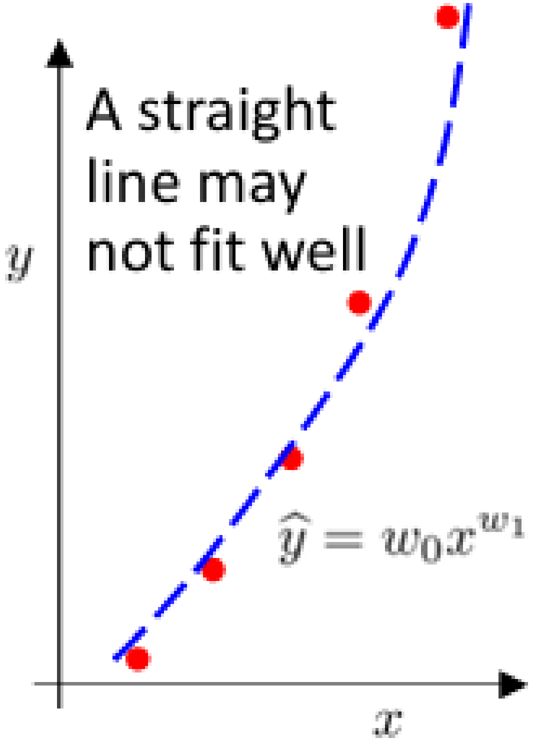
One approach to avoid nonlinearity:

modifying $\widehat{y} = w_0 x^{w_1}$

Example: **linearization**

Let's consider this dataset

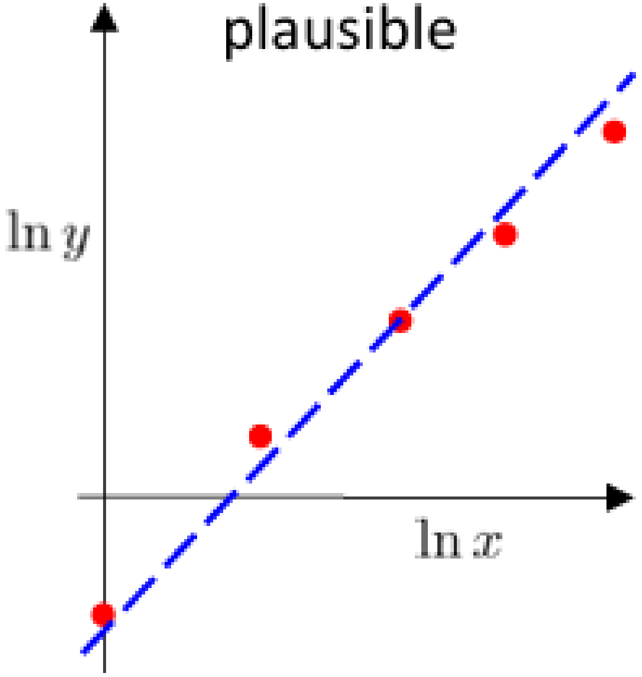
x	y
1	0.4
2	1.7
3	3
4	5.5
5	8.4



Linearization

Transforming

$\ln x$	$\ln y$
0	-0.916
0.693	0.531
1.099	1.099
1.386	1.705
1.609	2.128



A straight-line fit seems plausible

We are trying to fit $\widehat{y} = w_0 x^{w_1}$ $\boxed{\Rightarrow \ln \widehat{y} = \ln (w_0) + w_1 \ln x}$

Simple linear regression seems now applicable

Limitations of linearization

Use of a hypothesis $\widehat{y} = f(x)$ assumes existence of a model $y = g(x)$

such that the experiments (observations) generate $y_i = g(x = x_i) + \epsilon_i$

use of least square regression facilitates $f(x) \to g(x)$ with more training data

Linearization tacitly assumes multiplicative noise $y_i = \epsilon_i \theta_0 x^{\theta_1}$

If the noise is additive $y_i = \theta_0 x^{\theta_1} + \epsilon_i$ linearization may not be acceptable

Linearization is not possible for all nonlinear models

For instance, a model $y \approx \theta_0 + \theta_1 x^{\theta_2} + \theta_3 x^{\theta_4}$

cannot be linearized using the procedure discussed