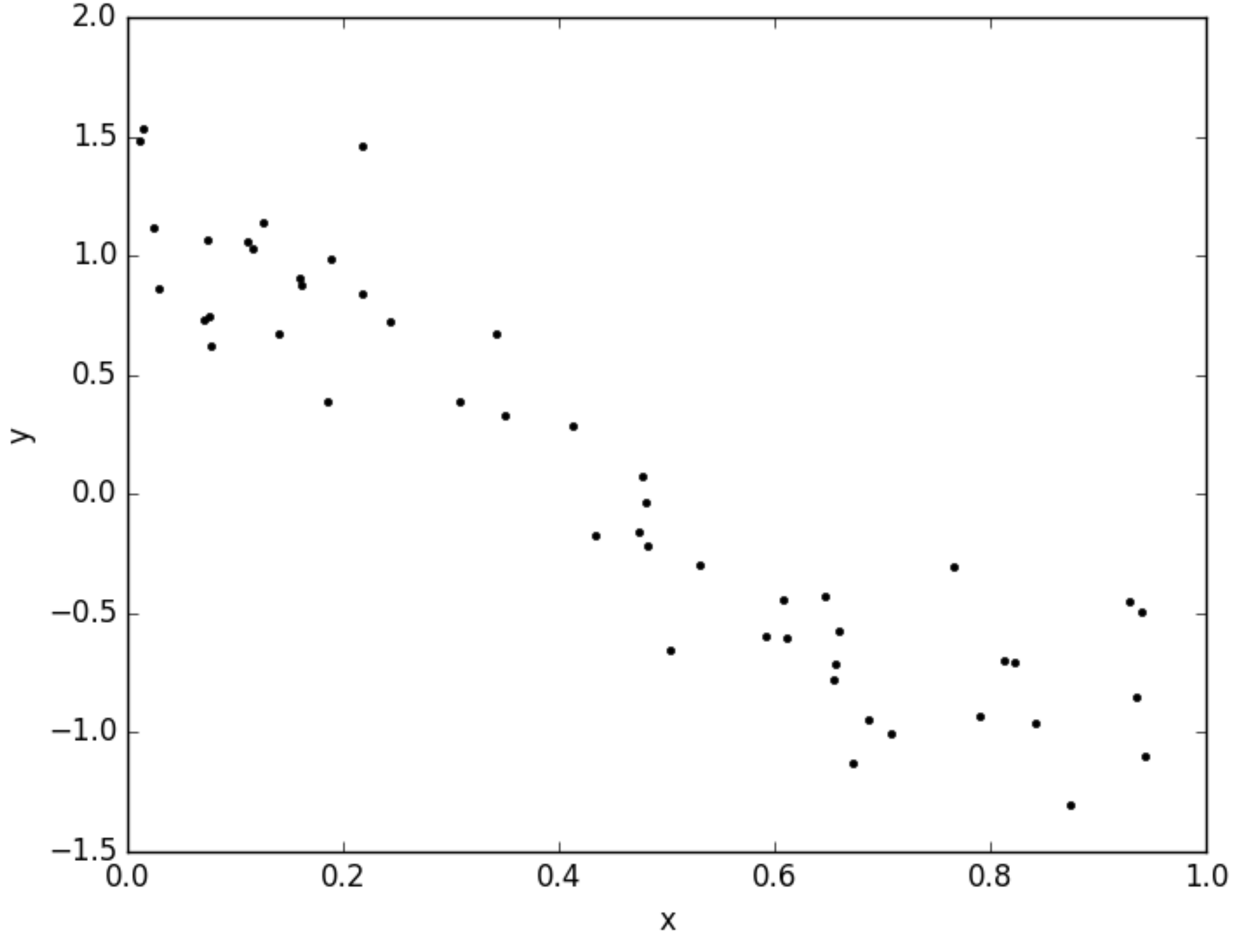
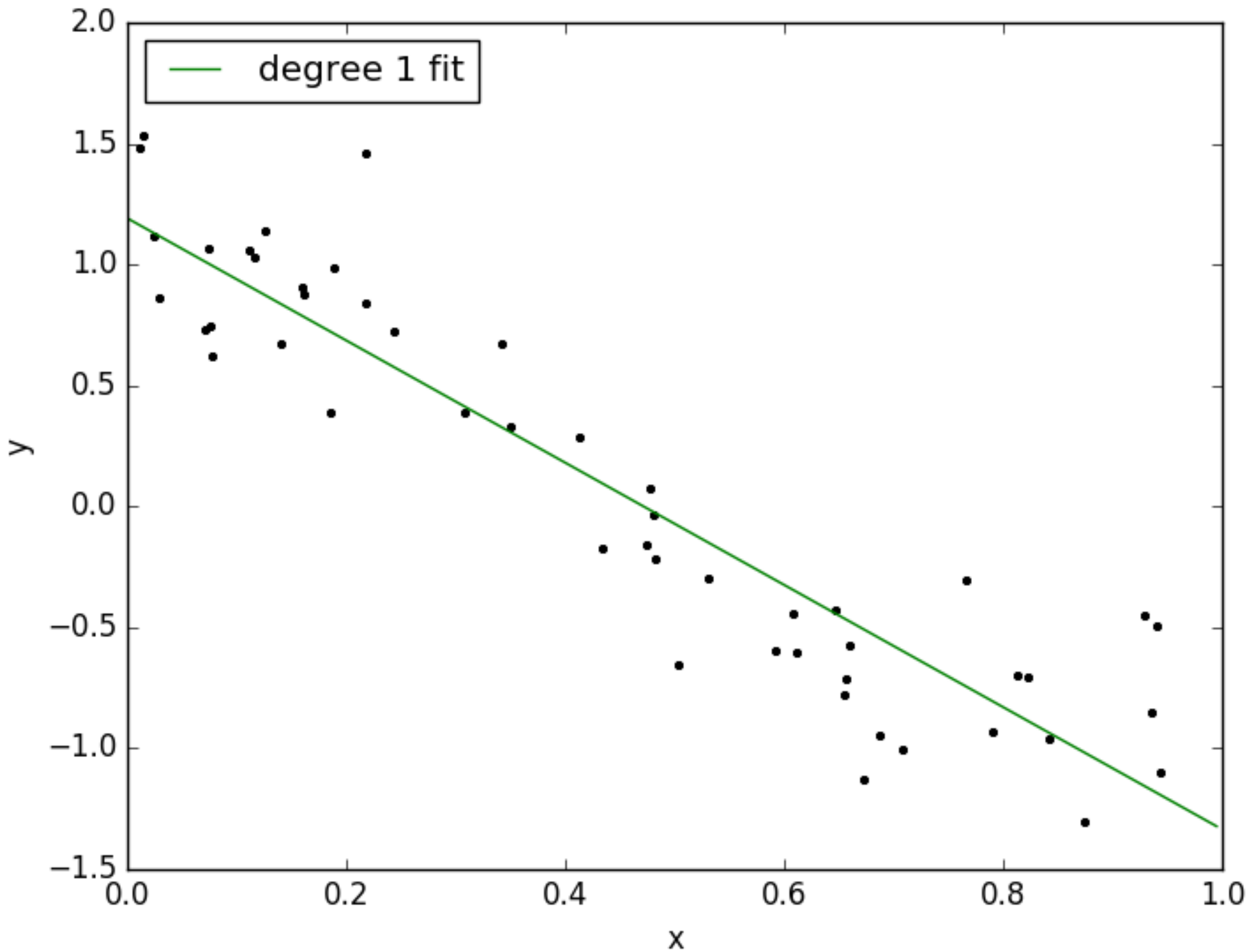
Ridge Regression

LINE FITTING



LINEAR REGRESSION

Linear Model `y = -2.526 x + 1.191`



RECAP - LEAST SQUARES

Recall the least squares optimization problem:

$$\text{minimize}_\theta \ \frac{1}{2}\|X\theta - y\|_2^2$$

What is the gradient of the optimization objective ????????

$$\nabla_\theta \frac{1}{2}\|X\theta - y\|_2^2 =$$

Chain rule:

$$\nabla_\theta f(X\theta) = X^T \nabla_{X\theta} f(X\theta)$$

$$X^T \nabla_{X\theta} \frac{1}{2}\|X\theta - y\|_2^2 =$$

Gradient of norm:

$$\nabla_\theta \|\theta - z\|_2^2 = 2(\theta - z)$$

$$\nabla_\theta \frac{1}{2}\|X\theta - y\|_2^2 = X^T(X\theta - y)$$

RECAP - LEAST SQUARES

Recall: points where the gradient equals zero are minima.

$$\nabla_\theta \frac{1}{2}\|X\theta - y\|_2^2 = X^T(X\theta - y)$$

So where do we go from here?????????

$$X^T(X\theta - y) = 0$$

Solve for model parameters θ

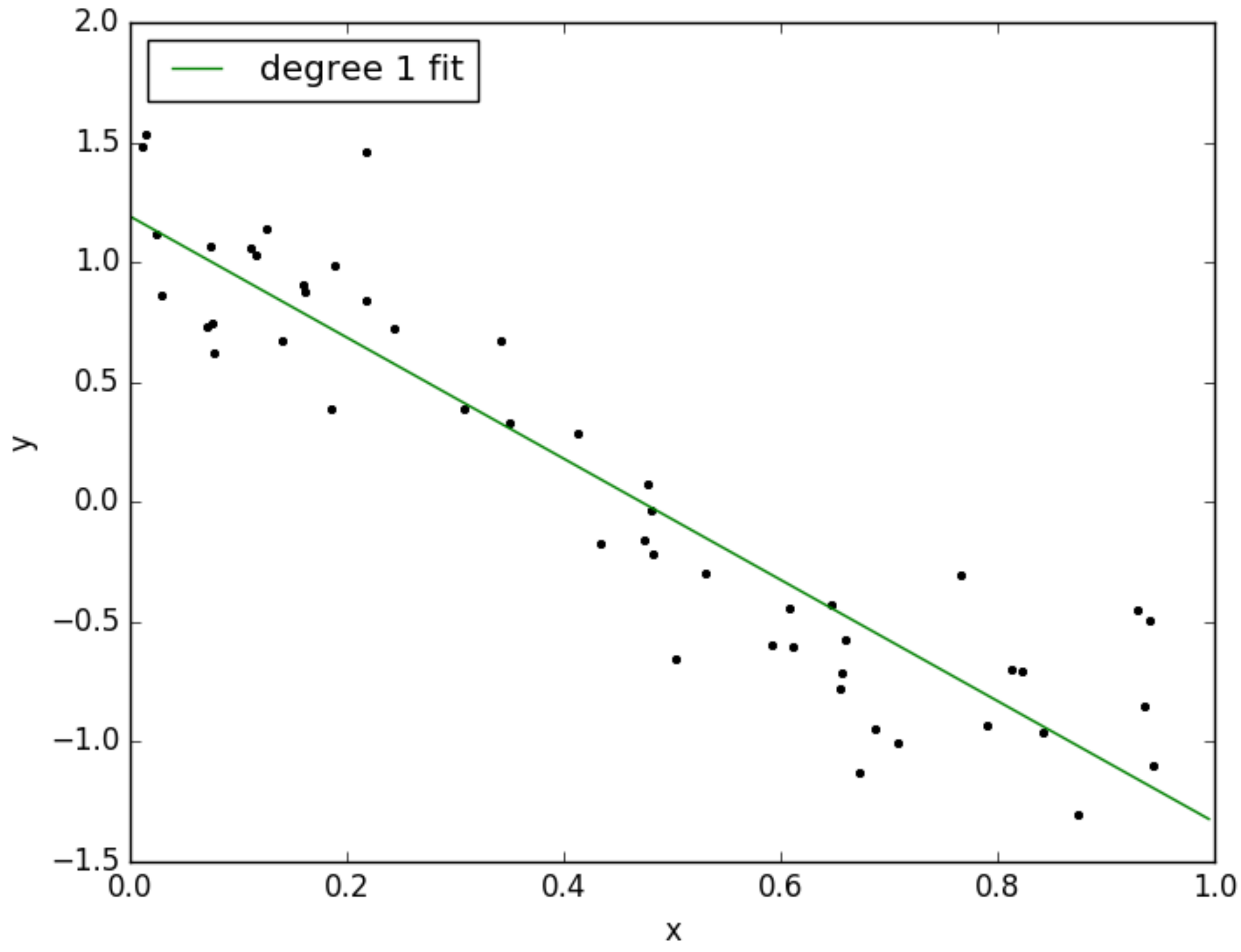
$$X^TX\theta - X^Ty = 0 \Rightarrow X^TX\theta = X^Ty$$

$$(X^TX)^{-1}X^TX\theta = (X^TX)^{-1}X^Ty$$

$$\theta = (X^TX)^{-1}X^Ty$$

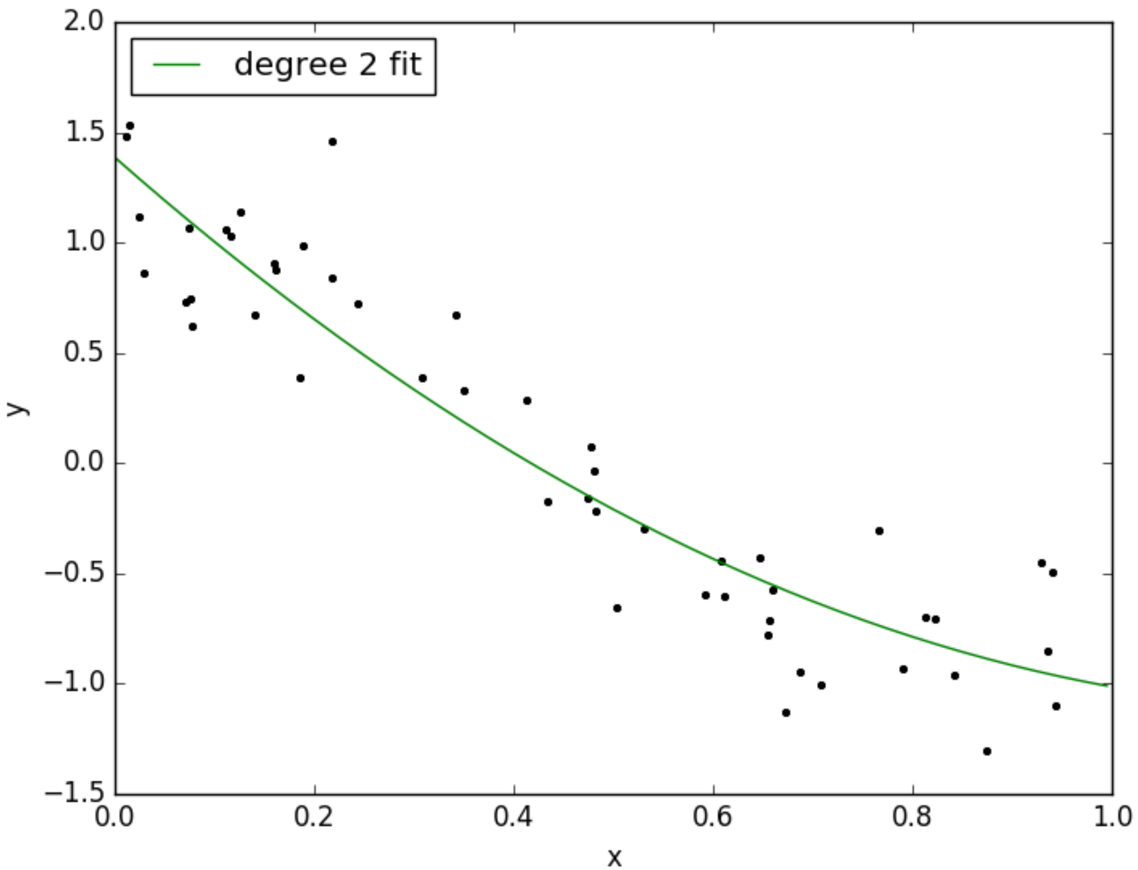
RIDGE REGRESSION - REGULARIZATION

Linear Model `y = -2.526 x + 1.191`



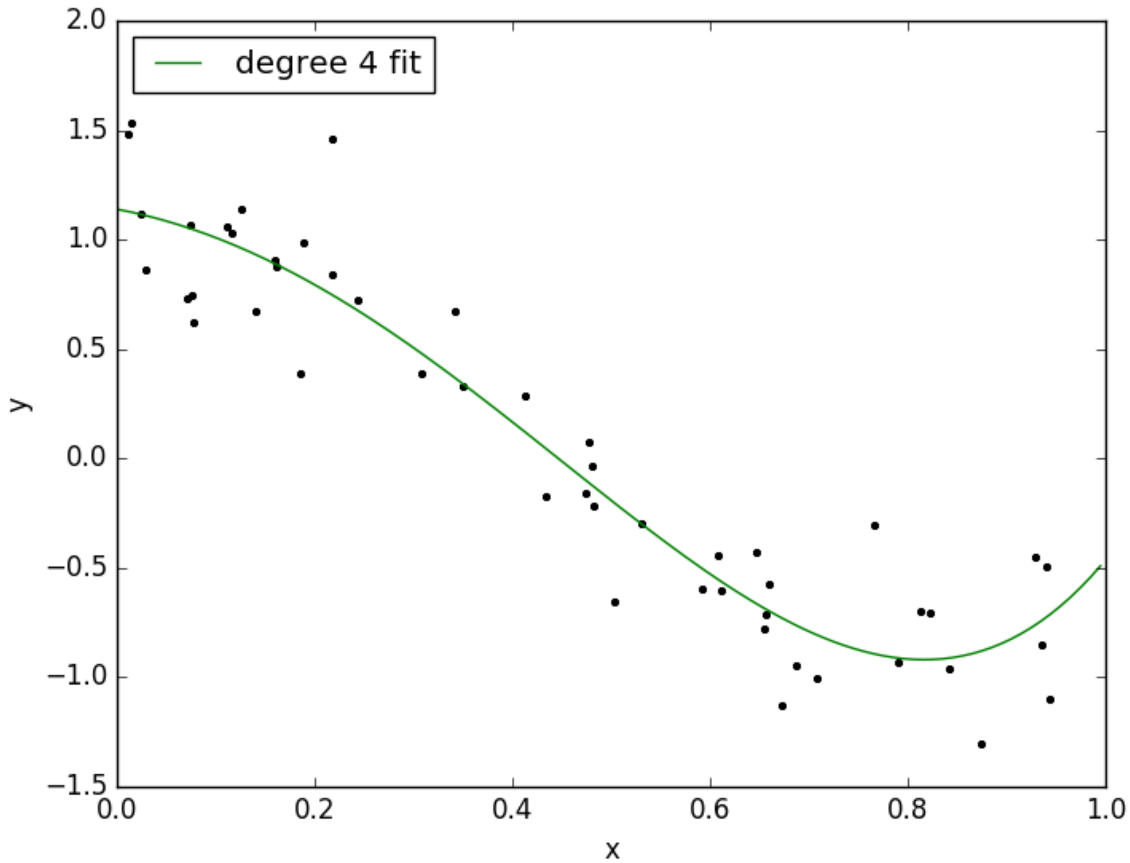
RIDGE REGRESSION - REGULARIZATION

Quadratic Model $y = 1.583\ x^2 - 3.983\ x + 1.386$



RIDGE REGRESSION - REGULARIZATION

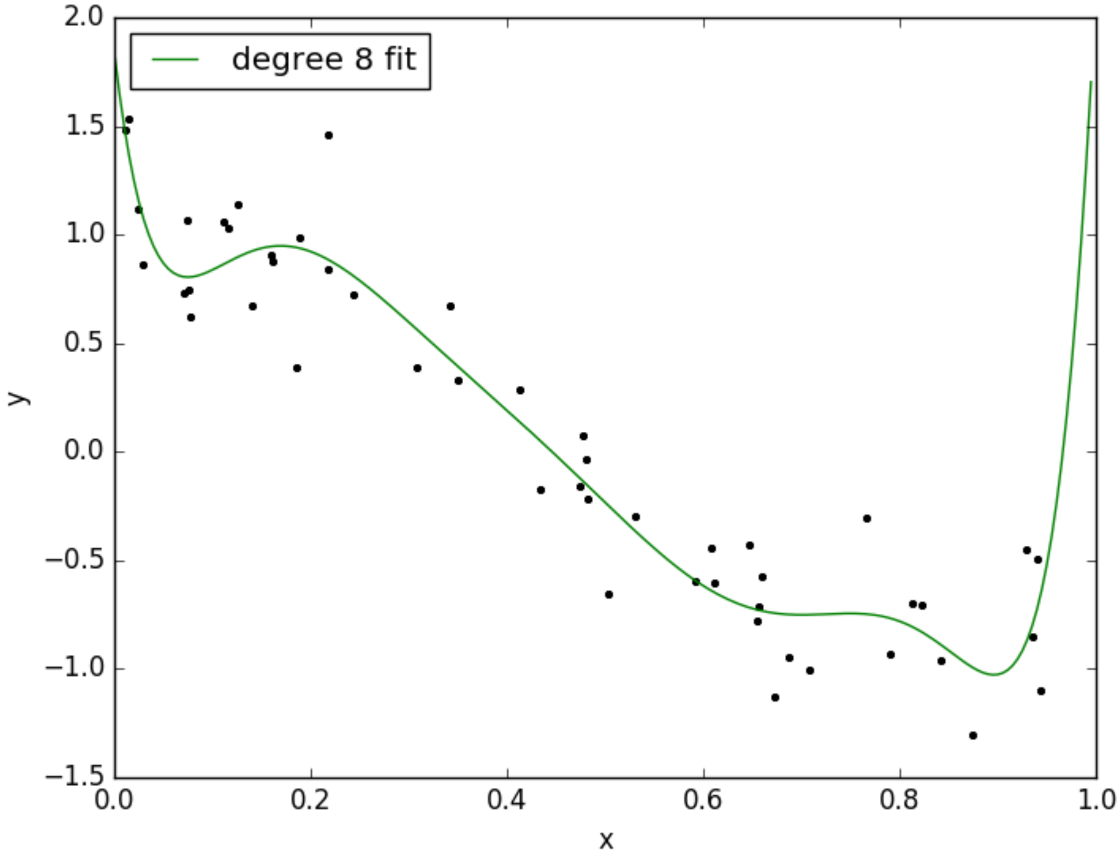
Degree 4 Model $y = 4.165 x^4 - 0.5359 x^3 - 4.369 x^2 - 0.8634 x + 1.139$



RIDGE REGRESSION - REGULARIZATION

Degree 8 Model

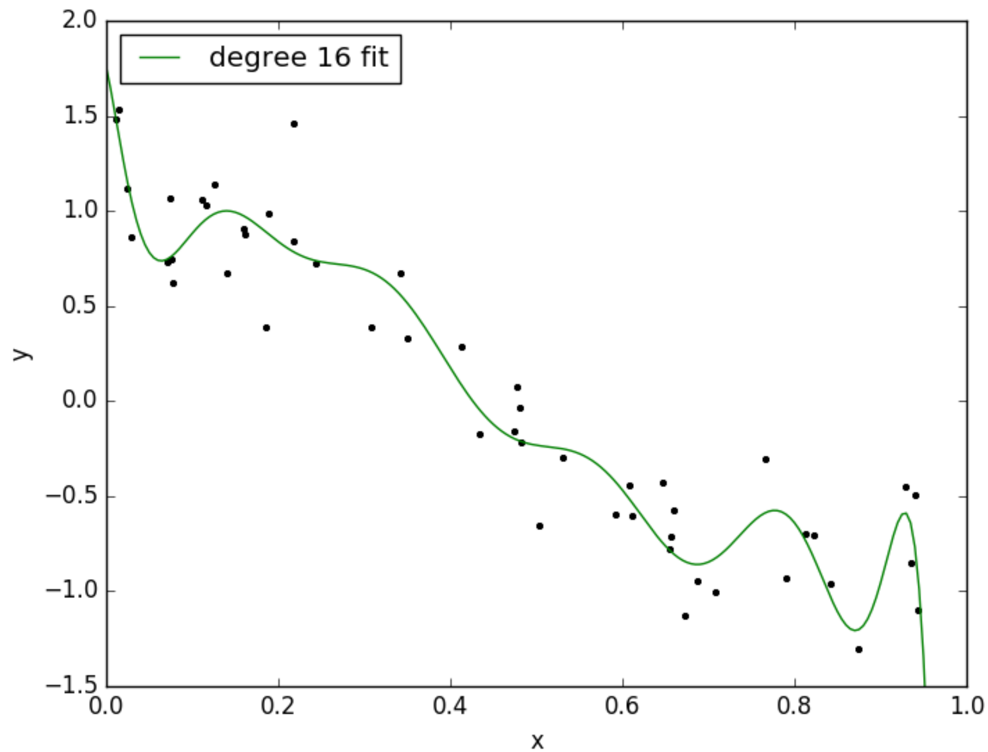
$$y = 4722x^8 - 1.81e{+}04x^7 + 2.872e{+}04x^6 - 2.444e{+}04x^5 + 1.206e{+}04x^4 - 3464x^3 + 537.3x^2 - 39.02x + 1.84$$



RIDGE REGRESSION - REGULARIZATION

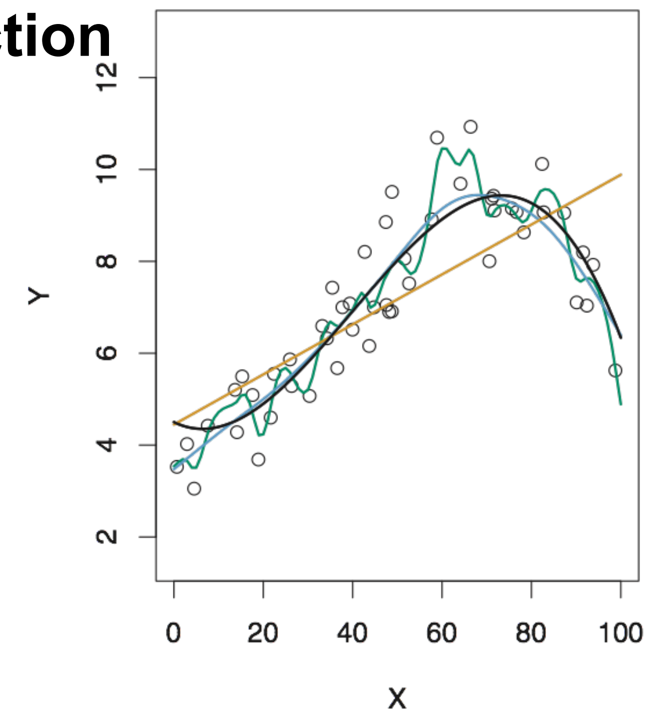
Degree 16 Model

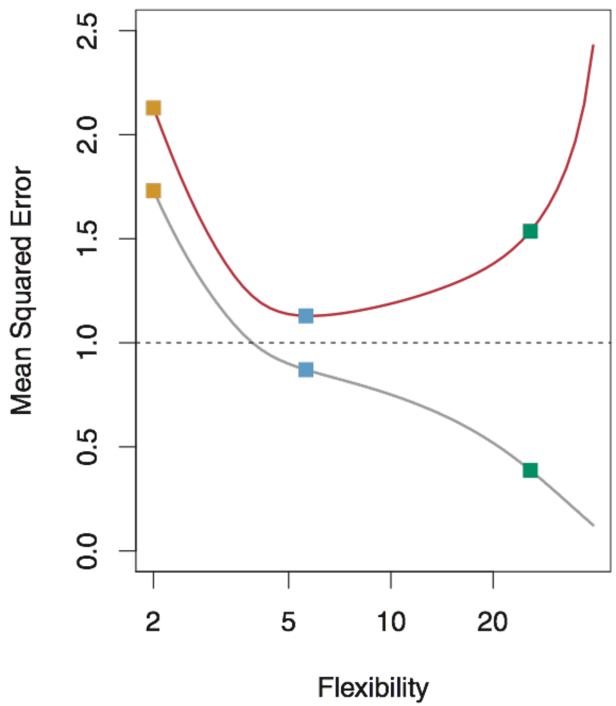
$$y = 1.33e{+}06x^{16} - 6.428e{+}06x^{15} + 1.268e{+}07x^{14} - 1.378e{+}07x^{13} + 1.019e{+}07x^{12} - 4.277e{+}06x^{11} - 6.472e{+}06x^{10} + 1.821e{+}07x^{9} - 2.086e{+}07x^{8} + 1.389e{+}07x^{7} - 5.775e{+}06x^{6} + 1.504e{+}06x^{5} - 2.35e{+}05x^{4} + 1.939e{+}04x^{3} - 516.8x^{2} - 20.56x + 1.757$$



OVERFITTING - SOLUTION

- **Model Selection**





- **Regularization (Ridge Regression)**

REGULARIZATION

- **Update the loss function**
- **For linear Regression, loss function**

$$L(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta(x_i) - y_i\right)^2$$

- **Parameters:** $\theta_0, \theta_1, \ldots, \theta_m$

REGULARIZATION

- **For linear Regression, loss function**

$$L(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta(x_i) - y_i \right)^2$$

- **Parameters:** $\theta_0, \theta_1, \ldots, \theta_m$

- **L2 norm or the sum of squares:**

$$\theta_1^2 + \theta_2^2 + \ldots + \theta_m^2 = \sum_{j=1}^{m} \theta_j^2 \approx ||\theta||_2^2 \text{–L2 norm}$$

REGULARIZATION

- **For linear Regression, loss function**

$$L(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta(x_i) - y_i\right)^2$$

- **L2 norm or the sum of squares:**

$$\theta_1^2 + \theta_2^2 + \ldots + \theta_m^2 = \sum_{j=1}^{m}\theta_j^2 \approx |\theta|_2^2 \text{ – L2 norm}$$

- **Updated loss function:**

$$\min_{\theta_1,\theta_2,\ldots,\theta_m} \qquad L(\theta) = \frac{1}{2n}\left[\sum_{i=1}^{n}\left(h_\theta(x_i) - y_i\right)^2 + \lambda\sum_{j=1}^{m}\theta_j^2\right]$$

λ – tuning/regularization parameter

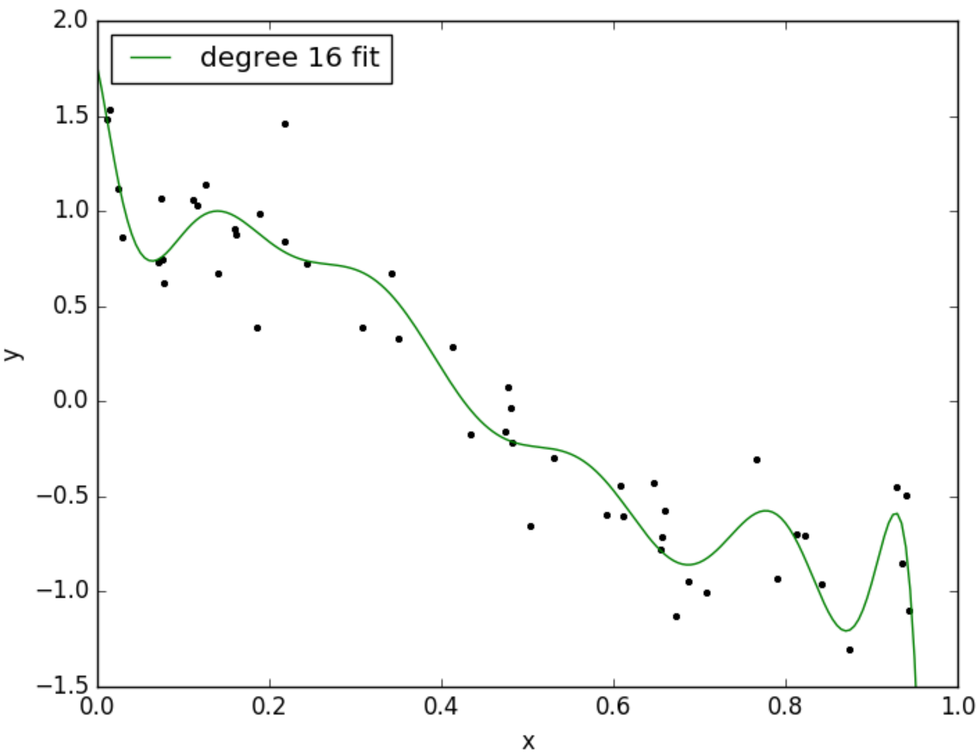
REGULARIZATION

- large λ, high bias and low variance
- Small λ, low bias and high variance
- Let us fit the higher order polynomial model using ridge regression

RIDGE REGRESSION - REGULARIZATION

Degree 16 Model

$y = 1.33e{+}06x^{16} - 6.428e{+}06x^{15} + 1.268e{+}07x^{14} - 1.378e{+}07x^{13} + 1.019e{+}07x^{12} - 4.277e{+}06x^{11} - 6.472e{+}06x^{10} + 1.821e{+}07x^{9} - 2.086e{+}07x^{8} + 1.389e{+}07x^{7} - 5.775e{+}06x^{6} + 1.504e{+}06x^{5} - 2.35e{+}05x^{4} + 1.939e{+}04x^{3} - 516.8x^{2} - 20.56x + 1.757$



RIDGE REGRESSION - REGULARIZATION

```
lambda = 1.00e-25
Learned polynomial for degree 16:
          16             15             14             13
1.33e+06 x  - 6.428e+06 x  + 1.268e+07 x  - 1.378e+07 x
              12             11             10             9
 + 1.019e+07 x  - 4.277e+06 x  - 6.472e+06 x  + 1.821e+07 x
              8             7             6             5            4
 - 2.086e+07 x + 1.389e+07 x - 5.775e+06 x + 1.504e+06 x - 2.35e+05 x
              3         2
 + 1.939e+04 x - 516.8 x - 20.56 x + 1.757
```

```
lambda = 1.00e-10
Learned polynomial for degree 16:
            16             15        14             13             12
-5.743e+04 x  + 1.191e+05 x  - 9716 x  - 7.902e+04 x  - 2.933e+04 x
              11             10             9             8        7
 + 4.548e+04 x  + 4.947e+04 x  - 1.401e+04 x - 4.864e+04 x + 5780 x
              6             5             4        3         2
 + 4.506e+04 x - 3.985e+04 x + 1.657e+04 x - 3990 x + 551.2 x - 38.04 x + 1.817
```

RIDGE REGRESSION - REGULARIZATION

```
lambda = 1.00e-06
Learned polynomial for degree 16:
        16       15         14         13         12         11
-173.1 x  + 272 x  + 129.2 x  - 117.3 x  - 210.6 x  - 102.3 x
          10         9         8         7         6        5         4
 + 97.19 x  + 206.6 x + 101.9 x - 149.9 x - 252.8 x + 63.8 x + 395.3 x
          3         2
 - 354.8 x + 108.1 x - 13.57 x + 1.46


lambda = 1.00e-03
Learned polynomial for degree 16:
       16         15       14         13         12         11
20.18 x  + 8.619 x  - 1.1 x  - 8.149 x  - 11.96 x  - 12.28 x
          10        9         8         7        6         5         4
 - 9.259 x  - 3.55 x + 3.512 x + 9.762 x + 12.2 x + 7.695 x - 4.032 x
          3         2
 - 13.02 x + 4.036 x - 2.023 x + 1.163
```

RIDGE REGRESSION - REGULARIZATION

```
lambda = 1.00e+02
Learned polynomial for degree 16:
         16            15            14            13             12
0.08259 x  + 0.06085 x  + 0.03886 x  + 0.01637 x  - 0.006929 x
            11            10            9          8          7
 - 0.03135 x  - 0.05728 x  - 0.08511 x - 0.1152 x - 0.1478 x
           6          5          4          3          2
 - 0.1827 x - 0.2188 x - 0.2526 x - 0.2764 x - 0.2731 x - 0.2057 x + 0.4131


lambda = 1.00e+03
Learned polynomial for degree 16:
          16            15            14            13            12
-0.07107 x  - 0.06893 x  - 0.06704 x  - 0.06541 x  - 0.06405 x
            11            10            9          8           7
 - 0.06293 x  - 0.06208 x  - 0.06148 x - 0.0611 x - 0.06089 x
            6           5           4          3           2
 - 0.06069 x - 0.06024 x - 0.05895 x - 0.0557 x - 0.04836 x - 0.03256 x + 0.17
```

NORMAL EQUATION - CLOSED FORM SOLUTION

$$L(\theta) = \frac{1}{2n}\Big[\sum_{i=1}^{n}\Big(h_\theta(x_i) - y_i\Big)^2 + \lambda\sum_{j=1}^{m}\theta_j^2\Big]$$

$$\min_{\theta_1,\theta_2,\ldots,\theta_m} \frac{1}{2n}\Big[\sum_{i=1}^{n}\Big(h_\theta(x_i) - y_i\Big)^2 + \lambda\sum_{j=1}^{m}\theta_j^2\Big]$$

$$\min_{\theta_1,\theta_2,\ldots,\theta_m} \frac{1}{2n}\Big[\Big(X\theta - y\Big)^T\Big(X\theta - y\Big) + \lambda\theta^T\theta\Big]$$

Minimize the loss function by taking its partial derivative and equate it to 0, to find the optimal set of parameters, θ

$$\nabla_\theta L(\theta) = \nabla_\theta\left(\frac{1}{2n}\Big[\Big(X\theta - y\Big)^T\Big(X\theta - y\Big) + \lambda\theta^T\theta\Big]\right) = 0$$

NORMAL EQUATION - CLOSED FORM SOLUTION

$$\min_{\theta_1,\theta_2,\ldots,\theta_m} \frac{1}{2n}\left[\left(X\theta - y\right)^T\left(X\theta - y\right) + \lambda\theta^T\theta\right]$$

$$\nabla_\theta L(\theta) = \frac{\partial}{\partial\theta}\left(\frac{1}{2n}\left[\left(X\theta - y\right)^T\left(X\theta - y\right) + \lambda\theta^T\theta\right]\right) = 0$$

Recall: For linear regression without regularization

$$\theta = (X^T X)^{-1} X^T y$$

For linear regression with regularization (ridge regression)

$$\theta = (X^T X + \lambda I)^{-1} X^T y$$

Now, some more Linear Algebra

Eigen Values and Eigen Vectors

FIRST, SOME BACK GROUND

- **Mean** $$\bar{x} = \frac{1}{n-1}\sum_{i=1}^{n} x_i$$
- **Variance** $$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$
- **Covariance:** $$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
- **Covariance matrix**

COVARIANCE MATRIX

- Covariance matrix for n-dimensional data set is an n by n matrix

- For example, for a 3 dimensional data set, using the dimensions p, q , r.

- The covariance matrix has 3 rows and 3 columns:

$$C = \begin{bmatrix} \text{cov}(p,p) & \text{cov}(p,q) & \text{cov}(p,r) \\ \text{cov}(q,p) & \text{cov}(q,q) & \text{cov}(q,r) \\ \text{cov}(r,p) & \text{cov}(r,q) & \text{cov}(r,r) \end{bmatrix}$$

MATRICES AND EIGEN VECTORS

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- **Scale**

$$2 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 24 \\ 16 \end{bmatrix} = 4 \times \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

EIGEN VECTORS AND EIGEN VALUES

- An *eigenvector* $\mathbf{x}$ of a linear transformation A is a non-zero vector that

  when A is applied to it, does not change its direction.

  Applying A to the eigenvector scales the eigenvector by a scalar value λ, called an *eigenvalue*

EIGEN VECTOR - PROPERTIES

- Eigen vectors can only be found for square matrices
- Not every square matrix has eigen vectors.
- Given an n x n matrix that does have eigenvectors, there are n of them

  for example, given a 3 x 3 matrix, there are 3 eigenvectors.
- Even if we scale the vector by some amount, we still get the same multiple

EIGEN VECTOR - PROPERTIES

- **Even if we scale the vector by some amount, we still get the same multiple**
- **Because all you’re doing is making it longer, not changing its direction.**
- **All the eigenvectors of a matrix are perpendicular or orthogonal.**
- **This means you can express the data in terms of these perpendicular eigenvectors.**
- **Also, when we find eigenvectors we usually normalize them to length one.**

EIGEN VALUES - PROPERTIES

- Eigenvalues are closely related to eigenvectors.
- These scale the eigenvectors
- eigenvalues and eigenvectors always come in pairs.

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 24 \\ 16 \end{bmatrix} = 4 \times \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

MORE PROPERTIES

- The trace of A is equal to the sum of its $eigenvalues$:

$$tr(A) = \sum_{i=1}^{n} \lambda_i$$

- The determinant of A is equal to the product of its $eigenvalues$:

$$|A| = \prod_{i=1}^{n} \lambda_i$$

- The rank of A is equal to the number of non-zero $eigenvalues$ of A

SPECTRAL THEOREM

Theorem: If $A \in \mathbb{R}^{m \times n}$ is symmetric matrix (meaning $A^T = A$),

then, there exist real numbers $\lambda_1, \ldots, \lambda_n$ (the eigenvalues)

and orthogonal, non-zero real vectors $\phi_1, \phi_2, \ldots, \phi_n$

(the eigenvectors) such that for each $i = 1,2,\ldots,n$:

$$A\phi_i = \lambda_i \phi_i$$

EIGEN VALUES AND VECTORS - MORE PROPERTIES

- An *eigenpair* is the pair of an eigenvalue and its associated *eigenvector*

- An *eigenspace* of A associated with λ is the space of vectors where:

  $(A - \lambda I) = 0$

- The *spectrum* of A is the set of all its *eigenvalues*:

  $\sigma(A) = \{\lambda \in C : \lambda I - A \text{ is singular}\}$

  where C is the space of all *eigenvalues* of A

- The spectral radius of A is the magnitude of its largest magnitude *eigenvalue*:

  $\rho(A) = max\{|\lambda_1|, \ldots, |\lambda_n|\}$

EXAMPLE

$$A = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

From spectral theorem:

$$A\phi = \lambda\phi$$

EXAMPLE

$$A = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

From spectral theorem:

$$A\phi = \lambda\phi \implies A\phi - \lambda I\phi = 0$$

$$(A - \lambda I)\phi = 0$$

$$\begin{bmatrix} 30-\lambda & 28 \\ 28 & 30-\lambda \end{bmatrix} = 0 \implies \lambda = 58 \text{ and } \lambda = 2$$