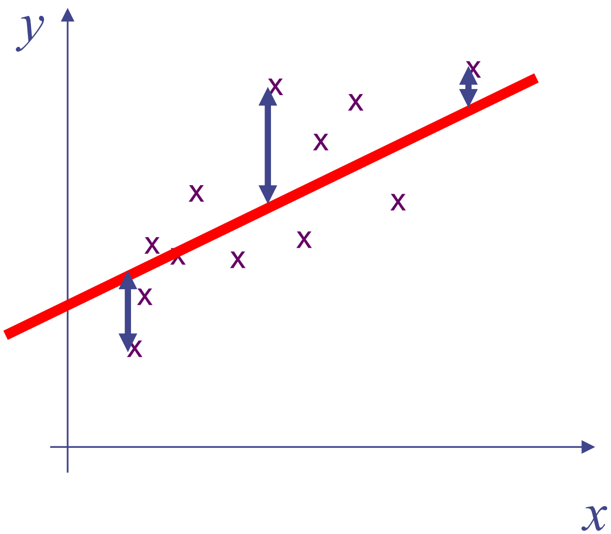
Machine Learning

4771

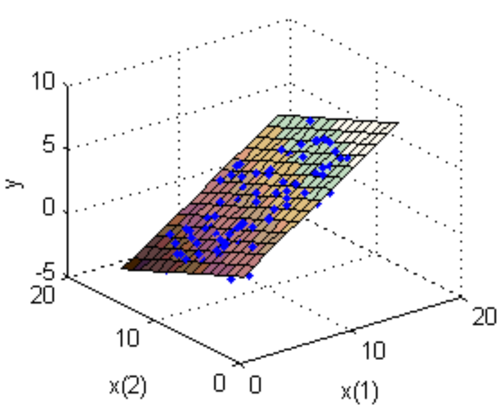
Instructor: Itsik Pe’er

Reminder: Regression

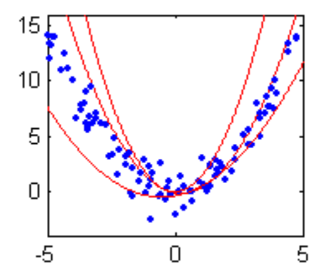


- 1D Linear:
  - Loss
  - Empirical risk
  - Least-squares

- Multi-D Linear: matrix form

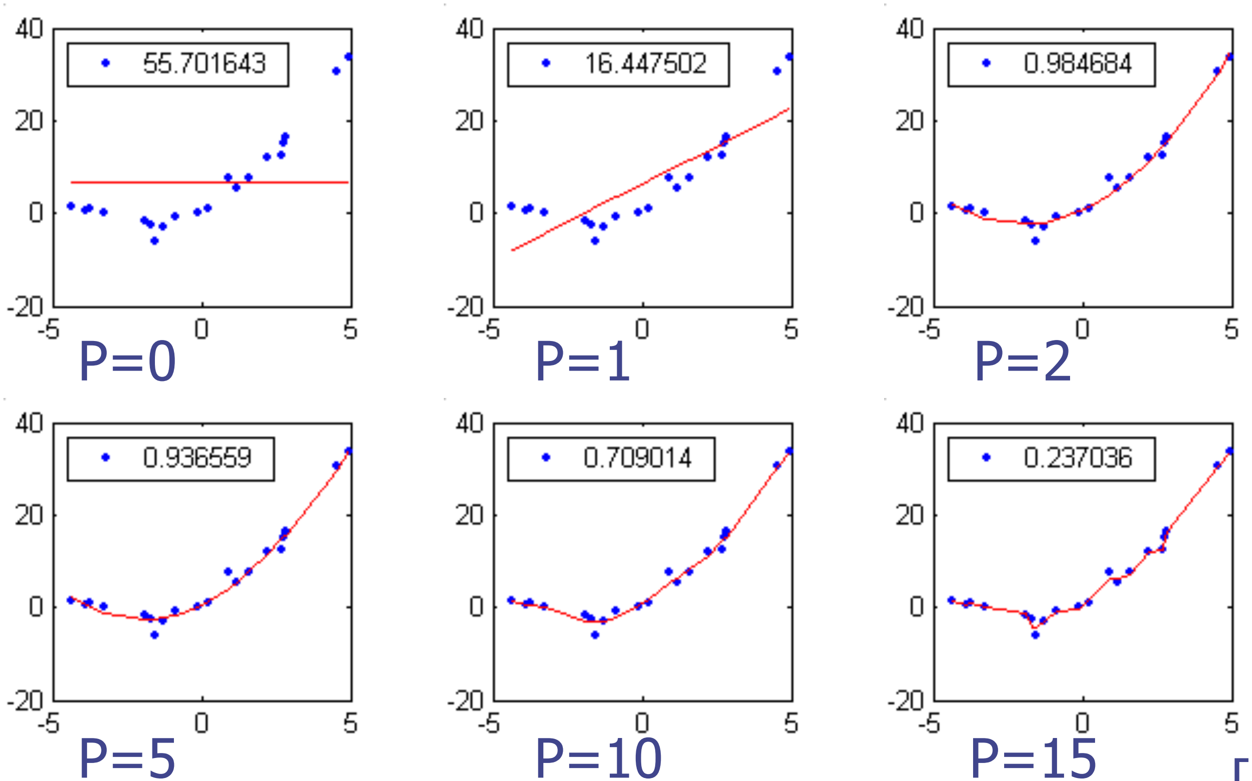


- Polynomial



Underfitting/Overfitting

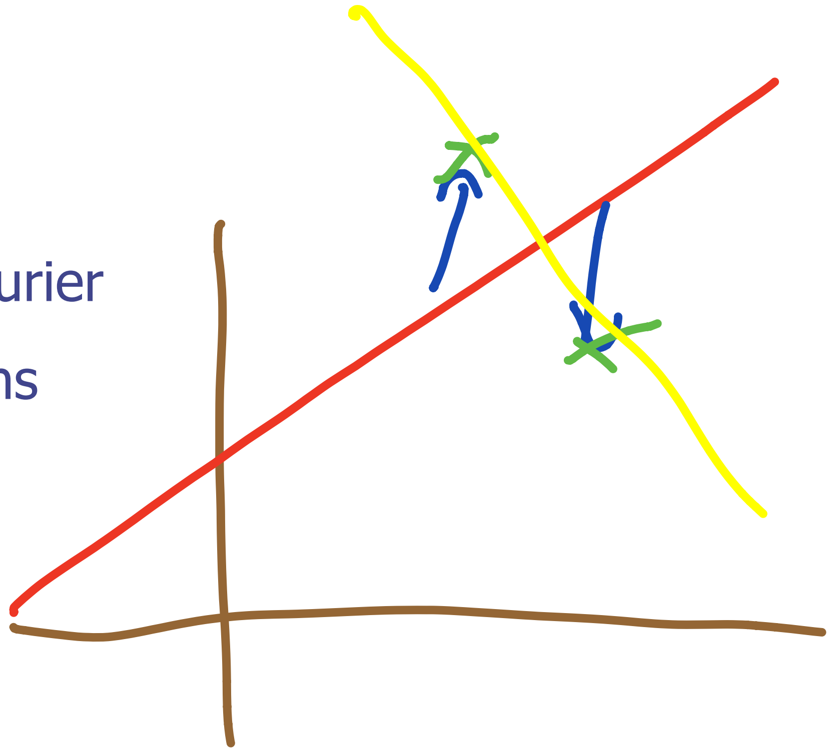
- Try varying P. Higher P fits a more complex function class
- Observe $R(\theta^*)$ drops with bigger P



[Bishop]

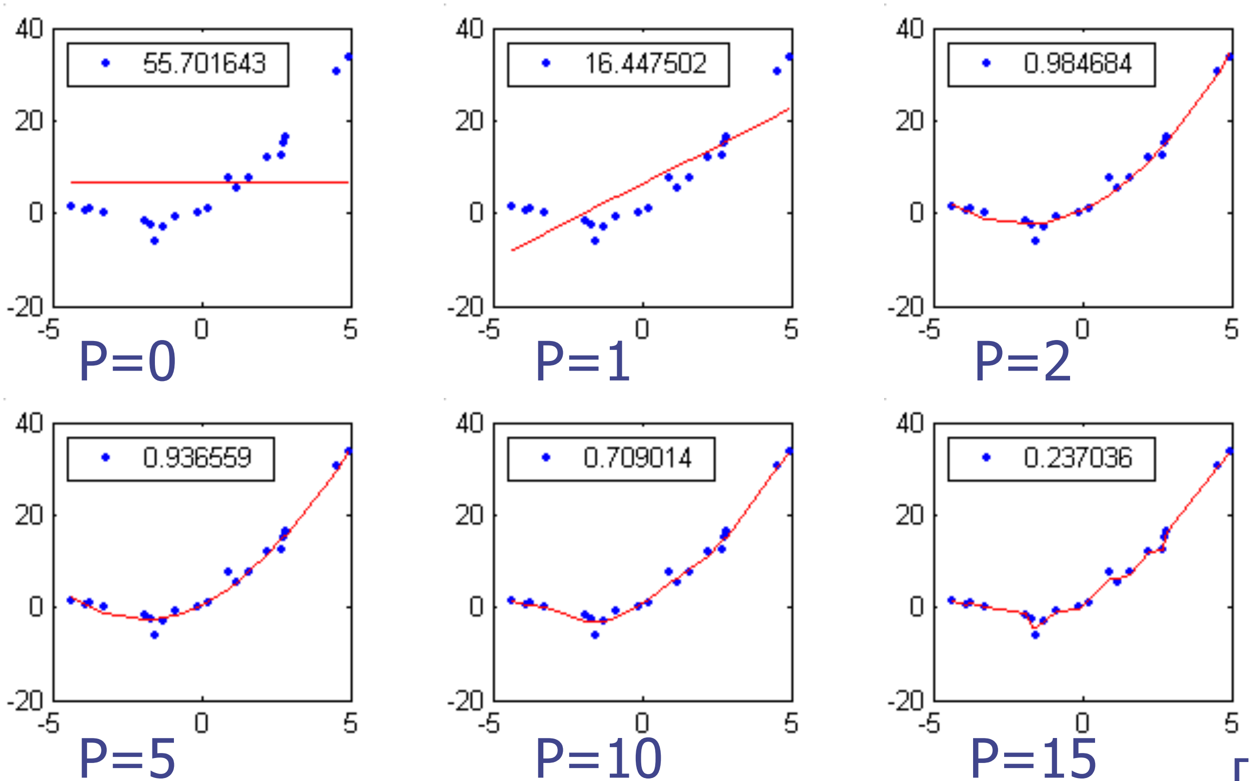
Class 4

- Overfitting
- Additive models: Fourier
- Radial basis functions



Underfitting/Overfitting

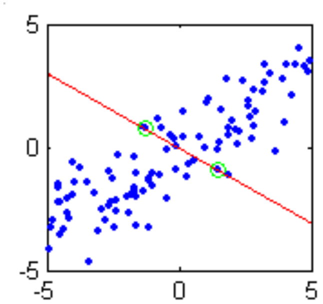
- Try varying P. Higher P fits a more complex function class
- Observe $R(\theta^*)$ drops with bigger P



[Bishop]

Evaluating The Regression

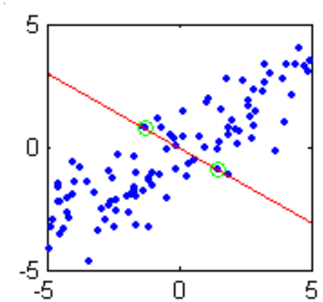
- Unfair to use empirical to find best order P
- High P (vs. N) can overfit, even linear case!
- min $R(\theta^*)$ not on training but on future data
- Want model to *Generalize* to future data



$$R(\theta) = \int (y - f(x,\theta))^2 p(x,y)\, dxy$$

Evaluating The Regression

- Unfair to use empirical to find best order P
- High P (vs. N) can overfit, even linear case!
- min $R(\theta^*)$ not on training but on future data
- Want model to *Generalize* to future data



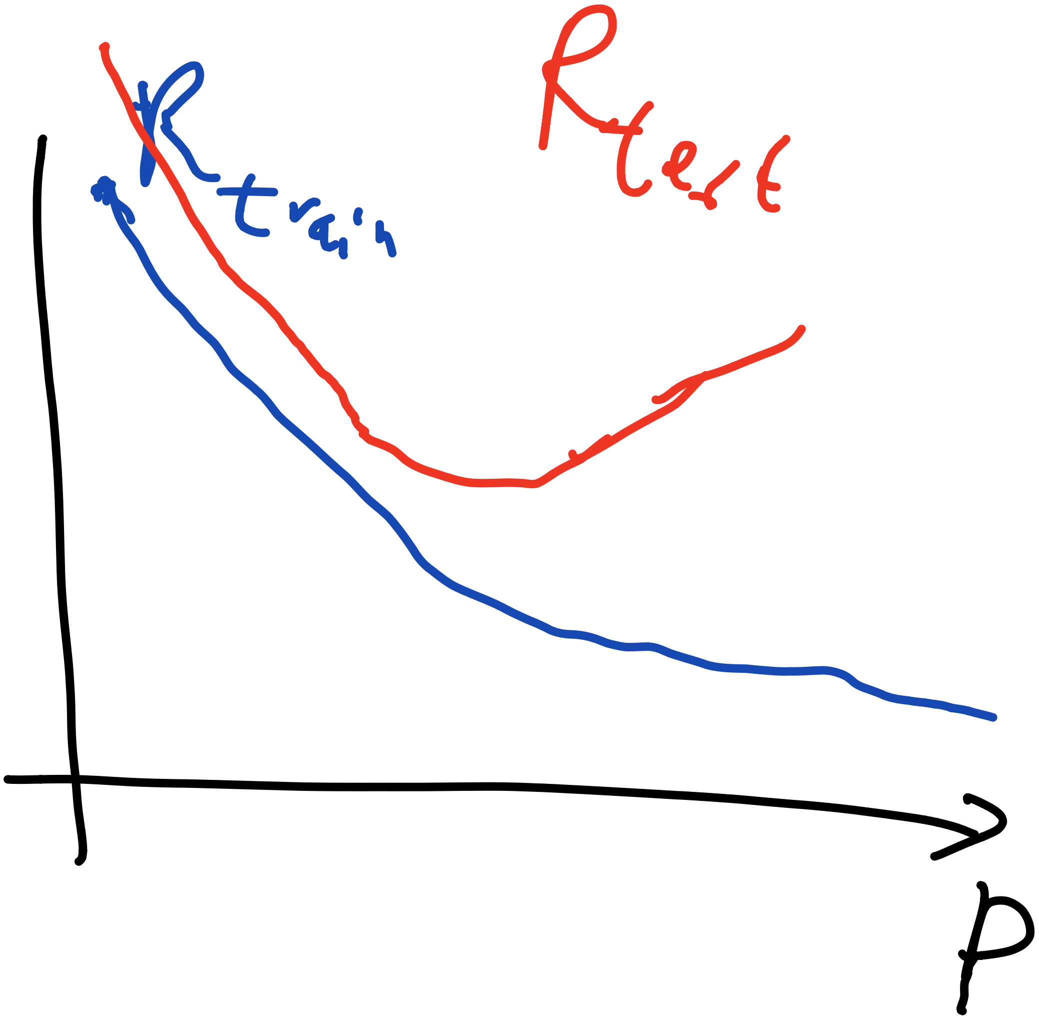
True loss: $R_{true}(\theta) = \int p(x,y)\frac{1}{2}(y-\theta^T x)^2 dx\, dy$

- One approach: split data into training / testing portion

$\{(x_1, y_1), \dots, (x_N, y_N)\}$ $\{(x_{N+1}, y_{N+1}), \dots, (x_{N+M}, y_{N+M})\}$

- Estimate θ* with training loss: $R_{train}(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \theta^T x_i)^2$
- Evaluate P with testing loss:

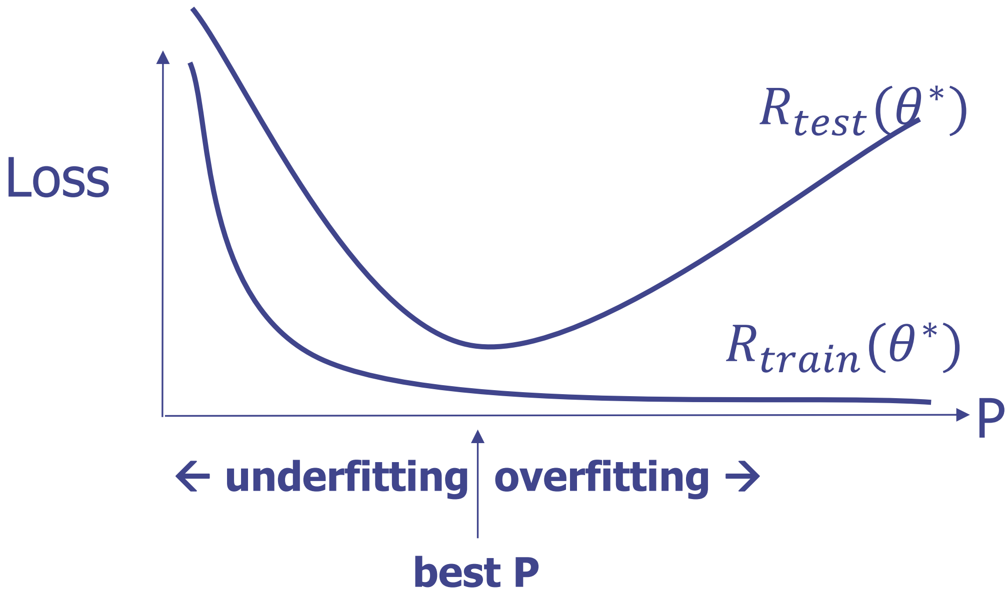
$$R_{test}(\theta^*) = \frac{1}{2M}\sum_{i=N+1}^{N+M}(y_i - \theta^{*T} x_i)^2$$



R_{train}
R_{test}
P

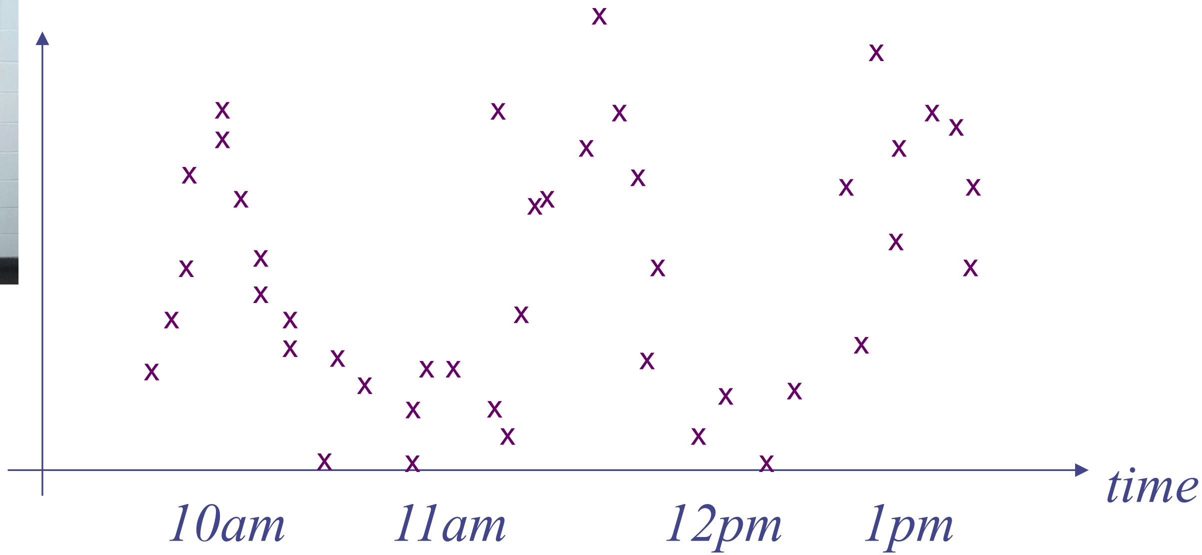
Crossvalidation

- Try fitting with different polynomial order P
- Select P which gives lowest $R_{test}(\theta^*)$

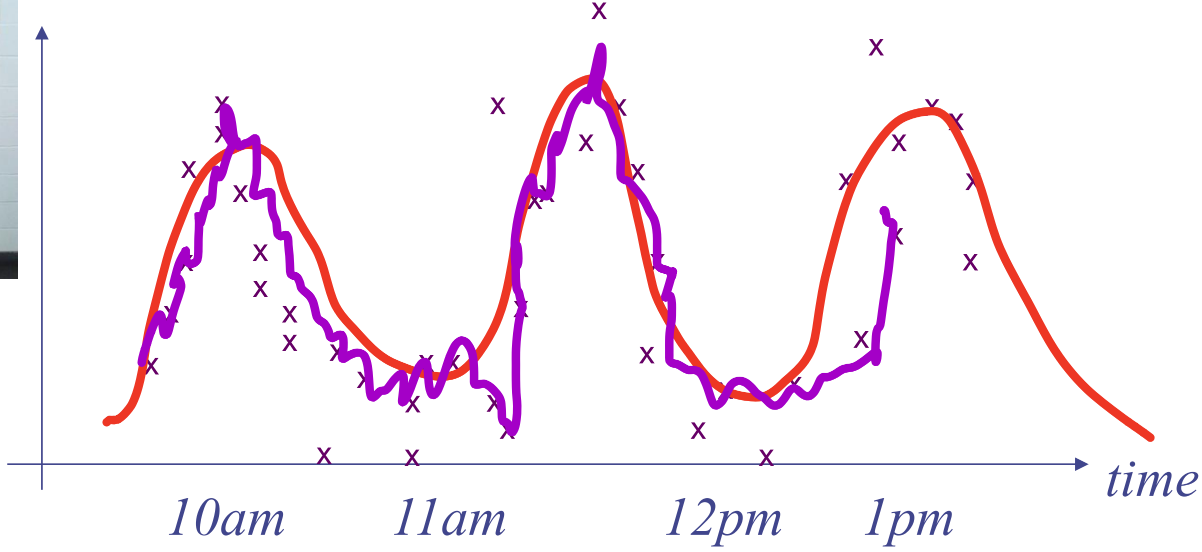


- Think of P as a measure of the complexity of the model
- Higher order polynomials are more flexible and complex

Example: Temporal data

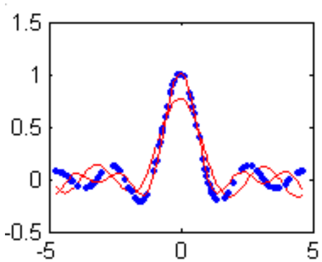


Example: Temporal data



- Need to fit periodic behavior
- Cycle: 90min, daily, weekly, annual

Sinusoidal Basis Functions



- General functions, not just polynomials:

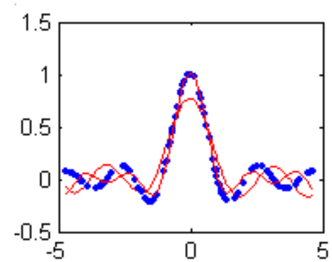
$$f(x;\theta) = \sum_{p=1}^{P} \theta_p \phi_p(x) + \theta_0$$

- These are generally called Additive Models
- Regression adds linear combinations of the basis fn's

$$\phi_{2p}(x) = \sin px$$

$$\phi_{2p+1}(x) = \cos px$$

Sinusoidal Basis Functions



- General functions, not just polynomials:

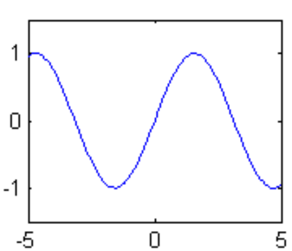
$$f(x;\theta) = \sum_{p=1}^{P} \theta_p \phi_p(x) + \theta_0$$

- These are generally called Additive Models
- Regression adds linear combinations of the basis fn's
- For example: Fourier (sinusoidal) basis

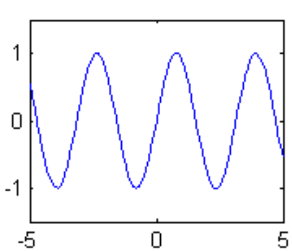
$$\phi_{2k}(x_i) = \sin(kx_i) \qquad \phi_{2k+1}(x_i) = \cos(kx_i)$$

- Note, don't have to be a basis per se, usually subset

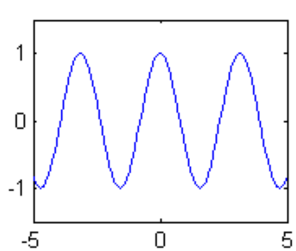
$\theta_1 \times$



$+\theta_2 \times$



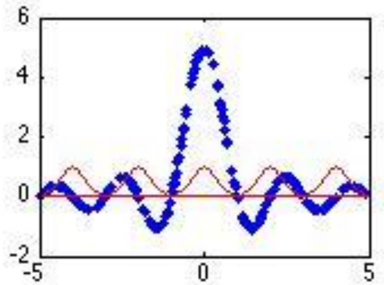
$+\theta_3 \times$



Patterson
Gimlin

Example: Bigfoot Sightings

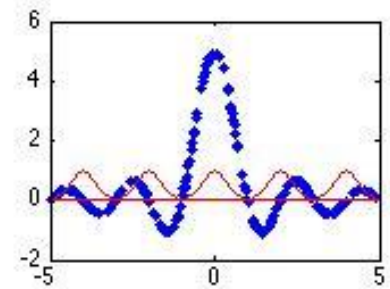
Radial Basis Functions (RBF)



- Can act as prototypes of the data itself

$$f(\boldsymbol{x};\theta) = \sum_{k=1}^{N} \theta_k \exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2\right)$$

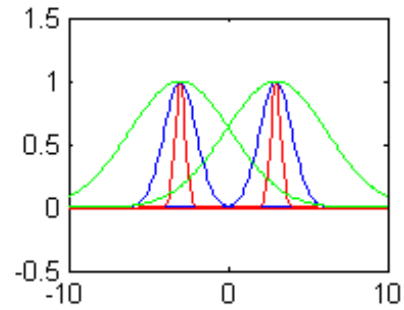
Radial Basis Functions (RBF)



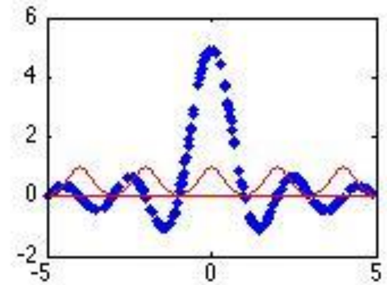
- Can act as prototypes of the data itself

$$f(\boldsymbol{x};\theta) = \sum_{k=1}^{N} \theta_k \exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2\right)$$

- Parameter σ = standard deviation controls how wide bumps are

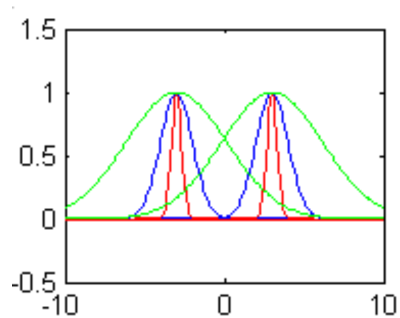


Radial Basis Functions (RBF)



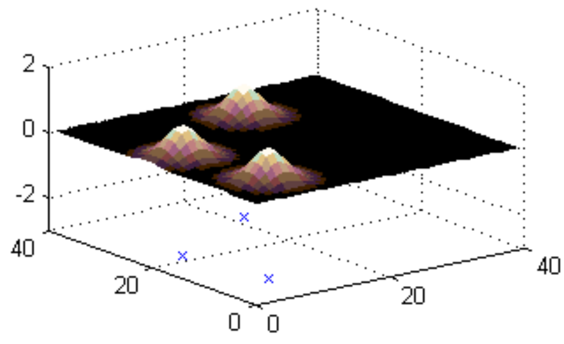
- Can act as prototypes of the data itself

$$f(\boldsymbol{x};\theta) = \sum_{k=1}^{N} \theta_k \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2)$$



- Parameter σ = standard deviation controls how wide bumps are

- Also works in multi-dimensions



Radial Basis Functions

- Training point $\rightarrow$ bump function

$$f(\boldsymbol{x};\theta) = \sum_{k=1}^{N} \theta_k \overbrace{\exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2\right)}^{\phi_k(x)}$$

- Reuse solution from linear regression: $\theta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$
- Can view the data instead as X, a big matrix of size N x N

Radial Basis Functions

- Training point $\rightarrow$ bump function

$$f(\boldsymbol{x};\theta) = \sum_{k=1}^{N} \theta_k \overbrace{\exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2)}^{\phi_k(x)}$$

- Reuse solution from linear regression: $\theta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$
- Can view the data instead as $\boldsymbol{X}$, a big matrix of size N x N

$$\boldsymbol{X} = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_k(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_k) & \phi_2(x_k) & \cdots & \phi_k(x_k) \end{bmatrix}$$

Radial Basis Functions

- Training point $\rightarrow$ bump function

$$f(\boldsymbol{x};\theta) = \sum_{k=1}^{N} \theta_k \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2)$$

- Reuse solution from linear regression: $\theta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$
- Can view the data instead as X, a big matrix of size N x N

$$\boldsymbol{X} = \begin{bmatrix} \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_1-\boldsymbol{x}_1\|^2) & \cdots & \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_1-\boldsymbol{x}_k\|^2) \\ \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_2-\boldsymbol{x}_1\|^2) & \cdots & \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_2-\boldsymbol{x}_k\|^2) \\ \vdots & \ddots & \vdots \\ \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_k-\boldsymbol{x}_1\|^2) & \cdots & \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_k-\boldsymbol{x}_k\|^2) \end{bmatrix}$$

Radial Basis Functions

- training point $\rightarrow$ bump function

$$f(\boldsymbol{x};\theta) = \sum_{k=1}^{N} \theta_k \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2)$$

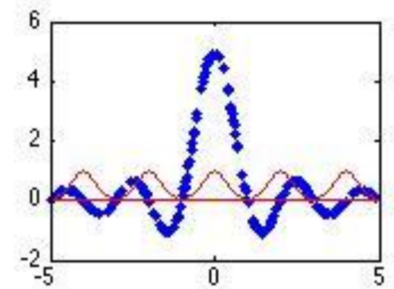
- Reuse solution from linear regression: $\theta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$
- Can view the data instead as X, a big matrix of size N x N

$$\boldsymbol{X} = \begin{bmatrix} \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_1-\boldsymbol{x}_1\|^2) & \cdots & \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_1-\boldsymbol{x}_k\|^2) \\ \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_2-\boldsymbol{x}_1\|^2) & \cdots & \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_2-\boldsymbol{x}_k\|^2) \\ \vdots & \ddots & \vdots \\ \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_k-\boldsymbol{x}_1\|^2) & \cdots & \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}_k-\boldsymbol{x}_k\|^2) \end{bmatrix}$$

- For RBFs, X is square and symmetric, so solution is just

$$\theta^* = \boldsymbol{X}^{-1}y$$

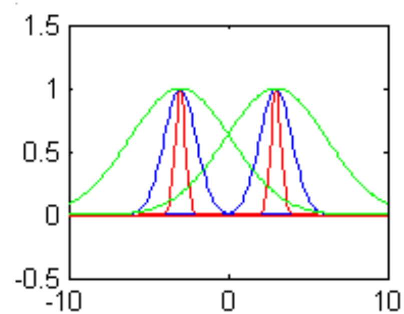
Bump Width for RBF



- Can act as prototypes of the data itself

$$f(\boldsymbol{x};\theta)=\sum_{k=1}^{N}\theta_k \exp(-\frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2)$$

- Parameter σ = standard deviation controls how wide bumps are



What happens if too big/small?

How would we know that?

Evaluating Our Learned Function

- We minimized empirical risk to get θ^*
- How well does $f(x;\theta^*)$ perform on future data?
- It should *Generalize* and have low True Risk:

$$R_{true}(\theta) = \int P(x,y)\frac{1}{2}(y-\theta^T x)^2 dx\, dy$$

- Can't compute true risk, instead use Testing Empirical Risk
- We randomly split data into training and testing portions

$$\{(x_1,y_1),\ldots,(x_N,y_N)\} \qquad \{(x_{N+1},y_{N+1}),\ldots,(x_{N+M},y_{N+M})\}$$

- Find θ^* with training data:

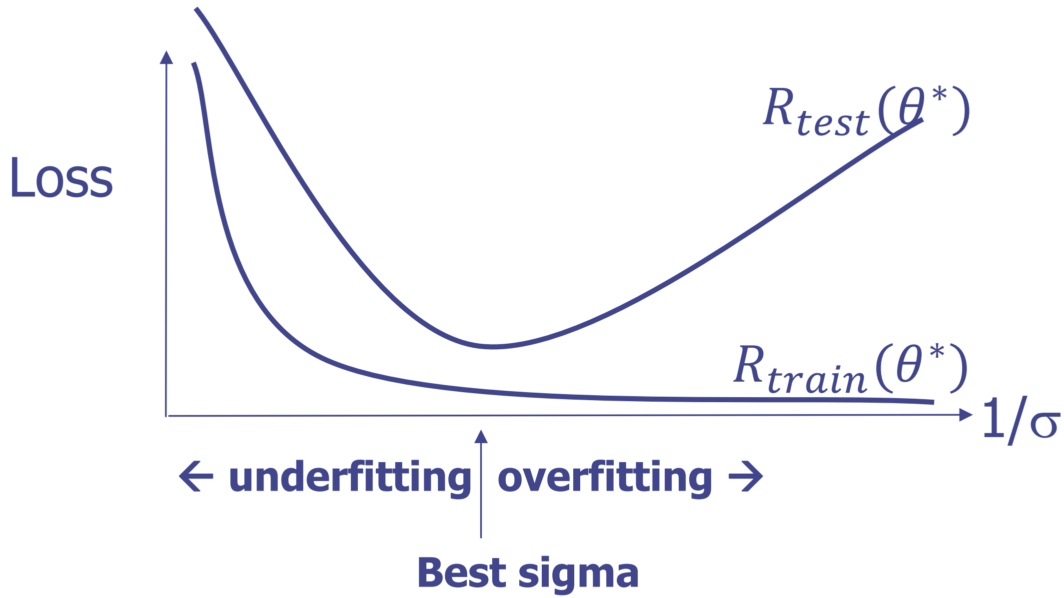
$$R_{train}(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i-\theta^T x_i)^2$$

- Evaluate it with testing data:

$$R_{train}(\theta) = \frac{1}{2M}\sum_{i=N+1}^{N+M}(y_i-\theta^T x_i)^2$$

Crossvalidation

- Try fitting with different sigma radial basis function widths
- Select sigma which gives lowest $R_{test}(\theta *)$



- Think of sigma as a measure of the simplicity of the model
- Thinner RBFs are more flexible and complex