MACHINE LEARNING

CHAPTER 2: LINEAR MODELS FOR REGRESSION

LINEAR MODELS FOR REGRESSION

1. The concept of regression
2. Maximum Likelihood and Least Square
3. Over-fitting and Regularization
4. The Bias-Variance Trade-off
5. Bayesian Linear Regression
6. Sparse regression

What is Regression?

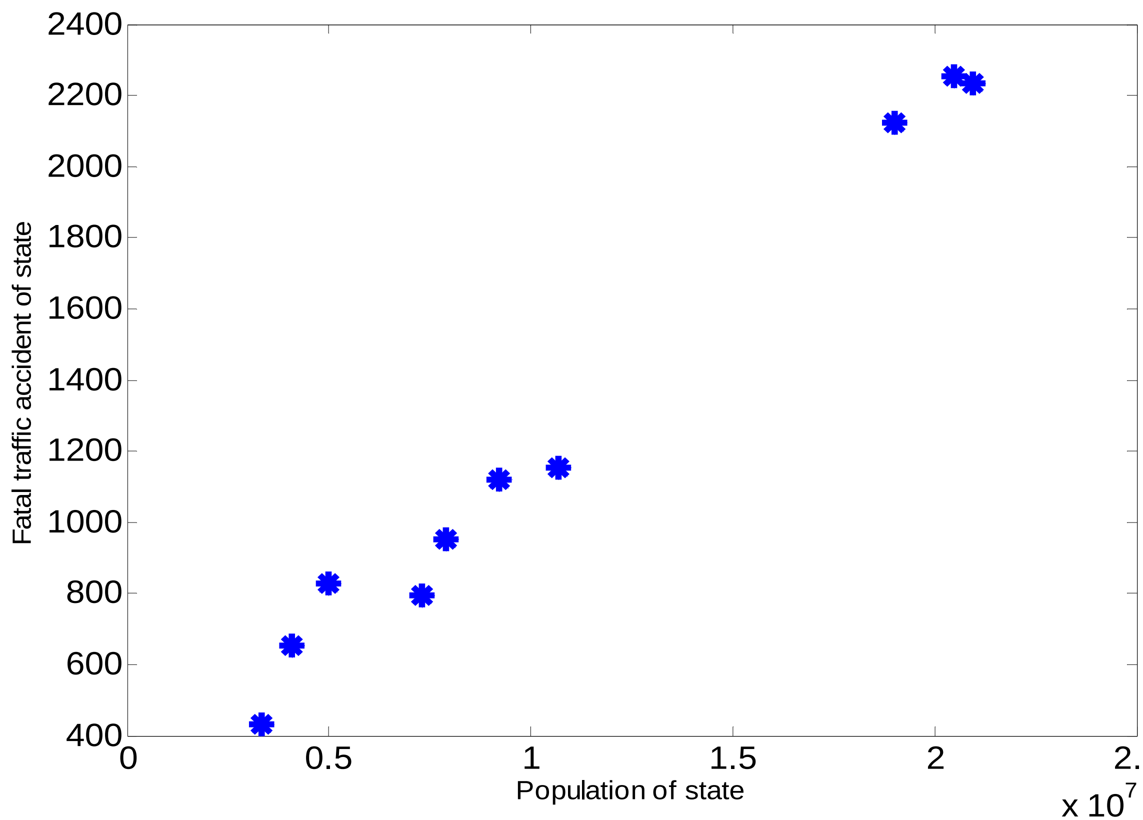
Regression is the process of learning the relationship between a set of input (explanatory) and output (response) variables, such that given new instances of input variables, their corresponding output variables can be predicted.

What is Regression?

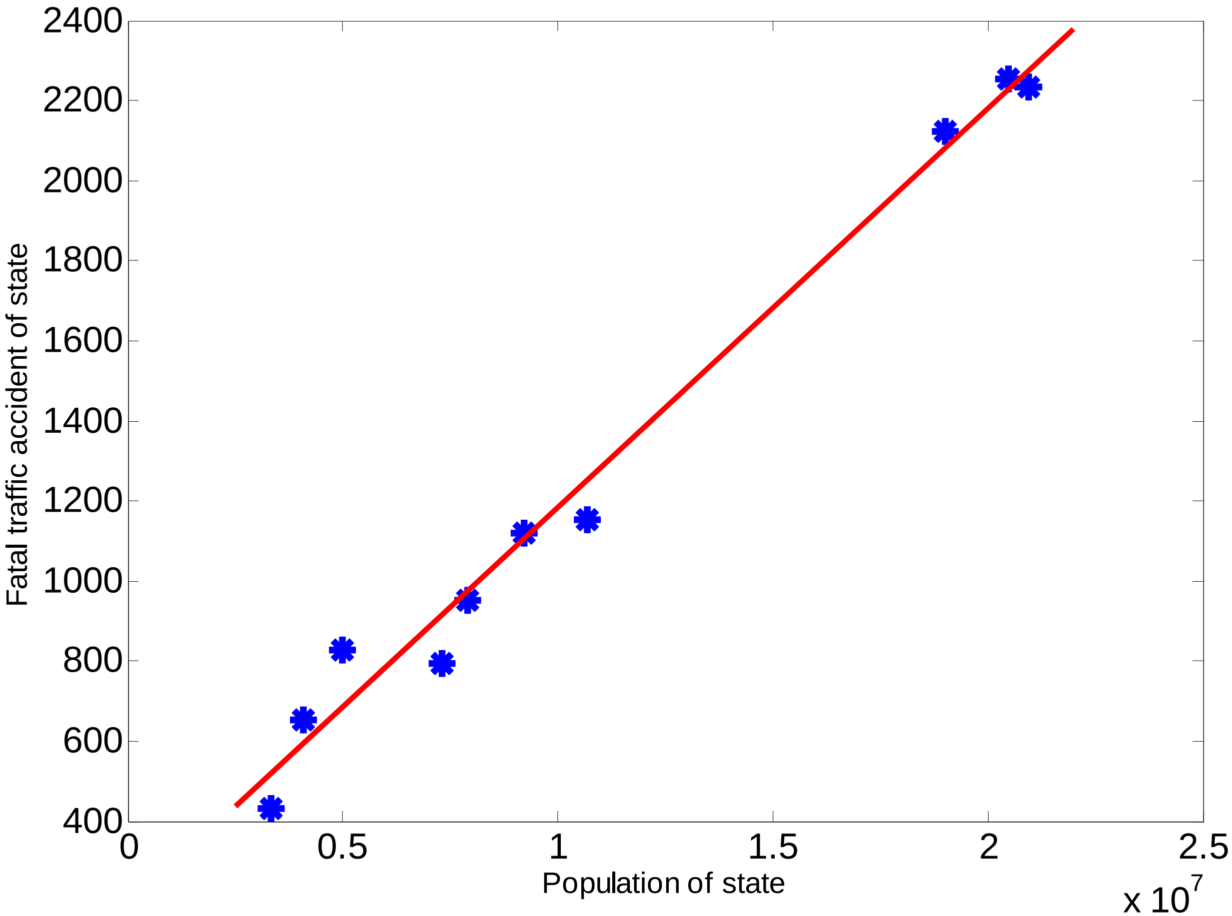
Example for linear basis functions

Input: population of state

Output: fatal traffic accident of state



What is Regression?



$$y(x, \mathbf{w}) = w_0 + w_1 x$$

$$w_0 = 177.3$$

$$w_1 = 0.0001$$

What is Regression?

Age estimation

$$\rightarrow \; 21$$

Head pose estimation

$$\rightarrow \begin{bmatrix} 20^\circ \\ 5^\circ \\ 0^\circ \end{bmatrix}$$

What is Linear Regression?

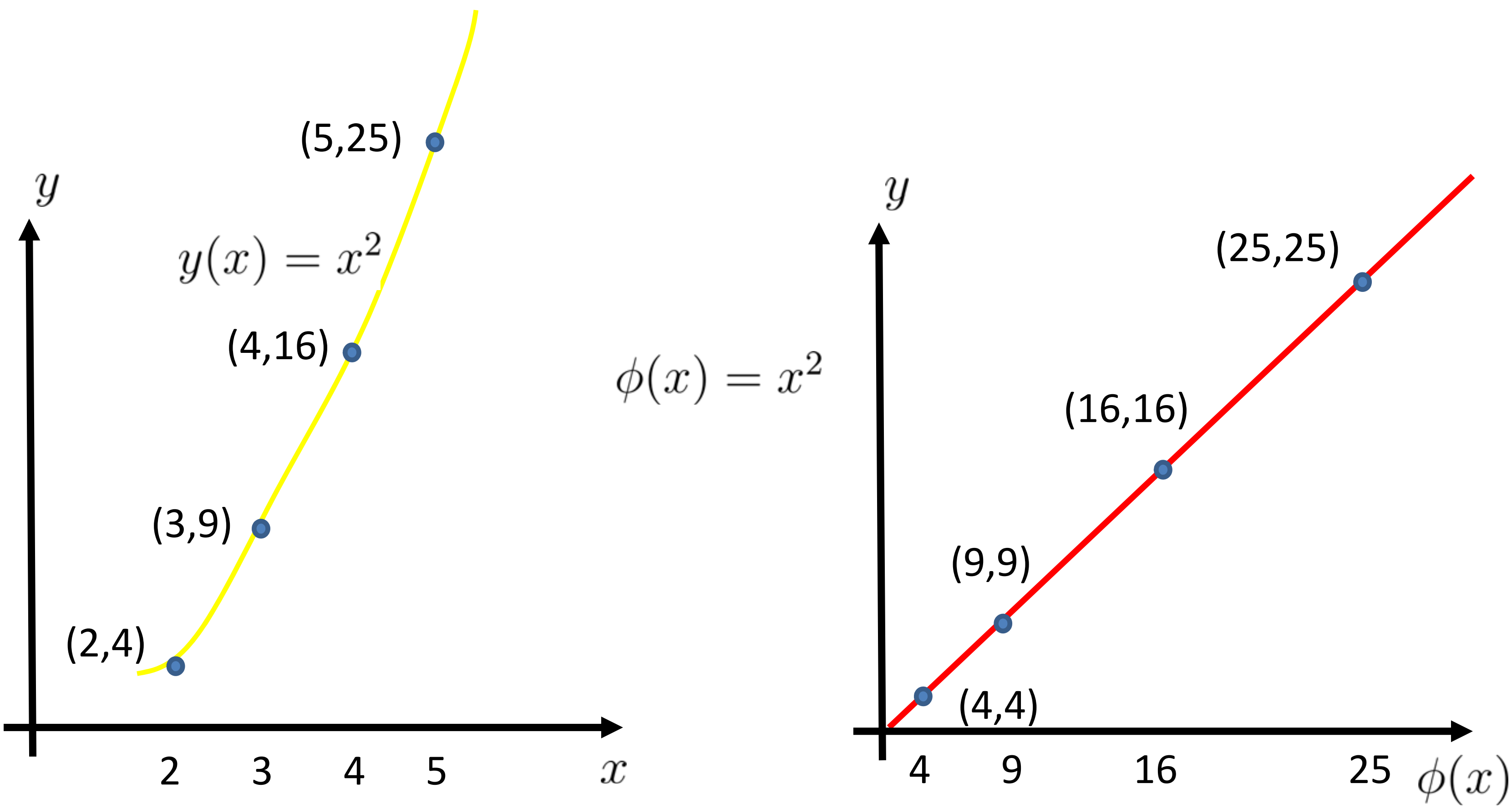
The simplest linear model for regression is one that involves a linear combination of the input variables

$$y\left(\mathbf{x},\mathbf{w}\right)=w_0+w_1x_1+...+w_Dx_D$$

where $\mathbf{x}=\left(x_1,...x_D\right)^T$. This is often simply known as ***linear regression***.

The key property of this model is that it is a linear function of the parameters $w_0...w_D$.

Linear Basis Function Models



Linear Basis Function Models

Generally

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

where $\phi_j(\mathbf{x})$ are known as *basis functions*.

Typically, $\phi_0(\mathbf{x}) = 1$, so that w_0 acts as a bias.

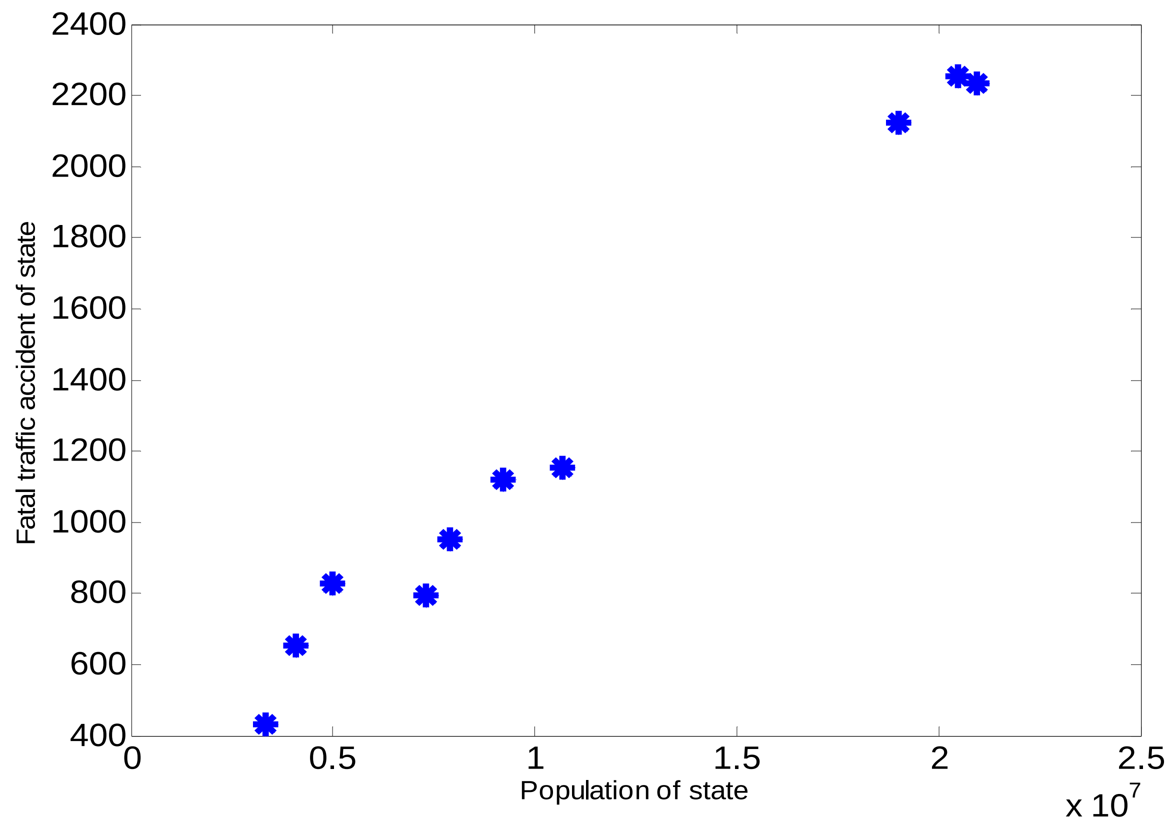
In the simplest case, we use linear basis functions : $\phi_d(\mathbf{x}) = x_d$.

Linear Basis Function Models

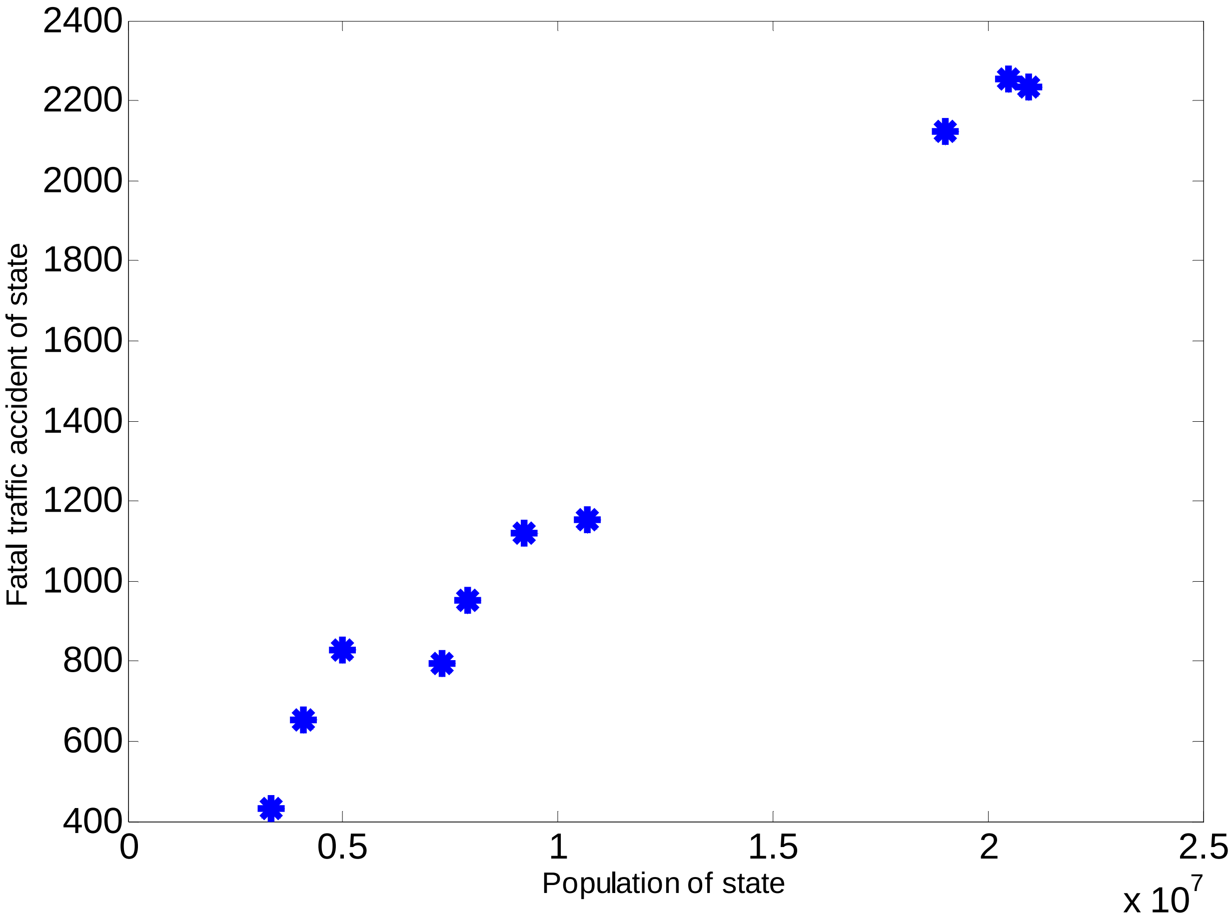
Example for linear basis functions

Input: population of state

Output: fatal traffic accident of state

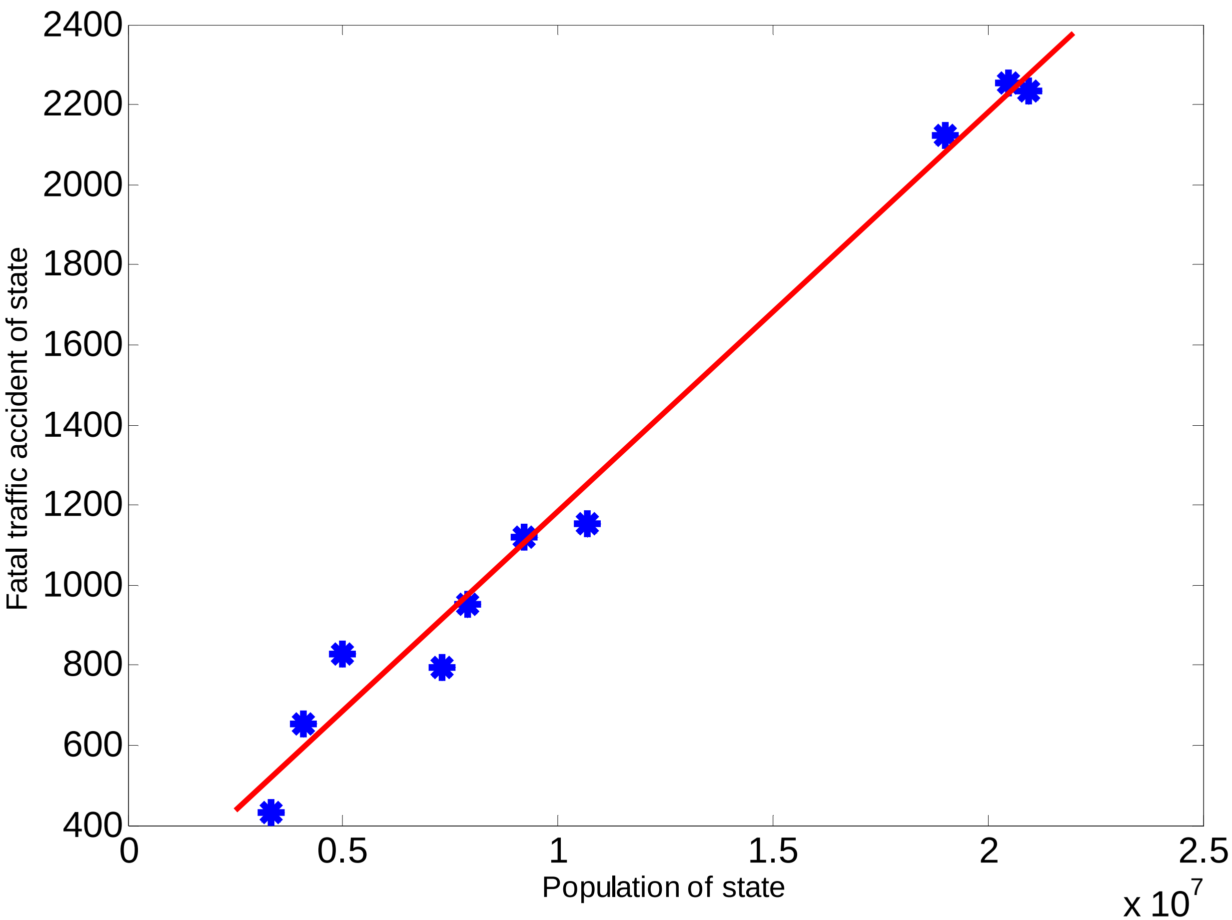


Linear Basis Function Models



$$y\left(x,\mathbf{w}\right)=w_0+w_1x$$

Linear Basis Function Models



$$y(x, \mathbf{w}) = w_0 + w_1 x$$

$$w_0 = 177.3$$

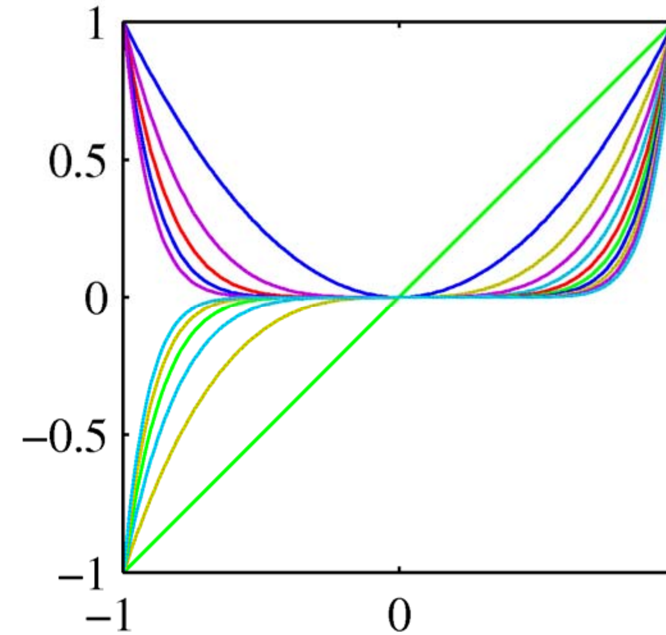
$$w_1 = 0.0001$$

Linear Basis Function Models

Polynomial basis functions:

$$\phi_j(x) = x^j.$$

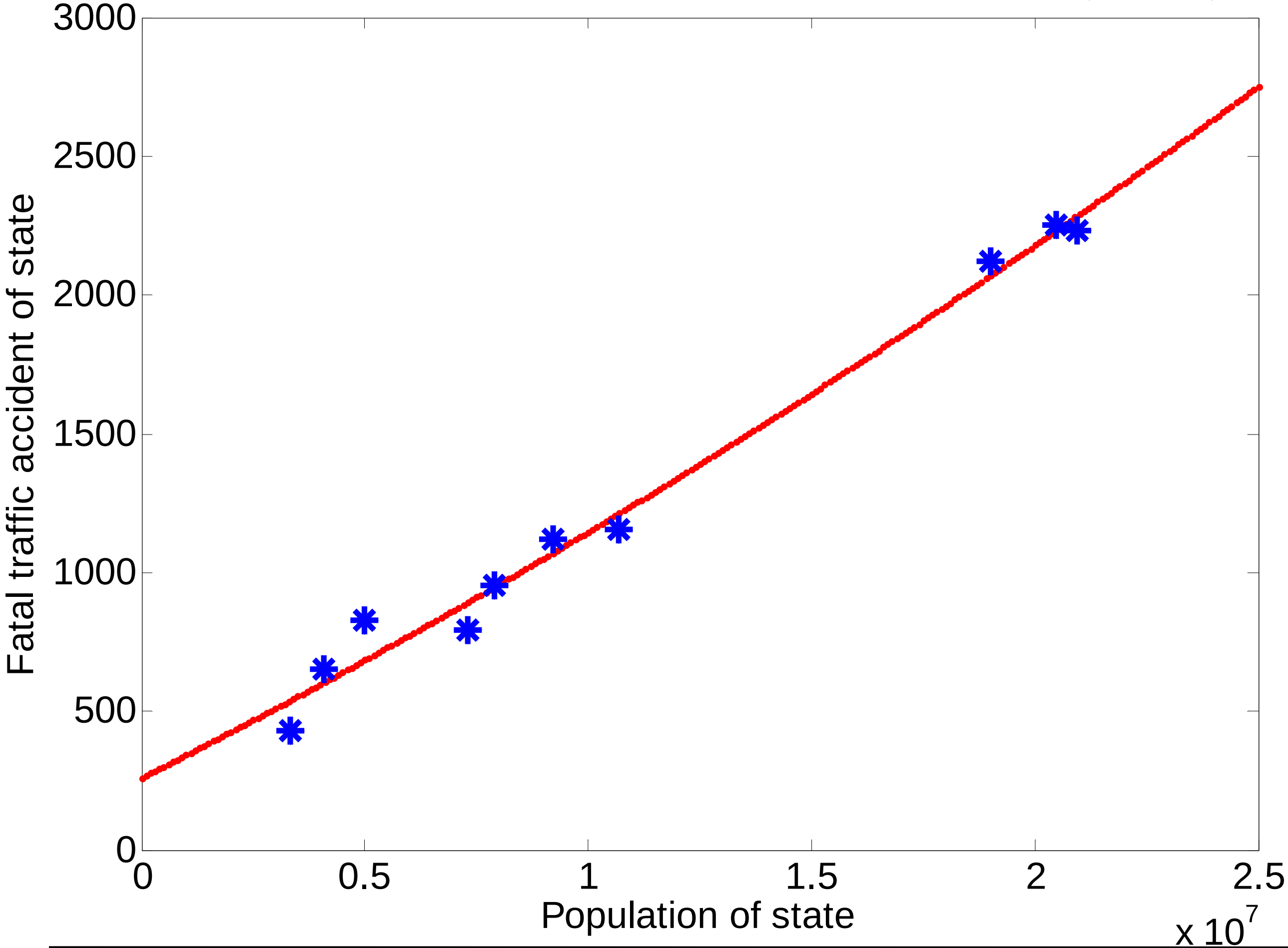
These are global; a small change in x affect all basis functions.



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Linear Basis Function Models

$$y\left(x, \mathbf{w}\right) = w_0 + w_1 x + w_2 x^2$$



$$w_0 = 260.8$$

$$w_1 = 8.1237 \times 10^{-5}$$

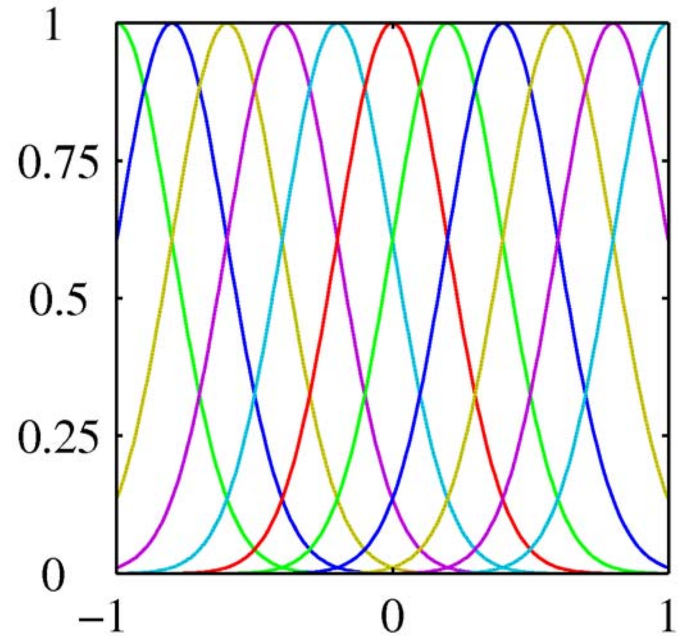
$$w_2 = 7.3671 \times 10^{-13}$$

Linear Basis Function Models

Gaussian basis functions:

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

These are local; a small change in x only affect nearby basis functions. μ_j and s control location and scale (width).



Linear Basis Function Models

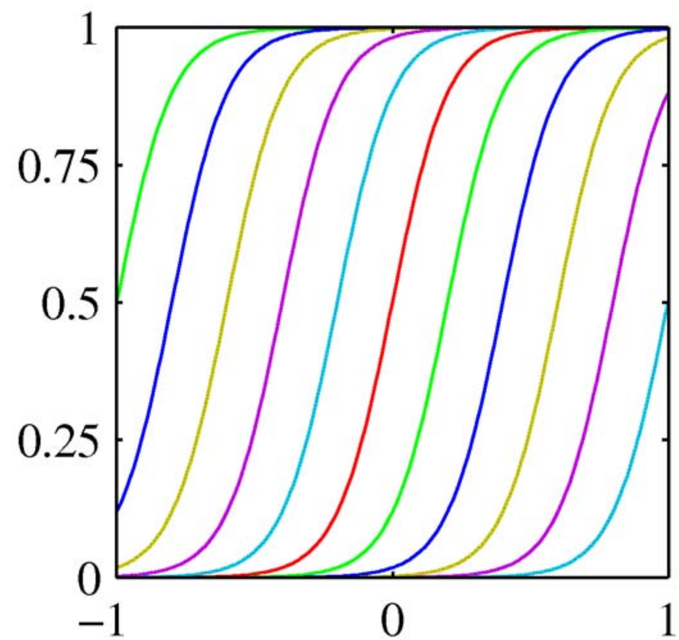
Sigmoidal basis functions:

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

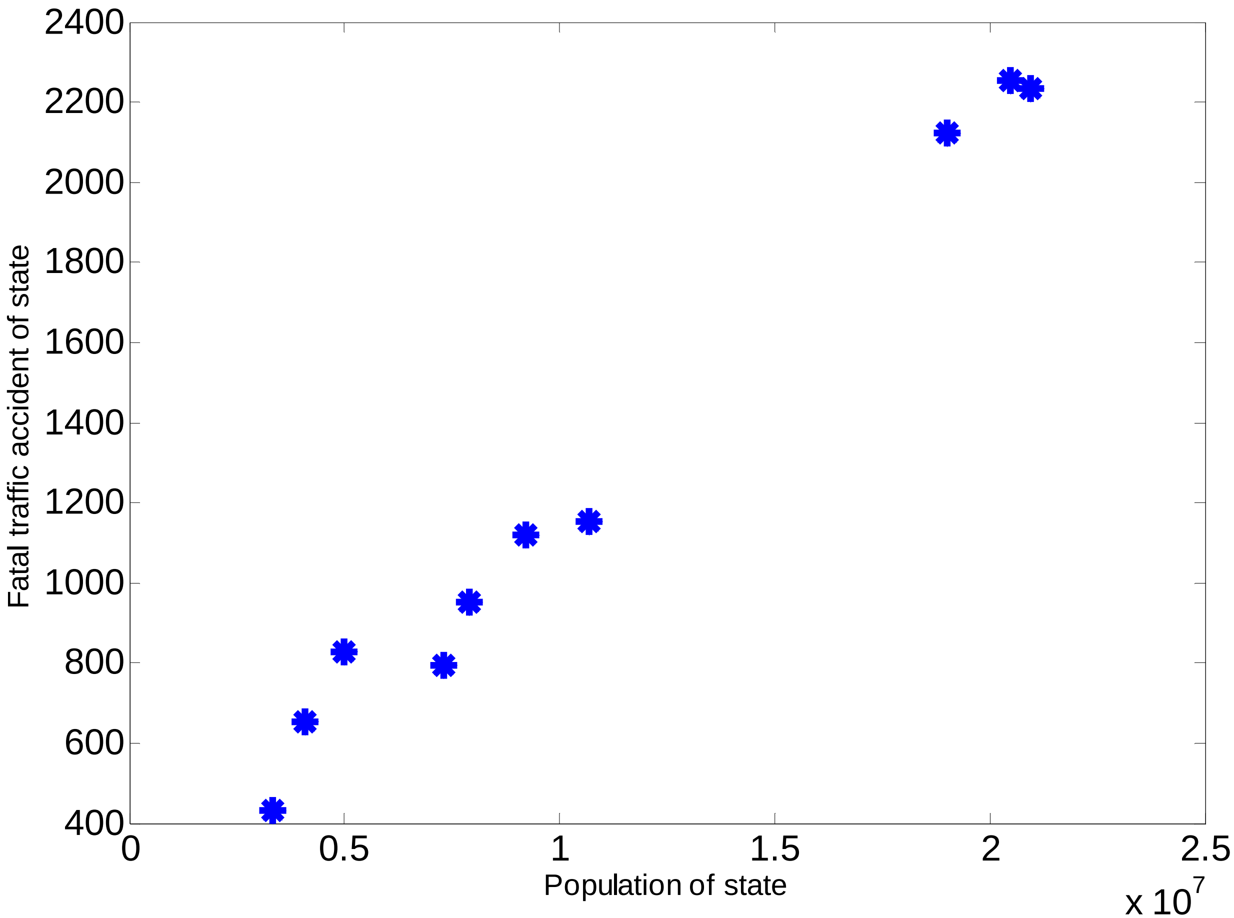
Also these are local; a small change in x only affect nearby basis functions. μ_j and s control location and scale (slope).



LINEAR MODELS FOR REGRESSION

1. The concept of regression
2. Maximum Likelihood and Least Square
3. Over-fitting and Regularization
4. The Bias-Variance Trade-off
5. Bayesian Linear Regression
6. Sparse regression

Maximum Likelihood and Least Square



$$y(x, \mathbf{w}) = w_0 + w_1 x$$

Maximum Likelihood and Least Square

Population	$\times 10^7$	Fatal traffic accident	$\times 10^3$
x_1	2.0493	t_1	2.2538
x_2	1.0676	t_2	1.1531
x_3	2.0939	t_3	2.2342
x_4	0.7897	t_4	0.9514
x_5	0.7299	t_5	0.7950
x_6	0.4996	t_6	0.8286
x_7	1.9002	t_7	2.1222
x_8	0.4083	t_8	0.6541
x_9	0.9201	t_9	1.1208
x_{10}	0.3346	t_{10}	0.4322

Maximum Likelihood and Least Square

Sum-of-squares error

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}(w_0 + w_1 x_n - t_n)^2$$

where $N = 10$.

Maximum Likelihood and Least Square

Sum-of-squares error in matrix form

$$E(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\right\|^2$$

Where

$$\mathbf{w} = (w_0, w_1)^T \qquad \mathbf{\Phi} = \begin{pmatrix} 1 & \mathrm{x}_1 \\ \vdots & \vdots \\ 1 & \mathrm{x}_{10} \end{pmatrix} \quad \mathbf{t} = (t_1, \ldots, t_{10})^T$$

Maximum Likelihood and Least Square

The sum-of-squares error

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|^2 = \frac{1}{2}(\mathbf{\Phi}\mathbf{w} - \mathbf{t})^T(\mathbf{\Phi}\mathbf{w} - \mathbf{t})$$

$$= \frac{1}{2}\left(\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - 2\mathbf{w}^T\mathbf{\Phi}^T\mathbf{t} + \mathbf{t}^T\mathbf{t}\right)$$

Maximum Likelihood and Least Square

The gradient of $E(\mathbf{w})$

$$\nabla E(\mathbf{w}) = \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} - \mathbf{\Phi}^T \mathbf{t}$$

Let

$$\nabla E(\mathbf{w}) = \mathbf{0}$$

the optimal solution for $\mathbf{w}$ is

$$\mathbf{w}^* = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{t}$$

Maximum Likelihood and Least Square

Population	$\times 10^7$	Traffic accident	$\times 10^3$
x_1	2.0493	t_1	2.2538
x_2	1.0676	t_2	1.1531
x_3	2.0939	t_3	2.2342
x_4	0.7897	t_4	0.9514
x_5	0.7299	t_5	0.7950
x_6	0.4996	t_6	0.8286
x_7	1.9002	t_7	2.1222
x_8	0.4083	t_8	0.6541
x_9	0.9201	t_9	1.1208
x_{10}	0.3346	t_{10}	0.4322

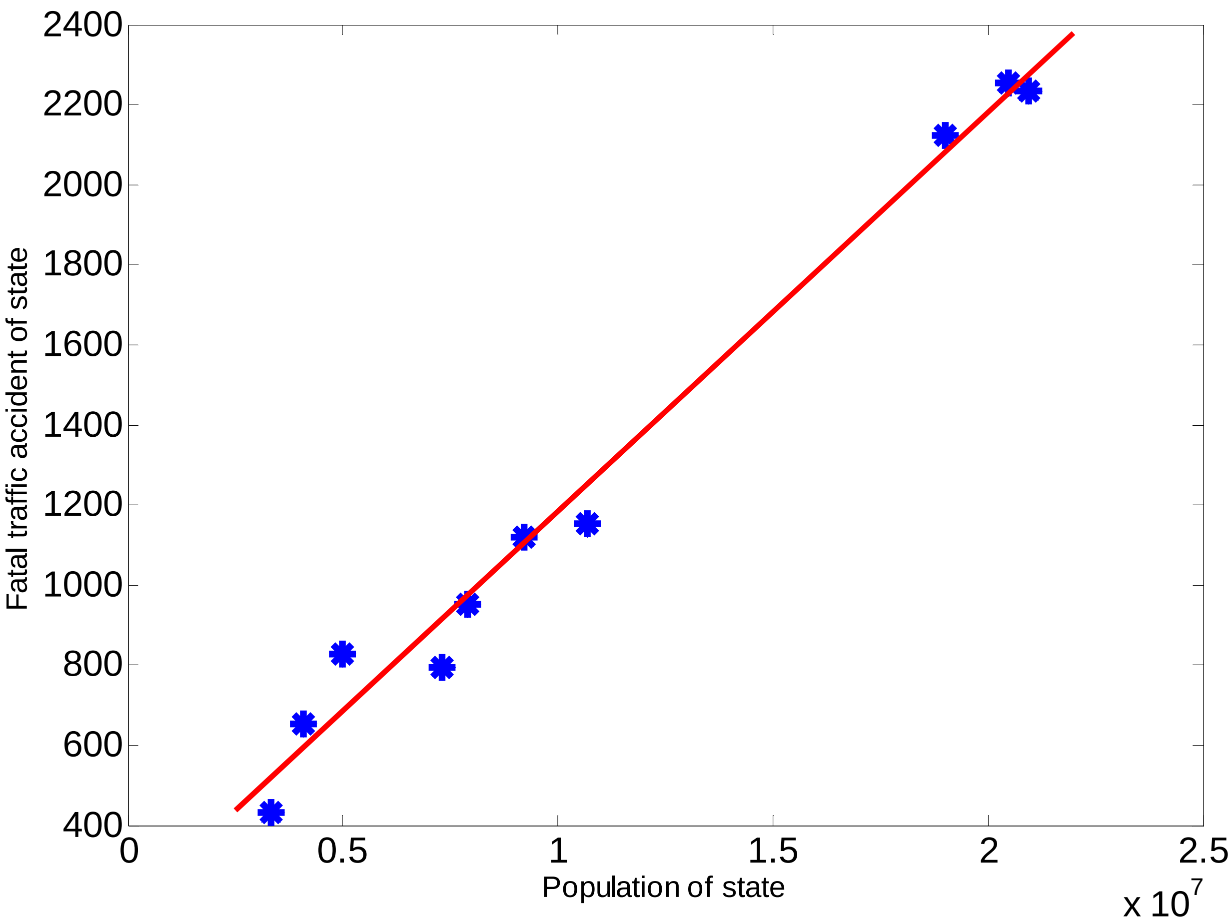
$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & \mathrm{x}_1 \\ \vdots & \vdots \\ 1 & \mathrm{x}_{10} \end{pmatrix}$$

$$\mathbf{t} = (t_1, \ldots, t_{10})^T$$

$$\boxed{\mathbf{w}^* = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \mathbf{t}}$$

$$= (177.3, 0.0001)^T$$

Maximum Likelihood and Least Square



$$y\left(x, \mathbf{w}\right) = w_0 + w_1 x$$

$$w_0 = 177.3$$

$$w_1 = 0.0001$$

Maximum Likelihood and Least Square

Least squares:

$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$: N observations of input variables

$\mathbf{t} = (t_1, ..., t_N)^T$: N observations of target variable t

Regression model:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Maximum Likelihood and Least Square

We minimize

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left\{\mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_{\boldsymbol{n}}\right) - t_n\right\}^2$$

Maximum Likelihood and Least Square

Write in matrix form

$$E(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right\|_2^2$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix} \qquad \mathbf{t} = (t_1, ..., t_N)^T$$

Maximum Likelihood and Least Square

The gradient of $E(\mathbf{w})$

$$\nabla E(\mathbf{w}) = \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} - \mathbf{\Phi}^T \mathbf{t}$$

Let

$$\nabla E(\mathbf{w}) = \mathbf{0}$$

the optimal solution for $\mathbf{w}$ is

$$\mathbf{w}^* = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{t}$$

Maximum Likelihood and Least Square

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

If N < M, $\mathbf{\Phi}^T\mathbf{\Phi}$ is underdetermined and $\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}$ does not exist.

If N > M, $\mathbf{\Phi}^T\mathbf{\Phi}$ is overdetermined and $\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}$ does exist.

Maximum Likelihood and Least Square

In statistics, **maximum-likelihood estimation** (**MLE**) is a method of estimating the parameters of a statistical model given data.

For parameters $\boldsymbol{\theta}$, the joint distribution for all observations is $p\left(x_1, ..., x_N | \boldsymbol{\theta}\right)$. We let $\mathcal{L}\left(\boldsymbol{\theta}\right)$ denote this joint distribution and name it as likelihood function of parameters $\boldsymbol{\theta}$.

Maximum Likelihood and Least Square

The maximum likelihood estimation of parameters $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}_{ML} = \mathrm{argmax}\mathcal{L}\left(\boldsymbol{\theta}\right)$$

Maximum Likelihood and Least Square

Assume observations from a deterministic function with added Gaussian noise:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \qquad \text{where} \qquad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

which is the same as saying,

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

Given observed inputs, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and targets, $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$, we obtain the likelihood function

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

Maximum Likelihood and Least Square

Observations:

$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$

Parameters:

$\mathbf{w}$ and β

the likelihood function:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

$$\boxed{\mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}} = \operatorname{argmax} p\left(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta\right)}$$

Maximum Likelihood and Least Square

Taking the logarithm, we get

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w},\beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

where

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

is the sum-of-squares error.

Maximum Likelihood and Least Square

Computing the gradient and setting it to zero yields

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = \mathbf{0}.$$

Solving for $\mathbf{w}$, we get

$$\mathbf{w}_{\mathrm{ML}} = \overbrace{\left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}}\mathbf{t}$$

The Moore-Penrose pseudo-inverse, $\boldsymbol{\Phi}^{\dagger}$.

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Maximum Likelihood and Least Square

Maximizing with respect to the bias, w_0, alone, we see that

$$
\begin{aligned}
w_0 &= \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j} \\
&= \frac{1}{N}\sum_{n=1}^{N} t_n - \sum_{j=1}^{M-1} w_j \frac{1}{N}\sum_{n=1}^{N} \phi_j(\mathbf{x}_n).
\end{aligned}
$$

We can also maximize with respect to β, giving

$$
\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2
$$

Geometry of Least Squares

Consider

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M]\,\mathbf{w}_{\mathrm{ML}}.$$

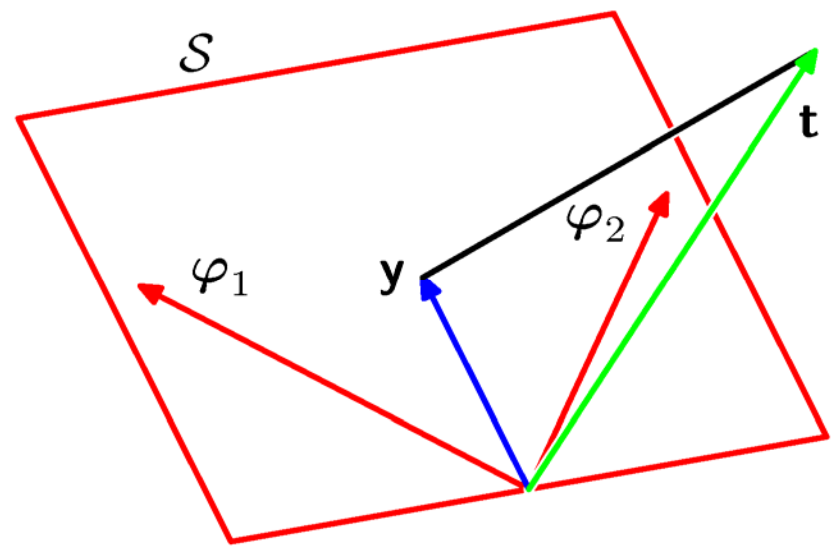
$$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T} \qquad \mathbf{t} \in \mathcal{T}$$

N-dimensional

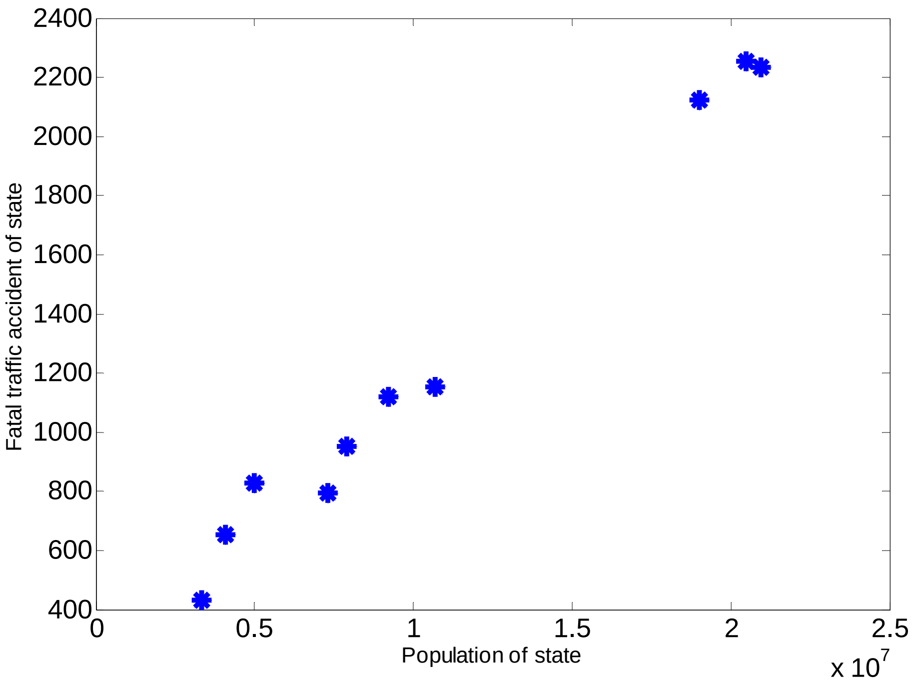
M-dimensional

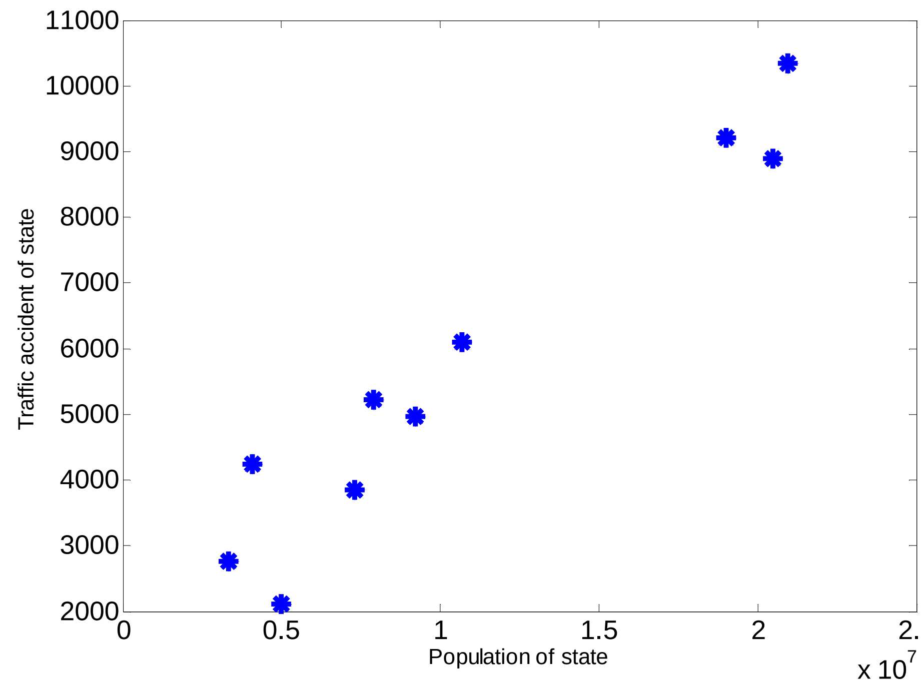
$\mathcal{S}$ is spanned by $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M$.

$\mathbf{w}_{\mathrm{ML}}$ minimizes the distance between **t** and its orthogonal projection on $\mathcal{S}$, i.e. **y**.



Multiple Outputs





$$y_1(x, \mathbf{w}_1) = \mathbf{w}_1^T \boldsymbol{\phi}(x)$$

$$y_2(x, \mathbf{w}_2) = \mathbf{w}_2^T \boldsymbol{\phi}(x)$$

Input: population of state

Output: 1. fatal traffic accident of state
2. accident of state

Multiple Outputs

Population	$\times 10^7$	Fatal traffic accident	$\times 10^3$	Traffic accident	$\times 10^4$
x_1	2.0493	t_{11}	2.2538	t_{12}	0.8891
x_2	1.0676	t_{21}	1.1531	t_{22}	0.6102
x_3	2.0939	t_{31}	2.2342	t_{32}	1.0346
x_4	0.7897	t_{41}	0.9514	t_{42}	0.5223
x_5	0.7299	t_{51}	0.7950	t_{52}	0.3851
x_6	0.4996	t_{61}	0.8286	t_{62}	0.2107
x_7	1.9002	t_{71}	2.1222	t_{72}	0.9206
x_8	0.4083	t_{81}	0.6541	t_{82}	0.4247
x_9	0.9201	t_{91}	1.1208	t_{92}	0.4927
x_{10}	0.3346	t_{101}	0.4322	t_{102}	0.2763

Maximum Likelihood and Least Square

Sum-of-squares error

$$E\left(\mathbf{W}\right) = \frac{1}{2}\sum_{n=1}^{N}\left\|\mathbf{y}\left(x_n, \mathbf{W}\right) - \mathbf{t}_n\right\|^2$$

where

$$\mathbf{y}\left(x_n, \mathbf{W}\right) = \left[\begin{array}{c} \boldsymbol{w}_1^T \boldsymbol{\phi}\left(x_n\right) \\ \boldsymbol{w}_2^T \boldsymbol{\phi}\left(x_n\right) \end{array}\right]$$

$$= \mathbf{W}^T \boldsymbol{\phi}\left(x_n\right)$$

where $N = 10$.

Multiple Outputs

Write in matrix form

$$E(\mathbf{W}) = \frac{1}{2}\|\mathbf{\Phi}\mathbf{W} - \mathbf{T}\|_F^2$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix} \qquad \mathbf{T} = \begin{bmatrix} \boldsymbol{t}_1^T \\ \vdots \\ \boldsymbol{t}_N^T \end{bmatrix}$$

Multiple Outputs

The gradient of $E\left(\mathbf{W}\right)$

$$\nabla E\left(\mathbf{W}\right) = \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{W} - \mathbf{\Phi}^T\mathbf{T}$$

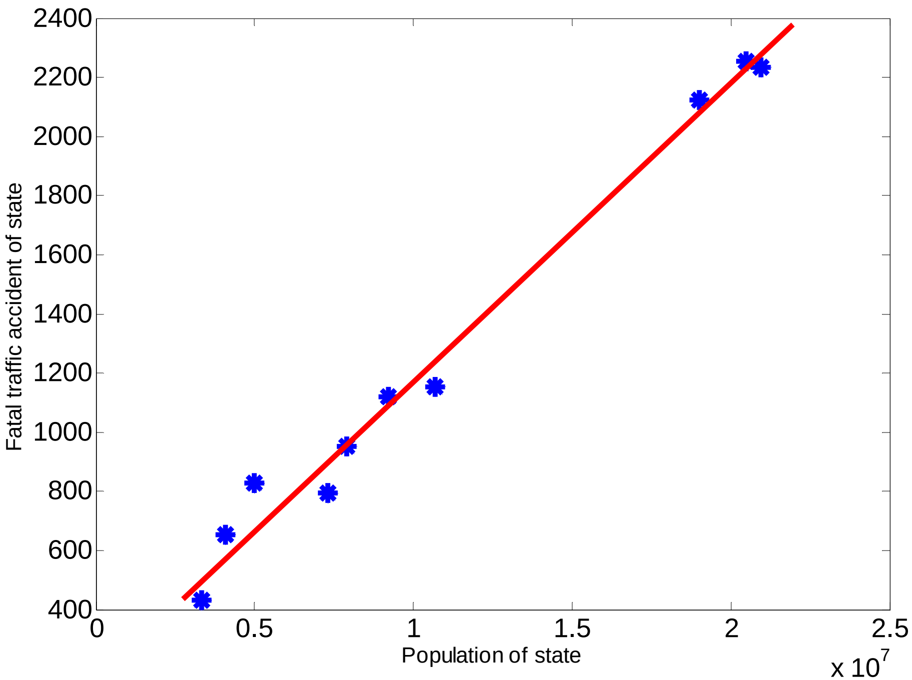
Let

$$\nabla E\left(\mathbf{W}\right) = \mathbf{0}$$

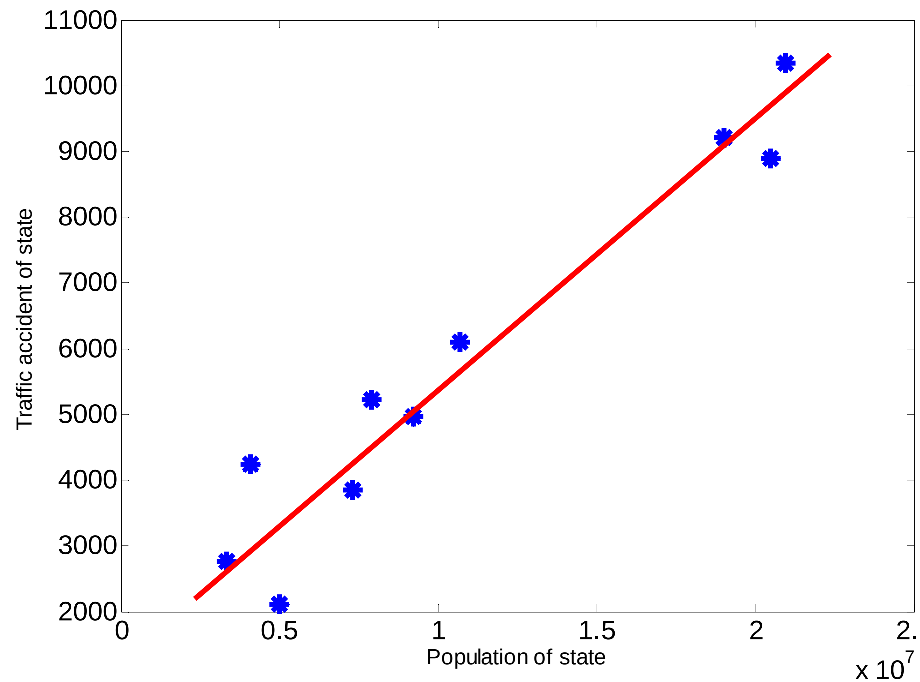
the optimal solution for $\mathbf{W}$ is

$$\mathbf{W}^* = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{T}$$

Multiple Outputs



$$y_1(x, \mathbf{w}_1) = \mathbf{w}_1^T \boldsymbol{\phi}(x)$$



$$y_2(x, \mathbf{w}_2) = \mathbf{w}_2^T \boldsymbol{\phi}(x)$$

$$\boxed{\mathbf{W}^* = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \mathbf{T}}$$

Multiple Outputs

Analogously to the single output case we have:

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) &= \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{W}, \mathbf{x}), \beta^{-1}\mathbf{I}) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I}). \end{aligned}$$

Given observed inputs, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and targets, $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_N]^{\mathrm{T}}$, we obtain the log likelihood function

$$\begin{aligned} \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I}) \\ &= \frac{NK}{2} \ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2} \sum_{n=1}^{N} \left\| \mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) \right\|^2 . \end{aligned}$$

Multiple Outputs

Maximizing with respect to $\mathbf{W}$, we obtain

$$\mathbf{W}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{T}.$$

If we consider a single target variable, t_k, we see that

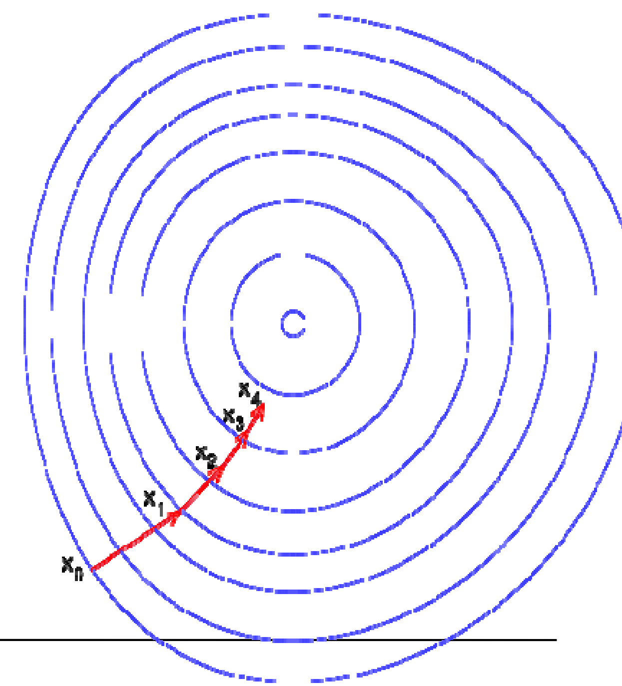
$$\mathbf{w}_k = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}_k = \mathbf{\Phi}^{\dagger}\mathbf{t}_k$$

where $\mathbf{t}_k = [t_{1k}, \ldots, t_{Nk}]^{\mathrm{T}}$, which is identical with the single output case.

Sequential Learning

Gradient descent

$$\nabla J(\boldsymbol{w}) = \begin{bmatrix} \dfrac{\partial J}{\partial w_1} \\ \vdots \\ \dfrac{\partial J}{\partial w_n} \end{bmatrix}$$



Sequential Learning

Data items considered one at a time (a.k.a. online learning); use stochastic (sequential) gradient descent:

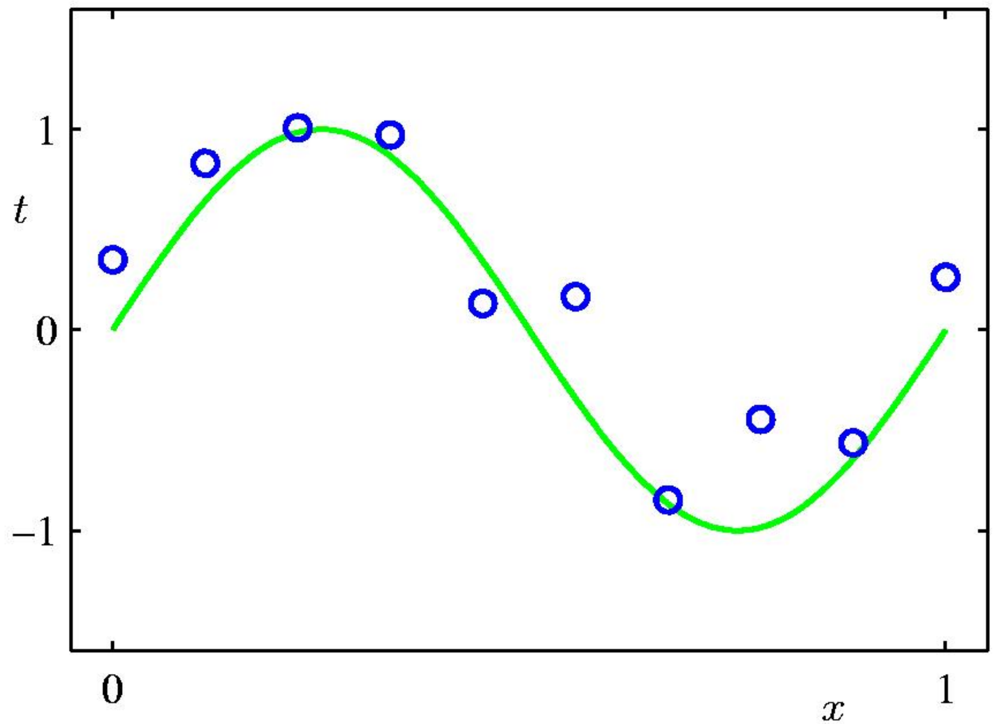
$$
\begin{aligned}
\mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} - \eta \nabla E_n \\
&= \mathbf{w}^{(\tau)} + \eta (t_n - \mathbf{w}^{(\tau)\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)) \boldsymbol{\phi}(\mathbf{x}_n).
\end{aligned}
$$

This is known as the *least-mean-squares (LMS) algorithm*. Issue: how to choose η?

LINEAR MODELS FOR REGRESSION

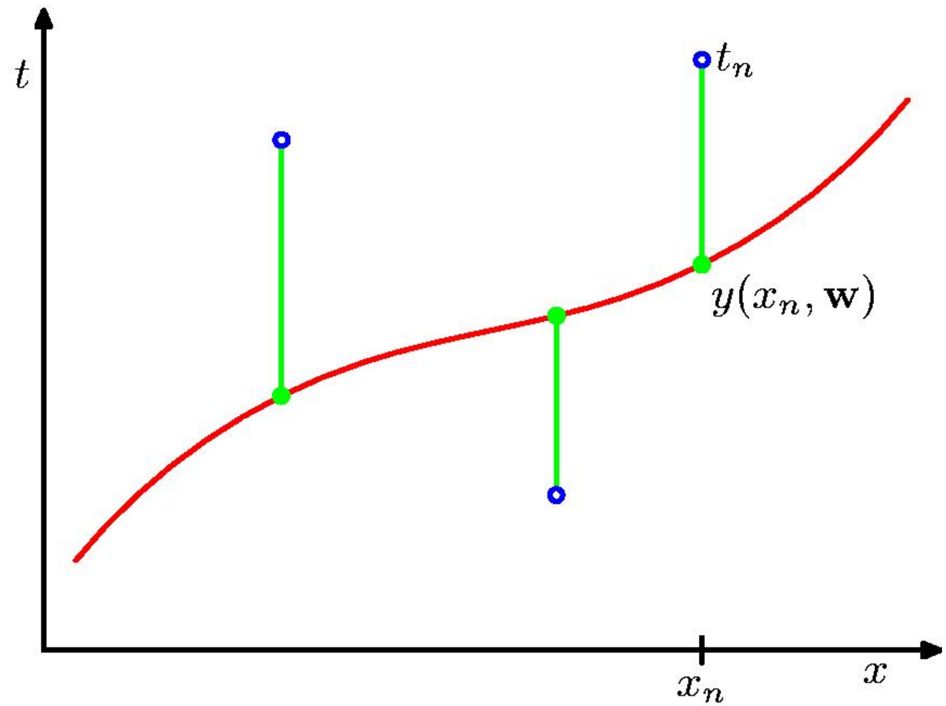
1. The concept of regression
2. Maximum Likelihood and Least Square
3. Over-fitting and Regularization
4. The Bias-Variance Trade-off
5. Bayesian Linear Regression
6. Sparse regression

Polynomial Curve Fitting



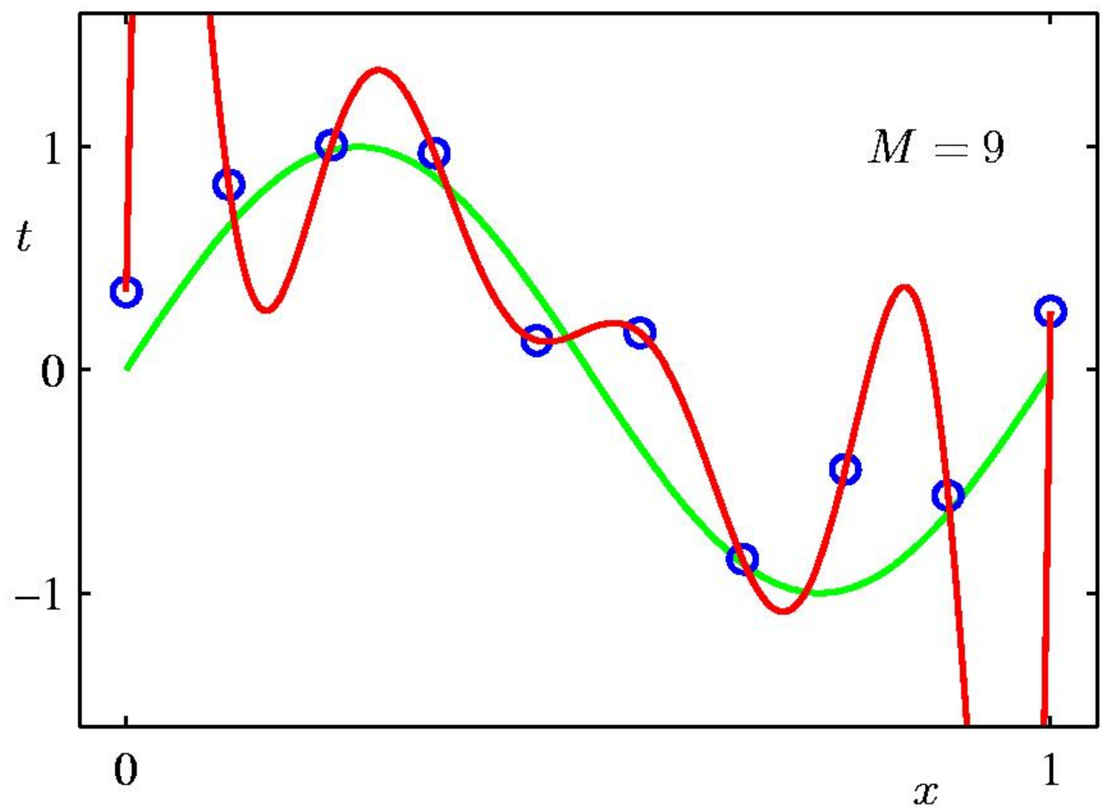
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

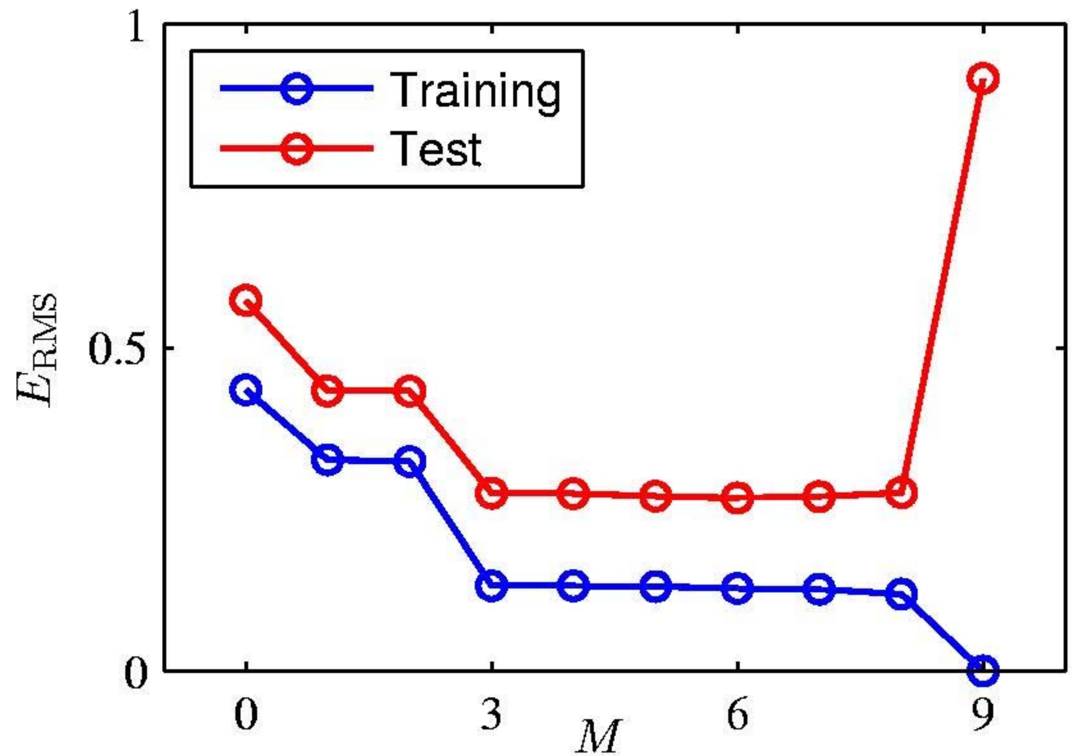


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

9th Order Polynomial



Over-fitting



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Polynomial Coefficients

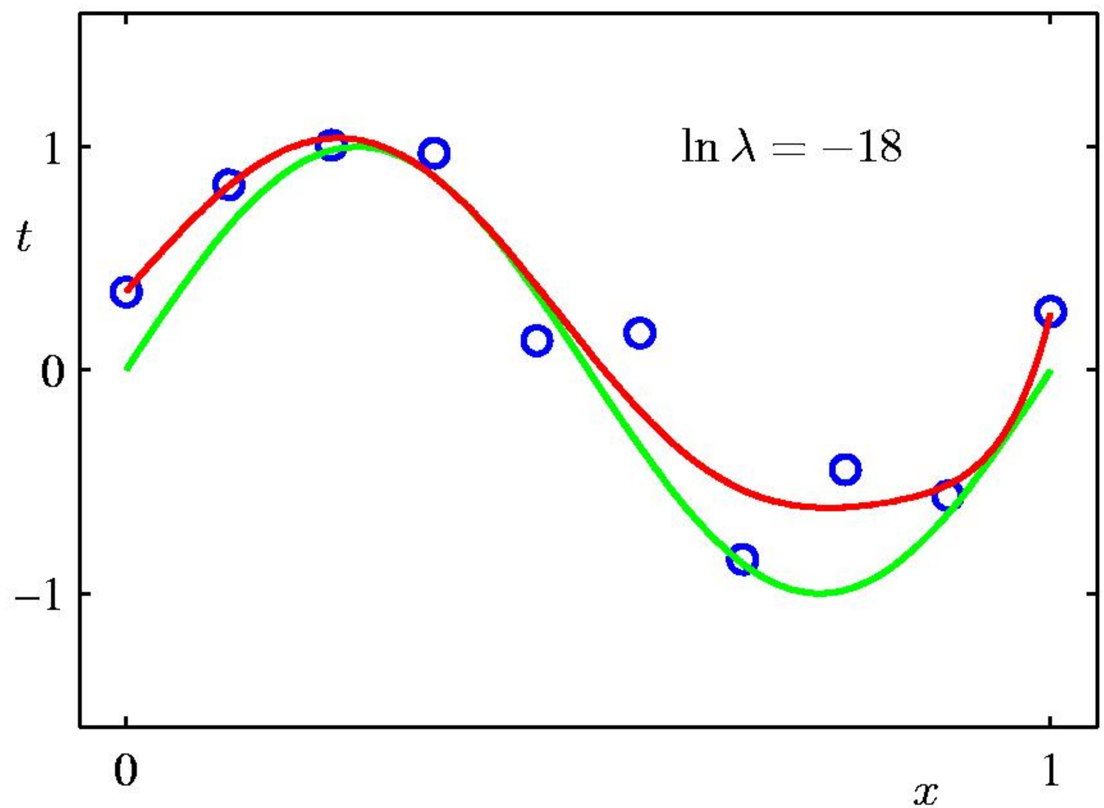
	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Regularization

Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Regularization: $\ln \lambda = -18$



Polynomial Coefficients

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Regularized Least Squares

Consider the error function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term

With the sum-of-squares error function and a quadratic regularizer, we get

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

λ is called the regularization coefficient.

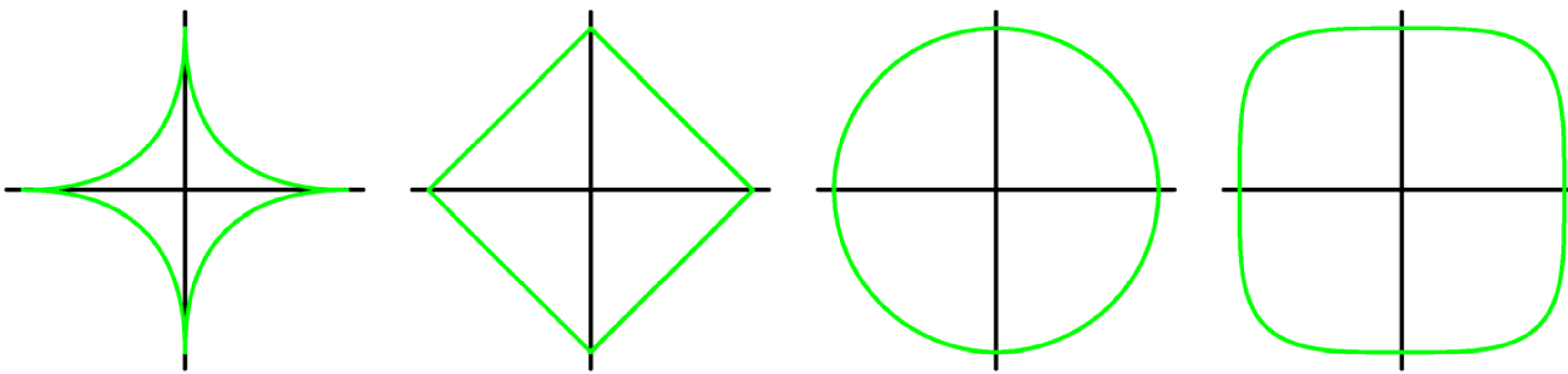
which is minimized by

$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$$

Regularized Least Squares

With a more general regularizer, we have

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$



$q = 0.5$ $q = 1$ $q = 2$ $q = 4$

Lasso Quadratic

Regularized Least Squares

Lasso tends to generate sparser solutions than a quadratic regularizer.