

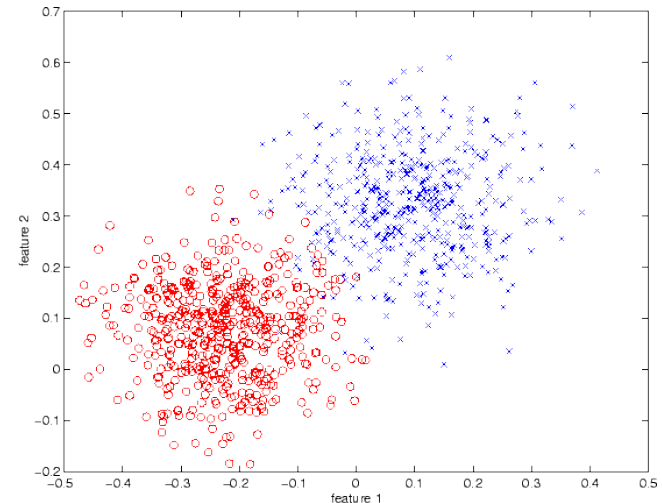
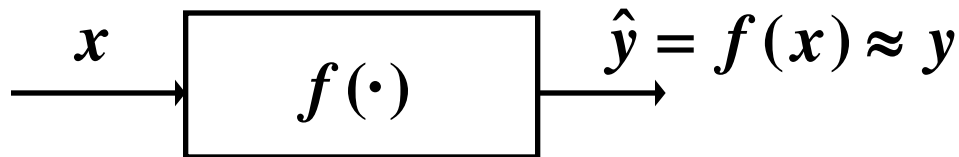
MLE & Regression

Nuno Vasconcelos
(Ken Kreutz-Delgado)

UCSD

Statistical Learning from Data

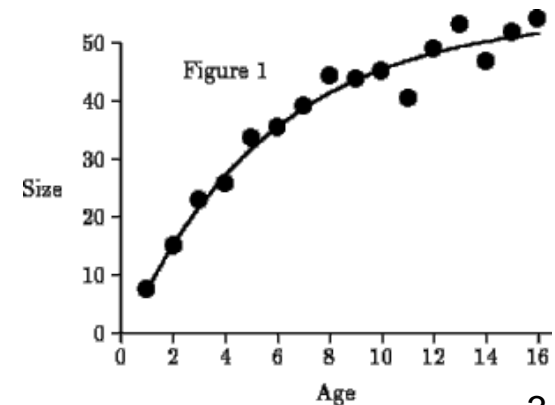
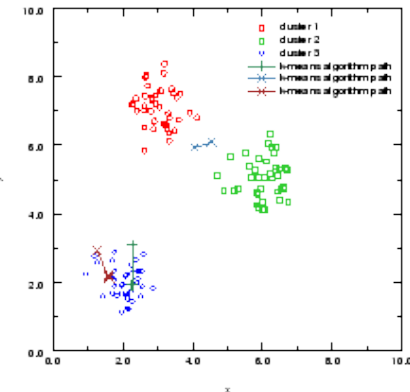
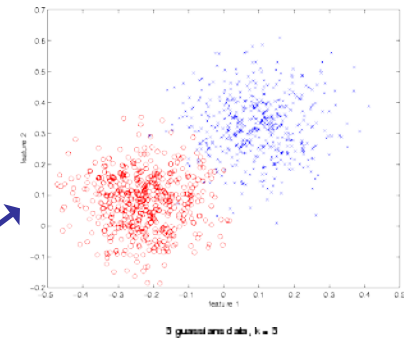
- **Goal:** Given a relationship between a feature vector x and a vector y , and iid data samples (x_i, y_i) , find an approximating function $f(x) \approx y$



- This is called **training or learning**.
- **Two major types** of learning:
 - **Unsupervised (aka Clustering)** : only X is known.
 - **Supervised (Classification or Regression)**: both X and target value Y are known during training, only X is known at test time.

Supervised Learning

- Feature Vector X can be anything, but the type of Y dictates the type of supervised learning problem
 - Y in $\{0,1\}$ is referred to as **detection**
 - Y in $\{0, \dots, M-1\}$ is referred to as **(M-ary) classification**
 - Y continuous is referred to as **regression**
- We have been dealing mostly with **classification**, now we will **emphasize regression**
- The regression problem provides a relatively easy setting to explain non-trivial MLE problems



The Standard Regression Model

- The regression problem is *usually* modeled as follow:
 - There are two random vectors. The independent (regressor) variable X and the dependent (regressed) variable Y .
 - An iid dataset of training examples $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$
 - An additive noise parametric model of the form

$$Y = f(X; \theta) + E$$

where $\theta \in \Theta \subset \mathbb{R}^p$ is a deterministic parameter vector, and E is an iid additive random vector that accounts for noise and model error.

- Two fundamental types of regression problems
 - *Linear regression*, where $f(\cdot)$ is linear in θ
 - *Nonlinear regression*, otherwise
 - What matters is *linearity in the parameter θ* , not in the data X !

Example Regression Models

- Linear Regression:

- Line Fitting

$$f(x; \theta) = \theta_1 x + \theta_0$$

- Polynomial Fitting

$$f(x; \theta) = \sum_{i=0}^k \theta_i x^i$$

- Truncated Fourier Series

$$f(x; \theta) = \sum_{i=0}^k \theta_i \cos(ix)$$

- Nonlinear Regression:

- Neural Networks

$$f(x; \theta) = \frac{1}{1 + e^{-\theta_1 x - \theta_0}}$$

- Sinusoidal Decompositions

$$f(x; \theta) = \sum_{i=0}^k \cos(\theta_i x)$$

- Etc.

- We often assume that E is *additive white Gaussian noise* (AWGN)
- We *always* assume that E and X are *independent*

Probabilistic Model of Y Conditioned on X

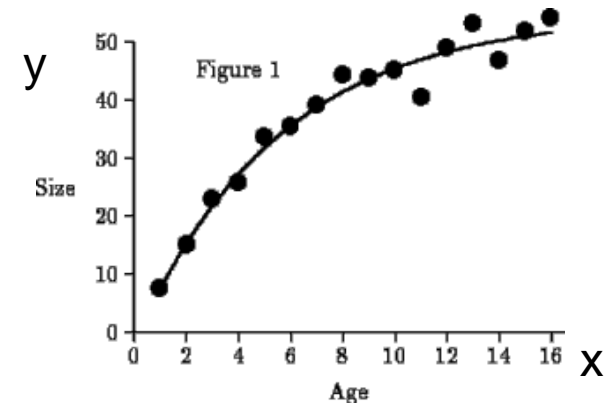
- A realization is $X = x$, $E = \varepsilon$, $Y = y$:

$$y = f(x; \theta) + \varepsilon$$

- x is always known, the goal is to *predict y given x*
- Thus, for each x , $f(x, \theta)$ is treated like a constant
- The realization $E = \varepsilon$ is added to $f(x, \theta)$ to form $Y = y$
- Hence, Y is *conditionally* distributed as E but with a constant added
- This only changes the mean of the distribution of E , $P_E(\varepsilon; \theta)$, yielding

$$P_{Y|X}(y | x; \theta) = P_E(y - f(x; \theta); \theta)$$

- The conditional probability model for $Y|X$ is determined from the distribution of the noise, $P_E(\varepsilon; \theta)$!



The (Conditional) Likelihood Function

- Consider a collection of iid training points $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$. If we define $\mathcal{X} = \{x_1, \dots, x_n\}$ $\mathcal{Y} = \{y_1, \dots, y_n\}$, we have $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$.
- Conditioned on \mathcal{X} , the likelihood of θ given \mathcal{D} is

$$\begin{aligned} P_{\mathcal{D}|\mathcal{X}}(\mathcal{D} | \mathcal{X}; \theta) &= P_{\mathcal{Y}|\mathcal{X}}(\mathcal{Y} | \mathcal{X}; \theta) = \prod_{i=1}^n P_{Y|X}(y_i | \mathcal{X}; \theta) \\ &= \prod_{i=1}^n P_{Y|X}(y_i | x_i; \theta) = \prod_{i=1}^n P_E(y_i - f(x_i; \theta); \theta) \end{aligned}$$

- This is also the \mathcal{X} -conditional likelihood of θ given \mathcal{Y}
- Note: we have used the facts that y_i is conditionally iid and depends only on x_i (both facts being a consequence of our modeling assumptions).

Maximum Likelihood Estimation

- This suggests that
 - Given a collection of iid training points $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, the natural procedure to estimate the parameter θ is ML estimation:

$$\begin{aligned}\hat{\theta}_{\text{ML}} &= \arg \max_{\theta \in \Theta} \prod_i P_{Y|X}(y_i | x_i; \theta) \\ &= \arg \max_{\theta \in \Theta} \prod_i P_E(y_i - f(x_i; \theta); \theta)\end{aligned}$$

Equivalently,

$$\begin{aligned}\hat{\theta}_{\text{ML}} &= \arg \max_{\theta \in \Theta} \sum_i \log P_{Y|X}(y_i | x_i; \theta) \\ &= \arg \max_{\theta \in \Theta} \sum_i \log P_E(y_i - f(x_i; \theta); \theta)\end{aligned}$$

AWGN MLE

- One frequently used model is the *scalar AWGN* case where the noise is zero-mean with variance σ^2

$$P_E(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

- In this case the conditional pdf for $Y|X$ is a Gaussian of mean $f(x; \theta)$ and variance σ^2

$$P_{Y|X}(y | x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y - f(x; \theta))^2}{2\sigma^2} \right\}$$

- If the variance σ^2 is unknown, *it is included in θ*

AWGN MLE

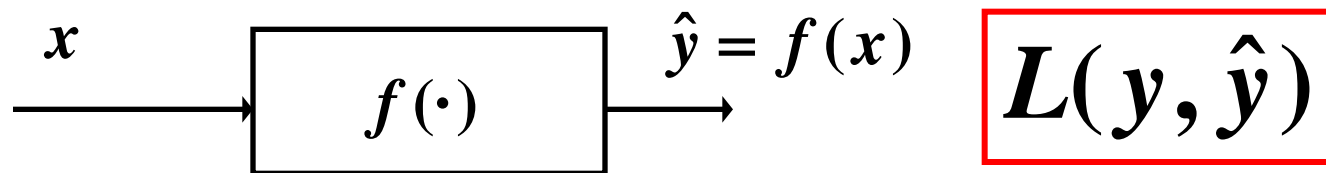
- Assume the variance σ^2 is known. Then the MLE is:

$$\begin{aligned}\hat{\theta}_{\text{ML}} &= \arg \max_{\theta \in \Theta} \sum_i \log P_E(y_i - f(x_i; \theta)) \\ &= \arg \min_{\theta \in \Theta} \sum_i \frac{(y_i - f(x_i; \theta))^2}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2) \\ &= \arg \min_{\theta \in \Theta} \sum_i (y_i - f(x_i; \theta))^2\end{aligned}$$

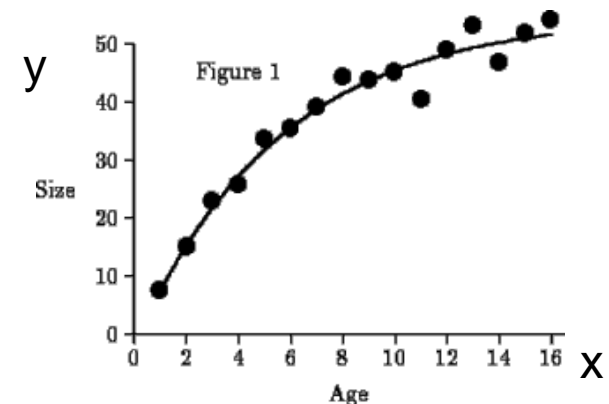
- Since this minimizes the squared Euclidean distance of the estimation error (or prediction error), it is also known as least squares curve fitting

MLE & Optimal Regression

- The above development can be framed in our initial formulation of optimizing the loss of the learning system



- For a regression problem this still applies
 - the interpretation of $f(\cdot)$ as a predictor even becomes more intuitive
- Solving by ML is equivalent to picking a loss identical to the negative of the log of the noise probability density



Loss for Scalar Noise with Known PDF

- Additive Error PDF:

- Gaussian (AWGN case)

$$P_E(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

- Laplacian

$$P_E(\varepsilon) = \frac{1}{2\sigma} e^{-\frac{|\varepsilon|}{\sigma}}$$

- Rayleigh

$$P_E(\varepsilon) = \frac{\varepsilon}{\sigma^2} e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

- Loss, $\varepsilon = (y - f(x; \theta))$:

- L_2 Distance

$$L(f(x; \theta), y) = (y - f(x; \theta))^2$$

- L_1 Distance

$$L(f(x; \theta), y) = |y - f(x; \theta)|$$

- Rayleigh Distance

$$L(f(x; \theta), y) = (y - f(x; \theta))^2 - \log(y - f(x; \theta))$$

Maximum Likelihood Estimation

- How do we find the optimal parameters?
- Recall that to obtain the MLE we need to solve

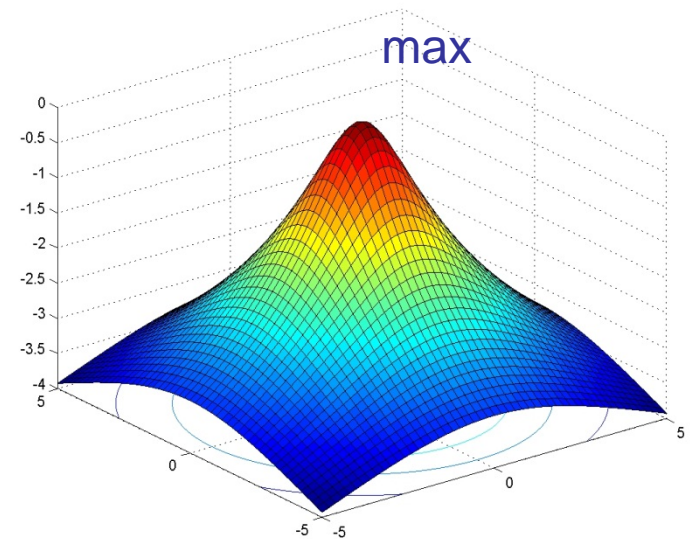
$$\theta^* = \arg \max_{\theta \in \Theta} P_D(D; \theta)$$

- The unique local solutions are the parameter values such that

$$\frac{\partial}{\partial \theta} P_D(D; \hat{\theta}) = 0$$

$$\theta^T H(D, \hat{\theta}) \theta < 0, \quad \forall \theta \neq 0$$

- Note that you always have to check the second-order Hessian condition!



Maximum likelihood Estimation

Recall some important results

- **FACT:** each of the following is a **necessary and sufficient condition** for a real symmetric matrix A to be (strictly) **positive definite**:
 - i) $x^T A x > 0, \forall x \neq 0$
 - ii) All **eigenvalues** of A are real and satisfy $\lambda_i > 0$
 - iii) All **upper-left submatrices** A_k have strictly positive determinant. (strictly positive leading principal minors).
 - iv) There exists a matrix R with independent rows such that $A = R R^T$. Equivalently, there exists a matrix Q with independent columns such that $A = Q^T Q$
- **Definition of upper left submatrices:**

$$A_1 = a_{1,1} \quad A_2 = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \quad A_3 = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \quad \dots$$

Matrix derivatives

- to compute the gradient and Hessian it is useful to rely on matrix derivatives
- some examples that we will use

$$\nabla_{\Theta} (A\Theta) = A^T$$

$$\nabla_{\Theta} (\Theta^T A \Theta) = (A + A^T) \Theta$$

$$\nabla_{\Theta} \|b - A\Theta\|^2 = -2A^T (b - A\Theta)$$

- there are various lists of the most popular formulas
- one example is

<http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html>

Maximum likelihood

- returning to our problem
- to obtain ML estimate we need to solve

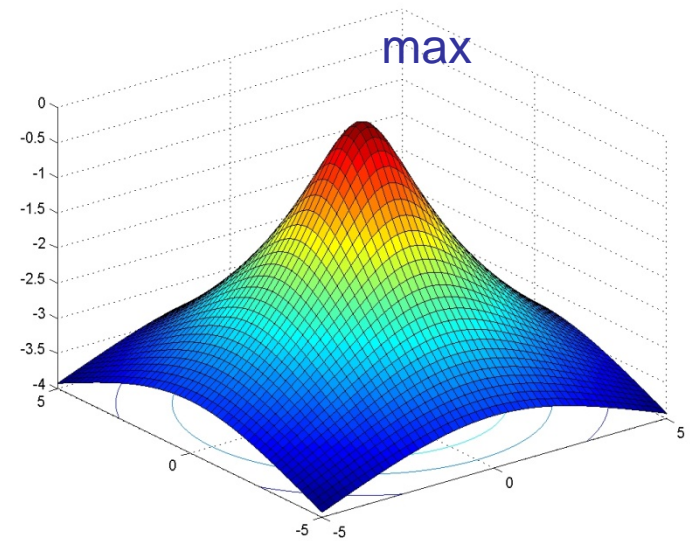
$$\Theta^* = \arg \max_{\Theta} P_X(D; \Theta)$$

- the solutions are the parameters such that

$$\nabla_{\Theta} P_X(x; \Theta) = 0$$

$$\theta^t \nabla_{\Theta}^2 P_X(x; \theta) \theta \leq 0, \quad \forall \theta \in \mathbb{R}^n$$

- note that you always have to check the second-order condition



Maximum likelihood

- for regression this becomes

$$\begin{aligned}\Theta^* &= \arg \max_{\Theta} \sum_i \log P_{Y|X}(y_i | x_i; \Theta) \\ &= \arg \min_{\Theta} \sum_i L(y_i, x_i; \Theta)\end{aligned}$$

- and the solution is given by

$$\nabla_{\Theta} \left[\sum_i L(y_i, x_i; \Theta) \right] = 0$$

$$\theta^T \nabla_{\Theta}^2 \left[\sum_i L(y_i, x_i; \Theta) \right] \theta \geq 0, \quad \forall \theta$$

Maximum likelihood

- noting that the gradient and Hessian are linear operators (derivatives are linear)
- these can be written as

$$\sum_i \nabla_{\Theta} [L(y_i, x_i; \Theta)] = 0$$

- and

$$\theta^T \left[\sum_i \nabla_{\Theta}^2 L(y_i, x_i; \Theta) \right] \theta \geq 0, \quad \forall \theta$$

Example

- Consider the problem of 2-D line fitting

- The model is

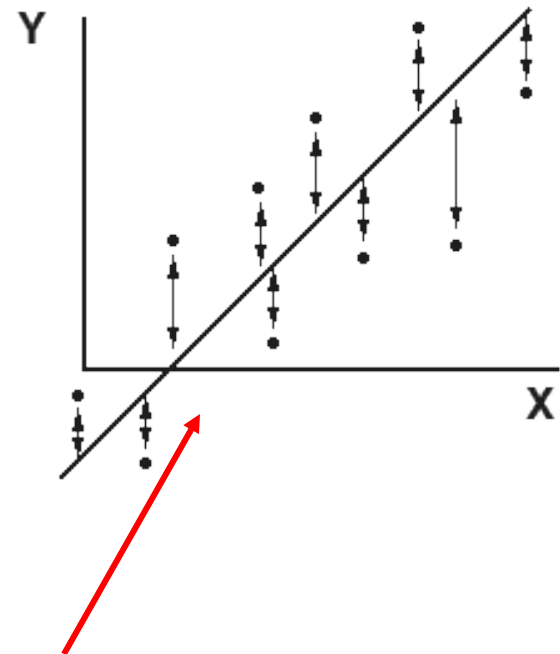
$$y = f(x; \theta) + \varepsilon = \theta_1 x + \theta_0 + \varepsilon$$

where ε is scalar AWGN of known variance

- The (effective) loss function is

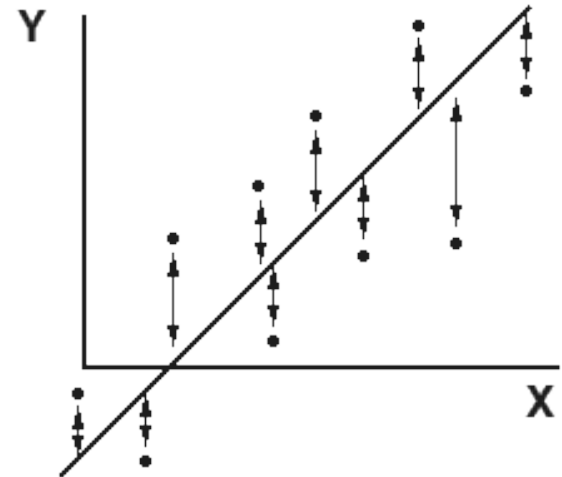
$$L = \sum_i (y_i - \theta_1 x_i - \theta_0)^2$$

- We are looking for the line that makes the square of these vertical distances as small as possible in an averaged sense.
- Our first step is to compute the zeros of the gradient
 - this amounts to solving a system of linear equations



Example

- $$\begin{cases} \frac{\partial L}{\partial \theta_0} = -2 \sum_i (y_i - \theta_1 x_i - \theta_0) = 0 \\ \frac{\partial L}{\partial \theta_1} = -2 \sum_i (y_i - \theta_1 x_i - \theta_0) x_i = 0 \end{cases}$$

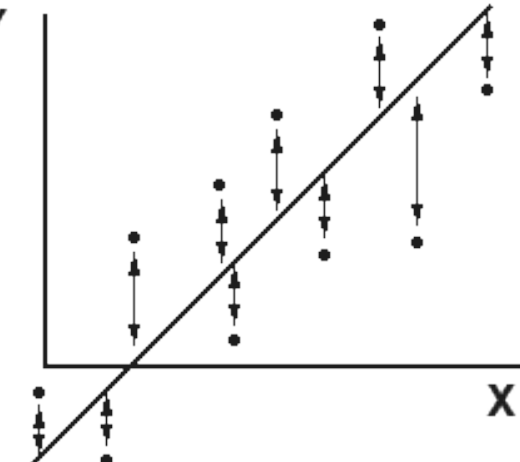


$$\begin{cases} \sum_i y_i = \theta_1 \sum_i x_i + n\theta_0 \\ \sum_i y_i x_i = \theta_1 \sum_i x_i^2 + \theta_0 \sum_i x_i \end{cases}$$

This can be written in matrix form as

Example

- $$\begin{bmatrix} \frac{1}{n} \sum_i y_i \\ \frac{1}{n} \sum_i y_i x_i \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{n} \sum_i x_i \\ \frac{1}{n} \sum_i x_i & \frac{1}{n} \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$



Defining the sample averaged quantities:

$$\bar{y} = \langle y \rangle = \frac{1}{n} \sum_i y_i, \quad \langle x^k \rangle = \frac{1}{n} \sum_i x_i^k, \quad \langle yx \rangle = \frac{1}{n} \sum_i y_i x_i$$

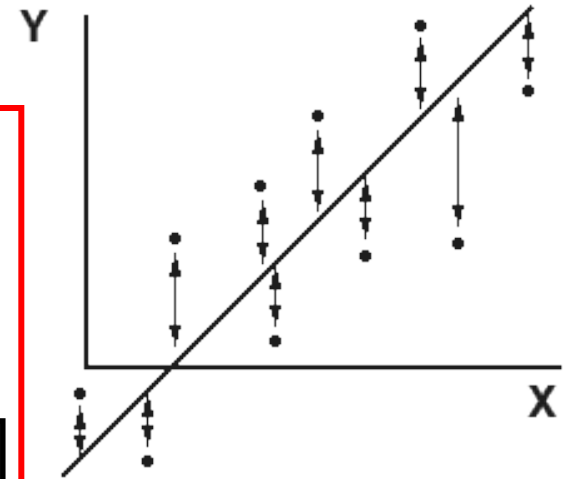
we get

$$\begin{bmatrix} \langle y \rangle \\ \langle xy \rangle \end{bmatrix} = \begin{bmatrix} 1 & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

Example

- The solution is

$$\begin{aligned} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix}_{\text{ML}} &= \begin{bmatrix} \mathbf{1} & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle y \rangle \\ \langle xy \rangle \end{bmatrix} \\ &= \frac{1}{\langle x^2 \rangle - \langle x \rangle^2} \begin{bmatrix} \langle x^2 \rangle & -\langle x \rangle \\ -\langle x \rangle & \mathbf{1} \end{bmatrix} \begin{bmatrix} \langle y \rangle \\ \langle xy \rangle \end{bmatrix} \\ &= \frac{1}{\langle x^2 \rangle - \langle x \rangle^2} \begin{bmatrix} \langle x^2 \rangle \langle y \rangle - \langle x \rangle \langle xy \rangle \\ \langle xy \rangle - \langle x \rangle \langle y \rangle \end{bmatrix} \end{aligned}$$

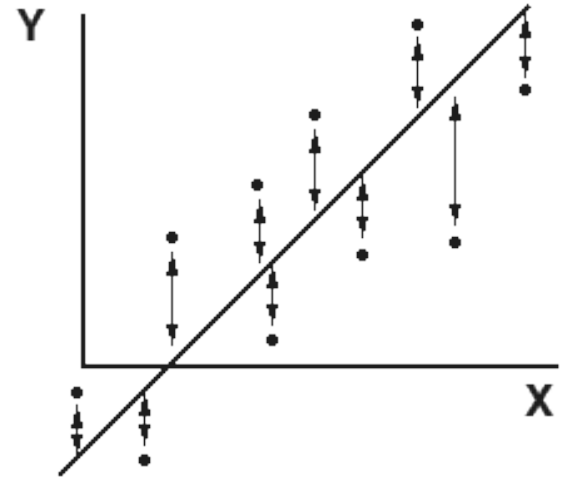


Example

- or, in a form that may be more familiar

$$\hat{\theta}_0 = \frac{1}{n} \sum_i y_i - \hat{\theta}_1 \frac{1}{n} \sum_i x_i$$

$$\hat{\theta}_1 = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_j y_j}{n \sum_i x_i^2 - \left(\sum_i x_i \right)^2}$$



$$\hat{\theta}_1 = \frac{\overline{\text{cov}(x, y)}}{\overline{\text{var}(x)}}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

- we also need to check that we have a minimum

$$\begin{aligned} \frac{\partial L}{\partial \theta_0} &= -2 \sum_i (y_i - \theta_1 x_i - \theta_0) \\ \frac{\partial L}{\partial \theta_1} &= -2 \sum_i (y_i - \theta_1 x_i - \theta_0) x_i \end{aligned} \Rightarrow \frac{\partial^2}{\partial \theta^2} L = 2n \begin{bmatrix} 1 & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{bmatrix}$$

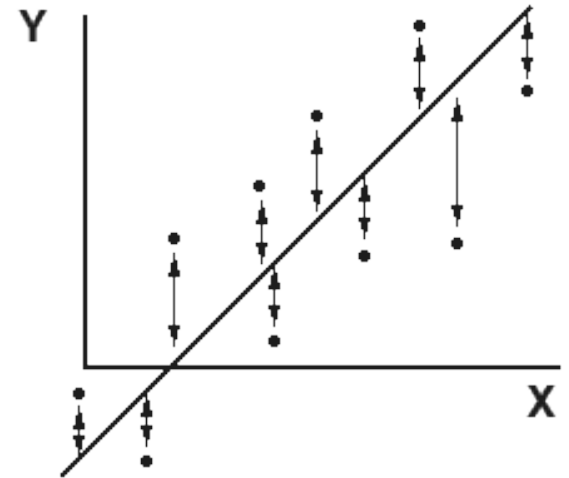
Example

- The Hessian

$$H(\hat{\theta}) = \frac{\partial^2}{\partial \theta^2} L(\hat{\theta}) = 2n \begin{bmatrix} 1 & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{bmatrix}$$

has to be **positive definite**

- Recall that one of the criteria is for the leading **principal minors** to be strictly positive
- Check:
 - $1 > 0$
 - $\langle x^2 \rangle - \langle x \rangle^2 = \overline{\text{var}(x)} = \text{sample variance of } x > 0$



$$\hat{\theta}_1 = \frac{\overline{\text{cov}(x, y)}}{\overline{\text{var}(x)}}, \quad \hat{\theta}_0 = \bar{y} - \theta_1 \bar{x}$$



Least Squares in General

- What if I have other models?
- Can we solve this more generally?
 - Note that we can write the model

$$f(x; \theta) = \theta_1 x + \theta_0$$

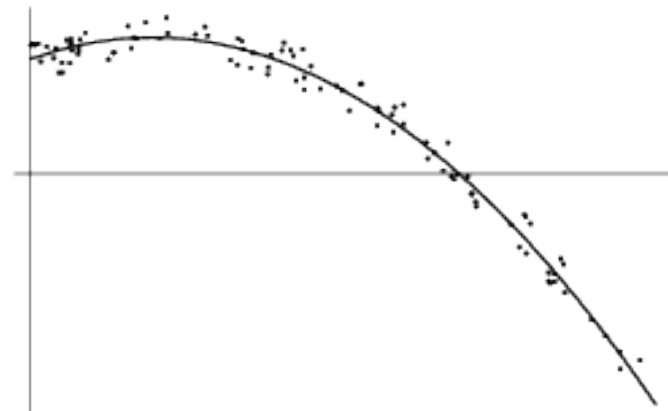
– as

$$f(x; \theta) = \gamma(x)^T \theta$$

$$\gamma(x) = \begin{bmatrix} 1 \\ x \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

- This can be generalized to any model if we exploit the assumption of linearity in the $(k+1)$ -vector θ to form

$$\gamma(x) = \begin{bmatrix} \gamma_0(x) \\ \vdots \\ \gamma_k(x) \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_k \end{bmatrix}$$



Examples

- Elements of $\gamma(x)$ can be arbitrary non-linear functions of x
 - Line Fitting

$$f(x; \theta) = \theta_1 x + \theta_0$$

$$\gamma(x)^T = [1 \quad x]$$

- Polynomial Fitting

$$f(x; \theta) = \sum_{i=0}^k \theta_i x^i$$

$$\gamma(x)^T = [1 \quad \cdots \quad x^k]$$

- Truncated Fourier Series

$$f(x; \theta) = \sum_{i=0}^k \theta_i \cos(ix)$$

$$\gamma(x)^T = [1 \quad \cdots \quad \cos(kx)]$$

Least Squares Parameter Estimation

- For scalar iid AWGN of known variance, we have the (unweighted) least squares loss function

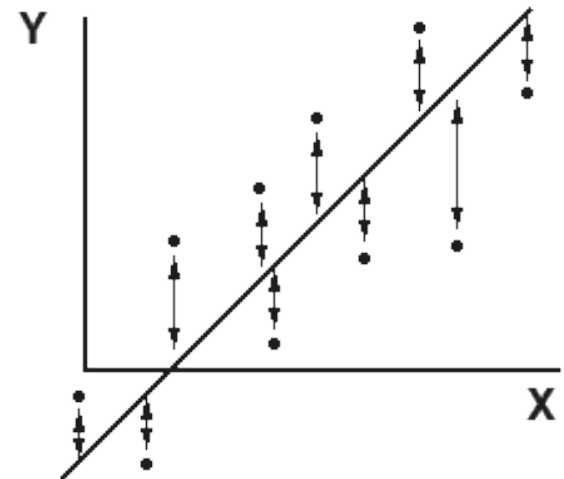
$$L = \sum_i (y_i - \theta_1 x_i - \theta_0)^2$$

which we can write as

$$L = \sum_i (y_i - \gamma(x_i)^T \theta)^2$$

- or

$$L = \|y - \Gamma(x)\theta\|^2$$



- where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \Gamma(x) = \begin{bmatrix} \gamma(x_1)^T \\ \vdots \\ \gamma(x_n)^T \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_k \end{bmatrix}$$

Examples

- The most important component is the matrix $\Gamma(x)$

- Line Fitting

$$\Gamma(x) = \begin{bmatrix} \mathbf{1} & x_1 \\ \vdots & \vdots \\ \mathbf{1} & x_n \end{bmatrix}$$

- Polynomial Fitting

$$\Gamma(x) = \begin{bmatrix} \mathbf{1} & \cdots & x_1^k \\ \vdots & \ddots & \vdots \\ \mathbf{1} & \cdots & x_n^k \end{bmatrix}$$

- Truncated Fourier Series

$$\Gamma(x) = \begin{bmatrix} \mathbf{1} & \cdots & \cos(k x_1) \\ \vdots & \ddots & \vdots \\ \mathbf{1} & \cdots & \cos(k x_n) \end{bmatrix}$$

Least squares

- in summary, we always have

$$L = \|y - \Gamma(x)\Theta\|^2$$

- to minimize this we simply have to find x such that

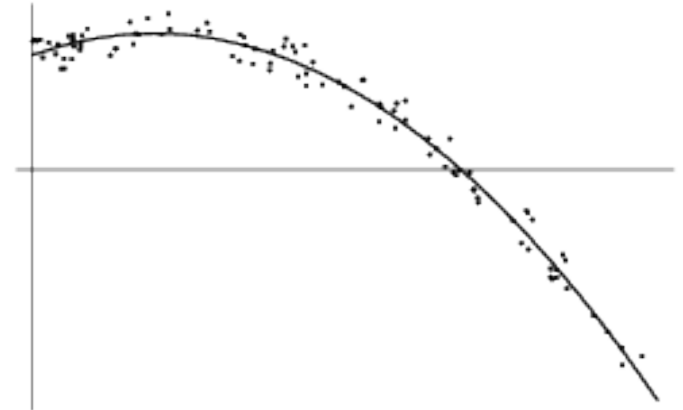
$$\nabla_{\Theta} L = -2\Gamma(x)^T [y - \Gamma(x)\Theta] = 0$$

or

$$\Gamma(x)^T \Gamma(x)\Theta = \Gamma(x)^T y$$

from which, as long as $\Gamma(x)^T \Gamma(x)$ is invertible,

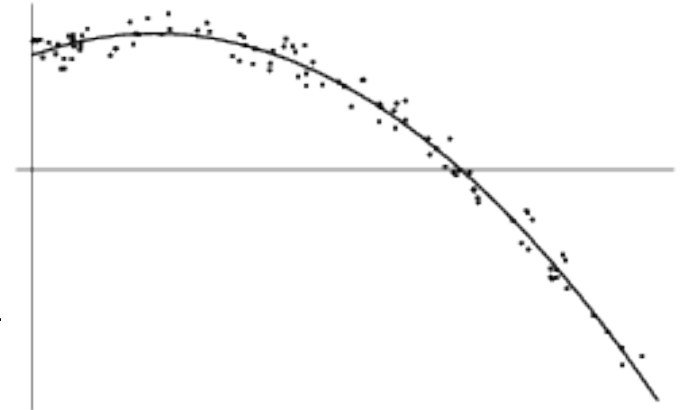
$$\Theta^* = [\Gamma(x)^T \Gamma(x)]^{-1} \Gamma(x)^T y$$



Least squares

- we next check the Hessian

$$\begin{aligned}\nabla_{\Theta}^2 L &= \nabla_{\Theta} (\nabla_{\Theta} L) \\ &= -2 \nabla_{\Theta} \left\{ \Gamma(x)^T [y - \Gamma(x)\Theta] \right\} \\ &= 2 \Gamma(x)^T \Gamma(x)\end{aligned}$$



- this is positive definite if the rows of $\Gamma(x)$ are independent
- which turns out to be
 - the condition for $\Gamma(x)^T \Gamma(x)$ to be invertible,
 - which is the necessary condition for the solution to be feasible
- note that we design $\Gamma(x)$, so we can always make this happen
- usually we only have to make sure all the x_i are different

(Unweighted) Least Squares

- In summary

- A problem of the type

$$\min_{\theta} \mathbf{L}(\theta) = \|y - \Gamma(x)\theta\|^2$$

has a least squares solution

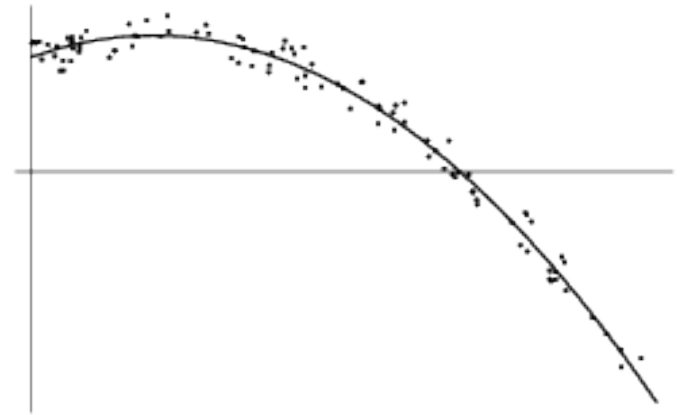
$$\hat{\theta}_{\text{LS}} = [\Gamma(x)^T \Gamma(x)]^{-1} \Gamma(x)^T y = \Gamma(x)^+ y$$

iff $\Gamma(x)$ has full column rank.

- The matrix

$$\Gamma(x)^+ = [\Gamma(x)^T \Gamma(x)]^{-1} \Gamma(x)^T$$

is called the (Moore-Penrose) pseudo-inverse of $\Gamma(x)$



(Unweighted) Least Squares

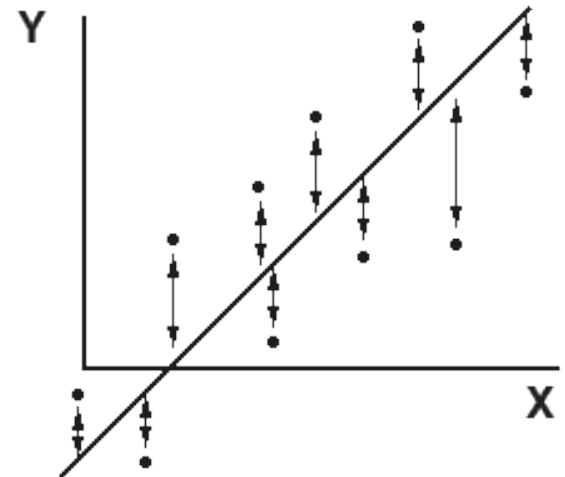
- Here is a way of thinking about this
 - we have an inconsistent system of equations

$$y \approx \Gamma(x)\theta$$

This can't be solved because although $\Gamma(x)$ has full (column) rank, it is “tall” (has more rows than columns) and thus is not invertible

E.g. consider the line

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$



- To make a consistent system, we multiply both sides by $\Gamma(x)^T$

$$\Gamma(x)^T y = \Gamma(x)^T \Gamma(x) \theta$$

(Unweighted) Least Squares

- This is now a solvable system. E.g.,

$$\begin{bmatrix} \mathbf{1} & \dots & \mathbf{1} \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \dots & \mathbf{1} \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} \mathbf{1} & x_1 \\ \vdots & \vdots \\ \mathbf{1} & x_n \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

whose solution is given by the pseudo-inverse

$$\hat{\theta}_{\text{LS}} = \Gamma^+(x)y = [\Gamma(x)^T \Gamma(x)]^{-1} \Gamma(x)^T y$$

We have just seen that this is the best approximate solution to the original problem in the (unweighted) least squares sense

$$\hat{\theta}_{\text{LS}} = \arg \min_{\theta \in \Theta} \|y - \Gamma(x)\theta\|^2$$

(Unweighted) Least squares

- In principle, assuming that the matrix $\Gamma(x)$ has full column rank, the least squares solution is straightforward to compute
- For example, let's redo the [line example](#)

$$\Gamma(x)^T \Gamma(x) = \begin{bmatrix} \mathbf{1} & \dots & \mathbf{1} \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} \mathbf{1} & x_1 \\ \vdots & \vdots \\ \mathbf{1} & x_n \end{bmatrix} = n \begin{bmatrix} \mathbf{1} & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{bmatrix}$$

$$\Gamma(x)^T y = \begin{bmatrix} \mathbf{1} & \dots & \mathbf{1} \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = n \begin{bmatrix} \langle y \rangle \\ \langle xy \rangle \end{bmatrix}$$

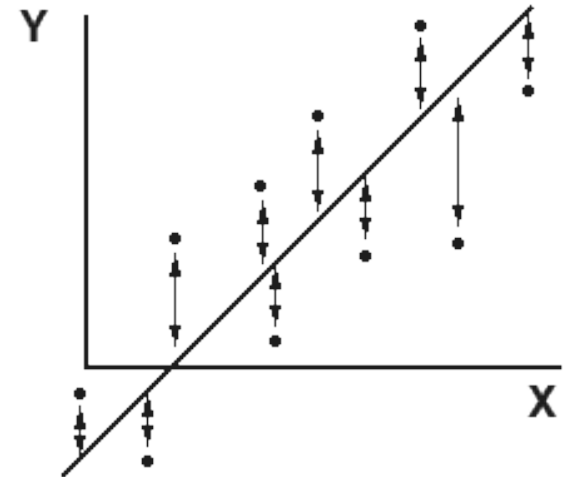
(Unweighted) Least squares

- So that

$$\hat{\theta}_{\text{LS}} = \Gamma^+(x) = [\Gamma(x)^T \Gamma(x)]^{-1} \Gamma(x)^T y$$

- leads to

$$\begin{aligned} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix}_{\text{LS}} &= \begin{bmatrix} 1 & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle y \rangle \\ \langle xy \rangle \end{bmatrix} \\ &= \frac{1}{\langle x^2 \rangle - \langle x \rangle^2} \begin{bmatrix} \langle x^2 \rangle \langle y \rangle - \langle x \rangle \langle xy \rangle \\ \langle xy \rangle - \langle x \rangle \langle y \rangle \end{bmatrix} \end{aligned}$$



which is the solution that we had obtained before, but now with less work. Of course, we know from ECE174 that there is a deep geometric formalism at play here.

Relationship to Probabilistic Model

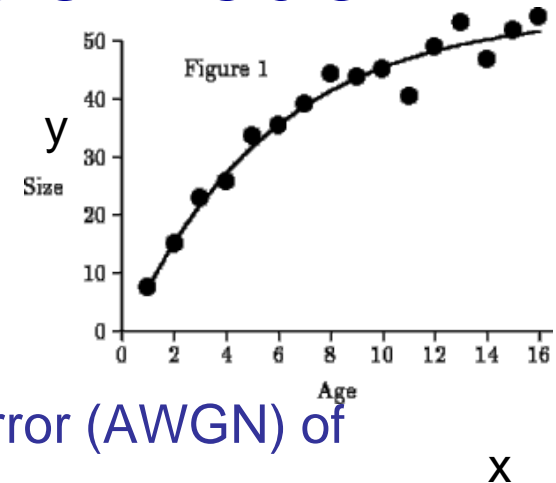
- The (unweighted) least square solution
 - Estimates the function $f(x; \theta)$ of maximum likelihood for the scalar model

$$y = f(x; \theta) + \varepsilon$$

- where ε is a scalar iid zero-mean Gaussian error (AWGN) of known variance,

$$P_E(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

- The method is general
- Other models $f(x, \theta)$ will lead to other least squares problems
- If variance is unknown, we don't have a pure LS problem
- If we have a vector model, in general we have weighted LS
- If the error is not Gaussian, problem is not least squares.



END