Machine Learning 1

- Lecture 3 -

Supervised Learning: Linear Regression

-Patrick Forré-

Slides created by:
Rianne van den Berg

Image credit: Kirillm | Getty Images

Overview

1. **Recap: Probability theory**

2. **Recap: Statistical learning principles**

3. **Linear Models for Regression**

4. **Model selection/supervised learning**

Overview

1. **Recap: Probability theory**
2. Recap: Statistical learning principles
3. Linear Models for Regression
4. Model selection/supervised learning

The rules of probability theory

For random variables *X* ∈ X and *Y* ∈ Y:

	Discrete	Continous
Additivity	$p(X \in A) = \sum_{x \in A} p(x)$	$p(X \in A) \simeq \int_A p(x) dx$
Positivity	$p(x) \geq 0$	$p(x) \geq 0$
Normalization	$\sum_x p(x) \simeq 1$	$\int_{\mathcal{X}} p(x)dx = 1$
Sum Rule	$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$	$p(x) = \int p(x,y) dy$
Product Rule	$p(x, y) = p(x\|y)p(y)$	$p(x, y) = p(x\|y)p(y)$

↳ generalizations → measure theory

Bayes Rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- *p(y)* : the prior probability of *Y* = *y*
- *p(y | x)* : the posterior probability of *Y* = *y*
- *p(x | y)* : the likelihood of *X* = *x* given *Y* = *y*
- *p(x)* : the evidence for X = x

Multivariate Gaussian Distribution

- *D*-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_D)^T$
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

  $|\boldsymbol{\Sigma}| = \det\boldsymbol{\Sigma}$

  (D x D matrix)

  quadratic in x and μ
- $\boldsymbol{\Sigma} = \mathrm{cov}(x, x)$
- $\mathbb{E}[\mathbf{x}] = \mu$

$$\int \exp\{-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\}\mathrm{d}^D x = \frac{(2\pi)^{D/2}}{|\mathbf{A}|^{1/2}}$$

Independent Random Variables

Two random variables X and Y are *independent* iff measuring X gives no information on Y, and vice versa.

- Formally: X and Y are called independent if

$p(x,y) = p(x)\,p(y)$ for all $x \in \mathcal{X},\ y \in \mathcal{Y}$

- Equivalent to

$$p(x|y) = p(x) \qquad \forall x\ \forall y\ p(y) > 0$$

Overview

1. Recap: Probability theory

2. **Recap: Statistical learning principles**

3. Linear Models for Regression

4. Model selection/supervised learning

Statistical Learning Principles

- Given: data set $D = (x_1, x_2, \ldots, x_N)$ of N observations
- Choice: (parametric statistical) model class:
  $p(x_1, x_2, \ldots \mid \mathbf{w})$ parameterised by $\mathbf{w} \in \mathscr{W}$ as "proposal distributions"

  e.g. $\mathcal{N}(x_1 .. x_N \mid \underbrace{\mu, \Sigma}_{w})$
- Questions: How do we choose a model to make predictions? Which is the best model to explain the data? etc..
- 3 general statistical learning principles to go from data to models:
  1. Maximum Likelihood Estimation (MLE)
  2. Maximum a Posteriori (MAP)
  3. Bayesian Model Averaging

Maximum Likelihood Estimation (MLE)

- Given: data set $D = (x_1, x_2, \ldots, x_N)$ of N observations.
- Likelihood of the dataset: $p(D|\mathbf{w})$ (as a fct of $\mathbf{w}$)
- Maximum likelihood principle: the most likely "explanation" of D is given by $\boldsymbol{w}_{\text{ML}}$ which maximizes the (log-)likelihood function: $\mathbf{w}_{\text{ML}} := \underset{\mathbf{w}}{\text{argmax}}\, p(D|\mathbf{w})$ (log $p(D|\mathbf{w})$)
- Under i.i.d. assumptions: each $x_i \in D$ is independently distributed according to the same underlying true distribution: $x_i \sim p_{\text{true}}(x)$ unknown
- so may assume the *independence conditioned on* $\boldsymbol{w}$: $x_i \sim p(x|\mathbf{w})$ iid.
- For i.i.d. samples: $p(D|\mathbf{w}) = p(x_1, x_2, \ldots, x_N|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|\mathbf{w})$ iid
- Predictions via: new data point x predict: $p(x|\mathbf{w}_{ML})$

(Bayesian point estimator)

Maximum A Posteriori Estimates (MAP)

- Given: Data set $D = (x_1, x_2, \ldots, x_N)$ of N observations.
- Likelihood of data: $p(D|w)$
- Prior belief in **w**: $p(w)$
- MAP estimate: choose most probable ***w*** given the data, i.e. a posterior mode:

$$\mathbf{w}_{\mathrm{MAP}} = \operatorname*{argmax}_{w} p(w|D)$$

$$= \operatorname*{argmax}_{w} p(D|w) \cdot p(w)$$

$$= \operatorname*{argmin}_{w} -\log p(D|w) - \log p(w)$$

- For i.i.d. data:

$$\operatorname*{argmin}_{w} \left[\sum_{n=1}^{N} -\log p(x_n|w) - \log p(w) \right]$$

- Predictions via:

$$p(x|w_{MAP})$$

$$p(w|D) = \frac{p(D|w) \cdot p(w)}{p(D)}$$

$$p(D) = \int p(D|w) p(w) dw$$

Bayesian Model Averaging (Full Bayes)

- Given: Data set $D = (x_1, x_2, \ldots, x_N)$ of N observations.
- Likelihood of data: $p(D|w)$
- Prior belief in **w**: $p(w)$ should represent some prior knowledge/belief of the plausibility of ***w***.

- After observing data $D = (x_1, x_2, \ldots, x_N)$, posterior distribution:

$$p(w|D) = \frac{p(D|w)}{p(D)} \cdot p(w)$$

- Predictive distribution via posterior model averaging:

$$p(x|D) := \int p(x|w') \, p(w'|D) \, dw'$$

Overview

1. Recap: Probability theory

2. Recap: Statistical learning principles

3. **Linear Models for Regression**

4. Model selection/supervised learning

& curve fitting
supervised learning
for continuous RV.

Linear Regression

- Regression: $D = \{(\mathbf{x}_1, t_1), ..., (\mathbf{x}_N, t_N)\}$
  - Input variables
  - Target variables

$x_n \in \mathbb{R}^D$

$= (x_{n1})_{d=1...D}^T$

$x_i \in \mathbb{R}^D$

$t_i \in \mathbb{R}$

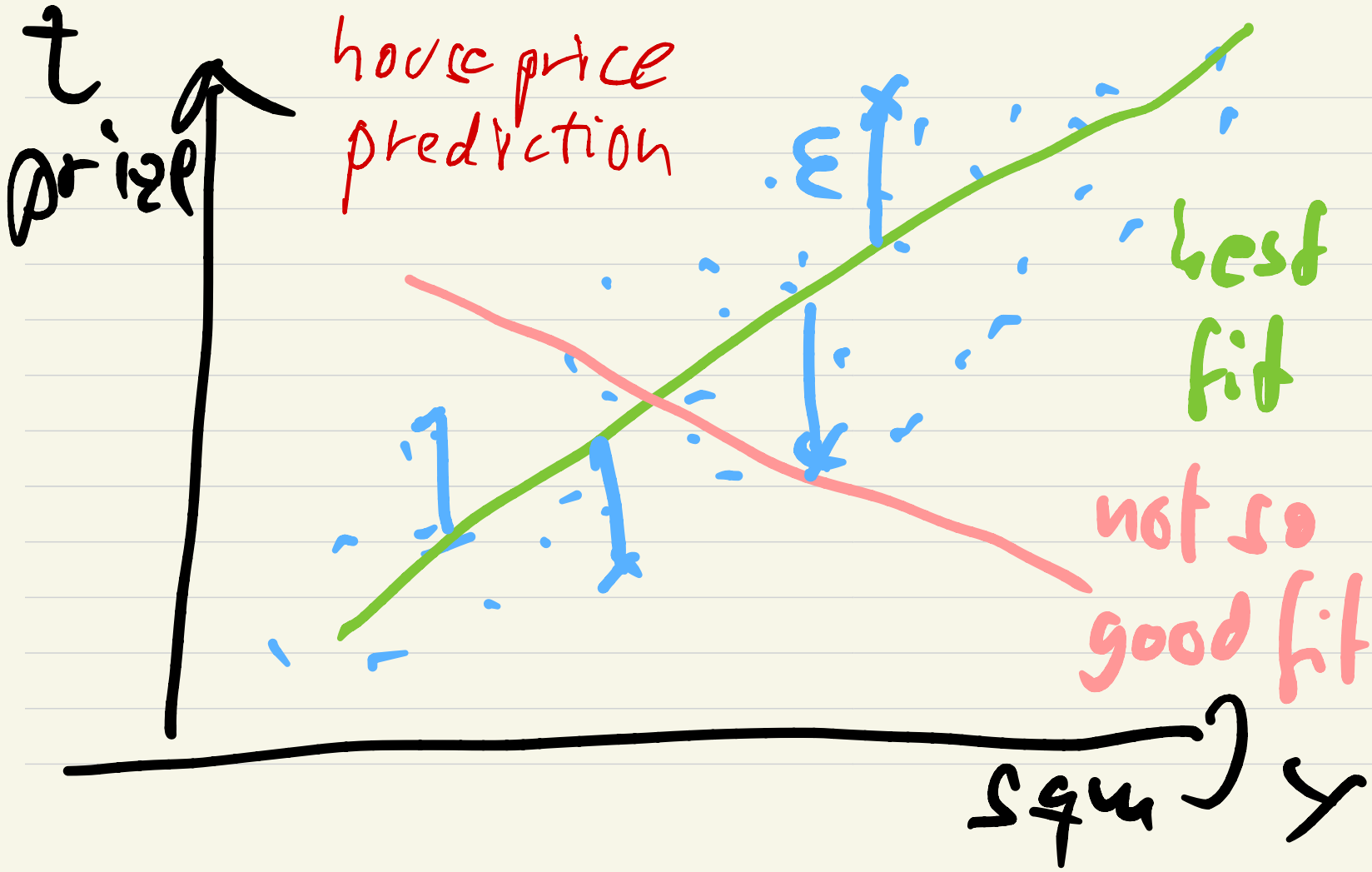
- Simplest linear model:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_{\cdot 1} + w_2 x_{\cdot 2} + w_3 x_{\cdot 3} + \dots + w_{m-1} x_{\cdot m-1}$$

$$= w^T x$$ components

$$w = \begin{pmatrix} w_0 \\ \vdots \\ w_{m-1} \end{pmatrix}$$

$$X = \begin{pmatrix} x_{11} \\ x_{10} \end{pmatrix}$$


house price prediction
price
sqm
ε
best fit
not so good fit

Linear Basis Models

$x \in \mathbb{R}^D$

$\phi(x) \in \mathbb{R}^M$

- Fix number of parameters *M* s.t.

$w \in \mathbb{R}^M$

- Choose *M* - *1* basis functions/features of $\boldsymbol{x}$:

$\phi_1(x), \ldots, \phi_{M-1}(x)$

- Approximation:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 \phi_1(x) + w_2 \phi_2(x) + \ldots + w_{M-1} \phi_{M-1}(x)$$

$w = \begin{pmatrix} w_0 \\ \vdots \\ w_{M-1} \end{pmatrix}$

- w_0 : bias/offset

↳ not linear in x

But linear in w !!

- Set $\phi_0(\mathbf{x}) = 1$ such that $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})$

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

↳ linear in w

Example: Basis Functions (I)

- Projection on input components : $\phi_i(\mathbf{x}) = x_i$

$x = (x_1, \ldots, x_D)$

for $\mathbf{x} = (x_1, x_2, ..., x_D)^T$: $y(\mathbf{x}, \mathbf{w}) =$

$w_0 + w_1 x_1 + w_2 x_2 + \ldots$

↳ multidim linear regression

- i-power map for $x \in \mathbb{R}$: $\phi_i(x) = x^i$ (monomials)

$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \ldots$

↳ one-dim polynomial regression

Example: Basis Functions (III)

- Gaussian basis functions: $\phi_i(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)$

$\mathbf{x} \in \mathbb{R}^D$

$y(\mathbf{x}, \mathbf{w}) =$ $w_0 + w_1 e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)} + w_2 e^{-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)} + \dots$

- Logistic sigmoid functions: $\phi_i(x) = \sigma\left(\frac{x-\mu_i}{s_i}\right)$ $\quad x \in \mathbb{R}$

with $\sigma(x) := \frac{1}{1+\exp(-x)}$

$y(x, \mathbf{w}) = w_0 + w_1 \sigma\left(\frac{x-\mu_1}{s_1}\right) + w_2 \sigma\left(\frac{x-\mu_2}{s_2}\right) + \dots$

μ_i, s_i fixed.

Example: Basis Functions (IV)

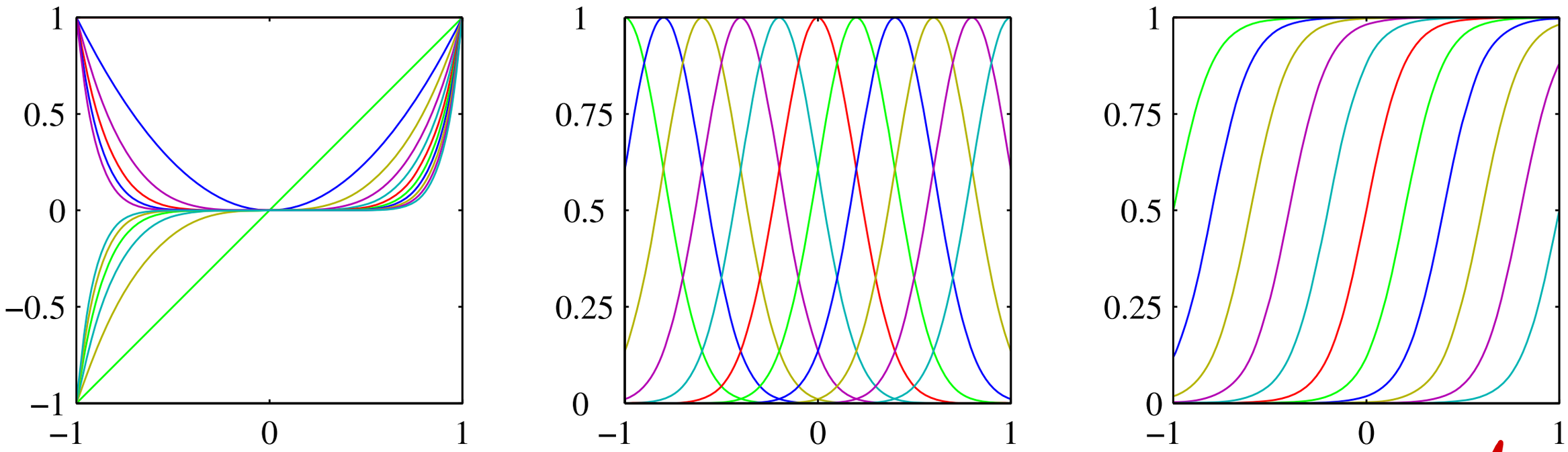


Figure: Example of basis functions (Bishop 3.1)

Maximum Likelihood

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

Hyperparameter
β precision
β^{-1} = variance

- Assume gaussian noise around the target

$$t = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + \varepsilon$$

determ. random

$\varepsilon \sim N(0, \beta^{-1})$

$\varepsilon \perp\!\!\!\perp x$ (indep)

- $p(t|\mathbf{x}, \mathbf{w}, \beta) = N(t | \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \beta^{-1})$

- Dataset: $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and $\mathbf{t} = (t_1, ..., t_N)^T$

- Likelihood function (iid)

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

ML: Sum-of-Squares Error

- Likelihood: $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i), \beta^{-1})$

$\sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(t_i - w^T\phi(x_i))^2\right)$

- Log likelihood $\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) =$

$N \log\sqrt{\frac{\beta}{2\pi}} - \frac{\beta}{2}\sum_{n=1}^{N}(t_i - w^T\phi(x_i))^2$

- Sum-of-squares error: $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}|t_i - w^T\phi(x_i)|^2$

- For comparison of different dataset sizes *N*

$$E_D^{\mathrm{RMSE}}(\mathbf{w}) = \sqrt{\frac{1}{N}\sum(t_i - w^T\phi(x_i))^2}$$

Example: Sum-of-Squares Error

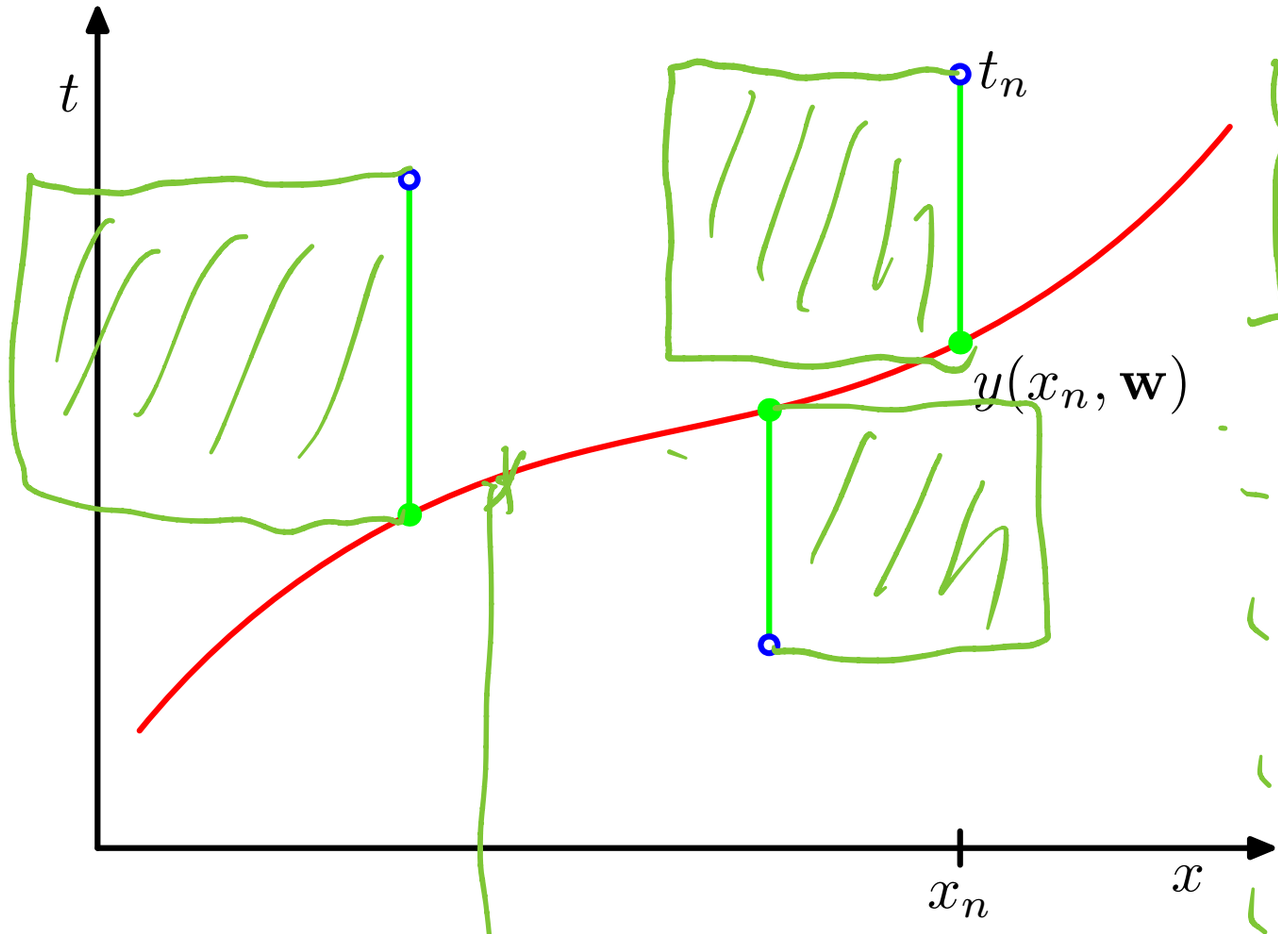


Figure: Errors are given by half the squares of green bars (Bishop 1.3)

$f : U \subseteq \mathbb{R}^d \longrightarrow \mathbb{R}$ differentiable (2-time cont.)

Thm

1) If x is local minimum of f

$$\Rightarrow \nabla f(x) = 0$$

2) If $x \in U$, $\nabla f(x) = 0$

and $\nabla^2 f(x)$ pos. def $\Rightarrow x$ local minimum

3) If f is convex

(e.g. $D^2 f$ pos. def. everywhere)

$\Rightarrow$ $Df(x) = 0$ implies

x global minimum

$\Rightarrow$ $x^* \in \arg\min_x f$

$\Rightarrow$ x^* is global min.

$\Rightarrow$ x^* is local min of f

$\Rightarrow$ $\nabla f(x^*) = 0$

so check these points to find global min.

Maximum Likelihood Estimates

- Maximize the log likelihood / Minimize the sum-of-squares error:

$$\arg\max_{\mathbf{w}} \log p(t|X,w,\beta) = \arg\min_{\mathbf{w}} E(w)$$

$$\frac{\partial}{\partial \mathbf{w}} \log p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) = -\beta \frac{\partial}{\partial \mathbf{w}} E_D(\mathbf{x}) = -\beta \frac{\partial}{\partial \mathbf{w}} \frac{1}{2} \sum_{i=1}^{N} \{t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i)\}^2$$

$$= -\frac{\beta}{2} \sum_{n=1}^{N} \frac{\partial}{\partial w} (t_i - w^T \phi(x_i))^2$$

$$= +\frac{\beta}{2} \sum 2 \cdot (t_i - w^T \phi(x_i)) \cdot \phi(x_i)^T \doteq 0$$

$$\Rightarrow \sum_{i=1}^{N} \phi(x_i) \cdot t_i = \sum_{i=1}^{N} \phi(x_i) \phi(x_i)^T w$$

$$\frac{\partial a}{\partial \mathbf{x}} = \left(\frac{\partial a}{\partial x_1}, \frac{\partial a}{\partial x_2}, \ldots \right)$$ ← ROW vector

Maximum Likelihood Estimates

- Optimal $\boldsymbol{w}^\star$ satisfies

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

$$\sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T\mathbf{w} = \sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i)t_i$$

$N \times M$

$\Phi^T \Phi w = \Phi^T t$

$t = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$

if $\Phi^T\Phi$ invertible $\Rightarrow$ $w_{ML} \triangleq w^* = (\Phi^T\Phi)^{-1}\Phi^T t$

else $w^* = \Phi^+ t$ $\quad \Phi^+ \leftarrow$ Moore-Penrose pseudoinverse

$$\mathbb{E}[t'|\mathbf{x}', \mathbf{w}_{\mathrm{ML}}] = y(x', w_{ML}) = w_{ML}^T \phi(x')$$