Tuesday, 9/1/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 3

Announcements

- Two surveys are open
  - 1 is optional (preferences)
  - 1 is required (availability)
    - Counts towards your participation grade
- First homework will be posted this week
- Office hours (Prof. and TAs) are now posted on D2L
  - Includes Zoom links
  - On D2L calendar

Today's Lecture

- Key concepts in ML (part 2)
- Notation
- Comment on definition of dataset D (yl x or y,x)
- Regression (part 1)

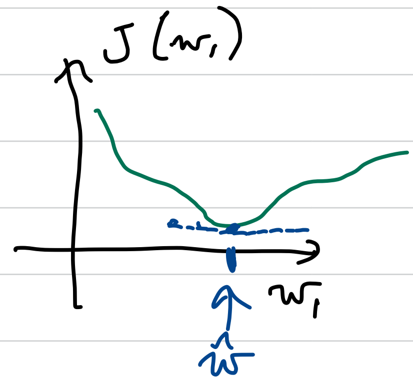
Key issues and concepts in ML (part 2)

3. Optimization method

Method to find extremal value of the objective fcn., J.

Ex:

- Gradient descent
  - > stochastic
  - > batch or mini-batch
  - > 2^{nd} order techniques (e.g., Newton's method)
- Lagrangian opt.
- Solving algebraically (e.g., $\nabla_{\underline{w}} J(\underline{w}, \mathcal{D}) = \underline{0}$)
  - → Pseudoinverse



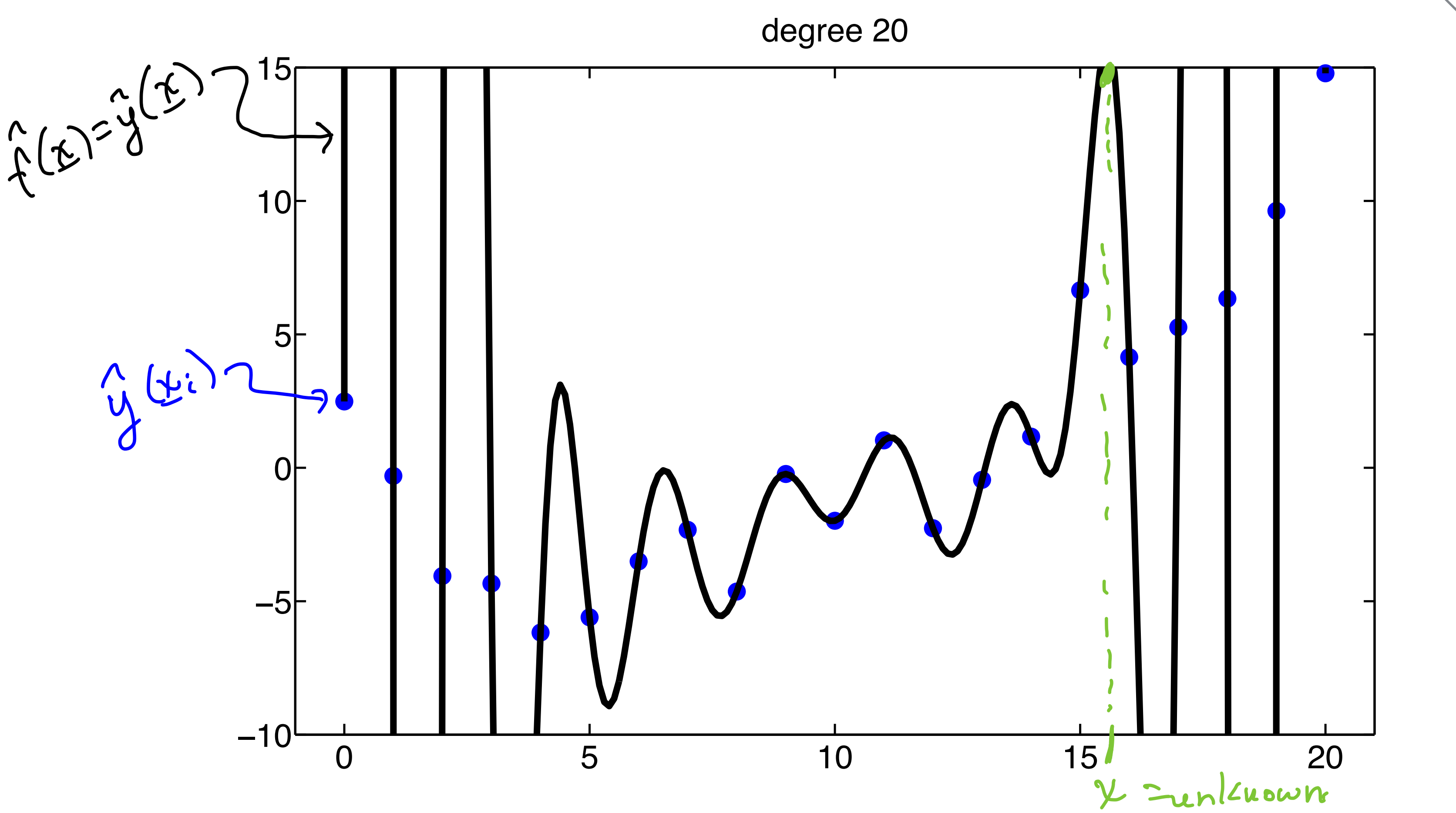
4. Complexity

Ex: Regression problem with 1D input.

Let $J(\underline{w}, \mathcal{D}) = MSE(\hat{y}(x_i), y_i)$

and let hypothesis set be:

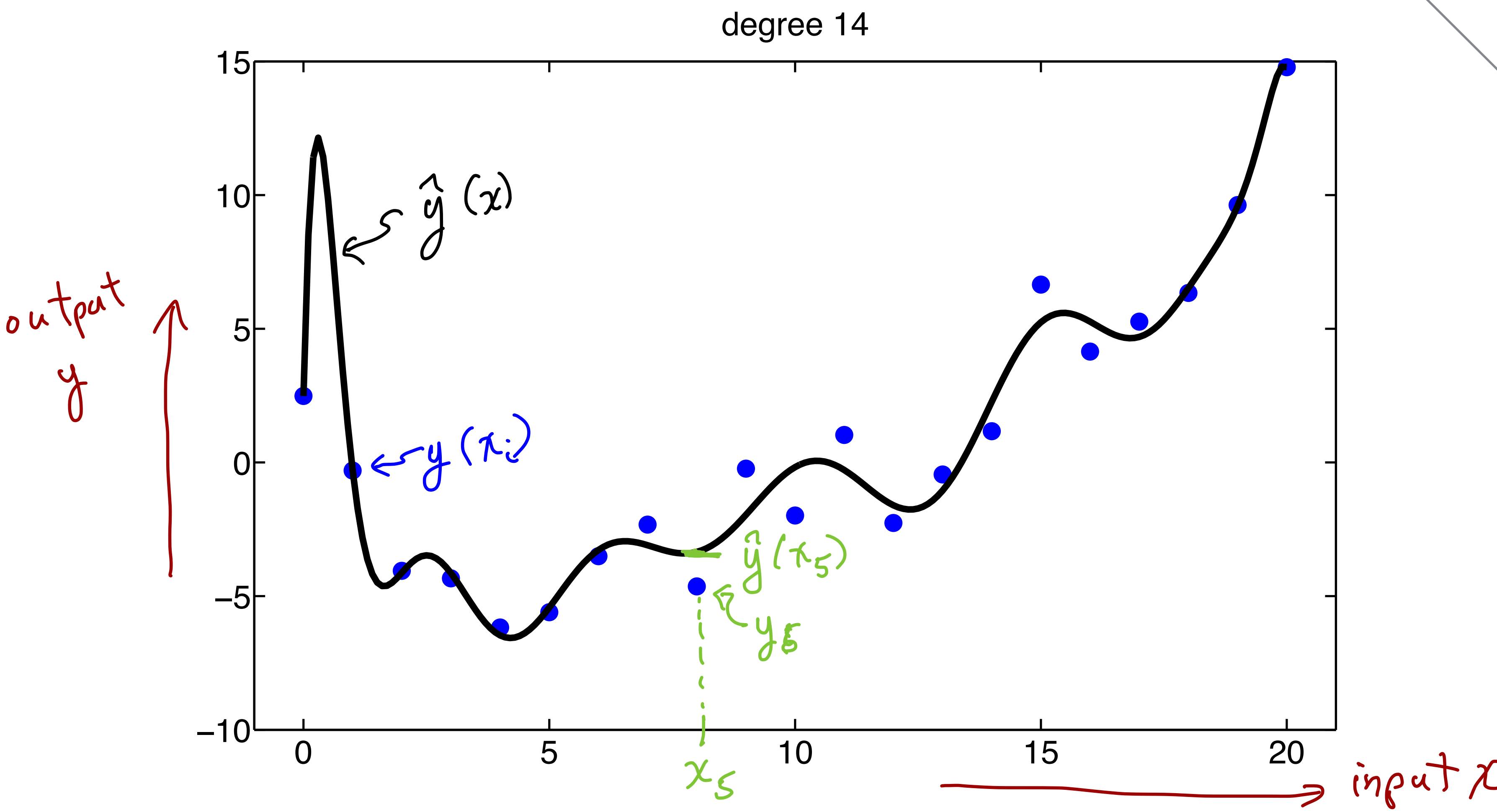
$$\mathcal{H} = \left\{ \hat{f}_d(x) = \sum_{i=0}^{d} w_i x^i \;\middle|\; 1 < d \leq 20,\ d \in \mathbb{Z},\ w_i \in \mathbb{R} \right\}$$



Murphy Fig. 1.18 (b). Regression to fit polynomial of degree 20, to 21 data points (minimizing MSE).

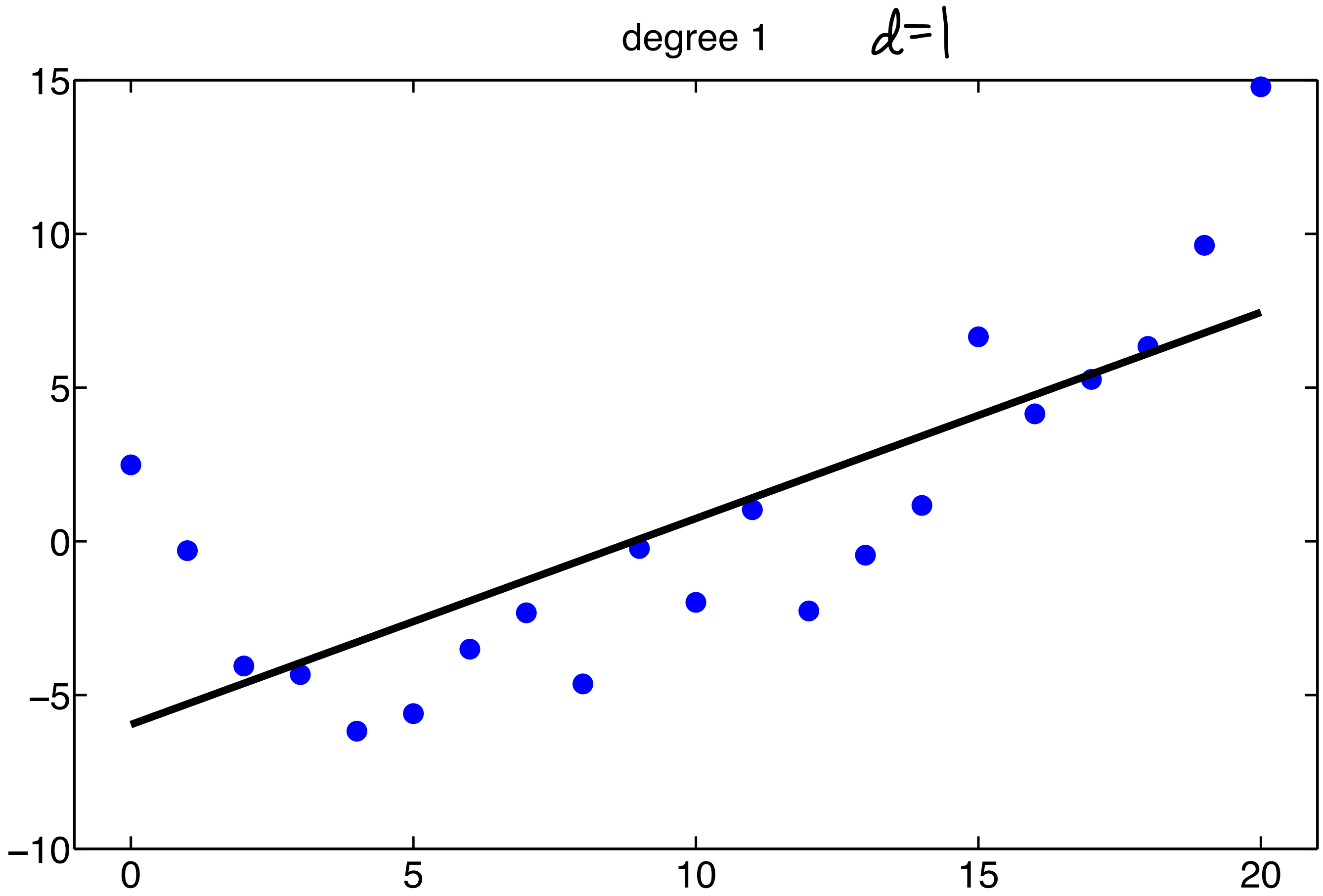
$$MSE_{d=20} = 0.$$

$\Rightarrow$ poor generalization.

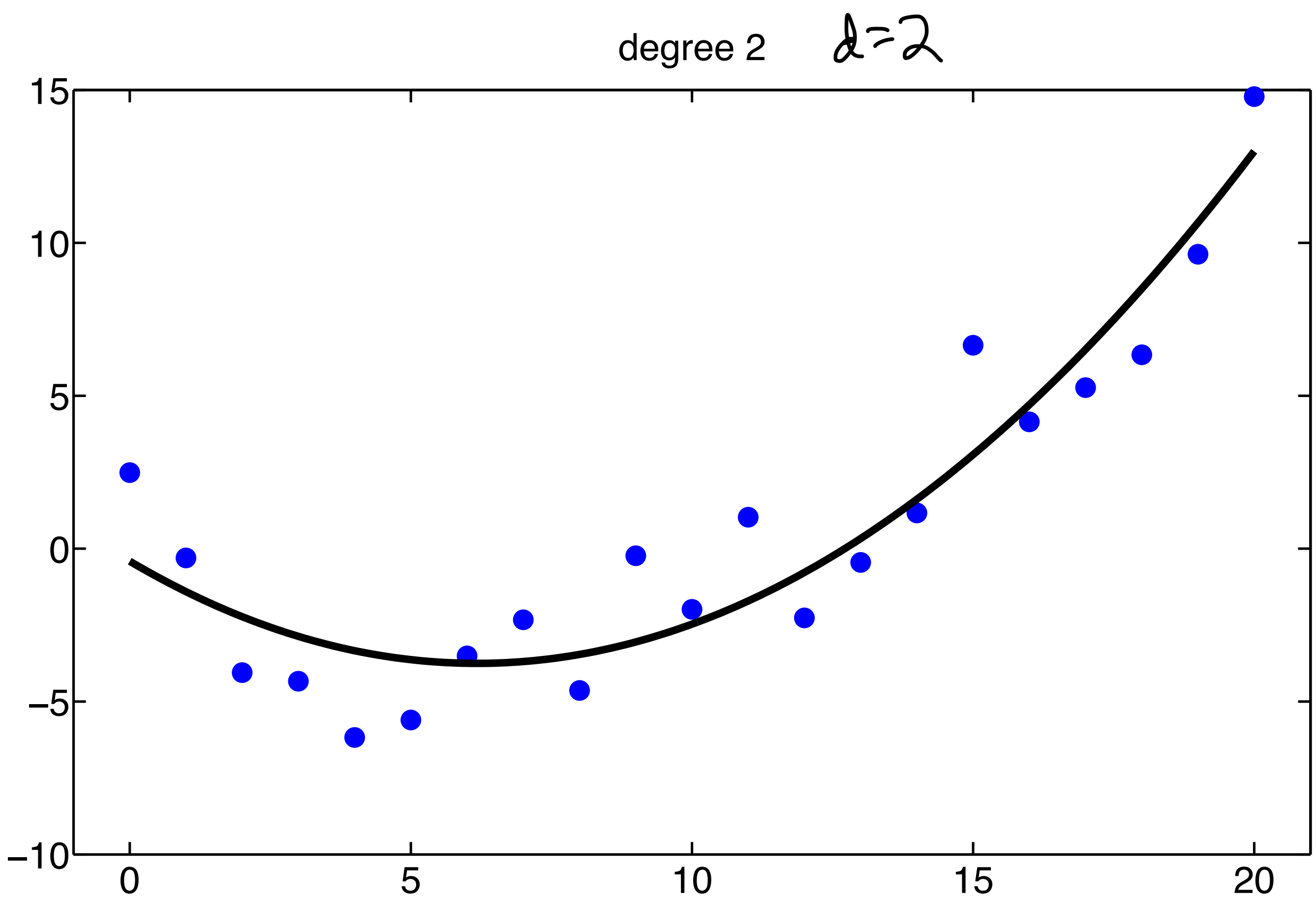


Murphy Fig. 1.18 (a). Regression to fit polynomial of degree 14, to 21 data points (minimizing MSE).

$MSE_{d=14} > 0.$



Murphy Fig. 1.7 (a). Linear regression on 1D data



Murphy Fig. 1.7 (b). Polynomial (degree 2) regression on same 1D data

On training data: $MSE_{d=20} \leq MSE_{d=14} \leq MSE_{d=2} \leq MSE_{d=1}$

<u>Definition of overfitting</u> – fitting a model too closely to the data, resulting in a function $\hat{f}(\underline{x})$ or $\hat{y}(\underline{x})$ that <u>unintentionally</u> models noise, errors, or idiosyncracies in the data.

Compare complexities of dataset and model:

<u>Dataset</u>

$N_{Tr} = 21$

<u>Model</u>

degrees of freedom (d.o.f.)

= # variables adjusted independently during training

= $d+1$

⇒ Complexity of hypothesis set, data, and problem important!

5. Assumptions and priors

Why do we think that $d=14$ is a better fit than $d=20$?

→ Assumption that $f(x)$ is smooth or slowly varying between neighboring data points.

→ Maybe prior knowledge of the problem.

Example:

Testing a patient's blood sugar once per hour.
If patient does not eat during the hour (between data points), then blood sugar will usually vary slowly and smoothly between data points.

How to input assumptions into a ML problem?

→ Balance complexity : N_{Tr}, d.o.f., problem.

→ Include priors or regularizers in the objective fcn.

- e.g., discourage large $|w_i|$.

Notation for augmented & unaugmented quantities

Non-augmented space

$$\underline{w} = \underline{w}^{(0)} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

$$\underline{x} = \underline{x}^{(0)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

Linear $\hat{f}(\underline{x}) = w_0 + \underline{w}^T \underline{x}$

Augmented space

$$\underline{w} = \underline{w}^{(+)} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}$$

$$\underline{x} = \underline{x}^{(+)} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$$

Linear $\hat{f}(\underline{x}) = \underline{w}^T \underline{x}$

(dropping superscripts)

Simarly for $\underline{\phi}^{(0)}(\underline{x})$, $\underline{\phi}^{(+)}(\underline{x})$, and $\underline{\phi}(\underline{x})$.

Regression [Murphy Ch. 7]

House-price prediction

Let $\underline{x} = \begin{bmatrix} 1 \\ \text{living area} \\ \text{No. of rooms} \\ \text{Age of house} \\ \text{Location 1} \\ \text{Location 2} \end{bmatrix}$ and $\underline{w} = \underline{w}^{(t)}$

Linear model: $\hat{f}(\underline{x}) = \underline{w}^T \underline{x}$

Nonlinear model: $\hat{f}(\underline{x}) = \underline{w}'^T \underline{\phi}(\underline{x}) = \sum_{i=1}^{D'} w_i' \phi_i(\underline{x})$

Each $\phi_i(\underline{x})$ = nonlinear fcn. of $\underline{x}$ (or of x_1, x_2, x_3, x_4, x_5)

e.g.: quadratic — each $\phi_i(\underline{x}) = x_1^j x_2^k x_3^l x_4^m x_5^n$

$j, k, l, m, n \in \mathbb{Z}^{\geq 0}$ and $j+k+l+m+n \leq 2$.

$\underline{\phi}$ is "basis set expansion" [Murphy]

"nonlinear transformation" [AML]

$\left.\begin{array}{l}\text{"}\Phi\text{ machine"} \\ \text{"nonlinear mapping"}\end{array}\right\}$ [EE559]

1. Hypothesis set [M 7.2] ($\underline{w} = \underline{w}^{(t)}$)

The o/p can't be exactly described by our $\hat{f}(\underline{x})$. [assumption]

$\Rightarrow$ make the model statistical.

Let's model y as: $p(y \mid \underline{x}, \underline{\theta})$

↑ unknown parameters, to be estimated from $\mathcal{D}$.

→ make <u>assumption</u> about $p(y \mid \underline{x}, \underline{\theta})$.

(i) Here $p(y \mid \underline{x}, \underline{\theta}) = \mathcal{N}(y \mid \underline{w}^T \underline{x}, \sigma^2)$ (linear)

(i)' or $= \mathcal{N}(y \mid \underline{w}^T \underline{\phi}(\underline{x}), \sigma^2)$ (nonlinear)

Equivalent to: $y(\underline{x}) = \underline{w}^T \underline{x} + n, \quad n \sim \mathcal{N}(n \mid 0, \sigma^2)$

↑ or $\underline{\phi}(\underline{x})$

Regression using Maximum Likelihood Estimate (MLE) [M 7.3]

$(\underline{w} = \underline{w}^{(t)})$

$p(\mathcal{D} \mid \underline{\theta})$ = likelihood of $\underline{\theta}$

Estimate $\underline{\theta}$ using MLE: $\quad \underline{\hat{\theta}}_{MLE} \triangleq \underset{\underline{\theta}}{\operatorname{argmax}} \left[\underbrace{\ln p(\mathcal{D} \mid \underline{\theta})}_{\text{or } p(\mathcal{D} \mid \underline{\theta})} \right]$

If we assume (i) or (i)', with σ, $\underline{\phi}$ known, then:

$$\underline{\theta} = \underline{w}$$

and $\underline{\hat{w}}_{MLE} = \underset{\underline{w}}{\operatorname{argmax}} \underbrace{\ln p(\mathcal{D} \mid \underline{w})}$

assume data pts. in $\mathcal{D}$ are independently distributed (i.d.)

$$\sum_{i=1}^{N} \ln p(y_i \mid \underline{x}_i, \underline{\theta})$$