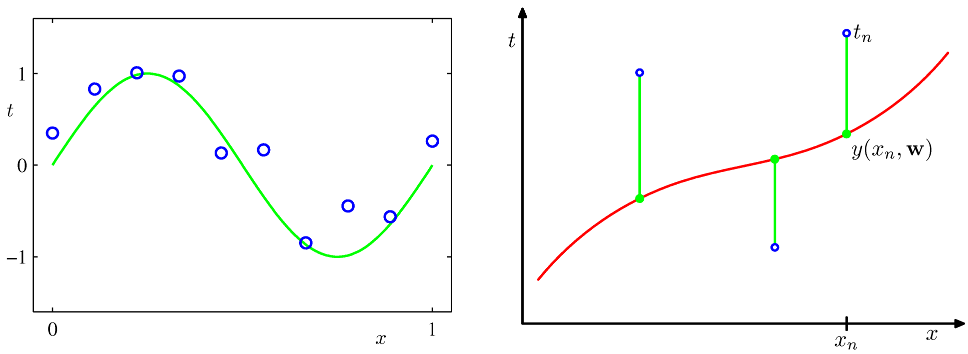
Introduction to Machine Learning
Estimating Parameters, Regularization

Mahesan Niranjan
University of Southampton

DiscNet Summer School

June 2025

Polynomial Curve Fitting



- One dimensioanl inputs and targets: $\boldsymbol{x} = [x_1, x_2, \dots x_N]$; $\boldsymbol{t} = [t_1, t_2, \dots t_N]$
- Fit a function (`Eqn 1.1`):

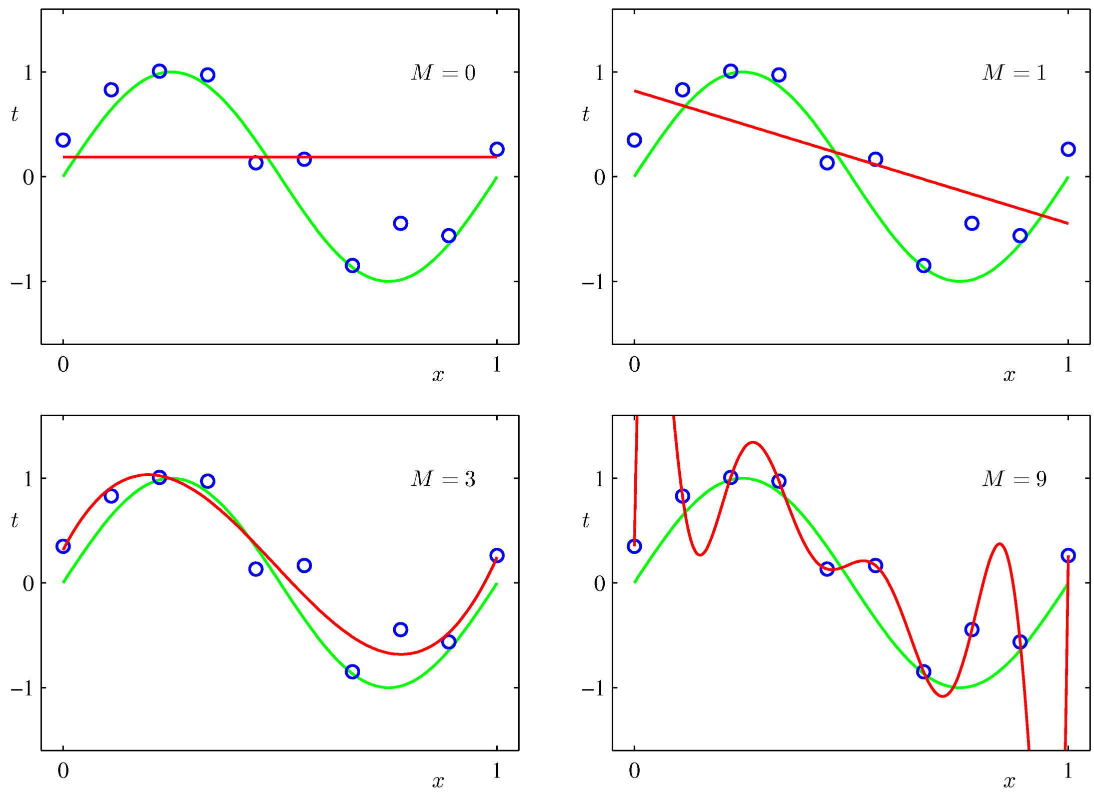
$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- Minimize error:

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

Polynomial Curve Fitting

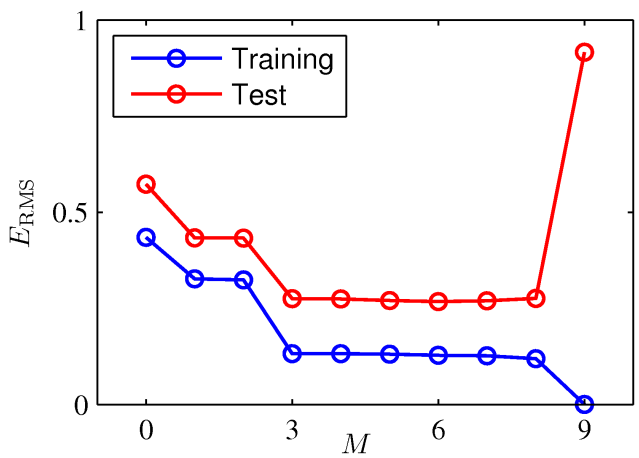
Illustrating Model Complexity



Error with Polynomial Order

- Root mean squared error (RMS)

$$E_{RMS} = \sqrt{2E(\boldsymbol{w}^*)/N}$$



Regularized Least Squares

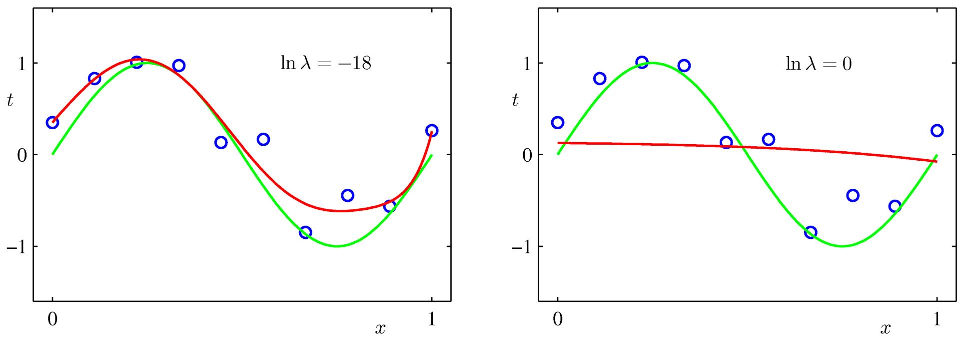
- As M increases, weight magnitudes get large
- Large positive and large negative weights cancelling out and fitting the data

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	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

Regularized Least Squares

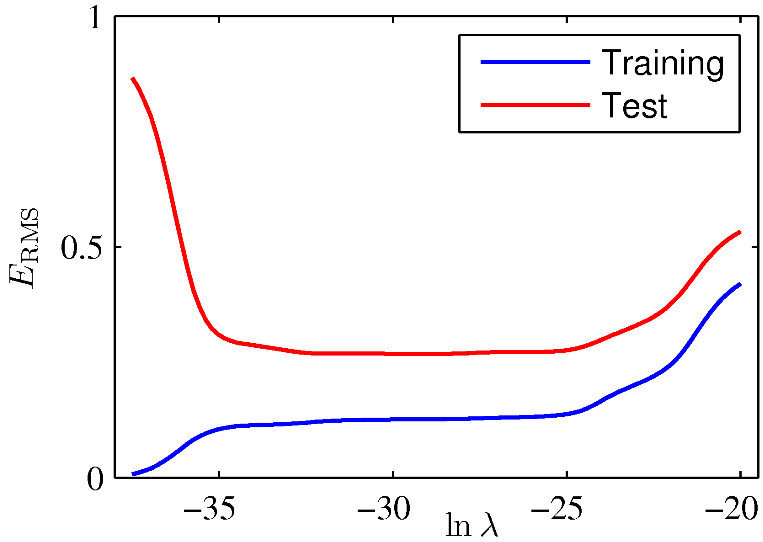
- The way to deal with this is *regularization*, a very important tool in machine learning.

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

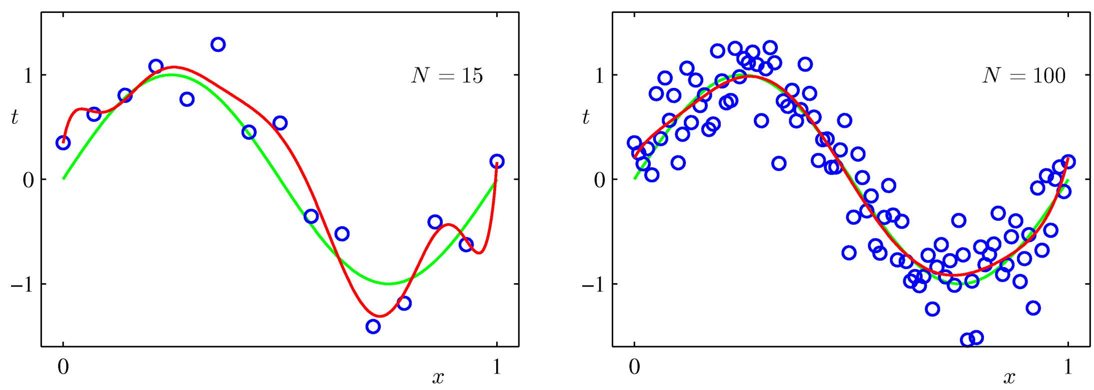


- Note the regulariation weight in the figures is ln λ
- Left hand figure is just about correct, right hand figure is over-doing it.

Choice of Regularization Parameter



Effect of dataset size



- Having a lot of data is good
- Lot of excitement around "*big data*"
- But is it always the case?

Probabilistic Inference

- In polynomial curve fitting earlier, we estimated the coefficients $\boldsymbol{w}$ by minimising the squared error.
- We will formulate it in a probabilistic setting.
- First we note *Bayesian* versus *Frequentist* arguments

$$p(\boldsymbol{w} \mid \mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})\, p(\boldsymbol{w})}{p(\mathcal{D})}$$

- Posterior $\propto$ Likelihood $\times$ Prior
- Denominator is integral over numerator:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{w})\, p(\boldsymbol{w})$$

- We will see this in classification, regression and estimation
- Bayesian vs Frequentist paradigms [Section 1.2.3]

Bayesian	Just one dataset, uncertainty through $p(\boldsymbol{w}) \rightarrow p(\boldsymbol{w}|\mathcal{D})$
Frequentist	$\mathcal{D}$ is just one realization of a process

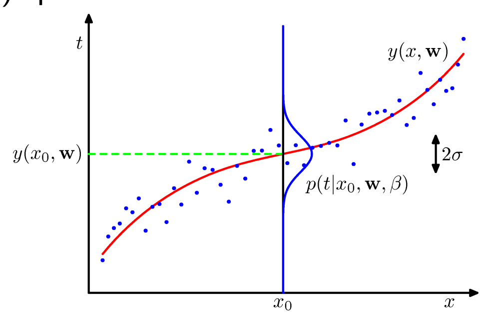
Polynomial Regression in a Probabilistic Setting

Section 1.2.5

- Data consists of Inputs $\boldsymbol{x} = (x_1, x_2, ..., x_N)^T$ and corresponding targets $\boldsymbol{t} = (t_1, t_2, ..., t_N)^T$
- Our probabilistic model is to assume that given x, the target is generated by a function $y(x; \boldsymbol{w})$ and is corrupted by zero mean Gaussian noise:

$$p(t|x; \boldsymbol{w}, \beta) = \mathcal{N}\left(t \mid y(x; \boldsymbol{w}), \beta^{-1}\right)$$

- β is precision parameter, inverse of variance.
- We assume (and usually this is the case) the data are independent, identically distributed (IID)samples from the (x, t) space.



- Likelihood of the data:

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}\right)$$

Probabilistic setting (cont'd)

- Likelihood (again) is a function of the parameters $\boldsymbol{w}$ and β

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \boldsymbol{w}), \beta^{-1}\right)$$

- What values of the unknowns maximises the likelihood?
- It is convenient to work with log likelihood
  - Product turns into summation
  - Exponent can be avoided
- Log likelihood:

$$\ln p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

- When we maximize this respect to $\boldsymbol{w}$, we see we are minimising the mean squared error, giving $\boldsymbol{w}_{ML}$ (ML Estimation)
- How exactly to solve for $\boldsymbol{w}_{ML}$, we will see later.
- When we maximize with respect to β, we get

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

Probabilistic setting (cont'd)

- With a probabilistic model, we now think of a *predictive distribution*

$$p\left(t|x, \boldsymbol{w}_{ML}, \beta_{ML}\right) = \mathcal{N}\left(t|y(x, \boldsymbol{w}_{ML}), \beta_{ML}^{-1}\right)$$

- We can also take a Bayesian inference perspective
- Assume a prior distribution over the parameters – an isotropic Gaussian

$$\begin{aligned} p(\boldsymbol{w}|\alpha) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \\ &= \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\} \end{aligned}$$

- Make sure you understand the above to be a special case of the general expression: $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ seen earlier!
- We can now write a posterior distribution over $\boldsymbol{w}$

$$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta)\, p(\boldsymbol{w}|\alpha)$$

- We can now maximize the posterior distribution (MAP Estimation); the maximum of this is the same as the minimum of

$$\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}$$

- Maximising the posterior is the same as minimising the *regularized* sum of squares.