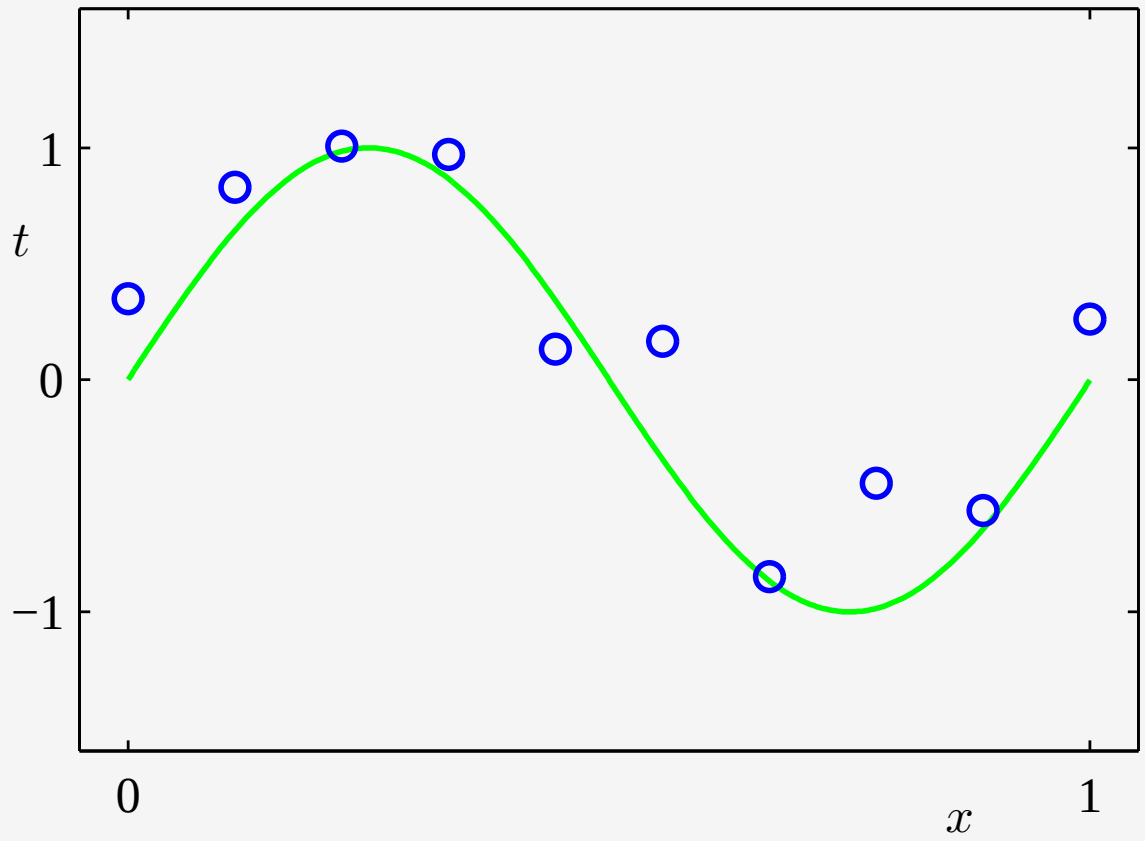
Machine Learning I
Week 2: Regression, Model Selection, Regularization and Validation

Martin Felder, Christian Osendorfer

Technische Universität München

29/30. October 2009

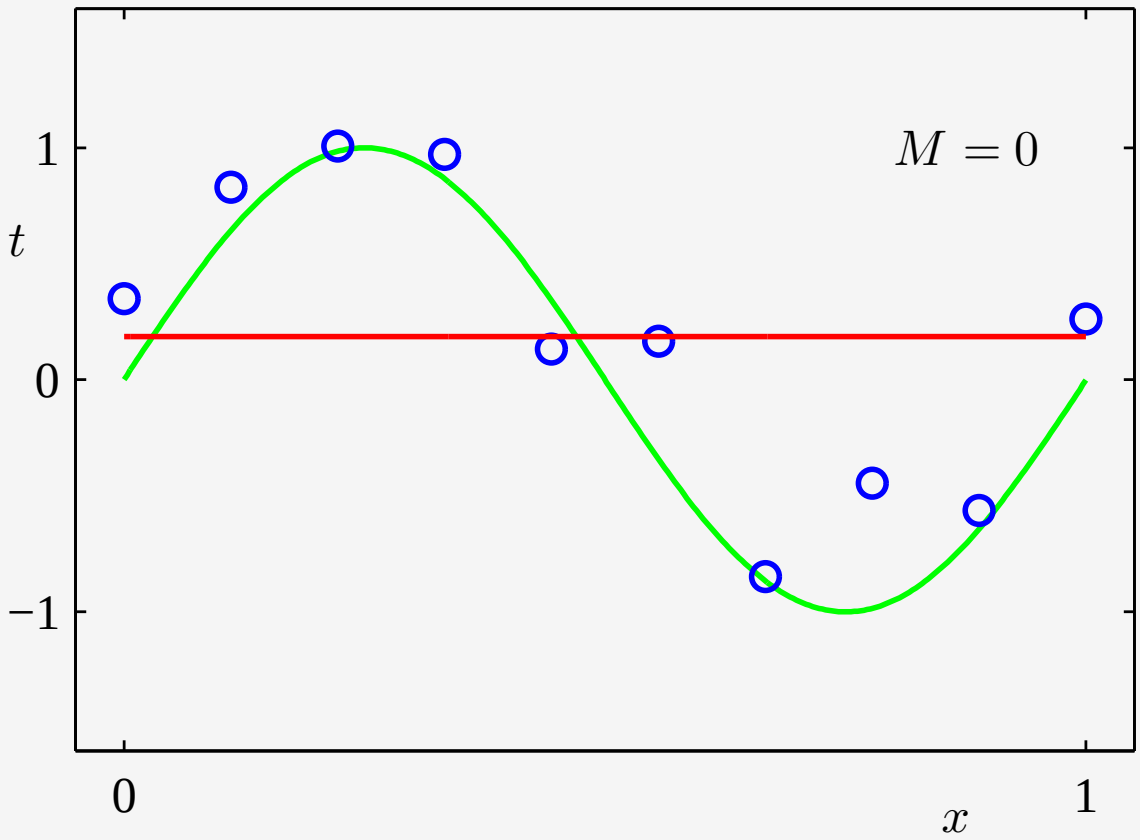
Consider noisy real-valued functions



$$\text{inputs: } \boldsymbol{X} = (x_1, \ldots, x_N)^{\mathsf{T}} \tag{1}$$

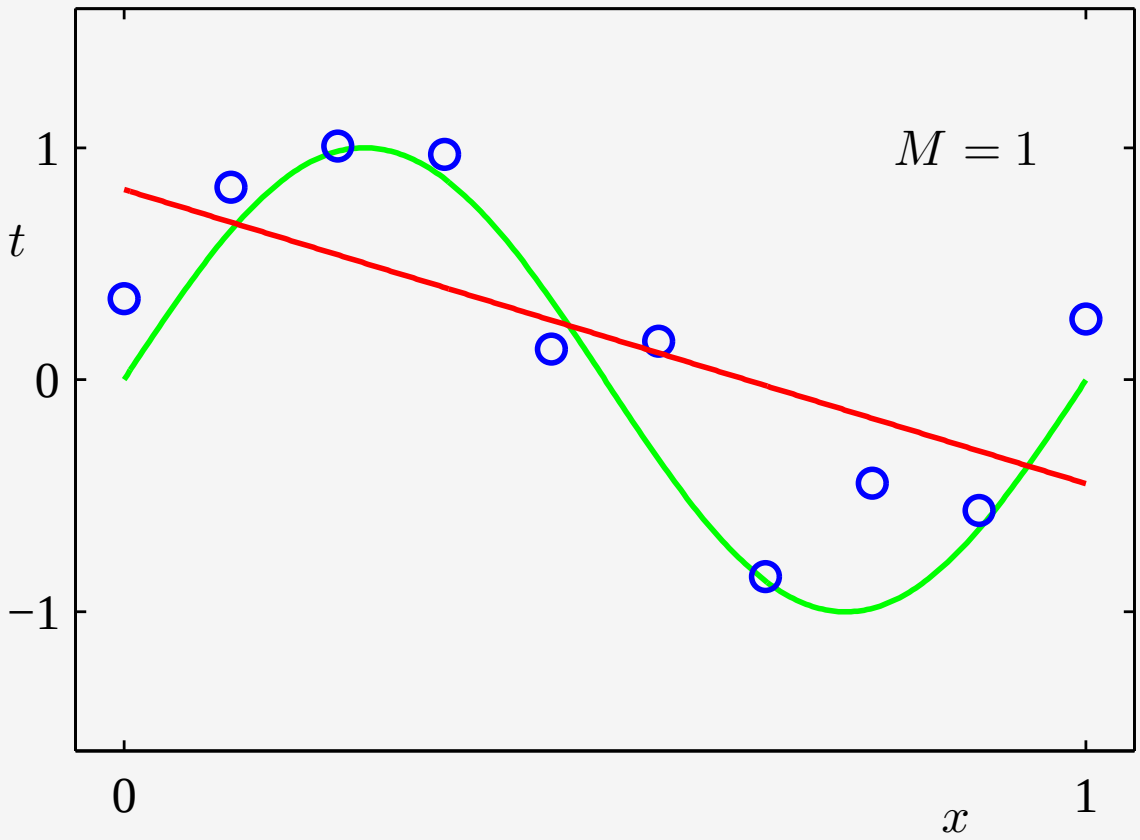
$$\text{targets: } \boldsymbol{T} = (t_1, \ldots, t_N)^{\mathsf{T}}, \quad t_i = h(x_i) + \epsilon = \sin(2\pi x_i) + \epsilon \tag{2}$$

Model: 0th order polynomial



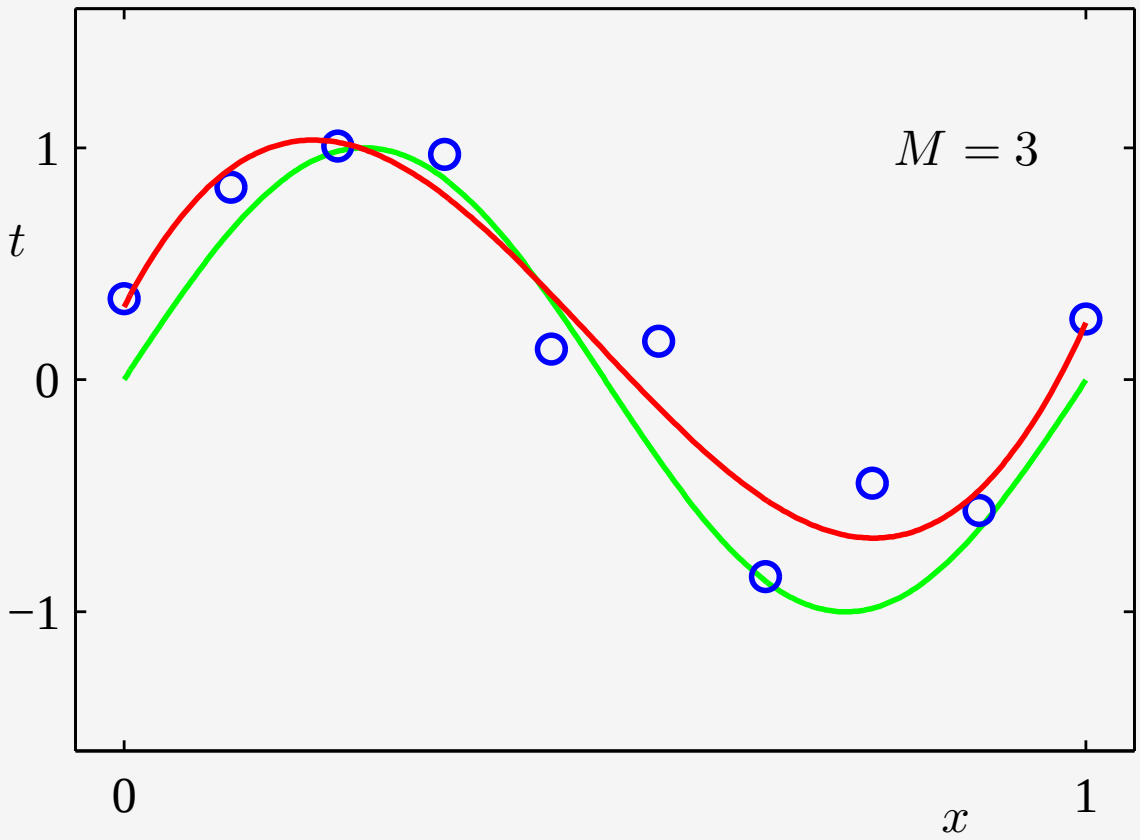
$$y(x, \boldsymbol{w}) = w_0$$

Model: 1st order polynomial



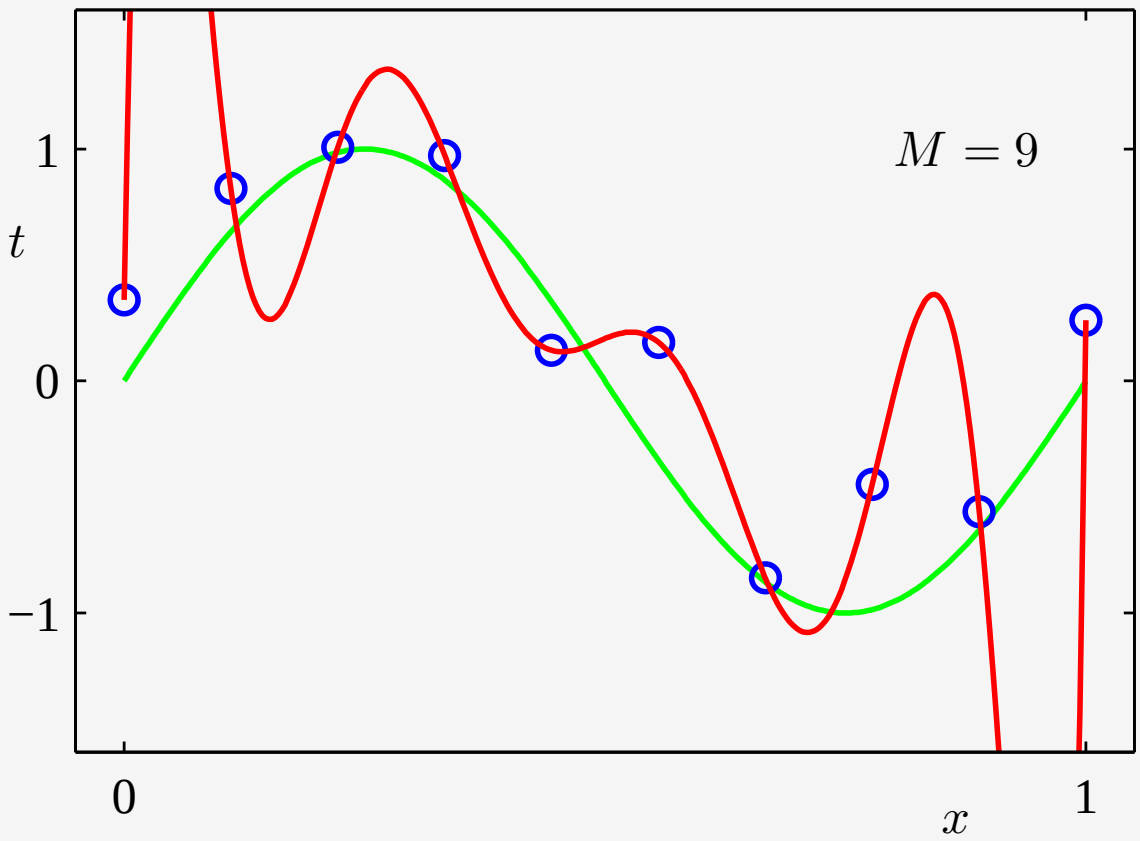
$$y(x, \boldsymbol{w}) = w_0 + w_1 x$$

Model: 3rd order polynomial

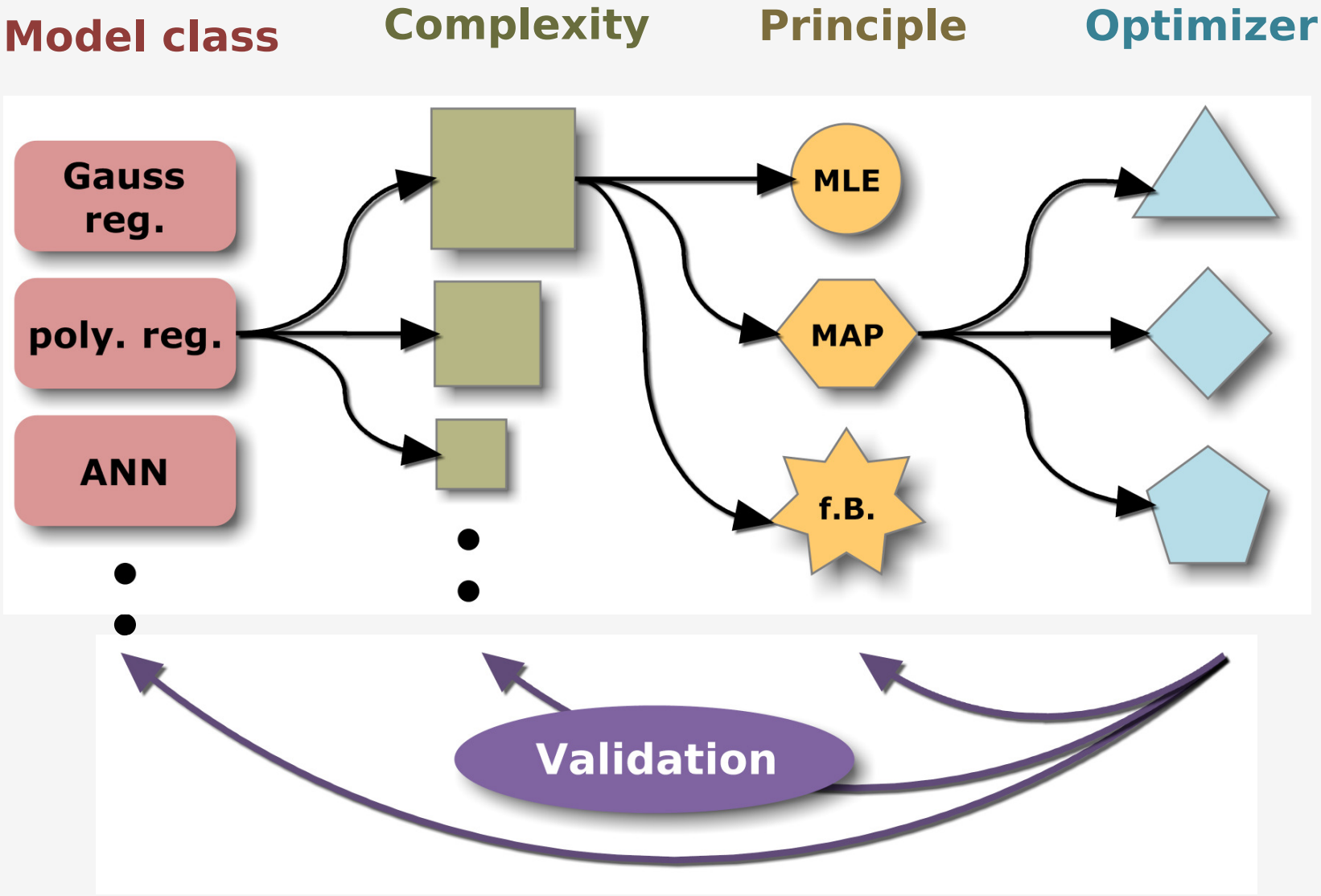


$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + + w_3 x^3$$

Model: 9th order polynomial



$$y(x, \boldsymbol{w}) = \sum_{j=0}^{M} w_j x$$



Problem Definition

Have: Input vector $\boldsymbol{x}$, model y with parameters $\boldsymbol{w}$
Question: What is output value t?

$\Longrightarrow$ construct model from M functions $\phi(\boldsymbol{x})$:

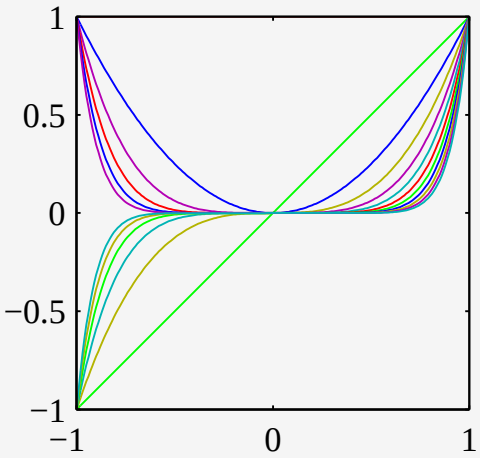
$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}} \phi(\boldsymbol{x}) \tag{3}$$

where

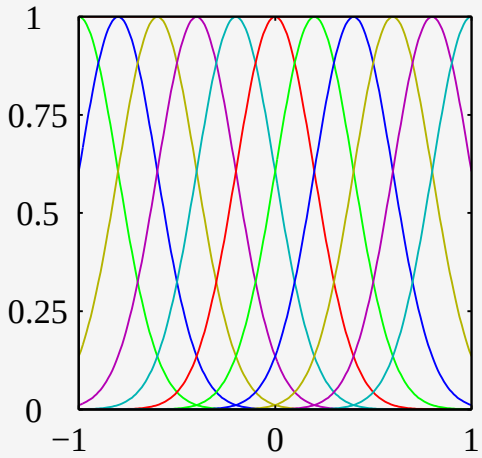
ϕ	basis function	–	many choices, can be nonlinear
w_0	bias	–	equivalent to defining $\phi_0 \equiv 1$

$\Longrightarrow$ still linear in $\boldsymbol{w}$
$\Longrightarrow$ compare to Taylor expansion, Fourier transform, wavelets...

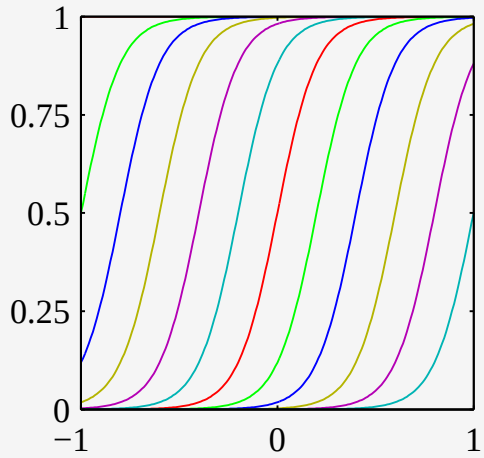
Typical Basis Functions



polynomials



Gaussians



“sigmoids”
(=S-shaped curves)

Now assume we measure random variable t as

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon \qquad [\epsilon\text{: Gaussian, zero mean}] \tag{4}$$

$$\Rightarrow \mathcal{E}[t|\boldsymbol{x}] = y(\boldsymbol{x}, \boldsymbol{w}) \qquad \text{for least squares loss fct.}$$

Let dataset $\mathcal{D}$ consist of a series of observations $\boldsymbol{T} = (t_1, t_2, \ldots, t_N)$ and corresponding input vectors $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$. Then the likelihood function is

$$p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod^{N} \mathcal{N}(t_n|\boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n), \beta^{-1}) \qquad [\beta^{-1}\text{: precision}] \tag{5}$$

As usual, it's easier to deal with logarithmized probabilities (leave out $\boldsymbol{X}$ for brevity):

$$\ln p(\boldsymbol{T}|\boldsymbol{w}, \beta) = \frac{N}{2}\ln\beta - \underbrace{\frac{N}{2}\ln(2\pi)}_{const.} - \beta \underbrace{\frac{1}{2}\sum^{N}\left(t_n - \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n)\right)^2}_{E_{\mathcal{D}}(\boldsymbol{w})} \tag{6}$$

Maximum Likelihood Solution

Now find max likelihood by setting gradient w.r.t. $\boldsymbol{w}$ to zero:

$$\nabla_{\boldsymbol{w}} \ln p(\boldsymbol{T}|\boldsymbol{w}, \beta) = \sum_{n}^{N} \left(t_n - \boldsymbol{w}^{\mathrm{T}} \phi(\boldsymbol{x}_n)\right) \phi(\boldsymbol{x}_n) \tag{7}$$

$$\Rightarrow \boldsymbol{w}_{\mathrm{ML}} = \underbrace{(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}}_{=\boldsymbol{\Phi}^{\dagger}} \boldsymbol{T} \qquad \leftarrow \text{normal equations of least sq. prob.} \tag{8}$$

$\boldsymbol{\Phi}^{\dagger}$ is called Moore-Penrose pseudo-inverse of $\boldsymbol{\Phi}$ (because for a square matrix, $\boldsymbol{\Phi}^{\dagger} = \boldsymbol{\Phi}^{-1}$). Furthermore,

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

What else can we say about the ML solution?

1.) Usually, $\phi_0(\boldsymbol{x}) \equiv 1$ (bias!) $\hookrightarrow$ pull w_0 out of the sum $\hookrightarrow$ solve $\nabla \ln p = 0$ again $\hookrightarrow$

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j} \tag{9}$$

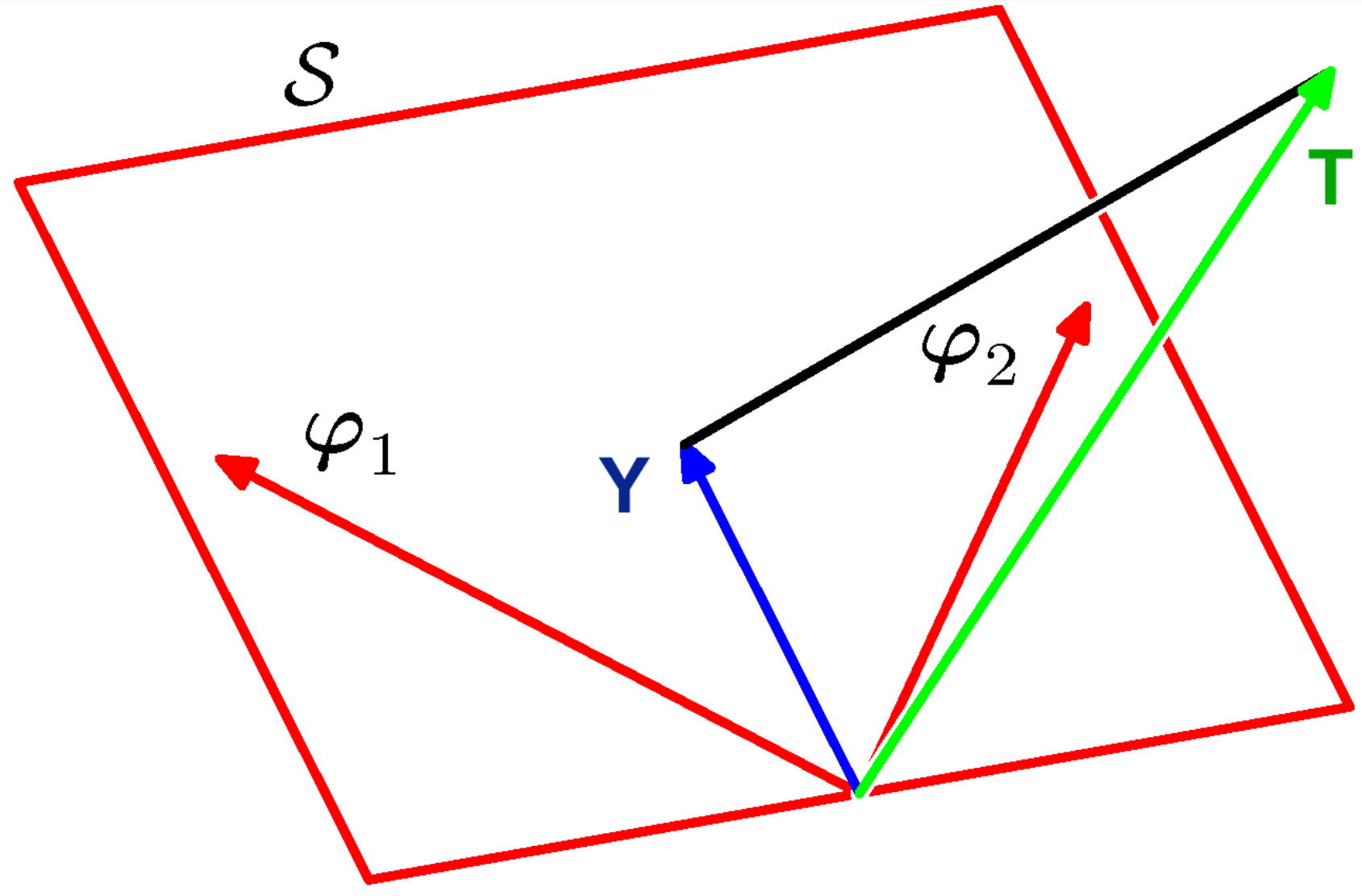
$\Rightarrow$ bias compensates for difference between (training set) averages of targets and weighted linear reconstruction.

2.) Maximize likelihood w.r.t. β $\hookrightarrow$

$$\beta_{\text{ML}}^{-1} = \frac{1}{N} \sum^{N} \left(t_n - \boldsymbol{w}_{\text{ML}}^{\text{T}} \phi(\boldsymbol{x}_n)\right)^2 \tag{10}$$

$=$ inverse variance for residual scatter of weighted linear reconstruction around targets.
$\Rightarrow$ if $\boldsymbol{w}_{\text{ML}}^{\text{T}} \phi$ is perfect fit to generating fct. $y(\boldsymbol{x}, \boldsymbol{w})$, then
$\beta_{\text{ML}} \rightarrow$ inverse variance of error ϵ.

Geometrical Interpretation



The Bias-Variance Dilemma

We will follow the derivation of Bishop (1995) here. Consider again the quadratic error function, and let $N \to \infty$:

$$E_{\mathcal{D}} = \lim_{N\to\infty} \frac{1}{2N} \sum_n (y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n)^2 \tag{11}$$

$$= \frac{1}{2} \int (y(\boldsymbol{x}, \boldsymbol{w}) - t)^2 \, p(t, x) \, \mathrm{d}t \, \mathrm{d}x \tag{12}$$

Now define

$$\langle t|x \rangle \equiv \int t \, p(t|x) \, dt \tag{13}$$

$$\langle t^2|x \rangle \equiv \int t^2 \, p(t|x) \, dt. \tag{14}$$

The Bias-Variance Dilemma

By using the product rule, and including a zero expressed as $+\langle t|x\rangle - \langle t|x\rangle$ into the quadratic term, it can be shown that

$$\begin{aligned} E_{\mathcal{D}} &= \frac{1}{2}\int (y(\boldsymbol{x},\boldsymbol{w}) - \langle t|x\rangle)^2\, p(x)\, dx \\ &\quad + \frac{1}{2}\int (\langle t^2|x\rangle - \langle t|x\rangle^2)\, p(x)\, dx \end{aligned} \tag{15}$$

Only the first integral is relevant for minimization of $E_{\mathcal{D}}$, the second one merely sets a lower limit to the achievable error. Note that the absolute minimum $\boldsymbol{w}^*$ of $E_{\mathcal{D}}$ is given by

$$y(x,\boldsymbol{w}^*) = \langle t|x\rangle, \tag{16}$$

i.e. the average model output is equal to the average of the training data. In practice, this state will be reached first where the training data density $p(x)$ is highest.

The Bias-Variance Dilemma

However, the size of $(y(\boldsymbol{x}, \boldsymbol{w}) - \langle t|x \rangle)^2$ depends on the exact choice of training data $\mathcal{D}$ – there are infinitely many realizations! To keep the notation uncluttered, define

$$y_{\mathcal{D}} \equiv y(\boldsymbol{x}, \boldsymbol{w}|\mathcal{D}). \tag{17}$$

This dependence can be eliminated by considering the ensemble average $\mathcal{E}[\cdot] = \mathbb{E}_{\mathcal{D}}[\cdot]$ over many different training data sets $\mathcal{D}$:

$$\mathcal{E}\left[(y_{\mathcal{D}} - \langle t|x \rangle)^2\right] \tag{18}$$

Again, ideally this would be zero, but in practice no model $y_{\mathcal{D}}$ will perform equally well on all different training data sets, since it does not contain the full information of the infinite generator.

The Bias-Variance Dilemma

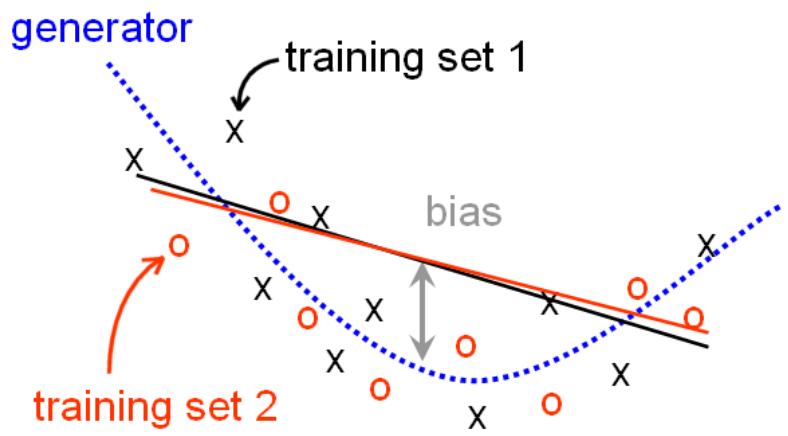
By introducing another zero made from $+\mathcal{E}[y_\mathcal{D}] - \mathcal{E}[y_\mathcal{D}]$, one obtains

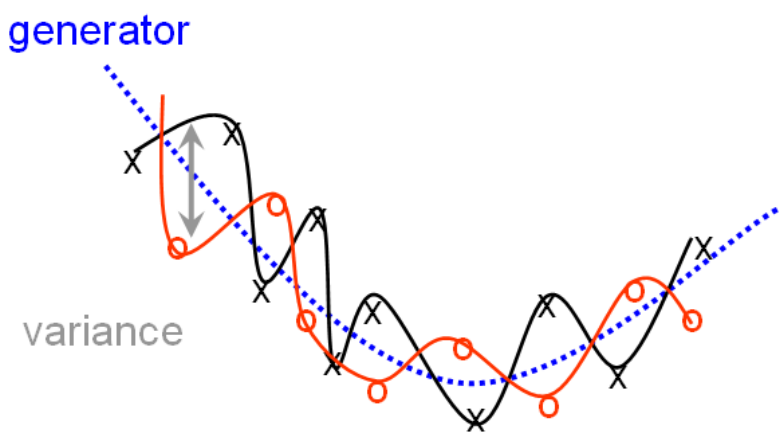
$$\mathcal{E}\left[(y_\mathcal{D} - \langle t|x\rangle)^2\right] = \tag{19}$$
$$\underbrace{(\mathcal{E}[y_\mathcal{D}] - \langle t|x\rangle)^2}_{\text{bias}^2} + \underbrace{\mathcal{E}\left[(y_\mathcal{D} - \mathcal{E}[y_\mathcal{D}])^2\right]}_{\text{variance}}$$

Bias = model pays not enough heed to the data, or is insufficiently complex.

Variance = model too strongly data-driven, cannot detect underlying generator.

$$\underbrace{\left(\mathcal{E}[y_{\mathcal{D}}] - \langle t|x\rangle\right)^2}_{\text{bias}^2} + \underbrace{\mathcal{E}\left[(y_{\mathcal{D}} - \mathcal{E}[y_{\mathcal{D}}])^2\right]}_{\text{variance}}$$





$\longrightarrow$ Regularization $\longleftarrow$
mediates between these extremes.

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	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

MLE often suffers from overfitting $\hookrightarrow$ use regularization:

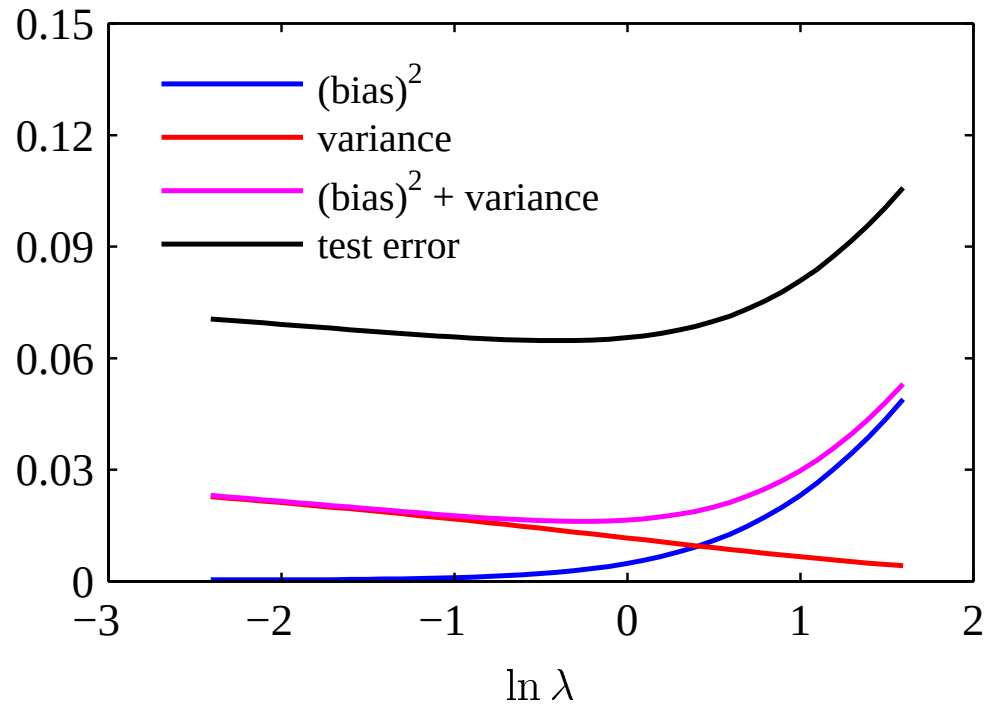
$$E_{\mathcal{D}} = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n)\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{20}$$

$\Rightarrow$ this is like a Lagrange term specifying an additional constraint:

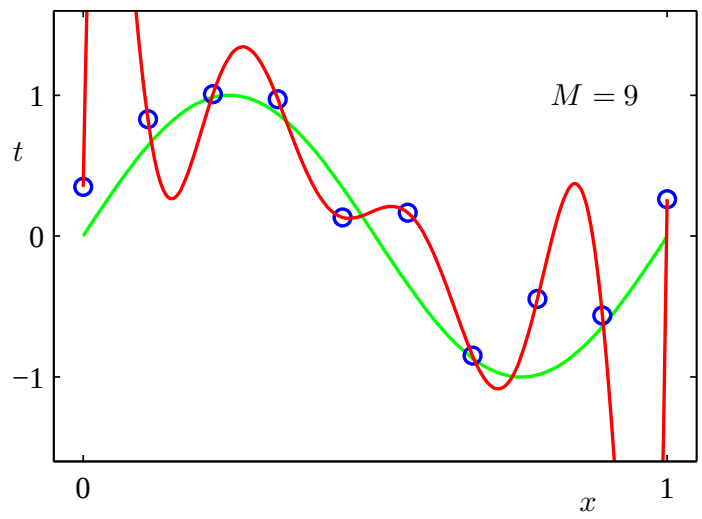
$$\sum_{j=1}^{M}|w_j|^q \leq \varepsilon$$

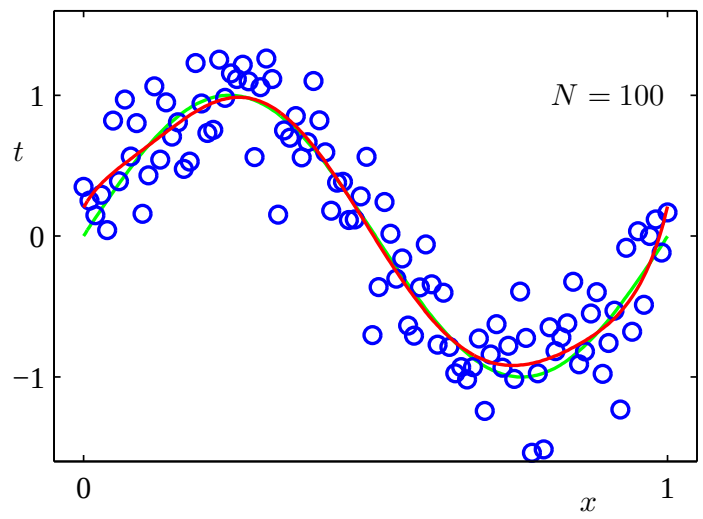
$\hookrightarrow$ most often, use quadratic regularizer: $q = 2$, i.e.

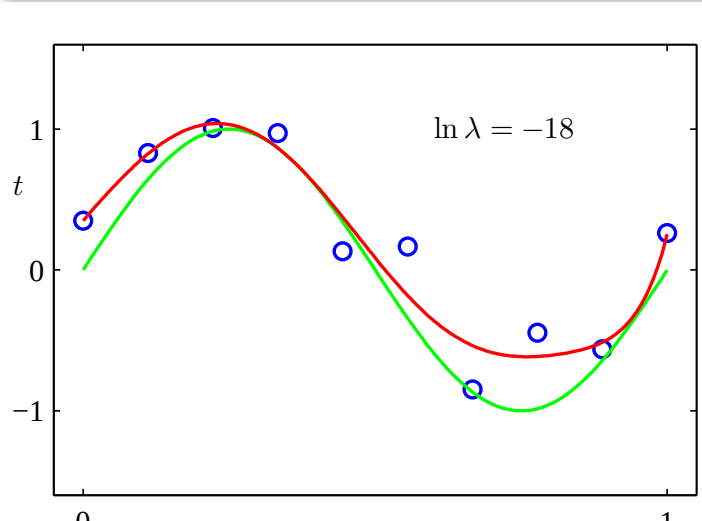
$$\sum_{j=1}^{M}|w_j|^q = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}$$

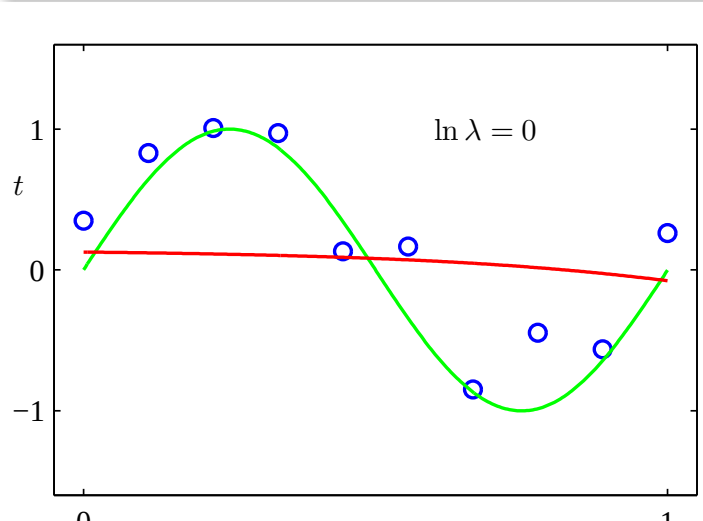


Regularization and Model Complexity









MAP Solution

Recap: For the coin flip experiment, we introduced prior information to prevent overfitting. By analogy:

	train data	likelihood	prior	posterior
coin:	$\mathcal{D} = \boldsymbol{X}$	$p(\mathcal{D}\|\theta)$	$p(\theta\|a, b)$	$p(\theta\|\mathcal{D})$
regr.:	$\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{T}\}$	$p(\boldsymbol{T}\|\boldsymbol{X}, \boldsymbol{w}, \beta)$	$p(\boldsymbol{w}\|\cdot)$	$p(\boldsymbol{w}\|\boldsymbol{X}, \boldsymbol{T}, \cdot)$

Note: As mentioned earlier, $\boldsymbol{X}$ is usually dropped for clarity.

Prior: How to find a good one?

- recall (from Eq. 5) that our likelihood function $p(\boldsymbol{T}|\boldsymbol{w}, \beta)$ is a Gaussian,
- treat precision $\beta = 1/\sigma^2$ as a known parameter (for now),
- know that the conjugate prior for a Gaussian with known variance is also a Gaussian.

MAP Solution

Hence introduce the following prior:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0) \qquad [\boldsymbol{m}_0\text{: mean, } \boldsymbol{S}_0\text{: covariance}] \tag{21}$$

Often we don't know much about the prior distribution anyway. For a suitably designed model with independent parameters $\boldsymbol{w}$, the following prior is usually reasonable:

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0 = 0, \boldsymbol{S}_0 = \alpha^{-1}\boldsymbol{I}) \tag{22}$$

This results in the posterior:

$$p(\boldsymbol{w}|\boldsymbol{T}, \alpha, \beta) \propto p(\boldsymbol{T}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha) \tag{23}$$

Applying the log yields again an error function E_{MAP} to minimize:

$$\ln p(\boldsymbol{w}|\boldsymbol{T}, \alpha, \beta) = -\frac{\beta}{2}\sum^{N}\left(t_n - \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n)\right)^2 - \frac{\alpha}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} + \text{const.} \tag{24}$$

$\Rightarrow$ just like squared error minimization with a quadratic regularizer weighted with $\lambda = \frac{\alpha}{\beta}$ (Eq. 20).

Posterior Distribution

We can actually find a closed expression for the posterior!

Posterior parameter distribution

$$p(\boldsymbol{w}|\boldsymbol{T}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N) \tag{25}$$

with

$$\boldsymbol{m}_N = \boldsymbol{S}_N \left(\boldsymbol{S}_0^{-1} \boldsymbol{m}_0 + \beta \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{T} \right) \tag{26}$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{\Phi} \tag{27}$$

Properties of the posterior:

- Since we again have a Gaussian, the maximum posterior solution equals the mode: $\boldsymbol{w}_{\text{MAP}} = \boldsymbol{m}_N$.
- In the limit of an infintely broad prior, $\boldsymbol{S}_0^{-1} \rightarrow 0$, therefore $\boldsymbol{w}_{\text{MAP}} \rightarrow \boldsymbol{w}_{\text{ML}} = \boldsymbol{\Phi}^{\dagger} \boldsymbol{T}$ (Eq. 8).
- For $N = 0$, i.e. no data points, we get the prior back.

Posterior Distribution for a Simple Prior

If we look at our simplified case:

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0 = 0, \boldsymbol{S}_0 = \alpha^{-1}\boldsymbol{I}) \tag{28}$$

the posterior parameters simplify to:

$$\boldsymbol{m}_N = \beta \boldsymbol{S}_N \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{T} \tag{29}$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{\Phi} \tag{30}$$

A simple example

Bayesian regression for the target values

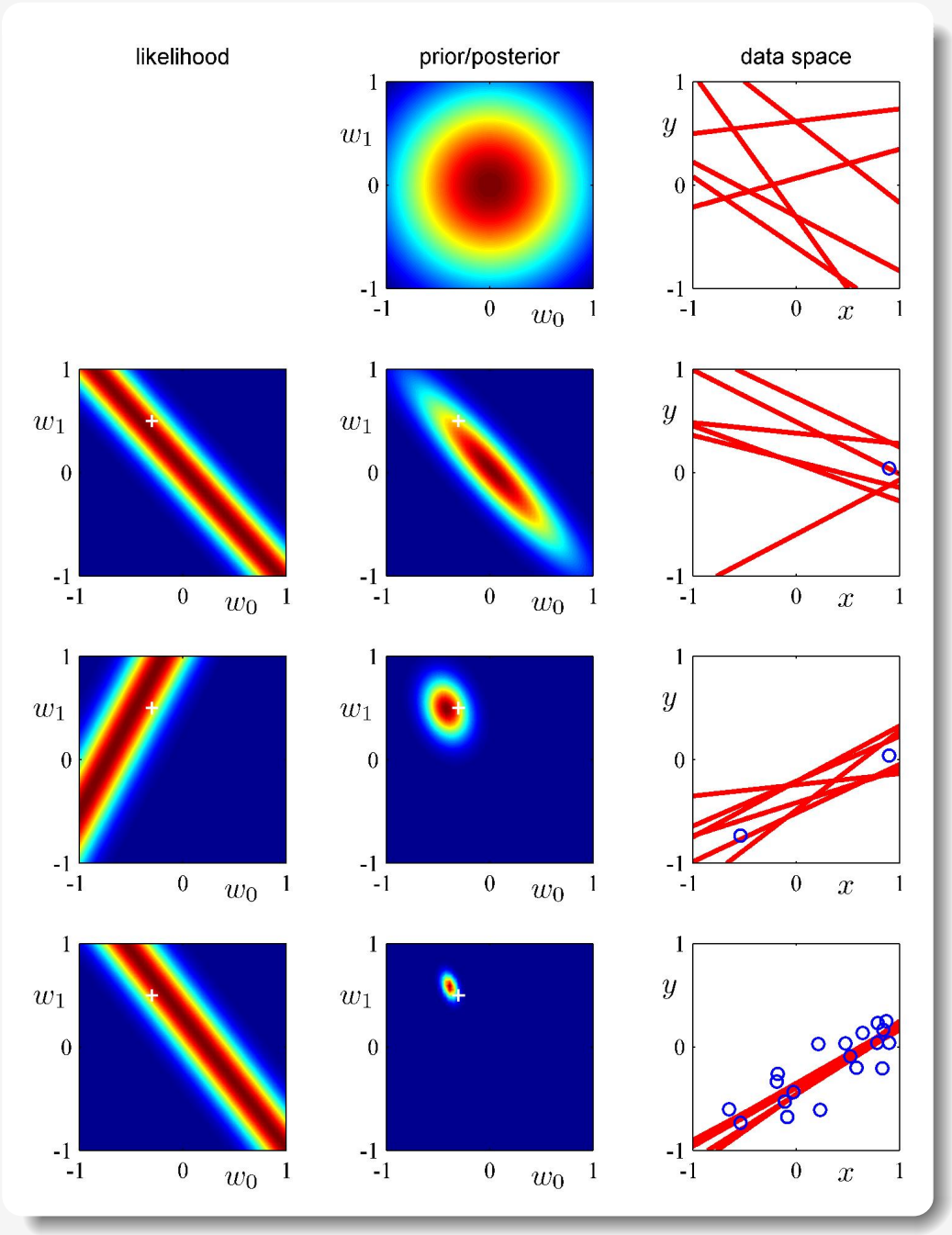
$$t_n = -0.3 + 0.5x_n + \epsilon$$

where ϵ is a Gaussian noise term ($\sigma = 0.2$).

To model this, we set $\phi(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}$ and thus

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

Sequential Estimation: The demo shows how the posterior's breadth gets smaller as more and more points t are taken into account, and how its mode converges to the optimum (=correct) values of the weights (white cross).



Predictive Distribution

Usually, we want to know output t for new values of $\boldsymbol{x}$ – the model parameters $\boldsymbol{w}$ are just a means to achieve this. To predict t, evaluate

$$p(t|\boldsymbol{x}, \boldsymbol{T}, \alpha, \beta) = \int \underbrace{p(t|\boldsymbol{x}, \boldsymbol{w}, \beta)}_{\text{likelihood (5)}} \underbrace{p(\boldsymbol{w}|\boldsymbol{T}, \alpha, \beta)}_{\text{posterior (25)}} \, \mathrm{d}\boldsymbol{w} \tag{31}$$

(coin flip analogy: $p(x|\mathcal{D}, a, b) = \int_0^1 p(x|\theta) p(\theta|\mathcal{D}, a, b) \, \mathrm{d}\theta$)

Predictive distribution

$$p(t|\boldsymbol{x}, \boldsymbol{T}, \alpha, \beta) = \mathcal{N}(t|\boldsymbol{m}_N^\mathsf{T} \phi(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})) \tag{32}$$

with variance

$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \phi(\boldsymbol{x})^\mathsf{T} \boldsymbol{S}_N \phi(\boldsymbol{x}). \tag{33}$$

Example (using 9 radial basis functions)