

# Linear regression

---

Course of Machine Learning  
Master Degree in Computer Science  
University of Rome "Tor Vergata"

Giorgio Gambosi

a.a. 2017-2018

- Linear combination of input features

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D$$

with  $\mathbf{x} = (x_1, \dots, x_D)$

- Linear function of parameters  $\mathbf{w}$
- Linear function of features  $\mathbf{x}$ . Extension to linear combination of **base functions**

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

- Let  $\phi_0(\mathbf{x}) = 1$ , then  $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$

# Base functions

- Many types:
  - Polynomial (global functions)

$$\phi_j(x) = x^j$$

- Gaussian (local)

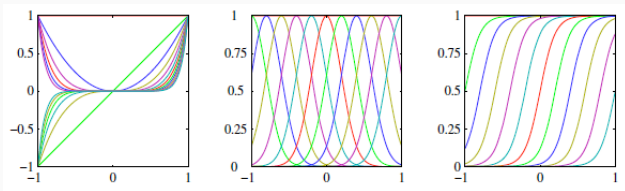
$$\phi_j(x) = \exp\left(-\frac{(x - \mu_j)^2}{2s^2}\right)$$

- Sigmoid (local)

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right) = \frac{1}{1 + e^{-\frac{x - \mu_j}{s}}}$$

- Hyperbolic tangent (local)

$$\phi_j(x) = \tanh(x) = 2\sigma(x) - 1 = \frac{1 - e^{-\frac{x - \mu_j}{s}}}{1 + e^{-\frac{x - \mu_j}{s}}}$$



# Maximum likelihood and least squares

- Assume an additional gaussian noise

$$t = y(\mathbf{x}, \mathbf{w}) + \varepsilon$$

with

$$p(\varepsilon) = \mathcal{N}(\varepsilon|0, \beta^{-1})$$

$\beta = \frac{1}{\sigma^2}$  is the **precision**.

- Then,

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

and the expectation of the conditional distribution is

$$E[t|\mathbf{x}] = \int t p(t|\mathbf{x}) dt = y(\mathbf{x}, \mathbf{w})$$

- The likelihood of a given training set  $\mathbf{X}, \mathbf{t}$  is

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

- The corresponding log-likelihood is then

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

where

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N \left( t_i - \mathbf{w}^T \phi(\mathbf{x}_i) \right)^2 = \frac{1}{2} (\Phi \mathbf{w} - \mathbf{y})^T (\Phi \mathbf{w} - \mathbf{y})$$

# Maximum likelihood and least squares

- Maximizing the log-likelihood w.r.t.  $\mathbf{w}$  is equivalent to minimizing the error function  $E_D(\mathbf{w})$
- Maximization performed by setting the gradient to 0

$$\begin{aligned}\mathbf{0} &= \frac{\partial}{\partial \mathbf{w}} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^N \left( t_i - \mathbf{w}^T \phi(\mathbf{x}_i) \right) \phi(\mathbf{x}_i)^T \\ &= \sum_{i=1}^N t_i \phi(\mathbf{x}_i)^T - \mathbf{w}^T \left( \sum_{i=1}^N \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^T \right)\end{aligned}$$

- Result

$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

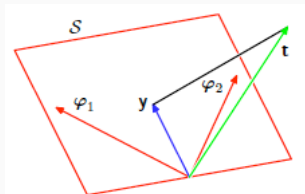
normal equations for least squares

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

$$\beta_{ML}^{-1} = \frac{1}{N} \sum_{i=1}^N \left( t_i - \mathbf{w}_{ML}^T \phi(\mathbf{x}_i) \right)^2$$

## Least squares geometry

- $\mathbf{t} = (t_1, \dots, t_N)^T$  is a vector in  $\mathbb{R}^N$
- Each basis function  $\phi_j$  applied to  $\mathbf{x}_1, \dots, \mathbf{x}_N$  is a vector  $\varphi_j = (\phi_j(\mathbf{x}_1), \dots, \phi_j(\mathbf{x}_N))^T \in \mathbb{R}^N$
- If  $M < N$ , vectors  $\varphi_0, \dots, \varphi_{M-1}$  define a subspace  $\mathcal{S}$  of dimension (at most)  $M$
- $\mathbf{y} = (y(\mathbf{x}_1, \mathbf{w}), \dots, y(\mathbf{x}_N, \mathbf{w}))^T$  is a vector in  $\mathbb{R}^N$ : it can be represented as linear combination  $\mathbf{y} = \sum_{i=0}^{M-1} w_i \phi(\mathbf{x}_i)$ . Hence, it belongs to  $\mathcal{S}$
- Given  $\mathbf{t} \in \mathbb{R}^N$ ,  $\mathbf{y} \in \mathbb{R}^N$  is the vector in subspace  $\mathcal{S}$  at minimal squared distance from  $\mathbf{t}$
- Given  $\mathbf{t} \in \mathbb{R}^N$  and vectors  $\phi_0, \dots, \phi_{M-1}$ ,  $\mathbf{w}_{ML}$  is such that  $\mathbf{y}$  is the vector on  $\mathcal{S}$  nearest to  $\mathbf{t}$



- The minimum of  $E_D(\mathbf{w})$  may be computed numerically, by means of **gradient descent** methods
- Initial assignment  $\mathbf{w}^{(0)} = (w_0^{(0)}, w_1^{(0)}, \dots, w_D^{(0)})$ , with a corresponding error value

$$E_D(\mathbf{w}^{(0)}) = \frac{1}{2} \sum_{i=1}^N \left( t_i - (\mathbf{w}^{(0)})^T \phi(\mathbf{x}_i) \right)^2$$

- Iteratively, the current value  $\mathbf{w}^{(i-1)}$  is modified in the direction of **steepest descent** of  $E_D(\mathbf{w})$
- At step  $i$ ,  $w_j^{(i-1)}$  is updated as follows:

$$w_j^{(i)} := w_j^{(i-1)} - \eta \left. \frac{\partial E_D(\mathbf{w})}{\partial w_j} \right|_{\mathbf{w}^{(i-1)}}$$



- In matrix notation:

$$\mathbf{w}^{(i)} := \mathbf{w}^{(i-1)} - \eta \frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}^{(i-1)}}$$

- By definition of  $E_D(\mathbf{w})$ :

$$\mathbf{w}^{(i)} := \mathbf{w}^{(i-1)} - \eta(t_i - \mathbf{w}^{(i-1)}\phi(\mathbf{x}_i))\phi(\mathbf{x}_i)$$

# Regularized least squares

- Regularization term in the cost function

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

$E_D(\mathbf{w})$  dependent from the dataset (and the parameters),  $E_W(\mathbf{w})$  dependent from the parameters alone.

- The **regularization coefficient** controls the relative importance of the two terms.
- Simple form

$$E_W(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} = \frac{1}{2} \sum_{i=0}^{M-1} w_i^2$$

- Sum-of squares cost function: **weight decay**

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N \{t_i - \mathbf{w}^T \phi(\mathbf{x}_i)\}^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} = \frac{1}{2} (\Phi \mathbf{w} - \mathbf{y})^T (\Phi \mathbf{w} - \mathbf{y}) + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$$

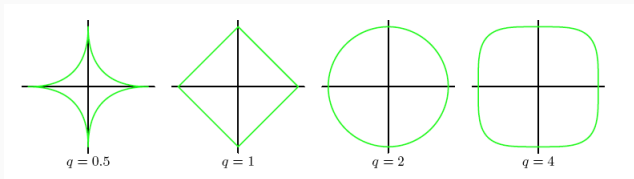
with solution

$$\mathbf{w} = (\lambda \mathbf{I} + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

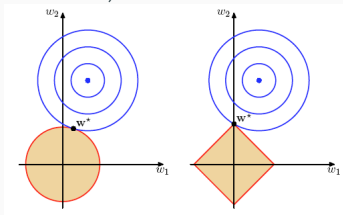
# Regularization

- A more general form

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N \{t_i - \mathbf{w}^T \phi(\mathbf{x}_i)\}^2 + \frac{\lambda}{2} \sum_{j=0}^{M-1} |w_j|^q$$



- The case  $q = 1$  is denoted as **lasso**: sparse models are favored (in blue, level curves of the cost function)



In considering the regression problem:

- we are looking for an estimator  $y(\mathbf{x})$  of the value  $y$  for any  $\mathbf{x}$
- we wish to minimize a given cost function  $\mathcal{C}(y, y(\mathbf{x}))$

A common choice for the cost function is the squared difference

$$C(y, y(\mathbf{x})) = (y(\mathbf{x}) - y)^2$$

## Bias-variance decomposition

We want to choose an estimator  $y(\mathbf{x})$  which minimizes the expected cost. It can be shown that the expected cost can be seen as

$$E_{\mathbf{x},y}[\mathcal{C}(y, y(\mathbf{x}))] = E_{\mathbf{x}}[(y(\mathbf{x}) - E_y[y|\mathbf{x}])^2] + E_{\mathbf{x}}[\sigma_{y|\mathbf{x}}^2]$$

where

- $E_{\mathbf{x}}[(y(\mathbf{x}) - E_y[y|\mathbf{x}])^2]$  is the contribution to the cost deriving by the chosen estimator  $y(\mathbf{x})$ : it results 0 if the (unknown) conditional mean  $E_y[y|\mathbf{x}]$ , a.k.a. the **regression function**, is chosen as estimator
- the second term is the expectation wrt  $\mathbf{x}$  of the conditional variance

$$\sigma_{y|\mathbf{x}}^2 = \int_y (y - E_y[y|\mathbf{x}])^2 p(y|\mathbf{x}) dy$$

it is a measure of the variability of  $y$  when  $\mathbf{x}$  is known and it is only related to data, not to the estimator

- The regression function  $E_y[y|\mathbf{x}]$  can be (approximately) derived from training data.
- Assume a collection of training set is available, each of them of same size  $n$  and independently sampled from distribution  $p(\mathbf{x}, y)$ . When one such set  $\mathcal{T}$  is submitted to the learning algorithm applied, an estimator function  $y(\mathbf{x}, \mathcal{T})$  is obtained

# Measuring the quality of the learning algorithm

Observe that

- $(y(\mathbf{x}, \mathcal{T}) - E_y[y|\mathbf{x}])^2$  is the cost of the estimator learned from  $\mathcal{T}$  when applied on point  $\mathbf{x}$
- $E_{\mathcal{T}}[(y(\mathbf{x}, \mathcal{T}) - E_y[y|\mathbf{x}])^2]$  is the (expected) cost of the learned estimator (wrt the available training set) when applied to  $\mathbf{x}$

The overall quality of the learning algorithm can be modeled in terms of expectation (on the input space) of the expected cost (on the training set used for learning)

$$E_{\mathbf{x}}[E_{\mathcal{T}}[(y(\mathbf{x}) - E_y[y|\mathbf{x}])^2]]$$

This measure should be minimized.

It can be easily shown that the measure can indeed be decomposed as

$$\begin{aligned} E_{\mathbf{x}}[E_{\mathcal{T}}[(y(\mathbf{x}) - E_y[y|\mathbf{x}])^2]] = \\ E_{\mathbf{x}}[E_{\mathcal{T}}[(y(\mathbf{x}, \mathcal{T}) - E_{\mathcal{T}}[y(\mathbf{x}, \mathcal{T})])^2]] + E_{\mathbf{x}}[(E_{\mathcal{T}}[y(\mathbf{x}, \mathcal{T})] - E_y[y|\mathbf{x}])^2] \end{aligned}$$



The term

$$E_{\mathbf{x}}[(E_{\mathcal{T}}[y(\mathbf{x}, \mathcal{T})] - E_y[y|\mathbf{x}])^2]$$

is denoted as (bias)<sup>2</sup>

- expectation over all  $\mathbf{x}$ , of the squared difference between the expected (wrt training set) value returned by the learned estimator and the value returned by the regression function
- measure of the correspondence between the derived estimator and the regression function

The term

$$E_{\mathbf{x}}[E_{\mathcal{T}}[(y(\mathbf{x}, \mathcal{T}) - E_{\mathcal{T}}[y(\mathbf{x}, \mathcal{T})])^2]]$$

is said **variance**

- expectation on the input space of the variance of the value returned by a learned estimator wrt the training set available
- measure of the variability of the behavior of different estimator functions learned by the algorithm when different training sets are used

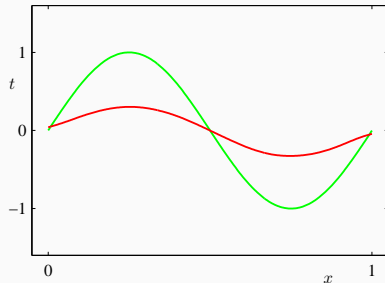
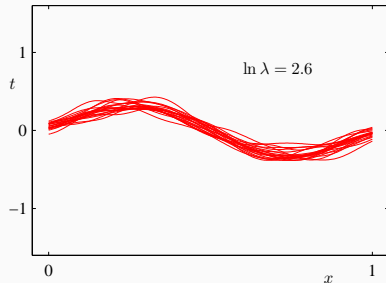
Aim: minimize the expected cost, that is the sum **(bias)<sup>2</sup>+variance**.

Unfortunately, there exists a tradeoff between them.

- Consider the case of function  $y = \sin 2\pi x$  and assume  $L = 100$  training sets  $\mathcal{T}_1, \dots, \mathcal{T}_L$  are available, each of size  $n = 25$ .
- Given  $M = 24$  gaussian basis functions  $\phi_1(x), \dots, \phi_M(x)$ , from each training set  $\mathcal{T}_i$  a prediction function  $y_i(x)$  is derived by minimizing the regularized cost function

$$E_D(\mathbf{w}) = \frac{1}{2}(\Phi\mathbf{w} - \mathbf{t})^T(\Phi\mathbf{w} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

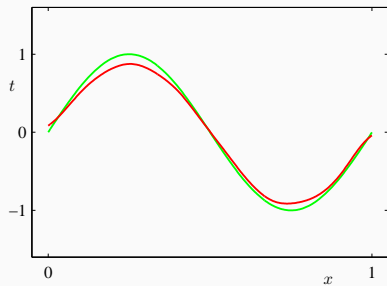
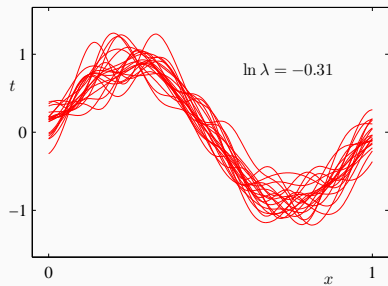
## An example



Left, a possible plot of prediction functions  $y_i(\mathbf{x})$  ( $i = 1, \dots, 100$ ), as derived, respectively, by training sets  $\mathcal{T}_i$ ,  $i = 1, \dots, 100$  setting  $\ln \lambda = 2.6$ . Right, their expectation, with the unknown function  $y = \sin 2\pi x$ .

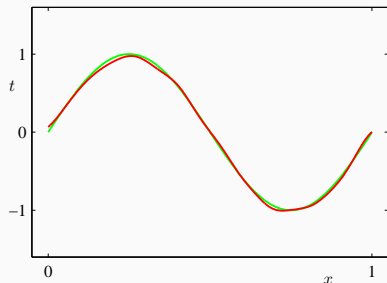
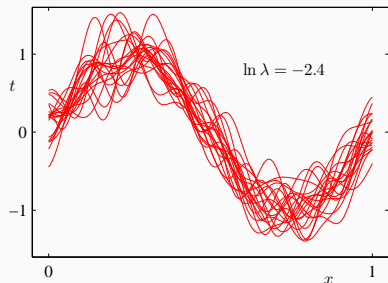
The prediction functions  $y_i(\mathbf{x})$  do not differ much between them (small variance), but their expectation is a bad approximation of the unknown function (large bias).

## An example



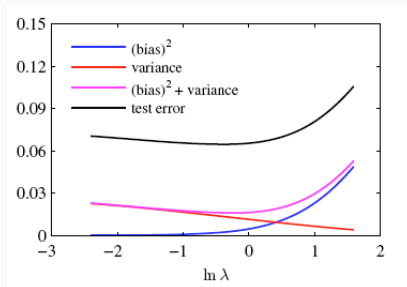
Plot of the prediction functions obtained with  $\ln \lambda = -0.31$ .

## An example



Plot of the prediction functions obtained with  $\ln \lambda = -2.4$ . As  $\lambda$  decreases, the variance increases (prediction functions  $y_i(\mathbf{x})$  are more different each other), while bias decreases (their expectation is a better approximation of  $y = \sin 2\pi x$ ).

## An example



- Plot of  $(\text{bias})^2$ , variance and their sum as functions of  $\lambda$ : as  $\lambda$  increases, bias increases and variance decreases. Their sum has a minimum in correspondence to the optimal value of  $\lambda$ .
- The term  $E_{\mathbf{x}}[\sigma_{y|\mathbf{x}}^2]$  shows an inherent limit to the approximability of  $y = \sin 2\pi x$ .

- Applying maximum likelihood to determine the values of model parameters is prone to overfitting: need of a regularization term  $\mathcal{E}(\mathbf{w})$ .
- In order control model complexity, a bayesian approach assumes a prior distribution of parameter values.



## Prior distribution

Posterior proportional to prior times likelihood: likelihood is gaussian (gaussian noise).

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

Conjugate of gaussian is gaussian: choosing a gaussian prior distribution of  $\mathbf{w}$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)$$

results into a gaussian posterior distribution

$$p(\mathbf{w} | \mathbf{t}, \Phi) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N) \propto p(\mathbf{t}, \Phi | \mathbf{w}) p(\mathbf{w})$$

where

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t}) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \Phi^T \Phi\end{aligned}$$

A common approach: zero-mean isotropic gaussian prior distribution of  $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{M-1} \left( \frac{\alpha}{2\pi} \right)^{1/2} e^{-\frac{\alpha}{2} w_i^2}$$

- Parameters in  $\mathbf{w}$  are assumed independent e identically distributed, according to a gaussian with mean  $\mathbf{0}$ , uniform variance  $\sigma^2 = \alpha^{-1}$  and null covariance.
- Prior distribution defined with a **hyper-parameter**  $\alpha$ , inversely proportional to the variance.

Given the likelihood

$$p(\mathbf{t}|\Phi, \mathbf{w}, \beta) = \prod_{i=1}^n e^{-\frac{\beta}{2}(t_i - \mathbf{w}^T \phi(x_i))^2}$$

the posterior distribution for  $\mathbf{w}$  derives from Bayes' rule

$$p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \sigma) = \frac{p(\mathbf{t}|\Phi, \mathbf{w}, \sigma)p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\Phi, \alpha, \sigma)} \propto p(\mathbf{t}|\Phi, \mathbf{w}, \sigma)p(\mathbf{w}|\alpha)$$

## In this case

It is possible to show that, assuming

$$\begin{aligned}p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \\p(\mathbf{t}|\mathbf{w}, \Phi) &= \mathcal{N}(\mathbf{t}|\mathbf{w}^T\Phi, \beta^{-1}\mathbf{I})\end{aligned}$$

the posterior distribution is itself a gaussian

$$p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \sigma) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

with

$$\begin{aligned}\mathbf{S}_N &= (\alpha\mathbf{I} + \beta\Phi^T\Phi)^{-1} \\ \mathbf{m}_N &= \beta\mathbf{S}_N\Phi^T\mathbf{t}\end{aligned}$$

Note that if  $\alpha \rightarrow 0$  the prior tends to have infinite variance, and we have minimum information on  $\mathbf{w}$  before the training set is consider. In this case,

$$\mathbf{m}_N \rightarrow (\Phi^T\beta\mathbf{I}\Phi)^{-1}(\Phi^T\beta\mathbf{I}\mathbf{t}) = (\Phi^T\Phi)^{-1}(\Phi^T\mathbf{t})$$

that is  $\mathbf{w}_{ML}$ , the ML estimation of  $\mathbf{w}$ .

- Given the posterior distribution  $p(\mathbf{w}|\Phi, \mathbf{t}, \alpha, \beta)$ , we may derive the value of  $\mathbf{w}_{MAP}$  which makes it maximum (the **mode** of the distribution)
- This is equivalent to maximizing its logarithm

$$\log p(\mathbf{w}|\Phi, \mathbf{t}, \alpha, \beta) = \log p(\mathbf{t}|\mathbf{w}, \Phi, \beta) + \log p(\mathbf{w}|\alpha) - \log p(\mathbf{t}|\Phi, \beta)$$

and, since  $p(\mathbf{t}|\Phi, \beta)$  is a constant wrt  $\mathbf{w}$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}} \log p(\mathbf{w}|\Phi, \mathbf{t}, \alpha, \beta) = \underset{\mathbf{w}}{\operatorname{argmax}} (\log p(\mathbf{t}|\mathbf{w}, \Phi, \beta) + \log p(\mathbf{w}|\alpha))$$

that is,

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmin}} (-\log p(\mathbf{t}|\Phi, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha))$$

By considering the assumptions on prior and likelihood,

$$\begin{aligned} w_{MAP} &= \underset{\mathbf{w}}{\operatorname{argmin}} \left( \frac{\beta}{2} \sum_{i=1}^n (t_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2 + \frac{\alpha}{2} \sum_{i=0}^{M-1} w_i^2 + \text{constants} \right) \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \left( \sum_{i=1}^n (t_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2 + \frac{\alpha}{\beta} \sum_{i=0}^{M-1} w_i^2 \right) \end{aligned}$$

this is equivalent to considering a cost function

$$E_{MAP}(\mathbf{w}) = \sum_{i=1}^n (y_i - \mathbf{w}^T \phi(x_i))^2 + \frac{\alpha}{\beta} \mathbf{w}^T \mathbf{w}$$

that is to a regularized min square function with  $\lambda = \frac{\alpha}{\beta}$

- The bayesian approach can be applied when the training set is acquired incrementally (sequential learning).
- Fundamental property: if prior and likelihood are gaussians, the posterior is gaussian too.
- Let  $T_1, T_2$  be two independent training sets, acquired one after the other. Then

$$p(\mathbf{w}|T_1, T_2) \propto p(T_1, T_2|\mathbf{w})p(\mathbf{w}) = p(T_2|\mathbf{w})p(T_1|\mathbf{w})p(\mathbf{w}) \propto p(T_2|\mathbf{w})p(\mathbf{w}|T_1)p(\mathbf{w})$$

- The posterior after observing  $T_1$  can be used as a prior for the next training set acquired.
- In general, for a sequence  $T_1, \dots, T_n$  of training sets,

$$p(\mathbf{w}|T_1, \dots, T_n) \propto p(T_n|\mathbf{w})p(\mathbf{w}|T_1, \dots, T_{n-1})$$

$$p(\mathbf{w}|T_1, \dots, T_{n-1}) \propto p(T_{n-1}|\mathbf{w})p(\mathbf{w}|T_1, \dots, T_{n-2})$$

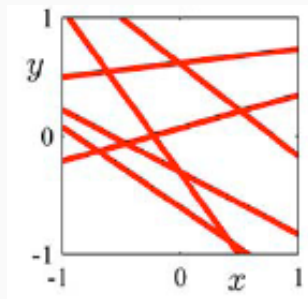
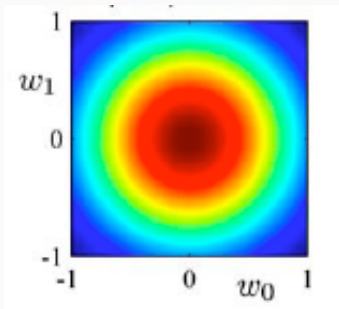
...

$$p(\mathbf{w}|T_1) \propto p(T_1|\mathbf{w})p(\mathbf{w})$$



## Example

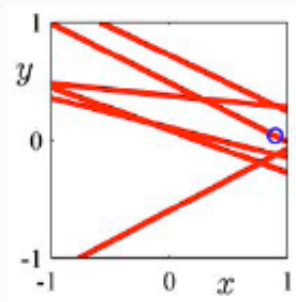
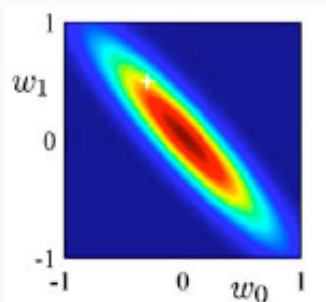
- Input variable  $x$ , target variable  $t$ , linear regression  
 $y(x, w_0, w_1) = w_0 + w_1x$ .
- Dataset generated by applying function  $y = a_0 + a_1x$  (with  $a_0 = -0.3$ ,  $a_1 = 0.5$ ) to values uniformly sampled in  $[-1, 1]$ , with added gaussian noise ( $\mu = 0$ ,  $\sigma = 0.2$ ).
- Assume the prior distribution  $p(w_0, w_1)$  is a bivariate gaussian with  $\mu = \mathbf{0}$  and  $\Sigma = \sigma^2 \mathbf{I} = 0.04 \mathbf{I}$



Left, prior distribution of  $w_0, w_1$ ; right, 6 lines sampled from the distribution.

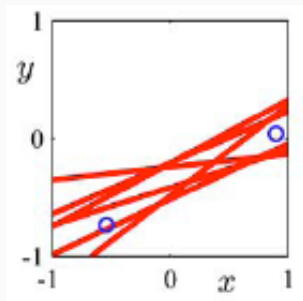
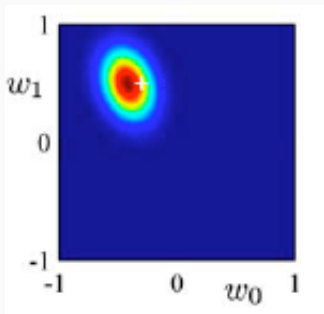
## Example

After observing item  $(x_1, y_1)$  (circle in right figure).



Left, posterior distribution  $p(w_0, w_1 | x_1, y_1)$ ; right, 6 lines sampled from the distribution.

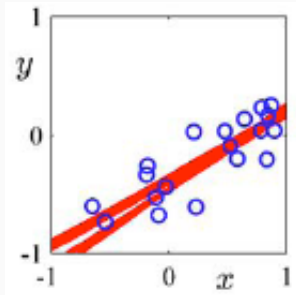
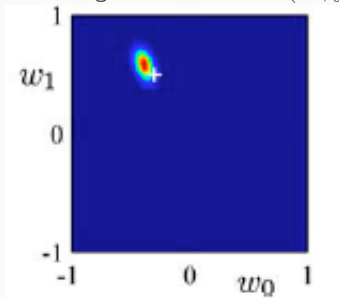
After observing items  $(x_1, y_1), (x_2, y_2)$  (circles in right figure).



Left, posterior distribution  $p(w_0, w_1 | x_1, y_1, x_2, y_2)$ ; right, 6 lines sampled from the distribution.

## Example

After observing a set of  $n$  items  $(x_1, y_1), \dots, (x_n, y_n)$  (circles in right figure).



Left, posterior distribution  $p(w_0, w_1 | x_i, y_i, i = 1, \dots, n)$ ; right, 6 lines sampled from the distribution.

- As the number of observed items increases, the distribution of parameters  $w_0, w_1$  tends to concentrate (variance decreases to 0) around a mean point  $a_0, a_1$ .
- As a consequence, sampled lines are concentrated around  $y = a_0 + a_1x$ .

## Classical

- A value  $\mathbf{w}_{LS}$  for  $\mathbf{w}$  is learned through a point estimate, performed by minimizing a quadratic cost function, or equivalently by maximizing likelihood (ML) under the hypothesis of gaussian noise; regularization can be applied to modify the cost function to limit overfitting
- Given any  $\mathbf{x}$ , the obtained value  $\mathbf{w}_{LS}$  is used to predict the corresponding  $t$  as  $y = \bar{\mathbf{x}}^T \mathbf{w}_{LS}$ , where  $\bar{\mathbf{x}}^T = (1, \mathbf{x})^T$ , or, in general, as  $y = \phi(\mathbf{x})^T \mathbf{w}_{LS}$

## Bayesian point estimation

- The posterior distribution  $p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta)$  is derived and a point estimate is performed from it, computing the mode  $\mathbf{w}_{MAP}$  of the distribution (MAP)
- Equivalent to the classical approach, as  $\mathbf{w}_{MAP}$  corresponds to  $\mathbf{w}_{LS}$  if  $\lambda = \frac{\alpha}{\beta}$
- The prediction, for a value  $\mathbf{x}$ , is a gaussian distribution  $p(y|\phi(\mathbf{x})^T \mathbf{w}_{MAP}, \beta)$  for  $y$ , with mean  $\phi(\mathbf{x})^T \mathbf{w}_{MAP}$  and variance  $\beta^{-1}$
- The distribution is not derived directly from the posterior  $p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta)$ : it is built, instead, as a gaussian with mean depending from the expectation of the posterior, and variance given by the assumed noise.

## Fully bayesian

- The real interest is not in estimating  $\mathbf{w}$  or its distribution  $p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta)$ , but in deriving the predictive distribution  $p(y|\mathbf{x})$ . This can be done through expectation of the probability  $p(y|\mathbf{x}, \mathbf{w}, \beta)$  predicted by a model instance wrt model instance distribution  $p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta)$ , that is

$$p(y|\mathbf{x}, \mathbf{t}, \Phi, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta) d\mathbf{w}$$

- $p(y|\mathbf{x}, \mathbf{w}, \beta)$  is assumed gaussian, and  $p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta)$  is gaussian by the assumption that the likelihood  $p(\mathbf{t}|\mathbf{w}, \Phi, \beta)$  and the prior  $p(\mathbf{w}|\alpha)$  are gaussian themselves and by their being conjugate

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|\mathbf{w}^T \phi(\mathbf{x}), \beta)$$

$$p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\beta \mathbf{S}_N \Phi^T \mathbf{t}, \mathbf{S}_N)$$

where  $\mathbf{S}_N = (\alpha \mathbf{I} + \beta \Phi^T \Phi)^{-1}$



# Approaches to prediction in linear regression

## Fully bayesian

Under such hypothesis,  $p(y|\mathbf{x})$  is gaussian

$$p(y|\mathbf{x}, \mathbf{y}, \Phi, \alpha, \beta) = \mathcal{N}(y|m(\mathbf{x}), \sigma^2(\mathbf{x}))$$

with mean

$$m(\mathbf{x}) = \beta \phi(\mathbf{x})^T \mathbf{S}_N \Phi^T \mathbf{t}$$

and variance

$$\sigma^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x})$$

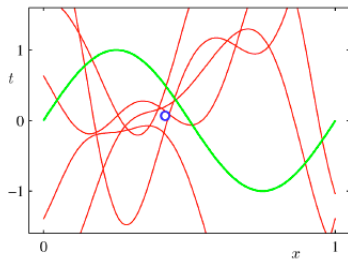
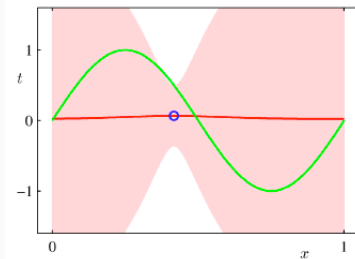
- $\frac{1}{\beta}$  is a measure of the uncertainty intrinsic to observed data (noise)
- $\phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x})$  is the uncertainty wrt the values derived for the parameters  $\mathbf{w}$
- as the noise distribution and the distribution of  $\mathbf{w}$  are independent gaussians, their variances add
- $\phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}) \rightarrow 0$  as  $n \rightarrow \infty$ , and the only uncertainty remaining is the one intrinsic into data observation

## Example

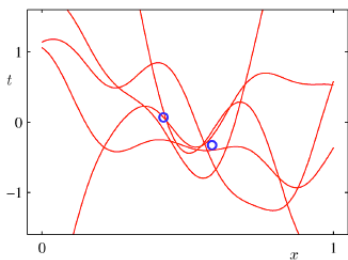
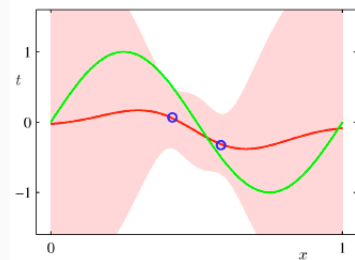
- predictive distribution for  $y = \sin 2\pi x$ , applying a model with 9 gaussian base functions and training sets of 1, 2, 4, 25 items, respectively
- left: items in training sets (sampled uniformly, with added gaussian noise); expectation of the predictive distribution (red), as function of  $x$ ; variance of such distribution (pink shade within 1 standard deviation from mean), as a function of  $x$
- right: items in training sets, 5 possible curves approximating  $y = \sin 2\pi x$ , derived through sampling from the posterior distribution  $p(\mathbf{w}|\mathbf{t}, \Phi, \alpha, \beta)$

# Example

$n = 1$

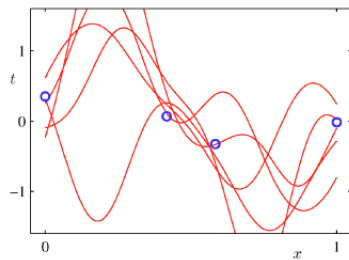
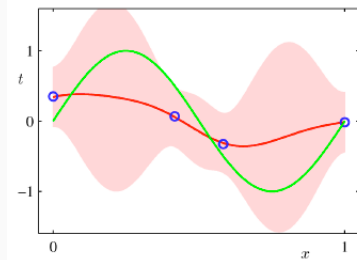


$n = 2$

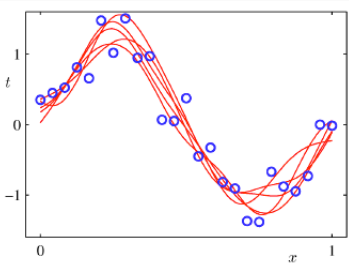
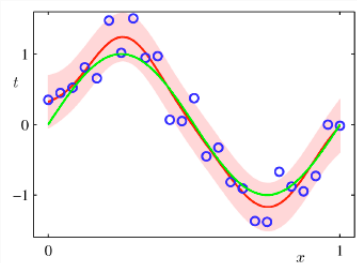


# Example

$n = 4$



$n = 25$



- The expectation of the predictive distribution can be written also as

$$y(\mathbf{x}) = \beta \phi(\mathbf{x})^T \mathbf{S}_N \Phi^T \mathbf{t} = \sum_{i=1}^n \beta \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}_i) t_i$$

- The prediction can then be seen as a linear combination of the target values  $t_i$  of items in the training set, with weights dependent from the item values  $\mathbf{x}_i$  (and from  $\mathbf{x}$ )

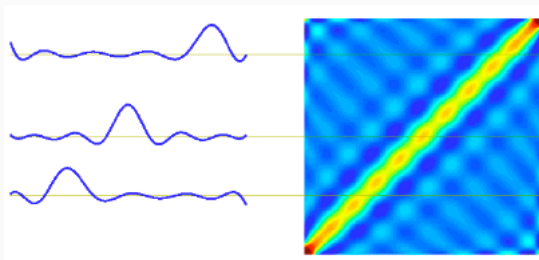
$$y(\mathbf{x}) = \sum_{i=1}^n \kappa(\mathbf{x}, \mathbf{x}_i) t_i$$

The weight function  $\kappa(\mathbf{x}, \mathbf{x}') = \beta \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}')$  is said *equivalent kernel* or **linear smoother**

## Equivalent kernel

Right: plot on the plane  $(x, x_i)$  of a sample equivalent kernel, in the case of gaussian basis functions.

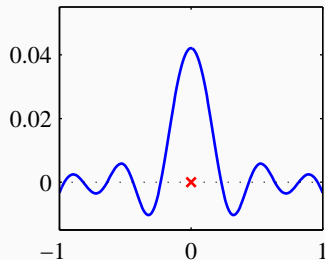
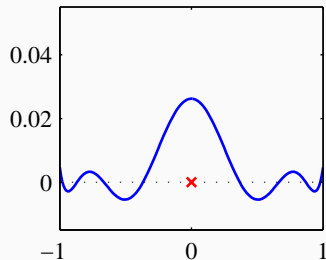
Left: plot as a function of  $x_i$  for three different values of  $x$



In deriving  $y$ , the equivalent kernel tends to assign greater relevance to the target values  $t_i$  corresponding to items  $x_i$  near to  $x$ .

## Equivalent kernel

The same localization property holds also for different base functions.



Left,  $\kappa(0, x')$  in the case of polynomial basis functions.

Right,  $\kappa(0, x')$  in the case of Gaussian basis functions.

Some properties:

- It is possible to prove that  $\sum_{i=1}^n \kappa(\mathbf{x}, \mathbf{x}_i) = 1$  for any  $\mathbf{x}$
- The covariance between  $y(\mathbf{x})$  and  $y(\mathbf{x}')$  is given by

$$\text{cov}(\mathbf{x}, \mathbf{x}') = \text{cov}(\boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}, \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}')) = \boldsymbol{\Phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}') = \frac{1}{\beta} \kappa(\mathbf{x}, \mathbf{x}')$$

predicted values are highly correlated at nearby points.

- Instead of introducing base functions which results into a kernel, we may define a localized kernel directly and use it to make predictions (gaussian processes)
- The equivalent kernel can be expressed as inner product  $\kappa(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^T \boldsymbol{\psi}(\mathbf{x}')$  of a suitable set of functions

$$\boldsymbol{\psi}(\mathbf{x}) = \beta^{1/2} \mathbf{S}_N^{1/2} \boldsymbol{\phi}(\mathbf{x})$$



- First approach: define a set of base functions
  - used to derive  $\mathbf{w}$
  - or (by means of the resulting equivalent kernel) to directly computing  $y(\mathbf{x})$  as a linear combination of training set items
- New approach: a suitable kernel is defined and used to compute  $y(\mathbf{x})$