IDA 2016.4/1

Linear regression

Training data $\{\underline{x}_i, d_i\}_{i=1}^{P}$

$y_i = \underline{w}^T \cdot \underline{x}_i$

for all training data $\underline{y} = \underline{\underline{X}} \cdot \underline{w}$ model's output

the goal $\underline{d} = \underline{\underline{X}} \cdot \underline{w}$

ANALYTIC solutions

simple solution

$\underline{w}^* = \underline{\underline{X}}^{-1} \cdot \underline{d}$ inverse

$\underline{w}^* = (\underline{\underline{X}}^T \cdot \underline{\underline{X}})^{-1} \cdot \underline{\underline{X}}^T \cdot \underline{d}$ pseudo inverse

equivalent with the LS solution

$$\underline{w}^*_{LS} = \arg\min_{\underline{w}} (\underline{d} - \underline{\underline{X}} \cdot \underline{w})^T (\underline{d} - \underline{\underline{X}} \cdot \underline{w})$$

if $\underline{\underline{X}}$ (or $\underline{\underline{X}}^T \cdot \underline{\underline{X}}$) is singular regularization

$\underline{\underline{X}}^{-1} \rightarrow (\underline{\underline{X}} + \lambda \underline{\underline{I}})^{-1}$ λ regularization coefficient

$(\underline{\underline{X}}^T \cdot \underline{\underline{X}})^{-1} \rightarrow (\underline{\underline{X}}^T \cdot \underline{\underline{X}} + \lambda \underline{\underline{I}})^{-1}$

Maximum likelihood solution

$d(x) = g(x) + n$ observation noise

Gaussian $\mathcal{N}(0, \sigma_n^2)$; $\mathcal{N}(0, \beta^{-1})$

conditional density function

$$p(d \mid x, \underline{w}) = \frac{1}{\sqrt{2\pi \sigma_n^2}} \exp\left[-\frac{1}{2} (d - \underline{w}^T \cdot \underline{x})^2 \frac{1}{\sigma_n^2} \right]$$

for the all training data (P data points)

$$p(\underline{d} \mid \underline{\underline{X}}, \underline{w}) = \frac{1}{2\pi^{P/2} |\Sigma_{nn}|^{1/2}} \cdot \exp\left[-\frac{1}{2} (\underline{d} - \underline{\underline{X}} \cdot \underline{w})^T \cdot \underline{\underline{\Sigma}}_{nn}^{-1} (\underline{d} - \underline{\underline{X}} \cdot \underline{w}) \right]$$

log likelihood function

$$\mathcal{L} = -\log\{p(\underline{d} \mid \underline{\underline{X}}, \underline{w})\} = \text{const} + \frac{1}{2}\,(\underline{d} - \underline{\underline{X}}\cdot\underline{w})^T \cdot \underline{\underline{\Sigma}}_{nn}^{-1}\,(\underline{d} - \underline{\underline{X}}\cdot\underline{w})$$

Maximum likelihood solution

$$\underline{w}^*_{ML} = \arg\min_{\underline{w}} \left[(\underline{d} - \underline{\underline{X}}\cdot\underline{w})^T (\underline{\underline{\Sigma}}_{nn}^{-1} \cdot (\underline{d} - \underline{\underline{X}}\cdot\underline{w}) \right]$$

$$\underline{w}^*_{ML} = (\underline{\underline{X}}^T \cdot \underline{\underline{\Sigma}}_{nn}^{-1} \underline{\underline{X}})^{-1} \cdot \underline{\underline{X}}^T \cdot \underline{\underline{\Sigma}}_{nn}^{-1} \cdot \underline{d}$$

for isotrope Gaussian noise (white noise) $\underline{\underline{\Sigma}}_{nn} = \sigma_n^2 \cdot \underline{\underline{I}}$

$$\underline{w}^*_{ML} = (\underline{\underline{X}}^T \cdot \underline{\underline{X}})^{-1} \cdot \underline{\underline{X}}^T \cdot \underline{d}$$

egzactly the same as $\underline{w}^*_{LS}$

for ML solution the noise variance can also be estimated

$$\sigma^2_{ML} = \frac{1}{\beta_{ML}} = \frac{1}{P}\sum_{i=1}^{P}(d_i - \underline{w}^T \underline{x}_i)^2 \quad ; \quad \sigma^2_{ML} = \arg\max_{\sigma^2} \mathcal{L}$$

Regularization :

open question : how to select λ reg. coeff

illustrative figures

MSE decomposition : noise variance + bias^2 + variance

noise variance

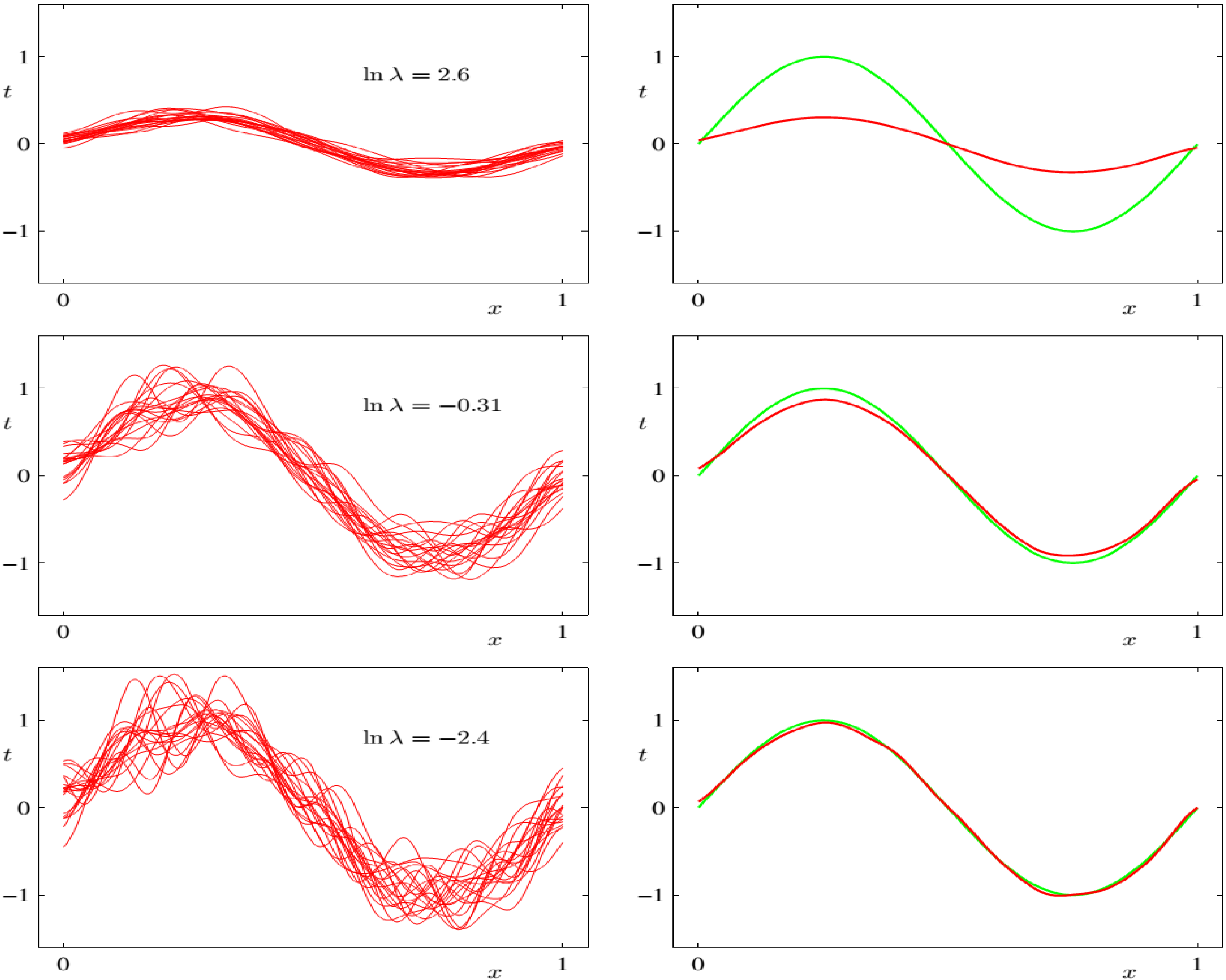
$$E\{[d(x) - g(x)]^2\}$$

bias

$$g(x) - E\{y(\underline{x}, \underline{w})\}$$

variance of the model's output

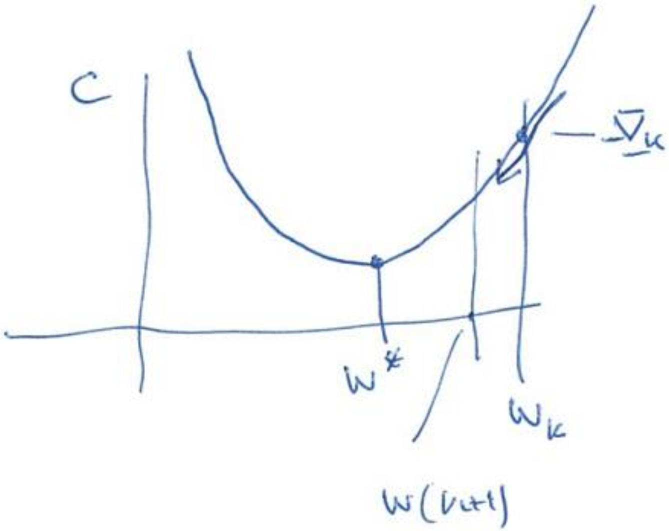
$$E\left[y(\underline{x}, w) - E(y(\underline{x}, \underline{w}))\right]^2$$



$\ln \lambda = 2.6$
$\ln \lambda = -0.31$
$\ln \lambda = -2.4$
t
x
0
1
-1

Iterative solution

$$\underline{w}(k+1) = \underline{w}(k) + \mu(-\underline{\nabla}_k) \qquad \underline{\nabla}_k = \frac{\partial C}{\partial \underline{w}} \quad \text{gradient}$$



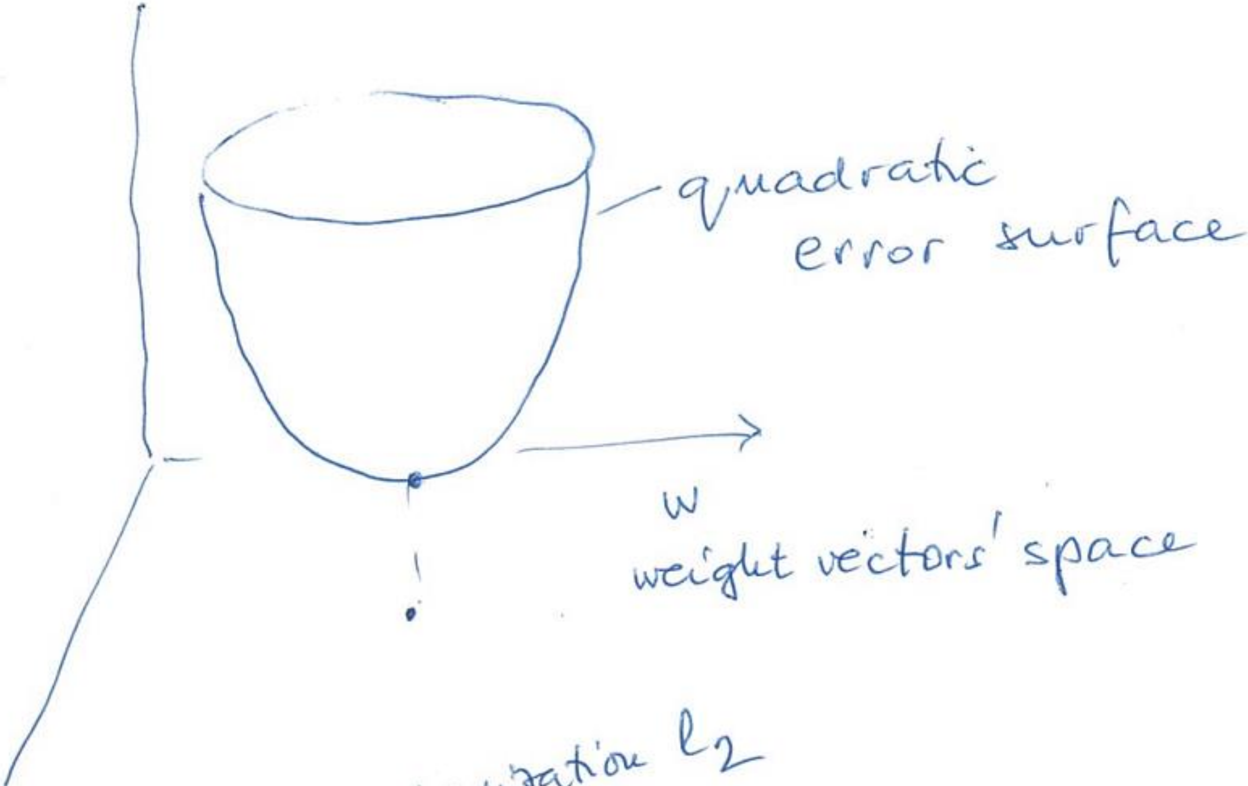
it is convergent if μ is properly selected

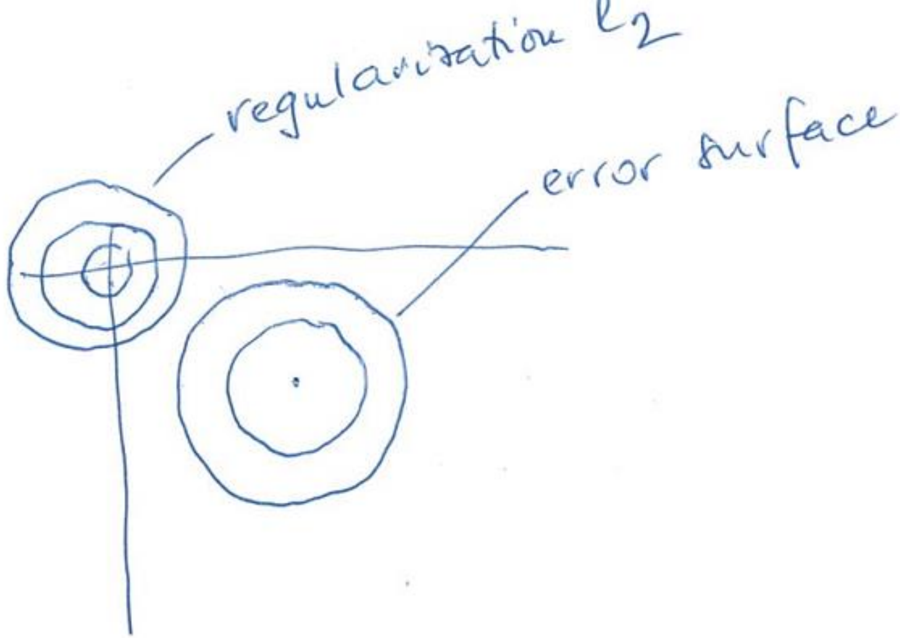
LMS algorithm

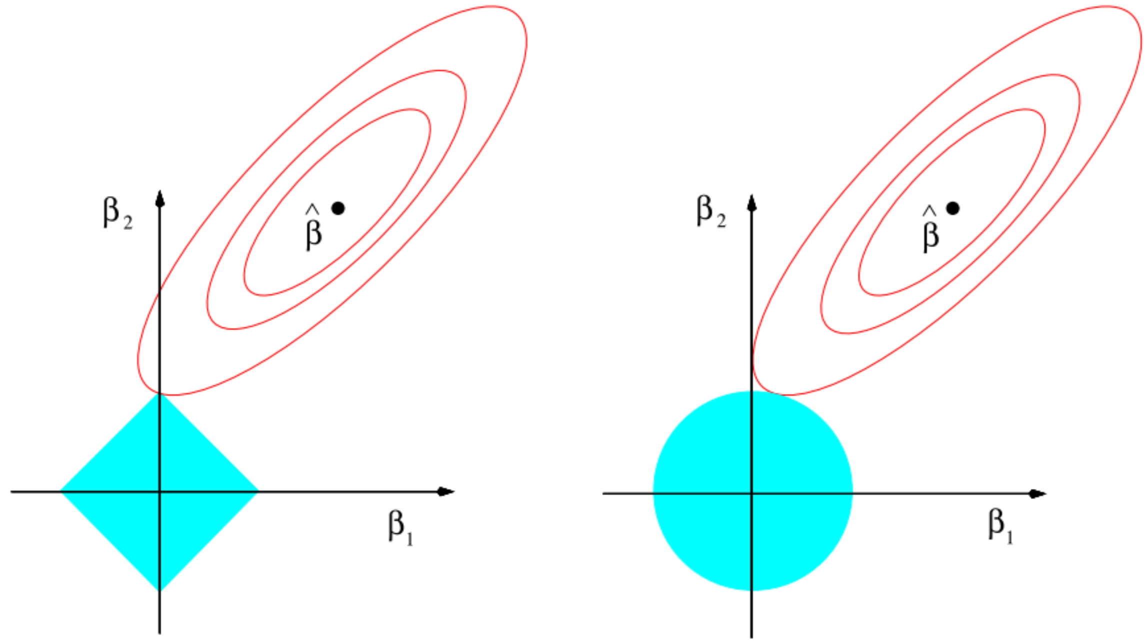
$$\underline{w}(k+1) = \underline{w}(k) + 2\mu\, \varepsilon(k) \cdot \underline{x}(k) \; ; \quad \varepsilon(k) = d(k) - y(k)$$

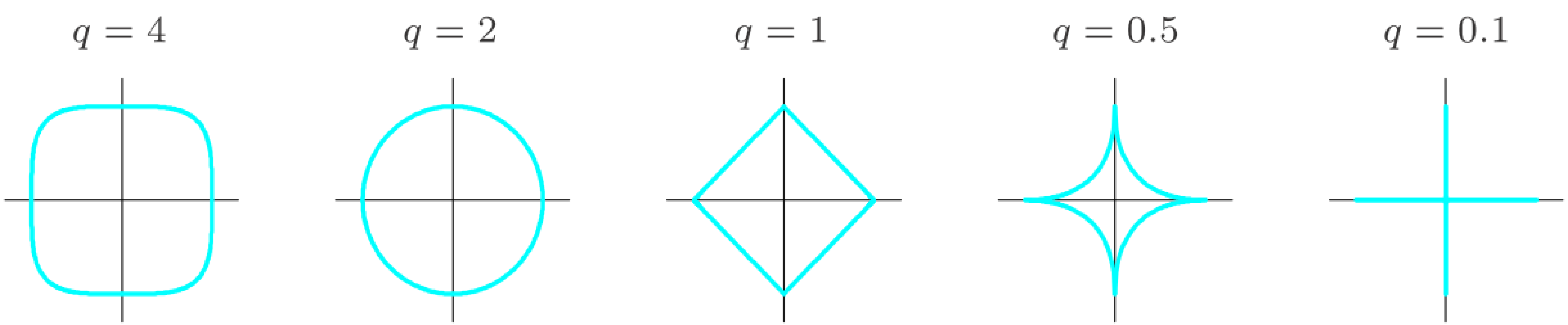
convergent if $\quad 0 < \mu < \dfrac{1}{\lambda_{max}}$; λ_{max} max. eigenvalue of $E[\underline{x} \cdot \underline{x}^T]$

covariance matrix


quadratic
error surface
w
weight vectors' space


regularization l_2
error surface





$$\frac{\lambda}{2} \sum_{j=1}^{M} |w_j|^q$$

IDA 2016. 4/4.

Bays linear regression

Regularized LS, and ML solution: how to determine λ?

General solution cross validation

This is not a direct way, it is a trial-and-error way.

The goal is to avoid overfitting: to find the proper model complexity

Bays linear regression try to avoid overfitting, to determine appropriate regularization coefficient using the training data.

Basic principle

modell parameter vector $\underline{w}$ is a random variable

its starting (prior) probability distribution is known

Most often $p(\underline{w})$ Gaussian with $\underline{m}_0$ mean and $\underline{\underline{\Sigma}}_{w_0}$ covariance matrix

<u>prior</u>

$$p(\underline{w}): N(\underline{m}_0, \underline{\underline{\Sigma}}_{w_0})$$

$$p(\underline{w}) = \frac{1}{(2\pi)^{N/2} |\underline{\underline{\Sigma}}_{w_0}|^{1/2}} \exp\left[-\frac{1}{2}(\underline{w}-\underline{m}_0)^T \underline{\underline{\Sigma}}_{w_0}^{-1} (\underline{w}-\underline{m}_0)\right]$$

<u>Bayes-rule</u> to determine the posterior

$$p(\underline{w}|\underline{d}) = \frac{p(\underline{d}|\underline{w}) \cdot p(\underline{w})}{p(\underline{d})} = \frac{p(\underline{d}|\underline{w}) \cdot p(\underline{w})}{\int p(\underline{d}|\underline{w}) \cdot p(\underline{w}) \, d\underline{w}}$$

here $p(\underline{d}|\underline{w})$ is the conditional density function of the observations (likelihood function)

For isotropic noise (white noise) and for such $\underline{w}$ prior

where $\underline{\underline{\Sigma}}_{w_0} = \frac{1}{\alpha} \underline{\underline{I}}$

$$\underline{m}_p = \underline{\underline{\Sigma}}_{w_p} \left[\alpha \cdot \underline{m}_0 + \beta \cdot \underline{\underline{X}}^T \cdot \underline{d} \right]$$

$$\underline{\underline{\Sigma}}_{w_p}^{-1} = \left[\underline{\underline{\Sigma}}_{w_0}^{-1} + \beta \cdot \underline{\underline{X}}^T \cdot \underline{\underline{X}} \right] = \alpha \cdot \underline{\underline{I}} + \beta \cdot \underline{\underline{X}}^T \cdot \underline{\underline{X}}$$

The posterior for the simple case

$$\ln p(\underline{w} \mid \underline{d}) = -\frac{\beta}{2} \sum_{i=1}^{P} \left(d_i - \underline{w}^T \cdot \underline{x}_i \right)^2 - \frac{\alpha}{2} \underline{w}^T \cdot \underline{w} + \text{const.}$$

This corresponds to a regularized LS(ML) solution but the regularization coefficient is determined

$$\lambda = \frac{\alpha}{\beta}$$

$$p(\underline{w}|\underline{d}) \propto p(\underline{d}|\underline{w}) \cdot p(\underline{w})$$

As both are Gaussians, the posterior will be Gaussian too.

$$p(\underline{w}|\underline{d}) : \mathcal{N}(\underline{m}_p, \underline{\underline{\Sigma}}_{\underline{w}_p})$$

To determine $\underline{m}_p$ and $\underline{\underline{\Sigma}}_{\underline{w}_p}$ use logarithm

$$\ln p(\underline{w}|\underline{d}) \propto \ln p(\underline{d}|\underline{w}) + \ln p(\underline{w})$$

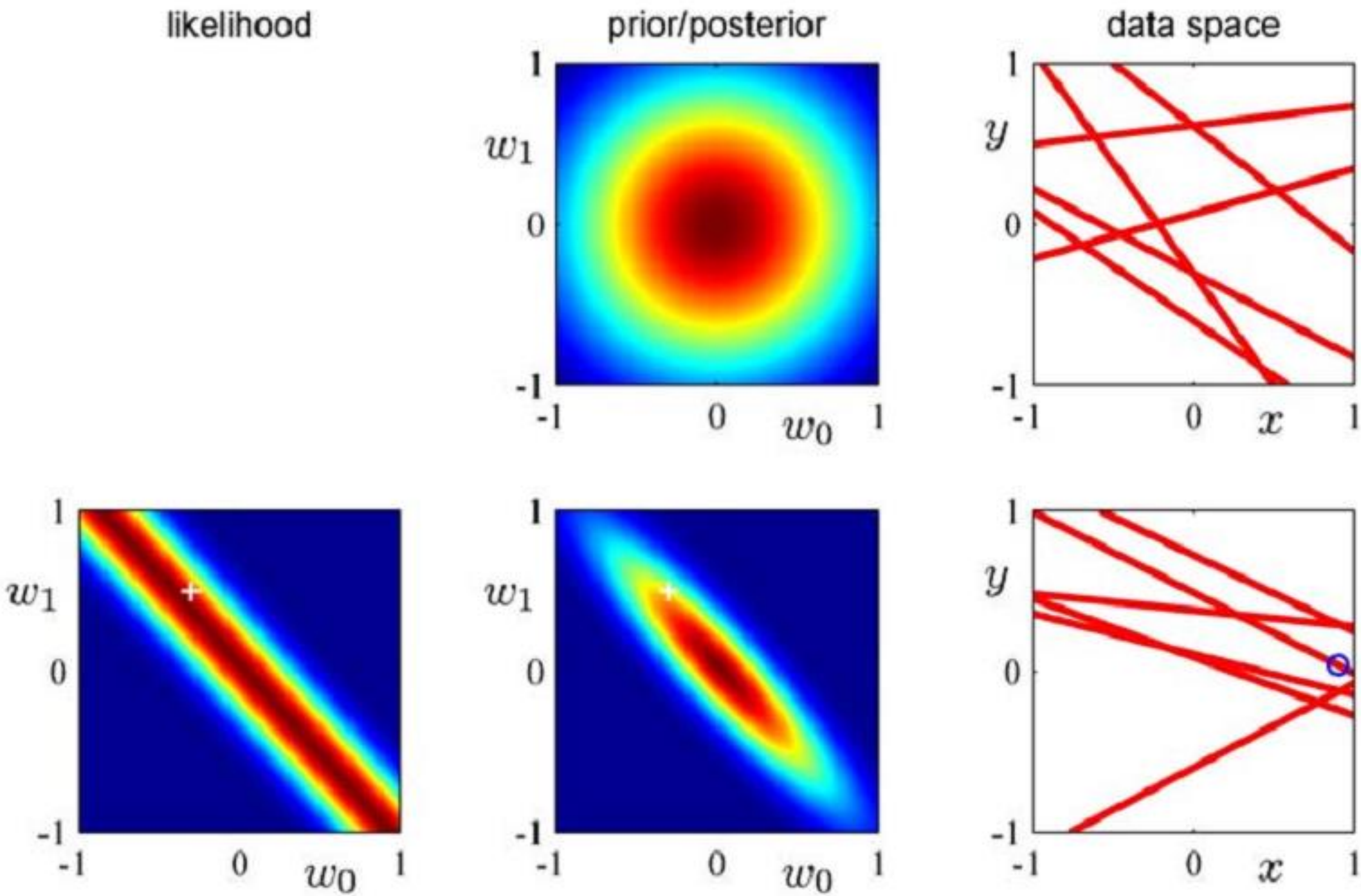
After some steps

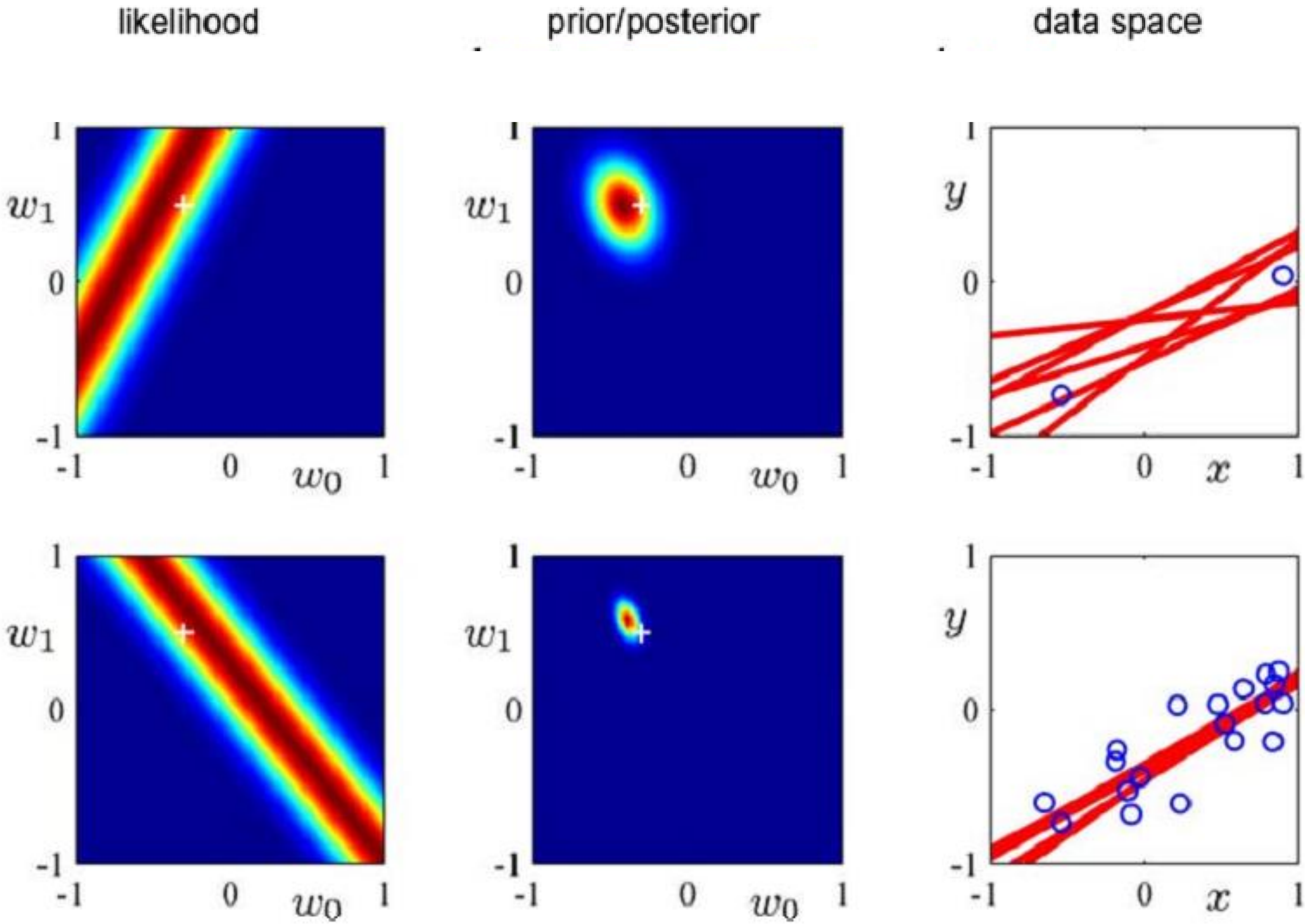
$$\underline{m}_p = \underline{\underline{\Sigma}}_{\underline{w}_p}\left[\underline{\underline{\Sigma}}_{w_0}^{-1} \cdot \underline{m}_0 + \underline{\underline{X}}^T \cdot \underline{\underline{\Sigma}}_{nn}^{-1} \cdot \underline{d}\right]$$

$$\underline{\underline{\Sigma}}_{w_p}^{-1} = \underline{\underline{\Sigma}}_{\underline{w}_0}^{-1} + \underline{\underline{X}}^T \cdot \underline{\underline{\Sigma}}_{nn}^{-1} \cdot \underline{\underline{X}}$$

; $\underline{\underline{\Sigma}}_{nn}$ noise covariance matrix

A simple practical example

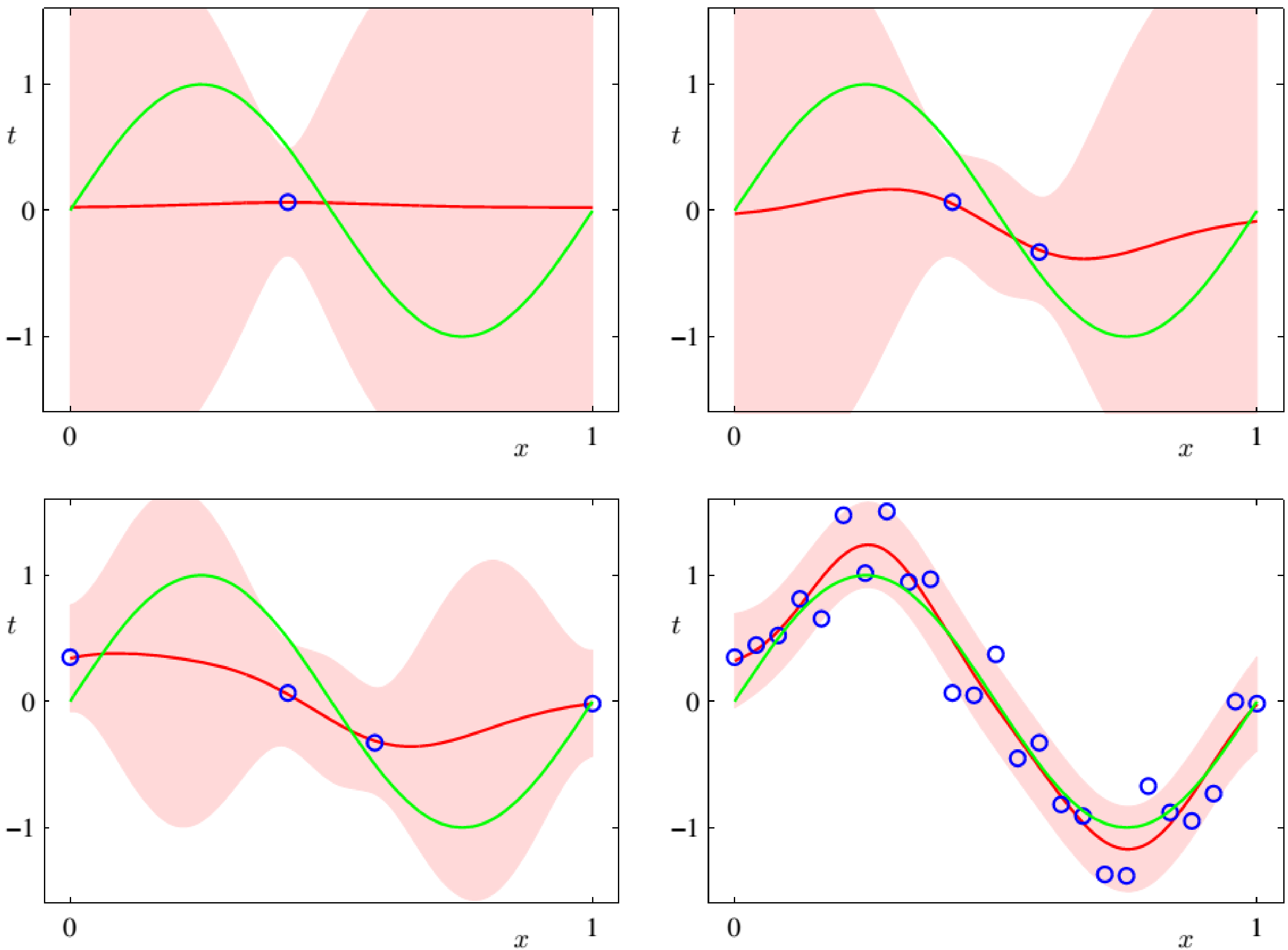



likelihood
prior/posterior
data space
w_1
w_0
y
x
-1
0
1

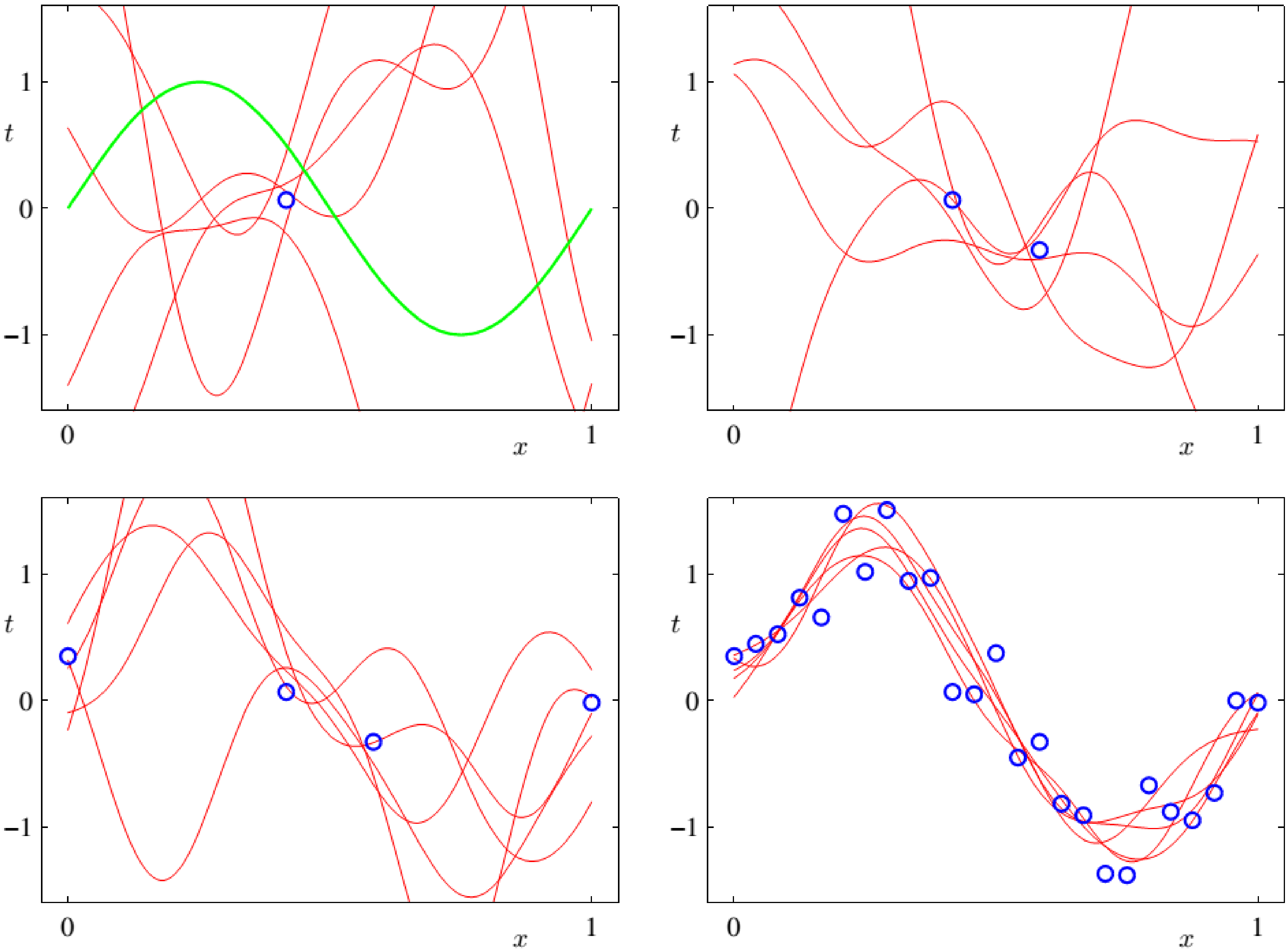
Further questions. ① predictive distribution

If $p(\underline{w}|\underline{d})$ posterior is know the predictive distribution of the model's output can be determined for a new input (and after using all training data)

$$p(d|\underline{d},\underline{X},\alpha,\beta,\underline{x}) = \int p(d|\underline{w},\beta) \cdot p(\underline{w}|\underline{d},\underline{X},\alpha,\beta)\, d\underline{w}$$



1
t
0
−1
0
x
1



t
x
0
1
−1

(2) estimation of α and β hyperparameters

it is assumed that α and β hyperparameters are random variables and their prior density functions are known.

In this case the marginal density function can be determined (at least in principle)

No analytic solution exists; approximate solution can be determined

$$p(d \mid \underline{d} \ldots) = \int p(d \mid \underline{w}, \beta) \cdot p(\underline{w} \mid \underline{d}, \alpha, \beta) \cdot p(\alpha, \beta \mid \underline{d})\, d\underline{w}\, d\alpha\, d\beta$$

$$\alpha = \frac{\gamma}{\underline{m}_p^T \cdot \underline{m}_p} \quad , \quad \text{but} \quad \gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i} \quad \text{where } \lambda_i$$

the ith eigenvalue of

$$\beta \cdot \underline{\underline{X}}^T \cdot \underline{\underline{X}}$$

$$\beta = \frac{1}{P - \gamma} \sum_{i=1}^{P} \left[d_i - \underline{m}_p^T \cdot \underline{x}_i \right]^2$$

Extensions

Nonlinear regression

- nonlinear, but linear-in-the parameters
- general nonlinear (nonlinear in the parameters)

- linear-in-the parameters

natural extension of linear regression

$\underline{x} \rightarrow \underline{\varphi}(\underline{x})$ nonlinear basis functions

$$y = \underline{w}^T \cdot \underline{x} \longrightarrow y = \underline{w}^T \cdot \underline{\varphi}(\underline{x}) = \sum_{i=1}^{M} w_i \cdot \varphi_i(\underline{x})$$

$$M \leftrightarrow N$$

for all data (for all P training data)

$$\underline{y} = \underline{\underline{\Phi}} \cdot \underline{w} \quad ; \quad \text{the goal is} \quad \underline{d} = \underline{\underline{\Phi}} \cdot \underline{w}$$

LS $\quad \underline{w}^* = \underline{\underline{\Phi}}^{-1} \cdot \underline{d} \quad ; \quad \underline{w}^* = (\underline{\underline{\Phi}}^T \cdot \underline{\underline{\Phi}})^{-1} \cdot \underline{\underline{\Phi}}^T \cdot \underline{d}$

regularization . . . .

ML solution . . .

Bayes solution . . .

<u>Questions</u>

how to select the φ_i basis functions

M (the number of basis functions)