Statistical Machine Learning
Bayesian Linear Regression

Luca Bortolussi

Department of Mathematics and Geosciences
University of Trieste

Office 238, third floor, H2bis
lbortolussi@units.it

Data Science and Scientific Computing

Outline

1. Linear regression
2. Bayesian linear regression
3. Dual representation and kernels

MAXIMUM LIKELIHOOD REGRESSION

- Observations $(\mathbf{x}_i, t_i)$, $i = 1, \ldots, N$
- $M + 1$ Generalised basis functions $\phi_j : \mathbb{R}^n \to \mathbb{R}$, with $\phi_0(\mathbf{x}) = 1$ (polynomials, Radial Basis Functions, sigmoids)
- Gaussian noise: $t = y(\mathbf{x}, \mathbf{w}) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \beta^{-1})$
- Likelihood is $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_i}), \beta^{-1})$
- Maximum likelihood solution computable in closed form
- Regularization by penalising large weights (Lasso and Ridge regression)

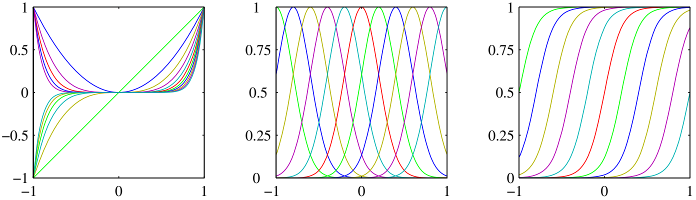
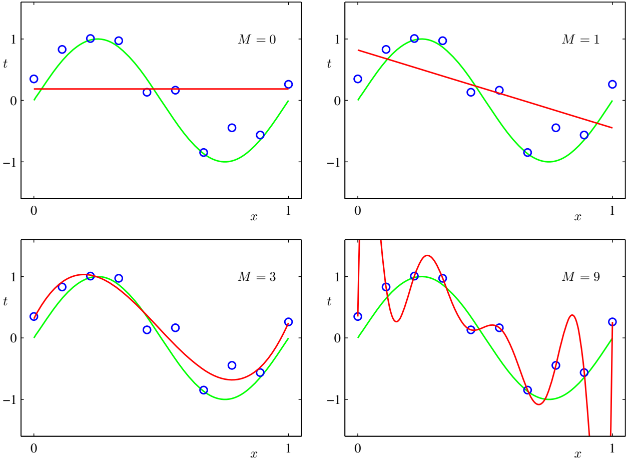


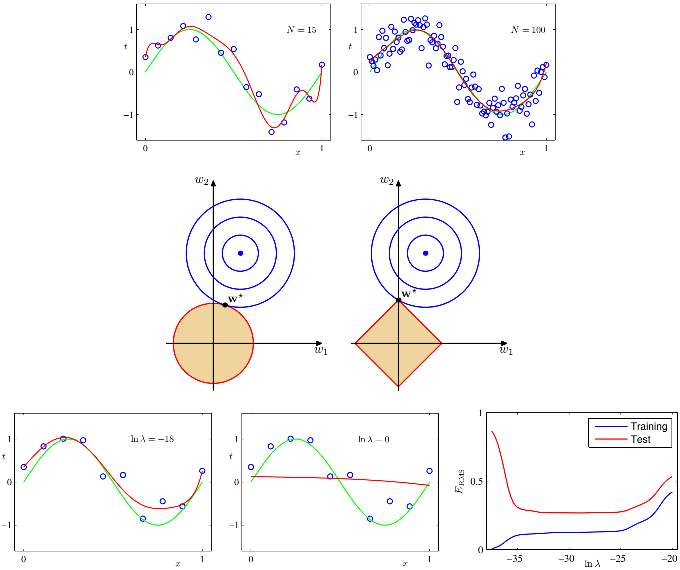
Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

AN EXAMPLE (BISHOP)

- Max likelihood solution for different max degree of monomial *M*



REGULARIZATION



Outline

THE BAYESIAN APPROACH

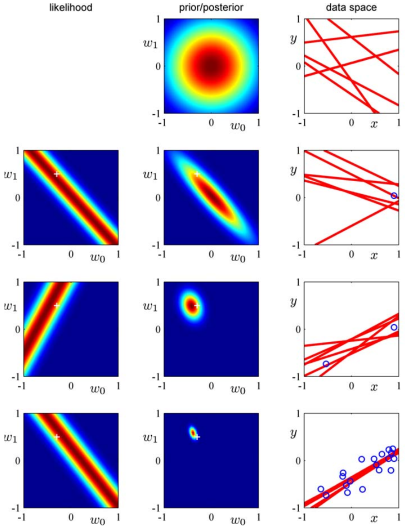
- Regularisation works by biasing
- One way to bias estimators is to have prior beliefs and being Bayesian
- Gaussian prior for regression weights: $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$
- Compute posterior by by Bayes theorem:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \alpha, \beta) p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\mathbf{X}, \alpha, \beta)}$$

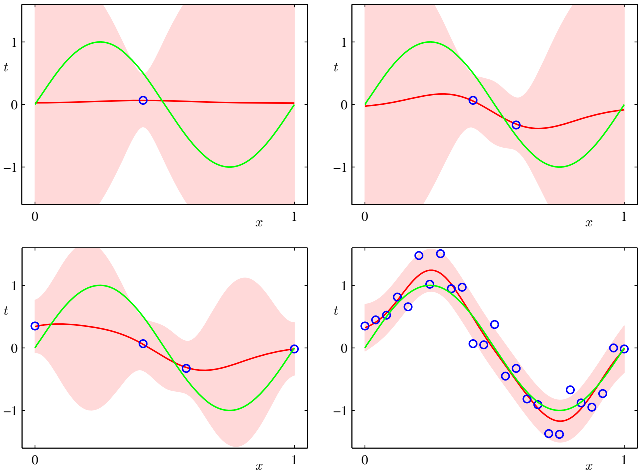
- Predictive distribution:

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

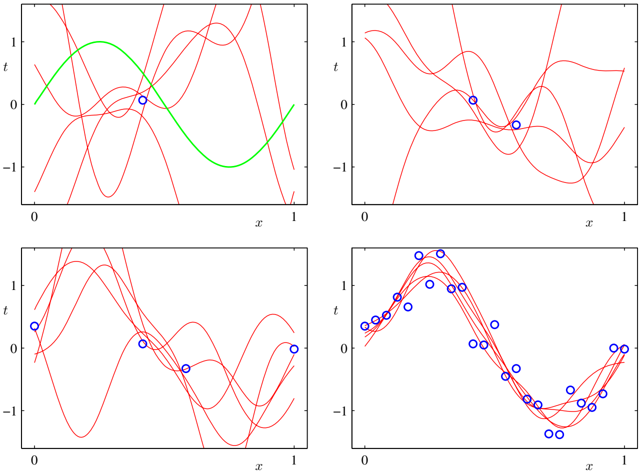
POSTERIOR UPDATE



EXAMPLE



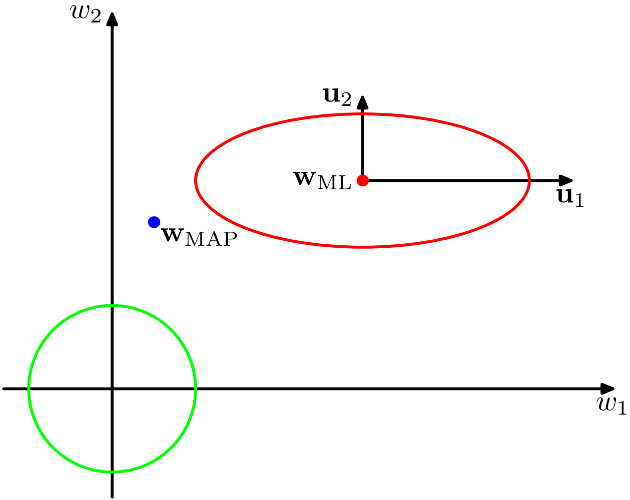
EXAMPLE



Marginal likelihood

- The marginal likelihood or evidence is $p(\mathbf{t}|\alpha, \beta)$.
- It can be used to identify good hyperparameters α and β
- If we have more models, e.g. $\mathcal{M}_1$ and $\mathcal{M}_2$, the evidence $p(\mathbf{t}|\mathcal{M}_j)$ can be used for Bayesian model comparison (via Bayes factors) or to compute posterior model support $p(\mathcal{M}_j|\mathbf{t})$

Effective number of parameters γ



Outline

1. Linear regression
2. Bayesian linear regression
3. Dual representation and kernels

KERNELS AND DUAL FORMULATION

- Dual variables $\mathbf{a}$ are defined via input data projection:

$$\mathbf{w} = \sum_{j=1}^{N} a_j \boldsymbol{\phi}(\mathbf{x_j})$$

- The kernel is $k(\mathbf{x_i}, \mathbf{x_j}) := \boldsymbol{\phi}(\mathbf{x_i})^T \boldsymbol{\phi}(\mathbf{x_j})$
- The Gram matrix $\mathbf{K}$ is $K_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$
- The dual regression problem

$$E_d(\mathbf{a}) + \lambda E_W(\mathbf{a}) = \sum_{i=1}^{N} (t_i - \mathbf{a}^T \mathbf{K^i})^2 + \lambda \mathbf{a}^T \mathbf{K} \mathbf{a}$$

  has also closed form solution
- The kernel trick avoids direct reference to basis functions.