Bayesian Linear Regression

Evgeny Burnaev

Skoltech, Moscow, Russia

Outline

Linear Model

- Linear Basis Function Models

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

  where $\phi_j(\mathbf{x})$ are known basis functions
- Typical basis fucntions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$

$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1}{}^{\mathrm{T}}\mathbf{x} + \mu_{j,0}\right),\ \sigma(a) = \frac{1}{1 + e^{-a}}$$

- We assume that parameters of basis functions are fixed to some known values

Maximum Likelihood and Least Squares

- Data model for the target t: $t = y(\mathbf{x}, \mathbf{w}) + \epsilon$

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}),\ \mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt = y(\mathbf{x}, \mathbf{w})$$

- Data likelihood (here $\mathbf{t} = \{t_1, \ldots, t_N\}$ and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$)

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

- Data log-likelihood has the form

$$\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{n=1}^{N} \log \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

$$= \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \beta E_D(\mathbf{w}),$$

where $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$

Lest Squares = MLE

- Optimizing log-likelihood:

$$\mathbf{w}_{ML} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t},\ \mathbf{\Phi} = \{(\phi_i(\mathbf{x}_j))_{j=0}^{M-1}\}_{i=1}^{N}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathbf{w}_{ML}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

- Regularized Least Squares

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

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \to \min_{\mathbf{w}}$$

$$\mathbf{w}_{LS} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

The Bias-Variance Decomposition

- Data generation process: $t = h(\mathbf{x}) + \epsilon$
- Optimal prediction: $h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt$
- The expected squared loss:

$$\mathbb{E}[(t-y(\mathbf{x}))^2] = \int \{y(\mathbf{x})-h(\mathbf{x})\}^2 p(\mathbf{x})d\mathbf{x} + \int \{h(\mathbf{x})-t\}^2 p(\mathbf{x},t)d\mathbf{x}dt$$

- In our case

$$y(\mathbf{x}) = y(\mathbf{x};\mathcal{D})$$

The Bias-Variance Decomposition

- The Bias-Variance Decomposition

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise},$$

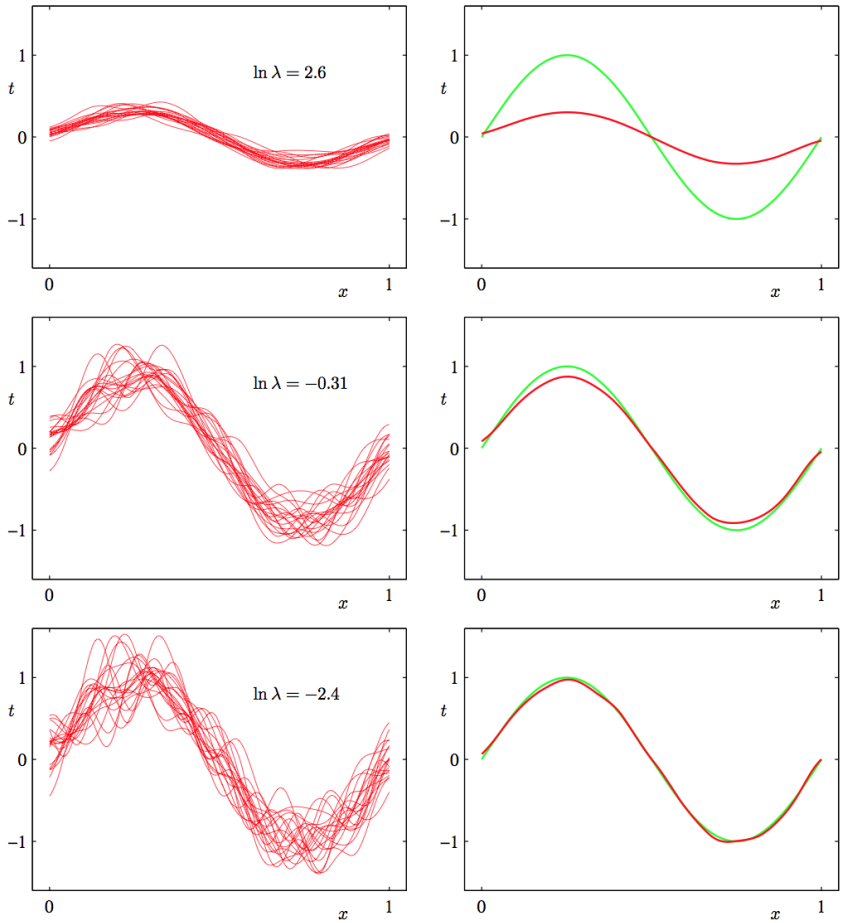
where

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2] p(\mathbf{x}) d\mathbf{x}$$

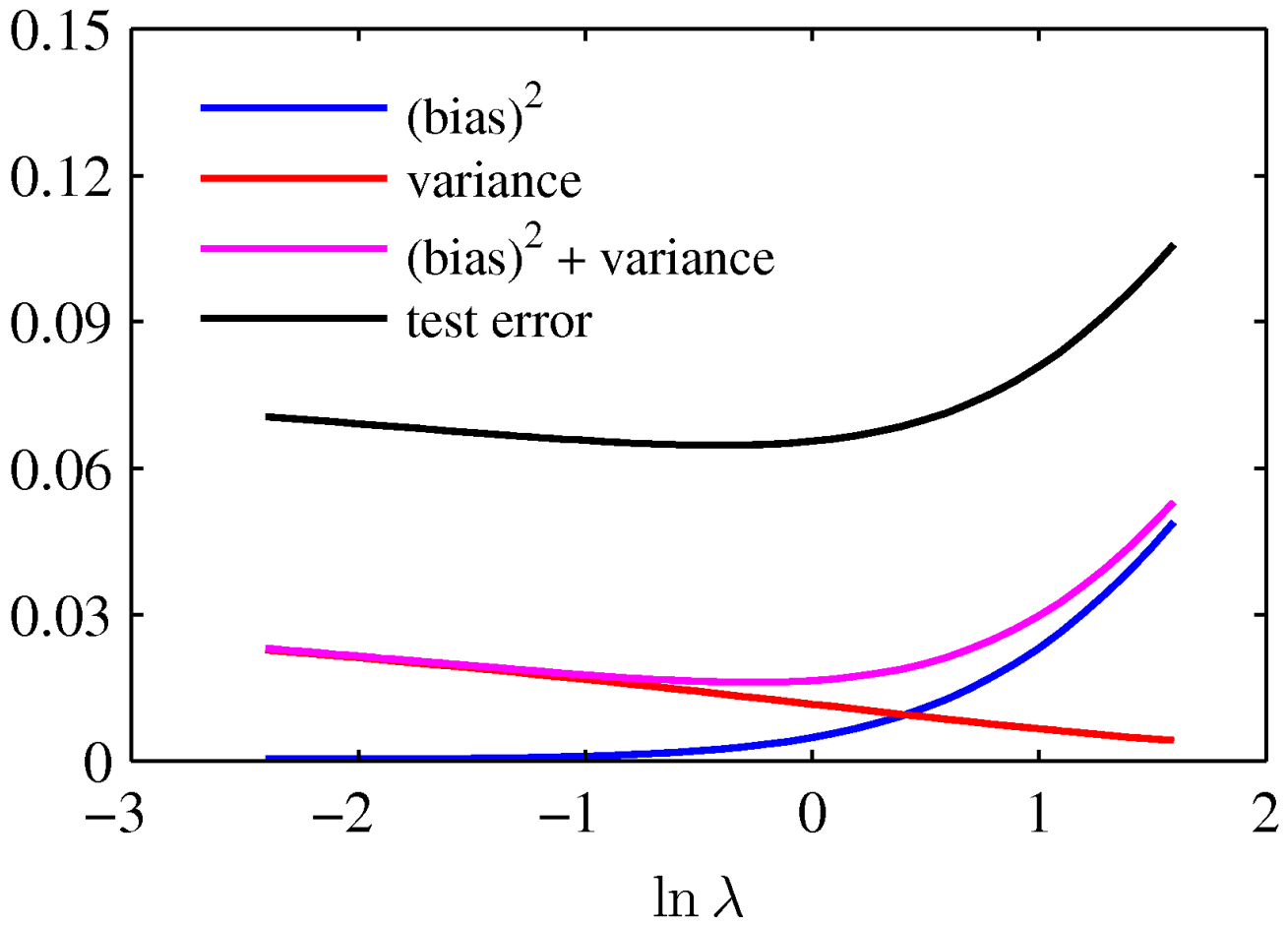
$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

The Bias-Variance Decomposition



Dependence of bias/variance on model complexity (regularization): $L = 100$ data sets, $N = 25$ data points each, $M = 25$ Gaussian basis functions. The right column: average of the 100 fits (red)

The Bias-Variance Decomposition



Parameter distribution

- The likelihood $p(\mathbf{t}|\mathbf{w})$ is the exponential of a quadratic function of $\mathbf{w}$. Thus the conjugate prior is

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

- The posterior w.r.t. data $\mathcal{D} = \{\mathbf{t}, \mathbf{X}\}$

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

where

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right)$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$$

Parameter distribution

- The typical prior is

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

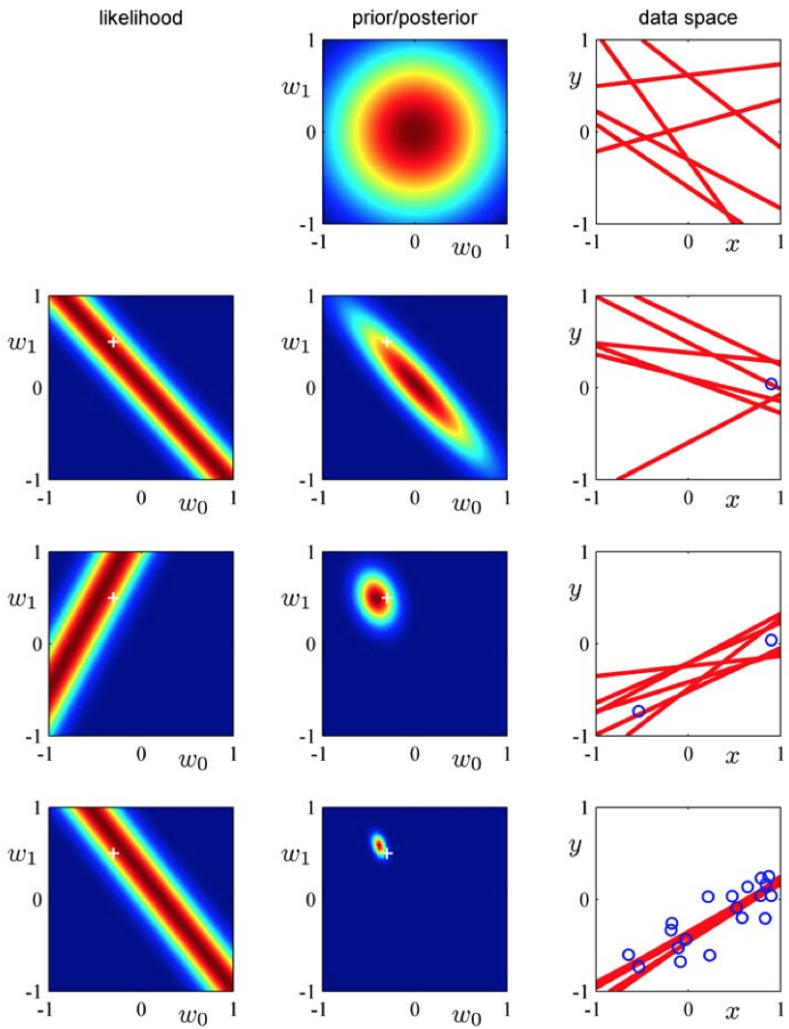
- The posterior is defined by

$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$
$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha^{-1}\mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}$$

- The log posterior

$$\log p(\mathbf{w}|\mathcal{D}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const}$$

Sequential Bayesian Learning



The Model $y(x, \mathbf{w}) = w_0 + w_1 x$

Predictive Distribution

- Make prediction of t for new value of $\mathbf{x}$:

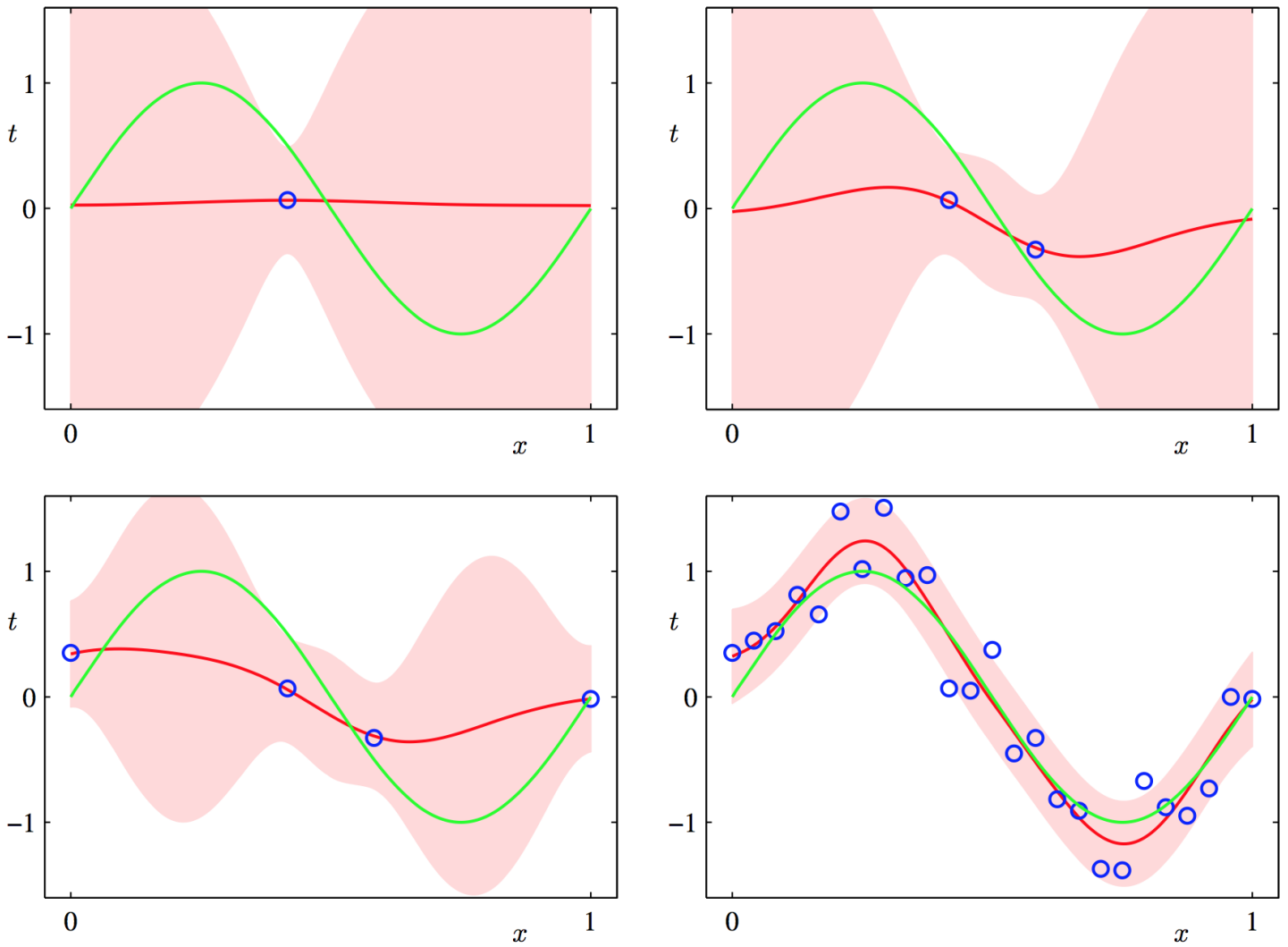
$$p(t|\mathbf{x}, \mathcal{D}, \alpha, \beta) = \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}, \alpha, \beta) d\mathbf{w}$$

- Since $p(t|\mathbf{x}, \mathbf{w}, \beta)$ is Gaussian and the posterior $p(\mathbf{w}|\mathcal{D}, \alpha, \beta)$ is Gaussian, then

$$p(t|\mathbf{x}, \mathcal{D}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})),$$
$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}),\ \mathbf{S}_N^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

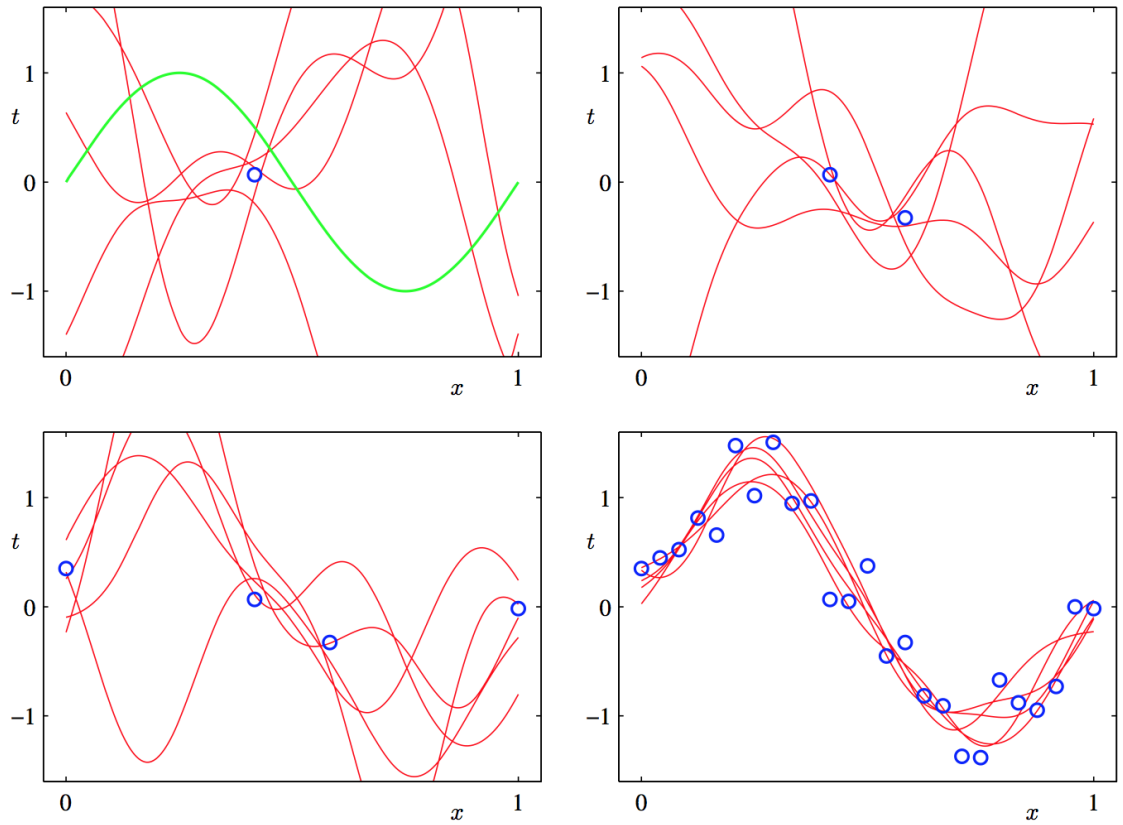
- $p(t|\mathbf{x}, \mathcal{D}, \alpha, \beta)$ depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach!

Predictive Distribution



$M = 9$ Gaussian functions

Predictive Distribution



Plots of $y(\mathbf{x}, \mathbf{w})$ using samples from the posterior distributions over $\mathbf{w} \sim p(\mathbf{w}|\mathcal{D}, \alpha, \beta)$ for some α and β

Bayesian Model Comparison

- We select among models $\{\mathcal{M}_i\}_{i=1}^{L}$. Here a model refers to a probability distribution over the observed data $\mathcal{D}$
- In case of input-output data $\mathcal{D} = \{\mathbf{X}, \mathbf{t}\}$ we assume $\mathbf{X}$ to be known and fixed
- $p(\mathcal{M}_i)$ is a prior. The posterior is

$$p(\mathcal{M}_i|\mathcal{D}) \sim p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i)$$

- If $p(\mathcal{M}_i) \sim \frac{1}{L}$, then the main term is the model evidence $p(\mathcal{D}|\mathcal{M}_i)$. The Bayes factor for two models is $\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$
- Predictive distribution

$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D})$$

Parametric Model

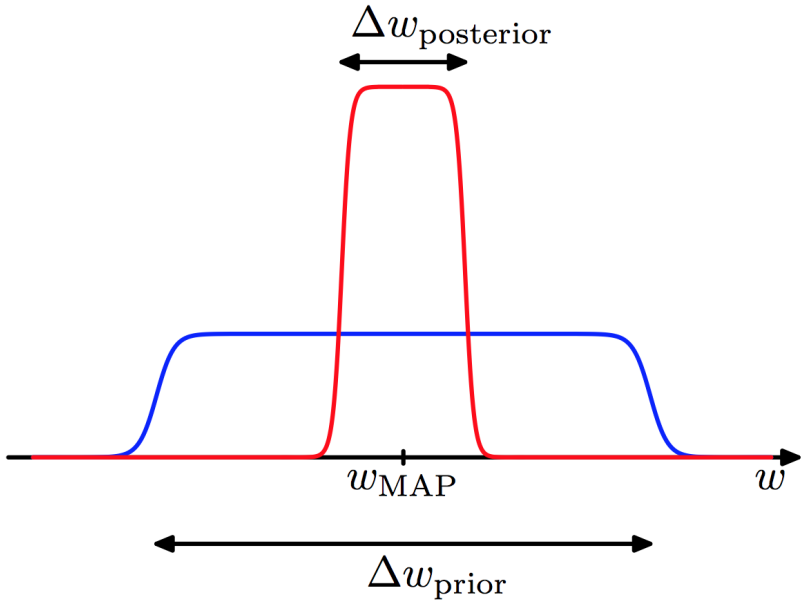
- For a model with parameters $\mathbf{w}$ the model evidence

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}$$

- Cf. with the posterior distribution

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}, \mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)}$$

Model Selection



- Let us consider an approximation: $p(w) = \frac{1}{\Delta w_{\text{prior}}}$
- The posterior

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|w, \mathcal{M}) p(w|\mathcal{M}) dw \approx p(\mathcal{D}|w_{MAP}, \mathcal{M}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

Model Selection

- The log-posterior

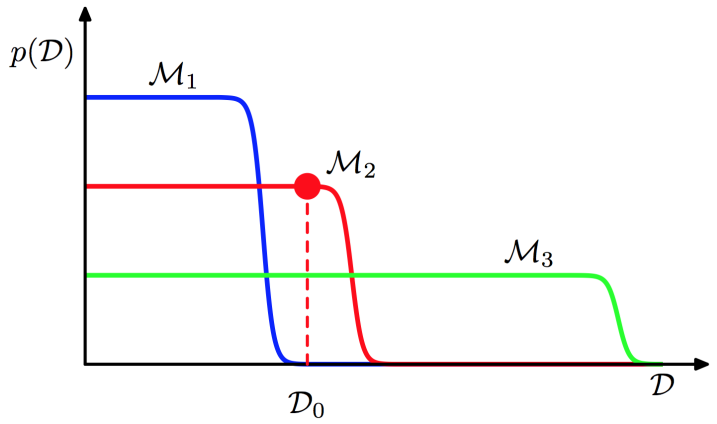
$$\log p(\mathcal{D}|\mathcal{M}) \approx \log p(\mathcal{D}|w_{MAP}, \mathcal{M}) + \log \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$

- In case of M parameters

$$\log p(\mathcal{D}|\mathcal{M}) \approx \log p(\mathcal{D}|w_{MAP}, \mathcal{M}) + M \log \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$

- Thus if parameters are finely tuned to the data in the posterior distribution, then the penalty term is large, since $\Delta w_{\text{posterior}} \ll \Delta w_{\text{prior}}$

Model Complexity vs. Data



- A simple model (for example, based on a first order polynomial) has little variability and so will generate data sets that are fairly similar to each other. Its distribution $p(\mathcal{D})$ is therefore confined to a relatively small region of the horizontal axis
- By contrast, a complex model (such as a ninth order polynomial) can generate a great variety of different data sets, and so its distribution $p(\mathcal{D})$ is spread over a large region of the space of data sets
- The more complex model spreads its predictive probability over too broad a range of data sets and so assigns relatively small probability to any one of them

Predictive Distribution

- Make prediction of t for new value of $\mathbf{x}$:

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

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

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}),\ \mathbf{S}_N^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

- $p(t|\mathbf{x}, \mathcal{D}, \alpha, \beta)$ depends on α and β! We introduce hyperpriors over α and β!

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

Predictive Distribution

- We introduce hyperpriors over α and β

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

- If the posterior distribution $p(\alpha, \beta|\mathcal{D})$ is sharply peaked around values $\hat{\alpha}$ and $\hat{\beta}$, then we simply marginalize over $\mathbf{w}$, where α and β are fixed to the values $\hat{\alpha}$ and $\hat{\beta}$, so that

$$p(t|\mathbf{x}, \mathcal{D}) \approx p(t|\mathbf{x}, \mathcal{D}, \hat{\alpha}, \hat{\beta}) = \int p(t|\mathbf{x}, \mathbf{w}, \hat{\beta})p(\mathbf{w}|\mathcal{D}, \hat{\alpha}, \hat{\beta})d\mathbf{w}$$

- The posterior for α and β is given by

$$p(\alpha, \beta|\mathcal{D}) \sim p(\mathcal{D}|\alpha, \beta)p(\alpha, \beta)$$

If the prior is relatively flat, then in the evidence framework $(\hat{\alpha}, \hat{\beta}) = \arg\max_{\alpha,\beta} p(\mathcal{D}|\alpha, \beta)$

Evaluation of the evidence function

Let us calculate the evidence for (α, β)

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

Let us denote by $E(\mathbf{w})$ the sum of the fit and the regularization on coefficients $\mathbf{w}$

$$E(\mathbf{w}) = \beta E_D(\beta) + \alpha E_W(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \mathbf{\Phi w}\|^2 + \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w}$$

then since $p(\mathcal{D}|\mathbf{w}, \beta)$ and $p(\mathbf{w}|\alpha)$ are Gaussians with quadratic forms $E_D(\beta)$ and $E_W(\mathbf{w})$, we get that

$$p(\mathcal{D}|\alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

Evaluation of the evidence function

For

$$\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} \in \mathbb{R}^{M\times M},\ \mathbf{m}_N = \beta\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

we get that

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N),$$

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N,\ \mathbf{A} = \nabla\nabla E(\mathbf{w})$$

Evaluation of the evidence function

Thus

$$\int \exp\{-E(\mathbf{w})\}d\mathbf{w}$$
$$= \exp\{-E(\mathbf{m}_N)\} \int \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\} d\mathbf{w}$$
$$= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}$$

Therefore the evidence is equal to

$$\log p(\mathcal{D}|\alpha,\beta) = \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E(\mathbf{m}_N) - \frac{1}{2}\log|\mathbf{A}| - \frac{N}{2}\log(2\pi)$$

Maximizing the evidence function

— Let us maximize $p(\mathcal{D}|\alpha,\beta)$ w.r.t. α

$$\log p(\mathcal{D}|\alpha,\beta) \sim \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E(\mathbf{m}_N) - \frac{1}{2}\log|\mathbf{A}| \to \max_{\alpha}$$

— Let us consider the eigenvector equation

$$(\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

$\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ has eigenvalues $\alpha + \lambda_i$

$$\frac{d}{d\alpha}\log|\mathbf{A}| = \frac{d}{d\alpha}\log\prod_i(\lambda_i+\alpha) = \frac{d}{d\alpha}\sum_i\log(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha}$$

— The stationary points of $\log p(\mathcal{D}|\alpha,\beta)$ w.r.t. α satisfy

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N - \frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha}$$

$$\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N = M - \alpha\sum_i\frac{1}{\lambda_i+\alpha} = \gamma$$

Maximizing the evidence function

$$\alpha \mathbf{m}_N^{\mathrm{T}} \mathbf{m}_N = M - \alpha \sum_i \frac{1}{\lambda_i + \alpha} = \gamma$$

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}$$

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\mathrm{T}} \mathbf{m}_N}$$

We adopt an iterative process:

- We make an initial choice for α
- We use this to find $\mathbf{m}_N$
- We evaluate γ and re-estimate α, etc.

Maximizing the evidence function

— Since for eigenvalues of $\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ we get that $\frac{d\lambda_i}{d\beta} = \frac{\lambda_i}{\beta}$, then

$$\frac{d}{d\beta}\log|\mathbf{A}| = \frac{d}{d\beta}\sum_i \log(\lambda_i + \alpha) = \frac{1}{\beta}\sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$$

— The stationary points of $\log p(\mathcal{D}|\alpha, \beta)$ w.r.t. α

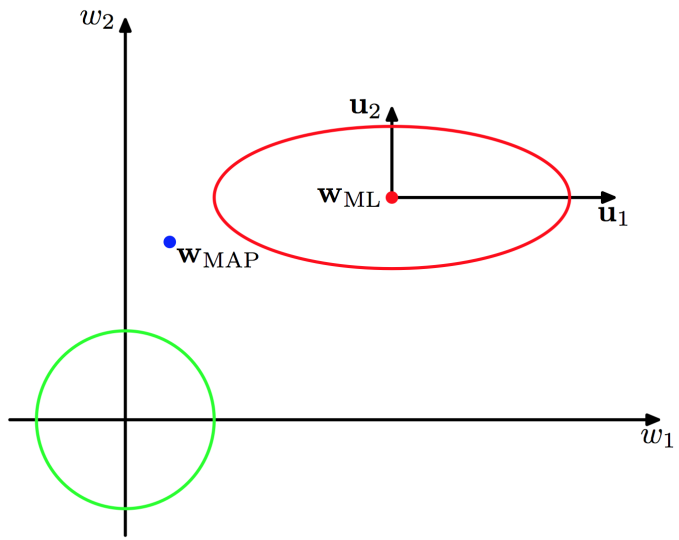
$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\gamma}{2\beta}$$

$$\frac{1}{\beta} = \frac{1}{N - \gamma}\sum_{n=1}^{N}\{t_n - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

— We adopt an iterative process:

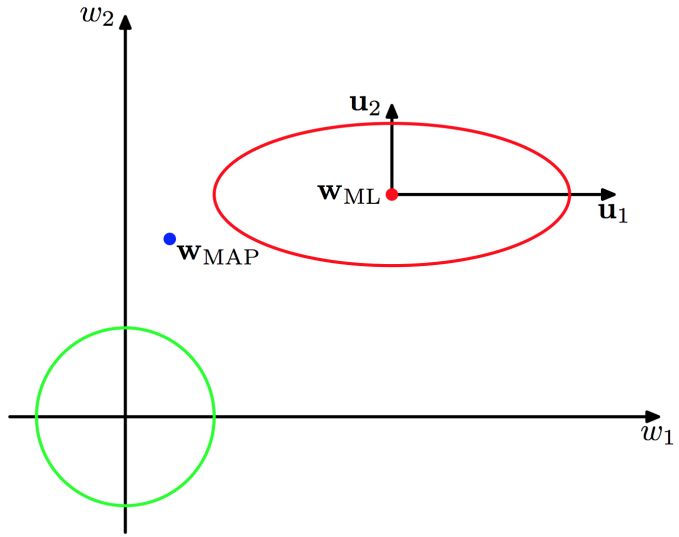
- We make an initial choice for β
- We use this to find $\mathbf{m}_N$ and γ
- We re-estimate β, etc.

Effective number of parameters



- Contours of the likelihood function (red) and the prior (green) in which the axes in parameter space have been rotated to align with the eigenvectors $\mathbf{u}_i$ of the Hessian
- For $\alpha = 0$ the mode of the posterior $\mathbf{w}_{MAP} = \mathbf{w}_{ML}$; for non-zero α the mode is at $\mathbf{w}_{MAP} = \mathbf{m}_N$

Effective number of parameters



- In the direction w_1 the eigenvalue λ_1 is small compared with α and so the quantity $\lambda_1/(\lambda_1 + \alpha)$ is close to zero, and so $w_{1,MAP} \approx 0$
- By contrast, in the direction w_2 the eigenvalue $\lambda_2 \gg \alpha$ is large and so the quantity $\lambda_2/(\lambda_2 + \alpha) \approx 1$, i.e. $w_{2,MAP} \approx w_{2,MLE}$
- Thus $0 \leq \gamma \leq M$. Since not all parameters are tuned to the data: the effective number of parameters determined by the data is γ, with remaining $M - \gamma$ param. set to small values by

Effective number of parameters

- Let us consider the limit $N \gg M$
- Since $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ involves an implicit sum over data points, so λ_i increase with the size of the data set. In this case $\gamma = M$ and re-estimation equations

$$\alpha = \frac{M}{2E_W(\mathbf{m}_N)}$$
$$\beta = \frac{N}{2E_D(\mathbf{m}_N)}$$