# **Evidence Approximation**

Sargur Srihari srihari@cedar.buffalo.edu

# Background

- Bayesian methods need priors with hyperparameters
- Analytical treatment with distributions over hyper-parameters is intractable
- Approximate methods are used
- Also called as:
  - Empirical Bayes (Statistics literature)
  - Type 2 Maximum Likelihood
  - Generalized Maximum Likelihood
  - Evidence Approximation (ML literature)

# **Topics**

- Linear Regression: Bayesian Treatment
  - Hyper-parameters for noise and weights
  - Predictive distribution
    - Marginalize over hyper-parameters and weights
- Need for Approximation
  - Evaluation of evidence function
    - Maximizing the evidence function
- Interpretation: Effective no. of parameters

## Linear Regression with Basis Functions

- Target variable is a scalar t given by
  - deterministic function y(x,w) with additive Gaussian noise

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

 $\varepsilon$  is zero-mean Gaussian with precision  $\beta$ 

Linear regression

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

- where  $w=(w_0, w_1, ..., w_{M-1})$  and  $\phi=(\phi_0, \phi_1, ..., \phi_{M-1})^T$
- $\phi(x)$  are called Basis functions
- There are M parameters

# Bayesian treatment

- Prior distribution of parameter p(w)
- Since likelihood p(t|w) with Gaussian noise has an exponential form
  - Conjugate prior is chosen to be Gaussian  $p(\mathbf{w}) = N(\mathbf{w} | \mathbf{m}_{o}, S_{o})$  with mean  $\mathbf{m}_{o}$  and covariance  $S_{o}$
- Posterior is a Gaussian  $p(\mathbf{w}|\mathbf{t}) = N(\mathbf{w}|\mathbf{m}_N, S_N)$ where  $m_N = S_N (S_0^{-1} m_0 + \beta \Phi^T t)$  and  $S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \dots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & & & \\ \phi_0(\mathbf{x}_N) & & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix} \quad N \times M \text{ Design matrix}$$

# Parameters and Hyper-parameters

- Parameters w, a vector of M elements
  - Where *M* is model complexity
- Two hyper-parameters:
  - $-\alpha$  is weight precision

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

- Zero-mean isotropic Gaussian
- Has single parameter  $\alpha$
- $\beta$  is noise precision

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

### **Predictive Distribution**

- Usually not interested in the value of w itself
  - But predicting t for new values of x
- Introducing hyper-priors over  $\alpha$  and  $\beta$  and
  - marginalizing over w,  $\alpha$  and  $\beta$

$$p(t \mid t) = \iiint p(t, w, \alpha, \beta \mid t) dw d\alpha d\beta$$
$$= \iiint p(t \mid w, \beta) p(w \mid t, \alpha, \beta) p(\alpha, \beta \mid t) dw d\alpha d\beta$$

Conditioning variable x has been left out for convenience.

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

$$p(\mathbf{w} \mid \mathbf{t}) = N(\mathbf{w} \mid \mathbf{m}_{N}, S_{N})$$

- Marginalizing wrt  $\alpha$ ,  $\beta$ , w is analytically intractable
- If  $p(\alpha, \beta | t)$  is sharply peaked around  $\hat{\alpha}$ ,  $\hat{\beta}$  then  $p(t | t) \approx p(t | t, \hat{\alpha}, \hat{\beta}) = \int p(t | w, \hat{\beta}) p(w | t, \hat{\alpha}, \hat{\beta}) dw$

# Strategy of Approximation

- Goal: determine  $\hat{\alpha}$ ,  $\hat{\beta}$  where  $p(\alpha, \beta | t)$  is maximum
- From Bayes theorem

$$p(\alpha, \beta \mid t) \quad \alpha \quad p(t \mid \alpha, \beta) p(\alpha, \beta)$$

- If prior  $p(\alpha, \beta)$  is flat, values  $\hat{\alpha}$  and  $\hat{\beta}$  obtained by maximizing marginal likelihood  $p(t|\alpha, \beta)$
- Can analytically determine maximum of  $p(t|\alpha,\beta)$  for linear basis model

### Evidence Function Evaluation

### Obtained by integrating over weight parameter w

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

#### Where we substitute

$$\begin{aligned} p(\mathbf{t} \mid w, \beta) &= \prod_{i=1}^{N} N\Big(\mathbf{t}_{n} \mid \mathbf{w}^{T} \phi(\mathbf{x}_{n}), \beta^{-1}\Big) \\ p(w \mid \alpha) &= N\Big(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}I\Big) \end{aligned}$$

### And get

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

$$where$$

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w})$$

$$= \frac{\beta}{2} || t - \Phi \mathbf{w} ||^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$

$$E(\mathbf{w}) \text{ is the regularize sum of squares error Upto a constant of properties of the properties of the$$

 $E(\mathbf{w})$  is the regularized Upto a constant of proportionality

## Final Expression for Model Evidence

$$\ln p(\mathsf{t} \mid \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathsf{m}_N) - \frac{1}{2} \ln |A| - \frac{N}{2} \ln(2\pi)$$

 Where M is the dimensionality of w, N is no of samples, and

$$\begin{split} & A = \alpha I + \beta \Phi^T \Phi \quad \text{which implies that A} = \nabla \nabla E(\mathbf{w}) \text{ is the Hessian matrix of Error} \\ & E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^T A(\mathbf{w} - \mathbf{m}_N) \\ & E(\mathbf{m}_N) = \frac{\beta}{2} \mid \mid t - \Phi \mathbf{m}_N \mid \mid^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N \\ & \mathbf{m}_N = \beta A^{-1} \Phi^T \mathbf{t} \end{split}$$

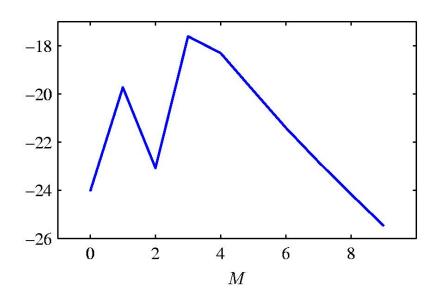
• We wish to maximize wrt M,  $\alpha$ ,  $\beta$ 

## Maximize Model Evidence over M

$$\ln p(\mathsf{t} \mid \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathsf{m}_N) - \frac{1}{2} \ln |A| - \frac{N}{2} \ln(2\pi)$$

We need  $\alpha$  and  $\beta$  to maximize this

# Polynomial Regression Model Evidence vs *M*



Order of Polynomial

### Evidence favors model with M=3

Data has a poor fit for M=2 since the polynomial expansion for sinusoid is an odd function and has no even terms

### Maximize Evidence over α

• Maximization of  $p(t|\alpha,\beta)$  wrt  $\alpha$  yields

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

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

$$\left| \gamma = \sum_{i} \frac{\lambda_{i}}{\alpha + \lambda_{i}} \right|$$

- where  $\lambda_i$  are  $\beta$  times the eigen values of  $\Phi^T\Phi$ 
  - Defined by eigenvector equation  $(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$

# Eigen Value tutorial

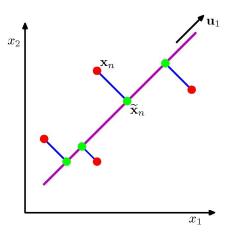
- For a square matrix A of size M x M
  - Eigen vector equation is  $A\mathbf{u}_i = \lambda_i \mathbf{u}_i$  for i = 1,...M where  $\mathbf{u}_i$  is an eigenvector and  $\lambda_i$  the corresponding eigenvalue
    - Can be viewed as M simultaneous equations which are solutions to characteristic equation  $|A-\lambda_i I|=0$

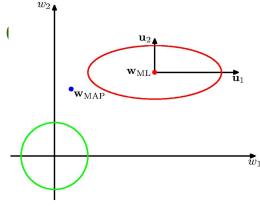
# Eigen Value Interpretation

- Eigen values are real-valued for symmetric matrices
- For a covariance matrix Σ
  - Eigen vector corresponding to largest eigen value indicates direction of maximum variance
    - Principal Components



large eigen values indicate direction
 high curvature of likelihood function





## Iterative Solution for α

• Maximization of  $p(t|\alpha,\beta)$  wrt  $\alpha$  yields

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

$$\gamma = \sum_{i} \frac{\lambda_{i}}{\alpha + \lambda_{i}}$$

- An implicit solution
  - since  $\gamma$  and mode  $m_N$  of posterior depend on  $\alpha$
- Iterative procedure:

- $\mathbf{m}_{N} = \beta A^{-1} \Phi^{T} \mathbf{t}$  $A = \alpha I + \beta \Phi^{T} \Phi^{T}$
- Initial choice for  $\alpha$ , find  $m_N$  etc until convergence
- α determined purely from training data
  - No independent data set required to optimize model

Srihari Machine Learning

# Maximize Evidence over $\beta$

• Maximization of  $\ln p(t|\alpha,\beta)$  wrt noise precision  $\beta$  yields

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^{N} \{t_n - m_N \phi(x_n)\}^2$$

Again, an implicit solution

$$\gamma = \sum_{i} \frac{\lambda_{i}}{\alpha + \lambda_{i}}$$

- Iterative Solution:
  - Choose initial value of  $\beta$ , calculate  $m_N$ ,  $\gamma$  and re-estimate  $\beta$  until convergence 16

## Effective No of Parameters

Interpretation provides insight into Bayesian solution for  $\alpha$ 

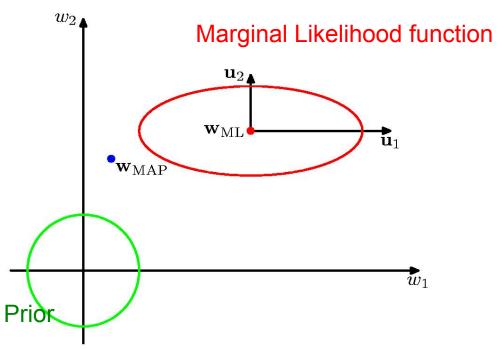
$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N} \qquad \gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}$$

where  $\lambda_i$  are eigen values of Hessian  $\Phi^T\Phi$ 

 $\lambda_i | \lambda_i + \alpha$  lies between 0 and 1. When  $\lambda_i$  is small, corresponding weight is close to zero. Here  $w_I$  is small and  $w_2$  is larger. Thus

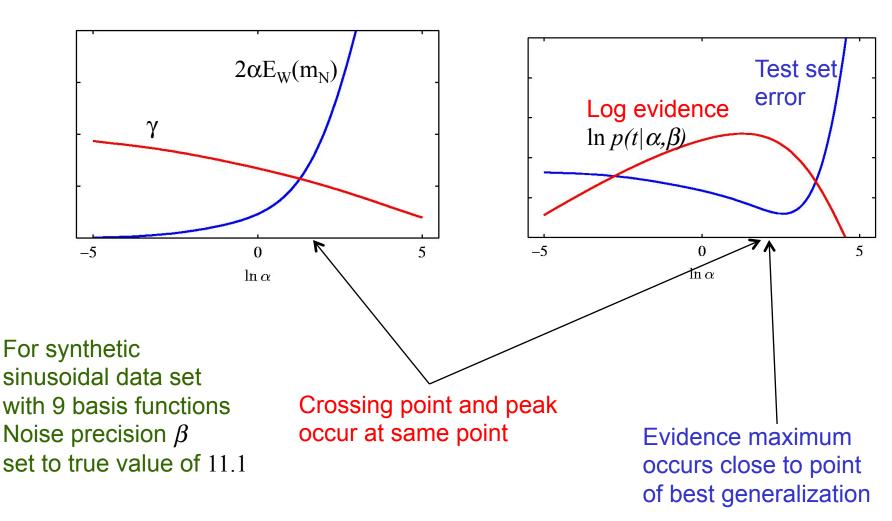
*γ* measures the effective total no of parameters

Axes in parameter space rotated To align with Eigen vectors  $\mathbf{u}_i$  of Hessian  $A = \alpha I + \beta \Phi^T \Phi$ 

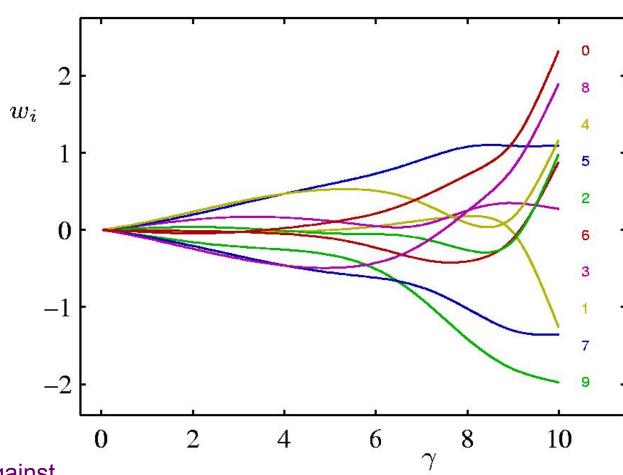


For  $\alpha=0$ ,  $w_{ML}$  is the mode (max likelihood solution) For nonzero  $\alpha$ ,  $w_{MAP}=m_N$  is the mode

# Determining optimum value of $\alpha$



# How $\alpha$ controls parameters $w_i$



Plot of 10 parameters from Gaussian basis function model:

Individual parameters against effective no of parameters

### Limitations of Fixed Basis Functions

 Although we replaced D parameters with M basis functions, the number of basis functions needs to grow exponentially with D

### Solutions

- Radial basis functions networks and SVMs exploit intrinsic dimensionality of data
- Neural networks, with sigmoidal nonlinearity adapt to the data manifold

# Summary of Linear Regression

- Goal: predict value of one or more target variables *t* given value of *D*-dimensional vector **x** of input variables
  - Polynomial is a simple form
- We looked at probabilistic methods
- Linear regression models are linear in parameter values, but use nonlinear functions of input variables (called basis functions)
- Maximum likelihood formulation is equivalent to Least Squares
  - Bias Variance tradeoff exists
  - Partitioning data set is not satisfactory use of data
- Bayesian formulation is most general solution
  - Weighted average over parameter values
  - Hyper-parameters are estimated using evidence approximation
- Limitations when the dimensionality is large