Learning Objectives

1、How to achieve linear regression using basis functions?

2、What are the relationships between maximum likelihood and least squares, between maximum a posterior and regularization, and among expected loss, bias, variance, and noise?

3、What are the common regularization methods for regression?

4、How to achieve Bayesian linear regression?

5、What is the kernel for regression?

6、How to choose the model complexity?

7、What are the evidence approximation and maximization?

Outlines

- Linear Basis Function Models
- Maximum Likelihood and Least Squares
- Bias Variance Decomposition
- Bayesian Linear Regression
- Predictive Distribution
- Bayesian Model Comparison
- Evidence Approximation and Maximization

Bayesian Model Comparison (1)

- How do we choose the ‘right’ model?
- Assume we want to compare models $\mathrm{M}_i, \quad i=1, \cdots, L,$ using data D; this requires computing

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

Posterior Prior *Model evidence* or *marginal likelihood*

- *Bayes Factor*: ratio of evidence for two models

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

Bayesian Model Comparison (2)

- Having computed $p(\mathcal{M}_i|\mathcal{D})$, we can compute the predictive (mixture) distribution

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

- A simpler approximation, known as *model selection*, is to use the model with the highest evidence.

Bayesian Model Comparison (3)

- For a model with parameters $\mathbf{w}$, we get the model evidence by marginalizing over $\mathbf{w}$

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

- Note that

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

Bayesian Model Comparison (4)

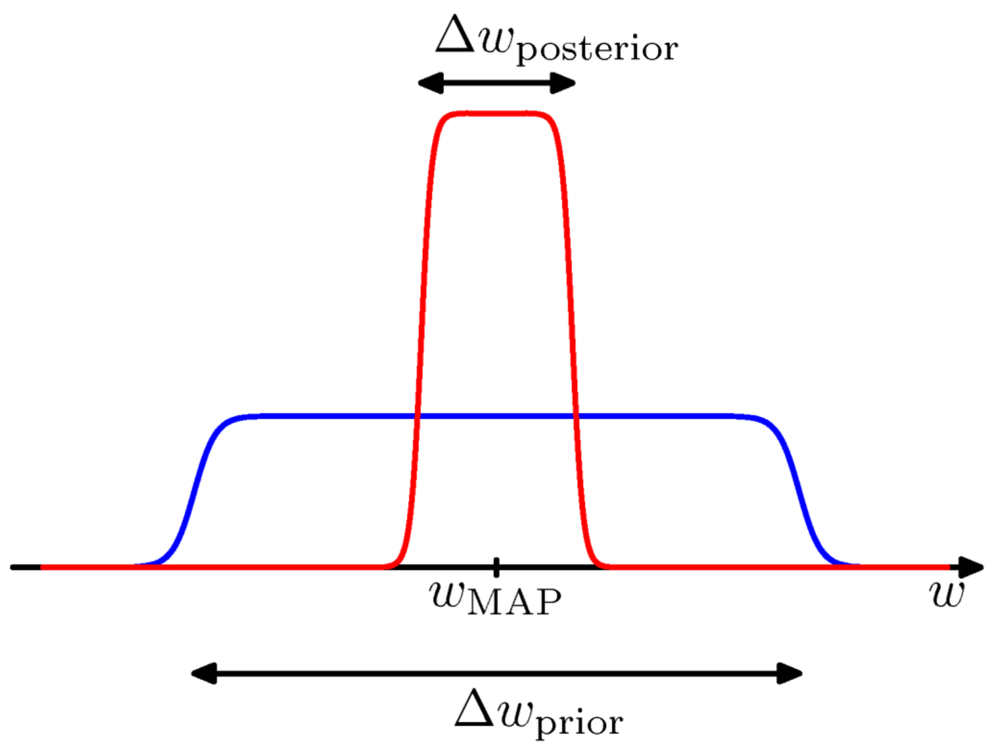
For a given model with a single parameter, w, consider the approximation

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\,\mathrm{d}w$$

$$\simeq \quad p(\mathcal{D}|w_{\mathrm{MAP}})\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}$$

where the posterior is assumed to be sharply peaked.

$$p(w) = \frac{1}{\Delta w_{prior}}$$



Bayesian Model Comparison (5)

- Taking logarithms, we obtain

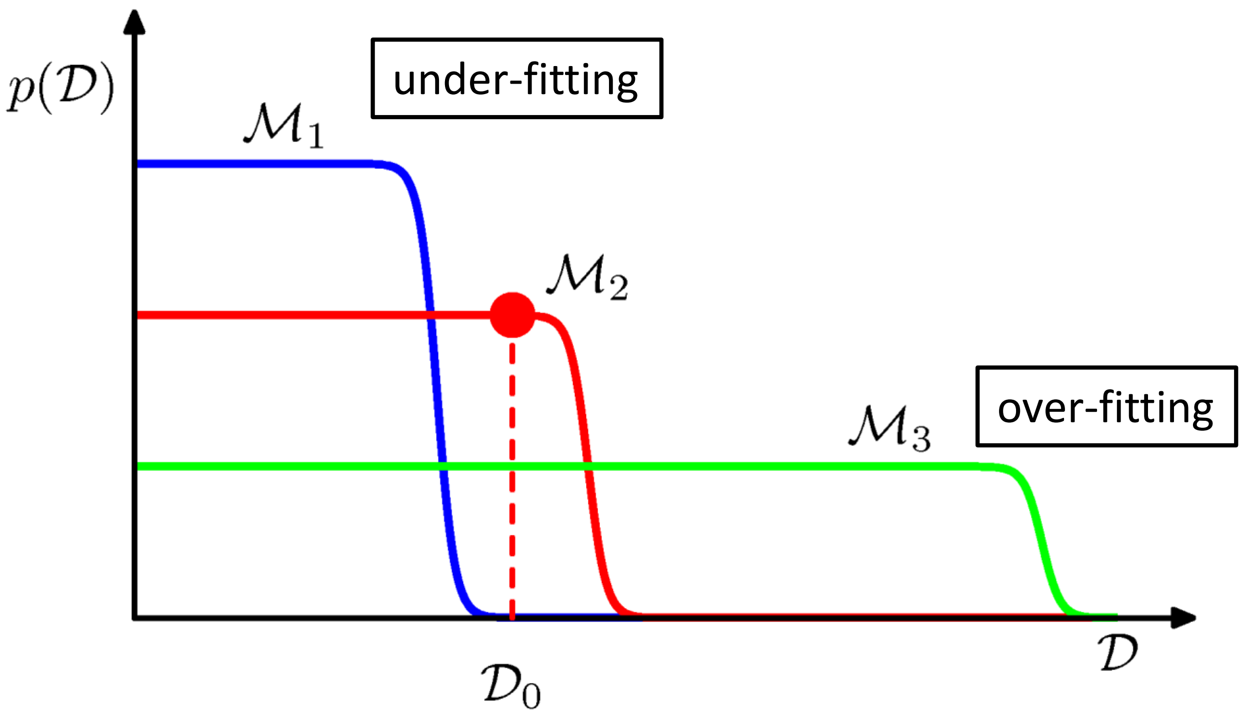
$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \underbrace{\ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)}_{\text{Negative}}.$$

- With M parameters, all assumed to have the same ratio $\Delta w_{\mathrm{posterior}}/\Delta w_{\mathrm{prior}}$, we get

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}}) + \underbrace{M \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)}_{\text{Negative and linear in } M.}.$$

Bayesian Model Comparison (6)

Matching data and model complexity



Outlines

- Linear Basis Function Models
- Maximum Likelihood and Least Squares
- Bias Variance Decomposition
- Bayesian Linear Regression
- Predictive Distribution
- Bayesian Model Comparison
- Evidence Approximation and Maximization*

The Evidence Approximation (1)*

The fully Bayesian predictive distribution is given by

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

but this integral is intractable. Approximate with

$$p(t|\mathbf{t}) \simeq p\left(t|\mathbf{t}, \widehat{\alpha}, \widehat{\beta}\right) = \int p\left(t|\mathbf{w}, \widehat{\beta}\right) p\left(\mathbf{w}|\mathbf{t}, \widehat{\alpha}, \widehat{\beta}\right) \, \mathrm{d}\mathbf{w}$$

where $\left(\widehat{\alpha}, \widehat{\beta}\right)$ is the mode of $p(\alpha, \beta|\mathbf{t})$, which is assumed to be sharply peaked; a.k.a. *empirical Bayes, type II* or *generalized maximum likelihood,* or *evidence approximation*.

The Evidence Approximation (2)*

From Bayes' theorem we have

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta)$$

and if we assume $p(\alpha, \beta)$ to be flat we see that

$$\begin{aligned} p(\alpha, \beta|\mathbf{t}) &\propto p(\mathbf{t}|\alpha, \beta) \\ &= \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}. \end{aligned}$$

General results for Gaussian integrals give

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

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

The Evidence Approximation (3)*

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

Precision: $\boldsymbol{A} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \quad \boldsymbol{A} = \boldsymbol{S}_N^{-1}$

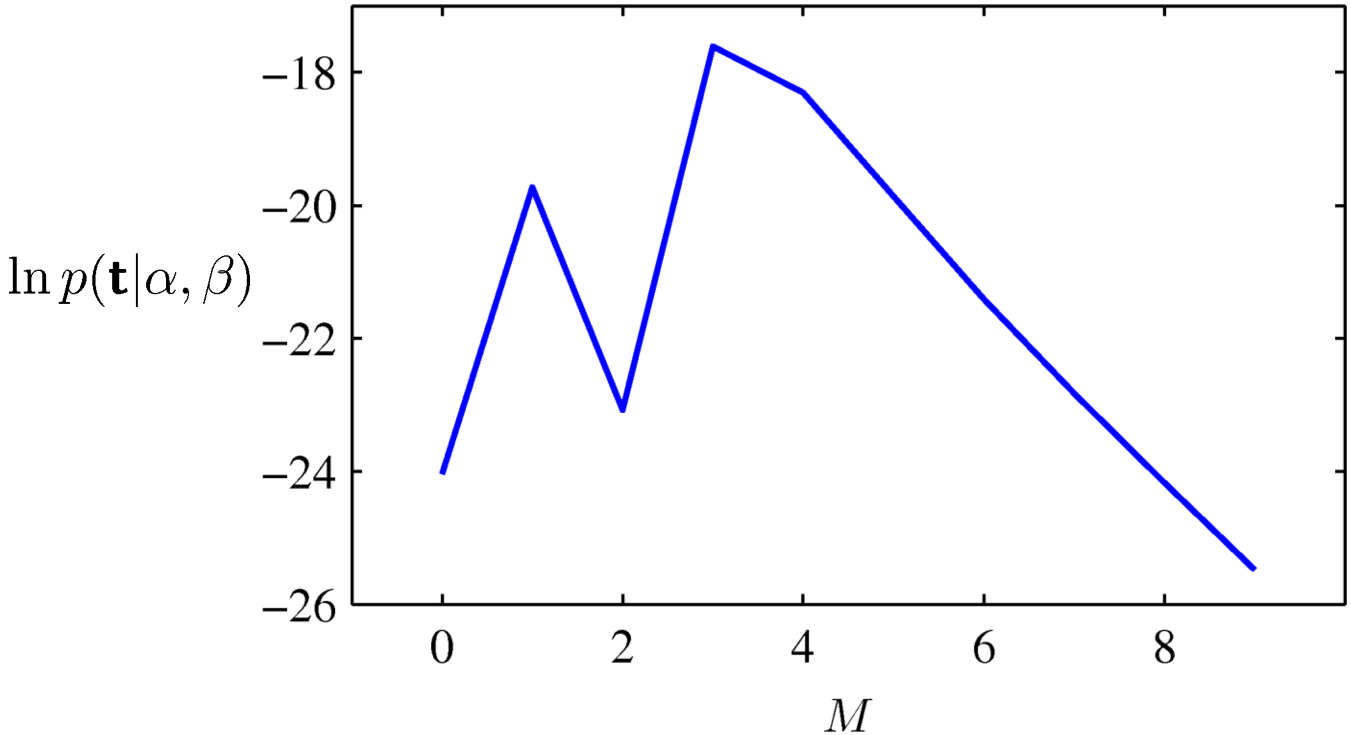
$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}. \end{aligned}$$

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

$$\begin{aligned} &\int \exp\{-E(\boldsymbol{w})\}\,\mathrm{d}\,\boldsymbol{w} \\ &= \exp\{-E(\boldsymbol{m}_N)\} \int \exp\left\{-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_N)^T \boldsymbol{A}(\boldsymbol{w} - \boldsymbol{m}_N)\right\} \mathrm{d}\boldsymbol{w} \\ &= \exp\{-E(\boldsymbol{m}_N)\}(2\pi)^{\frac{M}{2}} |\boldsymbol{A}|^{-\frac{1}{2}} \end{aligned}$$

The Evidence Approximation (4)*

- Example: sinusoidal data, M^{th} degree polynomial, $\alpha = 5 \times 10^{-3}$



Maximizing the Evidence Function (1)*

- To maximise $\ln p(\mathbf{t}|\alpha, \beta)$ w.r.t. α and β, we define the eigenvector equation

$$\boxed{\text{Precision:}} \quad \left(\beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right) \mathbf{u}_i = \lambda_i \mathbf{u}_i.$$

- Thus

$$\boxed{\text{Precision:}} \quad \mathbf{A} = \mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

has eigenvalues $\lambda_i + \alpha$.

Maximizing the Evidence Function (2)*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \ln |\boldsymbol{A}| = \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_i \ln(\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha}$$

$$\frac{\partial \ln p(\mathbf{t}|\alpha, \beta)}{\partial \alpha} = \; 0 = \frac{M}{2\alpha} - \frac{1}{2} \boldsymbol{m}_N^T \boldsymbol{m}_N - \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha}$$

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \ln |\boldsymbol{A}| = \frac{\mathrm{d}}{\mathrm{d}\beta} \sum_i \ln(\lambda_i + \alpha) = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$$

$$\frac{\partial \ln p(\mathbf{t}|\alpha, \beta)}{\partial \beta} = 0 = \frac{N}{2\beta} - \frac{1}{2} \sum_{n=1}^{N} \{t_n - \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 - \frac{\gamma}{2\beta}$$

Maximizing the Evidence Function (3)*

- We can now differentiate $\ln p(\mathbf{t}|\alpha,\beta)$ w.r.t. α and β, and set the results to zero, to get

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

$$\boxed{\frac{1}{\beta_{\mathrm{MAP}}}:} \quad \frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2$$

where

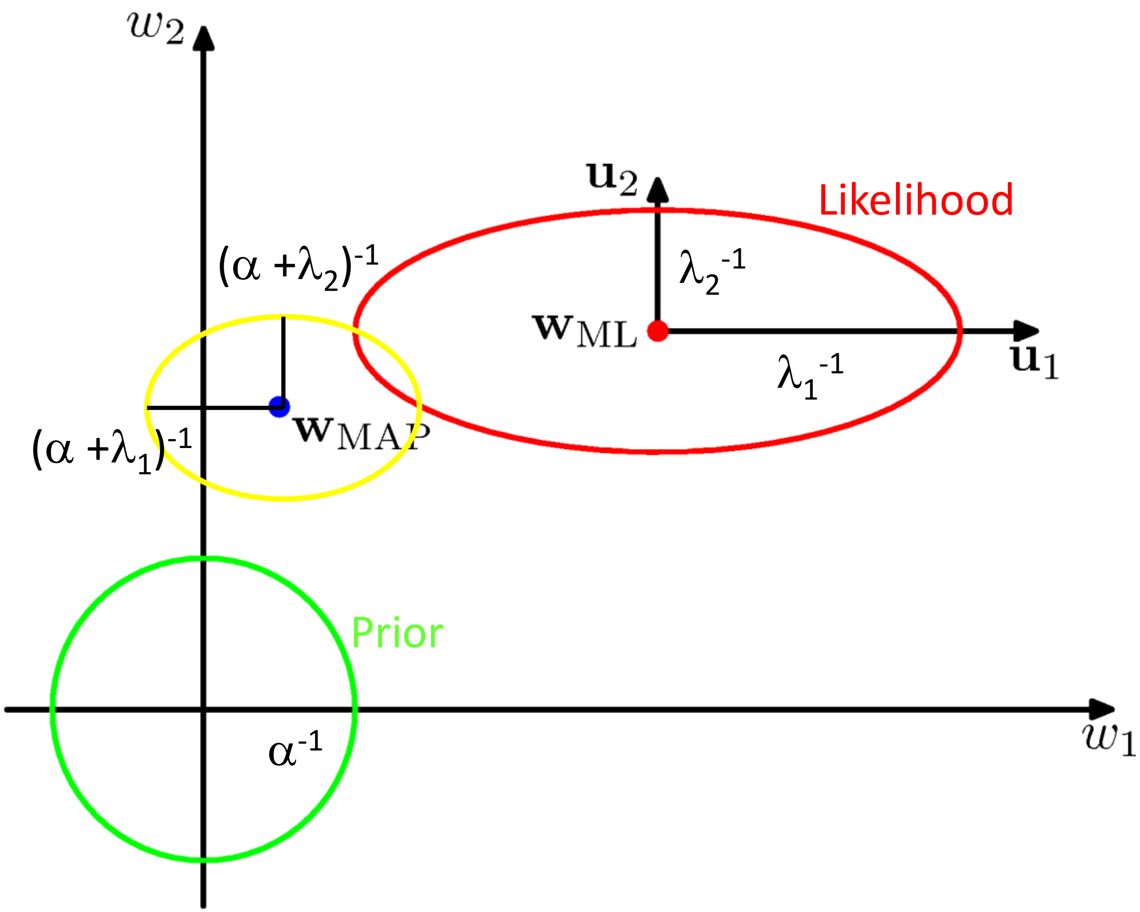
$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$$

γ depends on both α and β.

recall

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

Effective Number of Parameters (1)*



$\lambda_1 \ll \alpha$

w_1 is not well determined by the likelihood when more disturbed from β

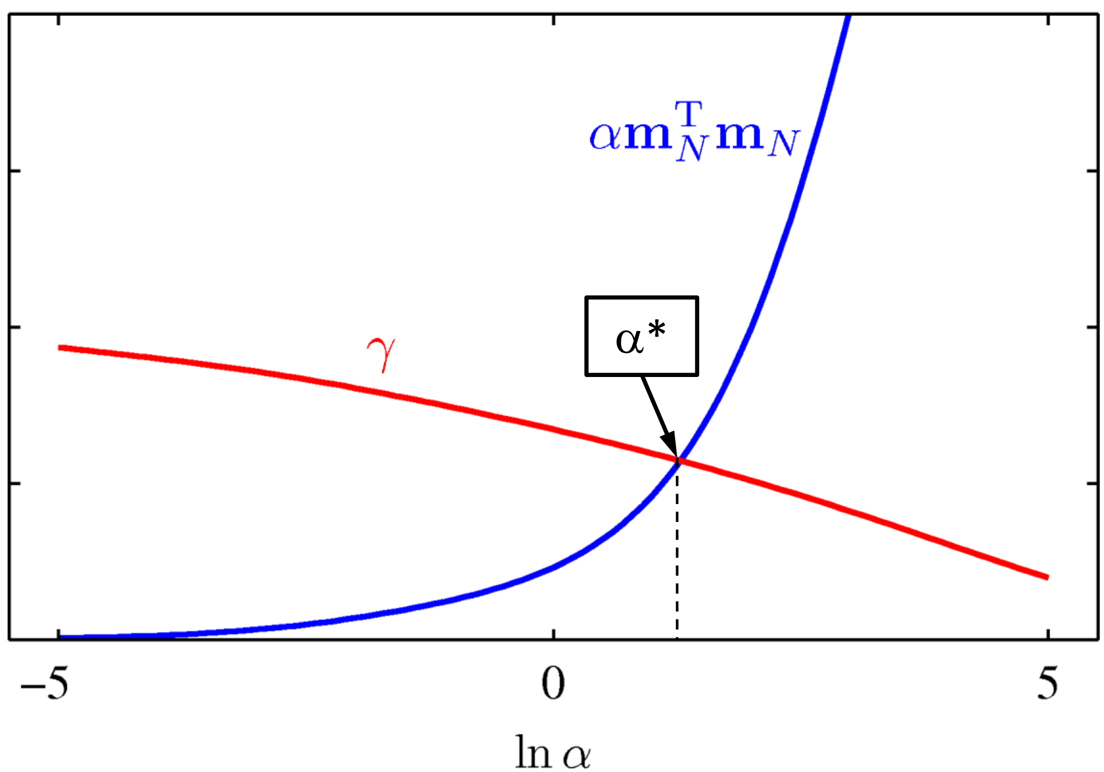
$\lambda_2 \gg \alpha$

w_2 is well determined by the likelihood when less disturbed from β

γ is the number of well determined parameters

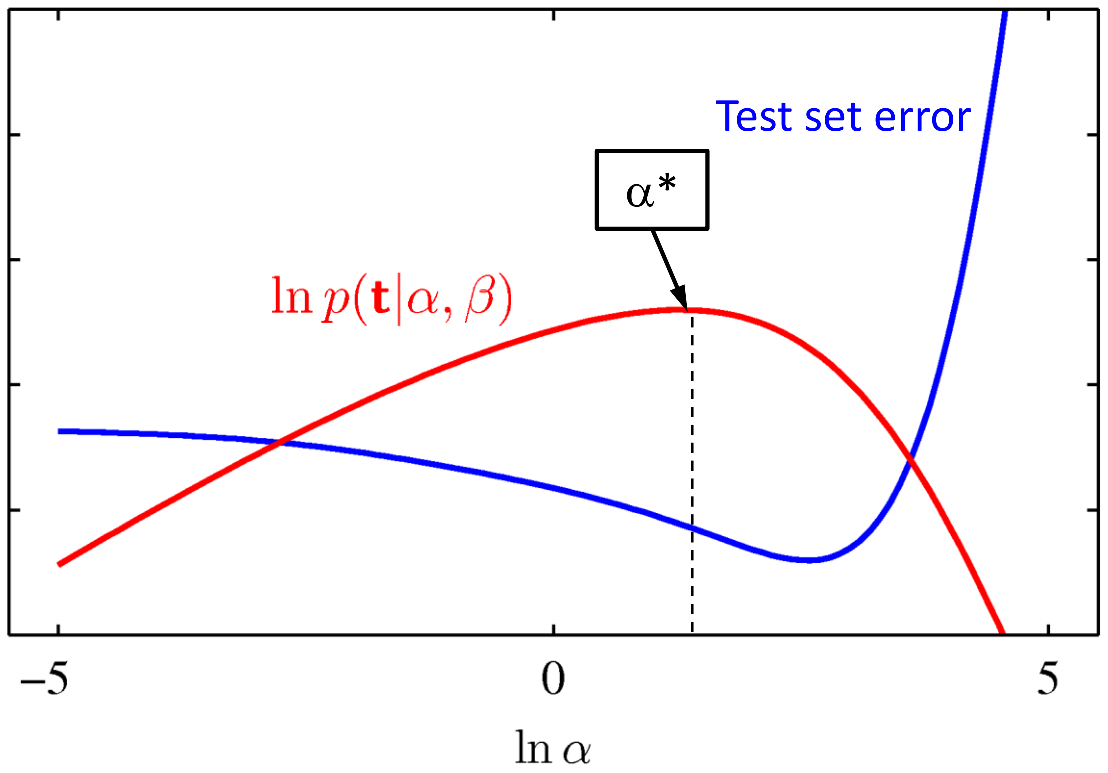
Effective Number of Parameters (2)*

- Example: sinusoidal data, 9 Gaussian basis functions, $\beta = 11.1$ (true value β*).



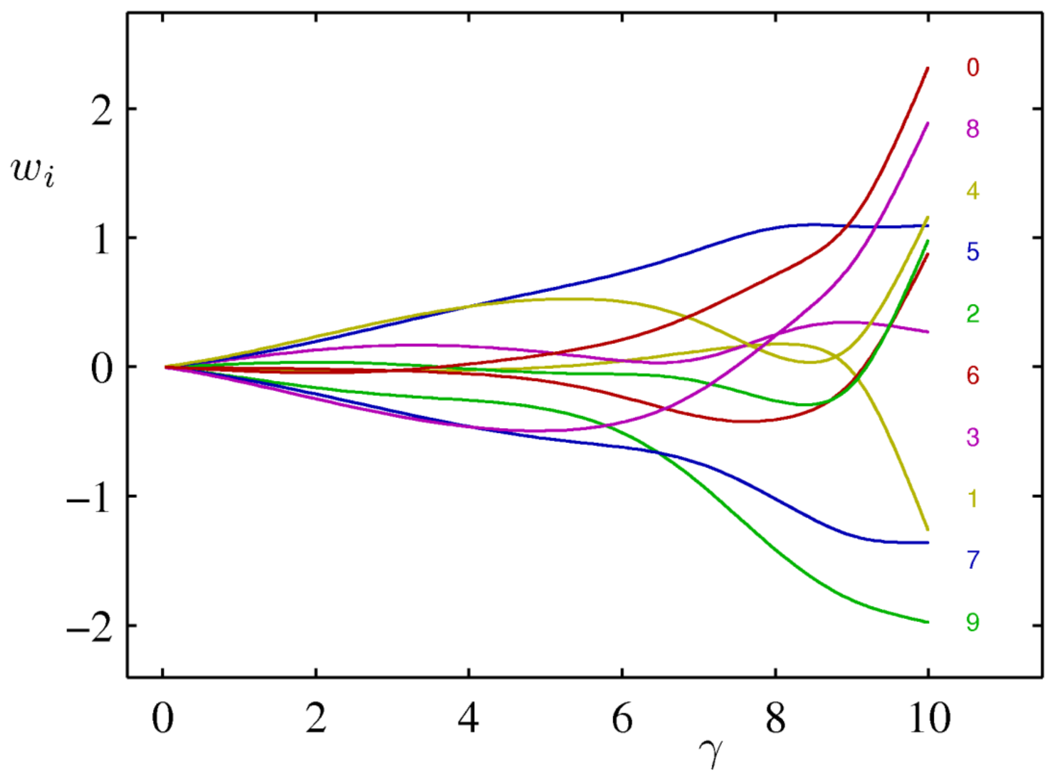
Effective Number of Parameters (3)*

- Example: sinusoidal data, 9 Gaussian basis functions, $\beta = 11.1$ (true value β*).



Effective Number of Parameters (4)*

- Example: sinusoidal data, 9 Gaussian basis functions, $\beta = 11.1$ (true value β*).



$\infty > \alpha \geq 0$

$0 \leq \gamma \leq 10$

Effective Number of Parameters (5)*

- In the limit $N \gg M$, $\gamma = M$ and we can consider using the easy-to-compute approximation

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

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

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

Limitations of Fixed Basis Functions

- M basis function along each dimension of a D-dimensional input space requires M^D basis functions: the curse of dimensionality.
- In later chapters, we shall see how we can get away with fewer basis functions, by choosing these using the training data.