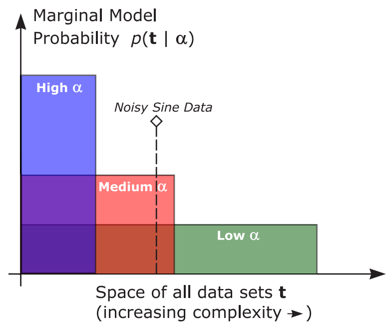
# Bayesian Inference: Principles and Practice

## 2. Bayesian Inference: Marginalisation

Mike Tipping




Microsoft Research Cambridge, UK

# Marginalisation

- It would be incorrect to assume that since the *maximum a posteriori* (MAP) and penalised least-squares (PLS) estimates are equivalent, the Bayesian framework is simply a probabilistic re-interpretation of classical methods

- This is not the case!

- The distinguishing element of Bayesian methods is *marginalisation*, where we attempt to integrate out all ‘nuisance’ variables

- As we will now see, this is a powerful component of the Bayesian framework

# Lecture 2: Overview

- Marginalisation: the Bayesian way to make predictions
- Approximate Bayesian prediction for our example model
- The marginal likelihood and Ockham's razor
- Bayesian model selection

# Making Predictions

- Consider, having 'learned' from the training values $\mathbf{t}$, how we make a prediction for data $t_*$ given a new input datum $x_*$:

| Framework | Learned Quantity | Prediction |
|---|---|---|
| Classical | $\mathbf{w}_{PLS}$ | $y(x_*; \mathbf{w}_{PLS})$ |
| MAP Bayesian | $p(\mathbf{w}\|\mathbf{t}, \alpha, \sigma^2)$ | $p(t_*\|\mathbf{w}_{MAP}, \sigma^2)$ |
| True Bayesian | $p(\mathbf{w}\|\mathbf{t}, \alpha, \sigma^2)$ | $p(t_*\|\mathbf{t}, \alpha, \sigma^2)$ |

- Where, following the 'true Bayesian' way, we *marginalise* to obtain:

$$p(t_*|\mathbf{t}, \alpha, \sigma^2) = \int p(t_*|\mathbf{w}, \sigma^2)\, p(\mathbf{w}|\mathbf{t}, \alpha, \sigma^2)\, d\mathbf{w}$$

- The *predictive distribution* $p(t_*|\mathbf{t}, \alpha, \sigma^2)$ incorporates our uncertainty over the weights $\mathbf{w}$ by taking all likely values, having seen $\mathbf{t}$, into account

# The General Bayesian Predictive Framework

- In general, for any model, if we wish to predict $t_*$ given some training data $\mathbf{t}$, what we really, really want is: $p(t_*|\mathbf{t})$

- So far, we've only placed a prior over the weights $\mathbf{w}$ — to be truly, truly, Bayesian, we should define $p(\alpha)$, a *hyperprior*, and $p(\sigma^2)$

- Then the full posterior over 'nuisance' variables becomes:

$$p(\mathbf{w}, \alpha, \sigma^2|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\alpha)p(\alpha)p(\sigma^2)}{p(\mathbf{t})}$$

- The highlighted normalising factor is called the *marginalised likelihood*:

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\alpha)p(\alpha)p(\sigma^2)\, d\mathbf{w}\, d\alpha\, d\sigma^2$$

  and is nearly always analytically intractable!

- Nevertheless, as we'll soon see, $p(\mathbf{t})$ is a very useful quantity

# Practical Bayesian Prediction (1)

- So, full Bayesian inference in our model would be:

$$p(t_*|\mathbf{t}) = \int p(t_*|\mathbf{w}, \sigma^2)\, p(\mathbf{w}, \alpha, \sigma^2|\mathbf{t})\; d\mathbf{w}\; d\alpha\; d\sigma^2$$

- We can't compute either $p(\mathbf{w}, \alpha, \sigma^2|\mathbf{t})$ or $p(t_*|\mathbf{t})$ analytically

- Procedure:

  1. Perform analytically computable integrations

  2. Approximate remainder, perhaps by:
     - Type-II maximum likelihood
     - Laplace's method
     - Variational techniques
     - Sampling

# Practical Bayesian Prediction (2)

- Ideally, we desire the full posterior $p(\mathbf{w}, \alpha, \sigma^2|\mathbf{t})$, which can be written as:

$$p(\mathbf{w}, \alpha, \sigma^2|\mathbf{t}) \equiv p(\mathbf{w}|\mathbf{t}, \alpha, \sigma^2)\, p(\alpha, \sigma^2|\mathbf{t})$$

- First term is our earlier weight posterior: $p(\mathbf{w}|\mathbf{t}, \alpha, \sigma^2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- Second term $p(\alpha, \sigma^2|\mathbf{t})$ we will approximate by a $\delta$-function at its mode. *i.e.* we find "most probable" values $\alpha_{\mathrm{MP}}$ and $\sigma^2_{\mathrm{MP}}$ which maximise:

$$p(\alpha, \sigma^2|\mathbf{t}) = \frac{p(\mathbf{t}|\alpha, \sigma^2)\, p(\alpha)\, p(\sigma^2)}{p(\mathbf{t})}$$

- If we (sensibly) assume flat, *uninformative*, priors over $\log \alpha$ and $\log \sigma$, then we equivalently maximise $p(\mathbf{t}|\alpha, \sigma^2)$ — "Type-II maximum likelihood"

# Practical Bayesian Prediction (3)

- Having found $\alpha_{\text{MP}}$ and $\sigma^2_{\text{MP}}$, our approximation to the predictive distribution is:

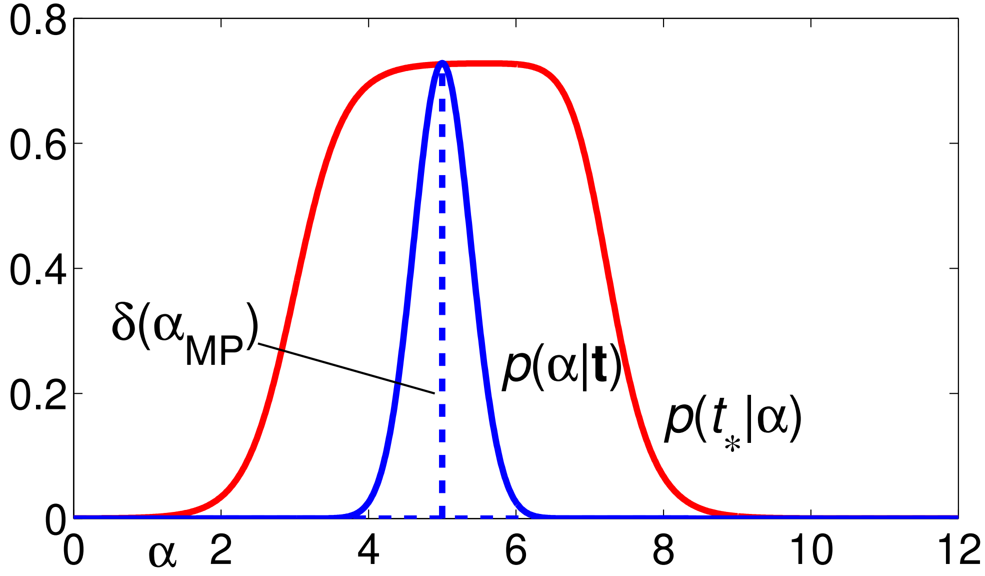
$$\int p(t_*|\mathbf{t}) = \int p(t_*|\mathbf{w},\sigma^2)\; p(\mathbf{w}|\mathbf{t},\alpha,\sigma^2)\; p(\alpha,\sigma^2|\mathbf{t})\; d\mathbf{w}\; d\alpha\; d\sigma^2$$

$$\approx \int p(t_*|\mathbf{w},\sigma^2)\; p(\mathbf{w}|\mathbf{t},\alpha,\sigma^2)\; \delta(\alpha_{\text{MP}},\sigma^2_{\text{MP}})\; d\mathbf{w}\; d\alpha\; d\sigma^2$$

$$= \int p(t_*|\mathbf{w},\sigma^2_{\text{MP}})\; p(\mathbf{w}|\mathbf{t},\alpha_{\text{MP}},\sigma^2_{\text{MP}})\; d\mathbf{w}$$

- Note that we don't require that $p(\alpha,\sigma^2|\mathbf{t}) \approx \delta(\alpha_{\text{MP}},\sigma^2_{\text{MP}})$ but:

True $p(t_*|\mathbf{t}) = 0.727$
Approximation $p(t_*|\alpha_{\text{MP}}) = 0.725$



# Practical Bayesian Prediction (4)

- Recall that $p(\mathbf{w}|\mathbf{t}, \alpha_{\text{MP}}, \sigma^2_{\text{MP}}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, from which the approximate predictive distribution is:

$$p(t_*|\mathbf{t}) \approx \int p(t_*|\mathbf{w}, \sigma^2_{\text{MP}})\, p(\mathbf{w}|\mathbf{t}, \alpha_{\text{MP}}, \sigma^2_{\text{MP}})\, d\mathbf{w} = N(\mu_*, \sigma^2_*)$$

  with:

$$\mu_* = y(x_*; \boldsymbol{\mu})$$

$$\sigma^2_* = \sigma^2_{\text{MP}} + \mathbf{f}^{\text{T}} \boldsymbol{\Sigma} \mathbf{f}$$

  where $\mathbf{f} = [\phi_1(x_*), \ldots, \phi_M(x_*)]^{\text{T}}$

- Intuitively:
  - the mean predictor $\mu_*$ is the model function evaluated with the posterior mean weights (the same as the MAP prediction)
  - the predictive variance $\sigma^2_*$ is the sum of variances associated with both the noise process and the uncertainty of the weight estimates

# Marginal Likelihood

- To find $\alpha_{\text{MP}}$ and $\sigma^2_{\text{MP}}$ we maximise the "marginal likelihood" $p(\mathbf{t}|\alpha, \sigma^2)$

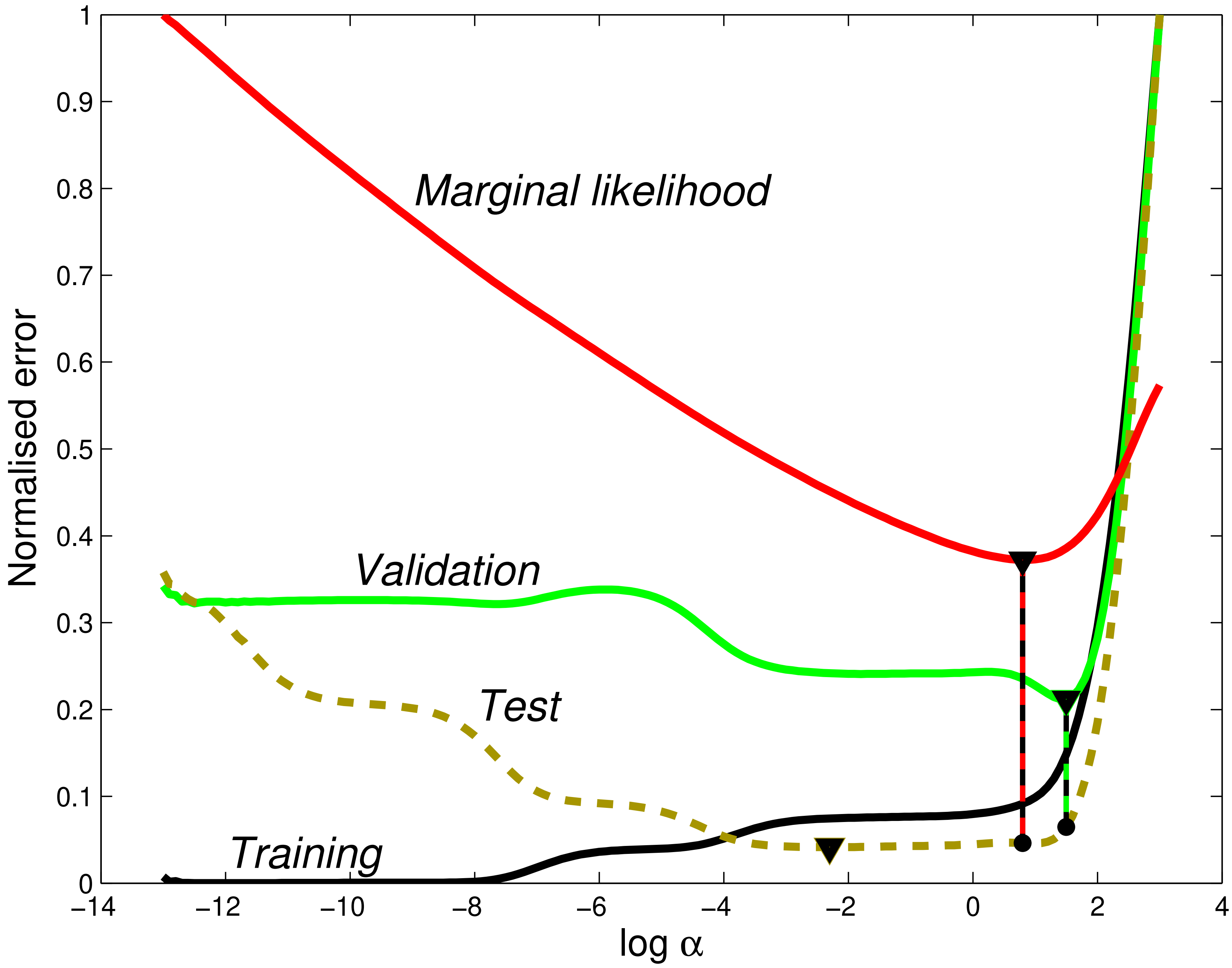
- This is given by:

$$p(\mathbf{t}|\alpha, \sigma^2) = \int p(\mathbf{t}|\mathbf{w}, \sigma^2)\, p(\mathbf{w}|\alpha)\, d\mathbf{w}$$

$$= (2\pi)^{-N/2}|\sigma^2\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathsf{T}}|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{t}^{\mathsf{T}}(\sigma^2\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathsf{T}})^{-1}\mathbf{t}\right\}$$

- This is a Gaussian distribution over the single $N$-dimensional dataset vector $\mathbf{t}$

- Note: we can use *all the data* to directly determine $\alpha_{\text{MP}}$ and $\sigma^2_{\text{MP}}$ — we don't need to reserve a separate data set to validate their values



Marginal likelihood
Validation
Test
Training
Normalised error
log α
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
−14
−12
−10
−8
−6
−4
−2
0
2
4

# The Bottom Line

- *Using only 15 examples and no validation data*, the Bayesian approach for setting $\alpha$ finds a closer model to the 'truth':

| | **Classical** | **Bayesian** | **Truth** |
|---|---|---|---|
| **Error** | 2.33 | 1.66 | 1.49 |

- The marginal likelihood criterion successfully rejects models that are either too simple or too complex — this is "Ockham's razor", a popular Bayesian concept

# Ockham's Razor

- In the fourteenth century, William of Ockham proposed:

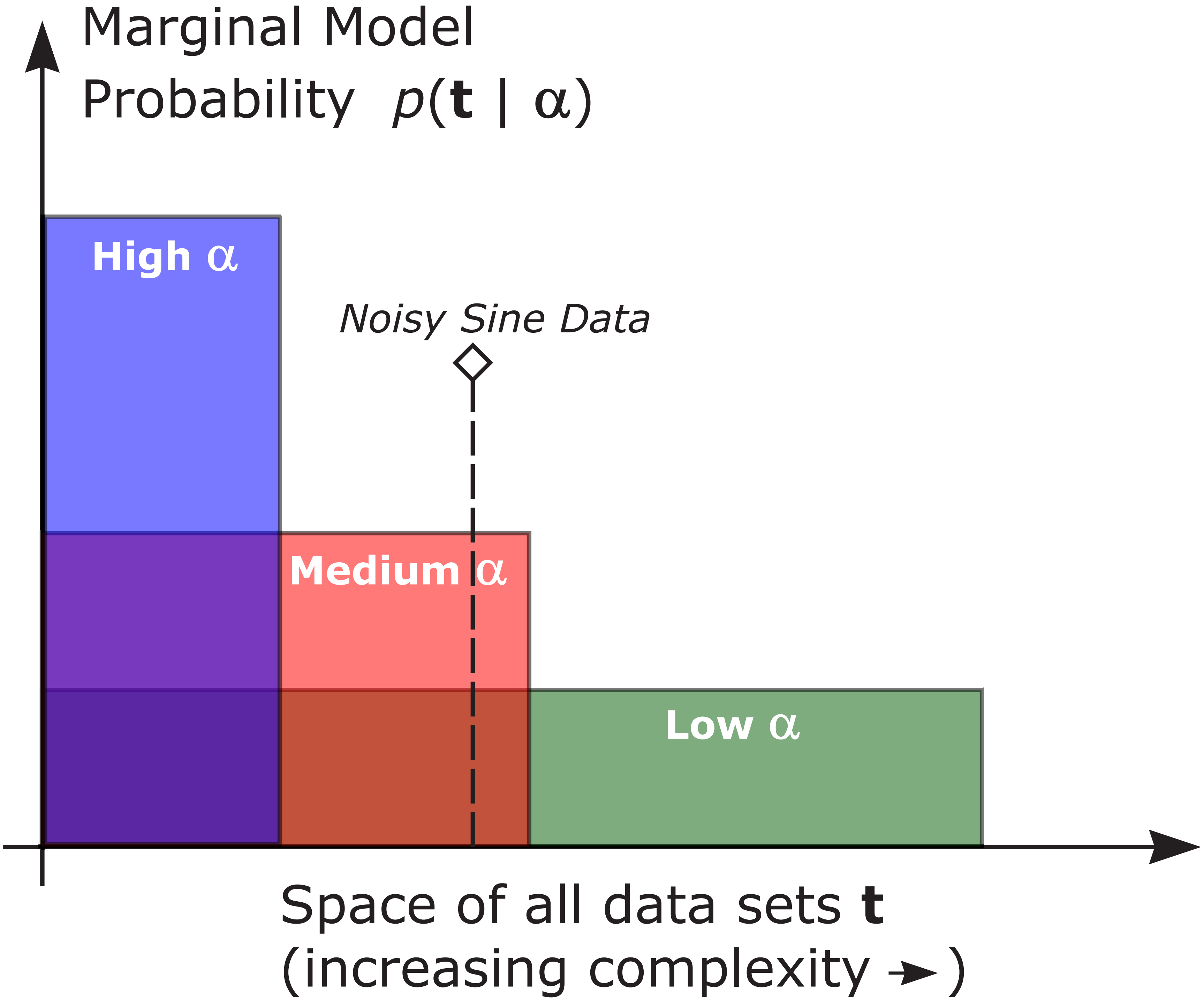
  "*Pluralitas non est ponenda sine neccesitate*"

  which literally translates as "entities should not be multiplied unnecessarily"

- In the context of machine learning, this translates as "models should be no more complex than is sufficient to explain the data"

- The Bayesian procedure is effectively implementing "Ockham's Razor" by assigning lower probability *both* to models that are too simple *and* too complex

- Why is an intermediate value of $\alpha$ preferred?



Marginal Model
Probability $p(\mathbf{t} \mid \alpha)$
High α
Noisy Sine Data
Medium α
Low α
Space of all data sets **t**
(increasing complexity →)

# Model Selection

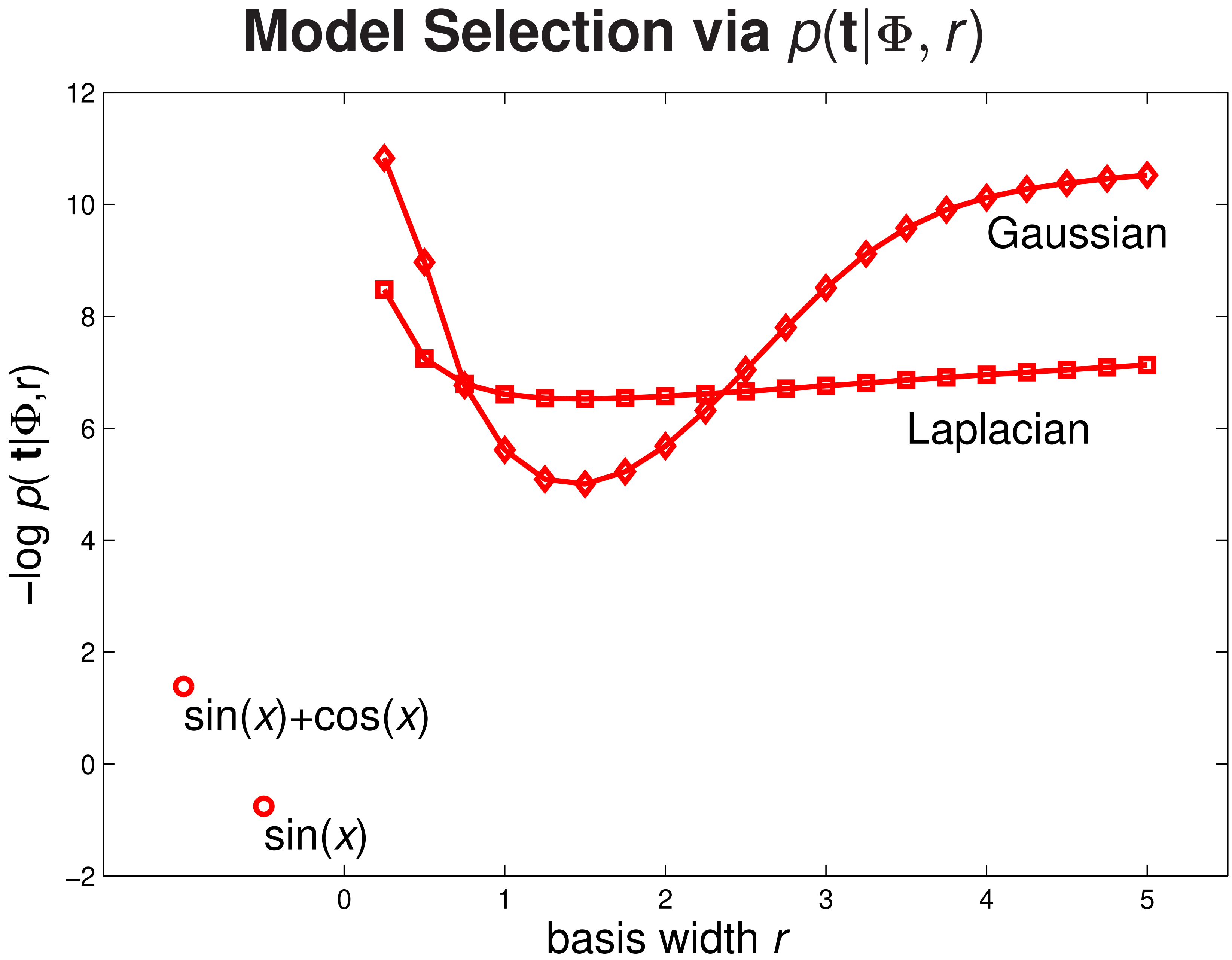
- Our models are also conditioned on other variables we have so far ignored: the choice of basis set $\Phi$ and, for our Gaussian RBF, the width parameter $r$

- We should define priors $P(\Phi)$ and $p(r)$, and integrate out those variables when making predictions

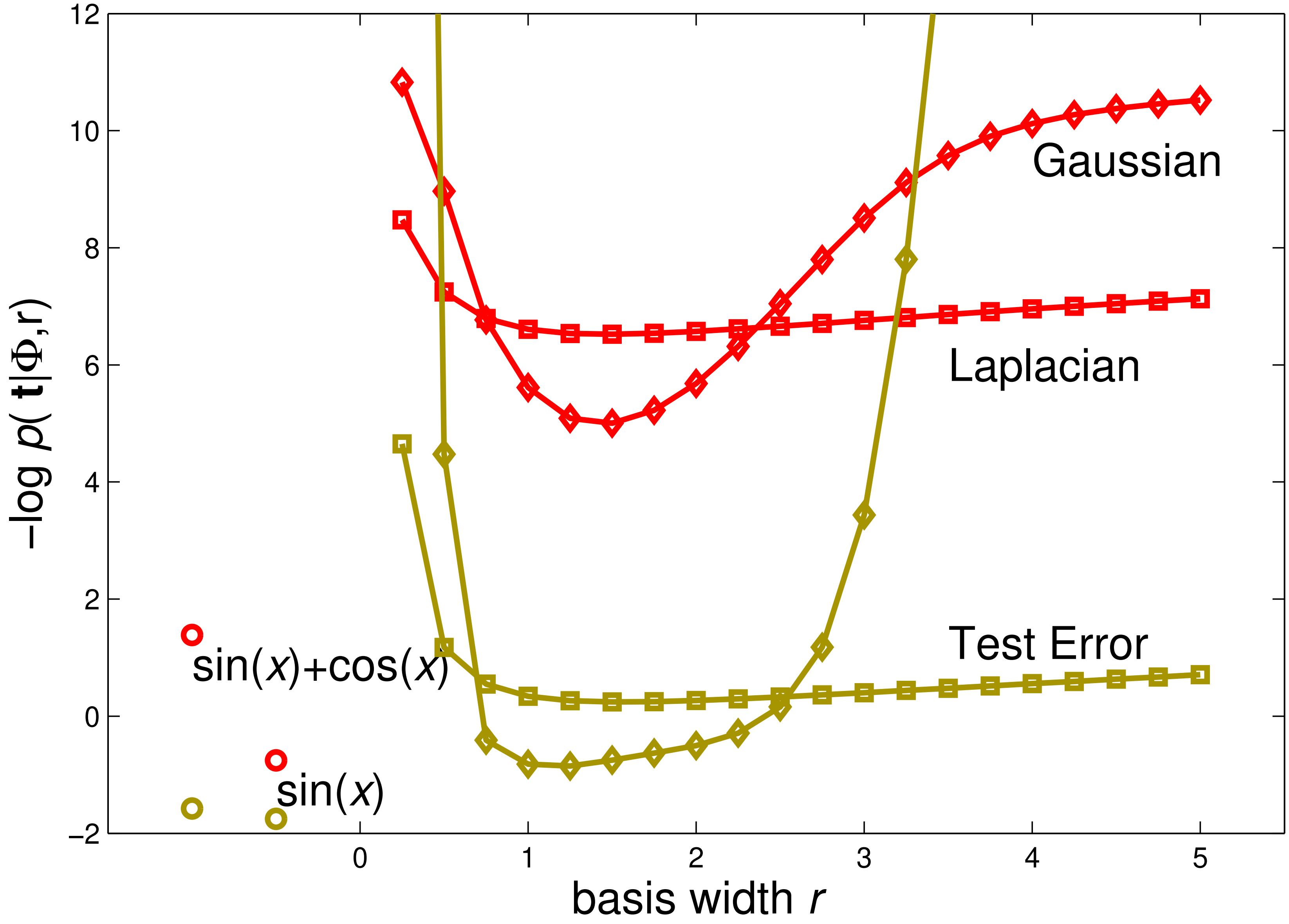
- More practically, we could use $p(\mathbf{t}|\Phi, r)$ as a criterion for *model selection*

- For this model, it is feasible to integrate out $\alpha$ and $\sigma^2$ numerically

- We compute the integral $p(\mathbf{t}|\Phi, r) = \int p(\mathbf{t}|\alpha, \sigma^2, \Phi, r)\, p(\alpha)\, p(\sigma^2)\, d\alpha\, d\sigma^2$ by sampling log-uniformly from $\alpha \in [10^{-12}, 10^{12}]$ and $\sigma \in [10^{-4}, 10^{0}]$


Model Selection via $p(\mathbf{t}|\Phi, r)$
Gaussian
Laplacian
sin(x)+cos(x)
sin(x)
$-\log p(\mathbf{t}|\Phi,r)$
basis width $r$
12
10
8
6
4
2
0
−2
0
1
2
3
4
5

# Correlation between $p(\mathbf{t}|\Phi, r)$ and test error



# The Story So Far...

- Marginalised likelihoods within the Bayesian framework allow us to:
  - estimate hyperparameters
  - choose between models

  such that we can determine 'good' models without needing validation data

- Additional benefits of the Bayesian framework are:
  - It is straightforward to estimate the noise variance
  - We can sample from both prior and posterior models of the data
  - The exact parameterisation of the model is irrelevant
  - Lecture 3: incorporation of other priors of interest (*i.e.* sparsity) and functional optimisation of hyperparameters