# List of Figures


1.1 Abstraction of the data modelling process . . . . . 9

2.1 Where Bayesian inference fits into the data modelling process . . . . . 15
2.2 Why Bayes embodies Occam's razor . . . . . 16
2.3 The Occam factor . . . . . 20
2.4 How the best interpolant depends on $\alpha$ . . . . . 24
2.5 Choosing $\alpha$ . . . . . 27
2.6 Good and bad parameter measurements . . . . . 29
2.7 The Evidence for data set X . . . . . 33
2.8 Data set 'Y,' interpolated with splines . . . . . 35
2.9 Typical samples from the prior distributions of six models . . . . . 36

3.1 Typical neural network output . . . . . 48
3.2 Data error versus number of hidden units . . . . . 49
3.3 Test error versus number of hidden units . . . . . 49
3.4 Test error vs. data error . . . . . 50
3.5 Log Evidence for solutions using the first regulariser . . . . . 50
3.6 The number of well–determined parameters . . . . . 51
3.7 Data misfit versus $\gamma$ . . . . . 51
3.8 Log Evidence versus Test error for the first regulariser . . . . . 52
3.9 Comparison of two test errors . . . . . 52
3.10 The three classes of weights under the second prior . . . . . 55
3.11 Log Evidence versus number of hidden units for the second prior . . . . . 55
3.12 Log Evidence for the second prior versus test error . . . . . 56

4.1 Demonstration of total and marginal information gain . . . . . 70

5.1 Approximation to the moderated probability . . . . . 76
5.2 Comparison of most probable outputs and moderated outputs . . . . . 77
5.3 Moderation is a good thing! . . . . . 78
5.4 Test error versus data error . . . . . 79
5.5 Test error versus evidence . . . . . 80
5.6 Correlation between Test error and evidence as the amount of data varies . 80
5.7 Demonstration of expected mean marginal information gain . . . . . 84

**7 Postscript** **89**
7.1 The closed hypothesis space . . . . . 89
7.2 For approximation, are probabilities relevant? . . . . . 90
7.3 Having to make too much explicit . . . . . 91
7.4 An alternative interpretation of weight decay . . . . . 91
7.5 Future tasks, open problems . . . . . 92

**Bibliography** **94**

# Contents


**1 Summary** **8**
- 1.1 The need for Occam's razor . . . 8
- 1.2 What is Bayesian modelling? . . . 8
- 1.3 What are neural networks and why do they need Occam's razor? . . . 11

**2 Bayesian Interpolation** **14**
- 2.1 Data modelling and Occam's razor . . . 14
- 2.2 The evidence and the Occam factor . . . 17
- 2.3 The noisy interpolation problem . . . 21
- 2.4 Selection of parameters $\alpha$ and $\beta$ . . . 23
- 2.5 Model comparison . . . 30
- 2.6 Demonstration . . . 31
- 2.7 Conclusions . . . 39

**3 A Bayesian Framework for Backprop Networks** **41**
- 3.1 The gaps in backprop . . . 41
- 3.2 Review of Bayesian regularisation and model comparison . . . 45
- 3.3 Adapting the framework . . . 46
- 3.4 Demonstration . . . 48
- 3.5 Discussion . . . 57

**4 Information–based objective functions** **60**
- 4.1 Introduction . . . 60
- 4.2 Choice of information measure . . . 62
- 4.3 Maximising total information gain . . . 64
- 4.4 Maximising information about the interpolant in a region of interest . . . 65
- 4.5 Maximising the discrimination between two models . . . 67
- 4.6 Demonstration and Discussion . . . 68
- 4.7 Conclusion . . . 69

**5 The Evidence applied to Classification** **72**
- 5.1 Introduction . . . 72
- 5.2 Every classifier should have two sets of outputs . . . 74
- 5.3 Evaluating the evidence . . . 78
- 5.4 Active learning . . . 81
- 5.5 Discussion . . . 83

**6 Inferring an input–dependent noise level** **86**

## Abstract

The Bayesian framework for model comparison and regularisation is demonstrated by studying interpolation and classification problems modelled with both linear and non–linear models. This frameworkquantitatively embodies 'Occam's razor.' Over–complex and under–regularised models are automatically inferred to be less probable, even though their flexibility allows them to fit the data better.

When applied to 'neural networks,' the Bayesian framework makes possible (1) objective comparison of solutions using alternative network architectures; (2) objective stopping rules for network pruning or growing procedures; (3) objective choice of type of weight decay terms or additive regularisers; (4) on–line techniques for optimising weight decay magnitude; (5) a measure of the effective number of well–determined parameters in a model; (6) quantified estimates of the error bars on network parameters and on network output. In the case of classification models, it is shown that the careful incorporation of error bar information into a classifier's predictions yields improved performance.

Comparisons of the inferences of the Bayesian framework with more traditional cross–validation methods help detect poor underlying assumptions in learning models.

The relationship of the Bayesian learning framework to 'active learning' is examined. Objective functions are discussed which measure the expected informativeness of candidate data measurements, in the context of both interpolation and classification problems.

The concepts and methods described in this thesis are quite general and will be applicable to other data modelling problems whether they involve regression, classification or density estimation.

## Acknowledgements

During the last three years, my work has benefitted greatly from discussions with Ron Benson, John Bridle, Peter Cheeseman, Sidney Fels, Steve Gull, Andreas Herz, John Hopfield, Doug Kerns, Allen Knutsen, David Koerner, Mike Lewicki, Tom Loredo, Steve Luttrell, Ronny Meir, Ken Miller, Marcus Mitchell, Radford Neal, Steve Nowlan, David Robinson, Ken Rose, Sibusiso Sibisi, John Skilling, Haim Sompolinsky and Nick Weir. The comments of referees on chapter 4 were also helpful.

I especially thank my advisor, John Hopfield, for his support, criticism and advice.

I am very grateful to Brooke Anderson, Dawei Dong, Brian Fox and Tom Tromey for their expert management of the Hopfield and CNS computers.

I would like to thank Dr. R. Goodman and Dr. P. Smyth for funding my trip to Maxent 90, where I learnt the final tools needed for this research.

This work was supported by a Caltech Fellowship and a Studentship from SERC, UK.

I hope that the ideas in this dissertation will only be used towards peaceful ends.

## Address from January 1st 1992

Darwin College,
Cambridge,
CB3 9EU, U.K.
mackay@eng.cam.ac.uk

Typeset with LaTeX.

©1992
David J.C. MacKay
All rights reserved

# Bayesian Methods for Adaptive Models

Thesis by

David J.C. MacKay

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1992
(Submitted December 10, 1991)