The Statistical Mechanics of Image Restoration

Jonathan Michael Pryce

PhD
University of Edinburgh
1993

Dedication

To Dad—I only begin to understand all you've done for me.

Declaration

I declare that I composed this thesis myself. The work contained herein is my own, and was carried out in collaboration with Dr. A.D. Bruce.

[0] An abridged version of this work is presented in:
J.M. Pryce and A.D. Bruce. The Statistical Mechanics of Image Restoration.
To be submitted to *Journal of Physics A*.

Acknowledgements

It is not possible to express adequately the debt that I owe my supervisor Alastair Bruce. Without his encouragement this work would not have developed in the worthwhile directions that it did. His ludicrously high standards have been a source of much heartache, but they raised the quality of this thesis an order of magnitude above what might have been.

A special word of thanks to Martin Simmen who has devoted many hours to reading and commenting on the final draft, even taking it on holiday to the United States.

To all those, in 4408, the department, and the university, whom I have worked and played with these past four years. Hail and well met!

This work was partially supported by a studentship from the UK Science and Engineering Research Council.

Abstract

Image restoration is concerned with the recovery of an 'improved' image from a corrupted picture, utilizing a prior model of the source and noise processes. We present a Bayesian derivation of the posterior probability distribution, which describes the relative probabilities that a certain image was the original source, given the corrupted picture. The ensemble of such restored images is modelled as a Markov random field (Ising spin system). Using a prior on the density of edges in the source, we obtain the cost function of Geman and Geman via information theoretic arguments.

Using a combination of Monte Carlo simulation, the mean field approximation, and series expansion methods, we investigate the performance of the restoration scheme as a function of the parameters we have identified in the posterior distribution. We find phase transitions separating regions in which the posterior distribution is data-like, from regions in which it is prior-like, and we can explain these sudden changes of behaviour in terms of the relative free energies of metastable states. We construct a measure of the quality of the posterior distribution and use this to explore the way in which the appropriateness of the choice of prior affects the performance of the restoration. The data-like and prior-like characteristics of the posterior distribution indicate the regions of parameter space where the restoration scheme is effective and ineffective respectively.

We examine the question of how best to *use* the posterior distribution to prescribe a single 'optimal' restored image. We make a detailed comparison of two different estimators to determine which better characterizes the posterior distribution. We propose that the TPM estimate, based on the mean of the posterior, is a more sensible choice than the MAP estimate (the mode of the posterior), both in principle and in practice, and we provide several practical and theoretical arguments in support.

We then address the issue of parameter estimation from the corrupted picture alone. We apply the evidence formalism of Gull, Skilling and MacKay to the problem of making the 'optimal' choice of restoration parameters in the posterior. For the purposes of measuring the evidence by numerical simulation, we explore and develop the 'method of expanded ensembles' for free energy measurement, in the context of the Ising model. Ultimately our results suggest that parameters identified by the evidence framework provide effective priors, leading to optimal restoration, only to the extent that the priors are well matched to the processes they claim to represent.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 New Applications of Statistical Mechanics

Statistical Mechanics is a branch of theoretical physics distinguished by its wide applicability across such diverse fields as physics, astronomy, chemistry, materials science and biology. At an advanced level it deals with the interactions between the constituents of a large assembly and the cooperative phenomena that result. Traditionally, most of the work of this nature has been confined to the field of condensed matter physics where there are numerous phenomena to be investigated and experimental techniques have become increasingly sophisticated. In addition, the physical systems to be investigated may be simple enough to permit a theoretical analysis in many cases. As a result, condensed matter research has led the way in statistical physics methods and much of our understanding is derived from successes in this field.

Lately, however, statistical physicists have been branching into many other subjects that lend themselves to a statistical mechanics approach. This is nowhere more apparent than in complexity research [64, 116]: recent years have seen the emergence of 'complex systems' as a distinct field of scientific research in its own right. Such systems all exhibit what may be described as surprising or unusual behaviour that is in some way a property of the system as a whole—treat the constituents in isolation and the unexpected behaviour vanishes. This collective behaviour has long been recognized in condensed matter research. Now such expertise is being applied extensively in subjects ranging from neurobiology and network computation, through fluid turbulence and climate modelling, to population growth and economics. In particular, the study of network computation models [8, 56] has found its way into the bread and butter research of many theoretical physics groups, exploiting the close similarity between such models and lattice models of magnetic systems in condensed matter.

In the rest of this introduction we briefly review some of the notable successes of statistical mechanics applied to network computation before moving on to examine the background of the image restoration problem. We are then in a position to expound this thesis: that statistical mechanics can significantly improve our *understanding* of image restoration.

The Hopfield Model

The Hopfield model [58] is a recurrent network of (McCulloch and Pitts) binary threshold units [80] analogous to a simple model of a magnet, with the 'units' playing the role of spins and the 'synaptic connections' the role of the spin-couplings. Such a physical model has an energy function defining a complex energy surface with many local minima. Hopfield was the first to realize the potential of this energy function, identifying the local minima with the stored states of an auto-associative memory [2], where the memory state is recalled by partially specifying the contents.

There is a simple algorithm after Hebb [55] that will sculpt the energy surface in such a way as to store a desired set of patterns. But we must expect that saturation problems will occur if we try to store too many patterns in the same network. There are other *spurious* states in addition to the desired memories: mixture states [4] which are a linear combination of the desired states; and entirely unrelated states [5], given the name **spin glass** states because of a close correspondence to spin glass models in statistical mechanics.

Further physical insights can be obtained by using stochastic units [57, 96]. The random noise in such units corresponds to the thermal fluctuations experienced by spins at finite temperature. Using a statistical mechanics approximation called **mean field theory** Amit *et al.* [5, 7] obtained the phase diagram of the Hopfield model, which identifies the different **phases** of the model, distinct regions where there are qualitative differences in the behaviour of the associative memory. The phase diagram is the key to understanding any statistical mechanics system.

The Gardner Theory

Another successful application of statistical mechanics is the Gardner [32, 33] calculation of the capacity of the simple perceptron [104]. How many randomly chosen input-output patterns can we expect to store successfully in a network of a given size? The basic idea is to consider the fraction of weight space (the space of all possible connection strengths) that implements a particular input-output function. The expression we obtain resembles a statistical mechanics **partition function**, and the quenched average over the patterns is obtained using the **replica** method. Building on Gardner's methods [34, 47, 48, 94] it is possible to calculate the ability of the network to *infer* a rule from a set of examples. For discrete networks a **phase transition** from poor to satisfactory generalization is observed when a critical number of examples is reached.

The Statistical Mechanics of Learning

Learning theory continues to attract the attention of statistical physicists. The training examples constitute the **quenched disorder** of the problem, and it is possible to define a **self-averaging** analogue of the free energy. A whole battery of statistical mechanics techniques have been brought to bear on the analysis of learning algorithms: see e.g. [120, 124]. For a binary perceptron implementing a realizable rule, learning has been completely and exactly solved by these methods.

The learning of an unrealizable rule (where the network architecture is too simple to be able to implement the complexity of the rule, even given an infinite training set) leads to a phenomenon analagous to frustration in spin glasses. This problem is currently being tackled with one of the latest additions to the statistical physicist's armoury: **replica symmetry breaking** [89].

1.2 Image Analysis

Data reconstruction—the inference of underlying structure from experimental data—is one of the key problems in modern science. In general we have an observed function g that has been generated by a (possibly unknown) process from a function f. The task is to estimate the function f, given the observed function $g = \kappa f$. Before proceeding we must ask what other information we have. Do we know the transfer function κ accurately, or at all? Do we have any *a priori* information about the original function f? Depending on the answers to these questions, a multitude of techniques may be applied. The field is enormous: the books [42, 105, 115] offer an introduction to the subject.

1.2.1 The Inverse Problem

In image analysis, the observed function is the image: a two-dimensional array of picture elements, or pixels. The image *synthesis* problem is to determine this observable image g given a complete representation of the true scene f and the imaging process κ. This direct problem is encountered in computer graphics applications such as ray-tracing [41], while the study of the practical issues involved lies in the domain of experimental optics.

The *inverse* problem is fundamentally more difficult. The observed image is generally an incomplete representation of the scene or object that we are viewing. The task may be to extract information about the scene from the image (the work of computer vision applications), or to remove blur or other degradation from the image (the image restoration problem that is the focus of this thesis).

Such inverse problems are made difficult by the information loss inherent in the image synthesis process: we collapse the continuous dimensions of the physical world onto the few degrees of freedom of a sampled and quantized image. The dimensions of the true scene that are represented in the image constitute the measurement space. Orthogonal to this is the **null space**—measurements of attributes that lie in the null space will yield no information [51]. Thus there will be numerous true scenes f, often quite dissimilar, that could be responsible for a given observed image g. We say that the problem is **ill-posed** [50].

Biological Motivation

Given the complexity of the problem it is quite remarkable that animals seem able to overcome many of these difficulties. The success of biological systems over conventional signal-processing techniques is often attributed to the massive parallelism of the brain, and certainly a large part of the mammalian brain is given over to visual processing. We are thus able, in ways that are little understood, to process an enormous quantity of information from a multitude of sources. We integrate this with our prior knowledge of the world we live in to interpret an image in a mostly unambiguous fashion (although there are many examples that fool the eye-brain combination into an incorrect interpretation). Artificial systems rarely even approach the success of biological ones, and there are always lessons to be learnt from nature's example.

The explicit use of prior knowledge of the world is one such example. Only the simplest maximum likelihood methods [51, 61] fail to assume some prior knowledge of the true scene, but this is seldom acknowledged. The prior model tells us what we might *expect* to see in the image, and mathematically this is expressed as a probability distribution. What we *actually* see is a representation of the true scene,modified by the observation process or noise in the environment. These processes too can be modelled by a probability distribution. Bayesian inference then allows us to write down a posterior distribution for the possible true scenes, based upon the prior model, the observation process, and the image actually observed. Bayesian inference is not new in image analysis (see e.g. [49, 61, 90, 102]), but there has been renewed interest in prior models based on discrete Markov random fields [16, 21, 36, 37, 85, 113].

Markov Random Fields

Markov Random Fields on finite lattices [127, 128] are just one approach to the implementation of a prior model. The idea is to model each pixel in the image as an element of a random field. A Markov process is a prescription for updating states where the transition probability is independent of the previous history of the system. In a Markov Random Field, or MRF, the update of an element depends only upon the current state of some neighbourhood of local sites. Thus they provide a flexible mechanism for modelling spatial dependence. Our interest in MRFs arises through their equivalence to the Gibbs distribution in statistical mechanics [53, 73].

There have been a number of successful applications of MRFs to image analysis problems.

- Texture Synthesis [22, 54]: an attempt to reproduce the regularity in the visual appearance of some materials by modelling a local spatial process.

- Classification of Satellite Data [117]: the land use type of an area is assessed using primarily local, contextual information.
- Surface Reconstruction and Boundary Detection [36]: MRFs are useful for representing *unobserved* image attributes such as discontinuities.

Monte Carlo Methods

Theoretical analysis of MRFs usually proves difficult especially if the model includes the sort of line processes introduced in [36]. However Monte Carlo methods (see e.g. [18]) are ideally suited to exploring such Markov processes. The stochastic relaxation algorithms used are themselves Markov processes [52] and this has led to much numerical simulation work, with only a bare minimum of theoretical analysis [31]. Such simulations may rely heavily on the use of parallel computing: Markov processes, being essentially local, are ideally suited to parallel implementation. This is another nod in the direction of the biological solution—massive parallelism. More importantly it presents the possibility of genuinely parallel hardware implementations (e.g. a silicon retina [87, 110]).

1.2.2 Image Restoration

Image restoration, the reconstruction of an image from incomplete and noisy data, is the particular aspect of image analysis that we investigate in this thesis. It forms a subset of the more general inverse problems described previously since the the functions f and g lie in the same space—they are *both* two-dimensional images. The true image f is operated on by some noise process to give a corrupted image g. The restoration problem is to estimate the true image from the corrupted version (see e.g. [60, 105]).

The Pseudo Inverse

In the simplest case we know the form of the corruption process. We search the space of all possible uncorrupted images, calculating for each one what image would result from the corruption process. We compare this calculated image with the actual corrupted image and find the best match (by finding the least squared error). In effect we are trying to find the inverse of the transfer function or corruption process, and we call the inverse found by this least squares method the 'pseudo inverse' [42]. Due to the ill-posed nature of the problem, small perturbations in the corrupted image g will in general give rise to unacceptably large changes in the solution. We cannot choose between such wildly differing solutions without some prior knowledge of what we expect the solution to be. Attempts to mitigate the ill-posed nature of this problem by altering the modelled transfer function κ are the subject of **regularization theory** [98, 119].

Bayesian Image Restoration

The image restoration problem is ill-posed because the data is incomplete: there is insufficient information to determine the source image uniquely from the data. Even if the transfer function is modelled perfectly, any inverse function will be unable to represent the null space [51] of the source image. The essence of the Bayesian approach is the assumption that the image to be reconstructed may be modelled as a random selection from an identifiable ensemble of similar images. If there *are* many source images we may be able to estimate this ensemble simply from a large number of previous observations. Otherwise we may construct an ensemble of images that satisfies certain prior beliefs about the source. Provided we model this prior adequately, the Bayesian method supplies a meaningful estimate of the null space component of the source. This is the approach we take in this thesis, and we defer further discussion to Chapter 2.

Maximum Entropy

The term 'Maximum Entropy' means many things to many people. In its most general form it is an information-theoretic method for calculating the most probable prior distribution given some limited information about the distribution. Indeed we will make use of 'maximum entropy' in exactly this way in the next chapter. However, in the engineering literature it refers to a particular model of image restoration first proposed by Frieden [28], and described by Skilling [111] as 'Classic MaxEnt.'

Classic MaxEnt is equivalent to Bayesian inference with the simplest possible prior [51, 121]: that the original image is formed by randomly distributing units of intensity across the frame subject only to a constraint on the total intensity of the final image. The noise process is a similar random distribution superimposed on this image, again subject to a constraint on the total noise present. The estimate we obtain is the most random estimate consistent with the presumed overall intensity of the original image and the noise.

This method has been enormously successful for some problems in image restoration—notably for images of randomly pulsed objects, such as starfields [29, 45]. Some proponents insist, with an almost evangelical fervour, that MaxEnt is the *only* method of regularization that can be rationally supported—based no doubt on the information theoretic aspects—and decry the use of any other prior whatsoever.

As Skilling shows [111], the MaxEnt prior is in fact evaluated relative to a model of sorts (the measure in the entropy integral—see [108]). In Classic MaxEnt this prior is taken to be completely flat; no prior knowledge is assumed. However successful this method may be for randomly pulsed objects, it gives very poor results [46] on real images with spatial correlations.

This has led Gull [46], in what he calls 'New MaxEnt', to modify the model used in the entropy integral when the MaxEnt prior is determined. The model is no longer flat but is now able to account for spatial correlations as it is itself an image obtained by blurring some set of hidden variables. It is the functional form of the model blur that corresponds to the prior used in the Bayesian MRF approach. It is not clear what the hidden variables should be, but they are usually modelled as the corrupted image itself, or alternatively as the image resulting from an earlier restoration attempt. In Classic MaxEnt the prior is uniform across the image and simply models the mean intensity of the overall corrupted image. New MaxEnt calculates the mean intensity in the neighbourhood of each point in the image and uses this as the local model in the MaxEnt reconstruction at that point.

With this formulation of MaxEnt it is now possible to argue that the Bayesian MRF approach is a special case of MaxEnt [46], rather than the other way around [51, 121]. However, the modifications required to successfully restore images with spatial correlations cause New MaxEnt to lose much of the compelling simplicity of the original formulation that led to its espousal as the *only* axiomatic approach to image restoration.

Real World Problems

Although this thesis is not concerned with the kind of images and noise processes present in the real world, we hope that a better theoretical understanding will ultimately lead to improvements in current methods. To this end, we briefly review the kinds of problem encountered in real applications of image restoration [122].

It is not possible to say with any degree of certainty how much may be gained from attempts to improve a real image—nor what techniques will prove useful. In practice, much depends on the human experience and expertise available when attempting the restoration. Images are vetted to assess whether any useful restoration is possible and what techniques are likely to work best with respect to the specific question being asked of the image. Any answer will usually be phrased in terms of relative probabilities.

The common degradations encountered in, for example, video camera images are: non-linearities in the recording medium; simple random noise; motion blur; and a lack of resolution. It is usually sufficient to use statistical methods such as principal component analysis—the Karhunen-Loeve transform, see e.g. [71]—and any of a number of super resolution processing techniques [29] to increase the apparent resolution of the images. It is possible to factor out both random noise and motion blur by analysing the correlations in a temporal sequence of images.

1.3 Statistical Mechanics of Image Restoration

There is a method of scientific research, best described as phenomenological, where the outcome of experiments are merely noted, and the results used perhaps to interpolate the behaviour of other experiments. There is, however, little attempt to gain a deeper understanding of the processes that are taking place. Without such understanding it is almost impossible to *design* modifications to these processes with any confidence in the likely outcome: improvements result only from 'hit or miss' alterations.

Although such research is necessary to *begin* building our knowledge of a new field, and perhaps for speedy commercial implementation of new discoveries, for sustained progress to be achieved it must be balanced by a similar effort aimed at theoretical understanding. It appears that too much research falls between these two stools: the model studied is frequently inadequate for the basis of a commercial product, and yet is too complex to be properly understood. Hence, we get 'bandwagon' research: minor variations of established techniques are applied to a range of often similar problems. The outcome is reported, but there has been no real progress in our understanding.

This thesis is devoted to improving our understanding.

The literature on image restoration is enormous and, fragmented across many disciplines, the language is frequently alien. [References [38, 60] provide an introduction to the literature.] We make no attempt to compete with the level of complexity seen in many of the earlier references and concentrate on building an understanding of the simplest models.

1.3.1 The Work of Geman and Geman

In 1984 Geman and Geman [36] (hereafter referred to as GG) proposed a model of image restoration, based on Bayesian inference, that included priors not only for spatial coherence but also for the presence of discontinuities (the so-called line processes). The idea is to construct a MRF consisting of two processes, one accounting for the intensity values and the other for the discontinuities or edges. They enjoyed much success with this model and the paper initiated a huge amount of similar work [citations in more than 500 publications by 1992].

The model is unfortunately analytically intractable. In this thesis we retreat from some of the complexities of GG in order to make analytic calculation feasible. Such simplification allows us to address some issues that seem so far to have been neglected or, at best, poorly understood due to the lack of a systematic treatment.

1.3.2 Uncharted Territory

The equivalence of Markov random fields and the Gibbs distribution of statistical mechanics was made quite explicit in the original GG paper. This equivalence motivated the use of Monte Carlo methods from statistical physics in the simulation of MRFs and models of image restoration. To date, however, few have taken the next logical step, and applied the *analytic* methods of statistical mechanics to the problem of image restoration. This thesis fills that gap.

The Gibbs distribution describes the behaviour of the 'spins' that represent the result of the restoration, relative to a particular instance of the noise. This noise constitutes a **quenched** disorder. We want results for the general case, so we must average over this disorder; we calculate quenched averages. In statistical mechanics such quenched averages are derivatives of the free energy. We will establish analogues of the free energy for the restoration scheme which will allow us to calculate these quenched averages.

Phase Transitions in Hypothesis Space

Bayesian inference requires that we construct a prior model. We make some estimate of the mechanism that generated the original image, and the performance of the reconstruction scheme is sensitive to this choice of prior. Since we may not have 'good' information when we choose the prior, we analyse two distinct cases.

- If we can construct a prior model that accurately reflects the image generation process, how well can we do in this optimal case?
- What happens when, for whatever reason, we choose a prior that is nothing like the real process that generated the original image? How well can we do when the reconstruction scheme tries to solve the wrong problem?

In any practical application we may not know the correct prior to use, so a comparison of these two cases provides some insight into the success or failure of the restoration. Is there simply not enough information to do a good job of restoration, or have we chosen a fundamentally poor prior? These two questions recur many times in this work.

Whatever prior we choose, Bayesian inference yields a parametric result for the posterior distribution. It is well known that for some parameter choices the restoration process will yield complete nonsense (see e.g. [44, 84]), but it is not always clear why, or for what regions of parameter space this failure occurs. These different regions are the **phases** of the model. As we modify the parameters the success of the restoration scheme varies continuously. However, at particular points there is a phase transition where the performance of the restoration changes discontinuously. Our understanding of phase transitions in condensed matter allows explanation of these sudden changes in behaviour. The questions we address are:

- What is the sensitivity of the restoration to the choice of parameters? Where do we get optimal restoration? Where in parameter space does the method break down?
- In what ways are the answers to the above questions altered by an incorrect choice of prior?

We address these questions through the use of analytic methods and simulation. The restoration scheme is too complex to be susceptible to exact theoretical analysis, but we can make progress by the use of simplifications and approximations. We make use of a statistical mechanics technique called mean field theory in order to obtain an approximation to the phase diagram of the model. This mean field theory is the standard analytic method from statistical mechanics, used to calculate approximations to the order parameters that describe the overall performance of the restoration scheme. It is not to be confused with the so-called mean field technique used by Geiger and Girosi [35] which, like the renormalization group approach of Gidas [40], is a deterministic algorithm for generating the actual reconstructions. Both are useful efforts to find more efficient ways of generating the reconstructions using ideas from statistical mechanics, but there is no attempt to use these techniques to *understand* the results. The other statistical mechanics technique we bring to bear on the analysis of the restoration problem is the small coupling expansion. Like the mean field approximation this method cannot provide exact results, but within the small coupling regime it does provide further insight into the details of the restoration scheme.

The Optimal Estimator?

Frequently the choice of estimator—the final image that is used as the reconstruction—is not even recognized as an issue. In the majority of other cases alternative estimators may be mentioned, but results will only be presented for the authors' *favourite* estimate. Such choices are not usually justified beyond a statement that the chosen estimator gives *satisfactory* results. A notable exception to this is the report by Marroquin [84,85] which usefully restates some general results on optimal Bayesian estimators, see e.g. [1]. However, few seem to have taken notice.

GG, and much subsequent work, use the mode of the posterior distribution as the estimate of the original image. This maximum *a posteriori* (MAP) estimate seems initially reasonable because we consider it in everyday terms. It is the *single* image that was most likely to have generated the corrupted image given all of the information that has been included by the Bayesian calculation of the posterior distribution. However, this is not a case of everyday probabilities. The space of potential original images is enormous. To choose the single most probable image and discard all of the rest seems foolish. A more reasonable approach might be to perform some average over the space of possible images. Marroquin [84] shows that the mean and mode of the posterior distribution are the optimal Bayes estimators for minimum mean-squared error and zero-one loss respectively.

If there is zero tolerance of errors—classifying one pixel incorrectly is as bad as getting all of them wrong—then the MAP estimate maximizes the (very small) chance of success; the probability that we get the image *exactly* right. If, however, we can tolerate some errors in the reconstruction—we are satisfied with images that are sufficiently *close* to the source—then it is better to minimize the misclassification rate. This can be achieved by thresholding the mean value of each pixel to obtain the thresholded posterior mean or TPM estimate.

In statistical physics terms, finding the mean value of the posterior distribution is similar to calculating the **observables** of a physical system. The MAP estimate, on the other hand, corresponds to the **ground state** of the system and neglects much of the available information associated with the **entropy** of the system at finite temperatures. Recent work seems still to focus upon the MAP estimate despite recognition that the TPM estimate performs quite comparably and is far less demanding to compute [69].

The literature is full of anecdotal accounts of the failure of the MAP estimate in certain parameter regimes. It is never clear, however, whether this failure is due to the choice of the MAP estimate and could be alleviated by an alternative estimate, or whether the model is simply unusable in this parameter regime. Other than a rather limited analysis by Marroquin on a modest-sized 2x2 pixel image [84] there has been no systematic comparison of the TPM and MAP estimates.

Another issue that affects this choice is the presence of local minima. This makes it difficult to establish the true MAP estimate. Greig *et al.* [44] investigate the difference between the MAP estimate obtained by the Monte Carlo methods advocated in GG, and the genuine mode of the posterior distribution. The problem of freezing is well-known in condensed matter research, and we are able to explain some of the phenomena associated with these **metastable states.** Greig *et al.* also compare their results with the method of iterated conditional modes [16] which specifically seeks these local minima.

Parameter Estimation

Parameter estimation is one of the unsolved (and probably insoluble) problems in image restoration. Any method of restoration *must* make some use of a prior model, and any such prior model will have certain key parameters that must be determined.

The degree of information available when we decide on the values of these parameters varies according to the problem. We know that a poor choice of parameters can lead to a nonsensical estimate (for both mode and mean). Indeed, the lack of any reliable method of choosing the parameters is often cited as the fatal flaw in Markov random field models of image restoration. Certainly it is this indeterminacy *as well as* the issue of the choice of prior that leads to the assertion that "much personal experience with the Bayesian method is required before one can rely on it" [51].

A great deal of work skirts the problem of parameter estimation and assigns the parameter values on an *ad hoc* basis (e.g. [36, 69]) or, as in Chapter 3, compares different choices of parameters with no attempt to estimate the parameters *a priori.*

Most of the image restoration work that *does* address the issue of parameter estimation assumes that an ensemble of prototype uncorrupted pictures is available which can be analysed in an attempt to correctly parameterize the source [37, 72, 109, 123]. There is a large body of work in the statistics literature on such parameter estimation from complete or fully observed data, not usually in the context of image restoration [11, 12, 13, 14, 95].

There are various techniques commonly used for estimation from fully observed data. All seek the **maximum likelihood** estimate. The likelihood function of the data is the probability of generating a particular set of data, given various parameters. The maximum likelihood estimate is the set of parameters that maximizes this probability for the given data. However, for large data sets this maximum must be found in a large multidimensional space, necessitating the use of **coding techniques** [11, 12].

An attractive alternative to conventional maximum likelihood estimation is maximum **pseudo-likelihood** estimation [13, 14]. Here the likelihood function is a product of the local likelihood at each site, which depends only upon a neighbourhood of the site and so can be computed directly. This method is now widely used for parameter estimation from complete data.

The idea of parameter estimation from the *incomplete data* was taken up in the statistics literature as an errors-in-variables problem [15]. Also from the field of statistics came the iterative EM algorithm for parameter estimation [23] which is now being applied to image restoration in the engineering literature [129]. A similar method found in both the engineering [76] and statistics [101] literature involves simultaneous image restoration and parameter estimation. These converge to the maximum likelihood parameters estimated from the reconstruction and the optimal reconstruction given these estimated parameters. However, there is no guarantee that such a re-estimation process will converge to even a *local* maximum of the parameters and the reconstruction simultaneously. Certainly the methods are unlikely to find the global maximum and in general the results depend upon the initial choice of parameters.

The specific problems of maximum likelihood parameter estimation for the classical Ising model were considered in detail by Pickard [97] but the analysis was restricted to the fully observed data case. When the Ising field has been corrupted by noise Frigessi and Piccioni [30] show a way to find the optimal parameters, assuming the knowledge that the source *was* an Ising field

But what happens if we get the prior wrong? How can we choose parameter values to optimize the restoration in the absence of firm knowledge of the prior? There is an approach called the evidence approximation [81], closely related to generalized maximum likelihood. The evidence was initially introduced by Gull [46] as a method for estimating the free parameter in conventional maximum entropy restoration. This has been applied to the Bayesian training problem for back-propagation neural networks [20, 82], although recently the approximations involved have come under some attack [126]. Neal [91] recognized that comparing the evidence for different parameter choices corresponds to calculating free energy differences in statistical mechanics systems. We have established these free energies in the Bayesian image restoration model, and are able to investigate the success of the evidence approximation applied to image restoration.

1.4 Thesis Outline

In the next chapter we set out all of the basic theory of the restoration process. To keep the later analysis tractable we restrict ourselves to the case of binary images. We derive the posterior distribution from first principles using a prior on the density of edges alone, and recover the intensity prior proposed by Geman and Geman. This is not unreasonable, as the MRF model was proposed as a way of controlling the edge-density in the reconstruction. We show rigorously by maximum entropy methods that, given *only* the information about the density of edges, this is the only rational prior distribution. We go on to derive the TPM estimate that we will use, and discuss the comparison with MAP, to be picked up in Chapter 4. We determine the quantities required to calculate the evidence. Finally we establish the links between the Bayesian statistics expounded in the chapter and the statistical physics approach to be used thereafter.

In Chapter 3 we systematically investigate the effect of parameter choice on the success of the restoration. We are able to explain the behaviour of the model in terms of well-understood characteristics of statistical mechanics models. There are regions of parameter space where the model performs well, and other regions where the model fails to behave like a reconstruction scheme of any sort. These differing regions correspond to separate phases in physical systems and we are able to explore the nature of the **phase transitions** between these regions using analytic as well as Monte Carlo techniques.

From two extensive analytic calculations—the mean field approximation and a small coupling expansion—we construct the phase diagram of the model. By reference to this we are able to explain the failure of the method in certain parameter regimes and analytically model the qualitative performance of the restoration. Monte Carlo simulation of the reconstruction scheme provides confirmation of the analytic results, and permits quantitative evaluation of the performance. Simulation also allows us to display the images that result from the restoration, including the cases of catastrophic failure and the milder failures caused by the metastable states.

Chapter 4 is concerned with the choice of the best estimate that can be obtained from the posterior distribution. A comprehensive comparison of the MAP and TPM estimates is carried out by simulation, with reference to the work of Greig *et al.* [44]. We show, as expected, that the TPM estimate is always optimal in cases where the prior is well chosen. Although we are unable to calculate these estimates analytically, our understanding of the phase diagram of the model gleaned from the mean field approximation in Chapter 3, allows us to explain the differences in the parameter sensitivity of the two estimates.

Finally, in Chapter 5, we consider a method for parameter estimation—the evidence approximation. This requires the calculation of free energy differences. The small-coupling calculation provides analytic results for the evidence approximation, but calculation of free energy differences by simulation is a different, more difficult problem. We verify the success of a recent method [77] by calculating the free energy of a simple Ising model, before proceeding with measurements of the evidence itself. The evidence calculation correctly identifies the optimal parameters when the prior is correctly chosen. When we get the prior wrong, things are not so simple.

CHAPTER 2

The Bayesian Formulation of Image Restoration

2.1 Introduction

Imagine that we are presented with an image which is the result of the superposition of noise on an original picture which we wish to recover. We are provided with some information which characterizes the class of image that the original belongs to. We use this information along with our knowledge of the noise process to develop a reconstruction scheme. Then, with the additional information provided us by the presented noisy image, we attempt a reconstruction of the particular source image underlying the noise.

2.1.1 The Image Coordinates

For expediency we restrict analysis to the simplest two-colour pictures. We regard the presented image as a two-dimensional array of N pixels, and we represent this as an array of binary variables $D_i = \pm 1, i = 1 \ldots N$, each corresponding to a black or white pixel in the image. Then we can denote the entire image using the vector notation $\mathbf{D} \equiv \{D_1 \ldots D_N\}$, and we will henceforth call this the **data.**

Now there is a set of 2^N pictures that can be composed from N binary pixels. Since we are dealing with a random noise process, it is conceivable that any member of this universe of pictures could have been the original source image. Therefore the set of all possible source pictures covers this universe and is identical to the set of all possible data pictures. As with the data, we represent the source images as arrays of binary variables $S_i = \pm 1, i = 1 \ldots N$. Our task is to find the **source** image $\mathbf{S} \equiv \{S_1 \ldots S_N\}$ that was most likely to have been corrupted to give the data $\mathbf{D}$. Or, rather more usefully, we wish to find the probability distribution over the *whole* ensemble of source images that gives for each image in $\{\mathbf{S}\}$ the probability that it was indeed the original source from which the data $\mathbf{D}$ was generated.

Since we are in the business of investigation and analysis, we want to be able to *compare* the images that result from the restoration process with the images that make up the source ensemble. Although these images share the same space, the underlying probability distributions are of course different. Therefore, we introduce a further ensemble $\{\mathbf{R}\}$, distinct from $\{\mathbf{S}\}$, of **reconstructions** which are the restored images generated by the restoration process. Once again $\mathbf{R}$ is an array of binary variables $R_i = \pm 1, i = 1 \ldots N$, representing a binary pixel array.

2.1.2 The True Probability Distributions

The above definition of the variables introduced the idea of probability distributions underlying the ensembles of images. We now *identify* the probability distributions that interest us. Throughout we will not attempt to label the distributions explicitly but will allow, where possible, the arguments of the function to specify implicitly which distribution we are referring to.

First there is the *true* ensemble of source pictures $\{\mathbf{S}\}$. The original source picture S is drawn from this ensemble with a probability given by $P(\mathbf{S})$: the **source distribution**. For the purposes of the restoration, we do not know the explicit form of $P(\mathbf{S})$ and we will make an estimate of this (the prior distribution). Of course, for experimental purposes we can control the source process, and we choose a distribution appropriate for the property we wish to investigate.

The mechanism that takes the source and generates a corrupted image is a stochastic process, and is expressed as a conditional probability. The probability of observing a particular corrupted image **D**, given a source image **S**, is expressed by the **likelihood** function $P(\mathbf{D}|\mathbf{S})$. Once again we control the choice of this true likelihood function in our experiments, but when determining the restoration process we must imagine that this distribution is unknown to us and is to be estimated.

Now that we are given the corrupted picture **D** we try to recover the source S. The probability distribution that we seek is the **true posterior** distribution $P(\mathbf{S}|\mathbf{D})$—the probability that an image S is the original source image, given that we have observed a particular data image **D**. We call the *actual* probability distribution that we determine the **restored** distribution: given a particular data picture $\mathbf{D}$ we generate reconstructions $\mathbf{R}$ with a conditional probability given by $P(\mathbf{R}|\mathbf{D})$. This restored distribution characterizes the restoration scheme and is determined by the estimates we make of the source and likelihood distributions.

2.1.3 The Screens

Now let us sketch the environment we have defined in a more tangible form. This will also provide a framework for the experimental simulation in Chapter 3, when we come to test the results of the theory. Consider a series of three screens capable of portraying a picture of the type we are discussing, an N pixel binary image. Label these individually as the source, data, and reconstruction screens. On the source screen will be displayed a picture from the source ensemble $\{\mathbf{S}\}$, selected with a probability given by $P(\mathbf{S})$. The data screen will display a picture from the data ensemble $\{\mathbf{D}\}$ with a probability given by the true likelihood $P(\mathbf{D}|\mathbf{S})$. Finally the reconstruction screen will display a picture drawn from the $\{\mathbf{R}\}$ ensemble with a probability given by the restored probability distribution $P(\mathbf{R}|\mathbf{D})$. We see in Figure 2.1 how this screen analogy depicts the restoration scheme as it might actually be used: there is a picture on the source screen which is hidden from us; via the noise process the picture on the data screen, which is all we *can* see, is generated with a probability $P(\mathbf{D}|\mathbf{S})$; then given this particular data picture we carry out the restoration process and generate a sequence of restored pictures which appear on the reconstruction screen with a sampling probability given by $P(\mathbf{R}|\mathbf{D})$. We will consider the second row of Figure 2.1 when we discuss, in §2.3, how best to interpret the pictures on the restoration screen.

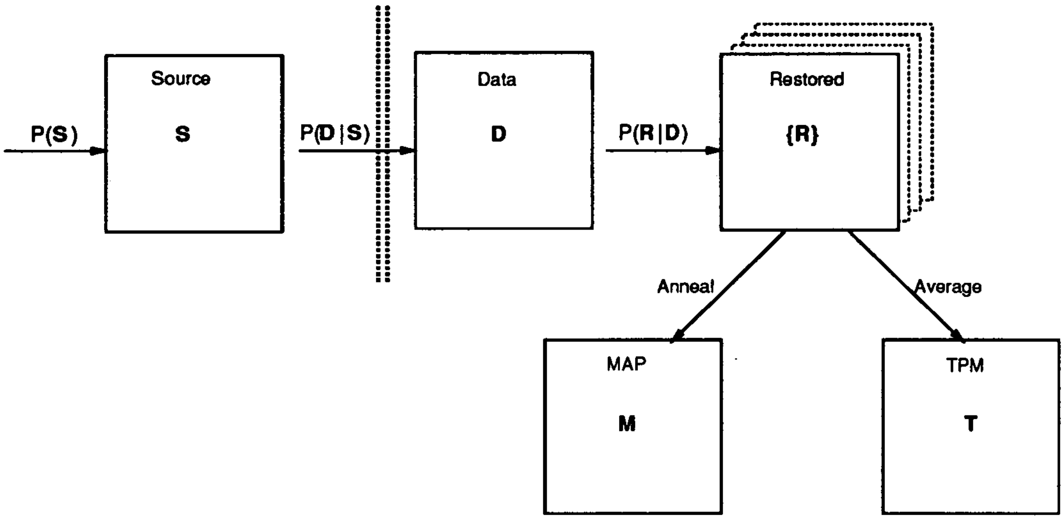


Figure 2.1. The Bayesian view of image restoration. The picture on the source screen is selected with a probability $P(\mathbf{S})$. A corrupted picture is generated with probability $P(\mathbf{D}|\mathbf{S})$ and displayed on the data screen. The restored screen displays an ensemble of pictures generated with probability $P(\mathbf{R}|\mathbf{D})$. The thresholded mean of this ensemble provides the TPM estimate, while the mode corresponds to the MAP estimate (to be discussed in Chapter 4). In a real world problem we would not be able to observe any of the processes behind the curtain that separates the source and data screens. However, for the purposes of experiment we control these processes too.

2.1.4 The Model Distributions

We want to investigate the effects of choosing a poor prior model of the processes that generated the data. This requires that we distinguish the *model* distributions from the *true* distributions described in §2.1.2.

We imagine that we do not know explicitly the true distribution $P(\mathbf{S})$ of the source images, but we will have *some* information about the source. We may for example know the expectation values of certain observable properties of the source pictures, averaged over the whole source ensemble $\{\mathbf{S}\}$. We will use any such information available to determine our best guess at the source distribution, the **prior** $\tilde{P}(\mathbf{S})$. Henceforth we will use the notation $\tilde{P}()$ to denote the model distributions— *approximations* to things we might measure. This distinguishes them from the true distributions $P()$ that we may assume we *know* when testing the restoration scheme.

In the same vein, we do not have complete information about the noise process—the true likelihood function $P(\mathbf{D}|\mathbf{S})$. We will use whatever knowledge we have about the statistics of the corruption process to construct our best guess at this distribution, which we will call the **model likelihood** $\tilde{P}(\mathbf{D}|\mathbf{S})$.

In fact we may not even have adequate information about the observables of the source ensemble. If we are fortunate we may know these from observations of a set of uncorrupted pictures drawn from the source ensemble. Alternatively, if we can accurately determine the noise process, we may be able to calculate these observables from measurements of a set of corrupted images. [For a simple statistical measure and Gaussian noise the inverse problem is not necessarily ill-posed, cf. §3.3.5.] Most likely, we have no certain knowledge of either the source distribution or the noise level. In this case we must resort to the parameter estimation techniques discussed later in §2.6.

We seek the true posterior distribution $P(\mathbf{S}|\mathbf{D})$, but since we do not have complete information we cannot determine it exactly; therefore we will use the limited information that we are given to assign the most rational values possible to the **model posterior** distribution $\tilde{P}(\mathbf{S}|\mathbf{D})$. This yields an explicit function of the coordinates $\mathbf{D}$ which we use to *define* the restoration process, setting $P(\mathbf{R}|\mathbf{D}) = \tilde{P}(\mathbf{S}|\mathbf{D})_{\mathbf{S}\to\mathbf{R}}$; we obtain the restored distribution by replacing $\mathbf{S}$ with $\mathbf{R}$ in the model posterior. When our last piece of information arrives—we are presented with a given data picture—we realize a probability distribution over $\{\mathbf{R}\}$ that will generate reconstructed pictures.

2.2 Priors and Posteriors

2.2.1 Introduction

In this section we make use of ideas drawn from the fields of information theory and Bayesian statistics. The two disciplines have a tremendous amount in common.

Information theory is a branch of the mathematical theory of probability and statistics [75], and is relevant to statistical inference. The communication theory aspects have led to a resurgence in interest with the dawning of the information age, concerned particularly with signal processing, compression techniques, and various ways of 'encoding' information. The information theoretic approach to statistical mechanics is another useful application, which offers careful and precise mechanisms for describing the amount of information available.

The field of Bayesian statistics sometimes more resembles a religion than a discipline. It concerns itself with the assignment of probabilities by inference, based upon the information available *to the agent assigning the probabilities.* In this respect, Bayesian inference models the kind of decision making processes that we use in everyday life [67].

Both disciplines are careful and useful in determining the available information: Bayes concentrates on the agent, while information theory analyses the object and embraces both frequentist and Bayesian probability theories.

Bayesian Statistics

The basic problem is one of inverse probability. We calculate a prior probability distribution that describes our knowledge of the ensemble of pictures appearing on the source screen, *before* observation of the data. We want to know how to update our degree of belief about what we think is displayed on the source screen *after* the data arrives (when we are shown what is displayed on the data screen). Bayes theorem [9] gives a mathematical procedure for updating our prior belief about the value of a set of coordinates, to produce the posterior distribution for these coordinates reflecting our increased knowledge after observation of the data.

In its most general form, Bayes theorem is easily derived from the definition of conditional probability. The joint probability of two events A and B both occurring may be expressed as the product of the probability of event B and the conditional probability of event A given event B. But we can equally well consider the events in the reverse order; then

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A).$$

Rearranging this equation gives us Bayes theorem in its commonly stated form

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, \tag{2.1}$$

or, in words, that the posterior probability distribution is proportional to the likelihood times the prior.

- $P(B|A)$ is the posterior probability distribution for event B, once we know that event A has in fact occurred.
- $P(B)$ is the prior distribution—the probability of B before we know anything about A.
- $P(A|B)$ is the likelihood of the event A given that the event B does occur.
- $P(A)$ is in fact the prior distribution of A, but clearly acts as a normalization, with $P(A) = \sum_B P(A|B)P(B)$.

Information Theory

The Bayesian statistics discussed above tells us how to derive the posterior distribution from the prior distribution, given the likelihood function of the data. However, we have not yet chosen the prior distribution and we need a consistent method to determine this. Here we call on ideas from information theory and use these to justify our claim that we will make the most rational assignment possible of the probabilities. The crux of our claim is that when assigning the prior distribution we should use all of the information available to us, but equally, not make any unfounded assumptions. If we do not use all available information we will fail to capture some potentially accessible behaviour of the distribution. In this case the results are often counter-intuitive [68]. If we unintentionally make hidden assumptions about the prior distribution we will introduce artefacts into the solution. Since we are operating in the absence of complete information, either failure may result in an adequate or improved solution in some special case. However, *averaged* over all possible cases, both errors will reduce the effectiveness of the restoration process.

If we have no information at all then our best prediction can only be that all possible outcomes will be equally likely, and our most rational assignment is simply that each outcome has the same probability. How then do we fold in any extra information that we may have? To aid us in this task, we desire a function that will give some quantitative measure of the amount of information we have assumed in our assignment of the probabilities. Such a function is the missing information function defined as

$$S = -\sum_{\mu} p_{\mu} \log p_{\mu}, \tag{2.2}$$

where the p_{μ} are discrete probabilities, and the sum is over the entire probability distribution. The expression (2.2) is also given the name 'Shannon entropy' [106]: entropy measures the degree of disorder, and hence a high entropy is consistent with a large amount of missing information.

The function measures the amount of information implicitly assumed by any choice of the probabilities. The following rationale enables us to use this property to make the most logical assignment of the probabilities.

1. We assume that any information we have is expressible as a constraint on the values of the probabilities. Any choice of probabilities that violates these constraints is failing to use the corresponding information.
2. We want to choose these probabilities in such a way as to obtain the maximal value of the missing information function, subject to the constraints. A choice that gives a value for S less than its maximal constrained value is making assumptions that are invalid given the available information.
3. We can find this maximal value using the method of Lagrange undetermined multipliers, where we vary the probabilities in order to maximize the function S subject to the various constraints. Because of the entropic form of S, this procedure for determining the distribution is often given the name 'maximum entropy' [66].

Shore and Johnson [108] proved that the principle of maximum entropy proposed by Jaynes in [66] is a consistent method of inference when given new information in terms of expectation values. Thus we can use this information theory technique to make the best possible assignment of the prior distribution, which we can then feed into our Bayesian scheme to calculate the posterior distribution.

2.2.2 The Posterior Distribution: The Bayesian Result

We now apply Bayes theorem to our specific problem; the determination of the best choice of probabilities for the model posterior $\tilde{P}(\mathbf{S}|\mathbf{D})$. Writing (2.1) in terms of the estimated distributions that represent our incomplete state of knowledge, we get:

$$\tilde{P}(\mathbf{S}|\mathbf{D}) = \frac{\tilde{P}(\mathbf{D}|\mathbf{S})\tilde{P}(\mathbf{S})}{\tilde{P}(\mathbf{D})}. \tag{2.3}$$

Consider in detail the terms in (2.3):

- $\tilde{P}(\mathbf{S}|\mathbf{D})$ is what we seek to prescribe, the model **posterior** distribution of **S** after the data arrives. It is our best shot at a reconstruction scheme using all our available information at this time as input to the right hand side of (2.3).

- $\tilde{P}(\mathbf{D}|\mathbf{S})$ is the model **likelihood** of getting the data **D** for a particular source **S** given no explicit knowledge of the true data distribution. We will make a rational assignment of its form using our information about the corruption process.

- $\tilde{P}(\mathbf{S})$ is the **prior** probability of a particular source picture **S**, given our limited information. It is our best guess at what is appearing on the source screen in the absence of explicit knowledge of the source distribution $P(\mathbf{S})$, and before we are shown what picture is currently on the data screen. We will use whatever prior information we have about the source pictures to guide us in making the most rational assignment of this function.

- $\tilde{P}(\mathbf{D})$ is the prior probability of the data given the same limited knowledge. Given our assignment of the model likelihood and prior, it is constrained by the normalization condition

$$\sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{S}|\mathbf{D}) = 1.$$

Its crucial feature in (2.3) is that it is by definition a function of the coordinates $\mathbf{D}$ alone, and will have no impact on the S-dependence of the posterior. We have no extra information about $\tilde{P}(\mathbf{D})$ that will not be used in the assignment of $\tilde{P}(\mathbf{D}|\mathbf{S})$ and $\tilde{P}(\mathbf{S})$—therefore our best guess for this distribution has to be:

$$\tilde{P}(\mathbf{D}) = \sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{D}|\mathbf{S})\tilde{P}(\mathbf{S}). \tag{2.4}$$

$\tilde{P}(\mathbf{D})$ is also known as the **evidence** [46, 81], as it measures the 'evidence' provided by the data $\mathbf{D}$ for our particular choice of prior and likelihood. We will pick up this idea later in §2.6.

As argued above, $\tilde{P}(\mathbf{S}|\mathbf{D})$ is our best guess at a solution to the original problem we posed back in §2.1.1, and therefore we *assign* the distribution $P(\mathbf{R}|\mathbf{D})$, which defines our restoration process, the same functional form as this model posterior, $\tilde{P}(\mathbf{S}|\mathbf{D})$.

Thus we write:

$$P(\mathbf{R}|\mathbf{D}) \stackrel{\text{def}}{=} \tilde{P}(\mathbf{S}|\mathbf{D})_{\mathbf{S}\to\mathbf{R}} = \left.\frac{\tilde{P}(\mathbf{D}|\mathbf{S})\tilde{P}(\mathbf{S})}{\tilde{P}(\mathbf{D})}\right|_{\mathbf{S}\to\mathbf{R}}. \tag{2.5}$$

2.2.3 The Prior Distribution: Information About the Source

Now we want to assign the most rational values possible to the prior distribution $\tilde{P}(\mathbf{S})$ using the information theory/maximum entropy technique described in §2.2.1. We proceed by imposing on the prior distribution, as constraints, all of the information we have about the source distribution.

In particular, let us imagine that we know, or at least think that we can estimate, the mean value O of some observable property of the source images, defined by the operator $O(\mathbf{S})$. Then we will impose this value as a constraint on the probability values of the prior distribution $\tilde{P}(\mathbf{S})$. We require that the average of the operator over the prior distribution $\langle O\rangle_p$ take on the value of our estimate O:

$$\langle O\rangle_p \overset{\text{def}}{=} \sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{S})O(\mathbf{S}) = O. \tag{2.6}$$

Here we use the subscript p to denote a functional of the prior distribution $\tilde{P}(\mathbf{S})$.

The missing information function for the prior distribution is defined as

$$S_p = -\sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{S}) \log \tilde{P}(\mathbf{S}). \tag{2.7}$$

If we have measurements O_α of n different operators $O_\alpha(\mathbf{S})$ averaged over the source distribution, then we want to choose the prior probability distribution $\tilde{P}(\mathbf{S})$ such that S_p takes on its maximal value, subject to the n constraints

$$\langle O_\alpha\rangle_p \overset{\text{def}}{=} \sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{S})O_\alpha(\mathbf{S}) = O_\alpha, \qquad \alpha = 1\ldots n, \tag{2.8}$$

and the normalization constraint

$$O_0 \overset{\text{def}}{=} \sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{S}) = 1. \tag{2.9}$$

Introducing Lagrange multipliers $\lambda_0 \ldots \lambda_n$ for these $n+1$ constraints, a turning point in S_p will correspond to

$$dS_p + \lambda_0 dO_0 + \lambda_1 d\left\langle O_1 \right\rangle_p + \cdots + \lambda_n d\left\langle O_n \right\rangle_p = 0.$$

Now, differentiating (2.7)

$$dS_p = -\sum_{\{\mathbf{S}\}} [1 + \log \tilde{P}(\mathbf{S})] d\tilde{P}(\mathbf{S}), \tag{2.10}$$

and from (2.8)

$$d\left\langle O_\alpha \right\rangle_p = \sum_{\{\mathbf{S}\}} O_\alpha(\mathbf{S}) d\tilde{P}(\mathbf{S}).$$

Thus

$$\sum_{\{\mathbf{S}\}} \left[\lambda_0 + \lambda_1 O_1(\mathbf{S}) + \cdots + \lambda_n O_n(\mathbf{S}) - 1 - \log \tilde{P}(\mathbf{S})\right] d\tilde{P}(\mathbf{S}) = 0.$$

Since this condition must hold for arbitrary variations in $\tilde{P}(\mathbf{S})$ we may write for each and every configuration $\mathbf{S}$:

$$\log \tilde{P}(\mathbf{S}) = \lambda_0 + \lambda_1 O_1(\mathbf{S}) + \cdots + \lambda_n O_n(\mathbf{S}) - 1,$$

which gives

$$\tilde{P}(\mathbf{S}) = e^{\lambda_0 - 1} \exp\{\lambda_1 O_1(\mathbf{S}) + \cdots + \lambda_n O_n(\mathbf{S})\}.$$

Now differentiating (2.10) with respect to the probability of a particular configuration S gives:

$$\frac{\partial^2 S_p}{\partial \tilde{P}^2(\mathbf{S})} = \frac{-1}{\tilde{P}(\mathbf{S})} < 0.$$

Since there are no cross-terms in (2.10), off-diagonal terms in the Jacobian of S_p are necessarily zero and we are guaranteed a maximum in S_p for this choice of $\tilde{P}(\mathbf{S})$.

The value of the normalization $e^{\lambda_0 - 1}$ can be determined from the normalization property (2.9), while the other Lagrange multipliers $\lambda_1 \ldots \lambda_n$ are implicitly defined by the constraint equations (2.8). When we perform the sums over the prior distribution we get from (2.9)

$$\lambda_0 = 1 - \log \sum_{\{\mathbf{S}\}} \exp\left\{\lambda_1 O_1(\mathbf{S}) + \cdots + \lambda_n O_n(\mathbf{S})\right\}, \tag{2.11}$$

and from (2.8), n equations,

$$O_\alpha = e^{\lambda_0 - 1} \sum_{\{\mathbf{S}\}} \frac{\partial}{\partial \lambda_\alpha} \exp\left\{\lambda_1 O_1(\mathbf{S}) + \cdots + \lambda_n O_n(\mathbf{S})\right\}, \tag{2.12}$$

so we have $n+1$ simultaneous equations to be solved for the $n+1$ unknowns, $\lambda_0 \ldots \lambda_n$.

Thus this method allows us to make the best assignment of the prior distribution $\tilde{P}(\mathbf{S})$ with only the information we have about the value of certain observables. In practice, we may not be able to determine the exact values of the Lagrange multipliers in the event that we cannot calculate the sums in (2.11) and (2.12) analytically.

2.2.4 The Likelihood: Information About the Noise Process

In order to determine the posterior distribution we must also assign the model likelihood $\tilde{P}(\mathbf{D}|\mathbf{S})$. In the most general binary case we model the noise process as an independent probability of corruption for each pixel: $P(D_i \neq S_i) = \tilde{q}_i(\mathbf{S})$. As the corruption probabilities may vary depending on the local nature of the source image, we may have to specify very many (up to $N.2^N$) corruption probabilities and Lagrange multipliers.

The probability of the data is given by

$$P(D_i|\mathbf{S}) = [1 - \tilde{q}_i(\mathbf{S})]\delta_{D_i S_i} + \tilde{q}_i(\mathbf{S})[1 - \delta_{D_i S_i}], \tag{2.13}$$

with

$$\delta_{D_i S_i} = \begin{cases} 1, & \text{if } D_i = S_i, \\ 0, & \text{otherwise,} \end{cases}$$

or equivalently

$$\sum_{D_i = \pm S_i} P(D_i|\mathbf{S}) D_i S_i = 1 - 2\tilde{q}_i(\mathbf{S}). \tag{2.14}$$

Following the method of §2.2.3 we have N constraints per source configuration:

$$\langle D_i S_i \rangle_l \overset{\text{def}}{=} \sum_{\{\mathbf{D}\}} \tilde{P}(\mathbf{D}|\mathbf{S}) D_i S_i = 1 - 2\tilde{q}_i(\mathbf{S}), \tag{2.15}$$

where the subscript l indicates a functional of the likelihood distribution $\tilde{P}(\mathbf{D}|\mathbf{S})$. These constraints require N Lagrange multipliers, which along with the normalization condition yield the probability distribution:

$$\tilde{P}(\mathbf{D}|\mathbf{S}) = e^{\mu_0 - 1} \exp\left\{ \sum_j \mu_j(\mathbf{S}) D_j S_j \right\}. \tag{2.16}$$

We can determine the values of the Lagrange multipliers as a function of $\tilde{q}_i(\mathbf{S})$ from the constraint equations (2.15):

$$\begin{aligned} 1 - 2\tilde{q}_i(\mathbf{S}) &= \sum_{\{\mathbf{D}\}} e^{\mu_0 - 1} D_i S_i \exp\left\{ \sum_j \mu_j(\mathbf{S}) D_j S_j \right\} \\ &= 2^N e^{\mu_0 - 1} \sinh \mu_i(\mathbf{S}) \prod_{j \neq i} \cosh \mu_j(\mathbf{S}), \end{aligned}$$

and the normalization condition gives

$$1 = \sum_{\{\mathbf{D}\}} \tilde{P}(\mathbf{D}|\mathbf{S}) = e^{\mu_0 - 1} 2^N \prod_j \cosh \mu_j(\mathbf{S}).$$

Hence, we obtain

$$\begin{aligned} (1 - 2\tilde{q}_i(\mathbf{S})) &= \tanh \mu_i(\mathbf{S}), \\ \Rightarrow \mu_i(\mathbf{S}) &= \frac{1}{2} \log \left[\frac{1}{\tilde{q}_i(\mathbf{S})} - 1 \right], \end{aligned}$$

and finally we can write

$$\tilde{P}(\mathbf{D}|\mathbf{S}) = \prod_j \left[\tilde{q}_j(\mathbf{S})(1 - \tilde{q}_j(\mathbf{S}))\right]^{\frac{1}{2}} \exp\left\{ \frac{1}{2} \sum_i \log\left(\frac{1}{\tilde{q}_i(\mathbf{S})} - 1 \right) D_i S_i \right\}. \tag{2.17}$$

2.2.5 The General Result

We have now specified the prior and likelihood distributions required by Bayes theorem (2.3). Recall that the evidence $\tilde{P}(\mathbf{D})$ is taken care of by the normalization. Therefore let us combine the general results and write the restoration scheme as defined in (2.5) as:

$$P(\mathbf{R}|\mathbf{D}) = \frac{1}{Z}\exp\left\{\lambda_1 O_1(\mathbf{R}) + \cdots + \lambda_n O_n(\mathbf{R}) + \frac{1}{2}\sum_i \log\left(\frac{1}{\tilde{q}_i(\mathbf{R})} - 1\right) R_i D_i + \mu_o(\mathbf{R}) - 1\right\} \tag{2.18}$$

with the normalization:

$$Z = \sum_{\{\mathbf{R}\}}\exp\left\{\lambda_1 O_1(\mathbf{R}) + \cdots + \lambda_n O_n(\mathbf{R}) + \frac{1}{2}\sum_i \log\left(\frac{1}{\tilde{q}_i(\mathbf{R})} - 1\right) R_i D_i + \mu_o(\mathbf{R}) - 1\right\} \tag{2.19}$$

2.3 A Measure of Quality

We have now established the form of the probability distribution $P(\mathbf{R}|\mathbf{D})$ that presents the available information in the best form for the image reconstruction we shall attempt. We want to develop criteria to assess just how good is the result we will obtain with this strategy, and quite how we shall use this result to generate reconstructed pictures.

2.3.1 Displaying the Output

Operationally we are able to generate configurations from the ensemble of restored pictures $\{\mathbf{R}\}$ sampled with a probability given by the distribution $P(\mathbf{R}|\mathbf{D})$. In general, any one of these configurations will appear with only an infinitesimal probability, and this is the case even for the most probable configuration in the output. Since this configuration is the mode of the posterior distribution, it is called the maximum *a posteriori* (MAP) estimate of the true source: the single source picture that was most likely to have been corrupted to give the data. Determining the MAP estimate from the distribution is operationally very difficult (we will discuss this in Chapter 4), but in any case, we should be able to make better use of the posterior distribution than this. Rather than find the single most probable configuration, we want to characterize the *entire* distribution. The frequency with which configurations appear in the output reflects the confidence we may place in them as reconstructions. We wish to find the generic properties of the *typical* configurations that appear with greatest overall probability.

We perform an ensemble average over the output configurations to give the first moment of the output vectors, which distills information from the whole reconstruction distribution:

$$\langle \mathbf{R}[\mathbf{D}] \rangle_R \stackrel{\text{def}}{=} \sum_{\{\mathbf{R}\}} \mathbf{R} P(\mathbf{R}|\mathbf{D}). \tag{2.20}$$

This is a real valued vector quantity, and gives the average value of the site variable R_k for each pixel in the ensemble of restored images. From this we may determine the most probable colour for the pixel, given by $\text{sgn}(\langle R_k[\mathbf{D}] \rangle_R)$ with the corresponding confidence measure of this prediction being $| \langle R_k[\mathbf{D}] \rangle_R |$.

The most concise way to display all of this information is to use a single grey-scale picture: the shade of each pixel, determined by the value of $\langle R_k[\mathbf{D}] \rangle_R$, indicates our confidence in the prediction for that particular element. In practice, we want the output to be another binary picture, so we display the thresholded value $\text{sgn}(\langle R_k[\mathbf{D}] \rangle_R)$, and we do not explicitly represent the varying confidence levels across the image. This is the thresholded posterior mean (TPM) estimate of Marroquin [84].

The TPM method contrasts with the approach of Geman and Geman [36] and much subsequent work which uses the MAP estimate. We will discuss the MAP estimate in detail in Chapter 4. More recent work by Skilling, Robinson, and Gull [112] shows a 'movie' of configurations with a probability of appearance determined by the probability weighting of each configuration. Our own eyes and brain then do the processing to detect those pixels which the restoration scheme predicts with high confidence from those which it predicts with low confidence. This is very similar to the grey scale picture described above.

2.3.2 Measuring the Quality of Restoration

Now we want to measure *quantitatively* how successful the reconstruction process can be on a typical picture. There are really two questions here:

1. How near is our reconstruction to the best that we could possibly do on the average—*if we were told the correct source distribution and correct corruption probability.*
2. How useful is the information we get in this best possible case.

Before we begin we must define some measure of the 'closeness' of two images. The **overlap** is a suitable measure, defined as the normalized scalar product of the two configuration vectors:

$$\mathbf{A}.\mathbf{B} \overset{\text{def}}{=} \frac{1}{N}\sum_{k} A_k B_k. \tag{2.21}$$

We see that the overlap takes on value 1 when the two images are identical, zero if there is no correlation between the images, and -1 if one image is the negative of the other. For the simplest measure of the success of the reconstruction we may calculate the overlap between the source and our chosen reconstruction. This is the approach taken to-date by those authors who have gone beyond *ad hoc* assessments of image quality and presented quantitative results of any sort (e.g. [44, 103]).

However, we have more information available in the probability distributions. To use this, we must be able to deal with real valued vectors, so let us define a square distance between two real valued vectors as:

$$d\,[\mathbf{A},\mathbf{B}] = \frac{1}{N}\sum_k [A_k - B_k]^2 .$$

Now to answer the first question, consider what our aim was when determining the reconstruction scheme. We have tried to get the probability distribution $P(\mathbf{R}|\mathbf{D})$ to be as close a match as possible to the true posterior distribution $P(\mathbf{S}|\mathbf{D})$. A standard measure of the difference between two probability distributions is the **cross-entropy** [43, 75]. However, we seek a measure that can be calculated in terms of observables.

We first measure the average distance between the data (the image we have to begin with) and configurations from the source distribution:

$$\begin{aligned} Q_0[\mathbf{D}] &\overset{\text{def}}{=} d\,[\mathbf{D}, \langle \mathbf{S}[\mathbf{D}]\rangle_S] \\ &= \frac{1}{N}\sum_k [D_k - \langle S_k[\mathbf{D}]\rangle_S]^2 . \end{aligned} \tag{2.22}$$

Note that $\langle Q_0\rangle_D$ is a function of the noise and the true posterior distribution only. Therefore it is constant for a given source distribution and noise process.

We can then quantify the improvement of the reconstructions over the data by measuring the average distance between images from the restored distribution and members of the source distribution:

$$\begin{aligned} Q_A[\mathbf{D}] &\stackrel{\text{def}}{=} d[\langle \mathbf{R}[\mathbf{D}]\rangle_R, \langle \mathbf{S}[\mathbf{D}]\rangle_S] \\ &= \frac{1}{N}\sum_k [\langle R_k[\mathbf{D}]\rangle_R - \langle S_k[\mathbf{D}]\rangle_S]^2 . \end{aligned} \tag{2.23}$$

We want to compare the difference between $Q_A[\mathbf{D}]$ and $Q_0[\mathbf{D}]$, averaged over the noise—all possible configurations of the data—and we normalize this, dividing by $\langle Q_0\rangle_D$. This normalization scales the difference so that the more corrupted a data picture is, the greater the improvement we will require before considering the output to be useful. Thus we define this measure of performance, which we will call the **quality factor**, as:

$$Q \stackrel{\text{def}}{=} \frac{\langle Q_0 - Q_A\rangle_D}{\langle Q_0\rangle_D} \tag{2.24}$$

$$= \frac{\langle d[\mathbf{D}, \langle \mathbf{S}[\mathbf{D}]\rangle_S] - d[\langle \mathbf{R}[\mathbf{D}]\rangle_R, \langle \mathbf{S}[\mathbf{D}]\rangle_S]\rangle_D}{\langle d[\mathbf{D}, \langle \mathbf{S}[\mathbf{D}]\rangle_S]\rangle_D}. \tag{2.25}$$

This has the following properties:

- In the event that we perfectly model the true posterior distribution $[P(\mathbf{R}|\mathbf{D}) = P(\mathbf{S}|\mathbf{D})]$ then $Q_A[\mathbf{D}] = 0,\ \forall\mathbf{D}$ and $Q = 1$.

- If our model is incomplete, $\langle Q_A\rangle_D$ will nevertheless be minimized when the two probability distributions are matched most closely, given the model. Therefore, finding the maximum value of Q will determine the optimal choice of parameters for the model.

- If we make no improvement on the dataset given, and simply choose $\langle \mathbf{R}[\mathbf{D}]\rangle_R = \mathbf{D}$, then we get $Q_A[\mathbf{D}] = Q_0[\mathbf{D}]$, $\forall\mathbf{D}$, and therefore $Q = 0$.

- If we make a particularly poor choice of the parameters, giving results that are worse than the given dataset, then the quality factor indicates this by becoming large and negative.

We now return to the second question. How good is the best restoration that could be achieved, given that we perfectly model the source distribution and noise process. Since the quality factor measures how close the true posterior distribution lies to the restored distribution we have obtained, this other question is one of how precise the optimal result is: what are the error bounds? We hope that the particular original source image that generated the data $\mathbf{D}$ lies somewhere in the region of high probability in the restored distribution. We want some measure of the width of this probability distribution.

The Hamming distance is the number of bits that differ between two binary signals, or in this case the number of pixels that differ between the two images:

$$H[\mathbf{A},\mathbf{B}] \overset{\text{def}}{=} \frac{1}{2}\sum_k (1 - A_k B_k) . \tag{2.26}$$

It is simply related to the overlap:

$$H[\mathbf{A},\mathbf{B}] = \frac{N}{2}(1 - \mathbf{A}.\mathbf{B}) .$$

Now we can define the width of the restored distribution $P(\mathbf{R}|\mathbf{D})$ as:

$$\begin{aligned} W_R &\overset{\text{def}}{=} \frac{1}{N}\langle\langle H[\mathbf{R},\mathbf{R}']\rangle_R\rangle_{R'} \\ &= \frac{1}{2}\frac{1}{N}\sum_k \sum_{\{\mathbf{R}\}} \sum_{\{\mathbf{R}'\}} \{1 - R_k R'_k\}\, P(\mathbf{R}|\mathbf{D})P(\mathbf{R}'|\mathbf{D}) \\ &= \frac{1}{2}\left\{1 - \frac{1}{N}\sum_k \langle R_k[\mathbf{D}]\rangle_R^2\right\}, \end{aligned} \tag{2.27}$$

and the width of the true posterior distribution as

$$\begin{aligned} W_S &\overset{\text{def}}{=} \frac{1}{N}\langle\langle H[\mathbf{S},\mathbf{S}']\rangle_S\rangle_{S'} \\ &= \frac{1}{2}\left\{1-\frac{1}{N}\sum_k \langle S_k[\mathbf{D}]\rangle_S^2\right\}. \end{aligned} \tag{2.28}$$

Measuring the width W_S allows us to answer the second question. The narrower is the true posterior distribution, the smaller is the set of images that the true source is likely to have been drawn from, and hence the greater is the confidence we can place in our estimate.

To summarize:

- The quality factor, Q, measures how close the means of the two probability distributions lie — the true posterior distribution $P(\mathbf{S}|\mathbf{D})$ and the restored distribution $P(\mathbf{R}|\mathbf{D})$.
- W_S measures the width of the true posterior distribution, which indicates how large is the range of values that the original source **S** may have been drawn from, after we know the data picture **D**. It measures the confidence with which we may determine the source picture if we *know* the source and true likelihood distributions.
- W_R measures the width of the restored distribution. Like W_S it is a confidence measure of sorts, but since the estimates of the prior and model likelihood may be poor, it is possible to be confident but wrong.

With this last point in mind we ought never to specify the restored distribution with a greater confidence than the true posterior distribution could provide—it is unreasonable to have $W_R < W_S$. Evidently the three measures are linked. If we evaluate Q in terms of W_R and W_S we find:

$$\begin{aligned} Q &= \frac{\frac{1}{N}\sum_k \langle [\langle R_k[\mathbf{D}]\rangle_R - D_k]\,\langle S_k[\mathbf{D}]\rangle_S\rangle_D + W_R}{1-\frac{1}{N}\sum_k \langle D_k\,\langle S_k[\mathbf{D}]\rangle_S\rangle_D - W_S} \\ &= \frac{\left\langle\langle\langle \mathbf{R}.\mathbf{S}\rangle_S\rangle_R\right\rangle_D + W_R - \alpha}{1-\alpha-W_S}, \end{aligned} \tag{2.29}$$

where

$$\alpha \overset{\text{def}}{=} \frac{1}{N}\sum_k \langle D_k\,\langle S_k[\mathbf{D}]\rangle_S\rangle_D\,. \tag{2.30}$$

The average overlap between the source and data is constant for a given noise process and is not affected by the parameterization of the restoration scheme, so we call this α.

The achievement of maximum quality in the restoration requires a balance between competing terms in the quality factor. We want to maximize the simple overlap between source and reconstruction, but we must not do this at the expense of making the restored distribution artificially narrow, lest we reduce the quality factor as well.

Refer to Figure 2.2. If we can make Q_A small enough, then the quality factor will be large and we are allowed to have W_R of similar size to W_S. If, however, the distance Q_A between distributions is large, then we need the width of the restored distribution $P(\mathbf{R}|\mathbf{D})$ to be large in order that it should still encompass the true posterior distribution $P(\mathbf{S}|\mathbf{D})$. The quality factor demonstrates this behaviour correctly.

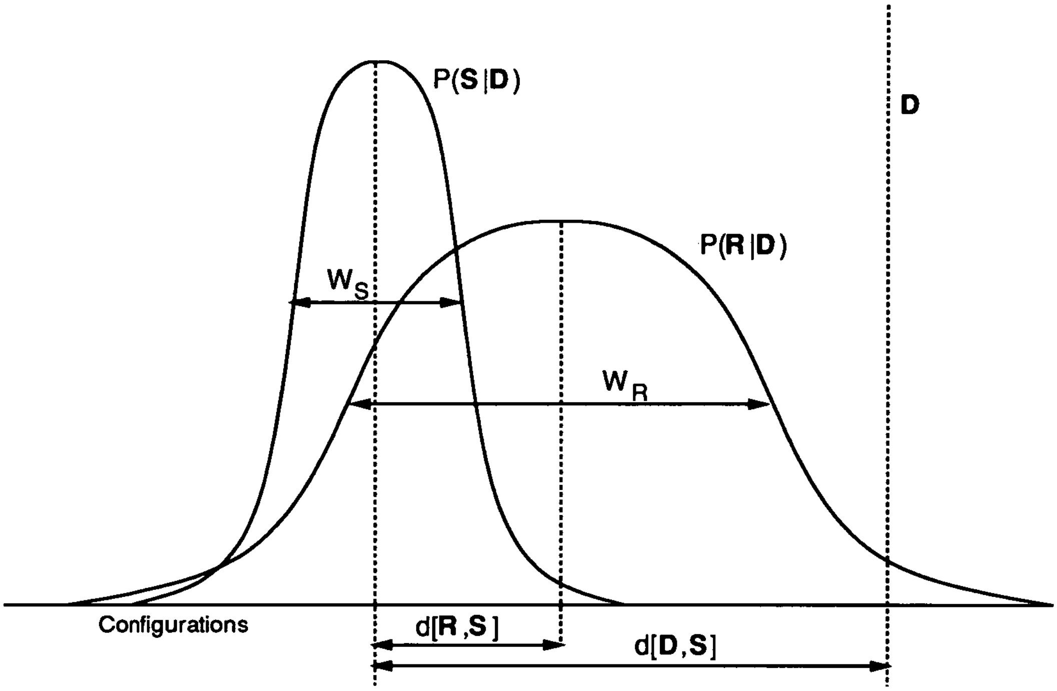


Figure 2.2. A comparison of the probability distributions. This is an attempt to show in one dimension the difference between the true posterior and model posterior distributions, and their relationship to the quality factor. The quality factor provides a measure of how close the restored distribution is to the true posterior, as a proportion of how close the data picture is to the true posterior. Notice that we need to widen the restored distribution, dependent upon how far its mean is from the true posterior, if we require most of the probability mass of the true posterior to be contained in the restored distribution.

2.3.3 Recovering a Single Thresholded Binary Output

Now that we have developed a consistent quality measure, we can return once again to the issue of the optimal estimator. Marroquin [84] states that if we correctly model the posterior distribution, the thresholded posterior mean (TPM) is the optimal Bayesian estimator that minimizes the average bitwise error between the estimate and the source.

In fact the proof of this is straightforward if we consider it in two parts. To begin with we are generating pictures $\mathbf{R}$ according to the distribution $P(\mathbf{R}|\mathbf{D})$. We want to find the binary picture $\mathbf{T}$ that has the maximal overlap (minimum bitwise error) with the $\mathbf{R}$, averaged over the entire ensemble of restored pictures. Thus we want to maximize

$$\frac{1}{N}\sum_{k}\sum_{\{\mathbf{R}\}} T_k R_k P(\mathbf{R}|\mathbf{D}) = \frac{1}{N}\sum_{k} T_k \left\langle R_k[\mathbf{D}]\right\rangle_R .$$

Since T_k can only take on values ± 1, this is clearly maximized when every term in the sum over sites is positive, and therefore:

$$T_k = \operatorname{sgn}\left\{\left\langle R_k[\mathbf{D}]\right\rangle_R\right\}, \tag{2.31}$$

recovering the intuitive result from §2.3.1.

Thus $\mathbf{T}$ is the single binary image that best characterizes the restored distribution in the sense that it minimizes the average bitwise error between $\mathbf{T}$ and the images in the restored distribution. It is then a trivial statement that if the restored distribution correctly models the true posterior distribution, then the TPM estimate minimizes the average bitwise error between T and the source picture.

If we want to measure how good the TPM estimate is in the general case (i.e. how close to the true posterior distribution) we should rewrite (2.23) replacing the reconstructions $\mathbf{R}$ by $\mathbf{T}$ and calculate the mean square distance between the thresholded image and the source:

$$\langle d[\mathbf{T},\langle\mathbf{S}[\mathbf{D}]\rangle_S]\rangle_D = 1-\frac{2}{N}\sum_k \langle\langle S_k[\mathbf{D}]\rangle_S \,\mathrm{sgn}\{\langle R_k[\mathbf{D}]\rangle_R\}\rangle_D+\frac{1}{N}\sum_k \left\langle\langle S_k[\mathbf{D}]\rangle_S^2\right\rangle_D \tag{2.32}$$

The minimum of the square distance (2.32) defines the maximum of the quality factor. Since only the second term in (2.32) depends on $\mathbf{T}$, the maximum of the quality factor Q coincides with the maximum of:

$$\begin{aligned} Q_T[\mathbf{D}] &\overset{\mathrm{def}}{=} \frac{1}{N}\sum_k \langle\langle S_k[\mathbf{D}]\rangle_S \,\mathrm{sgn}\{\langle R_k[\mathbf{D}]\rangle_R\}\rangle_D \\ &= \langle\langle \mathbf{T}.\mathbf{S}\rangle_S\rangle_D\,, \end{aligned} \tag{2.33}$$

which is simply the overlap between the TPM estimate $\mathbf{T}$ and the source $\mathbf{S}$, averaged over all possible values of the source and data.

2.4 A Specific Prior Model

We have so far been considering the most general case, without specifying any particular source distribution or noise process, nor the priors that we will use. We now restrict the general result to a simple noise process and a particular choice of prior distribution, which we can then investigate in greater detail.

We want to compare the performance of the model in two distinct cases: (i) where the prior model matches the true source distribution; and (ii) where the prior is a poor model of the true source. We will actually only consider a single prior for the source distribution. But we will effect the comparison by testing this prior against two distinct source types. We first model the priors for the source and noise processes.

2.4.1 The Prior on the Noise: Simple Gaussian

Imagine the simplest possible noise process: pure Gaussian noise, where each pixel has an equal probability $\tilde{q}$ of being inverted, and there is no dependence on the source configuration **S**. We assume that the noise *is* Gaussian, but we will only guess at its strength $\tilde{q}$. Then our constraint equation for the noise process is

$$\langle D_i S_i \rangle_l = 1 - 2\tilde{q} \tag{2.34}$$

and the average is over the likelihood distribution $\tilde{P}(\mathbf{D}|\mathbf{S})$. Our guess at the corruption process is then, from (2.17):

$$\tilde{P}(\mathbf{D}|\mathbf{S}) = \frac{1}{Z_l(\tilde{h})} \exp\left\{ \tilde{h} \sum_i D_i S_i \right\}, \tag{2.35}$$

where we have defined the field $\tilde{h}$, which couples the data to the source, as

$$\tilde{h} \stackrel{\text{def}}{=} \frac{1}{2} \log\left(\frac{1}{\tilde{q}} - 1 \right),$$

and $Z_l(\tilde{h})$ is determined by the normalization, $\sum_{\{\mathbf{D}\}} \tilde{P}(\mathbf{D}|\mathbf{S}) = 1$:

$$Z_l(\tilde{h}) = \sum_{\{\mathbf{D}\}} \exp\left\{ \tilde{h} \sum_i D_i S_i \right\}. \tag{2.36}$$

Using these results we can consider the success of the restoration scheme in the *absence* of any prior knowledge of the source. [With zero information, the only rational choice is a flat prior, $\tilde{P}(\mathbf{S})$ constant.] So given only the noise parameter $\tilde{q}$ and the data $\mathbf{D}$ the restored distribution (2.18) is simply

$$P(\mathbf{R}|\mathbf{D}) = \tilde{P}(\mathbf{D}|\mathbf{S})_{\mathbf{S}\rightarrow\mathbf{R}} = \frac{1}{Z_l(\tilde{h})} \exp\left\{\tilde{h}\sum_i D_i R_i\right\}.$$

This has the following properties.

- In the case of no noise, $\tilde{q} \rightarrow 0$ and $\tilde{h} \rightarrow \infty$, $\mathbf{R} = \mathbf{D}$ with probability 1: the restored image $\mathbf{R}$ is 'bound' to the data $\mathbf{D}$.
- When $\tilde{q} = 1$, every pixel in the data has been corrupted. Now as $\tilde{q} \rightarrow 1$, $\tilde{h} \rightarrow -\infty$, and the only $\mathbf{R}$ with significant probability is the negative of $\mathbf{D}$.
- When $\tilde{q} = \frac{1}{2}$, the data pictures will be completely random and will bear no relation to the source. In this case $\tilde{h} = 0$ and each and every possible $\mathbf{R}$ has the same low probability of $1/2^N$.

These results exhibit the behaviour on $\{\mathbf{R}\}$ that we would desire from the reconstruction scheme in these circumstances. This situation is somewhat unusual but is worth discussing further. In the absence of any knowledge about the prior we can never do better (with a single picture) than to take the reconstruction equal to the data. This is the same result as the maximum likelihood estimate for this model [i.e. finding the $\mathbf{S}$ that maximizes the likelihood function $P(\mathbf{D}|\mathbf{S})$]. However, our knowledge of the level of noise does allow us to make some statements about the degree of confidence we place in our estimate. As we increase our estimate of the noise level $\tilde{q}$, we are assuming that the true source is likely to be further and further away from the data.

2.4.2 The Prior on the Source: Edge Density

Now we fold in our prior information about the original image. We consider the case where we have an estimate of the mean density of edges ε_S, averaged over all pictures in the source distribution. We do not make any explicit statements about the bias (the excess of black pixels over white) in the source pictures, and this has the effect of implicitly modelling a zero bias. Using the edge-density prior, we can model our belief that edges are quite rare in most real world pictures and we argue that this is a sensible measure of the source distribution. To be a consistent measure requires that the spread of values of this observable over the source ensemble be not too great, otherwise it fails to characterize successfully the majority of the pictures and is a poor choice.

We impose the value of the mean density of edges ε_S as a constraint on the prior distribution:

$$\left\langle \frac{2}{\nu N} \sum_{<ij>} S_i S_j \right\rangle_p = 1 - 2\varepsilon_S, \tag{2.37}$$

where the sum is over nearest neighbour pairs of sites, and ν is the number of nearest neighbours ($\nu = 4$ for a square lattice). Following the method of Lagrange multipliers in §2.2.3 the calculation of the prior distribution yields:

$$\tilde{P}(\mathbf{S}) = \frac{1}{Z_p(\tilde{K})} \exp\left\{ \tilde{K} \sum_{<ij>} S_i S_j \right\}, \tag{2.38}$$

with the value of the coupling $\tilde{K}$ specified implicitly by the constraint

$$\left\langle \sum_{<ij>} S_i S_j \right\rangle_p = \frac{\partial \log Z_p(\tilde{K})}{\partial \tilde{K}} = \frac{\nu N}{2}(1 - 2\varepsilon_S), \tag{2.39}$$

and the normalization function is

$$Z_p(\tilde{K}) = \sum_{\{\mathbf{S}\}} \exp\left\{\tilde{K} \sum_{<ij>} S_i S_j\right\}. \tag{2.40}$$

We can place this result in the context of the parameter estimation problem to be discussed later. Given only an estimate of the mean density of edges in the source, there is only one *rational* value that may be assigned to the restoration parameter $\tilde{K}$. However, this value may not give *optimal* restoration if the mean density of edges is a poor measure with which to characterize the source.

2.4.3 The Resultant Posterior

Now we have expressions (2.38) for the prior distribution $\tilde{P}(\mathbf{S})$ and (2.35) for the likelihood $\tilde{P}(\mathbf{D}|\mathbf{S})$ and these define the posterior distribution

$$\tilde{P}(\mathbf{S}|\mathbf{D}) = \frac{\tilde{P}(\mathbf{D}|\mathbf{S})\tilde{P}(\mathbf{S})}{\tilde{P}(\mathbf{D})}, \tag{2.41}$$

with $\tilde{P}(\mathbf{D})$ given by the normalization:

$$\begin{aligned}\tilde{P}(\mathbf{D}) &= \sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{D}|\mathbf{S})\tilde{P}(\mathbf{S}) \\ &= \frac{1}{Z_p(\tilde{K})Z_l(\tilde{h})} \sum_{\{\mathbf{S}\}} \exp\left\{\tilde{K} \sum_{<ij>} S_i S_j + \tilde{h} \sum_i S_i D_i\right\}. \end{aligned} \tag{2.42}$$

Therefore the restoration process for this particular model is given by

$$P(\mathbf{R}|\mathbf{D}) = \frac{1}{Z(\tilde{K}, \tilde{h}; \mathbf{D})} \exp\left\{\tilde{K} \sum_{<ij>} R_i R_j + \tilde{h} \sum_i R_i D_i\right\}, \tag{2.43}$$

where the normalization is

$$\begin{aligned} Z(\tilde{K},\tilde{h};\mathbf{D}) &= \sum_{\{\mathbf{R}\}} \exp\left\{\tilde{K}\sum_{<ij>} R_iR_j + \tilde{h}\sum_i R_iD_i\right\} && (2.44) \\ &= \tilde{P}(\mathbf{D})Z_p(\tilde{K})Z_l(\tilde{h}). && (2.45) \end{aligned}$$

Examine the result (2.43). It sets forth the distribution for the ensemble of restored images given a particular data image. It provides a prescription for generating restored images from the data **D**, sampled with probability $P(\mathbf{R}|\mathbf{D})$. The images that appear with greatest probability minimize the cost function

$$\mathcal{H} = -\tilde{K}\sum_{<ij>} R_iR_j - \tilde{h}\sum_i R_iD_i. \tag{2.46}$$

The cost function consists of two terms:

- The first, $-\tilde{K}\sum_{<ij>} R_iR_j$, makes edges in the output **R** costly, with the magnitude of the cost dependent upon the value of the coupling constant $\tilde{K}$. This is the Lagrange multiplier and is determined by the estimated value of the density of edges ε_S in the source distribution, so the fewer edges in the source picture, the greater the cost of edges in the output. This term will tend to remove edges in the restored picture, but the level of competition with the second term in the cost function will determine just how many edges are removed from the data in generating the restored pictures.

- The second term, $-\tilde{h}\sum_i R_iD_i$, imposes a penalty for each pixel in the restored picture that differs from the presented data **D**. Therefore this term tends to align the restored picture with the data, but the strength of this tendency will be determined by the magnitude of $\tilde{h}$ which in turn depends on our estimate $\tilde{q}$ of the level of noise in the corruption process.

2.4.4 Discussion

The cost function (2.46) is equivalent to the cost function used by Geman and Geman in [36] except that we have neglected the line processes introduced there. GG concentrate on minimization of this function (to generate the MAP estimate), neglecting the more complete information available in the probability distribution (2.43). The width of this probability distribution is related to the degree of uncertainty in the prior model and the severity of the corruption process, and hence provides additional information about the level of confidence we may place in the output. We generate a whole ensemble of restored images sampled according to the probability distribution (2.43), each individually less probable than that obtained by minimizing (2.46), but nevertheless important because of the multiplicity of similar configurations.

We have arrived at (2.43) by using information theory to assign forms to the prior and estimated noise distributions. The formalism then *specifies* the values of the couplings, $\tilde{K}$ and $\tilde{h}$, that should be used for optimal restoration, based upon the edge density ε_S of the source images, and the pixel flip probability $\tilde{q}$, provided the assigned forms of the prior and likelihood are accurate. [This form of parameter estimation requires explicit knowledge of the source. Estimation of $\tilde{K}$ and $\tilde{h}$ without such information is discussed later in §2.6.] In contrast, the minimum of (2.46) depends only on the ratio of the couplings—there is one less degree of freedom. By ignoring the specified values of the couplings we are violating the constraints (2.39) and (2.34), and effectively ignoring some of the available information (recall §2.2.1).

2.5 The Test Distributions

We now specify the real source and noise distributions that we will use to test the restoration scheme. We are in effect drawing back the curtain that conceals the true source distribution and noise process in Figure 2.1. We may then consider a variety of different source distributions and noise processes, and observe the effect of a particular choice of prior distribution.

2.5.1 The Noise Process

The noise process we will consider is, as modelled in §2.4.1, pure Gaussian. We write (2.13) with $\tilde{q}_i(\mathbf{S}) = q \; \forall \mathbf{S}, i$, and the true conditional probability is:

$$\begin{aligned} P(\mathbf{D}|\mathbf{S}) &= \prod_i P(D_i|\mathbf{S}) \\ &= \exp \sum_i \log P(D_i|S_i). \end{aligned}$$

Now we can write

$$\begin{aligned} \log P(D_i|S_i) &= \delta_{D_i S_i} \log(1-q) + (1-\delta_{D_i S_i}) \log q \\ &= \log q + \frac{1}{2}(1 + D_i S_i) \log\left(\frac{1-q}{q}\right) \\ &= \frac{1}{2}\log(1-q)q + \frac{1}{2} D_i S_i \log\left(\frac{1-q}{q}\right), \end{aligned}$$

and therefore the noise process is described by

$$P(\mathbf{D}|\mathbf{S}) = \frac{1}{Z_l(h)} \exp\left\{ h \sum_i D_i S_i \right\}, \tag{2.47}$$

with

$$h = \frac{1}{2}\log\left(\frac{1-q}{q}\right), \tag{2.48}$$

and the normalization

$$Z_l(h) = [(1-q)q]^{-N/2}$$

is consistent with (2.36).

With this definition of the noise process we can calculate the average overlap of the source and data, defined as α in (2.30):

$$\begin{aligned}
\alpha = \langle D_k \langle S_k[\mathbf{D}]\rangle_S \rangle_D &= \sum_{\{\mathbf{D}\}} D_k P(\mathbf{D}) \sum_{\{\mathbf{S}\}} S_k P(\mathbf{S}|\mathbf{D}) \\
&= \sum_{\{\mathbf{S}\}} S_k P(\mathbf{S}) \sum_{\{\mathbf{D}\}} D_k P(\mathbf{D}|\mathbf{S}) \\
&= \sum_{\{\mathbf{S}\}} S_k P(\mathbf{S}) \frac{1}{[2\cosh(h)]^N} \sum_{\{\mathbf{D}\}} D_k \exp\left\{h \sum_i S_i D_i\right\} \\
&= \sum_{\{\mathbf{S}\}} S_k P(\mathbf{S}) S_k \tanh(h) \\
&= \tanh(h) = 1 - 2q.
\end{aligned} \tag{2.49}$$

If we compare (2.47) and (2.35) we see that they are, of course, equivalent: our model of the noise process $\tilde{P}(\mathbf{D}|\mathbf{S})$ correctly matches the true noise process $P(\mathbf{D}|\mathbf{S})$. However, we may not correctly guess the level of noise, i.e. $\tilde{q} \neq q$, so we will investigate the effect that the choice of the field $\tilde{h}$ has on the success of the restoration. The field h (determined by the noise level q) indicates how close the data is to the source, while the restoration parameter $\tilde{h}$ determines how close the restoration is to the data.

2.5.2 A Source Well Modelled by the Prior

We want to consider the case of a prior that is well-matched to the source distribution. Given the prior that we have chosen, we get this case if we generate the source image from the realization of a simple nearest neighbour Markov random field [128]. In this case we have:

$$P(\mathbf{S}) = \frac{1}{Z_p(K)} \exp\left\{ K \sum_{<ij>} S_i S_j \right\}, \tag{2.50}$$

with $Z_p(K)$ defined as in (2.40). Then

$$\begin{aligned} P(\mathbf{D}) &= \sum_{\{\mathsf{S}\}} P(\mathbf{D}|\mathsf{S}) P(\mathsf{S}) \\ &= \frac{1}{Z_l(h) Z_p(K)} \sum_{\{\mathsf{S}\}} \exp\left\{ K \sum_{<ij>} S_i S_j + h \sum_i S_i D_i \right\}. \end{aligned}$$

As with the noise process, we can, by varying $\tilde{K}$, investigate the effect of failing to estimate the parameter K correctly.

2.5.3 A Source Poorly Modelled by the Prior

It is of interest to determine the performance of the scheme when the prior is ill-matched to the source distribution. For these purposes we consider a single source picture, and so model the source distribution as a delta function. If we call the source picture in question $\mathsf{S}^{(0)}$, the distribution may be written:

$$P(\mathsf{S}) = \delta\left(|\mathsf{S} - \mathsf{S}^{(0)}|\right), \tag{2.51}$$

which then considerably simplifies the distribution of data pictures that may be generated:

$$\begin{aligned} P(\mathbf{D}) &= \sum_{\{\mathbf{S}\}} P(\mathbf{D}|\mathbf{S})P(\mathbf{S}) \\ &= P(\mathbf{D}|\mathbf{S}^{(0)}) \\ &= \frac{1}{Z_l(h)} \exp\left\{h \sum_i D_i S_i^{(0)}\right\}. \end{aligned}$$

This allows us to construct arbitrary synthetic pictures. For the analysis of this case in the next chapter we will consider simple chequerboard images with various edge-densities.

2.6 The Evidence for the Prior

In the early parts of this chapter we derived a framework for image restoration with a view toward conducting a systematic analysis of the performance of the restoration scheme. For the purposes of this investigation we allow ourselves complete knowledge of the source and noise processes so that we can objectively assess the success of the restoration process. Therefore, calculation of the quality factor (2.25) requires explicit knowledge of the processes that generated the data picture (i.e. we *know* the values of the generation parameters K and h).

This knowledge is only available in the context of the testing process. In any real restoration problem we will *not* have access to the values of the generation parameters. In effect we will be unable to take a peek behind the curtain in Figure 2.1. This leaves us with the problem of choosing the ‘best’ set of restoration parameters $\tilde{K}$ and $\tilde{h}$ given only access to the data picture. We require a prescription that will provide a consistent estimate of the optimal restoration parameters we should use in the absence of any knowledge of the source. If this estimation scheme is successful then we should find, when we disclose the information about the source, that these specified parameters maximize the quality factor and the overlap **T**.**S**.

The formalism that we will use for this determination is the **evidence** [46, 81]. This is a natural extension of the Bayesian methods we used earlier in the chapter to determine the reconstruction scheme initially. We first modify our notation a little, and recognize that so far we have effectively suppressed the parameter dependence of the reconstruction scheme when writing down probabilities. Our equation (2.5) for the restored distribution should, more completely, read:

$$P(\mathbf{R}|\mathbf{D};\tilde{K},\tilde{h}) = \left.\frac{\tilde{P}(\mathbf{D}|\mathbf{S};\tilde{h})\tilde{P}(\mathbf{S}|\tilde{K})}{\tilde{P}(\mathbf{D}|\tilde{K},\tilde{h})}\right|_{\mathbf{S}\to\mathbf{R}}. \tag{2.52}$$

The left-hand side represents the probability of getting reconstruction **R**, given the data **D** *and a particular choice of parameters* $\tilde{K}$ and $\tilde{h}$.

We now see that the denominator in equation (2.52) depends explicitly on the restoration parameters $\tilde{K}$ and $\tilde{h}$. This is the **evidence** provided by the data for this particular choice of $\tilde{K}$ and $\tilde{h}$. We can evaluate this, as in (2.4), in terms of the prior and model likelihood as follows:

$$\begin{aligned} P(\mathbf{D}|\tilde{K},\tilde{h}) &= \sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{D}|\mathbf{S};\tilde{h})\tilde{P}(\mathbf{S}|\tilde{K}) \\ &= \frac{1}{Z_p(\tilde{K})}\frac{1}{Z_l(\tilde{h})}\sum_{\{\mathbf{S}\}}\exp\left\{\tilde{K}\sum_{<ij>}S_iS_j + \tilde{h}\sum_i S_iD_i\right\} \\ &= \frac{Z(\tilde{K},\tilde{h};\mathbf{D})}{Z_p(\tilde{K})Z_l(\tilde{h})}. \end{aligned} \tag{2.53}$$

We can obtain the same result if we rearrange (2.52).

The simplest approach we can take to the assignment of the restoration parameters is a form of maximum likelihood estimation. According to Bayes we may write

$$P(\tilde{K},\tilde{h}|\mathbf{D}) \propto P(\mathbf{D}|\tilde{K},\tilde{h})P(\tilde{K},\tilde{h}). \tag{2.54}$$

Assuming we have no *a priori* information on the best choice of the parameters, $P(\tilde{K},\tilde{h})$ is constant, which leaves the evidence $P(\mathbf{D}|\tilde{K},\tilde{h})$ as the function we should maximize to find the most probable *a posteriori* (MAP) values $\tilde{K}^*,\tilde{h}^*$, estimated in the light of the data picture.

This idea is quite straightforward, however it is not a true Bayesian approach to simply set the values of the restoration parameters in equation (2.52) to these MAP estimates and to continue to call $P(\mathbf{R}|\mathbf{D};\tilde{K}^*,\tilde{h}^*)$ the posterior distribution. When we originally derived the posterior distribution (2.52), we assumed that the parameters $\tilde{K}$ and $\tilde{h}$ were set *a priori*. We now admit that they are not explicitly specified in the prior and model likelihood, but remain to be estimated *a posteriori* from the data. Any parameters that are not specified *a priori* cannot appear explicitly in the posterior: we obtain the posterior distribution by integrating over the unknown parameters. Thus

$$P(\mathbf{R}|\mathbf{D}) \stackrel{\text{def}}{=} \int P(\mathbf{R}|\mathbf{D};\tilde{K},\tilde{h})P(\tilde{K},\tilde{h}|\mathbf{D})d\tilde{K}d\tilde{h}. \tag{2.55}$$

As MacKay argues [81], provided the evidence, and therefore $P(\tilde{K},\tilde{h}|\mathbf{D})$, is sharply peaked around the maximum at $(\tilde{K}^*,\tilde{h}^*)$, the posterior distribution (2.55) will be dominated by (2.52) evaluated at the MAP values of the restoration parameters, i.e.

$$P(\mathbf{R}|\mathbf{D}) \simeq P(\mathbf{R}|\mathbf{D}; \tilde{K}^*, \tilde{h}^*). \tag{2.56}$$

There is an ongoing debate [83, 126] about the validity of this **evidence approximation,** and whether it is in the true spirit of Bayes.

Like the testing of the restoration scheme itself, we establish the success or failure of the method in a purely objective sense: does the approximation provide a reasonable estimate of the parameter values that maximize the quality factor? Does the success or failure depend on whether the prior is well-matched to the source?

2.7 The Statistical Mechanics Perspective

In preparation for the analytic work that follows in the next chapter, we place the results so far in the context and language of statistical physics.

We may think of the image as a statistical mechanics model of a magnetic system. Like the two-dimensional array of binary pixels in the image, we model the magnet as a two-dimensional lattice of atoms or spins that represent the crystal structure of the material. The term 'spin' arises from the quantum mechanical origin of the magnetic moment in the atom, and as we are dealing with binary site variables, what we have is known as a spin-$\frac{1}{2}$ system. The simplest and most widely studied example is the Ising model [62]. The partition function of the zero-field Ising model is just (2.40). Therefore, the source pictures generated by (2.50) are sample configurations of an Ising model.

The reconstruction system may be considered as an Ising model in a 'random' external field, i.e. the external field—the data term—is non-uniform across the lattice with regions of $D_k = +1$ interlaced with regions of $D_k = -1$. The cost function (2.46) is simply the energy function of the magnetic system and minimizing this cost function is equivalent to finding the ground state of the magnet. If we write (2.43) as

$$P(\mathbf{R}|\mathbf{D}) = \frac{1}{Z} \exp\{-\beta\mathcal{H}\}, \tag{2.57}$$

where β is an inverse temperature that scales the magnitude of the couplings $\tilde{K}$ and $\tilde{h}$, we can recover the Geman and Geman result (2.46) exactly in the zero-temperature limit ($\beta \rightarrow \infty$).

The quantities that we need in order to calculate the evidence turn out to be statistical mechanics partition functions. Finding the maximum of the evidence is equivalent to minimizing (in the sense of most negative) the free energy difference between the model posterior (restored) distribution and the prior distribution.

Numerous techniques have been developed for studying statistical systems like the Ising model—notably the mean field approximation, series expansion methods, and Monte Carlo simulation; all to be considered in the next chapter. In general it is necessary to find a way of describing the state of the system which is less tortuous than specifying the state of the spin variable at each and every site. One solution is to construct order parameters which succinctly describe the macroscopic properties of the system. For the basic Ising model the standard order parameter is the magnetization, and the behaviour of this order parameter clearly signals the phase transition in real magnetic materials between ferromagnetism at low temperatures and paramagnetism at higher temperatures. These order parameters may in general be calculated as log derivatives of the partition function, with respect to a particular conjugate field. In the case of the simple Ising model, the field conjugate to the magnetization is the external magnetic field. For the image model we may use the bias as an order parameter if we introduce an analogous uniform field into the partition function. However there are other useful order parameters for our model, namely the overlaps between the reconstruction and the source. It is these order parameters, along with the quality factor and the overlap of the TPM with the source that we will proceed to calculate and measure in the following chapter.

We have now stepped well into the field of disordered systems. We want to calculate the average value of these order parameters, but there are several ways of performing such averages where the average is over all source, data and reconstructed pictures. Reconsider the screens analogy: for each source picture selected from the source distribution we may generate a number of data pictures, and from each data picture we generate a number of reconstructions. Therefore we may not simply average over all source, data, and reconstructions simultaneously (what is known as an annealed average) but we must average over all reconstructions of a fixed data picture, before then averaging this result over all data configurations derived from a particular source and subsequently averaging over all possible source pictures. It is this 'quenching' of the disorder that leads to complications in the calculation of the order parameters and has required the introduction of various other techniques in disordered systems to deal with these quenched averages. Most notable of these is probably the replica trick [107], but this is in general useful only for models with long range interactions.

Distribution		Parameters	Normalization
Source	$P(\mathbf{S})$	K	$Z_p(K)$
True Likelihood	$P(\mathbf{D}\|\mathbf{S})$	h	$Z_l(h)$
True Posterior	$P(\mathbf{S}\|\mathbf{D})$	K, h	$Z(K, h; \mathbf{D})$
Prior	$\tilde{P}(\mathbf{S})$	$\tilde{K}$	$Z_p(\tilde{K})$
Model Likelihood	$\tilde{P}(\mathbf{D}\|\mathbf{S})$	$\tilde{h}$	$Z_l(\tilde{h})$
Model Posterior	$\tilde{P}(\mathbf{S}\|\mathbf{D})$	$\tilde{K}, \tilde{h}$	$Z(\tilde{K}, \tilde{h}; \mathbf{D})$
= Restored	$= P(\mathbf{R}\|\mathbf{D})$	ditto	ditto
Evidence	$P(\mathbf{D})$	$\tilde{K}, \tilde{h}$	

Table 2.1. A summary of the probability distributions discussed in this chapter, showing the distinction between the generation parameters K, h and the restoration parameters $\tilde{K}, \tilde{h}$.

2.8 Conclusion

We have now set out all of the basic theory we will use hereafter: see Table 2.1 for a summary of the notation. In the next chapter we will look at a software implementation of the restoration scheme and investigate the performance of the scheme on different source distributions and noise levels, as a function of the restoration parameters $\tilde{K}$ and $\tilde{h}$. We will attempt to improve our understanding of the results by analytic calculations using mean field theory and series expansion methods.

In Chapter 4 we will discuss the pros and cons of the different estimates of the source that we may derive from the restoration, and make some quantitative comparisons of the MAP and TPM estimates, again using theoretical techniques to explain some of the results.

Finally, in Chapter 5 we will develop the evidence formalism introduced in §2.6, both analytically and through the use of Monte Carlo simulations.

CHAPTER 3

Exploring the Prior: Phase Transitions in Hypothesis Space

3.1 Introduction

The restoration scheme has been constructed in such a way as to incorporate *prior* information, both on the density of edges in the source image, and on a random Gaussian noise process. These assumptions have determined the functional form of the posterior probability distribution for the restored images. We have yet to specify the values of the restoration parameters, which will be estimated based upon the likely density of edges in the source, and the assumed severity of the degradation. The particular values we choose, and their accuracy, reflect the prior knowledge available when we attempt the reconstruction.

In this chapter we explore the hypothesis space of the restoration parameters: we investigate how the success of the restoration scheme depends upon the appropriateness of the prior. There is freedom to choose the prior at two levels. First the functional form of the prior must be determined—we restrict ourselves to a prior on the density of edges in the source. Subsequently, whatever the functional form, there will be certain parameters to be fixed: $(\tilde{K}, \tilde{h})$. Therefore, we analyse two distinct cases: the **well-matched prior** where the functional form of the prior matches the generation process—we still have to choose the restoration parameters appropriately; and the **ill-matched prior**—we use the edge-density prior to restore pictures generated from a fixed chequerboard source.

We investigate the parameter dependence of a number of observable properties of the reconstructions: in particular, the quality factor and the overlap of the TPM estimate with the source. Using a Monte Carlo simulation of the restoration process, we first show that the optimal Bayesian choice of the restoration parameters *does* provide the optimal restoration (maximizes the quality factor and T.S) for the case of the well-matched prior. We then apply the techniques of mean field theory in order to predict and explain the qualitative behaviour of the model. We find **phase transitions** (discontinuous changes in the qualitative behaviour of the model) as we vary the restoration parameters, and we are able to explain these in terms of the relative free energies of metastable states. Subsequently we carry out a small coupling expansion of the quantities required to calculate the quality factor, which allows us to confirm some of the earlier results on the optimal parameter choice.

3.2 Simulation of the Restoration Scheme

In this section we describe how the model of image restoration developed in the previous chapter may be implemented in software, and the desired measurements made. First a word about the terminology. Computational physics work is usually concerned with the *simulation* of a model of a physical system. In this case the image restoration model does not match any physical process and it makes more sense to describe the software as an *implementation* of the scheme rather than a simulation. However, we are in the business of *testing* the restoration scheme, not implementing it, and the source picture we use is simulated rather than real. So with deference to our statistical mechanics background, we will continue to refer to 'simulation', and this also conveniently distinguishes it from any possible hardware implementation.

3.2.1 Monte Carlo Methods

Monte Carlo simulation is based upon the use of pseudo-random numbers to generate a Boltzmann distribution that satisfies a given energy function. There are two techniques commonly used in the restoration of binary images—namely the Metropolis algorithm [88] and the Gibbs sampler [36], although this latter method is better known as 'heat bath' in statistical mechanics (see e.g. [63]). Both methods rely upon the concept of **importance sampling.** In order to make measurements of macroscopic observable quantities we wish to average over all possible microscopic configurations of the system with a weighting factor proportional to the probability of finding the system in that configuration. So for an energy function $\mathcal{H}(\mathbf{R})$

we have a normalization, or partition function

$$Z = \sum_{\{\mathbf{R}\}} \exp\left[-\mathcal{H}(\mathbf{R})\right]. \tag{3.1}$$

[The Z stands for the German **Zustandsumme,** which means quite literally "sum over states".] The average value of a function f of the site variables is calculated as

$$\langle f \rangle = \frac{1}{Z} \sum_{\{\mathbf{R}\}} f(\mathbf{R}) \exp\left\{-\mathcal{H}(\mathbf{R})\right\}. \tag{3.2}$$

However, for any reasonably sized system (and in most condensed matter problems, the system size is of the order of Avogadro's number), the calculation of this sum is unfeasible due to the huge number of configurations to be considered.

The elegant way around this, first proposed by Metropolis *et al.* [88] is to generate the configurations with a probability of occurrence already given by the Boltzmannn distribution

$$P(\mathbf{R}) = \frac{1}{Z} \exp\left\{-\mathcal{H}(\mathbf{R})\right\}. \tag{3.3}$$

We may then obtain an average from a simple unweighted sum over configurations—the weighting factor has been taken care of by the way we *sample* configurations according to their relative *importance.*

How do we go about ensuring that each configuration occurs with the required probability? First, we write down the the balance equation, which states that for a system in equilibrium the rate of transition into a state must equal the transition rate out of that state. Thus considering a state labelled A:

$$P(A) \sum_{B} P(A \to B) = \sum_{B} P(B) P(B \to A). \tag{3.4}$$

A more restrictive condition which satisfies the balance equation (3.4) is the **detailed balance** condition, which states that for all states A and B, the rate of transition from state A to state B should equal the rate from B to A:

$$P(A)P(A \to B) = P(B)P(B \to A) \tag{3.5}$$

Therefore, in order that configurations A and B occur with the correct *relative* probabilities, we simply have to choose $P(A \to B)$ and $P(B \to A)$ so as to satisfy (3.5).

If $P(A)$ and $P(B)$ satisfy the Boltzmannn distribution, then

$$P(A) = \frac{1}{Z} \exp\{-\mathcal{H}(A)\}, \tag{3.6}$$

and

$$\frac{P(A)}{P(B)} = \exp\{-[\mathcal{H}(A) - \mathcal{H}(B)]\}. \tag{3.7}$$

Therefore for detailed balance (3.5) we require

$$\frac{P(B \to A)}{P(A \to B)} = \exp\{-[\mathcal{H}(A) - \mathcal{H}(B)]\}. \tag{3.8}$$

The Metropolis algorithm implements this condition by choosing

$$\begin{aligned} P(A \to B) &= \min\{1, \exp\{-[\mathcal{H}(B) - \mathcal{H}(A)]\}\}, \\ P(B \to A) &= \min\{1, \exp\{[\mathcal{H}(B) - \mathcal{H}(A)]\}\}. \end{aligned} \tag{3.9}$$

With these definitions $P(A \to B)$ and $P(B \to A)$ are guaranteed to meet equation (3.8): we say that the simulation satisfies detailed balance. Therefore the Metropolis algorithm consists of the following steps.

1. Choose a new configuration B.

2. Generate a random number p between zero and one.

3. Accept the change to the new configuration if the random number p is less than the transition probability $P(A \rightarrow B)$ given in (3.9).

Provided the method for choosing the new configuration in step 1 allows all possible states to be visited (i.e. the simulation is **ergodic**), then after very many iterations the probability of finding the system in a particular state will converge to the Boltzmann distribution.

On the other hand, the Gibbs sampler chooses the probabilities such that

$$P(B \rightarrow A) \propto P(A) \tag{3.10}$$

which trivially satisfies detailed balance. For simple update rules the calculation of $P(A)$ may be straightforward: e.g. for a single spin flip we choose the new value of the spin (independent of the old value) according to the probability

$$P(B \rightarrow A) = \frac{\exp\left[-\mathcal{H}(A)\right]}{2\cosh\left[\mathcal{H}(A)\right]}.$$

However, for more complex models such as continuous valued spins, the calculation of $P(A)$ can be arduous since it requires calculation of the normalization term involving a sum over all possible states. It is computationally simpler to implement Metropolis Monte Carlo dynamics since the update decision depends only on the calculation of $\Delta\mathcal{H} = \mathcal{H}(A) - \mathcal{H}(B)$. [See [31] for a discussion on the relative merits of the two methods.] Indeed, this flexibility allows entirely arbitrary changes in the system configuration. This is the basis for cluster updating methods [118] used to reduce the time spent in metastable regions and reduce the effects of critical slowing down. It is very much up to the experimenter to decide on the particular update rule. The Metropolis algorithm guarantees that detailed balance is satisfied, and it only remains for the experimenter to ensure that the simulation is ergodic. One can thus choose the update rule that provides the quickest equilibration. We choose to use the standard Metropolis dynamics, implemented using a chequerboard site visitation schedule to avoid the effective loss of ergodicity that may occur at low temperatures.

3.2.2 The Model

Following the idea of the screens introduced back in §2.1.3 and Figure 2.1, we have three two-dimensional binary arrays representing the source, data and restoration screens.

- The picture on the source screen may be generated in one of two ways. It may be a sample configuration from an Ising distribution, in which case we choose a value of the coupling K, and equilibrate the system before selecting a typical configuration. Alternatively, we may consider a fixed source, not generated by any statistical process; in practice we will use various sizes of chequerboard.

- The data screen contains the corrupted image generated from the source by the noise process. In a real application, it is only this screen which would be available to us—we exploit knowledge of the true source to enable us to *test* the restoration scheme, not to implement it. The picture on the data screen is generated from a particular source picture by simply flipping pixels randomly with a probability of corruption q at each site. This simulates random Gaussian noise.

- The restoration screen does the real work of the restoration scheme. It is upon this screen that appear configurations sampled according to the distribution $P(\mathbf{R}|\mathbf{D})$. The two restoration parameters $\tilde{K}$ and $\tilde{h}$ may be set arbitrarily, and the initial configuration set to an Ising ground state (i.e. all one colour), copied from the data screen, or simply generated randomly.

These three screens represent the probability distributions we consider. We next introduce a fourth screen on which we will display our best estimate of the source picture. Once the restoration distribution has reached equilibrium we perform a vector sum over a sequence of configurations on the restored screen. We generate the binary array $\mathbf{T}$ by thresholding this vector sum, so

$$\mathbf{T} = \vec{\mathrm{sgn}} \left\{ \sum_{\{\mathbf{R}\}} \mathbf{R} P(\mathbf{R}|\mathbf{D}) \right\} .$$

This is the TPM estimate, and we obtain a simple measure of the success of the restoration by evaluating the overlap $\mathbf{T}.\mathbf{S}$ of this estimate with the source.

3.2.3 Implementation Notes

The simulation code was implemented in C with various tricks used to promote maximum efficiency, notably:

- calculating all the transition probabilities once only, and storing the values in a look-up table;
- using pointers to pointers [*sic*] to implement the boundary conditions and chequerboard visitation schedule;

- using in-line code for the random number generator.

These techniques increase the initialization time (and memory requirements), but significantly improve the speed of the most frequently executed inner-loop code.

The random number generator used was RMARIN [65, 86], the first of a new generation of very long period generators with a period of 2^{141} and used extensively by the QCD Grand Challenge project at Edinburgh.

Many simulations were carried out on 24Mip Unix workstations with particularly intensive work performed on a 16-node i860 supercomputer.

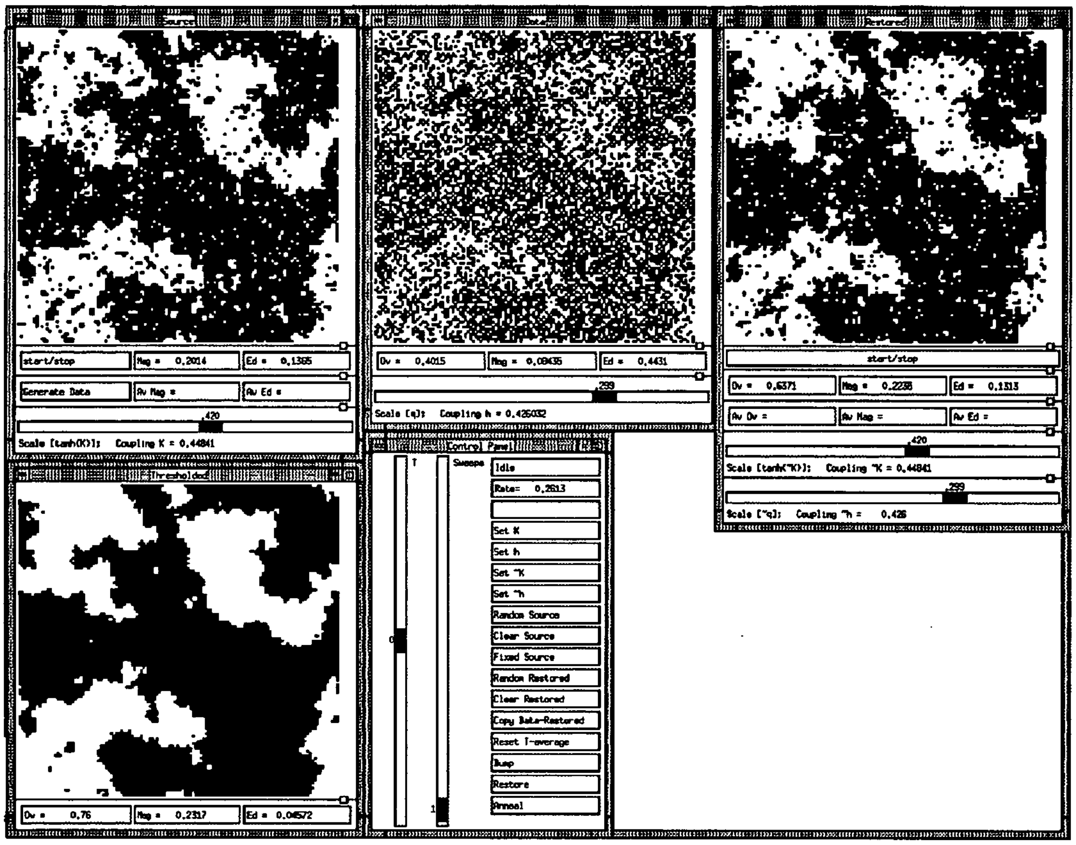
An X-window interface was implemented using the Motif Widget set and the Xt toolkit yielding graphical output such as is shown in Figure 3.1 [100] . This proved a most useful visualization tool. The calculation subroutines were common between both interactive X usage and overnight batch processing. In addition, the state of the 'screens' could be saved by the batch processes, on both the workstations and the supercomputer, and subsequently loaded during an interactive session for visualization purposes.

3.2.4 Results

Although we will perform a systematic analysis of the restoration process, others have produced much previous work that has relied on a rather *ad hoc* visual assessment of the reconstruction (e.g. [36, 40, 35]). When we try to visualize the processes involved in image restoration it is useful to be able to see the reconstructions that result in different parameter regimes, and for this reason we will present a number of sample pictures of the restoration process at work. However this is not a concise way of presenting the results, and a simple qualitative assessment of a small number of restored pictures will not enable one to perceive all of the trends and properties that may be discovered in a comprehensive set of quantitative results.

Figure 3.1 is a screen-dump of the interactive X-window interface, and shows the computional realization of the screens approach from Figure 2.1. We set the source coupling K as we wish and then run the Monte Carlo process on the 'Source' screen until the system has equilibrated and we have a representative picture from the source distribution. The other generation parameter, the noise level q, is set and the picture on the 'Data' screen is generated from the source by randomly flipping pixels with probability q. The restoration parameters $\tilde{K}$ and $\tilde{h}$ are set and the Monte Carlo process is run on the 'Restored' screen so that the pictures are displayed according to the posterior probability distribution. Finally the picture on the 'Thresholded' screen is generated by averaging over the pictures appearing on the restored screen and then applying a threshold to recover a binary image.

For each picture we measure the overlap with the source picture, the bias (magnetization) of the picture, and the density of edges in the image. If we examine the 'Restored' screen, we see the need for an estimator such as the TPM, shown on the 'Thresholded' screen: configurations appearing on the restored screen reflect the uncertainty (the width of the reconstruction distribution) in the level of 'entropic' noise (small-scale short-lived fluctuations). In fact, for these optimal restoration parameters, the width of the reconstruction distribution matches the width of the source distribution and the pictures on the 'Restored' screen agree qualitatively with the source. However these 'entropic' fluctuations, which give the image



Mag = 0.2014
Ed = 0.1365
Av Mag =
Av Ed =
Scale [tanh(K)]: Coupling K = 0.44841

Ov = 0.4015
Mag = 0.08435
Ed = 0.4431
Scale [q]: Coupling h = 0.425032

Rate= 0.2613

Ov = 0.6371
Mag = 0.2238
Ed = 0.1313
Av Ov =
Av Mag =
Av Ed =
Scale [tanh(~K)]: Coupling ~K = 0.44841
Scale [~q]: Coupling ~h = 0.426

Ov = 0.76
Mag = 0.2317
Ed = 0.04572

Figure 3.1. The X-Window interactive. Four screens are shown: the original image **S** is shown on the Source screen (top left) and is a sample configuration from an Ising distribution close to criticality (in order that there be a reasonable degree of long range structure in the image); the corrupted picture **D** was generated from the source with 30% noise and is displayed on the Data screen (top middle); a configuration from the restoration ensemble $\{\mathbf{R}\}$ is shown on the Restored screen (top right) [optimal restoration parameters $\tilde{K} = K$ and $\tilde{h} = h$ were used]; and the TPM estimate T obtained by averaging over the pictures in $\{\mathbf{R}\}$ is displayed on the Thresholded screen (bottom left). The overlap of the source and data is 0.4. The restored pictures **R** have an average overlap with the source of 0.64, while the TPM estimate has an overlap with the source of 0.76.

a textured appearance, are purely random and the TPM estimate shows a significant improvement of overlap with the source. Figure 3.1 shows just how effective the reconstruction process can be.

In Figure 3.2 we show some sample output of the restoration process on fixed chequerboard sources. We show three cases: one a relatively easy restoration with a low density of edges in the source and a low noise level; and two more difficult restorations, one with a high noise level and the other with a high density of edges in the source. As before the restoration process is quite successful in quantitative terms. Qualitatively, in the high noise case, the restoration does not much resemble a chequerboard, but one can make out the homogeneous regions that constitute the squares of the chequerboard.

Notice in the 4x4 chequerboard case that typical pictures from the posterior distribution are less like the source than the data, but that the mean of the distribution, shown on the thresholded screen, represents an improvement over the data. The restoration scheme is very successful in smoothing and removing random noise from large homogeneous regions. However, it has difficulty in successfully modelling the corners of the squares. This arises from the simple form of the nearest neighbour interaction in the prior. Consequently the restoration of small chequerboards is made more difficult.

Now that we have developed a qualitative 'feel' for what is going on in the reconstruction process, we move on to a systematic investigation of the model for different source types and restoration parameters.

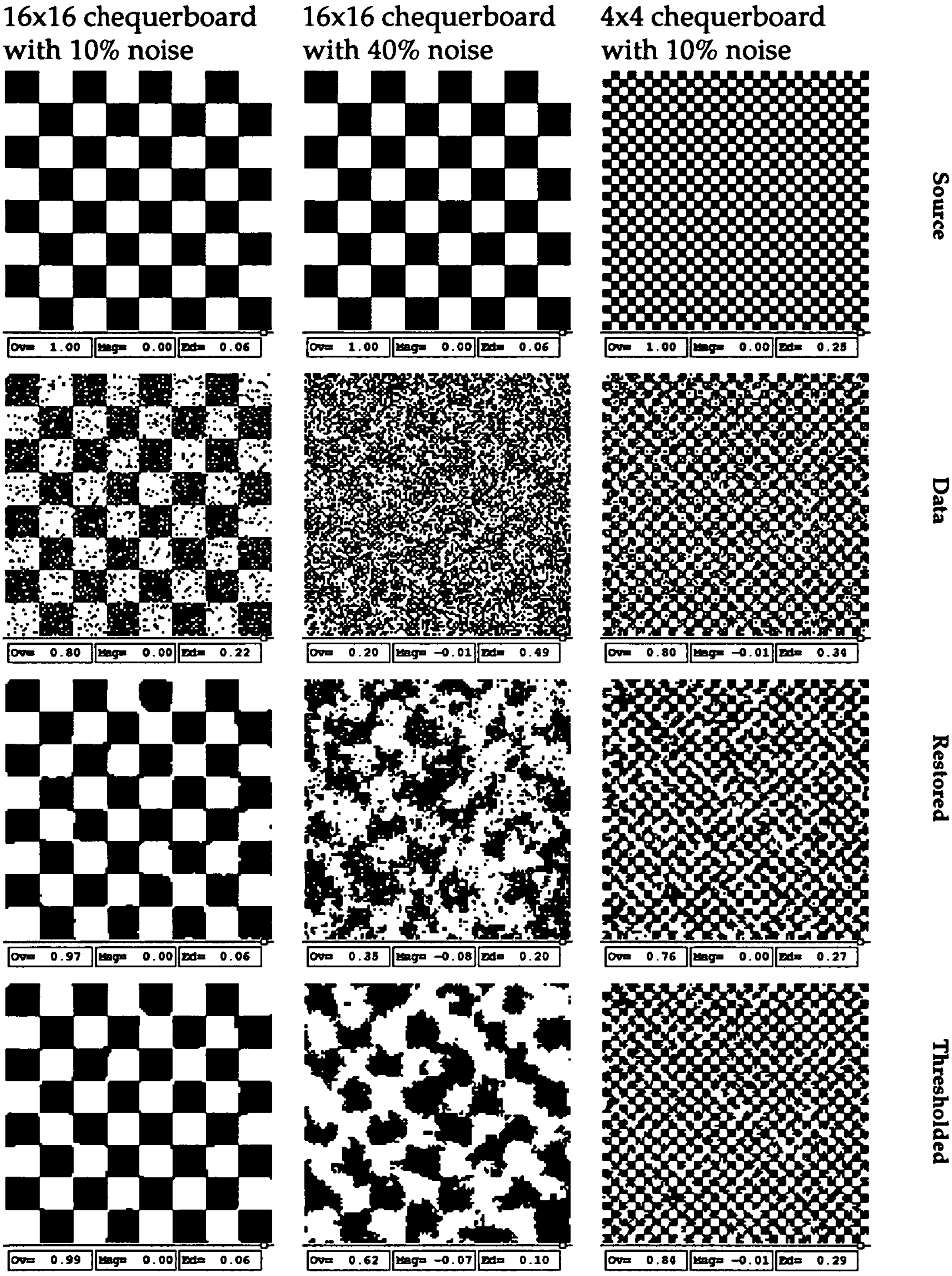

16x16 chequerboard with 10% noise
16x16 chequerboard with 40% noise
4x4 chequerboard with 10% noise
Source
Ov= 1.00 Mag= 0.00 Ed= 0.06
Ov= 1.00 Mag= 0.00 Ed= 0.06
Ov= 1.00 Mag= 0.00 Ed= 0.25
Data
Ov= 0.80 Mag= 0.00 Ed= 0.22
Ov= 0.20 Mag= -0.01 Ed= 0.49
Ov= 0.80 Mag= -0.01 Ed= 0.34
Restored
Ov= 0.97 Mag= 0.00 Ed= 0.06
Ov= 0.35 Mag= -0.08 Ed= 0.20
Ov= 0.76 Mag= 0.00 Ed= 0.27
Thresholded
Ov= 0.99 Mag= 0.00 Ed= 0.06
Ov= 0.62 Mag= -0.07 Ed= 0.10
Ov= 0.84 Mag= -0.01 Ed= 0.29

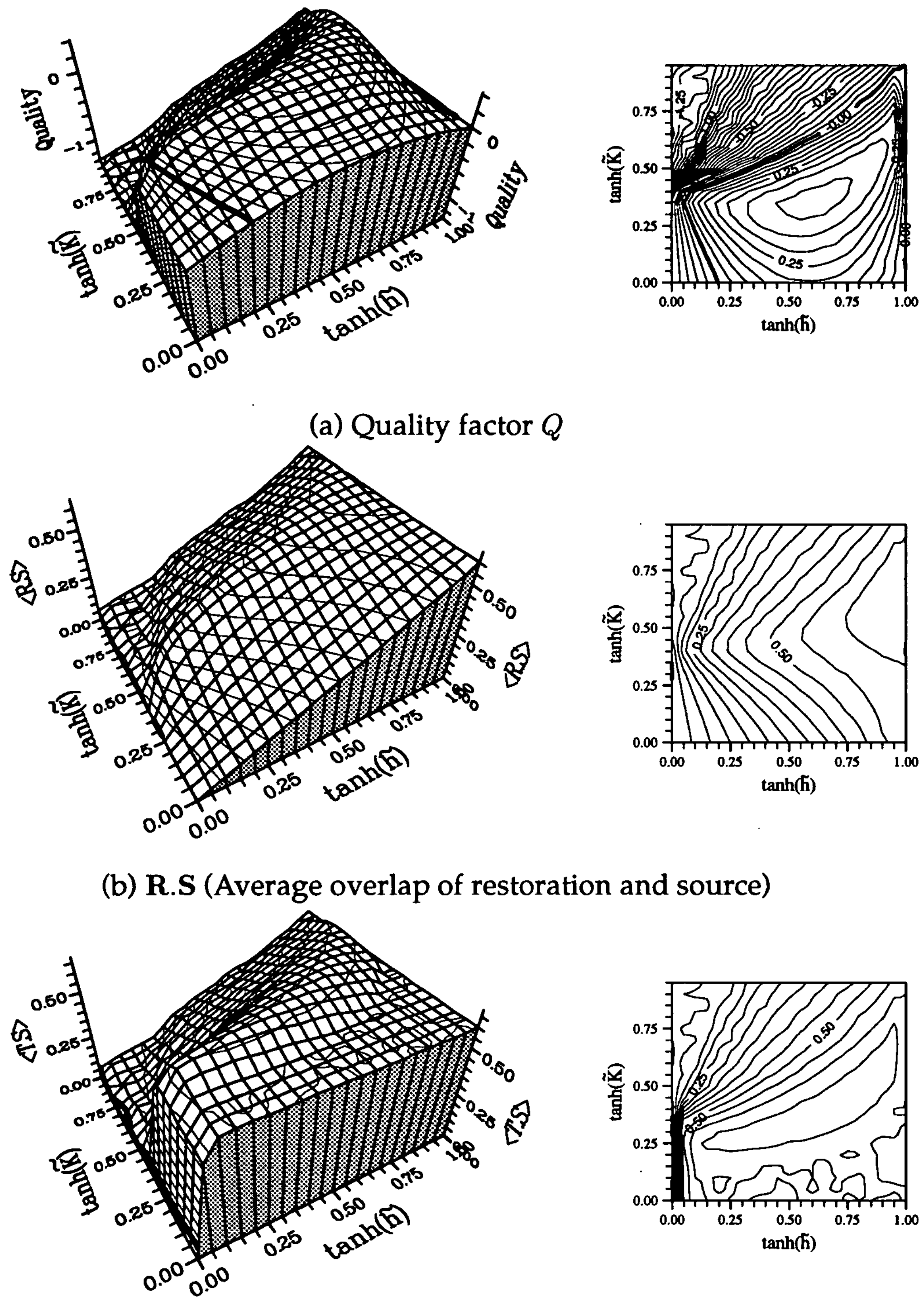
Figure 3.2. Some example chequerboard restorations. The restoration of the 16x16 chequerboard with 10% noise shown in the left-hand column is very successful. With 40% noise the quantitative performance is quite reasonable although it is fairly hard to see the chequerboard in the restoration. When the density of edges is increased, as in the 4x4 chequerboard in the right-hand column, the restoration problem is made more difficult.

A Prior Matched to the Source Distribution

The first question we ask is how good the restoration scheme is when the prior is well-matched to the source distribution. We always consider an edge-density prior (which leads to a nearest neighbour interaction in the prior probability function), so for this investigation we use a nearest neighbour MRF (or Ising model) to generate the source.

All simulation results show quenched averages, where we have averaged each of the quantities over 50 different instances of the source and data. Figure 3.3 shows the results we obtain for a range of prior parameters $\tilde{K}$ and $\tilde{h}$ with the source coupling and noise level (i.e. K and h) fixed at arbitrarily chosen values. Although we have chosen just a single point in the space of generation parameters, the qualitative results are typical of those obtained throughout the space.

The Bayesian derivation of the restoration scheme leads us to believe that we will see optimal performance when the prior exactly models the true generation process. Since we are using the correct functional form for the prior, this means that we should set the restoration parameters equal to the generation parameters, i.e. $\tilde{K} = K$ and $\tilde{h} = h$. It is clear from Figure 3.3(a) that the quality factor is maximized at this point, and furthermore from Figure 3.3(c) that the TPM estimate has maximal overlap with the source at this same point in parameter space: if the quality factor and TPM estimate are defined consistently the maxima of the quality factor and of the overlap $\mathbf{T}.\mathbf{S}$ must coincide.



(a) Quality factor Q

(b) R.S (Average overlap of restoration and source)

(c) T.S (Average overlap of TPM estimate and source)

Figure 3.3. Simulation results for Ising source with the density of edges in the source $\varepsilon_S = 0.25$ and 20% noise, $q = 0.2$. $[\tanh(K) \simeq 0.36, \tanh(h) = 0.6.]$ See the main text on page 82 for discussion.

Figure 3.3(b) shows the average overlap of the restoration pictures with the source. Notice that this quantity is maximized for large values of $\tilde{K}$ and $\tilde{h}$. When we discussed Figure 3.1 it was pointed out that the pictures on the restoration screen contained a degree of entropic, random statistical noise which is removed in the TPM estimate. However, as we increase the values of $\tilde{K}$, $\tilde{h}$ and walk up the ridge in Figure 3.3(b), we reduce the width of the restored distribution and remove this entropic noise, improving the average overlap of the distribution with the source picture. This property is utilized in order to generate the MAP estimate, but we believe that the TPM estimate provides a more consistent way of exploiting the posterior distribution. We will discuss this further in Chapter 4.

Notice that both overlaps **T.S** and **R.S** fall away to zero in the upper left of the parameter space (when $\tilde{K}$ is much larger than $\tilde{h}$) and that the quality factor is similarly poor in the same region. It has been noted before that restoration schemes fail for certain poor choices of parameter [84, 85], but these regions are usually avoided and no explanation has been attempted. Later in this chapter we will turn to analytic methods to explain this undesired behaviour.

How does the quality factor change around the maximum? It is clear from the definition of the quality factor that, whatever the values we choose for the generation parameters K and h, we will find that we get optimal restoration when $\tilde{K} = K \textit{ and } \tilde{h} = h$. However, when we get *one* aspect of the prior wrong, this affects the optimal choice of parameters for *other* aspects of the prior. Examine the 2D plot in Figure 3.3(a) and see that the principal axes of the closed contours around the optimal position are not parallel to the graph axes—the highest points lie diagonally across parameter space. This means that if we fix one of the parameters incorrectly (e.g. choose $\tanh(\tilde{K}) = 0.5$), we must adjust the other parameter from its optimal Bayesian value if we want to maximize the quality factor (in this case we must then choose $\tanh(\tilde{h}) \simeq 0.8$). This effective coupling between the parameters is exhibited again in the next example.

An Ill-matched Prior Model

What happens when the source image is synthetic, rather than drawn from an ensemble? Since the prior is still modelling an Ising source distribution, it is impossible to choose restoration parameters that exactly match the generation process. Nevertheless, we might hope that the optimal choice of the restoration parameter $\tilde{h}$ would continue to be determined by the level of noise in the corruption process, and that the optimal choice of the restoration parameter $\tilde{K}$ would be related to the density of edges in the source picture (since our derivation of the prior on the source back in §2.4.2 used only information about the density of edges). However, we have already been given a preview of the kind of behaviour to expect when the prior is ill-matched to the source.

The density of edges of various chequerboards is shown in Table 3.1. For each of these edge-densities we have calculated the corresponding Ising coupling: the value of K that would generate source pictures with the specified density of edges *if the source distribution were Ising.* Notice that there is a large variation in ε_S for a fairly small change in coupling $\tanh(K_{\text{eff}})$. In Figure 3.4 we plot the density of edges in a typical Ising configuration as a function of the coupling. Most source images we are interested in (i.e. having a small but finite density of edges) fall into a narrow band of coupling values.

Chequer Size	Edge Density	Ising Coupling	
	$[\varepsilon_S]$	$[K_{\text{eff}}]$	$\tanh(K_{\text{eff}})$
3x3	0.333	0.28768	0.280
4x4	0.25	0.37855	0.361
8x8	0.125	0.44841	0.421
16x16	0.0625	0.50154	0.463

Table 3.1. K_{eff} is the effective Ising coupling that would generate source pictures with the density of edges ε_S.

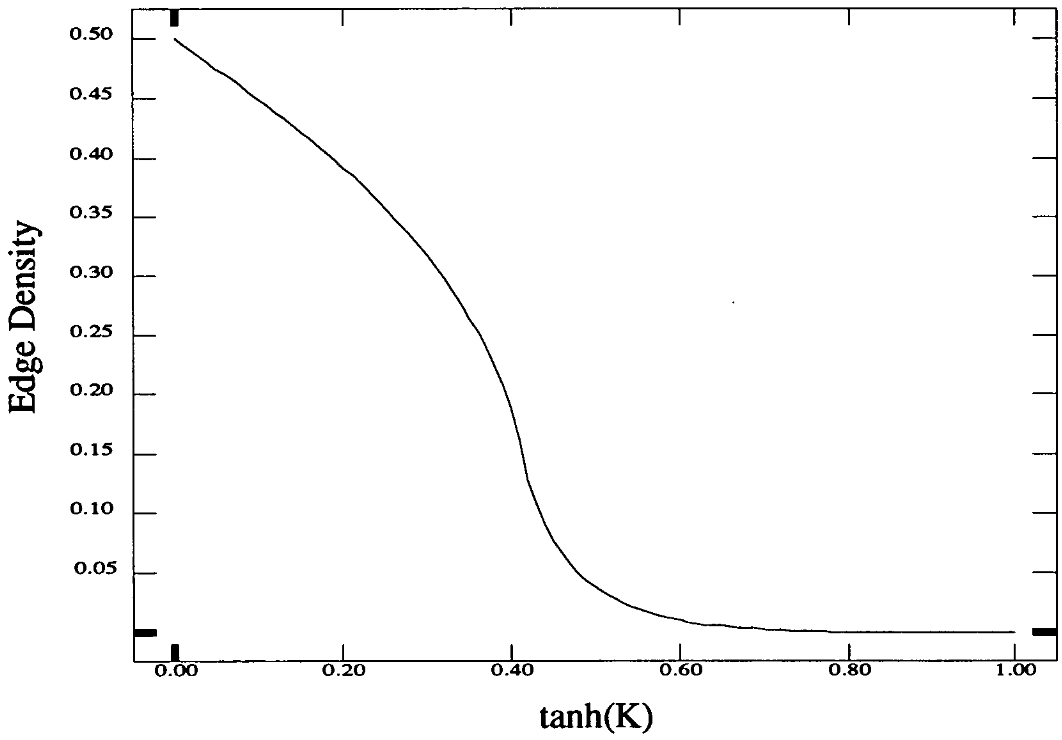
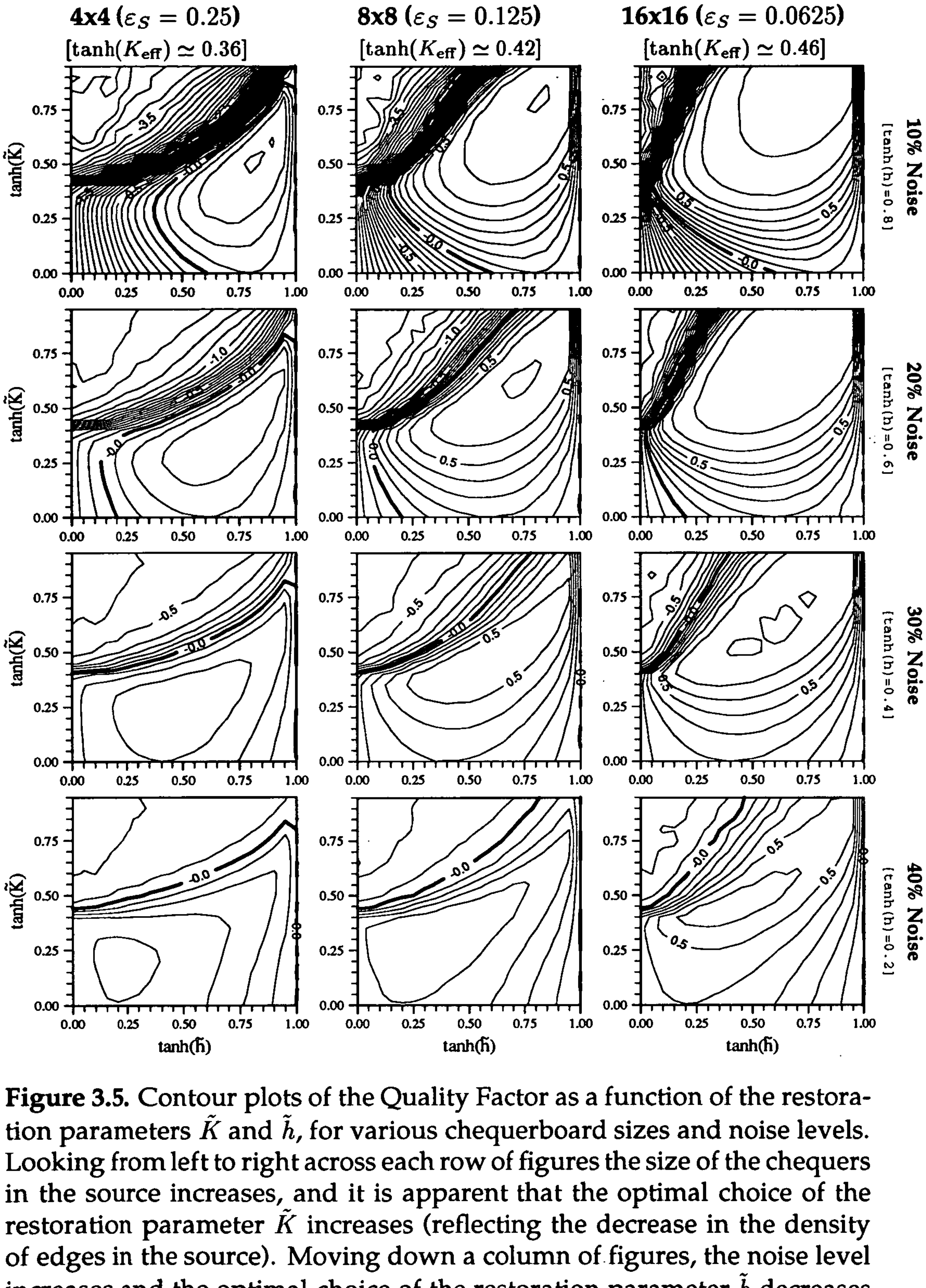


Figure 3.4. Plot of edge density versus Ising coupling. The data was generated using the exact solution of the 2D Ising model by Onsager [93] (see e.g. [59]). Notice the small range of couplings that produces images with edge densities in the domain of interest.

Figure 3.5 shows contour plots of the quality factor for three different chequer sizes: 4x4, 8x8, and 16x16 pixels square. The relevant edge-density and the corresponding effective coupling $K_{\rm eff}$ from Table 3.1 are shown at the head of each column. [Note that $K_{\rm eff}$ is presented for comparison only and is not a source generation parameter: the source image is simply a fixed chequerboard.] The level of noise and the corresponding value of h is shown at the right end of each row.

It is still the case that a low density of edges in the original source picture requires the use of a high value of the restoration coupling $\tilde{K}$, and vice versa. And similarly a low noise level requires a high value for $\tilde{h}$ and vice versa. However, it is apparent that the naive assignment of the restoration parameters, $\tilde{K} = K_{\rm eff}$ and $\tilde{h} = h$, does not maximize the quality factor. Indeed the optimal value for $\tilde{K}$ depends not only on the density of edges, but also on the level of noise. We know from the Ising source example that if one of the prior parameters is fixed to be suboptimal, then the best choice of the other parameter is dependent on the value of the first. In this case the simple prior based only upon the density of edges is inadequate to describe the chequerboard source and this inadequacy couples the optimal values of $\tilde{K}$ and $\tilde{h}$. They *both* depend on the chequerboard size *and* the level of noise.

Figure 3.5 can only summarize the comprehensive investigation of hypothesis space by simulation. Since we are exploring a four dimensional parameter space [K or ε_S, h, $\tilde{K}$ and $\tilde{h}$] there is an enormous amount of data to be represented, but for more detailed analysis we will pick a typical point in the space of generation parameters K and h. Further simulation results are presented alongside analytic results in Figures 3.8, 3.9, 3.14, and 3.15, following the analytic calculations to which we now turn.



4x4 ($\varepsilon_S = 0.25$) [$\tanh(K_{\text{eff}}) \simeq 0.36$]

8x8 ($\varepsilon_S = 0.125$) [$\tanh(K_{\text{eff}}) \simeq 0.42$]

16x16 ($\varepsilon_S = 0.0625$) [$\tanh(K_{\text{eff}}) \simeq 0.46$]

10% Noise [tanh(h)=0.8]

20% Noise [tanh(h)=0.6]

30% Noise [tanh(h)=0.4]

40% Noise [tanh(h)=0.2]

Figure 3.5. Contour plots of the Quality Factor as a function of the restoration parameters $\tilde{K}$ and $\tilde{h}$, for various chequerboard sizes and noise levels. Looking from left to right across each row of figures the size of the chequers in the source increases, and it is apparent that the optimal choice of the restoration parameter $\tilde{K}$ increases (reflecting the decrease in the density of edges in the source). Moving down a column of figures, the noise level increases and the optimal choice of the restoration parameter $\tilde{h}$ decreases in line with the decrease of the noise parameter h. However, still moving down a column the optimal choice of $\tilde{K}$ also decreases: the optimal choice of $\tilde{K}$ is affected by the noise parameter h.

3.3 The Mean Field Approximation

3.3.1 The Need for Approximation Methods

The first thing we must recognize is that calculation of the quantities that we have identified in the quality factor is a *hard* problem, requiring the computation of **quenched** averages. For example, we want to calculate the average value of, say, the overlap **R**.**D**, a quantity which depends upon both the restored picture and the data picture. To find this average we have to perform the weighted sum over all possible configurations of **R** and **D**. Since the distribution of restored pictures has an explicit dependence on the data, we must sum over all configurations of **R** for a fixed **D**, and subsequently average over all instances of the data. We say that the disorder in **D** is fixed, or quenched during the average over **R**. In algebraic terms:

$$\begin{aligned}\langle\langle f(\mathbf{R},\mathbf{D})\rangle_R\rangle_D &= \sum_{\{\mathbf{D}\}}\sum_{\{\mathbf{R}\}} P(\mathbf{R},\mathbf{D})f(\mathbf{R},\mathbf{D}) \\ &= \sum_{\{\mathbf{D}\}} P(\mathbf{D})\left[\sum_{\{\mathbf{R}\}} P(\mathbf{R}|\mathbf{D})f(\mathbf{R},\mathbf{D})\right].\end{aligned}$$

An analogous physical situation is found in a binary alloy [19] where at low temperatures the diffusion of the two atomic species occurs on a far longer timescale than other processes such as the evolution of the magnetization.

Calculations of such quenched averages have been carried out fully in only a very few cases (notably the Sherrington-Kirkpatrick spin glass [107]). Analytic progress therefore seems possible only with the aid of simplifications and approximations. For the purposes of the mean field calculation we:

1. analyse the case of a single fixed source rather than an ensemble of source images;
2. restrict exploration of the landscape of the model to the simplest non-trivial subspace; and
3. employ the mean field approximation in its variational formulation, to analyse the restricted space.

The first simplification allows us to dispense with one level of quenched average. Recall that the data depends explicitly on the source picture via the distribution $P(\mathbf{D}|\mathbf{S})$. Hence for the general quenched average of a function f we should write:

$$\langle\langle\langle f\rangle\rangle\rangle = \sum_{\{\mathbf{S}\}} P(\mathbf{S}) \sum_{\{\mathbf{D}\}} P(\mathbf{D}|\mathbf{S}) \sum_{\{\mathbf{R}\}} P(\mathbf{R}|\mathbf{D}) f.$$

The source variables are quenched with respect to the data in just the same way that the data is quenched with respect to the reconstructions—we have in effect three different timescales.

Therefore, we write the source distribution $P(\mathbf{S})$ as a delta function as in (2.51) and we can replace the general quenched free energy

$$\begin{aligned} F &= -\left\langle\left\langle \log Z(\tilde{K},\tilde{h};\mathbf{D})\right\rangle_D\right\rangle_S \\ &= -\sum_{\{\mathbf{S}\}} P(\mathbf{S}) \sum_{\{\mathbf{D}\}} P(\mathbf{D}|\mathbf{S}) \log Z(\tilde{K},\tilde{h};\mathbf{D}), \end{aligned}$$

by

$$\begin{aligned} F &= -\left\langle \log Z(\tilde{K},\tilde{h};\mathbf{D})\right\rangle_D \\ &= -\sum_{\{\mathbf{D}\}} P(\mathbf{D}|\mathbf{S}^0) \log Z(\tilde{K},\tilde{h};\mathbf{D}). \end{aligned}$$

Fixing S also allows us to specify the order parameters as functions of the source image. The quantities that we wish to measure are to be identified from these order parameters, which are computed from the free energy F by differentiating with respect to appropriate conjugate fields (introduced in the next section).

For the second simplification we employ the simplest set of **order parameters** that give meaningful results. [A set of order parameters is a simplified set of coordinates with which we may describe the macrostate of a statistical mechanics system, without resorting to the N coordinates of the full specification.] Consider for example the quenched average

$$\begin{aligned}\left\langle\langle\langle R_k S_k\rangle_R\rangle_D\right\rangle_S &\equiv \langle\langle R_k[\mathbf{D}]\rangle_R \langle S_k[\mathbf{D}]\rangle_S\rangle_D \\ &= \left\langle\left\langle R_k[\mathbf{S}^0]\right\rangle_R\right\rangle_D S_k^0,\end{aligned}$$

given the first simplification of the fixed source case. We must expect that there will be as many different values of this average as there are distinct sites in the source image—in general a multiplicity of order parameters. In order to keep the calculation tractable, we examine only the restricted space where

$$\left\langle\left\langle R_k[\mathbf{S}^0]\right\rangle_R\right\rangle_D = \begin{cases} R_+ & \text{when } S_k^0 = +1, \\ R_- & \text{when } S_k^0 = -1, \end{cases} \tag{3.11}$$

and to compute these order parameters we need to introduce conjugate field terms into the configurational energy.

Finally, the third simplification is the mean field approximation, originally proposed by Weiss [125] as a phenomenological theory of ferromagnetism. We may model a magnetic material as an array of atoms with magnetic moments or spins (corresponding to our pixel coordinates $\mathbf{S}, \mathbf{D}$, and $\mathbf{R}$) which interact with one another via a coupling K as in our model. In the theory as proposed by Weiss the interactions between spins on the lattice are approximated by an effective molecular field, proportional to the overall magnetization of the system. In other words, the effective field experienced by any particular spin is calculated by averaging over the fluctuations in all the other spins in the system.

Over the years, many approximation schemes have been suggested which ultimately reduce to mean field. They all share the property that they neglect fluctuations in the local molecular field at a site, but they frequently arise from quite different assumptions. The method we will present here is known as the variational approach (see e.g. [63]) and is based upon the minimization of a variational free energy.

3.3.2 The Variational Method

The full partition function, from which we could calculate all properties of interest is just (2.44) with an additional conjugate field term H_i (later set to zero), which allows us to calculate the quenched averages as derivatives of the free energy $F = -\log Z(\tilde{K}, \tilde{h}; \mathbf{D})$. Thus

$$Z(\tilde{K}, \tilde{h}; \mathbf{D}) = \sum_{\{\mathbf{R}\}} \exp\left\{ \tilde{K} \sum_{<ij>} R_i R_j + \tilde{h} \sum_i R_i D_i + \sum_i H_i R_i \right\}. \tag{3.12}$$

Then, for example:

$$\begin{aligned}\langle R_k[\mathbf{D}]\rangle_R &\overset{\text{def}}{=} \frac{1}{Z(\tilde{K},\tilde{h};\mathbf{D})}\sum_{\{\mathbf{R}\}} R_k \exp\left\{\tilde{K}\sum_{<ij>} R_iR_j + \tilde{h}\sum_i R_iD_i + \sum_i H_iR_i\right\} \\ &= \frac{\partial}{\partial H_k}\log Z(\tilde{K},\tilde{h};\mathbf{D}). \qquad (3.13)\end{aligned}$$

We begin our mean field calculation by first writing down an approximation for Z; a factorized expression, the only sort for which we can perform an explicit computation.

$$Z_V[\mathbf{S},\mathbf{D}] \overset{\text{def}}{=} \sum_{\{\mathbf{R}\}} \exp\left\{\sum_i R_i\left[H_i + g(D_i,S_i,\{H^\alpha\})\right]\right\}. \qquad (3.14)$$

The function $g(D_i,S_i,\{H^\alpha\})$ introduces the set of fields $\{H^\alpha\}$ conjugate to the order parameters we will calculate. At this stage we will keep the discussion quite general, and specify the explicit form of the function later. We then define

$$A = \tilde{K}\sum_{<ij>} R_iR_j + \tilde{h}\sum_i R_iD_i - \sum_i R_ig(D_i,S_i,\{H^\alpha\}), \qquad (3.15)$$

and write the identity

$$\begin{aligned}Z(\tilde{K},\tilde{h};\mathbf{D}) &= \sum_{\{\mathbf{R}\}}\exp\left\{\tilde{K}\sum_{<ij>} R_iR_j + \tilde{h}\sum_i R_iD_i + \sum_i H_iR_i\right\} \\ &= Z_V\left(\frac{1}{Z_V}\sum_{\{\mathbf{R}\}}\exp\left\{\sum_i R_i\left[H_i + g(D_i,S_i,\{H^\alpha\})\right] + A\right\}\right) \\ &= Z_V\left\langle \exp A\right\rangle_V,\end{aligned}$$

where $\langle\ldots\rangle_V$ represents an average over the factorized measure Z_V.

We next make use of the well known convexity property of the exponential function

$$\langle \exp A \rangle \geq \exp \langle A \rangle \tag{3.16}$$

which holds for any function A, and any measure that we may choose to average over.

Therefore we can write the true free energy:

$$\begin{aligned} F &= -\log Z(\tilde{K}, \tilde{h}; \mathbf{D}) \\ &\leq -\left[\log Z_V(\tilde{K}, \tilde{h}; \mathbf{S}, \mathbf{D}, \{H^\alpha\}) + \langle A \rangle_V\right] \\ &\overset{\text{def}}{=} \mathcal{F}(\{H^\alpha\}). \end{aligned}$$

The variational free energy $\mathcal{F}(\{H^\alpha\})$ is thus an upper bound on the value of the true free energy F. We may now take advantage of the arbitrariness of the set of variational fields $\{H^\alpha\}$ to find the infimum of this upper bound. The set of equations

$$\frac{\partial}{\partial H^\alpha} \mathcal{F}(\{H^\alpha\}) = 0, \tag{3.17}$$

determines the values of the $\{H^\alpha\}$ that minimize $\mathcal{F}(\{H^\alpha\})$. This procedure finds the optimal set of values for $\{H^\alpha\}$, the set that gives $\mathcal{F}(\{H^\alpha\})$ closest to the true free energy, given our initial choice of variational fields. Thus we define the mean field free energy

$$F_{MF} \overset{\text{def}}{=} \inf_{\{H^\alpha\}} \mathcal{F}(\{H^\alpha\}), \tag{3.18}$$

and we simply replace the true free energy by this 'best' estimate throughout the calculation of the order parameters. Hence

$$\langle R_k[\mathbf{D}] \rangle_R = -\frac{\partial F}{\partial H_k}$$

$$\begin{aligned}\to\ & -\frac{\partial F_{MF}}{\partial H_k}\\ =\ & \frac{\partial \log Z_V}{\partial H_k}+\frac{\partial \langle A\rangle_V}{\partial H_k}\\ =\ & \frac{\partial \log Z_V}{\partial H_k}\overset{\text{def}}{=}\langle R_k[\mathbf{S},\mathbf{D}]\rangle_V\,, \end{aligned}\tag{3.19}$$

and the relationship between the order parameters in the full theory (3.13) and in the mean field approximation (3.19) is made explicit. Now that it has served its purpose we may set $H_i = 0$ in (3.12).

3.3.3 The Choice of Order Parameters

Since Z_V is a factorized measure we can proceed with the explicit computation of these order parameters. We have already indicated in (3.11) the order parameters we will use, but let us step back for a moment and consider the expressions we could possibly choose for the variational partition function Z_V, [or rather, for $g(\mathbf{S},\mathbf{D},\{H^\alpha\})$]. We will choose the simplest form that allows us to make meaningful calculations of the quality factor (in the mean field approximation). Consider four possibilities:

$$\begin{aligned} g_1(D_i,H,\tilde{h}) &\overset{\text{def}}{=} \frac{1}{2}(1+D_i)(H+\tilde{h})+\frac{1}{2}(1-D_i)(H-\tilde{h})\\ g_2(D_i,H^+,H^-,\tilde{h}) &\overset{\text{def}}{=} \frac{1}{2}(1+D_i)(H^++\tilde{h})+\frac{1}{2}(1-D_i)(H^--\tilde{h})\\ g_3(D_i,S_i,H^+,H^-,\tilde{h}) &\overset{\text{def}}{=} \frac{1}{4}(1+S_i)\left[(1+D_i)(H^++\tilde{h})+(1-D_i)(H^+-\tilde{h})\right]\\ &\quad+\frac{1}{4}(1-S_i)\left[(1+D_i)(H^-+\tilde{h})+(1-D_i)(H^--\tilde{h})\right]\\ g_4(D_i,S_i,H^{++},\ldots,\tilde{h}) &\overset{\text{def}}{=} \frac{1}{4}(1+S_i)\left[(1+D_i)(H^{++}+\tilde{h})+(1-D_i)(H^{+-}-\tilde{h})\right]\\ &\quad+\frac{1}{4}(1-S_i)\left[(1+D_i)(H^{-+}+\tilde{h})+(1-D_i)(H^{--}-\tilde{h})\right] \end{aligned}$$

These are the only possible choices for g if we wish to parameterize the source by just its bias and density of edges.

The first two expressions, g_1 and g_2, turn out to be too restrictive to give any meaningful results. They are also not dependent on the source picture in any way, which makes the measurement of any of the quantities of interest, such as overlaps with the source, a hopeless task.

We choose g_3. We are effectively dividing the sites of the lattice into four different classes dependent upon the value of S_i and D_i at each site. But we introduce only two variational fields H^+ and H^-. Wherever the source image has $S_i = +1(-1)$ we approximate the effective molecular field in the reconstruction lattice by $H^+(H^-)$. Since there is also the external field $\tilde{h}$ coupled to the data at each site, we have four degrees of freedom in g_3.

For g_4 we introduce four variational fields depending on S_i *and* D_i at each site. However, we still have only four degrees of freedom and it turns out that the four equations for $H^{++}, H^{+-}, H^{-+}, H^{--}$ obtained using g_4 are not independent and reduce to the two equations obtained from g_3. Hence, we define

$$\begin{aligned}
Z_V(H^+, H^-, \tilde{h}; \mathbf{D}, \mathbf{S}) &= \sum_{\{\mathbf{R}\}} \exp\left\{\sum_i R_i \left[\frac{1}{4}(1+S_i)(1+D_i)(H^+ + \tilde{h})\right.\right. \\
&\quad + \frac{1}{4}(1+S_i)(1-D_i)(H^+ - \tilde{h}) + \frac{1}{4}(1-S_i)(1+D_i)(H^- + \tilde{h}) \\
&\quad \left.\left. + \frac{1}{4}(1-S_i)(1-D_i)(H^- - \tilde{h}) + H_i\right]\right\} \\
&= 2^N \prod_i \cosh\left[\frac{1}{4}(1+S_i)(1+D_i)(H^+ + \tilde{h})\right. \\
&\quad + \frac{1}{4}(1+S_i)(1-D_i)(H^+ - \tilde{h}) + \frac{1}{4}(1-S_i)(1+D_i)(H^- + \tilde{h}) \\
&\quad \left. + \frac{1}{4}(1-S_i)(1-D_i)(H^- - \tilde{h}) + H_i\right].
\end{aligned}$$

Now, due to the binary nature of S_i and D_i, when we take the logarithm of Z_V, each site picks out only one term in the argument of the cosh:

$$\begin{aligned}\log Z_V(H^+, H^-, \tilde{h}; \mathbf{D}, \mathbf{S}) = \frac{1}{4}\sum_i \Big[&(1+S_i)(1+D_i)\log\cosh(H^+ + \tilde{h} + H_i)\\ &+ (1+S_i)(1-D_i)\log\cosh(H^+ - \tilde{h} + H_i)\\ &+ (1-S_i)(1+D_i)\log\cosh(H^- + \tilde{h} + H_i)\\ &+ (1-S_i)(1-D_i)\log\cosh(H^- - \tilde{h} + H_i)\Big] + N\log 2. \end{aligned} \quad (3.20)$$

In order to calculate quenched averages we wish to average the value of the free energy over all instances of the data, and this requires that we find the average of $\log Z_V$,

$$\Big\langle \log Z_V(H^+, H^-, \tilde{h}; \mathbf{D}, \mathbf{S})\Big\rangle_D = \sum_{\{\mathbf{D}\}} P(\mathbf{D}) \log Z_V(H^+, H^-, \tilde{h}; \mathbf{D}, \mathbf{S}) \quad (3.21)$$

However, when we perform the sum over i in (3.20) and take the thermodynamic limit $N \rightarrow \infty$, we find that $\log Z_V$ does not in fact depend on the particular instance of the data $\mathbf{D}$, but only upon the noise level q. We say that $\log Z_V$ **self-averages**:

$$\begin{aligned}\lim_{H_i \to 0} \Big\langle \log Z_V(H^+, H^-, \tilde{h}; \mathbf{D}, \mathbf{S})\Big\rangle_D &= N\log 2\\ &+ \frac{N}{2}(1-q)\log\cosh(H^+ + \tilde{h}) + \frac{N}{2}q\log\cosh(H^+ - \tilde{h})\\ &+ \frac{N}{2}(1-q)\log\cosh(H^- - \tilde{h}) + \frac{N}{2}q\log\cosh(H^- + \tilde{h}), \end{aligned} \quad (3.22)$$

where we have assumed zero bias in the source. For such systems it is a standard tenet of statistical mechanics that observables are dominated by their most probable values, and that the average and most probable values are essentially the same. Similarly, we argue that the value of $\log Z_V$ for any $\mathbf{D}$ chosen randomly from the same probability distribution $P(\mathbf{D}|\mathbf{S}^0)$

should be independent of the actual choice of **D**. It is easy to see how this self-averaging property arises—as we go from site to site in the sum on i in (3.20) we are choosing N independent D_i from the probability distribution $P(\mathbf{D}|\mathbf{S}^0)$ which is uniform over the sites. Therefore, for large N, this sum over sites is equivalent to an average over the distribution.

An alternative explanation imagines that we divide the system into a large number of subsystems. Each of the subsystems may be imagined as a different instance of the data. Therefore, if we perform the spatial sum over each subsystem first, the full spatial sum, averaging over subsystems, gives us the average over the data for free.

With this definition of Z_V, we must declare A as

$$A = \tilde{K} \sum_{<ij>} R_i R_j - \frac{H^+}{2} \sum_i R_i(1 + S_i) - \frac{H^-}{2} \sum_i R_i(1 - S_i), \tag{3.23}$$

and to find the mean field free energy we must calculate $\langle A \rangle_V$. First:

$$\begin{aligned}\langle R_i[\mathbf{S}, \mathbf{D}] \rangle_V &= \lim_{H_i \to 0} \frac{\partial}{\partial H_i} \log Z_V(H^+, H^-, \tilde{h}; \mathbf{D}, \mathbf{S}) \\ &= \frac{1}{4}(1 + S_i)(1 + D_i) \tanh(H^+ + \tilde{h}) + \frac{1}{4}(1 + S_i)(1 - D_i) \tanh(H^+ - \tilde{h}) \\ &\quad + \frac{1}{4}(1 - S_i)(1 + D_i) \tanh(H^- + \tilde{h}) + \frac{1}{4}(1 - S_i)(1 - D_i) \tanh(H^- - \tilde{h}).\end{aligned} \tag{3.24}$$

We then define a pair of order parameters

$$\begin{aligned}R^+ &\stackrel{\text{def}}{=} \frac{1}{N} \sum_i (1 + S_i) \langle R_i \rangle_V \\ &= \frac{1}{N} \sum_i \frac{1}{2}(1 + S_i) \left[(1 + D_i) \tanh(H^+ + \tilde{h}) + (1 - D_i) \tanh(H^+ - \tilde{h})\right] \\ &= (1 - q) \tanh(H^+ + \tilde{h}) + q \tanh(H^+ - \tilde{h}),\end{aligned} \tag{3.25}$$

$$\begin{aligned} R^- &\overset{\text{def}}{=} \frac{1}{N}\sum_i (1 - S_i) \langle R_i \rangle_V \\ &= (1-q)\tanh(H^- - \tilde{h}) + q\tanh(H^- + \tilde{h}). \end{aligned} \tag{3.26}$$

R^+ and R^- are defined as the mean bias of two complementary subsets of sites: reconstruction sites where $S_i = +1$ and $S_i = -1$ respectively. We see from (3.25) and (3.26) that these quantities self-average and therefore that they also represent the quenched average $\langle\langle R_k[\mathbf{S}, \mathbf{D}]\rangle_V\rangle_D$ over the same two sets of sites.

Next we make use of the fact that Z_V is a factorized measure and therefore $\langle R_i R_j \rangle_V = \langle R_i \rangle_V \langle R_j \rangle_V$. So when we calculate $\sum_{<ij>} \langle R_i \rangle_V \langle R_j \rangle_V$ from (3.23) we get sixteen terms of the form

$$C_1 = \frac{1}{16} \sum_{<ij>} (1 + S_i)(1 + D_i)(1 + S_j)(1 + D_j) \tanh^2(H^+ + \tilde{h}). \tag{3.27}$$

Given that the density of edges in $\mathbf{S}$ is ε_S and the noise level is q at each site, C_1 self-averages to give

$$C_1 = (1 - \varepsilon_S)(1 - q)^2 \frac{\nu N}{2} \tanh^2(H^+ + \tilde{h}), \tag{3.28}$$

where ν is the coordination number of the lattice. Finally we can write

$$\begin{aligned} \langle A \rangle_V &= \frac{N\tilde{K}\nu}{2}\left[\frac{1}{2}(1 - \varepsilon_S)\left({R^+}^2 + {R^-}^2\right) + \varepsilon_S R^+ R^-\right] \\ &\quad - \frac{NH^+R^+}{2} - \frac{NH^-R^-}{2}, \end{aligned} \tag{3.29}$$

and we see that the mean field free energy does not depend explicitly on the data picture. Nor does it depend on the *particular* source picture $\mathbf{S}^0$ that generated the data, but only upon the density of edges in the source, ε_S.

We can now write down the variational free energy, combining equations (3.29) and (3.22):

$$\begin{aligned}
\mathcal{F}(H^+, H^-) &= -\frac{1}{N}\log Z_V - \frac{1}{N}\langle A\rangle_V \\
&= -\frac{1}{2}(1-q)\log\cosh(H^+ + \tilde{h}) - \frac{1}{2}q\log\cosh(H^+ - \tilde{h}) \\
&\quad - \frac{1}{2}(1-q)\log\cosh(H^- - \tilde{h}) - \frac{1}{2}q\log\cosh(H^- + \tilde{h}) \\
&\quad - \frac{\tilde{K}\nu}{2}\left[\frac{1}{2}(1-\varepsilon_S)\left({R^+}^2 + {R^-}^2\right) + \varepsilon_S R^+ R^-\right] \\
&\quad + \frac{H^+R^+}{2} + \frac{H^-R^-}{2} - \log 2, \qquad (3.30)
\end{aligned}$$

and we find the mean field free energy by minimizing $\mathcal{F}$ with respect to the variational fields H^+, H^-. The values of the order parameters R^+ and R^- at this minimum define the equilibrium values, and thus the observables of the system.

Figure 3.6 shows the typical free energy surfaces that (3.30) defines in terms of the order parameters R^+ and R^-, for a number of different points in the parameter space $(\tilde{K}, \tilde{h})$. The values of H^+ and H^- are given in terms of the order parameters by rearranging equations (3.25) and (3.26). Notice that the surface changes qualitatively for different values of $\tilde{K}$ and $\tilde{h}$ and there is in general more than one local minimum in this surface. These are the metastable states of the system that feature in some regions of parameter space.

From (3.25) we see that the order parameter, R^+ depends on the conjugate field H^+ but not on H^-. Therefore

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial H^+} &= \left\{ -\frac{\nu \tilde{K}}{2} \left[(1-\varepsilon_S)R^+ + \varepsilon_S R^- \right] + \frac{H^+}{2} \right\} \frac{\partial R^+}{\partial H^+}, & (3.31) \\ \frac{\partial \mathcal{F}}{\partial H^-} &= \left\{ -\frac{\nu \tilde{K}}{2} \left[(1-\varepsilon_S)R^- + \varepsilon_S R^+ \right] + \frac{H^-}{2} \right\} \frac{\partial R^-}{\partial H^-}. & (3.32) \end{aligned}$$

We find candidate minima of $\mathcal{F}$ by requiring $\partial \mathcal{F}/\partial H^+ = \partial \mathcal{F}/\partial H^- = 0$ which results in a pair of equations for the fields:

$$\begin{aligned} H^+ &= \nu \tilde{K} \left[(1-\varepsilon_S)R^+ + \varepsilon_S R^- \right], & (3.33) \\ H^- &= \nu \tilde{K} \left[(1-\varepsilon_S)R^- + \varepsilon_S R^+ \right]. & (3.34) \end{aligned}$$

We may now substitute these values for the conjugate fields H^+ and H^- into our equations (3.25) and (3.26) for the order parameters to get a pair of coupled self-consistent equations for the equilibrium values R^+, R^-

$$\begin{aligned} R^+ &= (1-q) \tanh \left\{ \tilde{K}\nu \left[(1-\varepsilon_S)R^+ + \varepsilon_S R^- \right] + \tilde{h} \right\} \\ &\quad + q \tanh \left\{ \tilde{K}\nu \left[(1-\varepsilon_S)R^+ + \varepsilon_S R^- \right] - \tilde{h} \right\}, & (3.35) \\ R^- &= (1-q) \tanh \left\{ \tilde{K}\nu \left[(1-\varepsilon_S)R^- + \varepsilon_S R^+ \right] - \tilde{h} \right\} \\ &\quad + q \tanh \left\{ \tilde{K}\nu \left[(1-\varepsilon_S)R^- + \varepsilon_S R^+ \right] + \tilde{h} \right\}. & (3.36) \end{aligned}$$

3.3.4 Limiting Behaviour in Special Cases

We can check that the above equations for the order parameters are reasonable for limiting values of $\tilde{K}$ (the coupling that specifies our prior on the density of edges), and $\tilde{h}$ (the field in the model likelihood that specifies how noisy we believe the data to be). Recalling our definitions of

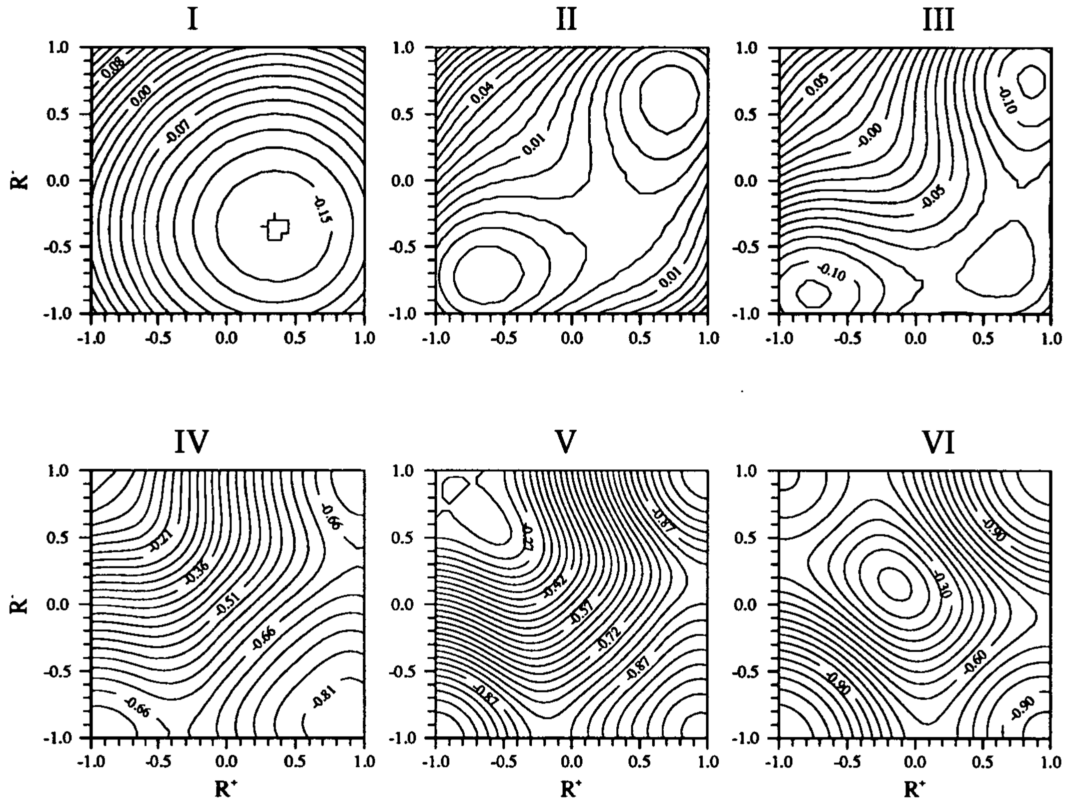


Figure 3.6. Free energy surfaces for the mean field model with density of edges $\varepsilon_S = 0.125$ and 30% noise. There is one local minimum in (I), two in (II), three in (III) and (IV), and four in (V) and (VI). In (I), (IV) and (V) the global minimum is data-like (lies on the $R^+ = -R^-$ line), while for the others there are two equal prior-like minima. The regions of parameter space that these correspond to will be indicated in Figure 3.7. Notice the symmetry about the $R^+ = -R^-$ line; this indicates that there is in fact only one degree of freedom in the equilibrium values of the order parameters, and we will use this fact to simplify the calculation of these values.

R^+ and R^- it is clear that the quantity $\frac{1}{2}(R^+ + R^-)$ is the overall bias in the reconstruction, while $\frac{1}{2}(R^+ - R^-)$ measures the overlap between the reconstruction and the source image $\mathbf{S}^0$.

- In the limit as $\tilde{K} \to \infty$ we find two stable solutions $R^+ = R^- = \pm 1$. These give a net bias of ± 1 and overlap of zero, and correspond to the two edge-free (i.e. single colour) ground states. This is easily explained: the infinite nearest neighbour coupling overwhelms any finite value of $\tilde{h}$ and the smoothing effect of the prior removes *all* edges from the image.

- In the limit of $\tilde{h} \to \infty$ we get, for any $\tilde{K}$,

$$\begin{aligned} R^+ &= (1-2q), \\ R^- &= -(1-2q). \end{aligned}$$

  Thus the bias remains zero as in the data, and the overlap with $\mathbf{S}$ retains the same value as the initial overlap of $\mathbf{D}$ with $\mathbf{S}$. This occurs because the restored picture is bound to the data by the infinite coupling $\tilde{h}$.

- At $\tilde{K} = 0$, for finite $\tilde{h}$ we have

$$\begin{aligned} R^+ &= (1-2q)\tanh(\tilde{h}), \\ R^- &= -(1-2q)\tanh(\tilde{h}). \end{aligned}$$

  This gives a bias of zero and overlap of $(1-2q)\tanh(\tilde{h})$. So for any finite $\tilde{h}$, this restoration scheme simply adds further noise to the corrupted picture, reducing the overlap.

- For $\tilde{h} = 0$ at finite $\tilde{K}$ we have

$$\begin{aligned} R^{+} &= \tanh\left\{\nu\tilde{K}\left[(1-\varepsilon_S)R^{+} + \varepsilon_S R^{-}\right]\right\}, \\ R^{-} &= \tanh\left\{\nu\tilde{K}\left[(1-\varepsilon_S)R^{-} + \varepsilon_S R^{+}\right]\right\}. \end{aligned}$$

One stable solution of these equations has $R^{+} = R^{-}$, in which case we have the implicit equation $R^{+} = \tanh(\nu\tilde{K}R^{+})$ This has either one solution at $R^{+} = R^{-} = 0$ (i.e. a completely random picture), or two prior-like solutions (i.e. few edges), depending upon the value of the prior coupling $\tilde{K}$. The result here is wholly determined by the prior.

3.3.5 Numerical Calculation of General Solutions

Reassured that the equations (3.35) and (3.36) for the order parameters are reasonable, we turn our attention to the numerical calculation of R^{+} and R^{-} and the corresponding mean field free energy (the minimum of the variational free energy).

We wish to solve the two coupled implicit equations for R^{+} and R^{-}. We can reduce these to a single implicit equation in H^{+} as follows. From (3.33) and (3.34) we find

$$H^{-} = \left(\frac{1-\varepsilon_S}{\varepsilon_S}\right)\left[H^{+} - \nu\tilde{K}(1-\varepsilon_S)R^{+}\right] + \nu\tilde{K}\varepsilon_S R^{+}. \qquad (3.37)$$

Then substituting this into (3.26) and the result back into (3.33) we get

$$\begin{aligned}
H^+ &= \nu\tilde{K}\varepsilon_S(1-q)\tanh\left\{\left(\frac{1-\varepsilon_S}{\varepsilon_S}\right)\left[H^+ - \nu\tilde{K}(1-\varepsilon_S)R^+\right] + \nu\tilde{K}\varepsilon_S R^+ - \tilde{h}\right\} \\
&+ \nu\tilde{K}\varepsilon_S q\tanh\left\{\left(\frac{1-\varepsilon_S}{\varepsilon_S}\right)\left[H^+ - \nu\tilde{K}(1-\varepsilon_S)R^+\right] + \nu\tilde{K}\varepsilon_S R^+ + \tilde{h}\right\} \\
&+ \nu\tilde{K}(1-\varepsilon_S)R^+ \qquad (3.38)
\end{aligned}$$

with R^+ given in terms of H^+ by (3.25). So we can solve for H^+ numerically using a simple, robust method such as bisection [99], and from this determine H^-, R^+ and R^-.

This method will find the fixed points of (3.38), but not all will be minima of the free energy. To ensure that we have found a value for H^+ that corresponds to a local minimum of $\mathcal{F}$ we calculate second derivatives. When the fixed point equations are satisfied we have $\partial\mathcal{F}/\partial H^+ = \partial\mathcal{F}/\partial H^- = 0$ and the second derivatives are, from (3.31) and (3.32),

$$\begin{aligned}
\left.\frac{\partial^2\mathcal{F}}{\partial {H^+}^2}\right|_{\partial\mathcal{F}/\partial H^+=0} &= \frac{1}{2}\frac{\partial R^+}{\partial H^+}\left\{1 - \nu\tilde{K}(1-\varepsilon_S)\frac{\partial R^+}{\partial H^+}\right\}, \\
\left.\frac{\partial^2\mathcal{F}}{\partial {H^-}^2}\right|_{\partial\mathcal{F}/\partial H^-=0} &= \frac{1}{2}\frac{\partial R^-}{\partial H^-}\left\{1 - \nu\tilde{K}(1-\varepsilon_S)\frac{\partial R^-}{\partial H^-}\right\}, \\
\frac{\partial^2\mathcal{F}}{\partial H^+\partial H^-} &= -\nu\tilde{K}\varepsilon_S\frac{\partial R^+}{\partial H^+}\frac{\partial R^-}{\partial H^-}.
\end{aligned}$$

A local minimum in $\mathcal{F}$ requires $\partial^2\mathcal{F}/\partial {H^+}^2 > 0$ and the Jacobian

$$J = \frac{\partial^2\mathcal{F}}{\partial {H^+}^2}\frac{\partial^2\mathcal{F}}{\partial {H^-}^2} - \left(\frac{\partial^2\mathcal{F}}{\partial H^+\partial H^-}\right)^2 > 0. \qquad (3.39)$$

Once we are satisfied that we have determined all of the minima of $\mathcal{F}$ we can then determine which of these is the global minimum corresponding to the equilibrium free energy, and which are metastable states. Although we will concentrate on the equilibrium state, we should note that metastable states do figure significantly in certain simulation regimes.

The expression for the mean field free energy, substituting (3.33) and (3.34) into (3.30), is

$$\begin{aligned} F_{MF} &= -\frac{1}{2}(1-q)\log\cosh(\nu\tilde{K}[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] + \tilde{h}) \\ &\quad - \frac{1}{2}q\log\cosh(\nu\tilde{K}[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] - \tilde{h}) \\ &\quad - \frac{1}{2}(1-q)\log\cosh(\nu\tilde{K}[(1-\varepsilon_S)R^- + \varepsilon_S R^+] - \tilde{h}) \\ &\quad - \frac{1}{2}q\log\cosh(\nu\tilde{K}[(1-\varepsilon_S)R^- + \varepsilon_S R^+] + \tilde{h}) \\ &\quad - \frac{\tilde{K}\nu}{2}\left\{\frac{1}{2}(1-\varepsilon_S)[{R^+}^2 + {R^-}^2] + \varepsilon_S R^+ R^-\right\} \\ &\quad + \frac{\tilde{K}\nu R^+}{2}[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] + \frac{\tilde{K}\nu R^-}{2}[(1-\varepsilon_S)R^- + \varepsilon_S R^+] \\ &\quad - \log 2 \end{aligned} \tag{3.40}$$

We can check (3.40) in the same limits of $\tilde{K}$ and $\tilde{h}$ as we did back in §3.3.4. First let us calculate the true free energy $F = -\frac{1}{N}\log Z$ for three special cases where

$$Z = \sum_{\{\mathbf{R}\}} \exp\left\{\tilde{K}\sum_{<ij>} R_i R_j + \tilde{h}\sum_i R_i D_i\right\}.$$

1. For $\tilde{K} \to \infty$ the minimum energy configuration dominates the configurational sum—the entropy term is zero and the free energy is just the configurational energy of the Ising ground state

$$F(\tilde{K} \to \infty) = -\frac{\nu\tilde{K}}{2}.$$

2. For $\tilde{h} \to \infty$, the entropy is zero once again. $\mathbf{R}$ is bound to $\mathbf{D}$ so, writing the density of edges in the data as ε_D, the internal energy and

hence the free energy is just

$$\begin{aligned} F(\tilde{h} \to \infty) &= -\frac{\tilde{K}}{N} \sum_{<ij>} D_i D_j - \tilde{h} \\ &= -\frac{\nu \tilde{K}}{2}(1 - 2\varepsilon_D) - \tilde{h}, \end{aligned}$$

3. For $\tilde{K} = 0$ and $\tilde{h} = 0$ the energy term is clearly zero and the number of arrangements is 2^N giving the free energy

$$F(\tilde{K} \to 0, \tilde{h} \to 0) = -\log 2.$$

Now we compare the behaviour of (3.40). We use the identity

$$\log \cosh(x) = |x| + \log(1 - e^{-2|x|}),$$

and we use the limiting values of R^+, R^- determined previously in §3.3.4.

1. In the limit $\tilde{K} \to \infty$, $R^+ = R^- = \pm 1$, which gives

$$F_{MF}(\tilde{K} \to \infty) = -\frac{\nu \tilde{K}}{2}.$$

2. In the limit $\tilde{h} \to \infty$, $R^+ = -R^- = 1 - 2q$, then

$$F_{MF}(\tilde{h} \to \infty) = -\tilde{h} - \frac{\nu \tilde{K}}{2}(1 - 2\varepsilon_S)(1 - 2q)^2.$$

3. When $\tilde{h} = \tilde{K} = 0$, all terms in (3.40) disappear except the constant, and so $F_{MF}(\tilde{K} = \tilde{h} = 0) = -\log 2$.

There is clearly agreement in the first and third cases, and a brief analysis of ε_D will show that the results agree in the second case also.

Relationship Between ε_S and ε_D

If the probability of an edge in the source is ε_S, then the probability of a corresponding edge in the data, given a probability q of flipping each pixel, is

$$\begin{aligned} \varepsilon_D &= \left[(1-q)^2+q^2\right]\varepsilon_S + 2(1-q)q(1-\varepsilon_S) \\ &= \varepsilon_S + 2q(1-q)(1-2\varepsilon_S). \end{aligned} \tag{3.41}$$

- When $q = 0.5$, $\varepsilon_D = 0.5$, irrespective of the value of ε_S, i.e. all information is lost.
- When $q = 0$ or 1, $\varepsilon_D = \varepsilon_S$ as expected since in these cases **S** maps deterministically onto **D**.
- For $\varepsilon_S < \frac{1}{2}$, $\varepsilon_D \geq \varepsilon_S$, i.e. the noise process always increases the density of edges in the picture.

From (3.41) we get

$$(1-2\varepsilon_D) = (1-2\varepsilon_S)\left[1-4q+4q^2\right], \tag{3.42}$$

which verifies the agreement of the second special case $\tilde{h} \rightarrow \infty$.

3.3.6 The Observables

We are now in a position to determine numerically all of the locally stable states of the mean field model for any choice of couplings $\tilde{K}, \tilde{h}$. Returning to equation (3.24) we can determine all of the quantities of interest in terms of H^+ and H^-. It is understood that $H^\pm$ are implicitly determined by the choice of couplings $\tilde{K}$ and $\tilde{h}$: the values we use are fixed point solutions of equations (3.37) and (3.38).

The overlap between source and reconstruction is

$$\begin{aligned}\langle \mathbf{S}.\mathbf{R}\rangle_D &= \frac{1}{N}\sum_k \left\langle\langle\langle R_k S_k\rangle_R\rangle_S\right\rangle_D \\ &\overset{\mathrm{MF}}{\to} \frac{1}{N}\sum_k \left\langle\langle R_k[\mathbf{S},\mathbf{D}]\rangle_V\, S_k^0\right\rangle_D . \qquad (3.43)\end{aligned}$$

Since the quantities are self-averaging we may replace the spatial sum with a quenched average over the data. Recalling that

$$\left\langle D_k[\mathbf{S}^0]\right\rangle_D = \sum_{\{\mathbf{D}\}} P(\mathbf{D}|\mathbf{S}^0)D_k = (1-2q)S_k^0, \qquad (3.44)$$

we get the average overlap of the source and reconstruction, from (3.24),

$$\begin{aligned}\mathbf{S}.\mathbf{R} &= \left\langle\langle R_k[\mathbf{S},\mathbf{D}]\rangle_V\, S_k^0\right\rangle_D \\ &= \frac{1}{2}(1-q)\tanh(H^+ + \tilde{h}) + \frac{1}{2}q\tanh(H^+ - \tilde{h}) \\ &\quad - \frac{1}{2}(1-q)\tanh(H^- - \tilde{h}) - \frac{1}{2}q\tanh(H^- + \tilde{h}). \qquad (3.45)\end{aligned}$$

The average overlap of the data and reconstruction is

$$\begin{aligned}\mathbf{D}.\mathbf{R} &= \langle\langle R_k[\mathbf{S},\mathbf{D}]\rangle_V\, D_k\rangle_D \\ &= \frac{1}{2}(1-q)\tanh(H^+ + \tilde{h}) - \frac{1}{2}q\tanh(H^- + \tilde{h}) \\ &\quad - \frac{1}{2}(1-q)\tanh(H^- - \tilde{h}) + \frac{1}{2}q\tanh(H^+ - \tilde{h}). \qquad (3.46)\end{aligned}$$

Under mean field

$$\begin{aligned}
&\left\langle \langle R_k[\mathbf{D}]\rangle_R^2 \right\rangle_D \rightarrow \left\langle \langle R_k[\mathbf{S},\mathbf{D}]\rangle_V^2 \right\rangle_D \\
&= \left\langle \frac{1}{4}(1+S_k)(1+D_k)\tanh^2(H^+ + \tilde{h}) + \frac{1}{4}(1+S_k)(1-D_k)\tanh^2(H^+ - \tilde{h}) \right. \\
&\quad \left. + \frac{1}{4}(1-S_k)(1+D_k)\tanh^2(H^- + \tilde{h}) + \frac{1}{4}(1-S_k)(1-D_k)\tanh^2(H^- - \tilde{h}) \right\rangle_D
\end{aligned}$$

and so the width of the restored distribution (2.27) is

$$\begin{aligned}
\langle W_R\rangle_D &= \frac{1}{2} - \frac{1}{4}(1-q)\tanh^2(H^+ + \tilde{h}) - \frac{1}{4}q\tanh^2(H^+ - \tilde{h}) \\
&\quad - \frac{1}{4}(1-q)\tanh^2(H^- - \tilde{h}) - \frac{1}{4}q\tanh^2(H^- + \tilde{h}). \quad (3.47)
\end{aligned}$$

These three results (3.45), (3.46) and (3.47) are what we require to calculate the quality factor (2.29).

In addition, the average bias in the reconstructions is

$$\begin{aligned}
\langle M_R\rangle_D &= \langle \langle R_k[\mathbf{S},\mathbf{D}]\rangle_V \rangle_D \\
&= \frac{1}{2}(1-q)\tanh(H^+ + \tilde{h}) + \frac{1}{2}q\tanh(H^+ - \tilde{h}) \\
&\quad + \frac{1}{2}(1-q)\tanh(H^- - \tilde{h}) + \frac{1}{2}q\tanh(H^- + \tilde{h}), \quad (3.48)
\end{aligned}$$

and the overlap of the thresholded posterior mean with the source is

$$\begin{aligned}
\mathbf{T}.\mathbf{S} &= \left\langle S_k^0 \mathrm{sgn}\,\langle R_k[\mathbf{S},\mathbf{D}]\rangle_V \right\rangle_D \\
&= \frac{1}{2}(1-q)\mathrm{sgn}(H^+ + \tilde{h}) + \frac{q}{2}\mathrm{sgn}(H^+ - \tilde{h}) \\
&\quad - \frac{1}{2}(1-q)\mathrm{sgn}(H^- - \tilde{h}) - \frac{q}{2}\mathrm{sgn}(H^- + \tilde{h}), \quad (3.49)
\end{aligned}$$

using the identity $\mathrm{sgn}[\tanh(x)] = \mathrm{sgn}(x)$, where

$$\mathrm{sgn}(x) = \begin{cases} +1 & \text{if } x \geq 0, \\ -1 & \text{if } x < 0. \end{cases}$$

3.3.7 Results

The first result we examine is the phase diagram of the mean field model shown in Figure 3.7. This is generated by first finding all of the stable solutions of the implicit equation (3.38) at each point in parameter space $(\tilde{K}, \tilde{h})$. Each stable solution corresponds to a minimum in the free energy surfaces shown in Figure 3.6. We then classify the point according to the number of stable solutions (or metastable states) and the character of the equilibrium solution [that which minimizes the free energy (3.40)]. The solution may be prior-like (non-zero bias and zero overlap with the source), or data-like (non-zero overlap with the source and data).

For small enough coupling $\tilde{K}$ we find the equilibrium solution is mostly aligned with the data and the overlap **R**.**S** is non-zero, while the bias remains zero. As we increase the value of $\tilde{K}$, two metastable prior-like states appear which have zero overlap with the source but a non-zero bias. Ultimately, for large $\tilde{K}$, further data-like metastable states appear. At some point, as $\tilde{K}$ increases, there is a phase transition where the ordered prior-like state becomes lower in energy than the data-like state: the prior wins.

This phase transition explains the failure of the restoration scheme in certain regions of parameter space. When the nearest-neighbour interaction $\tilde{K}$ is too strong, the collective behaviour overwhelms the field $\tilde{h}$ that binds

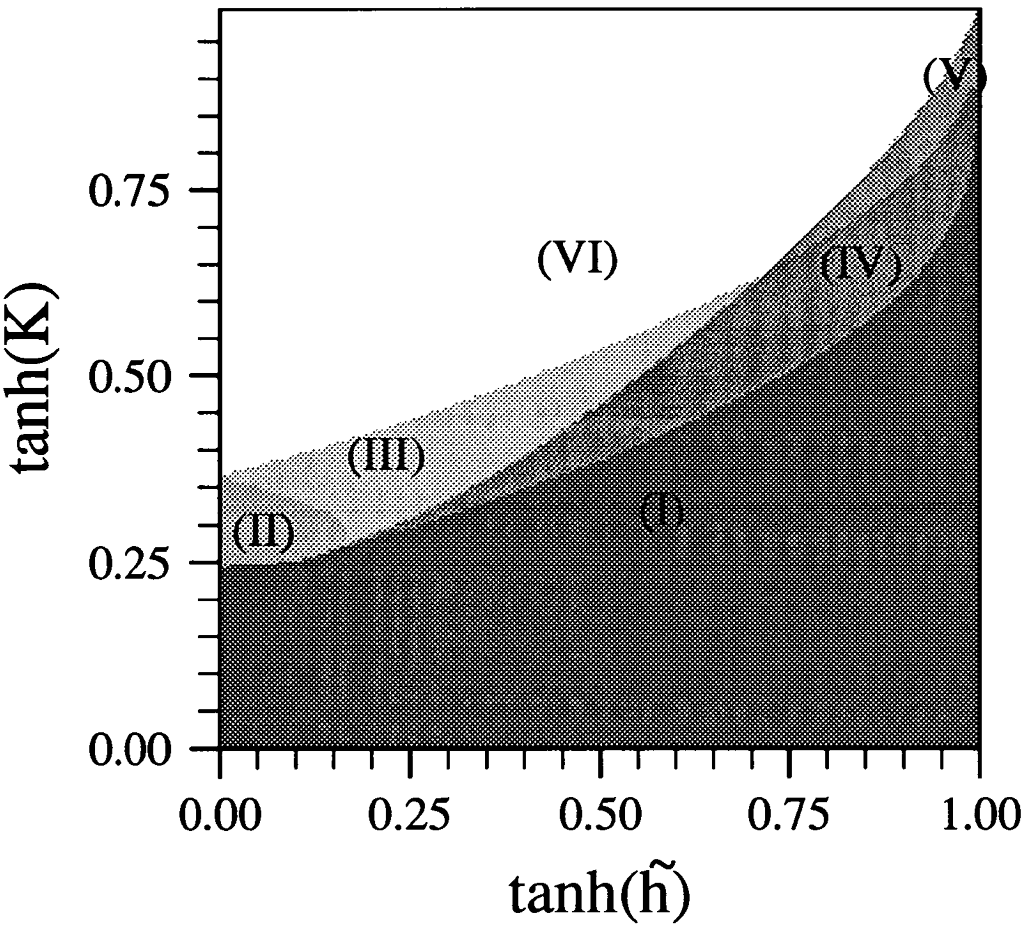


Figure 3.7. The phase diagram for the mean field model with density of edges $\varepsilon_S = 0.125$ and 30% noise $[q = 0.3]$. The regions of the phase diagram are distinguished by the number and character of metastable states as follows [refer to Figure 3.6]: (I), only one stable state with a non-zero overlap with the source; (II), two stable states with zero overlap but non-zero bias (prior-like states); (III), three metastable states, both data-like and prior-like states with the prior-like states being lower in free energy than the data-like state; (IV), three metastable states as in region (III) but with the data-like state lowest in free energy; (V), as region (IV) but there is now a further state, anti-aligned with the data; (VI), four metastable states as in region (V) but with the prior-like states lowest in free energy.
Therefore the restored pictures will have a non-zero overlap with the source only in regions (I), (IV) and (V). There is a phase transition line between these regions and the regions (II), (III) and (VI) where the restored pictures are overwhelmed by long range order and have zero overlap with the source. This notwithstanding it is only in region (II) where there are no data-like states whatsoever.

the restoration to the data, and unless one can access the metastable states rather than the global minimum of the free energy, this results in pictures of predominantly one colour, with no overlap with the source. There is a region of phase space (region (II) in Figure 3.7) where there are *no* data-like states and there is no possibility of finding suitable restorations. This region grows in size as the difficulty of the restoration problem grows (i.e. as the edge density ε_S or the noise level q increases).

The phase transition also explains why the quality factor and the overlaps presented in Figure 3.3 all show a marked decrease toward the top left of the diagrams (large $\tilde{K}$, small $\tilde{h}$).

Figure 3.8 compares the quality factor obtained from the mean field calculation with the simulation results for the comparable chequerboard. As we stated at the beginning of this section, the mean field calculation assumes a fixed source picture, characterized only by the density of edges in the image. Although we could compare this with a single image drawn from the Ising source distribution, such a source contains structure on many length scales, and is characterized by very many correlation functions. The plain chequerboard seems more typical of the simplest source picture modelled in the mean field calculation, and for the purposes of comparison we use this chequerboard source.

We see that there is excellent qualitative agreement between simulation and the mean field approximation.

- The phase transition is clearly defined in the mean field results, as we would expect since the calculation is carried out for an infinite system and we can exactly calculate the free energy and hence the

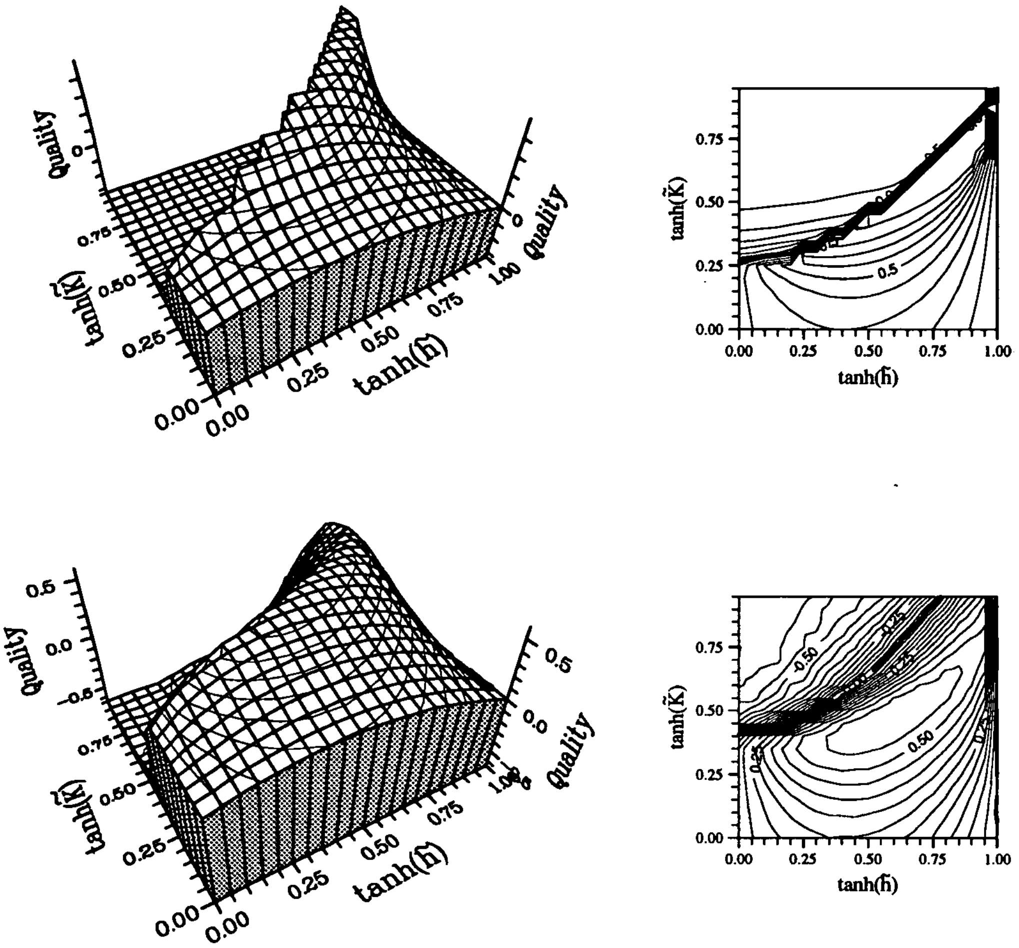


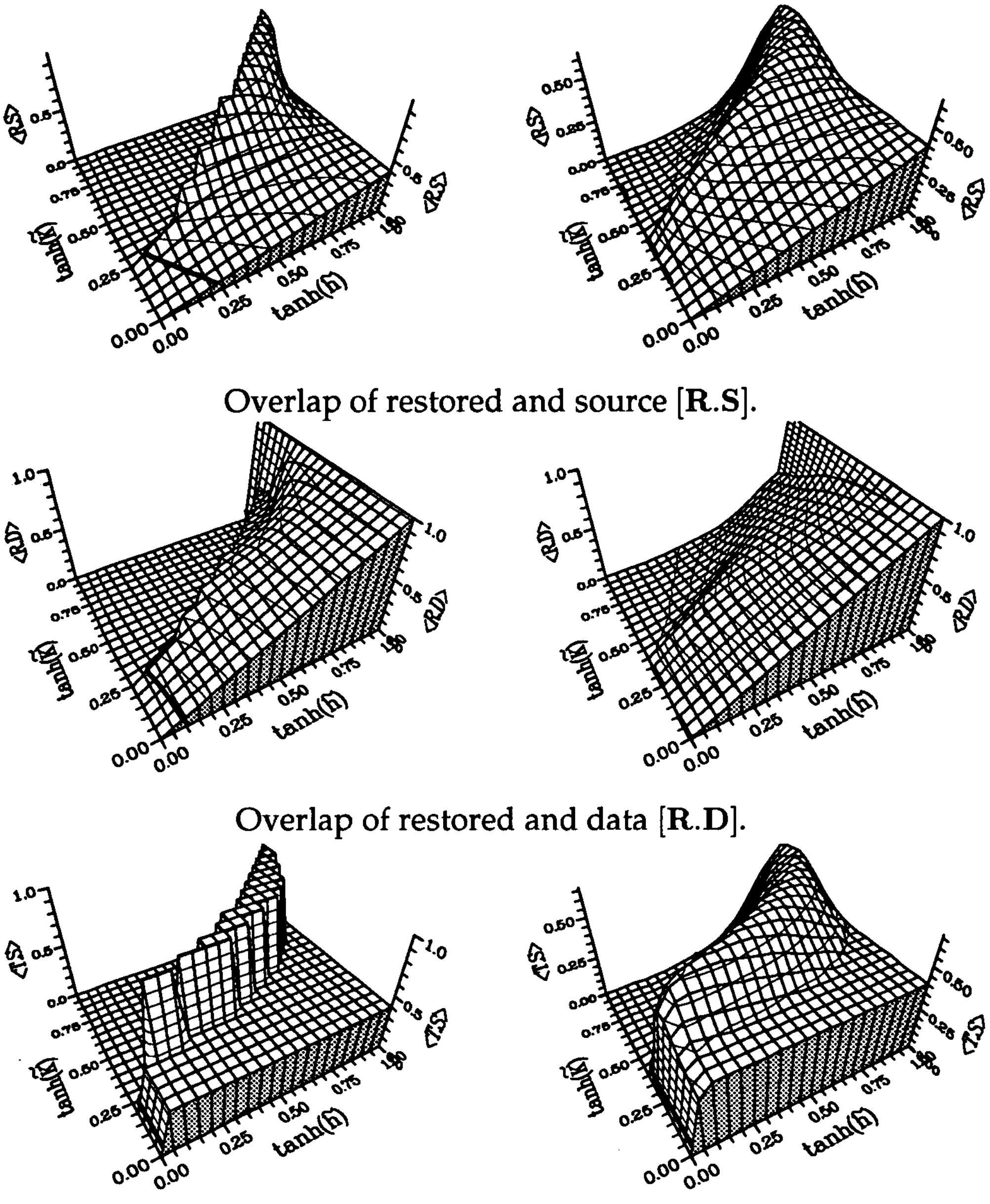
Figure 3.8. Comparison of mean field and simulation results for quality. The upper figures show the results for the mean field calculation with the density of edges $\varepsilon_S = 0.125$ and 30% noise $[q = 0.3, \tanh(h) = 0.4]$. The lower figures are the simulation results for the corresponding 8x8 chequerboard $[\varepsilon_S = 0.125]$ and 30% noise. There are quantitative differences in the position and sharpness of the phase transition but there is excellent qualitative agreement.

point at which a metastable state changes to become the global minimum of the system. The simulation does not exhibit significant finite size effects, but the transition is smeared out because the system gets caught in metastable states near the phase transition with the particular minimum found having a strong dependence on the starting configuration for the restoration.

- The position of the phase transition is also shifted. This is no surprise as we know that for the two-dimensional Ising model the phase transition occurs at $K_c = 0.44$, while the mean field calculation for the Ising model yields a critical coupling of $K_c = \frac{1}{4}$. Of course, for $\tilde{h} = 0$ the restoration scheme is simply an Ising model in zero field.

Figure 3.9 compares the results of the mean field calculation with simulation for three other quantities: **R**.**S** the average overlap of the restored pictures with the source; **R**.**D** the average overlap of the restored pictures with the data; and **T**.**S** the overlap of the TPM estimate with the source. The qualitative agreement these exhibit is quite remarkable given the approximations made in the calculation, and provides ample support for the earlier interpretation of the phase diagram.

Phase transitions are marked by a discontinuity in an observable. Although there is clear evidence of an ordered (prior-like) phase in the simulation results, the actual phase transition is softened and we see a continuous, if steep, change in the value of the quality factor. What happens is that when the ordered phase becomes the equilibrium state of the restoration model, data-aligned metastable states remain. Since we start the Monte Carlo process with the restored picture copied from the data, the system is trapped in one of these metastable states, and we still get useful restoration a little beyond the phase transition. However, as the nearest neighbour



Overlap of restored and source [**R**.**S**].

Overlap of restored and data [**R**.**D**].

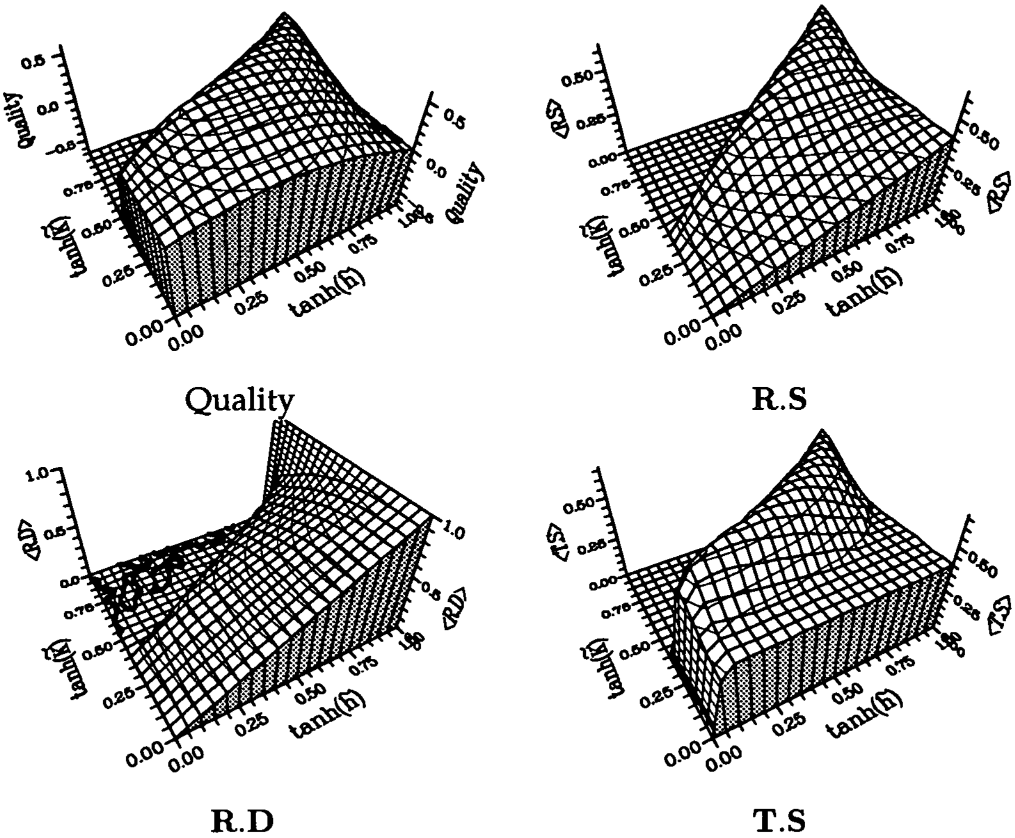
Overlap of TPM and source [**T**.**S**].

Figure 3.9. Comparison of mean field and simulation results. The left hand column shows results for the mean field calculation with the density of edges $\varepsilon_S = 0.125$ and 30% noise $[q = 0.3, \tanh(h) = 0.4]$. The figures on the right are the simulation results for the corresponding 8x8 chequerboard $[\varepsilon_S = 0.125]$ and 30% noise. As in Figure 3.8 there are quantitative differences in the position and sharpness of the phase transition but there is excellent qualitative agreement.

coupling is increased, the correlation length grows accordingly and large domains begin to be formed in spite of the underlying data. Imagine a chequerboard source. As the nearest neighbour coupling increases there comes a point where the surface to volume ratio of a single chequer is such that it is more favourable to remove the edges around the chequer than it is to retain the alignment of the chequer with the data. However, to reach this state requires the reversal of the whole cluster of pixels that make up the chequer. This is a metastable state since the chequer remains stable against single pixel flips. Therefore there is a small probability of a single chequer being lost in the restoration, but as the nearest neighbour coupling increases further this probability increases and the mean number of chequers inverted against the data increases. This leads to a finite rate of decrease in the overlap of the source and data as the coupling is increased. If, on the other hand, we begin the Monte Carlo process with the restored picture all one colour, i.e. close to the ordered phase, we observe a much sharper transition in the values of the observables as indicated in Figure 3.10.

Although starting from the edge-free state gives results that are much closer to mean field, and that are indeed a closer representation of the true equilibrium states, we continue to present results that were obtained by starting from the data-like state, since this is the more natural approach given the image restoration task.

Figure 3.11 provides further evidence of the success of the mean field calculation. It serves as a companion figure to Figure 3.5 which presented comparable results for simulation with chequerboard sources. The trends that were discovered by simulation are consistent with the mean field results:



Quality

R.S

R.D

T.S

Figure 3.10. Simulation results for 8x8 chequerboard and 30% noise, as in Figure 3.8 and Figure 3.9, but starting from the edge-free (one colour) state. Notice how much sharper the phase transition is in all cases, and the better qualitative agreement with the mean field results.

- The optimal choice of restoration coupling $\tilde{K}$ increases as the density of edges in the source decreases, *and* as the noise level decreases.
- The optimal choice of the restoration parameter $\tilde{h}$ increases as the noise level decreases, *and* as the density of edges in the source decreases.

Figure 3.11 also clearly shows how the position of the phase transition line depends upon the density of edges in the source and the noise level. As either the density of edges or the noise level increases the phase transition line cuts deeper into the phase diagram and the size of the ordered phase increases. In both cases the restoration problem is more difficult and this is reflected in the reduced volume of phase space that provides meaningful restoration.

In conclusion, we see excellent agreement between simulation and mean field, especially for the overlap **R**.**D**. For the overlap **R**.**S**, and the quality factor, the qualitative agreement is good, but in the mean field results the overlap approaches unity for large $\tilde{K}, \tilde{h}$ while it has a smaller upper bound in simulation. The most disconcerting mean field results are those for the overlap of the TPM estimate with the source, **T**.**S** (refer to the bottom row of Figure 3.9). There are three distinct regions. In the prior-like phase the overlap **T**.**S** is zero, while for low values of $\tilde{K}$ it is simply $1 - 2q$ and these both agree with the simulation results. However, in the region where we get best performance, the mean field results indicate *perfect* restoration for the TPM estimate. The qualitative results are not too bad since we do see a good correspondence between the region where in the mean field approximation $\mathbf{T}.\mathbf{S} = 1$ and the region of good performance (large **T**.**S** in the simulation).

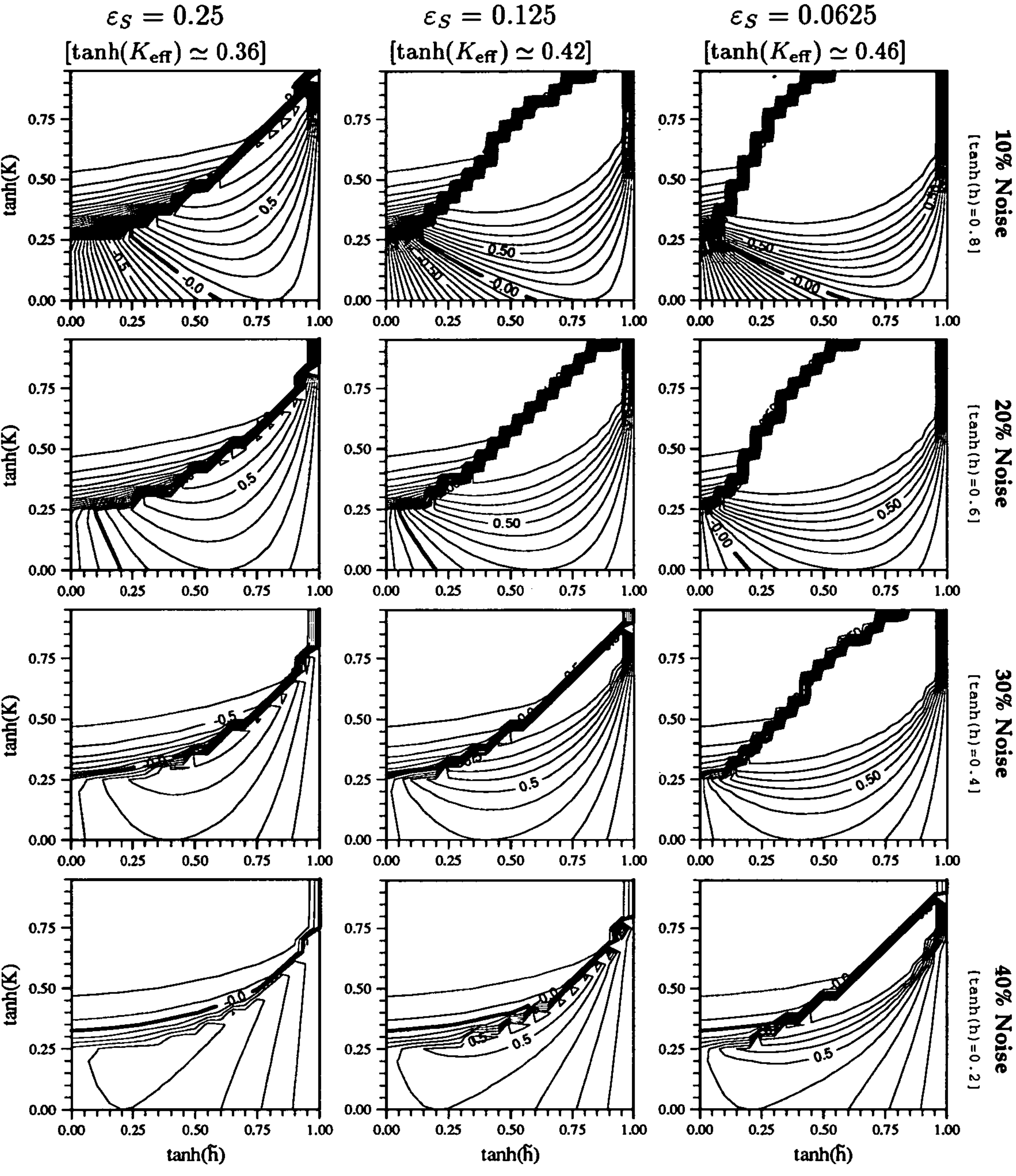


Figure 3.11.

Plot of the mean field quality factor for various values of ε_S and h. Compare with Figure 3.5. As the density of edges decreases the optimal choice of both $\tilde{K}$ and $\tilde{h}$ increases. As the noise level increases the optimal choice of both $\tilde{K}$ and $\tilde{h}$ decreases. The phase transition line cuts deeper into phase space the greater the density of edges and the higher the noise level.

The rather unlikely cases of perfect restoration indicated in the mean field results are related to the simplicity of the model. If we examine the equation for **T.S** (3.49) we see that the thresholded nature of this quantity has led to sgn functions. With only four terms on the right-hand side there is a very restricted set of values that the overlap could possibly take on, and only the three detailed above actually arise. A value of 1 indicates perfect restoration; $1 - 2q$ is the degree of overlap we begin with, [**S.D**]; while zero indicates a complete failure of the restoration scheme. These results lead us to experiment with a somewhat more complex model in the belief that this will more closely match the simulation results.

3.4 An Extension to the Mean Field Calculation

We claim that the behaviour of **T.S** arises from the restricted number of degrees of freedom that we allow the model—this corresponds to the way we have classified the sites into four types: $\{D_i = S_i = +1\}$, $\{D_i = -S_i = +1\}$, $\{D_i = S_i = -1\}$, $\{D_i = -S_i = -1\}$. The advantages are that it is a simple natural classification, and it allows us to derive the mean field equations with the source parameterized by merely the edge density, ε_S.

We may, however, consider more than four classes of site, provided we know more of the geometry of the source picture. [If we were to consider the classification on a site by site basis, with N classes, and hence N order parameters, then we would get very good results, but the calculation would be almost as intensive as computation of the partition function itself!]

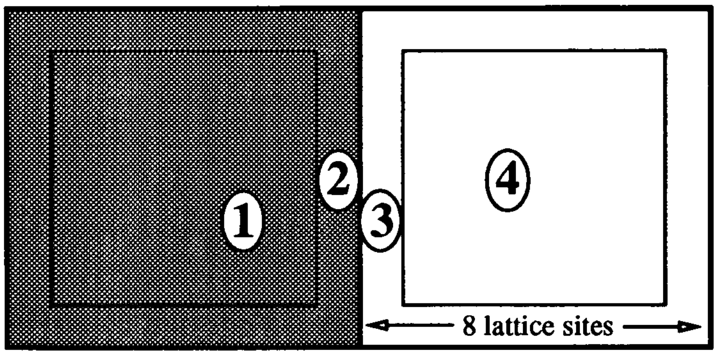


Figure 3.12. A schematic of the breakdown of the sites in an 8x8 chequerboard source into four regions. (1) and (4) are bulk sites: where all nearest neighbours are the same colour. (2) and (3) are edge sites where one or more nearest neighbours are a different colour and hence two neighbours may lie in a different class.

3.4.1 The Extended Calculation

We proceed with a more detailed classification in the case of the 8x8 chequerboard. We distinguish the sites of the source image as members of four classes, rather than as members of two [see Figure 3.12]:

1. sites where $S_i = +1$ and all nearest neighbours have $S_i = +1$, i.e. bulk sites with $S_i = 1$;

2. sites where $S_i = +1$ and one or more nearest neighbours have $S_i = -1$, i.e. edge sites with $S_i = +1$;

3. edge sites with $S_i = -1$;

4. bulk sites with $S_i = -1$.

This classification allows eight degrees of freedom in the mean field equations once we have taken account of the corruption process. We have four variational fields $H_{1...4}$ and four corresponding order parameters $R_{1...4}$.

We must first compute the size of each class, and the number and type of each nearest neighbour pair. The proportional area of each class is 1. $\frac{36}{128}$ 2. $\frac{28}{128}$ 3. $\frac{28}{128}$ 4. $\frac{36}{128}$. We can now write down the variational partition function for this particular geometry and calculate

$$\begin{aligned}\log Z_V &= \frac{36}{128}\Big[(1-q)\log\cosh(H_1+\tilde{h})+q\log\cosh(H_1-\tilde{h}) \\ &\quad +(1-q)\log\cosh(H_4-\tilde{h})+q\log\cosh(H_4+\tilde{h})\Big] \\ &\quad +\frac{28}{128}\Big[(1-q)\log\cosh(H_2+\tilde{h})+q\log\cosh(H_2-\tilde{h}) \\ &\quad +(1-q)\log\cosh(H_3-\tilde{h})+q\log\cosh(H_3+\tilde{h})\Big]. \qquad (3.50)\end{aligned}$$

Defining order parameters $R_{1\ldots4}$ as before we get

$$R_1 = (1-q)\tanh(H_1+\tilde{h})+q\tanh(H_1-\tilde{h}), \qquad (3.51)$$

$$R_2 = (1-q)\tanh(H_2+\tilde{h})+q\tanh(H_2-\tilde{h}), \qquad (3.52)$$

$$R_3 = (1-q)\tanh(H_3-\tilde{h})+q\tanh(H_3+\tilde{h}), \qquad (3.53)$$

$$R_4 = (1-q)\tanh(H_4-\tilde{h})+q\tanh(H_4+\tilde{h}). \qquad (3.54)$$

The number and type of each nearest neighbour pair is used to calculate

$$\begin{aligned}\langle A\rangle_V &= \frac{\nu\tilde{K}}{2}\Big[\frac{60}{256}\left({R_1}^2+{R_4}^2\right)+\frac{28}{256}\left({R_2}^2+{R_3}^2\right) \\ &\quad +\frac{24}{256}(R_1R_2+R_3R_4)+\frac{32}{256}R_2R_3\Big] \\ &\quad -\frac{36}{128}(R_1H_1+R_4H_4)-\frac{28}{128}(R_2H_2+R_3H_3). \qquad (3.55)\end{aligned}$$

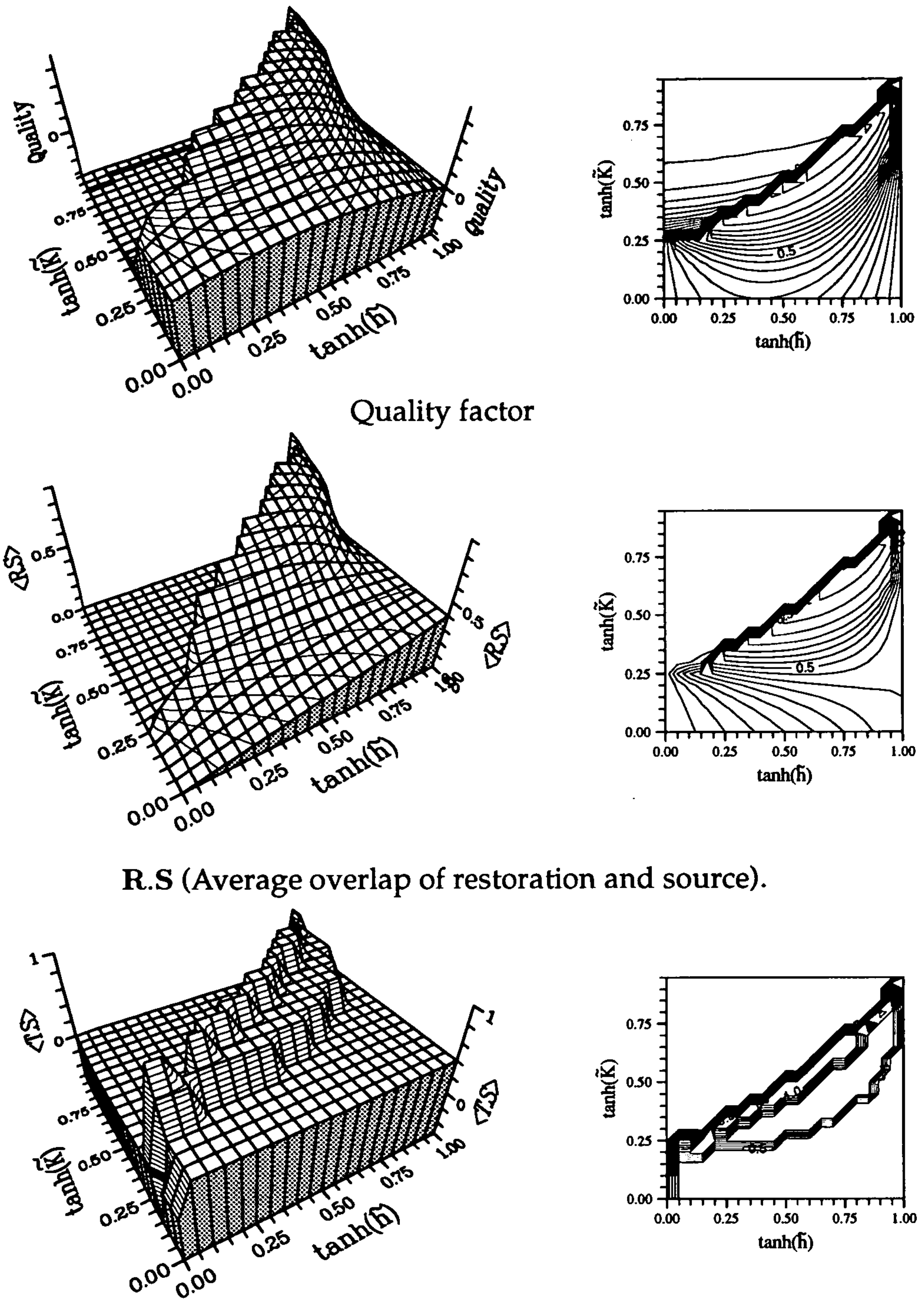
From (3.50) and (3.55) we can construct $\mathcal{F}(H_1,H_2,H_3,H_4)$ and then requiring $\nabla\mathcal{F}=0$ for a turning point in the variational free energy gives a set of four coupled equations

$$\begin{aligned} H_1 &= \frac{\nu\tilde{K}}{6}\left(5R_1 + R_2\right), && (3.56)\\ H_2 &= \frac{\nu\tilde{K}}{14}\left(7R_2 + 3R_1 + 4R_3\right), && (3.57)\\ H_3 &= \frac{\nu\tilde{K}}{14}\left(7R_3 + 3R_4 + 4R_2\right), && (3.58)\\ H_4 &= \frac{\nu\tilde{K}}{6}\left(5R_4 + R_3\right). && (3.59) \end{aligned}$$

with $R_{1\ldots4}$ given in terms of $H_{1\ldots4}$ by equations (3.51. . . 3.54). We now have a set of four coupled implicit equations in four unknowns. There is no straightforward simplification such as we used to find the single implicit equation (3.38) in the previous mean field case. In order to solve we apply a numerical algorithm from the NAG library [92, Subroutine C05NBF]. The computation requires considerable care in order to find all of the possible fixed points—for many sets of parameters and initial conditions the algorithm fails to converge. To check for minima requires calculation of second derivatives, and the Jacobian is a four by four determinant. The calculation is not detailed here.

3.4.2 Results

The results of this extended mean field calculation are presented in Figure 3.13 and should be compared with the results from the simpler mean field calculation shown in Figures 3.8 and 3.9. We see that there is very little qualitative change in the results, except for T.S which now exhibits a stepped effect. This more closely matches the simulation results. The stepped effect is due to the further degrees of freedom available in the calculation of T.S since we have four rather than two variational fields.



Quality factor

R.S (Average overlap of restoration and source).

T.S (Average overlap of TPM estimate and source).

Figure 3.13. Results for the extended mean field calculation for an 8x8 chequerboard and 30% noise. In the third figure we see that the extension to the mean field calculation has taken us one step towards the reality of the simulation results. T.S now takes on more than just a single value in the region of "good" restoration. In addition the maximal values of the quality factor and the overlap R.S are reduced from the perfect restoration indicated in the results from the simpler mean field calculation.

Examining **R.S** we see that the extra degrees of freedom allow a more accurate calculation and the values are reduced in the direction of the simulation results, with an upper bound on **R.S** of less than unity. The same effect is observed for the quality factor. We expect to see this improvement trend continue as we allow more and more degrees of freedom with the inclusion of further variational fields.

To summarize the mean field results, it is apparent that we get remarkably good qualitative results, even with just the simple model we first described, and we understand the reasons for the quantitative differences from the simulation results:

- It is well known that the position of the transition to the ordered phase is modified in the mean field approximation. Since mean field is equivalent to having an infinite interaction range (see e.g. [114]), we would expect a transition to long range order to occur at a smaller value of the coupling in the mean field approximation than a nearest neighbour interaction would suggest.
- The quantitative difference in the values of observables are caused, at least in part, by the simplification of the space of order parameters that we consider.

The utility of the mean field approximation in improving our understanding of the qualitative behaviour of the model far outweighs any reservations we may have about the quantitative results.

3.5 The Small Coupling Expansion

3.5.1 Introduction to Series Expansions

We now consider an alternative approach to mean field. This is the small coupling expansion, where we rewrite the exponential in the posterior probability distribution (2.43) as a series expansion. Then, provided the argument of the exponential is small—i.e. the coupling $\tilde{K}$ is small—we get an approximation to the true results by truncating the series and calculating only the first few terms. Of course such calculations are rarely simple, and the complexity of each higher order term is in general comparable to the total complexity of calculating all of the previous terms in the expansion. So there is invariably a trade-off between cost and accuracy, or the size of region in which the calculation gives meaningful results.

There is a wealth of literature on series expansions (see e.g. [25]). The most often quoted success is doubtless that of quantum electrodynamics. Feynman developed a graphical formalism which allows the calculations to be carried to very high orders using geometric arguments. In the work that follows we have not constructed a complete formalism, however it is still useful to use a diagrammatic notation at times to simplify the complex summations required.

Our aim is to calculate the quality factor (2.29), and to do this we need to calculate two quenched averages: $\langle\langle\langle \mathbf{R}.\mathbf{S}\rangle\rangle\rangle$ and $\langle|\mathbf{R}[\mathbf{D}]|^2\rangle$. We begin by considering source pictures drawn from an Ising distribution—for the small coupling calculation this is more straightforward than the fixed source case considered in the mean field approximation.

3.5.2 Calculation of R.S

Consider

$$\langle\langle\langle \mathbf{R}.\mathbf{S}\rangle\rangle\rangle = \sum_{\{\mathbf{S}\}}\sum_{\{\mathbf{R}\}} \mathbf{R}.\mathbf{S}P(\mathbf{S},\mathbf{R}), \tag{3.60}$$

where the joint probability distribution is

$$\begin{aligned} P(\mathbf{S},\mathbf{R}) &= \sum_{\{\mathbf{D}\}} P(\mathbf{R}|\mathbf{D})P(\mathbf{D}|\mathbf{S})P(\mathbf{S}) \\ &= \frac{1}{Z_p(K)Z_l(h)} \sum_{\{\mathbf{D}\}} \frac{1}{Z(\tilde{K},\tilde{h};\mathbf{D})} \exp\left\{ K \sum_{<ij>} S_i S_j \right. \\ &\quad \left. + \tilde{K} \sum_{<ij>} R_i R_j + h \sum_i S_i D_i + \tilde{h} \sum_i R_i D_i \right\}, \end{aligned} \tag{3.61}$$

and the normalizing partition functions are

$$Z(\tilde{K},\tilde{h};\mathbf{D}) = \sum_{\{\mathbf{R}\}} \exp\left\{ \tilde{K} \sum_{<ij>} R_i R_j + \tilde{h} \sum_i R_i D_i \right\}, \tag{3.62}$$

$$Z_l(h) = \sum_{\{\mathbf{D}\}} \exp\left\{ h \sum_i S_i D_i \right\} = [2\cosh(h)]^N, \tag{3.63}$$

$$Z_p(K) = \sum_{\{\mathbf{S}\}} \exp\left\{ K \sum_{<ij>} S_i S_j \right\}. \tag{3.64}$$

We define a normalized average

$$\langle \cdot \rangle_1 \stackrel{\text{def}}{=} \frac{1}{Z_l(h)Z_l(\tilde{h})2^N} \sum_{\{\mathbf{S}\}}\sum_{\{\mathbf{D}\}}\sum_{\{\mathbf{R}\}} (\cdot) \exp\left\{ \tilde{h} \sum_i R_i D_i + h \sum_i D_i S_i \right\}, \tag{3.65}$$

which allows us to write

$$\left\langle \langle\langle \cdot \rangle_S \rangle_D \right\rangle_R = \left\langle (\cdot) g_1(K,\tilde{K},\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R}) \right\rangle_1, \tag{3.66}$$

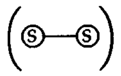
where we define

$$g_1(K, \tilde{K}, \tilde{h}; \mathbf{S}, \mathbf{D}, \mathbf{R}) = \frac{2^N Z_l(\tilde{h})}{Z_p(K) Z(\tilde{K}, \tilde{h}; \mathbf{D})} \exp\left\{ K \sum_{<ij>} S_i S_j + \tilde{K} \sum_{<ij>} R_i R_j \right\}. \tag{3.67}$$

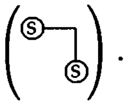
Written in this way, we have to expand $g_1(K, \tilde{K}, \tilde{h}; \mathbf{S}, \mathbf{D}, \mathbf{R})$ in powers of K and $\tilde{K}$, and then perform the average $\langle \cdot \rangle_1$.

In an effort to simplify the notation, we declare three functions of the configurations as follows:

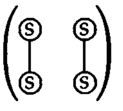
- The product of two nearest neighbour spins, summed over all nearest neighbour links on the lattice

$$A(\mathbf{S}) \stackrel{\text{def}}{=} \sum_{<ij>} S_i S_j \equiv$$

$$. \tag{3.68}$$

- The product of all second and third neighbour pairs—the sum over the lattice of all connected two link graphs

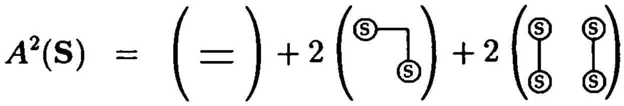
$$B(\mathbf{S}) \stackrel{\text{def}}{=} \sum_{<<ij>>} S_i S_j \equiv$$

$$. \tag{3.69}$$

- And a four-spin product over all disconnected two link graphs

$$C(\mathbf{S}) \stackrel{\text{def}}{=} \sum_{<ij><kl>} S_i S_j S_k S_l \equiv$$

$$. \tag{3.70}$$

In the graphical notation, a sum over all such graphs on the lattice is assumed.

Defined in this way we calculate the square of $A(\mathbf{S})$



$$= \frac{\nu N}{2} + 2B(\mathbf{S}) + 2C(\mathbf{S}),$$

since S_i are binary variables and $S_i^2 = 1\ \forall i$. This will greatly simplify the notation for an expansion to second order.

First we deal with the partition functions $Z(\tilde{K}, \tilde{h}; \mathbf{D})$ and $Z_p(K)$.

$$\begin{aligned} Z_p(K) &= \sum_{\{\mathbf{S}\}} \exp\{KA(\mathbf{S})\} \\ &= \sum_{\{\mathbf{S}\}} \left\{1 + KA(\mathbf{S}) + \frac{1}{2}K^2A^2(\mathbf{S}) + o(K^3)\right\} \\ &= 2^N \left\{1 + \frac{1}{2}K^2\frac{\nu N}{2} + o(K^3)\right\}, \end{aligned} \qquad (3.71)$$

where we have made use of the fact that $\sum_{\{\mathbf{S}\}}\sum_{<ij>} S_iS_j = 0$ as the S_i are binary variables and any sum over discrete sites will give zero when averaged over all configurations. This means that

$$\sum_{\{\mathbf{S}\}} A(\mathbf{S}) = \sum_{\{\mathbf{S}\}} B(\mathbf{S}) = \sum_{\{\mathbf{S}\}} C(\mathbf{S}) = 0. \qquad (3.72)$$

Next consider the main partition function

$$\begin{aligned} Z(\tilde{K}, \tilde{h}; \mathbf{D}) &= \sum_{\{\mathbf{R}\}} \exp\left\{\tilde{K} \sum_{<ij>} R_iR_j + \tilde{h}\sum_i R_iD_i\right\} \\ &= \sum_{\{\mathbf{R}\}} \left\{1 + \tilde{K}A(\mathbf{R}) + \frac{1}{2}\tilde{K}^2A^2(\mathbf{R}) + o(\tilde{K}^3)\right\} \exp\left\{\tilde{h}\sum_i R_iD_i\right\} \\ &= Z_l(\tilde{h}) \left\langle 1 + \tilde{K}A(\mathbf{R}) + \frac{1}{2}\tilde{K}^2A^2(\mathbf{R}) + o(\tilde{K}^3)\right\rangle_l, \end{aligned} \qquad (3.73)$$

where we define the normalized average

$$\langle . \rangle_l = \frac{1}{Z_l(\tilde{h})} \sum_{\{\mathbf{R}\}} (.) \exp\left\{\tilde{h} \sum_i R_i D_i\right\}. \tag{3.74}$$

In order to calculate the average in (3.73) we first find

$$\langle R_k \rangle_l = D_k \tanh(\tilde{h}), \tag{3.75}$$

and define

$$\tilde{\alpha} \stackrel{\text{def}}{=} \tanh(\tilde{h}). \tag{3.76}$$

Then, since $\exp\left\{\tilde{h} \sum_i D_i R_i\right\}$ in (3.74) is a factorized measure we easily find

$$\begin{aligned} \langle A(\mathbf{R}) \rangle_l &= \tilde{\alpha}^2 A(\mathbf{D}), && (3.77) \\ \langle B(\mathbf{R}) \rangle_l &= \tilde{\alpha}^2 B(\mathbf{D}), \\ \langle C(\mathbf{R}) \rangle_l &= \tilde{\alpha}^4 C(\mathbf{D}), \\ \left\langle A^2(\mathbf{R}) \right\rangle_l &= \frac{\nu N}{2} + 2\tilde{\alpha}^2 B(\mathbf{D}) + 2\tilde{\alpha}^4 C(\mathbf{D}). && (3.78) \end{aligned}$$

Substituting (3.77) and (3.78) into (3.73) gives

$$\begin{aligned} Z(\tilde{K}, \tilde{h}; \mathbf{D}) &= Z_l(\tilde{h}) \Big\{ 1 + \tilde{K}\tilde{\alpha}^2 A(\mathbf{D}) \\ &\quad + \frac{1}{2}\tilde{K}^2 \left[\frac{\nu N}{2} + 2\tilde{\alpha}^2 B(\mathbf{D}) + 2\tilde{\alpha}^4 C(\mathbf{D})\right] + o(\tilde{K}^3) \Big\}. \end{aligned} \tag{3.79}$$

The reciprocal of (3.79) is then

$$\begin{aligned} \frac{1}{Z(\tilde{K}, \tilde{h}; \mathbf{D})} &= \frac{1}{Z_l(\tilde{h})} \Big\{ 1 - \tilde{K}\tilde{\alpha}^2 A(\mathbf{D}) - \frac{1}{2}\tilde{K}^2 \Big[(1 - 2\tilde{\alpha}^4)\frac{\nu N}{2} \\ &\quad + 2\tilde{\alpha}^2(1 - 2\tilde{\alpha}^2) B(\mathbf{D}) - 2\tilde{\alpha}^4 C(\mathbf{D}) \Big] + o(\tilde{K}^3) \Big\}. \end{aligned} \tag{3.80}$$

Finally, expanding the exponential in (3.67) and using the small coupling expansions (3.80) and (3.71) gives

$$\begin{aligned} & g_1(K,\tilde{K},\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R}) = \\ & \quad 1+\tilde{K}A(\mathbf{R})+KA(\mathbf{S})-\tilde{K}\tilde{\alpha}^2A(\mathbf{D}) \\ & \quad +K\tilde{K}A(\mathbf{S})A(\mathbf{R})-\tilde{K}^2\tilde{\alpha}^2A(\mathbf{D})A(\mathbf{R})-K\tilde{K}\tilde{\alpha}^2A(\mathbf{D})A(\mathbf{S}) \\ & \quad -\frac{1}{2}\tilde{K}^2\left[(1-2\tilde{\alpha}^4)\frac{\nu N}{2}+2\tilde{\alpha}^2(1-2\tilde{\alpha}^2)B(\mathbf{D})-2\tilde{\alpha}^4C(\mathbf{D})\right] \\ & \quad +\frac{1}{2}\tilde{K}^2A^2(\mathbf{R})+\frac{1}{2}K^2A^2(\mathbf{S})-\frac{1}{2}K^2\frac{\nu N}{2}+o(K+\tilde{K})^3. \end{aligned} \tag{3.81}$$

We can check this result by confirming that $\left\langle g_1(K,\tilde{K},\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R})\right\rangle_1 = 1$.

It is not difficult to perform the average, since

$$\langle R_k\rangle_1 = \langle D_k\rangle_1 = \langle S_k\rangle_1 = 0, \tag{3.82}$$

and, once again $\langle\cdot\rangle_1$ is an average over a factorized measure so the average of any product where any site occurs only once will be zero:

$$\begin{aligned} \langle A(\mathbf{S})\rangle_1 &= \langle A(\mathbf{D})\rangle_1 = \langle A(\mathbf{R})\rangle_1 = 0 \\ \langle B(\mathbf{S})\rangle_1 &= \langle B(\mathbf{D})\rangle_1 = \langle B(\mathbf{R})\rangle_1 = 0 \\ \langle C(\mathbf{S})\rangle_1 &= \langle C(\mathbf{D})\rangle_1 = \langle C(\mathbf{R})\rangle_1 = 0 \end{aligned}$$

We have still to calculate averages such as $\langle A(\mathbf{S})A(\mathbf{D})\rangle_1$. The necessary site averages are

$$\begin{aligned} \langle S_kR_k\rangle &= \alpha\tilde{\alpha}, \\ \langle S_kD_k\rangle &= \alpha, \\ \langle D_kR_k\rangle &= \tilde{\alpha}, \end{aligned}$$

which produce the results

$$\begin{aligned}\langle A(\mathbf{S})A(\mathbf{R})\rangle_1 &= \frac{\nu N}{2}\alpha^2\tilde{\alpha}^2,\\ \langle A(\mathbf{S})A(\mathbf{D})\rangle_1 &= \frac{\nu N}{2}\alpha^2,\\ \langle A(\mathbf{R})A(\mathbf{D})\rangle_1 &= \frac{\nu N}{2}\tilde{\alpha}^2.\end{aligned}$$

while

$$\left\langle A^2(\mathbf{R})\right\rangle_1 = \left\langle A^2(\mathbf{D})\right\rangle_1 = \left\langle A^2(\mathbf{S})\right\rangle_1 = \frac{\nu N}{2}. \tag{3.83}$$

Finally, if we substitute these results back into (3.81) we confirm that

$$\left\langle g_1(K,\tilde{K},\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R})\right\rangle_1 = 1.$$

We can simplify the calculation further by rewriting the general equation (3.66) as

$$\left\langle \langle\langle\cdot\rangle_S\rangle_D\right\rangle_R = \langle\cdot\rangle_1 + \left\langle [(\cdot)-\langle\cdot\rangle_1]\, g_1(K,\tilde{K}\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R})\right\rangle_1. \tag{3.84}$$

This shows that any constants in $g_1(K,\tilde{K},\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R})$ will not contribute to the average. Furthermore since $\langle\cdot\rangle_1$ is an average over a factorized measure, (3.84) means that contributions arise only from the difference between two graphs in contact and the same two graphs disconnected.

The first quantity we want to calculate is the average overlap of the source pictures with the reconstructions

$$\begin{aligned}\left\langle \langle\langle \mathbf{S}.\mathbf{R}\rangle_S\rangle_R\right\rangle_D &= \left\langle \frac{1}{N}\sum_k S_k R_k g_1(K,\tilde{K},\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R})\right\rangle_1\\ &= \alpha\tilde{\alpha} + \left\langle \left[\frac{1}{N}\sum_k S_k R_k - \alpha\tilde{\alpha}\right] g_1(K,\tilde{K},\tilde{h};\mathbf{S},\mathbf{D},\mathbf{R})\right\rangle_1.\end{aligned} \tag{3.85}$$

As before, any unpaired site variable will give zero when the average is performed, so for example $\langle S_k R_k A(\mathbf{S})\rangle_1 = 0$. The only non-trivial calculation required is of second order terms such as $\langle S_k R_k A(\mathbf{S})A(\mathbf{R})\rangle_1$, where there is a contribution of $\alpha\tilde{\alpha}$ from the ν connected graphs containing site k, and a contribution of $\alpha^3\tilde{\alpha}^3$ from the $\nu(N-2)/2$ disconnected graphs. Using the graphical notation:

$$\begin{aligned}
\left\langle\left[\frac{1}{N}\sum_k S_k R_k - \alpha\tilde{\alpha}\right] A(\mathbf{S})A(\mathbf{R})\right\rangle_1 &= \frac{1}{N}\left\langle\begin{matrix}\textcircled{S}\text{---}\\ \textcircled{R}\text{---}\end{matrix}\right\rangle_1 - \frac{2}{N}\left\langle\begin{matrix}\textcircled{S}\text{---}\textcircled{S}\\ \textcircled{R}\text{---}\textcircled{R}\end{matrix}\right\rangle_1\left\langle\begin{matrix}\textcircled{S}\\ \textcircled{R}\end{matrix}\right\rangle_1\\
&= \nu\left[\alpha\tilde{\alpha} - \alpha^3\tilde{\alpha}^3\right],\\
\left\langle\left[\frac{1}{N}\sum_k S_k R_k - \alpha\tilde{\alpha}\right] A(\mathbf{R})A(\mathbf{D})\right\rangle_1 &= \frac{1}{N}\left\langle\begin{matrix}\textcircled{R}\text{---}\textcircled{S}\\ \textcircled{D}\text{---}\textcircled{D}\end{matrix}\right\rangle_1 - \frac{2}{N}\left\langle\begin{matrix}\textcircled{R}\text{---}\textcircled{R}\\ \textcircled{D}\text{---}\textcircled{D}\end{matrix}\right\rangle_1\left\langle\begin{matrix}\textcircled{S}\\ \textcircled{R}\end{matrix}\right\rangle_1\\
&= \nu\left[\alpha\tilde{\alpha} - \alpha\tilde{\alpha}^3\right],\\
\left\langle\left[\frac{1}{N}\sum_k S_k R_k - \alpha\tilde{\alpha}\right] A(\mathbf{S})A(\mathbf{D})\right\rangle_1 &= \frac{1}{N}\left\langle\begin{matrix}\textcircled{S}\text{---}\textcircled{R}\\ \textcircled{D}\text{---}\textcircled{D}\end{matrix}\right\rangle_1 - \frac{2}{N}\left\langle\begin{matrix}\textcircled{S}\text{---}\textcircled{S}\\ \textcircled{D}\text{---}\textcircled{D}\end{matrix}\right\rangle_1\left\langle\begin{matrix}\textcircled{R}\\ \textcircled{S}\end{matrix}\right\rangle_1\\
&= \nu\left[\alpha\tilde{\alpha} - \alpha^3\tilde{\alpha}\right].
\end{aligned}$$

Substituting these last three results back into (3.85) gives

$$\langle\langle\langle \mathbf{S}.\mathbf{R}\rangle\rangle\rangle = \alpha\tilde{\alpha}\left\{1 + \nu\tilde{K}(1-\tilde{\alpha}^2)(K - \tilde{K}\tilde{\alpha}^2) + o(K+\tilde{K})^3\right\}. \qquad (3.86)$$

3.5.3 Calculation of the Width

The other quantity we require to calculate the quality factor is $\left\langle\langle R_k[\mathbf{D}]\rangle_R^2\right\rangle_D$. Unfortunately, the averages required are subtly different from the overlap we have just calculated and we need to define a new average $\langle\cdot\rangle_2$ and function g_2.

$$\left\langle\langle R_k[\mathbf{D}]\rangle_R^2\right\rangle_D = \sum_{\{\mathbf{R}\}}\sum_{\{\mathbf{R}'\}} R_k R'_k P(\mathbf{R},\mathbf{R}'), \qquad (3.87)$$

where the joint probability

$$P(\mathbf{R},\mathbf{R}') = \sum_{\{\mathbf{D}\}}\sum_{\{\mathbf{S}\}} P(\mathbf{R}|\mathbf{D})P(\mathbf{R}'|\mathbf{D})P(\mathbf{D}|\mathbf{S})P(\mathbf{S}). \tag{3.88}$$

Therefore, if we define

$$\langle\cdot\rangle_2 \overset{\text{def}}{=} \frac{1}{2^N Z_l^2(\tilde{h})}\sum_{\{\mathbf{R}\}}\sum_{\{\mathbf{R}'\}}\sum_{\{\mathbf{D}\}}(\cdot)\exp\left\{\tilde{h}\sum_i (R_i + R_i')D_i\right\}, \tag{3.89}$$

we can write

$$\left\langle \langle R_k[\mathbf{D}]\rangle_R^2 \right\rangle_D = \left\langle R_k R_k' g_2(K,\tilde{K},\tilde{h};\mathbf{D},\mathbf{R},\mathbf{R}')\right\rangle_2, \tag{3.90}$$

where, using the equations for the probability distributions (3.62), (3.63), and (3.64), the function

$$\begin{aligned} g_2(K,\tilde{K},\tilde{h};\mathbf{D},\mathbf{R},\mathbf{R}') &= \frac{2^N Z(K,h;\mathbf{D})Z_l^2(\tilde{h})}{Z_p(K)Z_l(h)Z^2(\tilde{K},\tilde{h};\mathbf{D})}\exp\left\{\tilde{K}A(\mathbf{R}) + \tilde{K}A(\mathbf{R}')\right\} \\ &= 1 + (K\alpha^2 - 2\tilde{K}\tilde{\alpha}^2)A(\mathbf{D}) + \tilde{K}A(\mathbf{R}) + \tilde{K}A(\mathbf{R}') \\ &\quad + \tilde{K}^2\tilde{\alpha}^4 A^2(\mathbf{D}) - 2K\tilde{K}\tilde{\alpha}^2\alpha^2 A^2(\mathbf{D}) \\ &\quad + 2(K\tilde{K}\alpha^2 - 2\tilde{K}^2\tilde{\alpha}^2)A(\mathbf{R})A(\mathbf{D}) \\ &\quad - \tilde{K}^2\left\{2\tilde{\alpha}^2(1-2\tilde{\alpha}^2)B(\mathbf{D}) - 2\tilde{\alpha}^4 C(\mathbf{D})\right\} \\ &\quad + \frac{1}{2}K^2\left\{2\alpha^2 B(\mathbf{D}) + 2\alpha^4 C(\mathbf{D})\right\} \\ &\quad + \frac{1}{2}\tilde{K}^2\left[A(\mathbf{R}) + A(\mathbf{R}')\right]^2 + o(K+\tilde{K})^3 \end{aligned} \tag{3.91}$$

and we have used the series expansions (3.71), (3.79), and (3.80), replacing $\tilde{K}$ by K where appropriate, and we have neglected to write down any constant terms.

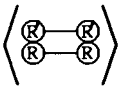
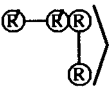
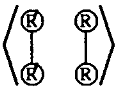
Most terms average to zero as before:

$$\begin{aligned}
\langle R_k\rangle_2 = \langle D_k\rangle_2 = \langle R'_k\rangle_2 &= 0,\\
\langle A(\mathbf{R}')\rangle_2 = \langle A(\mathbf{D})\rangle_2 = \langle A(\mathbf{R})\rangle_2 &= 0,\\
\langle B(\mathbf{R}')\rangle_2 = \langle B(\mathbf{D})\rangle_2 = \langle B(\mathbf{R})\rangle_2 &= 0,\\
\langle C(\mathbf{R}')\rangle_2 = \langle C(\mathbf{D})\rangle_2 = \langle C(\mathbf{R})\rangle_2 &= 0.
\end{aligned}$$

The other site averages we require are

$$\begin{aligned}
\langle R_k R'_k\rangle_2 &= \tilde{\alpha}^2,\\
\langle R_k D_k\rangle_2 &= \tilde{\alpha},
\end{aligned}$$

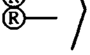
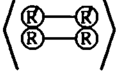
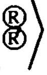
which give

$\langle A(\mathbf{R})A(\mathbf{R}')\rangle_2 = \Big\langle$  $\Big\rangle_2 + \Big\langle$  $\Big\rangle_2 + \Big\langle$  $\Big\rangle_2$

$$\begin{aligned}
&= \frac{\nu N}{2}\tilde{\alpha}^4,\\
\langle A(\mathbf{R})A(\mathbf{D})\rangle_2 &= \frac{\nu N}{2}\tilde{\alpha}^2,\\
\left\langle A^2(\mathbf{R})\right\rangle_2 = \left\langle A^2(\mathbf{R}')\right\rangle_2 &= \frac{\nu N}{2}.
\end{aligned}$$

These results confirm $\left\langle g_2(K,\tilde{K},\tilde{h};\mathbf{R},\mathbf{R}',\mathbf{D})\right\rangle_2 = 1$.

The only terms that figure in the expansion of (3.90) are:

$\left\langle\left[\frac{1}{N}\sum_k R_k R'_k - \tilde{\alpha}^2\right] A(\mathbf{R})A(\mathbf{R}')\right\rangle_2 = \frac{1}{N}\Big\langle$  $\Big\rangle_2 - \frac{2}{N}\Big\langle$  $\Big\rangle_2 \Big\langle$  $\Big\rangle_2$

$$= \nu\left[\tilde{\alpha}^2 - \tilde{\alpha}^6\right],$$

$$\begin{aligned}
\left\langle \left[\frac{1}{N}\sum_k R_k R'_k - \tilde{\alpha}^2\right] A(\mathbf{R})A(\mathbf{D}) \right\rangle_2 &= \frac{1}{N}\left\langle \begin{smallmatrix}\textcircled{R}\text{---}\textcircled{R}\\ \textcircled{D}\text{---}\textcircled{D}\end{smallmatrix} \right\rangle_2 - \frac{2}{N}\left\langle \begin{smallmatrix}\textcircled{R}\text{---}\textcircled{R}\\ \textcircled{D}\text{---}\textcircled{D}\end{smallmatrix} \right\rangle_2 \left\langle \begin{smallmatrix}\textcircled{R}\\ \textcircled{R}\end{smallmatrix} \right\rangle_2 \\
&= \nu\left[\tilde{\alpha}^2 - \tilde{\alpha}^4\right], \\
\left\langle \left[\frac{1}{N}\sum_k R_k R'_k - \tilde{\alpha}^2\right] A(\mathbf{R}')A(\mathbf{D}) \right\rangle_2 &= \frac{1}{N}\left\langle \begin{smallmatrix}\textcircled{R}\text{---}\textcircled{R}\\ \textcircled{D}\text{---}\textcircled{D}\end{smallmatrix} \right\rangle_2 - \frac{2}{N}\left\langle \begin{smallmatrix}\textcircled{R}\text{---}\textcircled{R}\\ \textcircled{D}\text{---}\textcircled{D}\end{smallmatrix} \right\rangle_2 \left\langle \begin{smallmatrix}\textcircled{R}\\ \textcircled{R}\end{smallmatrix} \right\rangle_2 \\
&= \nu\left[\tilde{\alpha}^2 - \tilde{\alpha}^4\right].
\end{aligned}$$

Finally, we obtain

$$\left\langle \langle R_k[D]\rangle_R^2 \right\rangle_D = \tilde{\alpha}^2\left\{1 + \nu\tilde{K}(1-\tilde{\alpha}^2)[2K\alpha^2 + \tilde{K}(1-3\tilde{\alpha}^2)] + o(K+\tilde{K})^3\right\}. \tag{3.92}$$

Similarly, we can calculate

$$\left\langle \langle S_k[D]\rangle_S^2 \right\rangle_D = \alpha^2\left\{1 + K^2\nu(1-2\alpha^2+\alpha^4) + o(K^3)\right\}. \tag{3.93}$$

3.5.4 Results

We now have all the results we need to write down the small coupling expansion of the quality factor (2.29).

$$\begin{aligned}
Q \;=\; & \Big[2\alpha\tilde{\alpha}\left\{1 + \nu\tilde{K}[K - \tilde{K}\tilde{\alpha}^2](1-\tilde{\alpha}^2)\right\} \\
& -\tilde{\alpha}^2\left\{1 + \nu\tilde{K}[2K\alpha^2 + \tilde{K}(1-3\tilde{\alpha}^2)](1-\tilde{\alpha}^2)\right\} + 1 - 2\alpha + o(K+\tilde{K})^3\Big] \\
& /\left[1 - 2\alpha + \alpha^2\left\{1 + K^2\nu(1-2\alpha^2)^2 + o(K^3)\right\}\right]
\end{aligned} \tag{3.94}$$

using the results (3.86) and substituting (3.92) and (3.93) into (2.27) and (2.28).

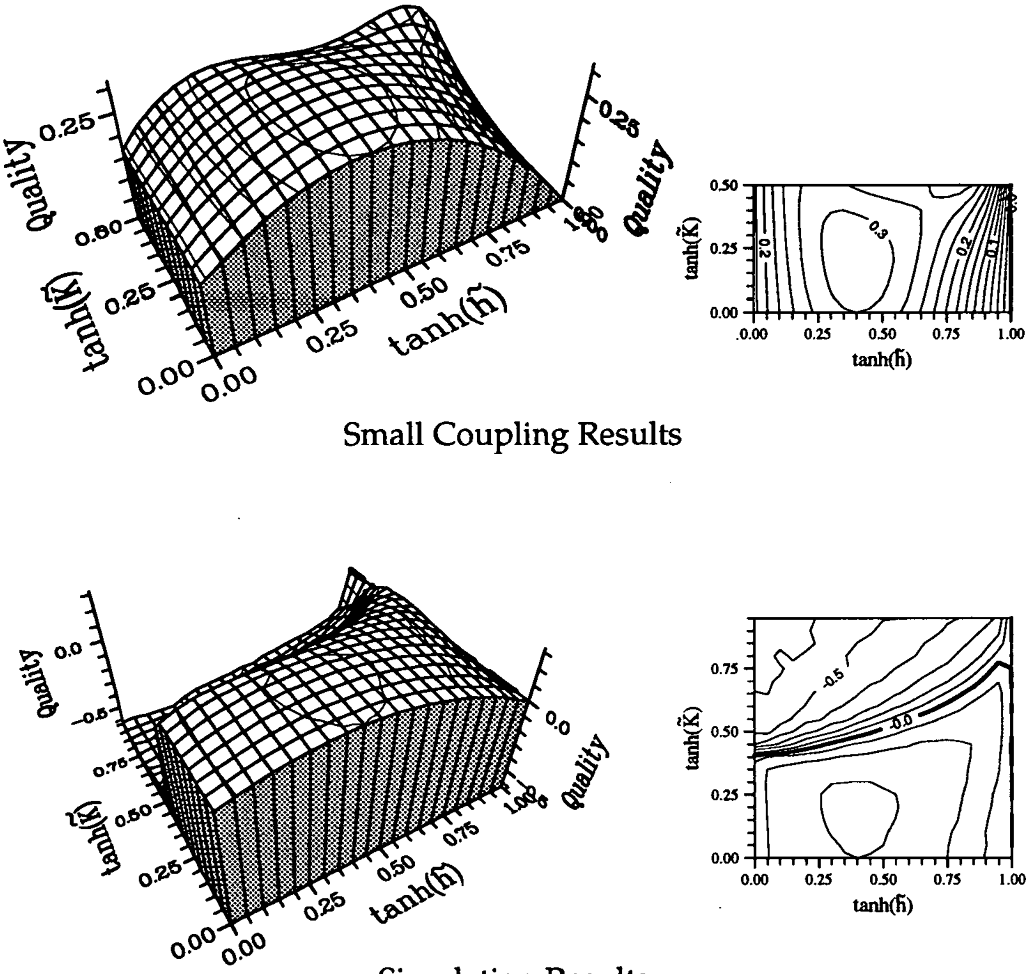
The denominator is a normalization which depends only on K and h. We can check certain properties of the quality factor with reference only to the numerator.

- If we bind the restored picture to the data with an infinite field $\tilde{h}$, then $\tilde{\alpha} = 1$ and $Q = 0$. We get no improvement.
- We can find the maximum value of Q as a function of $\tilde{K}$ and $\tilde{h}$. Differentiating (3.94) with respect to $\tilde{K}$ and $\tilde{h}$ we find a zero at $\tilde{K} = K$ *and* $\tilde{h} = h$. This verifies the Bayesian claim that the optimal restoration parameters will be those values used to generate the data.

Figure 3.14 compares the results of the small coupling expansion with simulation for an Ising source. We have chosen a small value of the near neighbour coupling K in the source picture since the expansion will only be accurate for small K and $\tilde{K}$. For this reason also, the results are presented for $\tanh(\tilde{K})$ only in the range $[0, 0.5]$. With these provisions we see excellent agreement with the simulation results and, as predicted by equation (3.94), the maximum of the quality factor occurs when $\tilde{K} = K$ and $\tilde{h} = h$.

3.6 Small Coupling Expansion for Fixed Source

We now attempt a small coupling expansion for the somewhat more complicated case of a fixed source image $\mathbf{S}^0$. This case is more difficult because many of the terms that averaged to zero in the previous calculation now contribute to the result.



Small Coupling Results

Simulation Results

Figure 3.14. Small coupling expansion results for an Ising source. The coupling used to generate the source picture was $\tanh(K) = 0.2$, with 30% noise $[\tanh(h) = 0.4]$. The upper figures show the results for the quality factor calculated by the small coupling expansion, while the lower figure show the corresponding results from a simulation of the same system. There is excellent agreement between the two, and as with the results in Figure 3.3(a) the maximum of the quality factor occurs when $\tilde{K} = K$ and $\tilde{h} = h$.

3.6.1 Calculation of the Overlaps

We retain the definitions of the lattice sums A, B, and C as given in equations (3.68),(3.69) and (3.70). And we want to calculate

$$
\begin{aligned}
\left\langle\left\langle \mathbf{R}.\mathbf{S}^0\right\rangle_R\right\rangle_D &= \sum_{\{\mathbf{R}\}} \mathbf{R}.\mathbf{S}^0 \sum_{\{\mathbf{D}\}} P(\mathbf{R}|\mathbf{D})P(\mathbf{D}|\mathbf{S}^0) \\
&= \sum_{\{\mathbf{R}\}} \mathbf{R}.\mathbf{S}^0 \sum_{\{\mathbf{D}\}} \frac{1}{Z_l(h)Z(\tilde{K},\tilde{h};\mathbf{D})} \exp\left\{ \tilde{K} \sum_{<ij>} R_i R_j \right. \\
&\quad \left. + \tilde{h} \sum_i R_i D_i + h \sum_i S_i^0 D_i \right\} \\
&= \left\langle \mathbf{R}.\mathbf{S}^0 g_3(\tilde{K},\tilde{h};\mathbf{D},\mathbf{R}) \right\rangle_3,
\end{aligned}
$$

where the average

$$
\langle\cdot\rangle_3 \stackrel{\text{def}}{=} \frac{1}{Z_l(h)Z_l(\tilde{h})} \sum_{\{\mathbf{R}\}} \sum_{\{\mathbf{D}\}} (\cdot) \exp\left\{ \tilde{h} \sum_i R_i D_i + h \sum_i S_i^0 D_i \right\}, \qquad (3.95)
$$

and the function

$$
\begin{aligned}
g_3(\tilde{K},\tilde{h};\mathbf{D},\mathbf{R}) &\stackrel{\text{def}}{=} \frac{Z_l(\tilde{h})}{Z(\tilde{K},\tilde{h};\mathbf{D})} \exp\left\{ \tilde{K} A(\mathbf{R}) \right\} \\
&= 1 + \tilde{K} A(\mathbf{R}) - \tilde{K}\tilde{\alpha}^2 A(\mathbf{D}) + \frac{1}{2}\tilde{K}^2 A^2(\mathbf{R}) \\
&\quad - \frac{1}{2}\tilde{K}^2 \left[2\tilde{\alpha}^2(1-2\tilde{\alpha}^2)B(\mathbf{D}) - 2\tilde{\alpha}^4 C(\mathbf{D}) \right] \\
&\quad - \tilde{K}^2\tilde{\alpha}^2 A(\mathbf{R})A(\mathbf{D}) + o(\tilde{K}^3), \qquad (3.96)
\end{aligned}
$$

and we have again neglected to write down any constant terms. The averages we require are:

$$
\begin{aligned}
\langle R_k \rangle_3 &= \alpha\tilde{\alpha} S_k^0, \\
\langle D_k \rangle_3 &= \alpha S_k^0, \\
\langle R_k D_k \rangle_3 &= \tilde{\alpha} S_k^0.
\end{aligned}
$$

Then

$$\begin{aligned}\sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] A(\mathbf{R}) \right\rangle_3 &= \left\langle \textcircled{S}\text{—}\textcircled{R} \right\rangle_3 - 2 \left\langle \begin{array}{c}\textcircled{S}\\ \textcircled{R}\end{array} \right\rangle_3 \left\langle \textcircled{R}\text{—}\textcircled{R} \right\rangle_3 \\ &= 2(\alpha\tilde{\alpha} - \alpha^3\tilde{\alpha}^3) A(\mathbf{S}^0), \\ \sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] A(\mathbf{D}) \right\rangle_3 &= \left\langle \textcircled{R}\textcircled{S}\textcircled{D}\text{—}\textcircled{D} \right\rangle_3 - 2 \left\langle \begin{array}{c}\textcircled{S}\\ \textcircled{R}\end{array} \right\rangle_3 \left\langle \textcircled{D}\text{—}\textcircled{D} \right\rangle_3 \\ &= 2(\alpha\tilde{\alpha} - \alpha^3\tilde{\alpha}) A(\mathbf{S}^0).\end{aligned}$$

Similarly

$$\begin{aligned}\sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] B(\mathbf{R}) \right\rangle_3 &= 2(\alpha\tilde{\alpha} - \alpha^3\tilde{\alpha}^3) B(\mathbf{S}^0), \\ \sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] B(\mathbf{D}) \right\rangle_3 &= 2(\alpha\tilde{\alpha} - \alpha^3\tilde{\alpha}) B(\mathbf{S}^0), \\ \sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] C(\mathbf{R}) \right\rangle_3 &= 4(\alpha^3\tilde{\alpha}^3 - \alpha^5\tilde{\alpha}^5) C(\mathbf{S}^0), \\ \sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] C(\mathbf{D}) \right\rangle_3 &= 4(\alpha^3\tilde{\alpha} - \alpha^5\tilde{\alpha}) C(\mathbf{S}^0).\end{aligned}$$

Therefore we have

$$\sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] A^2(\mathbf{R}) \right\rangle_3 = 4(\alpha\tilde{\alpha} - \alpha^3\tilde{\alpha}^3) B(\mathbf{S}^0) + 8(\alpha^3\tilde{\alpha}^3 - \alpha^5\tilde{\alpha}^5) C(\mathbf{S}^0), \tag{3.97}$$

and

$$\begin{aligned}&\sum_k \left\langle [R_k S_k^0 - \alpha\tilde{\alpha}] A(\mathbf{R}) A(\mathbf{D}) \right\rangle_3 = \left\langle \begin{array}{c}\textcircled{S}\text{—}\textcircled{R}\\ \textcircled{D}\text{—}\textcircled{D}\end{array} \right\rangle_3 - 2 \left\langle \begin{array}{c}\textcircled{R}\text{—}\textcircled{R}\\ \textcircled{D}\text{—}\textcircled{D}\end{array} \right\rangle_3 \left\langle \begin{array}{c}\textcircled{S}\\ \textcircled{R}\end{array} \right\rangle_3 \\ &\quad + \left\langle \begin{array}{r}\textcircled{S}\text{—}\textcircled{R}\textcircled{D}\\ |\\ \textcircled{D}\end{array} \right\rangle_3 + \left\langle \begin{array}{r}\textcircled{R}\text{—}\textcircled{S}\textcircled{D}\\ |\\ \textcircled{D}\end{array} \right\rangle_3 + \left\langle \begin{array}{r}\textcircled{R}\text{—}\textcircled{R}\textcircled{D}\\ |\\ \textcircled{D}\\ \textcircled{S}\textcircled{R}\end{array} \right\rangle_3 - 3 \left\langle \begin{array}{r}\textcircled{R}\text{—}\textcircled{R}\textcircled{D}\\ |\\ \textcircled{D}\end{array} \right\rangle_3 \left\langle \begin{array}{c}\textcircled{S}\\ \textcircled{R}\end{array} \right\rangle_3 \\ &\quad + 2 \left\langle \begin{array}{cc}\textcircled{S} & \textcircled{D}\\ | & |\\ \textcircled{R} & \textcircled{D}\end{array} \right\rangle_3 + 2 \left\langle \begin{array}{cc} & \textcircled{S}\textcircled{R}\\ \textcircled{R} & \textcircled{D}\\ | & |\\ \textcircled{R} & \textcircled{D}\end{array} \right\rangle_3 - 4 \left\langle \begin{array}{cc}\textcircled{R} & \textcircled{D}\\ | & |\\ \textcircled{R} & \textcircled{D}\end{array} \right\rangle_3 \left\langle \begin{array}{c}\textcircled{S}\\ \textcircled{R}\end{array} \right\rangle_3 \\ &= \nu N(\alpha\tilde{\alpha} - \alpha\tilde{\alpha}^3) + 2(\alpha^3\tilde{\alpha} + \tilde{\alpha}^3\alpha + \alpha\tilde{\alpha} - 3\alpha^3\tilde{\alpha}^3) B(\mathbf{S}^0) \\ &\quad + 4(\alpha^3\tilde{\alpha} + \tilde{\alpha}^3\alpha^3 - 2\alpha^5\tilde{\alpha}^3) C(\mathbf{S}^0).\end{aligned}$$

Finally, we obtain the average overlap between the reconstruction and the fixed source $\mathbf{S}^0$:

$$\begin{aligned}\left\langle\left\langle \mathbf{S}^0.\mathbf{R}\right\rangle_R\right\rangle_D &= \alpha\tilde{\alpha}\left\{1+\tilde{K}(1-\tilde{\alpha}^2)\frac{2A(\mathbf{S}^0)}{N} - \nu\tilde{K}^2\tilde{\alpha}^2(1-\tilde{\alpha}^2)\right. \\ &\left. + \tilde{K}^2\frac{2B(\mathbf{S}^0)}{N}(1-\tilde{\alpha}^2)\left[1-\tilde{\alpha}^2(1+\alpha^2)\right] + o(\tilde{K}^3)\right\}. \qquad (3.98)\end{aligned}$$

Notice that the configurational sum over disconnected two link graphs, $C()$, does not contribute to the result. We are calculating an intensive quantity that does not depend on the system size N, therefore we would not expect terms such as $\frac{1}{N}C(\mathbf{S}^0)$ to contribute as they are of order N. This is a general point when calculating physical quantities: only connected graphs contribute to the result.

3.6.2 Calculation of the Width W_R

We have still to calculate $\left\langle\langle R_k[\mathbf{D}]\rangle_R^2\right\rangle_D$ for the fixed source case.

$$\begin{aligned}\left\langle\langle R_k[\mathbf{D}]\rangle_R^2\right\rangle_D &= \sum_{\{\mathbf{R}\}}\sum_{\{\mathbf{R}'\}} R_k R_k' \sum_{\{\mathbf{D}\}} P(\mathbf{R}|\mathbf{D})P(\mathbf{R}'|\mathbf{D})P(\mathbf{D}|\mathbf{S}^0) \\ &= \sum_{\{\mathbf{R}\}}\sum_{\{\mathbf{R}'\}} R_k R_k' \sum_{\{\mathbf{D}\}} \frac{1}{Z_l(h)Z^2(\tilde{K},\tilde{h};\mathbf{D})}\exp\left\{\tilde{K}\sum_{<ij>} R_i R_j\right. \\ &\quad\left. + \tilde{K}\sum_{<ij>} R_i' R_j' + \tilde{h}\sum_i R_i D_i + \tilde{h}\sum_i R_i' D_i + h\sum_i S_i^0 D_i\right\} \\ &= \frac{1}{N}\sum_k \left\langle R_k R_k' g_4(\tilde{K},\tilde{h};\mathbf{D},\mathbf{R})\right\rangle_4,\end{aligned}$$

where the average

$$\langle\cdot\rangle_4 \overset{\text{def}}{=} \frac{1}{Z_l(h)Z_l^2(\tilde{h})} \sum_{\{\mathbf{R}\}} \sum_{\{\mathbf{R}'\}} \sum_{\{\mathbf{D}\}} (\cdot) \exp\left\{ \tilde{h} \sum_i R_i D_i + \tilde{h} \sum_i R_i' D_i + h \sum_i S_i^0 D_i \right\}, \tag{3.99}$$

and the function

$$\begin{aligned} g_4(\tilde{K}, \tilde{h}; \mathbf{D}, \mathbf{R}) &\overset{\text{def}}{=} \frac{Z_l^2(\tilde{h})}{Z^2(\tilde{K}, \tilde{h}; \mathbf{D})} \exp\left\{ \tilde{K}[A(\mathbf{R}) + A(\mathbf{R}')] \right\} \\ &= 1 + \tilde{K}\left[A(\mathbf{R}) + A(\mathbf{R}') - 2\tilde{\alpha}^2 A(\mathbf{D})\right] \\ &\quad + \frac{1}{2}\tilde{K}^2\left[A^2(\mathbf{R}) + 2A(\mathbf{R})A(\mathbf{R}') + A^2(\mathbf{R}')\right] \\ &\quad - 2\tilde{K}^2\tilde{\alpha}^2 A(\mathbf{D})[A(\mathbf{R}) + A(\mathbf{R}')] \\ &\quad + 2\tilde{K}^2\tilde{\alpha}^2(3\tilde{\alpha}^2 - 1)B(\mathbf{D}) + o(\tilde{K}^3). \end{aligned} \tag{3.100}$$

We have not included in the above expression any terms involving the disconnected graphs $C()$ nor any constants.

The averages we require are:

$$\begin{aligned} \langle R_k \rangle_4 &= \alpha\tilde{\alpha} S_k^0, \\ \langle D_k \rangle_4 &= \alpha S_k^0, \\ \langle R_k D_k \rangle_4 &= \tilde{\alpha}, \\ \langle R_k R_k' \rangle_4 &= \tilde{\alpha}^2, \\ \langle R_k R_k' D_k \rangle_4 &= \alpha\tilde{\alpha}^2 S_k^0. \end{aligned}$$

Then

$$\begin{aligned} \sum_k \left\langle [R_k R_k' - \tilde{\alpha}^2] A(\mathbf{R}) \right\rangle_4 &= \left\langle \textcircled{R}' \!-\!\!-\!\!-\! \textcircled{R} \right\rangle_4 - 2\left\langle \begin{matrix} \textcircled{R}' \\ \textcircled{R} \end{matrix} \right\rangle_4 \left\langle \textcircled{R} \!-\!\!-\!\!-\! \textcircled{R} \right\rangle_4 \\ &= 2(\alpha^2\tilde{\alpha}^2 - \alpha^2\tilde{\alpha}^4) A(\mathbf{S}^0), \end{aligned}$$

$$\begin{aligned}\sum_k \left\langle [R_k R'_k - \tilde{\alpha}^2] A(\mathbf{D}) \right\rangle_4 &= \left\langle \textcircled{R}\textcircled{R}\textcircled{D}\!-\!\textcircled{D} \right\rangle_4 - 2 \left\langle \genfrac{}{}{0pt}{}{\textcircled{R}}{\textcircled{R}} \right\rangle_4 \left\langle \textcircled{D}\!-\!\textcircled{D} \right\rangle_4 \\ &= 0.\end{aligned}$$

Similarly

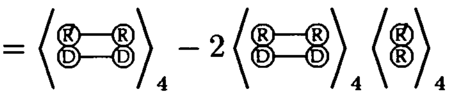
$$\begin{aligned}\sum_k \left\langle [R_k R'_k - \tilde{\alpha}^2] B(\mathbf{R}) \right\rangle_4 &= 2(\alpha^2 \tilde{\alpha}^2 - \alpha^2 \tilde{\alpha}^4) B(\mathbf{S}^0), \\ \sum_k \left\langle [R_k R'_k - \tilde{\alpha}^2] B(\mathbf{D}) \right\rangle_4 &= 0.\end{aligned}$$

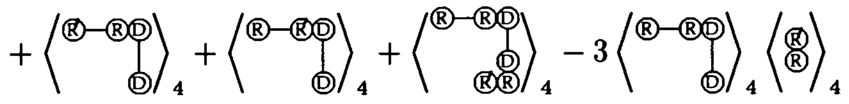
Therefore we have

$$\sum_k \left\langle [R_k R'_k - \tilde{\alpha}^2] A^2(\mathbf{R}) \right\rangle_4 = 4(\alpha^2 \tilde{\alpha}^2 - \alpha^2 \tilde{\alpha}^4) B(\mathbf{S}^0), \tag{3.101}$$

and

$$\sum_k \left\langle [R_k R'_k - \tilde{\alpha}^2] A(\mathbf{R}) A(\mathbf{D}) \right\rangle_4 =$$

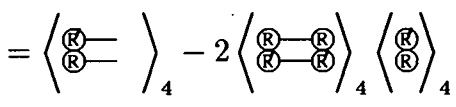


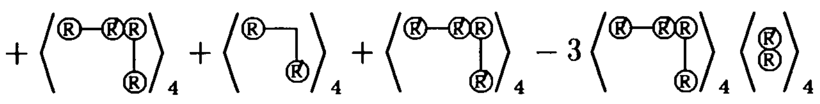


$$= \nu N(\tilde{\alpha}^2 - \tilde{\alpha}^4) + 4(\tilde{\alpha}^2 - \tilde{\alpha}^4)\alpha^2 B(\mathbf{S}^0),$$

and

$$\sum_k \left\langle [R_k R'_k - \tilde{\alpha}^2] A(\mathbf{R}) A(\mathbf{R}') \right\rangle_4 =$$





$$= \nu N(\tilde{\alpha}^2 - \tilde{\alpha}^6) + 2\left(2\tilde{\alpha}^4 + \tilde{\alpha}^2 - 3\tilde{\alpha}^6\right)\alpha^2 B(\mathbf{S}^0).$$

Chequersize	$A(\mathbf{S}^0)/N$	$B(\mathbf{S}^0)/N$
2x2	0	−2
3x3	1/3	−2/9
4x4	1/2	1
8x8	3/4	13/4

Table 3.2. Table of small coupling constants for fixed source chequerboards. Since $A(\mathbf{S}^0)$ and $B(\mathbf{S}^0)$ are two-point functions it is straightforward to calculate these same functions for the data picture: $A(\mathbf{D}) = (1-2q)^2 A(\mathbf{S}^0)$ and $B(\mathbf{D}) = (1-2q)^2 B(\mathbf{S}^0)$.

Finally we obtain

$$\begin{aligned}\left\langle \langle R_k[\mathbf{D}]\rangle_R^2 \right\rangle_D &= \tilde{\alpha}^2 \left\{ 1 + 4\tilde{K}\frac{A(\mathbf{S}^0)}{N}\alpha^2(1-\tilde{\alpha}^2) + \nu\tilde{K}^2(1-3\tilde{\alpha}^2)(1-\tilde{\alpha}^2) \right. \\ & \left. \tilde{K}^2\frac{2B(\mathbf{S}^0)}{N}\alpha^2(1-\tilde{\alpha}^2)(3-5\tilde{\alpha}^2) \right\} . \qquad (3.102)\end{aligned}$$

3.6.3 Results

Combining (3.102) with the result for the overlap of the restored distribution with the source (3.98), and recognizing that $W_S = 0$ for a fixed source, we can write down the small coupling expansion for the quality factor in the fixed source case:

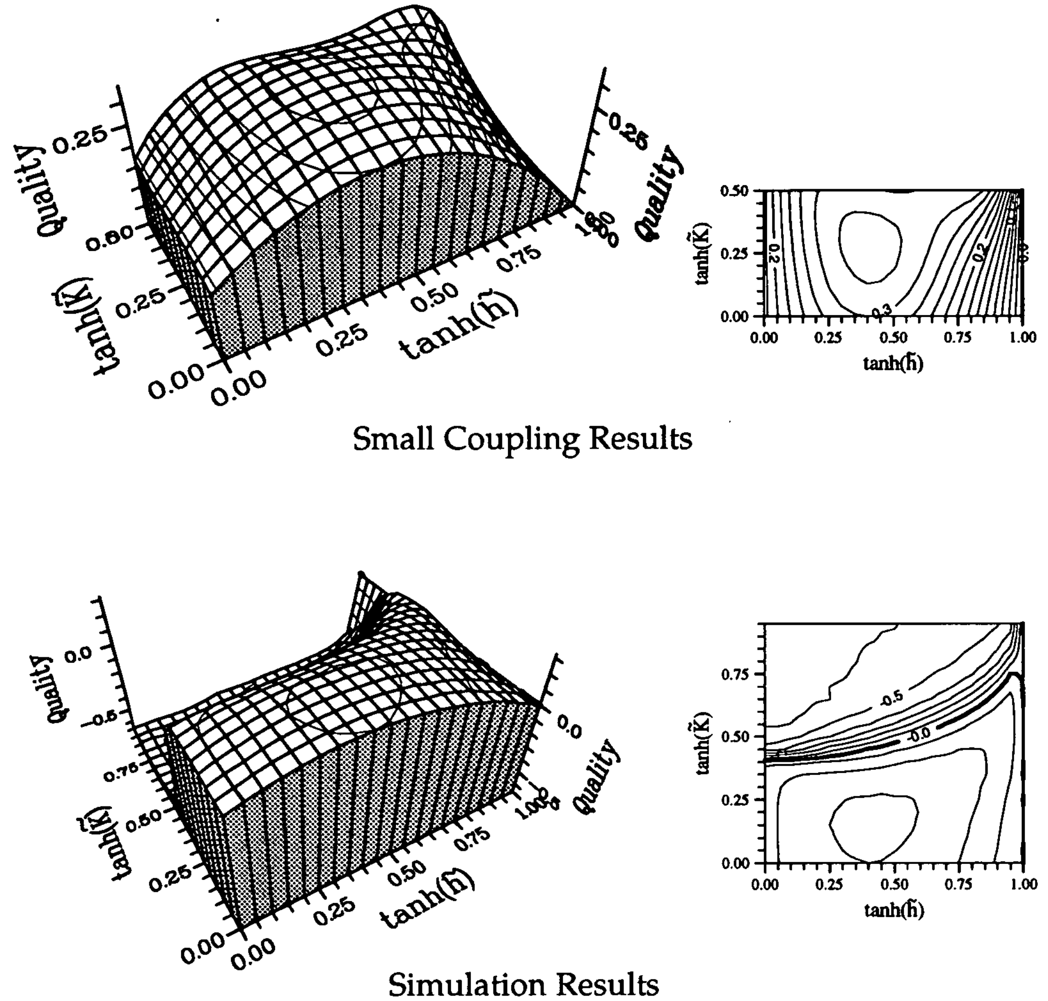
$$\begin{aligned} Q &= \left(2\alpha\tilde{\alpha} - \tilde{\alpha}^2 + 4\tilde{K}\frac{A(\mathbf{S}^0)}{N}(1-\tilde{\alpha}^2)\left[\alpha\tilde{\alpha} - \alpha^2\tilde{\alpha}^2\right] \right. \\ & + 2\tilde{K}^2\frac{B(\mathbf{S}^0)}{N}(1-\tilde{\alpha}^2)\left[2\alpha\tilde{\alpha}(1-\tilde{\alpha}^2(1+\alpha^2)) - \tilde{\alpha}^2\alpha^2(3-5\tilde{\alpha}^2)\right] \\ & \left. -\nu\tilde{K}^2\tilde{\alpha}^2(1-\tilde{\alpha}^2)\left[2\alpha\tilde{\alpha} + 1 - 3\tilde{\alpha}^2\right] \right) / \left(2(1-\alpha)\right) . \qquad (3.103)\end{aligned}$$

Figure 3.15 displays the results for a chequerboard source, obtained using this last calculation. Although the calculation is valid for any chequerboard size (see Table 3.2 for the values of $A(\mathbf{S}^0)$ and $B(\mathbf{S}^0)$ that should be used for different chequerboard sizes), we choose a small chequer size. This ensures that the interesting structure in the quality factor appears within the range of values of $\tilde{K}$ for which the approximation is reasonably accurate. Once again we see excellent agreement between simulation and theory. [However the quality factor is low, and we are not in a regime where the restoration scheme is very useful.]

3.6.4 Connections

There is a final consistency check that we can perform for these small coupling results. If we take a sample configuration from an Ising source and use this as our fixed source [i.e. measure $A(\mathbf{S})$ and $B(\mathbf{S})$], we should expect the fixed source result (3.103) to recover the Ising source result (3.94). What we require are the small coupling (in K) expansions of the fixed source measures $A(\mathbf{S})$ and $B(\mathbf{S})$ where the source $\mathbf{S}$ is a sample from an Ising distribution. These are easily calculated from the near-neighbour correlation functions of the Ising model (see e.g. [114]):

$$\begin{aligned} \frac{2}{\nu N} A(\mathbf{S}) &= \langle S_{0,0} S_{0,1} \rangle \qquad (3.104) \\ &= K + o(K^3), \\ \frac{2}{\nu(\nu-1)N} B(\mathbf{S}) &= \frac{2}{3} \langle S_{0,0} S_{1,1} \rangle + \frac{1}{3} \langle S_{0,0} S_{0,2} \rangle \\ &= \frac{5}{3} K^2 + o(K^4). \qquad (3.105) \end{aligned}$$



Small Coupling Results

Simulation Results

Figure 3.15. Small coupling expansion results for a fixed source. The system considered is a 3x3 chequerboard source with 30% noise. The upper row shows the quality factor calculated by the small coupling expansion, while the lower shows the same results from simulation. Although there is some small discrepancy in the position of the maximum due to the errors of order $(K + \tilde{K})^3$, overall we see very good agreement between simulation and the results of the calculation.

When we replace $\frac{1}{N}A(\mathbf{S})$ by K and neglect terms involving $B(\mathbf{S})$ as they are more than second order in $(K+\tilde{K})$ we find that the fixed source results (3.98) and (3.102) reduce to the Ising source results (3.86) and (3.92).

Given this agreement it is to be expected that, for a fixed source with a high density of edges, the optimal choice of $\tilde{K}$ and $\tilde{h}$ should match the value of the noise parameter h, and the value of the coupling K_{eff} that would generate Ising configurations with the correct density of edges. However, for lower densities of edges, corresponding to larger values of K_{eff} outside the regime of the small coupling expansion, this is not the case. The simulation results at the beginning of this chapter demonstrated the discrepancy between the optimal choice of restoration parameter $\tilde{K}$ and the effective coupling K_{eff} for larger chequerboards.

3.7 Conclusion

We have now completed our exploration of hypothesis space $(\tilde{K},\tilde{h})$. The aim was to investigate the manner in which the performance of the restoration scheme changes for different values of the restoration parameters, with particular interest in establishing the points in hypothesis space that provide the best restoration.

Throughout the work we used the edge-density prior. The question of how the performance of the restoration is affected by the appropriateness of the prior was addressed by considering two kinds of source process: various fixed chequerboard source pictures, and pictures sampled from a nearest-neighbour Markov random field.

In spite of some severe approximations, the mean field calculation produced quite remarkable qualitative results. The failure of the restoration scheme for large values of the nearest neighbour coupling was identified as the result of a transition to long range order in the restoration model: the smoothing effect of the prior wins over the data.

Finally, we have demonstrated the applicability of series expansion methods to this problem, and these have provided further insight into the criteria that determine the optimal point in hypothesis space.

CHAPTER 4

Exploiting the Posterior: Beyond the Ground State

4.1 Introduction

In the previous chapter we concerned ourselves with the search for the best match between two probability distributions: that generated by the restoration scheme, and the true posterior distribution we would have obtained had we known accurately the parameterization of the source and noise distributions. The quality factor that we used for this comparison was constructed from *averaged* functions of the restored distribution and the source distribution. We now focus our attention on a somewhat different task: that of finding the *single* binary image, constructed in some way from the information contained in the restoration distribution, that best matches the original source picture.

In Chapter 2 we decided that the **overlap** would be our measure of how 'good' the match was between the source picture and our estimate. If we choose to minimize the mean squared error between the pictures (which for binary images is equivalent to maximizing the overlap), we find the optimal Bayesian estimator to be the thresholded posterior mean, or TPM estimate. However, this optimal condition is only guaranteed when the restoration scheme exactly models the true posterior distribution, i.e. the source distribution *is* Ising, the noise process *is* simple Gaussian, and we have correctly chosen $\tilde{K} = K$ and $\tilde{h} = h$. In any other case—either we have guessed the restoration parameters $\tilde{K}$ and $\tilde{h}$ incorrectly, or the functional form of the prior is wrong—the situation is somewhat obscure. Early work by Hunt [61] suggested that the TPM estimate could not be calculated. Perhaps for this reason, in the original GG paper [36] and in much subsequent work (e.g. [24, 37, 40, 35, 103, 109]), the maximum *a posteriori* or MAP estimate is studied. There is only a small body of work that begins to recognize the utility of the TPM estimate [69, 70, 84, 85]. This chapter concentrates on an investigation and comparison of the two different estimates, the MAP and the TPM.

4.2 Finding the MAP Estimate

We first restate the probability distribution that characterizes the restoration scheme:

$$P(\mathbf{R}|\mathbf{D}) = \frac{1}{Z(\tilde{K},\tilde{h};\mathbf{D})} \exp\left\{ \tilde{K} \sum_{<ij>} R_i R_j + \tilde{h} \sum_i R_i D_i \right\}. \qquad (4.1)$$

The MAP estimate is exactly what its name suggests—the single binary image **R** that has maximum probability in the restored distribution (4.1) *a posteriori*, i.e. after the data has arrived. We can therefore simply describe the MAP estimate as the single image that minimizes the cost function

$$E = -\tilde{K} \sum_{<ij>} R_i R_j - \tilde{h} \sum_i R_i D_i. \qquad (4.2)$$

This estimate seems intuitively to be flawed since it discards much of the information available to us in the reconstruction distribution. The exact values of $\tilde{K}$ and $\tilde{h}$ are unimportant, it is merely the *ratio* of the two that determines the configuration of minimum energy, and it is only this *single* configuration that is used; all other images are ignored. This single configuration is simply the **ground state** of the restoration system.

There are many techniques for finding the minimum of a cost function. However, most deterministic methods such as gradient descent (or the greedy algorithm [109]) will get trapped in a metastable state—they find not the global minimum of the system but a point in configuration space that is merely locally stable. With this outcome the result is strongly dependent on the starting configuration used. The standard way to minimize the chances of finishing in a local rather than the global minimum is a stochastic technique known as simulated annealing [74]. Physical chemists use a process of heating followed by slow controlled cooling to remove defects from crystalline substances, and to temper metals. This annealing process provides a pathway for the substance to find a low energy state, free of defects. Simulated annealing emulates this cooling process in a similar attempt to find the ground state of a system (the state with the lowest energy).

To implement simulated annealing we rewrite the cost function (4.2) introducing an inverse temperature β:

$$\mathcal{H} = -\beta \left[\sum_{<ij>} R_i R_j + \frac{\tilde{h}}{\tilde{K}} \sum_i R_i D_i \right] . \tag{4.3}$$

Using this cost function with a small value of β corresponding to a high temperature, we simulate the spin system using the standard Metropolis Monte Carlo algorithm [see §3.2.1]. We then gradually lower the temperature, being careful to re-equilibrate the system at each larger value of β, until the system finds itself in the basin of attraction of the global minimum and settles into the ground state. Provided we cool the system slowly enough, this process enables the system to find its way out of any local minima it may be temporarily trapped in, utilizing the fluctuations in energy available at finite temperature. It is crucial to the success of this scheme that a stochastic technique (such as Metropolis Monte Carlo) is used to find equilibrium at each temperature. Such a stochastic relaxation scheme allows occasional increases in the internal energy of the system, and it is this that allows the system to escape from local minima (although escape becomes less likely at lower temperatures).

GG [36] present a proof that with a suitable annealing schedule (sequence of temperatures) this simulated annealing technique is guaranteed to find the global minimum and hence the exact MAP estimate. However, such a schedule would take a prohibitive length of time to complete (the total number of site updates required is exponential in the system size N). They claim that acceptable results are obtained using a schedule:

$$\beta = C \log(1 + k) \qquad k = 0, \ldots, k_{\max}. \tag{4.4}$$

This means that we start at $\beta = 0$, and at each temperature we allow the system to relax to equilibrium before incrementing k by one and altering the temperature accordingly. In this way the cooling process is complete in around 10^2–10^3 temperature steps, depending on the values of C and $k_{\mathbf{max}}$. Even with this faster schedule the annealing process is still far more computationally intensive than the calculation of the TPM estimate.

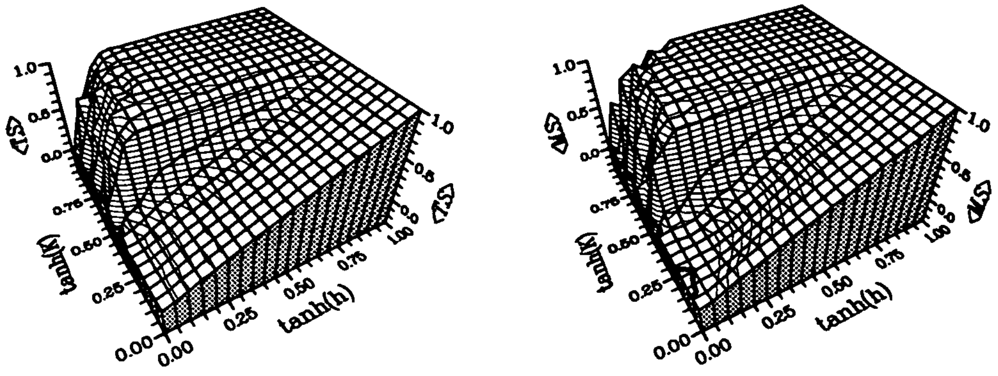
Examining equation (4.3) we can see our second point explicitly. Given a perfect annealing schedule that successfully finds the ground state, the MAP estimate depends only on the ratio of the parameters $\tilde{K}$ and $\tilde{h}$, and not upon their individual values—we have reduced the dimensionality of the parameter space, with all of the information loss that this entails.

4.3 Comparison of MAP and TPM

We carry out a number of experiments to compare the effectiveness of the MAP and TPM estimates and, returning to an earlier theme, we will consider the distinct cases of well-matched and ill-matched priors.

4.3.1 The Well-matched Prior

Initially, let us consider the optimal case discussed earlier. We generate a source picture according to the Ising distribution given by (2.50) and then corrupt this picture using the noise process (2.47) to generate the data. We then attempt the reconstruction using (4.1) with the optimal choice of parameters $\tilde{K} = K$ and $\tilde{h} = h$. From the reconstruction distribution (4.1)



The TPM Estimate

The MAP Estimate

Figure 4.1. The overlap of the source with the TPM and MAP estimates. Note that the coordinate axes represent *both* generation *and* restoration parameters—$K = \tilde{K}$ and $h = \tilde{h}$: we are investigating the *optimal* case of the well-matched prior, for a range of parameters. The figures are essentially similar. For low K the restoration scheme is ineffective resulting in an inclined plane in the TPM case. In the narrow mid-band of K, which covers a large range of source edge-densities, we see good improvement (the overlaps are raised above the inclined plane). Once K gets large the source pictures are almost entirely one colour and it is relatively easy to generate single colour pictures that have a high overlap with the source. Note the poorer performance of the MAP estimate indicated by the gully in the inclined plane. The erratic results observed for large K and small h occur because when there is a lot of noise present the restoration scheme finds it difficult to determine which colour the source was and is just as likely to guess incorrectly (overlap -1) as correctly (overlap $+1$).

we construct the TPM estimate (2.31) for this particular source and data picture, and calculate its overlap with the source [T.S]. Using the energy function given in (4.3) and following the annealing schedule given in (4.4) we obtain the MAP estimate for this source and data, and once again calculate its overlap with the source [M.S]. The annealing schedule we use sets $C = 0.25$ and $k_{\max} = 750$. These results for T.S and M.S are averaged over the source and data distributions for fixed K and h, and the results are presented as a function of $K = \tilde{K}$ and $h = \tilde{h}$ in Figure 4.1.

- For small values of K, the restoration scheme is generally ineffective—there is insufficient structure in the source picture itself for the restoration to do any better than to simply reproduce the data. This leads to the inclined plane seen at the front of the diagram since the overlap of source and data is simply $\tanh(h)$, as given in (2.49).
- Once K exceeds the critical coupling of the Ising model, large scale structure appears in the source pictures, and it is in this region that the scheme is most effective. [This narrow range of K corresponds to a large range of edge densities—see e.g. Table 3.1 and Figure 3.4.]
- As K increases further, we quickly reach a situation where the source pictures are almost entirely one colour. At this point the MAP and TPM estimates both result in a single colour picture, although it is noticeable that some authors select a source configuration before the ensemble has reached equilibrium [84, 85]. This has a large overlap with the source, so the scheme could be considered successful. However, the problem is rather easy and since there is no structure remaining in the estimates, it is questionable whether the restoration is at all useful.

It is in the low $K = \tilde{K}$ regime that the MAP estimate appears to perform poorly compared with the TPM estimate, indicated by a gully running up the inclined plane at the front of the MAP diagram. Figure 4.2, which illustrates the three different cases above, confirms this failure—the MAP estimate appears to oversmooth.

The other question to be asked of the data in Figure 4.1, anticipating the last section in this chapter, is to what extent the results for the MAP estimate have been affected by the simulated annealing process.

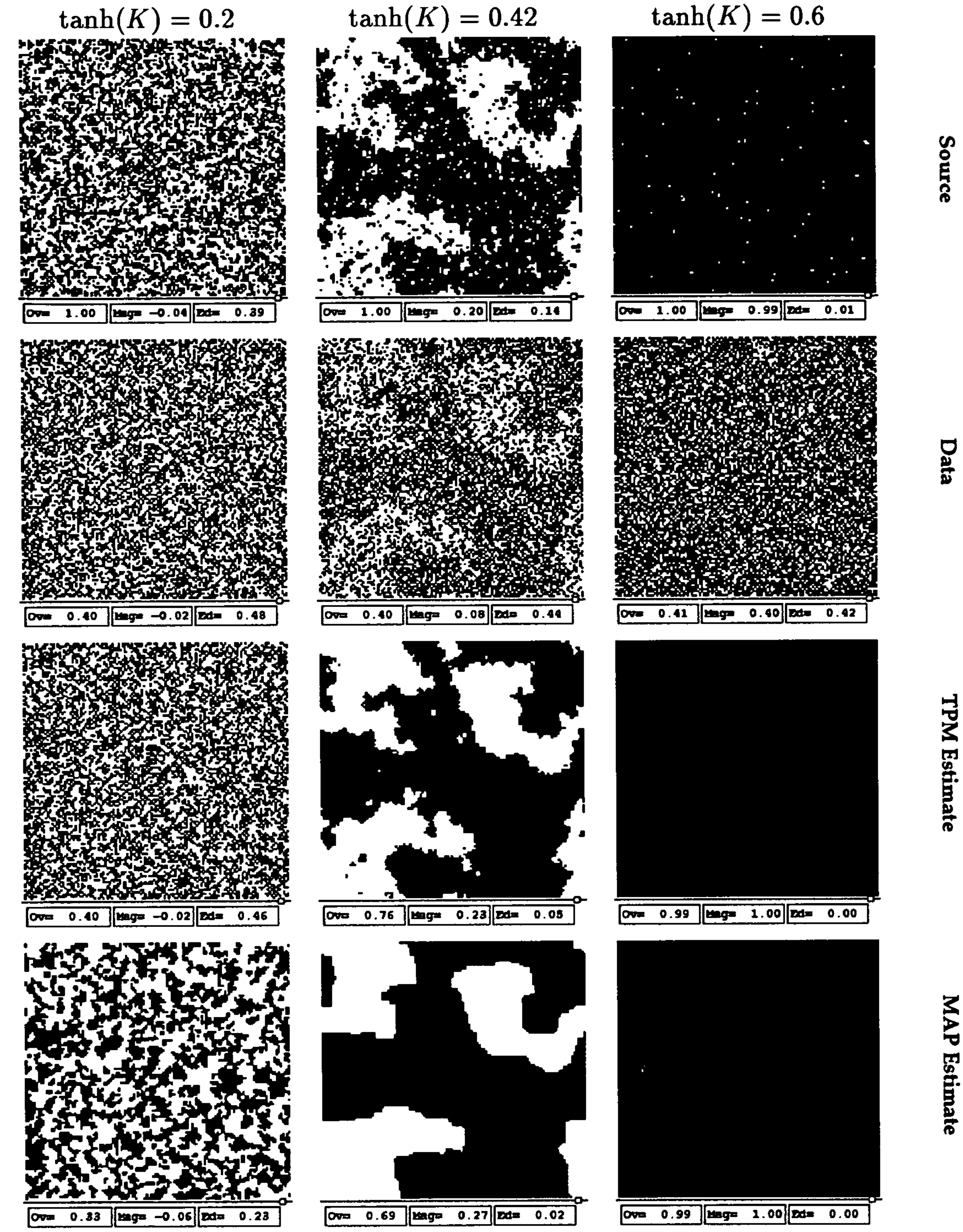
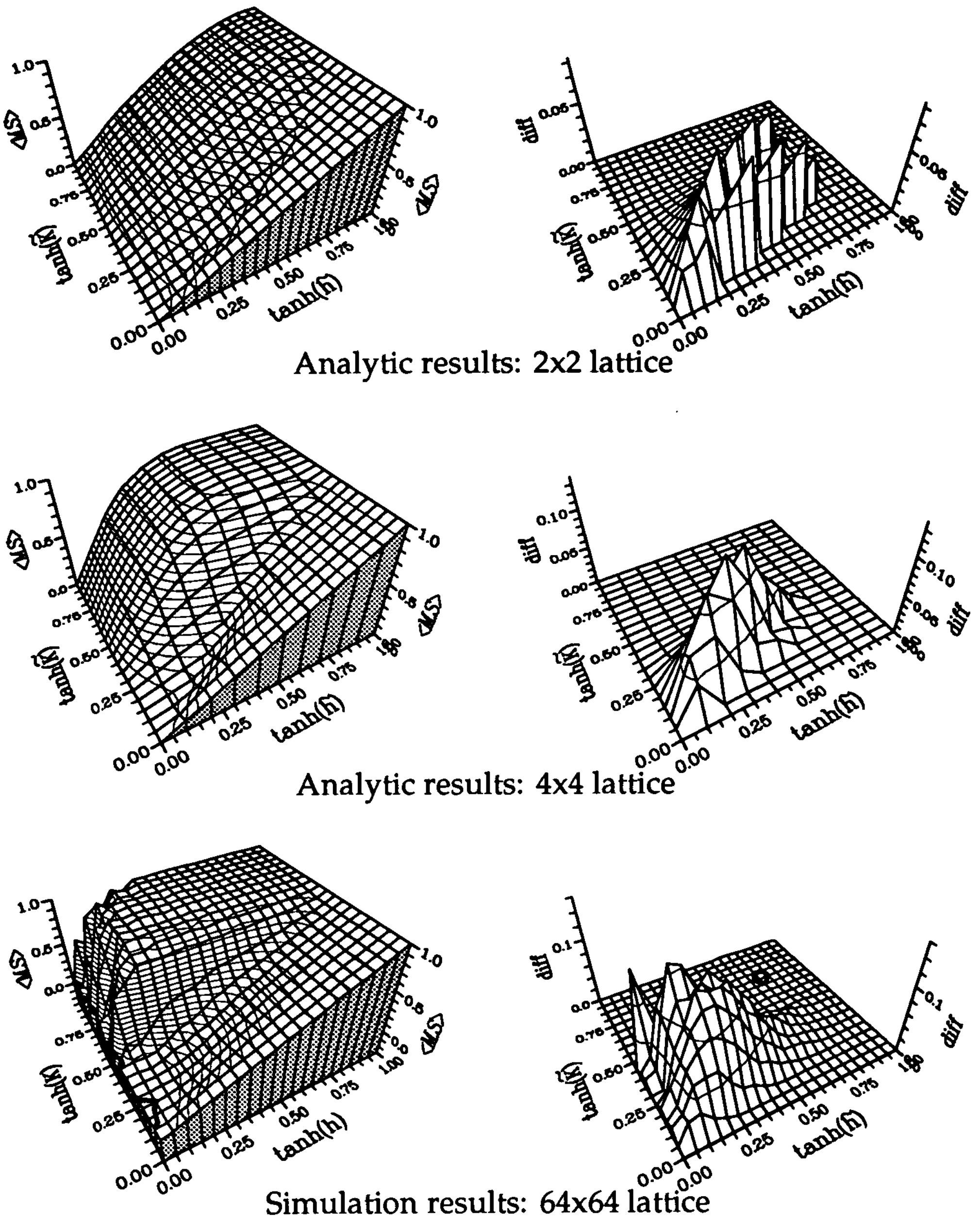

tanh(K) = 0.2
tanh(K) = 0.42
tanh(K) = 0.6
Source
Data
TPM Estimate
MAP Estimate
Ov= 1.00 Mag= -0.04 Ed= 0.39
Ov= 1.00 Mag= 0.20 Ed= 0.14
Ov= 1.00 Mag= 0.99 Ed= 0.01
Ov= 0.40 Mag= -0.02 Ed= 0.48
Ov= 0.40 Mag= 0.08 Ed= 0.44
Ov= 0.41 Mag= 0.40 Ed= 0.42
Ov= 0.40 Mag= -0.02 Ed= 0.46
Ov= 0.76 Mag= 0.23 Ed= 0.05
Ov= 0.99 Mag= 1.00 Ed= 0.00
Ov= 0.33 Mag= -0.06 Ed= 0.23
Ov= 0.69 Mag= 0.27 Ed= 0.02
Ov= 0.99 Mag= 1.00 Ed= 0.00

Figure 4.2. Pictures of optimal restoration using a well-matched prior. The top row shows the source picture generated using the indicated coupling K. The second row shows the corresponding corrupted picture after the application of 30% noise. The TPM and MAP estimates were calculated using the optimal couplings $\tilde{K} = K$ and $\tilde{h} = 0.4$. In general the MAP estimate oversmooths. In the low K case it has grown domains which are larger than any that exist in the source. In the medium K case the TPM estimate clearly outperforms the MAP estimate. At large K both estimates are single colour pictures: the few white pixels in the source are the result of entropic noise and cannot be rendered faithfully by either estimate.

After Marroquin [84, 85] we can calculate the MAP and TPM estimates exactly. For each data configuration **D** we explicitly sum over all possible restored pictures **R** to calculate the TPM estimate, and calculate the relative probability, in (4.1), of each configuration to find the MAP estimate—the configuration **R** that maximizes $P(\mathbf{R}|\mathbf{D})$. These results are then averaged over all data configurations for a particular source, weighting the average according to $P(\mathbf{D}|\mathbf{S})$, and the intermediate results then weighted according to $P(\mathbf{S})$. This process requires very intensive computation to enumerate all of the possible configurations. We extended the calculation from the simplest 2x2 square lattice treated by Marroquin, to 3x3 and 4x4 square lattices. We present these results in Figure 4.3 alongside those already discussed on a 64x64 square lattice using Monte Carlo methods.

The exact calculation results verify the results from Monte Carlo simulation, with some finite size scaling effects in evidence. The double ridge effect observed by Marroquin [84, 85] is seen to be an artefact of the 2x2 lattice only. However, the gully observed in the simulation results is clearly seen to be a true feature of the MAP estimate and is not simply an artefact of the annealing process. Notice that the magnitude of the difference between the two estimates increases with the size of the system.

In conclusion, for this optimal case, Figure 4.3 shows that the results for the MAP and TPM estimates are generally comparable. Where they do differ, however, we find as expected that the TPM estimate always does better than the MAP estimate.



Analytic results: 2x2 lattice

Analytic results: 4x4 lattice

Simulation results: 64x64 lattice

Figure 4.3. Comparison of MAP and TPM results for well-matched prior at different lattice sizes. The left hand column shows the average overlap of the MAP estimate with the source [M.S] while the right-hand column shows the average difference between the overlaps of the two estimates [T.S – M.S]. For brevity, the TPM results and the results for the exact 3x3 lattice are omitted. The finite-size effects can be seen clearly. In all cases the difference in overlaps is non-negative over the whole of parameter space: when the restoration parameters are chosen optimally the TPM estimate always beats the MAP estimate. The amount by which TPM beats MAP increases with the size of the lattice, which indicates that the shortcomings in MAP will be exacerbated in realistically sized pictures.

4.3.2 The Ill-matched Prior

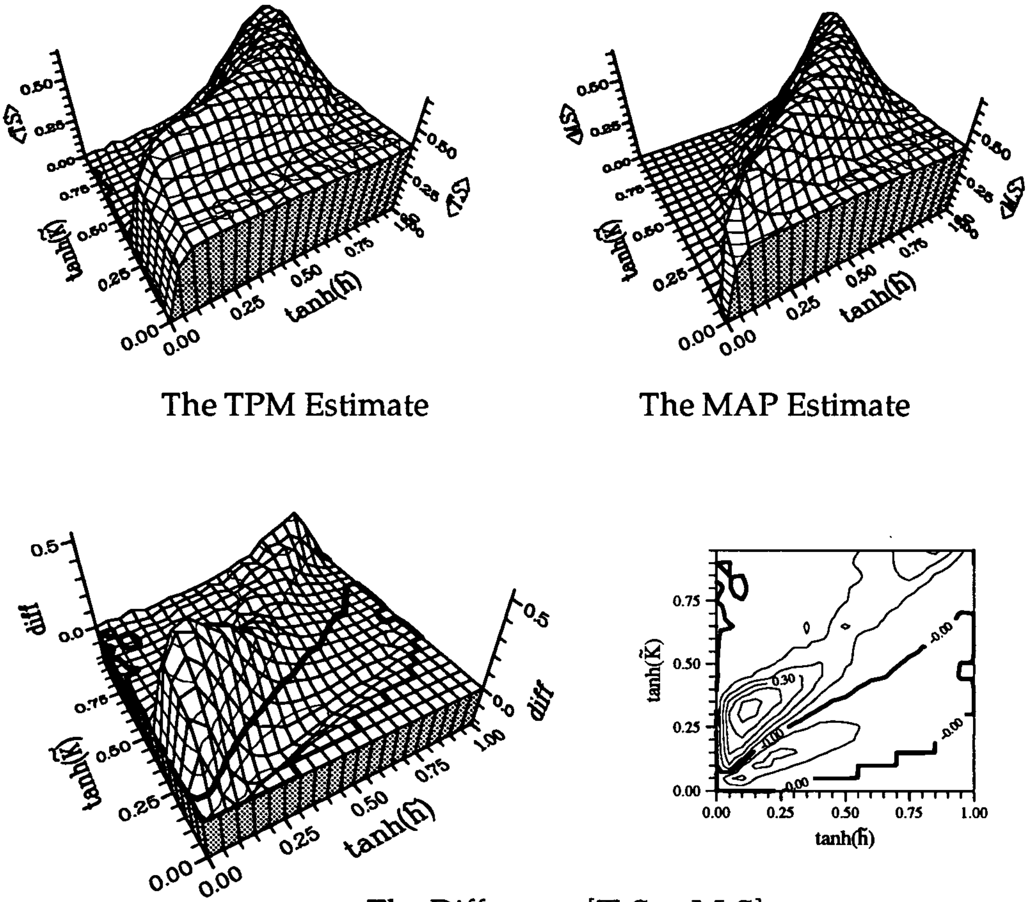
We have seen that in the case of a well-matched prior it is always better to use the TPM estimate than the MAP estimate. But what happens when we get the prior parameters wrong? Now $\tilde{K} \neq K$ or $\tilde{h} \neq h$. Or perhaps the source was not generated by an Ising process at all, as in the chequerboard source images we have considered.

Figure 4.4 shows a fairly typical example of the results obtained in such a case (again an 8x8 chequerboard with 30% noise). As we stated earlier, there is in fact only one free parameter determining the result of the MAP estimate—the ratio of $\tilde{h}/\tilde{K}$. However, for comparison purposes, we have expanded the parameter space into two dimensions to match the parameter space for the TPM estimate.

- For some choices of $\tilde{K}$ and $\tilde{h}$ it is evident that the MAP estimate does beat the TPM estimate—in the region of parameter space where this occurs, the TPM estimate finds insufficient information to make any improvement and simply returns the data as its best guess.

- There is a region where the TPM estimate is clearly better than the MAP estimate, and in much of this region the MAP estimate fails to return anything sensible whatsoever (the results are severely oversmoothed prior-like images).

Figure 4.5 shows a prime example of this oversmoothing. The restoration parameters $\tilde{K}$ and $\tilde{h}$ have been chosen to maximize the quality factor and hence provide the optimal TPM image. The corresponding MAP image obtained by simulated annealing looks nothing like the source, or the data.



The TPM Estimate

The MAP Estimate

The Difference [T.S − M.S]

Figure 4.4. Comparison of MAP and TPM for ill-matched prior (8x8 chequerboard with 30% noise). The upper row shows the qualitative difference in the results obtained using the TPM and MAP estimates. The bottom figures show the difference between the overlaps of the TPM and MAP estimates with the source. This shows that for the majority of parameter choices $(\tilde{K},\tilde{h})$ the TPM estimate does better, while in a smaller region the MAP estimate is preferred. However, the *best* results that can be obtained from each estimate are comparable.

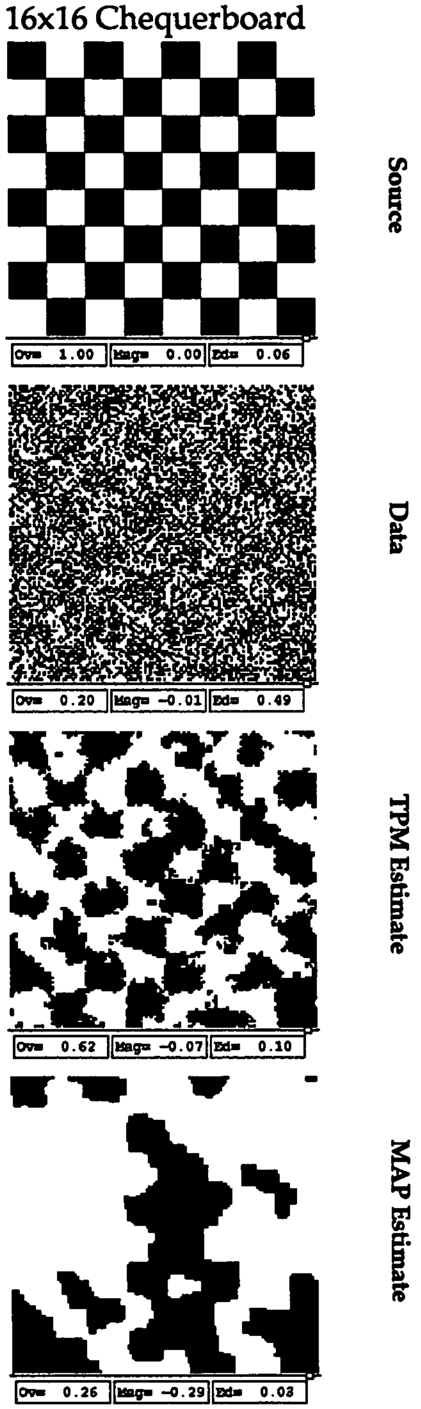


Figure 4.5. Pictures of optimal restoration using an ill-matched prior. The source is a 16x16 chequerboard, subsequently corrupted by 40% noise as in Figure 3.2. Compare with the middle column of Figure 4.2. When the restoration parameters are chosen to optimize the TPM estimate, the corresponding MAP estimate is badly over-smoothed. [But note that the MAP estimate can perform almost as well as this optimal TPM estimate for carefully chosen values of the restoration parameters.]

In spite of the adequate performance of the MAP estimate in certain regions of parameter space, we still claim that the TPM is the preferred estimate. For all source pictures we considered we found without exception that:

- the best achievable MAP estimate is never better (in terms of the overlap with the source) than the best TPM reconstruction;
- the TPM estimate beats the MAP estimate in a greater volume of parameter space than the reverse; and
- the volume in parameter space that yields any chosen level of improvement is alway greater in the TPM case.

The degree of these effects changes qualitatively with different noise levels and chequerboard sizes. With a 4x4 chequerboard the restoration task is far more difficult due to the smaller size of the coherent regions, and the TPM estimate is preferred over almost all of parameter space.

4.3.3 Other Issues in MAP Estimation

We now turn to a different issue surrounding the MAP estimate. More than one paper [44, 84, 85] has shown examples of cases where the MAP estimate is patent nonsense: the restoration parameters have placed us in the prior-like phase and we simply get the edge-free state as the solution. In particular, Greig *et al.* [44] use a version of the Ford-Fulkerson algorithm [27] to calculate the *exact* MAP estimate for some 64x64 scenes and compare them with the MAP estimate obtained by simulated annealing using various annealing schedules. They found that when the ratio $\tilde{K}/\tilde{h}$ was large, the exact MAP estimate was all one colour—the edge-free state—which is

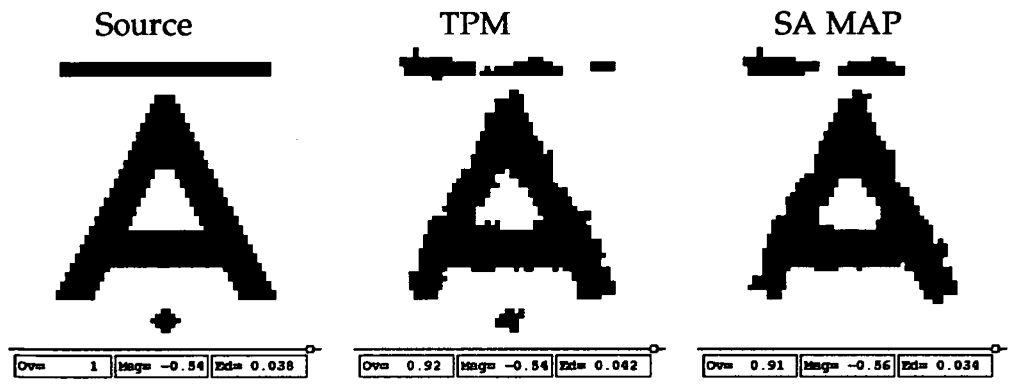


Figure 4.6. Sample restoration of the synthetic image from [44] after application of 25% noise. The restorations were performed using $\tilde{K} = 0.87$ and $\tilde{h} = 1.1$, detailed in the last row of Table 4.1, and found by maximizing the quality factor. The TPM estimate does marginally better than MAP.

		Overlap with source		
$\tilde{K}$	$\tilde{h}$	MAP		TPM
		Exact	annealed	
0.3	0.55	0.896	0.909	0.818
0.7	0.55	0.808	0.867	0.878
1.1	0.55	0.544	0.780	0.829
0.87	1.1	—	0.908	0.921

Table 4.1. TPM results for the synthetic image in Figure 4.6 compared with annealed and exact MAP results. The results for the exact MAP estimate are taken from Table 2 of [44]: the TPM estimate for the optimal restoration parameters, detailed in the last row above, outperforms all of the results presented there.

in general a very poor estimate of the true scene! However, using most annealing schedules, provided the initial configuration used is the data, the system becomes trapped in a metastable state closer to the data and the resulting image is a much better estimate than the exact MAP estimate. The *inadequacy* of the annealing technique with these schedules results in a better restoration than the true MAP estimate would provide.

We repeated the experiment on their synthetic image (shown in Figure 4.6) verifying their annealing result and also comparing both the exact and annealed MAP estimates with the TPM estimate. The results are shown in Table 4.1 and Figure 4.6. As before, the TPM estimate is overall most effective, although there are certain suboptimal choices of $\tilde{K}$ and $\tilde{h}$ where the annealed MAP estimate may do better.

4.4 A Discussion on the MAP estimate

We can go some way toward an explanation of the MAP estimate by making use of the phase diagrams that the mean field approximation provided in Chapter 3. We should first recognize what it is that we do when we anneal to find the MAP estimate—in effect we increase $\tilde{K}$ and $\tilde{h}$ simultaneously while maintaining a constant ratio $\tilde{K}/\tilde{h}$ [or we increase β in (4.3)]. Since the graphs we present have $[\tanh(\tilde{K})]$ and $[\tanh(\tilde{h})]$ as the axis variables, the isolines of constant ratio are not straight (except for the trivial case of $\tilde{K} = \tilde{h}$) and we show these isolines in Figure 4.7. These lines necessarily do not cross and, after we pick values of $\tilde{K}$ and $\tilde{h}$ to begin annealing from, the parameter values follow the isoline that passes through that point $(\tilde{h}, \tilde{K})$ up to the top right hand corner of the phase diagram.

Why MAP Fails

Now look at Figure 4.8, the mean-field phase diagram for the 8x8 chequerboard with 30% noise. [See Figure 3.7 for an explanation of the nature of the different phases.] Setting $(\tilde{h}, \tilde{K})$ somewhere near (A) places us in a phase where the restored screen is generating data-like pictures which have a non-żero overlap with the source—the TPM estimate will produce

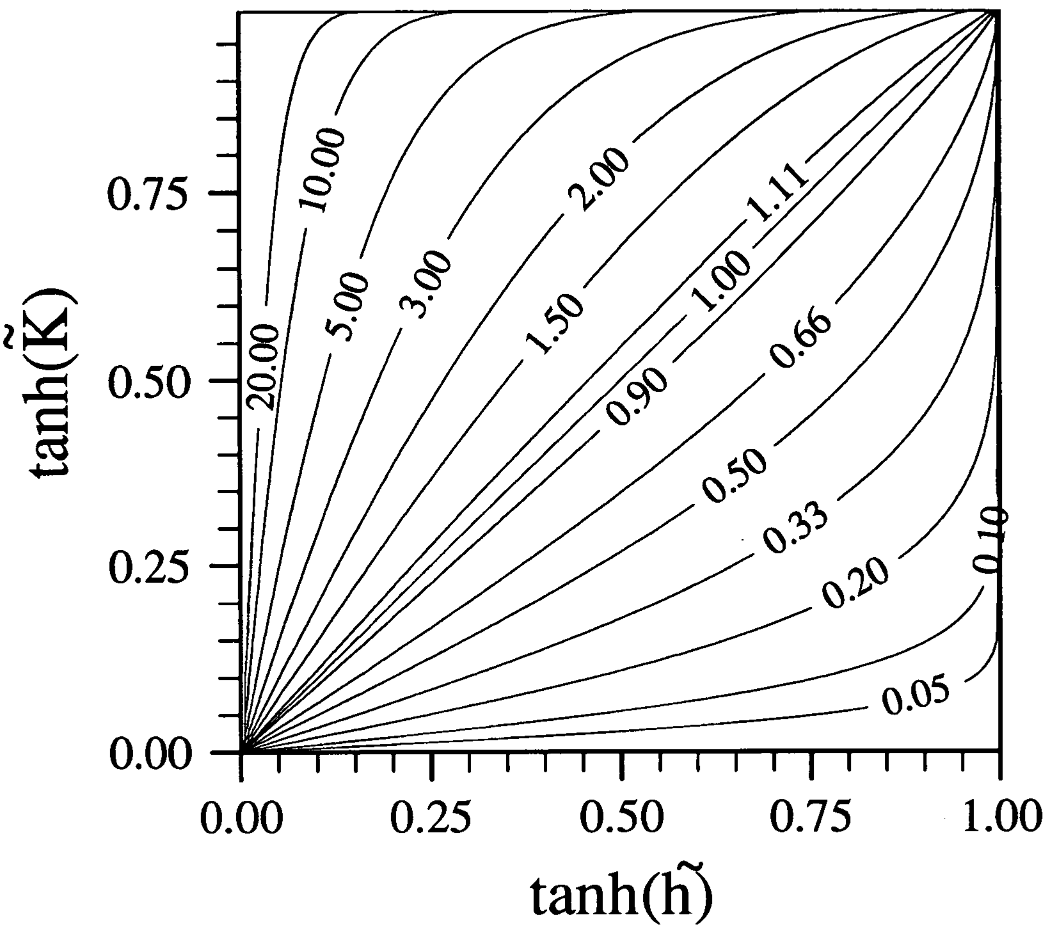


Figure 4.7. Annealing trajectories: lines of constant $\tilde{K}/\tilde{h}$. Whenever we anneal, the restoration parameters are varied together, moving through parameter space along one of the lines shown.

sensible results. When we anneal, however, the system undergoes a phase transition into the prior-like phase and therefore the MAP estimate that we obtain has very few edges. Any starting point that brings the annealing curve into the top right hand corner above the phase transition line will provide a useless prior-like MAP estimate. This is why we get regions where the MAP estimate fails catastrophically, but the TPM estimate is reasonable. The theory is confirmed in the lower diagrams of Figure 4.4 where there is a peak in the difference $[\mathbf{T}-\mathbf{M}].\mathbf{S}$ coincident with the region marked (A) in Figure 4.8.

It is also clear from this analysis of the phase diagram why the region of parameter space in which TPM performs well is always larger than the region in which MAP performs well. Any point in parameter space where the TPM estimate is 'bad' (where 'bad' means worse than the data) is guaranteed to also give a 'bad' MAP estimate. The phase transition line lies in parameter space in such a way that annealing paths only ever cross from the data-like phase to the prior-like phase. Following the annealing curves through parameter space, once the phase transition line has been crossed into the prior-like phase, there is no path back into the data-aligned phase. When the TPM estimate is 'bad' we are already in the prior-like phase so it is impossible for the annealing process to produce a data-like MAP estimate.

Why SA MAP Beats Exact MAP

We can also explain the discrepancy between the exact MAP estimate and what we obtain by simulated annealing. Beginning somewhere in region (B) of Figure 4.8 the TPM estimate will be fine; however, when annealing we cross the phase transition line so the result of the MAP estimate should

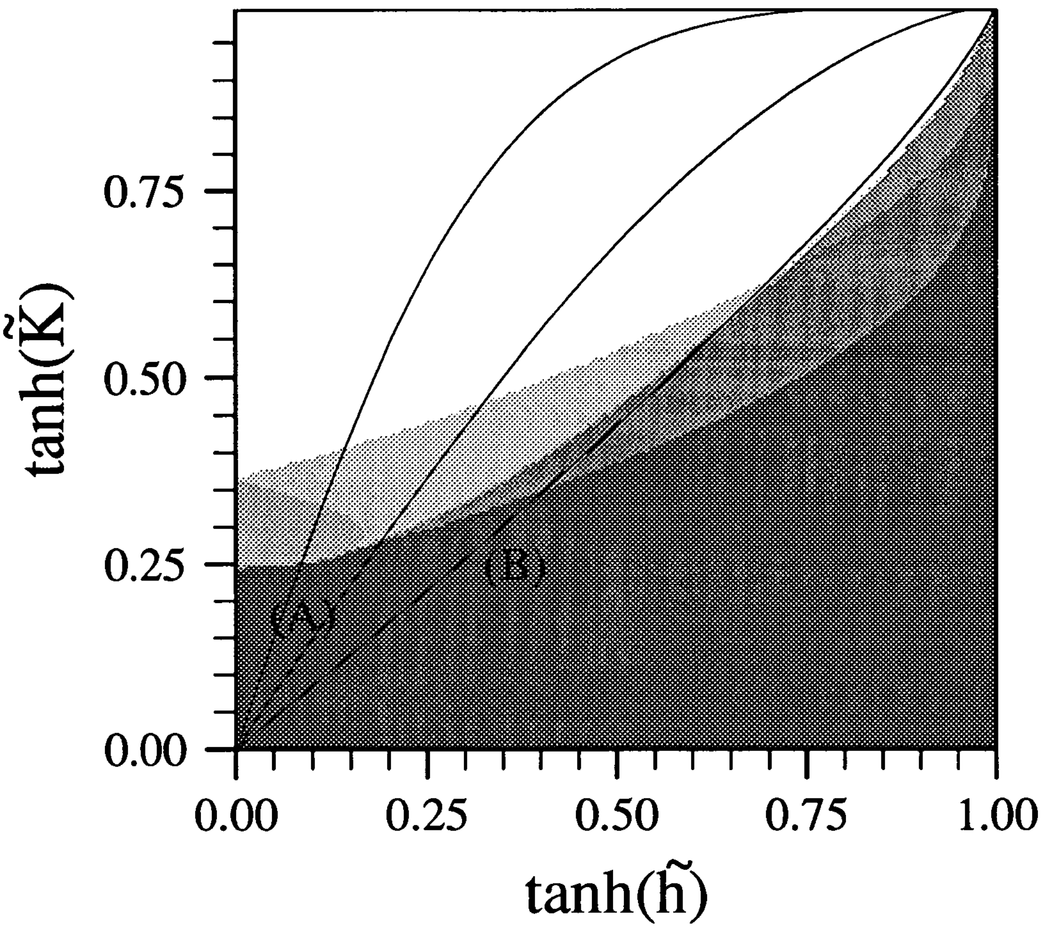


Figure 4.8. Example annealing trajectories through mean field phase space. The phase diagram is for an 8x8 chequerboard with 30% noise and is explained in the caption to Figure 3.7. In region (A) the TPM estimate is data-like, while, after following the annealing trajectories, the MAP estimate is in the ordered prior-like phase. Following the annealing trajectory from region (B) ought to find a MAP estimate in the prior-like phase. However, if the annealing is performed too quickly the system is caught in a metastable data-like state.

be prior-like. But notice that we cross the phase transition line at quite a low effective temperature, and move into a region of the phase diagram where metastable data-like states exist. Therefore the system falls into the nearest metastable state and is unable to attain the true ground state. The annealed MAP estimate is data-like and reasonable while the exact MAP estimate would be prior-like and useless.

4.5 Conclusion

We have carried out a systematic investigation into the efficacy of the MAP estimate in the image restoration problem. We find that in almost all cases the TPM estimate provides a more reliable estimate of the original source image. Although choosing the optimal values for the restoration parameters remains problematical, it is easy to avoid the region of parameter space where the TPM estimate offers no improvement, and it is only a part of this region where MAP may do better than TPM.

The failure of MAP can be understood in terms of the phase diagram. It seems foolish to carry out an annealing process that leads to phase transitions in the system. These phase transitions cannot be adequately controlled in the annealing process, and the results depend ultimately upon the phase the system is in as it approaches the ground state. All of the information gained prior to the phase transition is lost, and much of the compute time is wasted annealing in the wrong part of phase space.

The failure of simulated annealing to reproduce the exact MAP estimate may also be understood from the phase changes that occur during the annealing process. It is clear from these results that simulated annealing often fails to find the true MAP estimate. Again, it seems absurd to rely upon the metastable states for good restoration, when using a method specifically designed to avoid such local minima.

The TPM estimate suffers neither of these flaws, is consistently defined, and may be computed in a fraction of the Monte Carlo time. In conclusion, we believe that TPM should be the favoured estimate for image reconstruction problems. Although there may be other problems such as boundary detection where the MAP estimate *can* give better results [39], for image restoration the TPM estimate is always as good as and usually better than the MAP estimate, and does not demand the same level of compute resource as simulated annealing.

CHAPTER 5

Optimizing the Prior: The Thermodynamics of Hypothesis Evaluation

5.1 Introduction

In the work presented so far we have determined the optimal values of the restoration parameters, $\tilde{K}$ and $\tilde{h}$, in one of two ways. Given full knowledge of the true posterior and a matching prior model we could assign the optimal values, $\tilde{K} = K$ and $\tilde{h} = h$, exactly. When the prior was not well matched we could still determine the optimal values, in the sense that they maximize the quality factor, by a comparison of the restored distribution and the *known* source distribution. In the language of parameter estimation, we had access to **fully observed data.**

In this chapter we address the problem of parameter estimation from **partially observed data**. See the discussion after page 19 for an introduction to the literature. In the context of this thesis, 'partially observed' means that we have access to the data picture alone, with no knowledge of the source distribution when we choose the parameters $\tilde{K}$ and $\tilde{h}$. This is the situation one would encounter in a real restoration problem, where there is no explicit knowledge of the source picture or of the corruption process. In connection with hypothesis *evaluation* this does not preclude later access to the source distribution in order to measure the success of the parameter estimation scheme.

We investigate the use of the **evidence** [46, 81] as a criterion for choosing the restoration parameters. This method still requires that we choose a prior model against which we will evaluate the evidence. As we will see, a reasonable choice of prior is crucial to the success of the method as presented, but in the spirit of Bayes "a failure is an opportunity to learn" [81], and indicates a flaw in the chosen prior.

As presented in Chapter 2, the calculation requires the maximization of the evidence (2.53) or alternatively the log-evidence

$$\log P(\mathbf{D}|\tilde{K},\tilde{h}) = \log Z(\tilde{K},\tilde{h};\mathbf{D}) - \log Z_l(\tilde{h}) - \log Z_p(\tilde{K}). \qquad (5.1)$$

Therefore our task is to find a difference of free energies. As Neal [91] points out, the problem is essentially the same as the calculation of free energy differences in simulations of physical systems—and obtaining this information via Monte Carlo simulation is not straightforward. We can however discuss the following toy problem, which compares the evidence for different chequerboards. We will then go on to consider the complexities involved in the more general evidence calculations required for the edge-density prior we have considered so far.

5.2 An Exact Evidence Calculation

In our calculations we have always examined the edge-density prior introduced back in §2.4.2. We will see shortly the difficulties encountered in evidence measurements for such a prior, so in order to develop a feel for the general calculation, we begin with the simple idea of a chequerboard prior. Using a chequerboard prior means we guess that the source picture was drawn from a set of chequerboard pictures. The particular one chosen is labelled c, so the prior distribution is given by:

$$\tilde{P}(\mathbf{S}|c) \overset{\text{def}}{=} \delta\left(|\mathbf{S}-\mathbf{S}^c|\right), \tag{5.2}$$

where $\mathbf{S}^c$ is a chequerboard source with squares of side c.

Given this definition of the prior, the evidence is simply

$$\begin{aligned} P(\mathbf{D}|c,\tilde{h}) &= \sum_{\{\mathbf{S}\}} \tilde{P}(\mathbf{D}|\mathbf{S};\tilde{h})\tilde{P}(\mathbf{S}|c) \\ &= \frac{1}{Z_l(\tilde{h})}\exp\left\{\tilde{h}\sum_i D_i S_i^c\right\}, \end{aligned} \tag{5.3}$$

with $Z_l(\tilde{h})$ defined in (2.36).

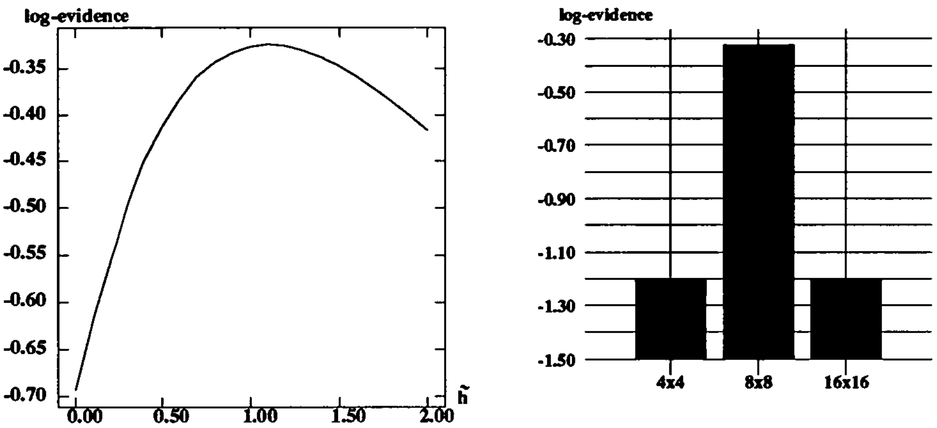
We can now calculate the evidence for a number of different chequerboards of side c, and as a function of the noise parameter $\tilde{h}$. If the source picture was indeed a chequerboard, and we choose the size c correctly so that $\mathbf{S}^c = \mathbf{S}^0$, then maximizing the evidence (5.3) with respect to the noise parameter $\tilde{h}$ correctly determines the noise level. Recall that $\langle \mathbf{D}.\mathbf{S}^0 \rangle = \tanh(h)$, so that if $\mathbf{S}^c = \mathbf{S}^0$ we get

$$\begin{aligned} \log P(\mathbf{D}|c,\tilde{h}) &= N\left[\tilde{h}\tanh(h) - \log\cosh(\tilde{h}) - \log 2\right], && (5.4)\\ \Rightarrow \frac{1}{N}\frac{\partial}{\partial \tilde{h}}\log P(\mathbf{D}|c,\tilde{h}) &= \tanh(h) - \tanh(\tilde{h}). && (5.5)\end{aligned}$$

where we have used self-averaging to write $\sum_i D_i S_i^c = N\tanh(h)$. Clearly the evidence is maximized at $\tilde{h} = h$ when we have chosen the prior correctly [see Figure 5.1(a)]. Note however that if $\mathbf{S}^c \neq \mathbf{S}^0$, then this procedure will *not* choose the optimal $\tilde{h} = h$.

We can use the evidence to determine what size of chequerboard the source was. Let us imagine that the source was an 8x8 chequerboard, i.e. $\mathbf{S}^0 = \mathbf{S}^8$. The log-evidence for the 8x8 chequerboard prior, $\log P(\mathbf{D}|\mathbf{S}^8, \tilde{h})$, is given by (5.4). The log-evidence for both the 4x4 and 16x16 chequerboard priors is reduced from this value by $N\tilde{h}\tanh(h)$, since in all cases the central limit theorem guarantees that $\sum_i D_i S_i^0 = 0$. Finding the size of the chequerboard is a rather trivial problem, but it demonstrates how the evidence procedure may be used to determine the parameterization of the source, provided the nature of the prior is chosen correctly. The evidence is maximized when the prior model $\mathbf{S}^c$ matches the source $\mathbf{S}^0$ that generated the data [see Figure 5.1(b)].

We are able to perform the previous calculation analytically but usually evidence calculations are computationally intensive; they are equivalent to the calculation of free energy differences. Free energy measurement by Monte Carlo simulation is a long standing problem. Before we move on to



(a) (b)

Figure 5.1. Analytic evidence results for a chequerboard prior. The source was an 8x8 chequerboard with 10% noise [corresponds to noise parameter $h = 1.1$]. (a) shows the log-evidence per site as a function of the noise parameter $\tilde{h}$, given that the chequerboard size c has been chosen correctly: it is maximized when $\tilde{h} = h$. (b) shows the log-evidence per site as a function of the chequerboard prior parameter c: it is maximized when the prior model matches the source $S^c = S^0$.

consider the evidence calculation itself, we spend some time investigating and developing a recent simulation method [77] for free energy measurement. There has been only limited work on such calculations and what follows is a lengthy digression on the issues and difficulties involved. The technique provides a powerful means of comparing different hypotheses, and merits detailed investigation.

5.3 Free Energy Measurement of the Ising Model

There has been renewed interest recently in the problem of free energy calculation, notably [10, 77] and see [17] for a review. The approach we will use here is the method of expanded ensembles [77]. The problem that one encounters when trying to calculate the free energy by Monte Carlo methods, is that there is no *microscopic* analogue of the free energy. In contrast, for example, to the **energy**, the free energy cannot be represented as a configurational average. When we want to calculate, say, the internal energy of the Ising model at equilibrium, we can measure the nearest neighbour correlation function for any microscopic configuration. Then, using the idea of importance sampling discussed in Chapter 3 to select configurations, we average this correlation function over many configurations and obtain an approximation of the equilibrium energy to arbitrary precision—greater accuracy simply requires more Monte Carlo time in order to generate better statistics. This approach is not available to us when we try to measure the free energy.

We embark on a measurement of the free energy for the two-dimensional Ising model for two reasons. First, the Ising model is the analogue of the edge-density prior on which we want to perform the evidence calculation. Second, we can verify the success of the simulation method by reference to the exact results for finite size systems [26]. The zero field Ising model partition function is

$$Z_I = \sum_{\{\mathbf{S}\}} \exp\left\{ K \sum_{<ij>} S_i S_j \right\}. \tag{5.6}$$

The free energy $F \stackrel{\text{def}}{=} -\log Z_I$ requires a comprehensive sum over *all* configurations **S**, with the value of the exponential fluctuating wildly between configurations. Since there is no straightforward function to be averaged, we cannot use the importance sampling trick to reduce the computational complexity.

An alternative way to see the distinction between the free energy calculation and the internal energy calculation is to note that all observables (things we may measure directly) are derivatives of the free energy and are averages of a function of the lattice variables. The free energy is *not* a simple function of the lattice variables. Indeed the converse view is useful—the free energy is an integral of an observable function, and it is this integral that we wish to calculate. As is always the case with any sort of numerical integration, the result we obtain will be a definite integral—the free energy difference. We may then determine the absolute value of the free energy at any desired point provided we know the absolute value at one point and use this as one of the limits of integration.

The method of expanded ensembles was originally introduced by Lyubartsev *et al.* [77] when they applied it to the free energy measurement of the restricted primitive model of electrolyte. The complexities of the method require a large number of parameters, the choice of which is crucial to the success of the measurement. They sketch the details of an iterative scheme that will improve the choice of parameters following several preliminary experiments, but they never indicate how one should make the initial choice of these parameters prior to the preliminary simulations. A significant portion of the following work addresses this open problem.

5.3.1 The Method of Expanded Ensembles

Following the method of Lyubartsev *et al.* [77] we introduce an expanded modified ensemble with partition function

$$Z_{\text{exp}} = \sum_{m=0}^{M} Z_m \exp(\eta_m), \tag{5.7}$$

where Z_m is the partition function of the Ising model for a fixed value of the coupling K_m, and η_m is a new parameter corresponding (in a fashion to be described later) to the particular value of K_m. Hence

$$Z_m = \sum_{\{S\}} \exp\left\{K_m \sum_{<ij>} S_i S_j\right\}, \tag{5.8}$$

and

$$Z_{\exp} = \sum_{m=0}^{M} \sum_{\{S\}} \exp\left\{K_m \sum_{<ij>} S_i S_j + \eta_m\right\}. \tag{5.9}$$

Now this expanded ensemble is amenable to Monte Carlo simulation by the Metropolis method discussed in Chapter 3. The distinction between this and conventional simulation is that calculation of the full partition function requires a sum over the M subensembles, each with a different value of coupling K_m. Therefore, we must allow the system to explore the space of couplings $\{K_m\}$ as well as the space of configurations $\{S\}$.

We use the standard Metropolis algorithm for the configurational updates as before. We use the same algorithm for updating the value of the coupling (i.e. shifting between subensembles) but this time the transition probability is chosen to be

$$P_{m\to k} = \min\left\{1, \exp\left[(K_k - K_m) \sum_{<ij>} S_i S_j + \eta_k - \eta_m\right]\right\}. \tag{5.10}$$

This choice of transition probability guarantees that detailed balance is satisfied and then we have just to ensure that the simulation is ergodic. Thus, the Metropolis algorithm allows us to mix configurational updates and coupling changes in any way we wish, provided we ensure ergodicity.

The probability p_m of finding the system in a particular state $(K_m, \mathbf{S})$ is given by

$$P(K_m, \mathbf{S}) = \frac{1}{Z_{\mathrm{exp}}} \exp\left\{ K_m \sum_{<ij>} S_i S_j + \eta_m \right\}. \tag{5.11}$$

Hence the probability of finding the system in the subensemble with coupling K_m is found by summing over configurations $\{\mathbf{S}\}$, yielding

$$p_m = \frac{Z_m \exp(\eta_m)}{Z_{\mathrm{exp}}}. \tag{5.12}$$

The ratio of these probabilities, which we can measure, then allows us to calculate the free energy difference between subensembles

$$\frac{p_m}{p_k} = \frac{Z_m \exp(\eta_m)}{Z_k \exp(\eta_k)}, \tag{5.13}$$

$$\Rightarrow \log Z_m - \log Z_k = \log\left(\frac{p_m}{p_k}\right) + \eta_k - \eta_m. \tag{5.14}$$

In the simulation, we measure the ratio p_m/p_k by comparing the length of MC time spent in each subensemble.

Of course, this only provides the *difference* in free energy between two ensembles. In order to obtain quantitative measurements of absolute free energy, it is necessary to 'connect' with an ensemble that is 'simple', (i.e. for which we *know* the value of the free energy). For the simple Ising model, we connect with the zero coupling system. This is a simple non-interacting system with a free energy of $-N \log 2$ which arises only from the entropy contribution.

This all seems straightforward enough. Operationally, however, there are many parameters to be chosen (the K_m and η_m for $m = 0 \ldots M$), and the success or failure of the measurement depends crucially on a suitable choice of these. A poor choice gives a simulation that is not ergodic, or one that moves through phase space so slowly that successive measurements are highly correlated. Lyubartsev *et al.* explain how to refine the selection of the $\{\eta_m\}$ given a suitable initial choice, but they do not indicate a reliable method for making this initial selection, nor do they disclose their prescription for choosing the set of couplings $\{K_m\}$.

We now consider the reason for introducing the parameters $\{\eta_m\}$. Without this modification to the ensemble, the MC steps that change the coupling according to (5.10) would always drive K upwards, and the expanded ensemble would equilibrate at the largest value of K—a low energy configuration. [We get this behaviour if we set $\eta_m = 0 \; \forall m$.] For equations (5.13) and (5.14) to hold we require that there be a finite probability of transition to any of the M subensembles, and in order that we generate reasonable statistics in the shortest MC time possible we require that the system spend a similar length of time in each of the subensembles. This last condition is satisfied if $p_m = M^{-1} \; \forall m$. This implies

$$\eta_k - \eta_m = \log Z_m - \log Z_k, \tag{5.15}$$

i.e. the parameters $\{\eta_m\}$, if chosen optimally, are just the free energies that we seek!

To summarize, our task is to calculate the free energies for each subensemble. We first guess the initial values of the $\{\eta_m\}$ which must be close to the correct values of the free energy (modulo a constant), and we then calculate a correction to these using Metropolis MC simulation of the expanded ensemble. It is essential that the $\{\eta_m\}$ be reasonable initial guesses, otherwise many of the transition probabilities between subensembles will be very small, and the statistics generated will be skewed by the slow evolution of the system through phase space.

5.3.2 Initial Choice of $\{\eta\}$

The transition probability between subensembles that we need to tune to avoid the pitfalls discussed above is given in (5.10). The prescription we use is to perform a number of standard spin-flip MC steps and then attempt a change in K using the transition probability (5.10). As long as both types of MC step are made regularly, the simulation should be ergodic and may explore all of the expanded ensemble of states. Therefore it is not especially crucial what the ratio of spin updates to K-change updates is. With this in mind we argue the following.

Let us imagine that we allow the system to equilibrate at the particular value of K_m it currently holds, before we attempt to update the value of K. We may then write, for a typical configuration

$$\sum_{<ij>} S_i S_j = -\mathcal{E}_m + \varepsilon, \tag{5.16}$$

where $\mathcal{E}_m$ is the equilibrium internal energy of the system (in units of the coupling K_m), and ε is a 'displacement' from this equilibrium energy, with the ε having a normal distribution. To a first approximation we neglect the displacement ε, since $\varepsilon/\mathcal{E}_m$ is of order $1/\sqrt{N}$. As discussed we require $p_m \sim p_k$ for good statistics, and this implies (from the detailed balance condition) that

$$P_{m\rightarrow k} \simeq P_{k\rightarrow m}. \tag{5.17}$$

Combining (5.10) and (5.16), this condition gives

$$\begin{aligned} -(K_k - K_m)\mathcal{E}_m + \eta_k - \eta_m &= -(K_m - K_k)\mathcal{E}_k + \eta_m - \eta_k, && (5.18) \\ \Rightarrow \eta_k - \eta_m &= \frac{1}{2}(K_k - K_m)(\mathcal{E}_k + \mathcal{E}_m). && (5.19) \end{aligned}$$

In this way we can assign the $\{\eta\}$ (relative to an arbitrary $\eta_0 = 0$) by calculating the equilibrium internal energy at each value of the coupling K. By assigning the $\{\eta\}$ in this way we are in fact simply performing a very crude numerical integration, using the trapezoidal rule (e.g. [99]). Lyubartsev *et al.* [77] do not suggest this as a method for choosing the $\{\eta\}$ but they do pick up on the idea when they discuss the difference between the internal energy and free energy. Choosing the $\{\eta\}$ as we have suggested shifts the energy distribution for each of the subensembles to approximately the same energy region (see Figure 3 in [77]) and therefore transitions between subensembles are equally likely to occur in *either* direction. [Without this shift, only transitions to subensembles of large K would be very probable, i.e. changes that reduce the energy.]

This choice of the $\{\eta\}$ does not give the exact free energies of course (else our task would be complete) and the choice that *will* give the optimal sampling distribution is in fact given by equation (5.15). This differs from our initial guesses in (5.19) by an entropy term which we neglected when we ignored the fluctuations of ε around the mean equilibrium energy in (5.16). However, the use of (5.19) provides a good initial guess to the optimal $\{\eta\}$, given a suitable set of $\{K\}$, and we are close enough to the optimal sampling distribution that we can calculate the correction to the free energy given by (5.14) fairly efficiently.

We are not yet guaranteed an efficient calculation however, since the number of subensembles M, and the particular choices of the coupling values $\{K\}$ are also extremely important. Given the prescription for choosing the $\{\eta\}$, consider the transition probabilities that result: putting (5.19) into (5.10) gives

$$P_{m \rightarrow k} = \exp \left[\frac{1}{2}(K_k - K_m)(\mathcal{E}_k - \mathcal{E}_m) \right]. \tag{5.20}$$

The first sanity check is that this *is* indeed a probability—the equilibrium energy varies inversely with the coupling K and so the argument of the exponential will always be negative. And clearly $P_{k \rightarrow m} = P_{m \rightarrow k}$ if we use (5.20). However, note that from the definition in (5.16) the equilibrium energy is an extensive quantity, and for a large system size this will lead to a large and negative argument in (5.20) and an accordingly small transition probability. The upshot of this will be a slow exploration of the expanded phase space and correspondingly poor generation of statistics. The solution is to choose the difference in coupling $\Delta K = K_m - K_k$ small enough that an appropriate transition probability results.

It is conventional in Monte Carlo simulations to aim for an acceptance ratio of around 50% in the belief that the optimal sampling distribution will be obtained when the transition probability is $\frac{1}{2}$, and we discuss this further in §5.3.4. We can arrange this but it requires a careful choice of ΔK, and evidently as the system size increases we will need to choose ΔK smaller to maintain an acceptably high transition rate.

5.3.3 Choosing the Couplings

Since we are considering a small difference in coupling we approximate the K dependence of the internal energy with a linear relationship

$$\mathcal{E}(K) \sim -\alpha K, \tag{5.21}$$

and we therefore write

$$\begin{aligned} \mathcal{E}_{k+1} - \mathcal{E}_k &\overset{\text{def}}{=} \Delta\mathcal{E}_k \\ &\sim -\alpha_k \Delta K_k. \end{aligned} \tag{5.22}$$

Substituting (5.22) into (5.20) gives

$$\begin{aligned} P_{k\to k+1} &\sim \exp\left(\frac{1}{2}\Delta K_k \Delta\mathcal{E}_k\right) \\ &= \exp\left(-\frac{1}{2}\alpha_k \Delta K_k^2\right), \end{aligned} \tag{5.23}$$

and if we require $P_{k\to k+1} \sim p$ (where p is the desired transition probability between adjacent couplings) we get

$$\Delta K_k \sim \left(\frac{2\log(1/p)}{\alpha_k}\right)^{\frac{1}{2}}, \tag{5.24}$$

and we need to calculate α_k at each point as

$$\alpha_k \sim -\frac{\Delta\mathcal{E}_k}{\Delta K_k}. \tag{5.25}$$

Operationally the procedure is as follows.

- Set initial values $K_0 = \eta_0 = \mathcal{E}_0 = 0, \alpha_0 = 1$.

- For $n = 1, 2 \ldots M$

1. Set
$$K_n = \left(\frac{2\log(1/p)}{\alpha_{n-1}}\right)^{\frac{1}{2}} + K_{n-1}. \tag{5.26}$$
2. Equilibrate the system at this coupling K_n, and then measure the average internal energy $\mathcal{E}_n$.
3. We can then choose
$$\eta_n = \frac{1}{2}(K_n - K_{n-1})(\mathcal{E}_n + \mathcal{E}_{n-1}) + \eta_{n-1}. \tag{5.27}$$
4. Finally we calculate α_n ready for the next iteration
$$\alpha_n = -\frac{\mathcal{E}_n - \mathcal{E}_{n-1}}{K_n - K_{n-1}}. \tag{5.28}$$

In practice this procedure is most effective in choosing the parameters $\{K\}$ and $\{\eta\}$. More importantly, the procedure is stable—if an error is made in choosing one ΔK_n, say too small, then $\Delta\mathcal{E}_n$ will be smaller, leaving α_n essentially unaffected and the error is not propagated to the next choice of K_{n+1}. The question that remains is whether we can do better than to choose $p = \frac{1}{2}$, and generate statistics more efficiently?

5.3.4 The Optimal Transition Probability

We mentioned earlier that it is conventional to aim for a transition probability of $p = \frac{1}{2}$ in order to generate statistics most efficiently. But why should we choose p this way? When making statistical measurements of observables, as we do in MC simulations, we make use of the central limit theorem (see e.g. [78]). This tells us that adding a very large number,

N, of basically independent random variables will result in a sum that has an essentially normal distribution with the width proportional to $\sqrt{N}$. Therefore the relative statistical error in the averages measured is of order $1/\sqrt{N}$, and we refine our calculation of the average by making the number of measurements N as large as possible. However, the requirement that the N measurements be essentially independent is quite crucial to the validity of the central limit theorem. Imagine transitions between just two states A and B with the transition probabilities equal, i.e. $P_{A\to B} = P_{B\to A}$. Then the detailed balance condition tells us that P_A/P_B, the ratio of the MC time spent in state A to the time spent in state B will converge to unity as the number of MC steps increases. But how quickly will this convergence occur?

If $P_{A\to B} = P_{B\to A} = \frac{1}{2}$, then after one step there is an equal probability of being in state A or state B, so consecutive measurements of the states are independent. However, if the transition probabilities are greater or less than $\frac{1}{2}$, then even after several steps there is an uneven probability of being in state A or state B (given that we know the initial state). Therefore, consecutive measurements cannot be considered independent. The larger the value of $|1-2P_{A\to B}|$, the greater the number of steps we must take between measurements before we can consider them to be essentially uncorrelated. So in this simple example it is correct to try to attain a transition rate of $p = \frac{1}{2}$.

Things are somewhat more complicated, however, in the expanded ensemble. Say that we want to measure the free energy difference of the Ising model between two values of the coupling K_A and K_B. We want to measure the ratio P_A/P_B, the relative proportion of MC time the system spends in the two subensembles A and B, and we want the number of uncorrelated measurements to be as large as possible. Ideally we would simply *choose* the transition rate $P_{A\to B}$ as close to $\frac{1}{2}$ as possible. But in a simple two state system $P_{A\to B}$ would be *defined* by (5.20), and if $|K_A - K_B|$ was at all large, this probability would be exceedingly small, as $\Delta\mathcal{E}$ is of order N. How then should we divide the interval $[K_A, K_B]$ into a number of subensembles so as to most quickly generate uncorrelated measurements? Clearly, even if the transition probability between adjacent K-states is $\frac{1}{2}$, we cannot consider each attempt to change the coupling as an independent measurement for the purposes of calculating P_A/P_B. We can only count measurements as essentially independent if the system has been able to evolve all the way from state A to state B in the intervening steps: it is the number of crossings that controls the statistics. Therefore we achieve the optimal sampling rate when we choose the $\{K\}$ so as to minimize the time taken for the system to traverse the coupling space between states A and B.

If the transition probability between adjacent K-states is p, then the time taken to make the transition is proportional to $\tau_{\text{single}} = 1/p$. Given that the system may move to an adjacent state either side of the original in time τ_{single} we have effectively a one-dimensional random walk. In the 1D random walk it takes n random steps for a particle to diffuse a distance of $\sqrt{n}$. Therefore the total time taken for the system to get from state A to state B is on average $\tau_{\text{full}} = \tau_{\text{single}} n^2$, where n is the number of subdivisions in $[K_A, K_B]$ and will depend upon the value of the transition probability p:

$$n \sim \frac{K_A - K_B}{\Delta K_k}, \tag{5.29}$$

and

$$\Delta K_k \sim \left(\frac{-2\log p}{\alpha_k}\right)^{\frac{1}{2}}. \tag{5.30}$$

For the moment we assume that α_k does not vary much in the range $[K_A, K_b]$, although this is not a strong constraint. Then

$$n^2 \propto \frac{1}{\log(1/p)}, \tag{5.31}$$

$$\Rightarrow \tau_{\text{full}} \propto \frac{1}{p\log(1/p)}. \tag{5.32}$$

To minimize τ_{full} we find the turning point in $(p\log p)^{-1}$, which occurs at $p = e^{-1}$. Therefore the desired ratio p_A/p_B should converge most quickly if we choose the transition probability $p < \frac{1}{2}$ in (5.28). To summarize the argument, it is better to reduce the number of subensembles n, even although this lowers the probability of making a transition between adjacent states, because the *overall* time taken to transit between the states that we wish to measure can be reduced.

5.3.5 Results

In Table 5.1 we present the average number of K-change steps taken to traverse from $K = 0$ to $K = 0.3$, for different values of the chosen probability parameter p. We see that in fact the *achieved* transition rate between adjacent states is always larger than the value p that was chosen, because we neglected the variations ε of the energies in (5.20). The energy values $\mathcal{E}_k$ are of order N, and the variations ε of order $\sqrt{N}$. In order to make the transition rates appreciable we have ensured that $\Delta\mathcal{E} = \mathcal{E}_k - \mathcal{E}_m$ is sufficiently small. However the variations ε remain of order $\sqrt{N}$ and so the value of $\Delta\mathcal{E}$ fluctuates widely around the average. These variations result in configurations S that are more typical of a different value of the coupling K and greatly increase the probability of transition to an adjacent K-state.

1	2	3	4	5	6
p	No (n) of Intervals	Traversal Time	Transition Rate (p_{ach})	n^2/p_{ach}	Traversal Time (HB)
0.0001	6	4112	0.031	1161	3859
0.001	7	3061	0.065	753	2443
0.01	9	2352	0.133	609	1744
0.05	11	2260	0.224	540	1571
0.1	13	2329	0.266	580	1638
0.17	15	2473	0.355	633	1605
0.2	16	2653	0.372	688	1567
0.3	18	2866	0.445	728	1633
0.4	21	3087	0.494	892	1628
0.5	24	3949	0.545	1056	1606

Table 5.1. Rate of exploration of K-space in the range $K = [0, 0.3]$, showing how the speed with which the space of couplings is explored varies with the selected acceptance probability p. (1) is the expected transition rate to adjacent K-states neglecting, as in (5.20), the energy fluctuations around the mean. (2) is the number of K-states that the algorithm divides the range into for this value of p. (3) is the number of attempted transitions required to traverse the entire range of K values. (4) is the achieved transition rate between adjacent K-states [note the discrepancy with (1)]. (5) is a predictive measure of the efficiency of the exploration—note the correspondence with the traversal time (3). (6) is the same measure as (3) but for the heat bath method discussed later in §5.3.6.

This makes the calculation of the optimal value for p far more complex—however, the general idea remains the same: an achieved transition rate of less than $\frac{1}{2}$ still proves optimal.

If we calculate the total statistical error based on the number of traversals of the range, we can predict the optimal p_{ach}. If the range is subdivided into n intervals, the expected number of traversals will be proportional to $p_{\mathrm{ach}}/(n^2)$, and hence the statistical error on the full measurement will be proportional to $n/(\sqrt{p_{\mathrm{ach}}})$. Minimizing this error yields the same optimal value for the transition rate, as shown in column (5) of Table 5.1.

We carried out a simulation of a 64x64 square Ising model to test the scheme. The results are presented in Table 5.2. They show the initial guess for the $\{\eta\}$, based on simple trapezoidal integration of the energy, and the value for the free energy after correction, alongside the exact analytic value for the free energy. The calculation of the exact free energy for a finite size Ising model is taken from Ferdinand and Fisher [26].

In conclusion, the method is very effective; however, for large systems the computational resource required for reasonably precise calculation is enormous. The refinement process achieved an order of magnitude reduction in the error over the initial choice of the $\{\eta\}$, but at the expense of an order of magnitude increase in the MC time required.

5.3.6 The Heat Bath Method

We can take advantage of the size of the variations ε of the energy in order to move through K-space more quickly if we use the heat bath algorithm

K	η	Free energy	Free energy	% error	%error
		(Simulation)	(Exact)	in η	in Sim.
0.066	-18.242	-17.780	-17.786	2.6e+00	3.9e-02
0.131	-71.692	-71.514	-71.494	2.8e-01	2.7e-02
0.194	-159.830	-159.924	-159.890	3.8e-02	2.1e-02
0.253	-277.110	-277.647	-277.626	1.9e-01	7.5e-03
0.307	-418.554	-418.818	-418.797	5.8e-02	5.0e-03
0.360	-597.269	-597.667	-597.621	5.9e-02	7.6e-03
0.404	-779.097	-779.672	-779.659	7.2e-02	1.7e-03
0.439	-960.848	-960.678	-960.639	2.2e-02	4.0e-03
0.472	-1170.336	-1169.920	-1169.939	3.4e-02	1.6e-03
0.522	-1524.066	-1523.810	-1523.879	1.2e-02	4.5e-03
0.595	-2077.409	-2076.680	-2076.848	2.7e-02	8.1e-03
0.708	-2977.731	-2977.070	-2977.346	1.3e-02	9.3e-03
0.938	-4847.249	-4846.560	-4846.956	6.0e-03	8.2e-03

Table 5.2. Results of the free energy measurement for the simple Ising model. The range of couplings $[0, 1]$ was divided into 65 intervals by the initialization algorithm: we present the results for every fifth value of K. The free energies are measured relative to an arbitrary zero at $K = 0$ (i.e. neglecting the usual $N \log 2$). The exact free-energies are calculated using the method in [26]. The initial guesses (η) at the free energies are quite accurate. The refined measurements of the free energies reduce the percentage error by around an order of magnitude.

to make transitions between subensembles. We no longer consider only transitions to *adjacent* states, with the transition rates governed by the Metropolis algorithm. The new state of the system, the subensemble that we move into, is chosen purely on the basis of the current configuration **S**, with no reference to the current value of the coupling K.

The heat bath algorithm (see e.g. [63]) is essentially the same as the Gibbs sampler we discussed in Chapter 3, except that we apply it here to changes in coupling K. We choose the transition probability

$$P_{k\rightarrow m} \overset{\text{def}}{=} \frac{\exp\left\{K_m \sum_{<ij>} S_i S_j + \eta_m\right\}}{\sum_n \exp\left\{K_n \sum_{<ij>} S_i S_j + \eta_n\right\}} = p_m. \tag{5.33}$$

The results are presented back in Table 5.1, column (6). We see that the heat bath method for changing the couplings always results in a faster exploration of the space of the couplings than the Metropolis method. Significantly, once the gap between couplings is small enough, there is no apparent penalty for increasing the number of K-states: the system traverses K-space at the same rate. There are costs, however, in the increased computation required to calculate the sum in the denominator of (5.33), and in the increased statistical spread resulting from the smaller proportion of the total MC time spent in each individual state. It is still necessary to choose the values of the $\{\eta_m\}$ carefully in the same way that we did for the Metropolis version.

Given the results presented in Table 5.1 we would strongly recommend the use of the heat bath algorithm over the Metropolis algorithm for the purposes of making transitions between subensembles.

5.4 The Evidence for a Hypothesis

We now return to the original motivation for this chapter, the calculation of the log-evidence in (5.1). We have just shown how to calculate $\log Z_p(\tilde{K})$ (or we can use the exact analytic calculation from [26]) so we simply need to calculate the value of $\log Z(\tilde{K}, \tilde{h}; \mathbf{D})$.

The method is precisely as we have described for the simple zero-field Ising model. We again attempt changes in the value of $\tilde{K}$ using the transition probability (5.10). The only distinction is that the traditional MC spin-flip updates are carried out using the energy function for the reconstruction distribution.

First a word about the notation. Although the point of the evidence calculation is to guide us in the assignment of the restoration parameters $(\tilde{K}, \tilde{h})$, we use these variables with a slightly different meaning in this chapter. The evidence, calculated at a particular value of $(\tilde{K}, \tilde{h})$, measures the likelihood that the particular data picture we are considering *could* have been generated using the nearest-neighbour function of the prior with coupling $\tilde{K}$, and with a noise level measured by $\tilde{h}$, regardless of the *actual* generation processes involved. Our knowledge of the source process has disappeared completely: now we truly appeal only to the data.

5.4.1 Method

Echoing the first part of this chapter we write the partition function of the expanded modified ensemble

$$Z_{\exp} = \sum_{m=0}^{M} \sum_{\{\mathbf{R}\}} \exp\left\{ \tilde{K}_m \sum_{<ij>} R_i R_j + \tilde{h} \sum_i R_i D_i + \eta_m \right\}, \tag{5.34}$$

and as before we carry out a Metropolis MC simulation of the expanded ensemble (for a fixed $\tilde{h}$) using conventional Metropolis updates within subensembles, each with partition function

$$Z_m = \sum_{\{\mathbf{R}\}} \exp\left\{ \tilde{K}_m \sum_{<ij>} R_i R_j + \tilde{h} \sum_i R_i D_i \right\}, \tag{5.35}$$

and then making changes in coupling space with transition probability

$$P_{m \to k} = \min\left\{ 1, \exp\left[(\tilde{K}_k - \tilde{K}_m) \sum_{<ij>} R_i R_j + \eta_k - \eta_m \right] \right\}. \tag{5.36}$$

Keeping $\tilde{h}$ fixed allows us to connect with the trivial limit of zero coupling, $\tilde{K} = 0$, which has free energy $-\log Z_l(\tilde{h}) = -2N \cosh(\tilde{h})$.

Note that we only alter the coupling $\tilde{K}$ and this means that there is no contribution from the data term to the change in energy. So the simulation is just as in the simple Ising case except that we use (5.35) instead of (5.8) to update the system. However, we could alternatively define an expanded ensemble where the field term $\tilde{h}$ was the varied parameter:

$$Z = \sum_{m=0}^{M} \sum_{\{\mathbf{R}\}} \exp\left\{ \tilde{K} \sum_{<ij>} R_i R_j + \tilde{h}_m \sum_i R_i D_i + \eta_m \right\}. \tag{5.37}$$

The spin flip dynamics would use (5.35) as before, but the subensembles would be at fixed coupling $\tilde{K}$, for different values of $\tilde{h}$, with the transition probability being

$$P_{m\rightarrow k} = \min\left\{1, \exp\left[(\tilde{h}_k - \tilde{h}_m)\sum_i R_i D_i + \eta_k - \eta_m\right]\right\}. \tag{5.38}$$

We would then connect with the zero field system which has free energy $-\log Z_p(\tilde{K})$ and is in the prior-like phase. Since we are most interested in measurements of the data-like phase, it makes more sense to connect to the zero coupling system, which is in the data-like phase on the relevant side of the phase transition.

Returning to the expanded ensemble (5.34) we are able to calculate free energy differences as in (5.14)

$$\log Z(\tilde{K}_m, \tilde{h}; \mathbf{D}) - \log Z(\tilde{K}_k, \tilde{h}; \mathbf{D}) = \log\frac{p_m}{p_k} + \eta_k - \eta_m. \tag{5.39}$$

In this way we can calculate the partition function of the restored distribution relative to the zero coupling case. But as $Z(\tilde{K} = 0, \tilde{h}; \mathbf{D})$ is independent of $\mathbf{D}$ and is just $Z_l(\tilde{h})$, the free energies we calculate already have $\log Z_l(\tilde{h})$ subtracted out. The log-evidence is therefore simply the difference between the free energy of the reconstruction system, and the free energy of the zero field Ising system at the same value of the coupling $\tilde{K}$.

So the log-evidence, calculated in this way is:

$$\log P(\mathbf{D}|\tilde{K}_k, \tilde{h}) = -\eta_k - \log\frac{p_0}{p_k} - \log Z_p(\tilde{K}_k). \tag{5.40}$$

The simulations of the zero field Ising model have taught us several things which we should now note:

- the calculation of the correction $\log(p_0/p_k)$ is very intensive computationally; and
- we have found a reliable method of choosing the $\{\eta\}$ well so that the correction term is very small.

Now it turns out that even on the scale of the *difference* in free energy between the restoration system and the zero coupling system, the correction is still small. So for some analyses it may not be efficient to expend the CPU time required to refine the result beyond the initial choice of the $\{\eta\}$.

Since we want to measure the *average* effectiveness of the evidence as a criterion for choosing the restoration parameters, we really want to calculate the quenched average of the log-evidence. Since this average requires that we calculate the log-evidence many times over for different data pictures, we used simple numerical integration of the internal energy to approximate the log-evidence. However, for a practical calculation from a single data picture, the most precise results are obtained using the expanded ensemble method described above.

5.4.2 The Well-Matched Prior

We first consider the case where we have used the correct functional form for the prior—i.e. the source is taken from an Ising distribution at a particular value of K, and the noise process is true Gaussian with a noise level q corresponding to the field parameter h. Frigessi and Piccioni [30] show that given the functional forms of the source and noise processes are known, it is possible by calculating first and second neighbour correlation functions of the data to determine exactly what parameters K and h were used to generate the data picture. The evidence procedure does not do these calculations explicitly, but we can conclude that the necessary information is contained in the statistics of the data picture alone.

Using the trapezoidal integration method to determine the free energy of the restored system as a function of $\tilde{K}$ and $\tilde{h}$, and then subtracting the free energy of the zero field Ising model, we obtain a plot of the log-evidence shown in Figure 5.2. This shows that the maximum of the evidence correctly indicates the optimal restoration parameters that will maximize the quality factor and the overlap of the TPM estimate with the source. [Compare with Figure 3.3.] For clarity, all of the following evidence graphs show $\log P(\mathbf{D}|\tilde{K},\tilde{h}) + N \log 2$ $[N = 64^2]$ with only positive contours plotted.

The plot also indicates a large negative evidence for the region where $\tilde{K}$ and $\tilde{h}$ are *both* large. We can be reassured that this makes sense by the following observations. For a source coupling much beyond the phase transition ($K > 0.44$) the source picture would have a large bias to one colour. However, the data picture (which has been generated with a smaller coupling) has approximately zero bias, and this could only be reconciled with a large source coupling if the noise level was large (and hence h small). Thus there is negative evidence for all large K but especially for large K and large h together.

Figure 5.3 shows further examples of the success of the evidence in finding the optimal parameters. When the prior is *well-matched* to the source, the evidence will find the optimal values of the restoration parameters that

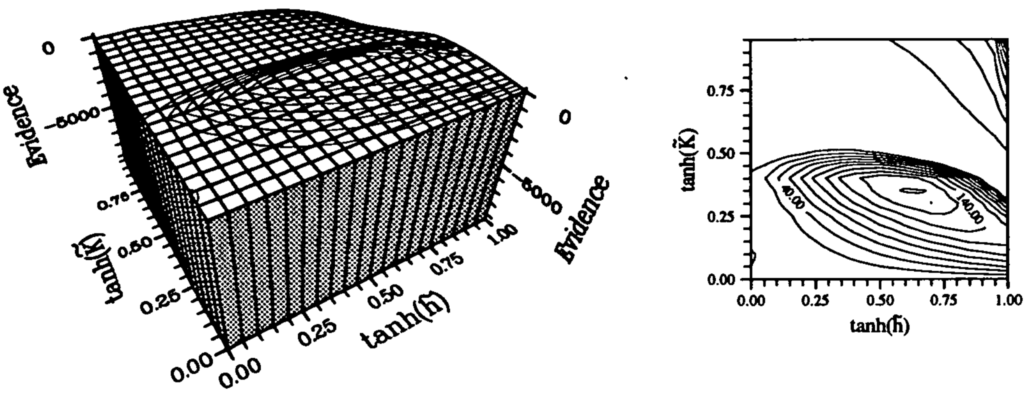


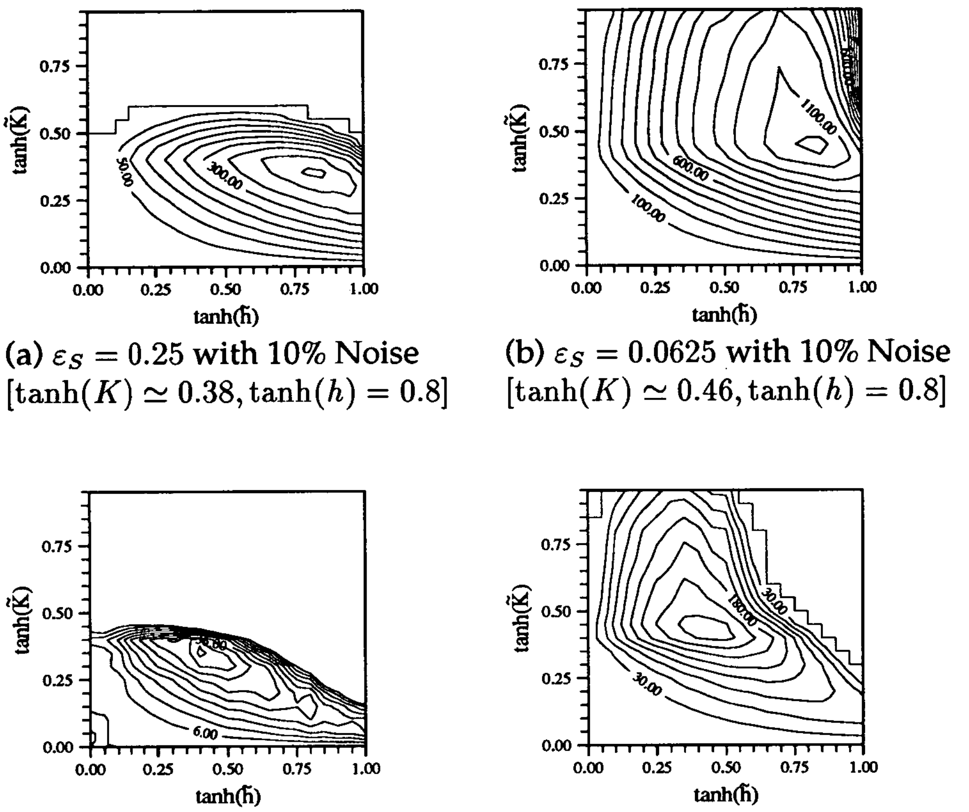
Figure 5.2. The results of the evidence measurement for an Ising source and 20% noise [$\tanh(K) \simeq 0.36, \tanh(h) = 0.6$], as in Figure 3.3. The evidence is very large and negative (-9000) for large $\tilde{K}$ and $\tilde{h}$. Therefore the 2D plot is more useful, showing only the contours around the maximum.

maximize the quality factor, and in this well-matched case these optimal parameters are $\tilde{K} = K$ and $\tilde{h} = h$.

Therefore we can conclude that the evidence procedure is successful at extracting the information on the K and h that generated D given only the data picture, *provided that the data was generated by the chosen prior.*

5.4.3 The Ill-matched Prior

We now investigate the success of the evidence procedure for guessing the optimal values for $\tilde{K}$ and $\tilde{h}$ that maximize the quality factor, when we have *not* chosen the correct functional form for the prior; again we realize this situation by modelling a fixed chequerboard source. These results are presented in Figure 5.4 for a range of chequerboard sizes and noise levels (compare with Figure 3.5 of the quality factor in Chapter 3).



(a) $\varepsilon_S = 0.25$ with 10% Noise $[\tanh(K) \simeq 0.38, \tanh(h) = 0.8]$

(b) $\varepsilon_S = 0.0625$ with 10% Noise $[\tanh(K) \simeq 0.46, \tanh(h) = 0.8]$

(c) $\varepsilon_S = 0.25$ with 30% Noise $[\tanh(K) \simeq 0.38, \tanh(h) = 0.4]$

(d) $\varepsilon_S = 0.0625$ with 30% Noise $[\tanh(K) \simeq 0.46, \tanh(h) = 0.4]$

Figure 5.3. Evidence results for a range of source parameters in the well-matched prior case. All examples show that the evidence correctly identifies the source parameters (and hence the optimal restoration parameters). For clarity, all of the evidence graphs show $\log P(\mathbf{D}|\tilde{K}, \tilde{h}) + N \log 2$ $[N = 64^2]$ with only positive contours plotted.

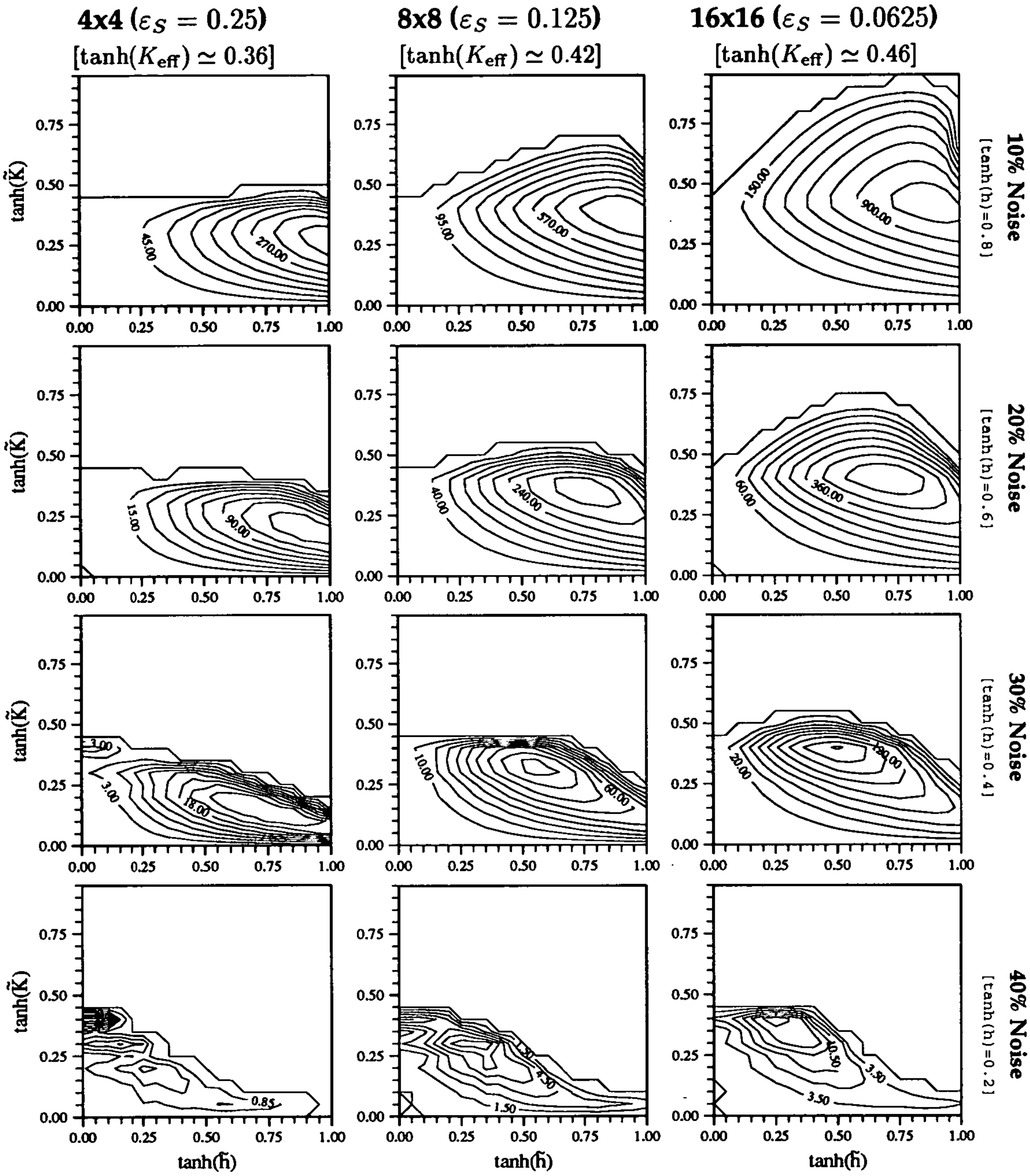


Figure 5.4. Evidence results for an ill-matched prior (chequerboard source). Moving across a row, the size of the chequers increases and the maximum of the evidence indicates a larger coupling $\tilde{K}$ should be used. Moving down a column, the noise level increases and the maximum indicates a small value of the field $\tilde{h}$ should be used. Compare these results with Figure 3.5. The trends in the restoration parameters are similar, but the maxima of the evidence and the quality factor do not coincide.

The qualitative behaviour of the maximum of the evidence does match that of the quality factor maximum. As the chequer size increases, the maximum of the evidence occurs at larger values of $\tilde{K}$. As the noise level increases the maximum occurs at smaller values of $\tilde{h}$. However these maxima do not coincide with the maxima of the quality factor shown in Figure 3.5. The figures show that the evidence is not a reliable method for determining the optimal restoration parameters when the original choice of the prior model is poor.

If there was any doubt in our minds, the mismatch between the maximum of the evidence and the maximum of the quality factor proves that the edge density prior is a particularly poor model for chequerboard source pictures. The quality factor results in Chapter 3 recommend the use of larger values of the restoration coupling $\tilde{K}$, since the source has large coherent regions (the chequers). However, the evidence is calculated based upon the assumption of an Ising source distribution. Since the bias in the data pictures is zero, the evidence for values of K much above the phase transition $K = 0.44$ becomes large and negative. Such a source coupling in a genuine Ising source process would generate pictures with an overwhelming bias to one colour. As we discussed in the previous section, since the bias in the data is zero (the source is a chequerboard), the evidence for a large value of K is large and negative. Of course, there *was* no large coupling used to generate the source—it is a fixed chequerboard and the evidence result is quite valid. It is simply the mismatch between the source and prior that undermines the result as a meaningful estimate of the optimal *restoration* parameters.

5.4.4 Model Comparison

The previous example demonstrated that the evidence procedure is not useful unless we have made an adequate choice of prior. But recall that we have already performed the evidence calculation for the chequerboard prior back in §5.2. If we compare the numerical value of the log-evidence for the chequerboard prior (5.4) [neglecting the constant $N \log 2$] with the results in the middle column of Figure 5.4, we find that the evidence is always greater for the chequerboard source than for the nearest neighbour prior. In particular, for the 8x8 chequerboard with 10% noise, the maximum of the evidence for the nearest neighbour prior is ~ 800, while the maximal evidence for the chequerboard source [using $N = 64^2$ in (5.4)] is ~ 1500. [Scale the results in Figure 5.1 by N and add $N \log 2$.] This is a clear indication that in this case, the chequerboard prior is more suited to the data than the edge-density prior.

This is not a rigorous comparison of the priors, but it does indicate the way in which one can use the evidence formalism to compare different forms of prior, as well as different parameter choices. If the results obtained from the evidence calculation are poor, this warns us that we have made a poor choice of prior and urges us to develop other priors against which we may test the data.

5.5 The Small Coupling Expansion of the Evidence

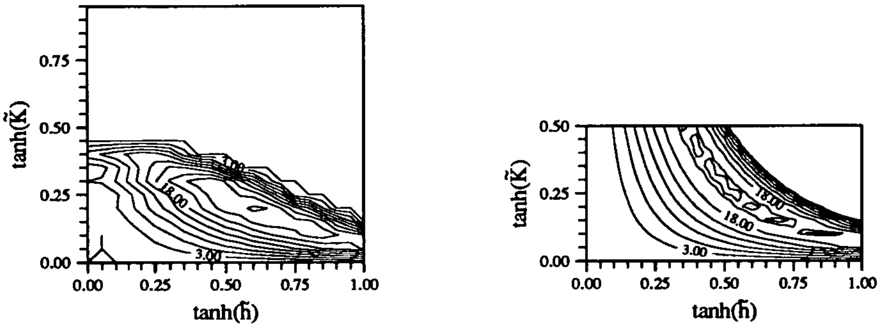
We set up the formalism for the small coupling expansions in Chapter 3. To conclude the work on the evidence we apply these analytic methods once again, and confirm the observations we have made in this chapter.

5.5.1 Method

The log-evidence is given in (5.1). Substituting in the small coupling results (3.71) and (3.79) gives

$$\begin{aligned}\log P(\mathbf{D}|\tilde{K},\tilde{h}) &= \log\Big\{1+\tilde{K}\tilde{\alpha}^2 A(\mathbf{D}) \\ &\quad +\frac{1}{2}\tilde{K}^2\left[2\tilde{\alpha}^2 B(\mathbf{D})+2\tilde{\alpha}^4 C(\mathbf{D})\right]+o(K+\tilde{K})^3\Big\} \\ &= \tilde{K}\tilde{\alpha}^2 A(\mathbf{D}) \\ &\quad +\frac{1}{2}\tilde{K}^2\left[-\frac{\nu N}{2}\tilde{\alpha}^4+2B(\mathbf{D})(\tilde{\alpha}^2-\tilde{\alpha}^4)\right] \\ &\quad +o(K+\tilde{K})^3. \qquad (5.41)\end{aligned}$$

Therefore, given a data picture we just need to make measurements of $A(\mathbf{D})$ and $B(\mathbf{D})$. For investigation purposes we find the quenched average of the log-evidence by calculating $A(\mathbf{S})$ and $B(\mathbf{S})$ and utilizing their self-averaging property. The corresponding results for the data are obtained by multiplying by $(1-2q)^2=\alpha^2$.



Simulation

Small Coupling

Figure 5.5. Small coupling evidence for the well-matched prior. The source coupling was $\tanh(K) = 0.2$ with 20% noise $[\tanh(h) = 0.6]$. Both the simulation result and the small coupling result indicate a maximum in the expected location $\tilde{K} = K$ and $\tilde{h} = h$.

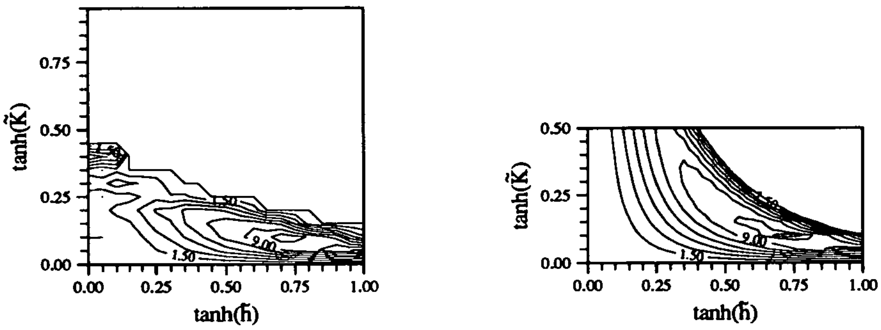
5.5.2 The Well-matched Prior

For the well-matched prior, the source is drawn from an Ising distribution and we calculate the values of $A(\mathbf{S})$ and $B(\mathbf{S})$ given in equations (3.104) and (3.105). Thus $\frac{1}{N}A(\mathbf{S}) = 2K\alpha^2 + o(K^3)$. Since $B(\mathbf{S})$ is of order K^2 we neglect the last term and obtain

$$\log P(\mathbf{D}|\tilde{K},\tilde{h}) = 2\tilde{K}\tilde{\alpha}^2 K\alpha^2 - \tilde{K}^2\tilde{\alpha}^4 + o(K+\tilde{K})^3. \qquad (5.42)$$

The gradient of (5.42) with respect to $\tilde{K}$ and $\tilde{h}$ is zero when $\tilde{K} = K$ and $\tilde{h} = h$ indicating a maximum at this point. This is confirmed in the comparison of simulation and small coupling results shown in Figure 5.5.

Notice the way the ridge of large evidence lies diagonally across the parameter space in Figure 5.5 and Figure 3.15. For small couplings it is difficult to determine whether the disorder evident in the data picture has arisen from little noise (large h) and a disordered source (small K), or from a larger coupling K with more noise (smaller h).



Simulation

Small Coupling

Figure 5.6. Small coupling evidence for an ill-matched prior (3x3 chequerboard). The small coupling results on the right indicate the same maximum of the evidence as simulation, but when compared with Figure 3.15 they clearly fail to find the optimal set of restoration parameters that maximize the quality factor.

5.5.3 The Ill-matched Prior

For the ill-matched prior, we again use a chequerboard source, and since this is a small coupling expansion we consider a 3x3 chequerboard, which has a suitably small effective coupling K_{eff} (see Table 3.1). The values for $A(\mathbf{S})$ and $B(\mathbf{S})$ and hence for $A(\mathbf{D})$ and $B(\mathbf{D})$ are obtained from Table 3.2. The results are shown in Figure 5.6. The success of the small coupling expansion for small chequerboards is proved once again, but the results confirm that the evidence is an inadequate basis for parameter estimation if the prior model is poorly chosen.

5.6 Conclusion

The method of expanded ensembles has been developed and refined in the context of free energy measurement of the Ising model. We have defined a prescription for setting the many parameters required, which was given only passing mention in the original paper [77]. In conclusion we have discovered that a modification of the method, to use the heat bath rather than Metropolis algorithm for making transitions between subensembles, provides a faster and more robust measurement process.

Such refinement does not prove necessary for our investigations of the evidence, which may be calculated approximately using numerical integration. The evidence approximation is successful at finding the correct values of the source and noise parameters provided the space of priors that we use contains the true source process. Otherwise we cannot say whether the evidence will provide any useful output—the further removed the true source image is from the prior distribution we use to calculate the evidence, the less reliable will be the result.

When the prior is ill-matched with the source, we are effectively finding the projection of one problem (the Ising source distribution) onto another (the fixed source chequerboard). We cannot be too surprised at failure in these circumstances. The evidence is calculated correctly, but the model that we evaluate the evidence against is flawed and does not represent the truth.

Given the opportunity to compare the evidence results with the quality factor results, the disparity acts as a warning that the prior model we have been using is inadequate. In fact, as we indicated in §5.4.4, it is possible to extend the evidence process a stage further and use it as a criterion for choosing different prior models (the functional form of the prior), as well as for parameter estimation [81].

CHAPTER 6

Conclusions

Image restoration, the recovery of an image that is in some way 'better' than the original noisy image, is a hard problem with many unsolved aspects. As we have seen, an enormous amount of research has been carried out across several scientific fields: notably in signal processing, and in applied statistics. The introduction of Markov random field models to this research has opened the way for a more theoretical treatment, making use of the similarity between such models and lattice models of magnetic systems in statistical mechanics.

The image restoration problem is distinct from image enhancement in that we build a prior model of the possible processes involved in the generation of the corrupted image, and use this to guide us when attempting the restoration. Thus we *infer* the source image from the data and our prior model. Bayesian statistics prescribes the tools required to make this inference in a consistent, logical manner, and we presented the Bayesian derivation of the restoration scheme in Chapter 2. By modelling the prior on the predicted density of edges in the source, information theory arguments delivered a prior probability function that was just a nearest-neighbour Markov random field.

With these basic arguments behind us, we were able to make use of analytic methods from statistical mechanics to investigate the performance of the restoration process. Our immediate interest was not the perfection of a practical restoration scheme, but the investigation of the factors that affect the performance of such a scheme, and the development of a better understanding of its successes and failures. The mean field approximation, presented in Chapter 3, explains the changing performance of the model in different regions of the space of restoration parameters. There is competition between the restoration parameters: the nearest-neighbour coupling in the prior that tries to smooth the image, and the field term that binds the restored image to the data. We wish the restored images to reflect the qualitative features of the data, but to be smoothed by the effect of the prior. However, beyond a critical value of the nearest-neighbour coupling we found a phase transition to a prior-like state: the prior wins over the data. This result indicates regions that should be avoided for the purposes of image restoration.

We considered the distinction between a prior that is well-matched to the source, and a prior that is poorly matched. Remember that for the purposes of measuring the success of the restoration scheme we have control over the parameters that generate the data as well as the restoration parameters that define the prior. Given the Bayesian arguments and the definition of our measure of quality, it is obvious that when the prior is well-matched to the source the optimal choice of restoration parameters (in the sense that they maximize the quality factor) is simply the values of the corresponding parameters used to generate the data. The small coupling expansion of the quality factor confirmed this. However, when we considered the ill-matched prior, we found that the optimal values of both parameters were modified from the values we might have naively assigned based simply on the edge-density of the source and the severity of the noise process. The most significant point to note is that even when we *correctly* model the noise process, a poor choice of prior will cause us to modify the optimal restoration parameter in the model likelihood. An incorrect choice of one aspect of the prior has implications for the choice of all other aspects of the model.

In Chapter 4 we attempted to lay to rest the debate over the optimal choice of estimate. Given the posterior probability distribution of restored images, what is the single image that best characterizes the distribution? With reference to the mean field phase diagram we argued that the MAP estimate obtained by simulated annealing is an absurd choice as many of the annealing trajectories cross the phase transition. This notwithstanding, the MAP estimate *can* provide results as good as the TPM estimate, but only with an order of magnitude increase in computational cost.

Finally, we returned to the thorny issue of parameter estimation. We wanted to find a prescription for choosing the 'best' restoration parameters in the sense that they maximize the quality factor, but with no prior knowledge of the actual generation parameters. The evidence formalism offers such a prescription for estimating the parameters from the data, but the measurement of the free energies involved is a non-trivial task. We extended the work on the 'method of expanded ensembles' to free energy measurement of the Ising model, and identified a number of improvements to the procedure. The method offers the possibility of direct comparison of arbitrary hypotheses; one has just to find a Monte Carlo path between the subensembles that represent the hypotheses.

Much as we discovered in the analysis of the quality factor, there is a distinct difference in the utility of the evidence, between the cases of well-matched and ill-matched priors. When the space of priors contains the source all is well and the maximum of the evidence coincides with the maximum of the quality factor. But when the prior is ill-matched the evidence maximum is in general in a different place to the maximum of the quality factor. The evidence chooses the 'optimal' parameters (in the sense that they maximize the evidence) by finding the parameters that were most likely to have generated the data *given the prior model.* The criteria are quite different when choosing parameters that will provide a restored picture most like the source when that source is nothing like the prior.

One avenue of research that merits further investigation would be to experiment explicitly with different priors, i.e. to introduce further couplings into the prior (and correspondingly the source). This would allow us to control a continuous variation between well-matched and ill-matched prior forms, rather than the disparate cases of Ising or chequerboard prior. This would also provide the scope for a more detailed analysis of the evidence procedure: can we, by this method, *quantitatively* evaluate different prior models in the light of the data?

In conclusion, this thesis has aimed to provide a better theoretical understanding of the issues surrounding the image restoration problem. In the words of Aristotle, "Those who wish to succeed must ask the right preliminary questions."

Bibliography

[1] K. Abend (1968). Compound Decision Procedures for Unknown Distributions and for Dependent States of Nature. In *Pattern Recognition*, ed. L. Kanal. Washington DC: Thompson Book Co., 207–249.

[2] J. A. Anderson (1968). A Memory Model Using Spatial Correlation Functions. *Kybernetic* **5**, 113–119.

[3] J. A. Anderson and E. Rosenfeld, eds. (1988). *Neurocomputing: Foundations of Research.* Cambridge: MIT Press.

[4] D. Amit, H. Gutfreund, and H. Sompolinsky (1985). Spin-Glass Models of Neural Networks. *Physical Review A* **32**, 1007–1018.

[5] D. Amit, H. Gutfreund, and H. Sompolinsky (1985). Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks. *Physical Review Letters* **55**, 1530–1533.

[6] D. Amit, H. Gutfreund, and H. Sompolinsky (1987). Statistical Mechanics of Neural Networks Near Saturation. *Annals of Physics* **173**, 30–67.

[7] D. Amit, H. Gutfreund, and H. Sompolinsky (1987). Information Storage in Neural Networks with Low Levels of Activity. *Physical Review A* **35**, 2293–2303.

[8] D. Amit (1989). *Modelling Brain Function.* Cambridge: Cambridge University Press.

[9] T. Bayes (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. *Phil. Trans. R. Soc. London* **53**, 370–418. Reprinted in *Biometrika* **45**(1958), 293–315.

[10] B.A. Berg and T. Neuhaus (1992). Multicanonical Ensemble—A New Approach to Simulate 1st-Order Phase Transitions. *Physical Review Letters* **68**, 9–12.

[11] J. Besag (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. Royal Statistical Society, B* **36**, 192–225.

[12] J.E. Besag and P.A.P. Moran (1975). On the Estimation and Testing of Spatial Interaction in Gaussian Lattice Processes. *Biometrika* **62**, 555–562.

[13] J. Besag (1975). Statistical Analysis of Non-lattice Data. *The Statistician* **24**, 179–195.

[14] J. Besag (1977). Efficiency of Pseudolikelihood Estimation for Simple Gaussian Fields. *Biometrika* **64**, 616–618.

[15] J. Besag (1987). Errors-in-variables Estimation for Gaussian Lattice Schemes. *J. Royal Statistical Society, B* **49**, 73–78.

[16] J. Besag (1986). On the Statistical Analysis of Dirty Pictures. *J. Royal Statistical Society, B* **48**, 259–302.

[17] D.L. Beveridge and F.M. Di Capua (1989). Free Energy Via Molecular Simulation—Applications to Chemical and Biomolecular Systems. *Annual Review of Biophysics and Biophysical Chemistry 1989*, **18**, 431–492.

[18] K. Binder and D. W. Heermann (1988). *Monte Carlo Simulation in Statistical Mechanics.* Berlin: Springer Verlag.

[19] R. Brout (1959). Statistical Mechanical Theory of a Random Ferromagnetic System. *Physical Review* **115**, 824–835.

[20] W.L. Buntine and A.S. Weigand (1991). Bayesian Back-propagation. *Complex Systems* **5**, 603–643.

[21] R. Chellappa and A. Jain eds. (1993). *Markov Random Fields: Theory and Application.* Boston, MA: London Academic.

[22] G.R. Cross and A.K. Jain (1983). Markov Random Field Texture Models. *IEEE Trans. Pattern Analysis and Machine Intelligence* **5**, 25–39.

[23] A.P. Dempster, N.M. Laird, and D.B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statistical Society, B* **39**, 1–38.

[24] H. Derin and H. Elliott (1987). Modelling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields. *IEEE Trans. Pattern Analysis and Machine Intelligence* **9**, 39–55.

[25] C. Domb and M.S. Green eds. (1974). *Phase Transitions and Critical Phenomena. Volume III: Series Expansions for Lattice Models.* London: Academic Press.

[26] A.E. Ferdinand and M.E. Fisher (1969). Bounded and Inhomogeneous Ising Models. I. Specific-Heat Anomaly of a Finite Lattice. *Physical Review* **185**, 832–846.

[27] L.R. Ford and D.R. Fulkerson (1962). *Flows in Networks.* Princeton: Princeton University Press.

[28] B.R. Frieden (1972). Restoring with Maximum Likelihood and Maximum Entropy. *J. Opt. Soc. Am.* **62**, 511–518.

[29] B.R. Frieden and J.J. Burke (1972). Restoring with Maximum Entropy, II: Superresolution of Photographs of Diffraction-Blurred Impulses. *J. Opt. Soc. Am.* **62**, 1202–1210.

[30] A. Frigessi and M. Piccioni (1989). Parameter Estimation for 2-D Ising Fields Corrupted by Noise. *Stochastic Processes and their Applications* **34**, 297–311.

[31] A. Frigessi, P. di Stefano, C-R. Hwang, and S-J. Sheu (1993). Convergence Rates of the Gibbs Sampler, the Metropolis Algorithm and Other Single-site Updating Dynamics. *J. Royal Statistical Society, B* **55**, 205–219.

[32] E. Gardner (1987). Maximum Storage Capacity in Neural Networks. *Europhysics Letters* **4**, 481–485.

[33] E. Gardner (1988). The Space of Interactions in Neural Network Models. *Journal of Physics A* **21**, 257–270.

[34] E. Gardner and B. Derrida (1989). Three Unfinished Works on the Optimal Storage Capacity of Networks. *Journal of Physics A* **22**, 1983–1994.

[35] D. Geiger and F. Girosi (1991). Parallel and Deterministic Algorithms from MRF's: Surface Reconstruction. *IEEE Trans. Pattern Analysis and Machine Intelligence* **13**, 401–412.

[36] S. Geman and D. Geman (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**, 721–741.

[37] S. Geman and C. Graffigne (1986). Markov Random Field Image Models and Their Applications to Computer Vision. In *Proceedings*

of the International Congress of Mathematicians, 1986 ed. A.M. Gleason. Providence: American Mathematical Society.

[38] D. Geman (1989). Random Fields and Inverse Problems in Imaging. *Ecole D'Ete de Probabilities de Saint-Flour XVII—1988* **1427**, 113–193. Springer Lecture Notes in Mathematics Vol. 1427.

[39] D. Geman, S. Geman, C. Graffigne, and P. Dong (1990). Boundary Detection by Constrained Optimization. *IEEE Trans. Pattern Analysis and Machine Intelligence* **12**, 609–628.

[40] B. Gidas (1989). A Renormalization Group Approach to Image Processing Problems. *IEEE Trans. Pattern Analysis and Machine Intelligence* **11**, 164–180.

[41] A.S. Glassner ed. (1989). *An Introduction to Ray Tracing*. London Academic.

[42] R.C. Gonzalez and P. Wintz (1977). *Digital Image Processing*. Reading MA: Addison Wesley Advanced Book Program.

[43] I.J. Good (1963). Maximum Entropy of Hypothesis Formulation, Especially for Multidimensional Contingency Tables. *Annals Math. Stat.* **34**, 911–934

[44] D.M. Greig, B.T. Porteous, and A.H. Seheult (1989). Exact Maximum A Posteriori Estimation for Binary Images. *J. Royal Statistical Society, B* **51**, 271–279.

[45] S.F. Gull and G.J. Daniell (1978). Image Reconstruction from Incomplete and Noisy Data. *Nature* **272**, 686–690.

[46] S.F. Gull (1989). Developments in Maximum Entropy Data Analysis. In *Maximum Entropy and Bayesian Methods, Cambridge 1988*, ed. J. Skilling. Dordrecht, London: Kluwer Academic Publishers.

[47] G. Györgyi (1990). Inference of a Rule by a Neural Network with Thermal Noise. *Physical Review Letters* **64**, 2957–2960.

[48] G. Györgyi and N. Tishby (1990). Statistical Theory of Learning a Rule. In *Neural Networks and Spin Glasses*, eds. W.K. Theumann and R. Koeberle. Singapore: World Scientific.

[49] A. Habibi (1972). Two-dimensional Bayesian Estimate of Images. *Proceedings IEEE* **60**, 878–883.

[50] J. Hadamard (1923). *Lectures on the Cauchy Problem in Linear Partial Differential Equations.* New Haven, CT: Yale University Press.

[51] K.M. Hanson (1987). Bayesian and Related Methods in Image Reconstruction from Incomplete Data. In [115], 79–125.

[52] J.M. Hammersley and D.C. Handscombe (1964). *Monte Carlo Methods.* London: Methuen.

[53] J.M. Hammersley and P. Clifford (1968). Markov Fields of Finite Graphs and Lattices. University of California-Berkeley, preprint.

[54] M. Hassner and J. Slansky (1980). The Use of Markov Random Fields as Models of Texture. *Computer Graphics and Image Processing* **12**, 357–370.

[55] D.O. Hebb (1949). *The Organization of Behavior.* New York: Wiley. Partially reprinted in Anderson and Rosenfeld [3].

[56] J. Hertz, A. Krogh, and R.G. Palmer (1991). *Introduction to the Theory of Neural Computation.* Redwood City: Addison-Wesley.

[57] G.E. Hinton and T.J. Sejnowski (1983). Optimal Perceptual Inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Washington 1983),* 448–453. New York: IEEE.

[58] J. J. Hopfield (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences, USA* **79**, 2554–2558. Reprinted in Anderson and Rosenfeld [3].

[59] K. Huang (1987). *Statistical Mechanics.* New York Chichester: Wiley.

[60] B.R. Hunt (1975). Digital Image Processing. *Proceedings IEEE* **63**, 693–708.

[61] B.R. Hunt (1977). Bayesian Methods in Nonlinear Digital Image Restoration. *IEEE Trans. Computers* **26**, 219–229.

[62] E. Ising (1925). *Zeitschrift Physik* **31**, 253.

[63] C. Itzykson and J-M. Drouffe (1989). *Statistical Field Theory.* New York: Cambridge University Press.

[64] E. Jen, ed. (1990). *Complex Systems 1989.* SFI Studies in the Sciences of Complexity, lecture volume 2. Redwood City: Addison-Wesley.

[65] F. James (1990). A Review of Pseudorandom Number Generators. *Computer Physics Communications* **60**, 329–344 and **69** 486.

[66] E.T. Jaynes (1957). Information Theory and Statistical Mechanics I. *Physical Review* **106**, 620–630.

[67] E.T. Jaynes (1984). Overview. In *Maximum Entropy and Bayesian Methods in Applied Statistics*, ed. J.H. Justice. Cambridge: Cambridge University Press.

[68] E.T. Jaynes (1984). Monkeys, Kangaroos and *N*. In *Maximum Entropy and Bayesian Methods in Applied Statistics*, ed. J.H. Justice. Cambridge: Cambridge University Press.

[69] F-C. Jeng and J.W. Woods (1991). Compound Gauss-Markov Random Fields for Image Estimation. *IEEE Trans. Signal Processing* **39**, 683–697.

[70] F-C. Jeng, J.W. Woods, and S. Rastogi (1993). Compound Gauss-Markov Random Fields for Parallel Image Processing. In [21] 11–38.

[71] I.T. Jolliffe (1986). *Principal Component Analysis*. New York: Springer-Verlag.

[72] R.L. Kashyap and R. Chellappa (1983). Estimation and Choice of Neighbours in Spatial-Interaction Models of Images. *IEEE Trans. Information Theory* **29**, 60–72.

[73] R. Kinderman and J.L. Snell (1980). *Markov Random Fields and Their Applications*. Providence, RI: American Mathematical Society.

[74] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi (1983). Optimization by Simulated Annealing. *Science* **220**, 671–680.

[75] S. Kullback (1959). *Information Theory and Statistics*. New York: Wiley.

[76] S. Lakshmanan and H. Derin (1989). Simultaneous Parameter Estimation and Segmentation of Gibbs Random Fields Using Simulated Annealing. *IEEE Trans. Pattern Analysis and Machine Intelligence* **11**, 799–813.

[77] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov, and P.N. Vorontsov-Velyaminov (1992). New Approach to Monte-Carlo Calculation of the Free Energy—Method of Expanded Ensembles. *Journal of Chemical Physics* **96**, 1776–1783.

[78] S-K. Ma (1985). *Statistical Mechanics*. Philadelphia: World Scientific.

[79] B.M. McCoy and T.T. Wu (1973). *The Two-dimensional Ising Model.* Cambridge MA: Harvard University Press.

[80] W. S. McCulloch and W. Pitts (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* **5**, 115–133. Reprinted in Anderson and Rosenfeld [3].

[81] D.J.C. MacKay (1992). Bayesian Interpolation. *Neural Computation* **4**, 415–447.

[82] D.J.C. MacKay (1992). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation* **4**, 448–472.

[83] D.J.C. MacKay (1992). Hyperparameters: optimize, or integrate out? University of Cambridge, preprint.

[84] J.L. Marroquin (1985). Optimal Bayesian Estimators for Image Segmentation and Surface Reconstruction. A.I. Lab. Memo 839, M.I.T.

[85] J.L. Marroquin, S. Mitter, and T. Poggio (1987). Probabilistic Solution of Ill-Posed Problems in Computational Vision. *J. American Statistical Association* **82**, 76–89.

[86] G. Marsaglia, A. Zaman, and W.W. Tsang (1990). Toward a Universal Random Number Generator. *Statistics and Probability Letters* **9**, 35–39.

[87] C. Mead (1989). *Analog VLSI and Neural Systems.* Reading: Addison Wesley.

[88] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953). Equation of State Calculations for Fast Computing Machines. *Journal of Chemical Physics* **21**, 1087–1092.

[89] M. Mézard, G. Parisi, and M.A. Virasoro (1987). *Spin Glass Theory and Beyond.* Singapore: World Scientific.

[90] N.E. Nahi and T. Assefi (1972). Bayesian Recursive Image Estimation. *IEEE Trans. Computers* **21**, 734–738.

[91] R.M. Neal (1992). Bayesian Training of Backpropagation Networks by the Hybrid Monte Carlo Method. University of Toronto, preprint CRG-TR-92-1.

[92] Numerical Algorithms Group (1987). *NAG FORTRAN Library Manual.* Oxford: Numerical Algorithms Group.

[93] L. Onsager (1944). Crystal Statistics I: A Two-Dimensional Model with an Order-Disorder Transition. *Physical Review* **65**, 117.

[94] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl (1990). On the ability of the Optimal Perceptron to Generalize. *Journal of Physics A* **23,** L581–L586.

[95] K. Ord (1975). Estimation Methods for Models of Spatial Interaction. *J. American Statistical Association* **70,** 120–126.

[96] P. Peretto (1984). Collective Properties of Neural Networks: A Statistical Physics Approach. *Biological Cybernetics* **50,** 51–62.

[97] D.K. Pickard (1977). Inference for Discrete Markov Fields: The Simplest Nontrivial Case. *J. American Statistical Association* **82,** 90–96.

[98] T. Poggio, V. Torre, and C. Koch (1985). Computational Vision and Regularization Theory. *Nature* **317,** 314–319.

[99] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling (1988). *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge, UK: Cambridge University Press.

[100] J.M. Pryce (1993). Ising Image Restoration Simulator: User Note. University of Edinburgh.

[101] W. Qian and D.M. Titterington (1991). Estimation of Parameters in Hidden Markov Models. *Phil. Trans. R. Soc. London A* **337,** 447–428.

[102] W.H. Richardson (1972). Bayesian-based Iterative Method of Image Restoration. *J. Optical Society of America* **62,** 55–59.

[103] B.D. Ripley (1986). Statistics, Images, and Pattern Recognition. *Canadian Journal of Statistics* **14,** 83–111.

[104] F. Rosenblatt (1962). *Principles of Neurodynamics.* New York: Spartan.

[105] Rosenfeld and Kac (1982). *Digital Picture Processing,* Second Edition. San Diego: London Academic Press.

[106] C.E. Shannon (1948). A Model Theory of Communication. *Bell Systems Technical Journal* **27,** 379, 623. Reprinted in *The Mathematical Theory of Communication,* C.E. Shannon and W. Weaver. Urbana: University of Illinois Press.

[107] D. Sherrington and S. Kirkpatrick (1975). Solvable Model of a Spin Glass. *Physical Review Letters* **35,** 1792–1796.

[108] J.E. Shore and R.W. Johnson (1980). Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Trans. Information Theory* **26,** 26–39 and **29,** 942–943.

[109] T. Simchony, R. Chellappa, and Z. Lichtenstein (1990): Relaxation Algorithms for MAP Estimation of Gray-Level Images with Multiplicative Noise. *IEEE Trans. Information Theory* **36**, 608–613.

[110] M.A. Sivilotti, M.A. Mahowald, and C.A. Mead (1987). Real-Time Visual Computations Using Analog CMOS Processing Arrays. In *Advanced Research in VLSI: Proceedings of the 1987 Stanford Conference*, ed. P. Losleben, 295–312. Cambridge: MIT Press. Reprinted in Anderson and Rosenfeld [3].

[111] J. Skilling (1989). Classic Maximum Entropy. In *Maximum Entropy and Bayesian Methods, Cambridge 1988*, ed. J. Skilling. Dordrecht, London: Kluwer Academic Publishers.

[112] J. Skilling, D.R.T. Robinson, and S.F. Gull (1991). Probabilistic Displays. In *Maximum Entropy and Bayesian Methods, Laramie 1990*, eds. W.T. Grandy, Jr. and L.H. Schick. Dordrecht, London: Kluwer Academic Publishers.

[113] A.F.M. Smith and G.O. Roberts (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *J. Royal Statistical Society, B* **55**, 3–23. Plus discussion, 53–102.

[114] H.E. Stanley (1971). *Introduction to Phase Transitions and Critical Phenomena.* New York: Oxford University Press.

[115] H. Stark, ed. (1987). *Image Recovery: Theory and Application.* Academic Press.

[116] D. Stein, ed. (1989). *Lectures in the Sciences of Complexity.* SFI Studies in the Sciences of Complexity, lecture volume 1. Redwood City: Addison-Wesley.

[117] P.H. Swain, S.B. Vardeman, and J.C. Tilton (1981). Contextual Classification of Multispectral Data. *Pattern Recognition* **13**, 429–441.

[118] R.H. Swendsen and J.S. Wang (1987). Nonuniversal Critical Dynamics in Monte Carlo Simulations. *Physical Review Letters* **58**, 86–88.

[119] A.N. Tikhonov and V.Y. Arsenin (1977). *Solutions of Ill Posed Problems.* Washington, DC: Winston & Sons.

[120] H.S. Seung, H. Sompolinsky, and N. Tishby (1992). The Statistical Mechanics of Learning From Examples. *Physical Review A* **45**, 6056–6091.

[121] H.J. Trussell (1980). The Relationship Between Image Restoration by the Maximum A Posteriori Method and a Maximum Entropy Method. *IEEE Trans. Accoustics, Speech and Signal Processing* **28**, 114–117.

[122] R. Turner, IBM Research Laboratories, Winchester. Personal Communication.

[123] G. Wahba (1977). Practical Approximate Solutions to Linear Operator Equations When the Data Are Noisy. *SIAM Journal of Numerical Analysis* **14**, 651–667.

[124] T.H.L. Watkin, A. Rau, and M. Biehl (1993). The Statistical Mechanics of Learning a Rule. *Reviews of Modern Physics* **65**, 499–556.

[125] P. Weiss (1907). *J. Phys. Radium, Paris* **6**, 667.

[126] D.H. Wolpert (1992). On the Use of Evidence in Neural Networks. In *Advances in Neural Information Processing Systems 5*, ed. Giles *et al.* San Mateo, CA: Morgan Kauffman.

[127] E. Wong (1968). Two-dimensional Random Fields and the Representation of Images. *SIAM Journal on Applied Mathematics* **16**, 756–770.

[128] J.W. Woods (1972). Two-dimensional Discrete Markovian Fields. *IEEE Trans. Information Theory* **18**, 232–240.

[129] Y. Zhauo, X. Zhuang, L. Atlas, and L. Anderson (1992). Parameter Estimation and Restoration of Noisy Images Using Gibbs Distributions in Hidden Markov Models. *CVGIP: Graphical Models and Image Processing* **54**, 187–197.