Inference and Experimental Design for Percolation and Random Graph Models

Andrei Iu. Bejan, PhD, MSc

Submitted for the degree of
Doctor of Philosophy
on completion of research in the
Department of Actuarial Mathematics and Statistics,
School of Mathematical and Computer Sciences,
Heriot-Watt University

June 2010

Abstract

The problem of optimal arrangement of nodes of a random weighted graph is studied in this thesis. The nodes of graphs under study are fixed, but their edges are random and established according to the so called *edge-probability function*. This function is assumed to depend on the weights attributed to the pairs of graph nodes (or distances between them) and a statistical parameter. It is the purpose of experimentation to make inference on the statistical parameter and thus to extract as much information about it as possible. We also distinguish between two different experimentation scenarios: *progressive* and *instructive* designs.

We adopt a utility-based Bayesian framework to tackle the optimal design problem for random graphs of this kind. Simulation based optimisation methods, mainly *Monte Carlo* and *Markov Chain Monte Carlo*, are used to obtain the solution. We study optimal design problem for the inference based on partial observations of random graphs by employing *data augmentation technique*.

We prove that the infinitely growing or diminishing node configurations asymptotically represent the worst node arrangements. We also obtain the exact solution to the optimal design problem for *proximity graphs* (*geometric graphs*) and numerical solution for graphs with *threshold* edge-probability functions.

We consider inference and optimal design problems for finite clusters from bond percolation on the integer lattice $\mathbb{Z}^d$ and derive a range of both numerical and analytical results for these graphs. We introduce *inner-outer* plots by deleting some of the lattice nodes and show that the 'mostly populated' designs are not necessarily optimal in the case of incomplete observations under both progressive and instructive design scenarios.

Finally, we formulate a problem of approximating finite point sets with lattice nodes and describe a solution to this problem.

To my grandparents.

Statement of Authorship

Some parts of this thesis have been published or submitted for publication in refereed journals, presented at conferences and used in teaching materials. Listed below is the information pertaining to these preliminary presentations of results.

1. Bejan, A. Iu. (2009) Inference and optimal design for percolation and random graph models. *Computer Laboratory Opera Group Seminars. The University of Cambridge.*

2. Bejan, A. Iu. (2009) Large clusters as rare events, their simulation and connection to critical percolation. *Networks (Operations Research) Seminar Series. The University of Cambridge.*

3. Bejan, A. Iu. (2008) Grid approximation of a finite set of points. Conference *Mathematics & IT: Research and Education 2008*, Chişinău, October 1-4.

4. Bejan, A. Iu., Gibson, G. J., Zachary, S. (2008) Inference and experimental design for some random graph models. Workshop *Designed Experiments: Recent Advances in Methods and Applications (DEMA2008)*, Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, 11-15 August 2008.

5. Bejan, A. (2008) Lecture notes "MCMC in modern applied mathematics". Center for Education and Research in Mathematics and Computer Science, Department of Mathematics and Computer Science, State University of Moldova. `http://www.cl.cam.ac.uk/~aib29/CECMI/MCMC/notes.pdf`

6. Cook, A. R., Gibson, G. J., Gilligan, C. A. (2008) Optimal observation times in experimental epidemic processes. *Biometrics*, **64**(3), pp. 860-868.

with Web Appendices at

http://www.biometrics.tibs.org/datasets/070104.pdf

Except where explicit reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis by which I have qualified for or been awarded another degree or diploma. No other person's work has been relied upon or used without due acknowledgement in the main text and bibliography of the thesis.

Acknowledgements

I was suggested to undertake this study by Professor Gavin Gibson in reply to my proposal for pursuing PhD research at Heriot–Watt University in 2005. Gavin's suggestion was to consider an abstract problem of identifying spatial locations of nodes of a random graph that make observation of the edge structure most informative about the statistical model underlying the formation of the graph and to look at the applications, particularly in plant epidemiology, where observations often tend to be a filtering of the above graph. Dr Stan Zachary, Heriot–Watt University, joined the supervising team with interests and expertise in probability theory and stochastic network analysis.

I would like to express my sincere gratitude in the first place to my supervisors for the enormous amount of time and support they have given to me. I would also like to thank my examiners, Professor Frank Ball, The University of Nottingham, and Dr George Streftaris, Heriot-Watt University for useful comments and critical observations which undoubtedly resulted in the improvement of the thesis.

I thank Professor Chris Gilligan, The University of Cambridge, for his hospitality in February 2007 and for permission to reproduce Figure 1.3 from Bailey, Otten and Gilligan (2000).

I thank Dr. Alex Cook (Heriot-Watt University, University of Cambridge, National University of Singapore) for permanent discussions on the topic and also for the plants he was so generous to give me. The plants are in good health and I can say that many people enjoy them!

My thanks and appreciation are also extended to the following people who have supported me in undertaking this research programme: Dr. Arkadii Semenţul, State University of Moldova, and Professor Gheorghe Mişcoi, Academy of Sciences

of Moldova.

These people made my social life in Scotland enjoyable: John Phillips, Michael Reidman, Jafar Fazilov, Wenny Chen, Eyad Al'Okke, Cornelius Schmidt-Colinet and Ben Hart. Edinburgh is a truly great city and can hardly be compared to any other city in the world! I thank all its tourists, especially those who leave the city at the end of August each year, letting it get back to normal life!

Perhaps one should agree with William Somerset Maugham, who said that *"Money is like a sixth sense without which you cannot make a complete use of the other five"*. PhD students need this sixth sense indeed to fully concentrate on their studies. The Overseas Research Student Awards Scheme (ORSAS) and James Watt scholarship provided me with financial help and I acknowledge this support. The British Government should be thanked for running the former, whereas the School of Mathematical and Computer Sciences of Heriot–Watt University should be thanked for providing me with the latter.

I am thankful to the organisers of the workshop *Design of Experiments 2008* and to the Isaac Newton Institute for Mathematical Sciences for their hospitality while attending the event. I am also thankful to EURANDOM (European Institute for Statistics, Probability, Stochastic Operations Research and its Applications) for organising the series of workshops *Young European Probabilists*, two of which I had the chance to attend.

Finally, I am deeply indebted to my family. I owe my persistence to my grandparents, Ivan Nikolaevich Bejan and Ol'ga Leont'evna Bodnar'. My parents, Liubov' and Yurii, and my brother, Serguei, should be thanked for their encouraging understanding and support. My wife, Kitty, deserves thanks for tolerating the combination of almost incompatible things—scientific research and family life.

Andrei Bejan

Cambridge, May 2010

This page is so the Research Thesis Submission Form can have the page number before the Contents page number.

Contents

Abstract **i**

Authorship **iii**

Acknowledgements **v**

1 Introduction **1**

1.1 Why inference and optimal design on random graphs? 1

1.2 General model description and further motivation 4

1.2.1 Model 4

1.2.2 Motivation: theoretical positions and practical aspects . . . 6

1.3 Related work on inference and experimental design problems for stochastic interaction and spatial response models 11

1.3.1 Spatial response models 12

1.3.2 Stochastic interaction models 13

1.4 Outline of the thesis 14

2 Tools and methodology **16**

2.1 Basic notions from the graph theory 16

2.2 Likelihood and Bayesian statistical inference 22

2.2.1 Data, likelihood and Fisher information 23

2.2.2 Bayesian concept 27

2.3 Monte Carlo methods and Markov Chain Monte Carlo 31

2.3.1 Monte Carlo methods 31

2.3.2 Markov Chain Monte Carlo 33

3 Utility-Based Optimal Designs within the Bayesian Framework 41

3.1 Introduction: from locally D-optimum to utility-based Bayesian designs . . . 41

3.1.1 Toy examples: three and four nodes . . . 41

3.1.2 Utility-based Bayesian optimal designs . . . 47

3.2 Shannon entropy, Lindley information measure and Kullback–Leibler divergence . . . 49

3.2.1 Bits of history . . . 49

3.2.2 Lindley information . . . 51

3.2.3 Comparing informativeness of experiments: expected Kullback–Leibler divergence and expected Lindley information gain as expected utility and their properties . . . 53

3.3 Progressive and Instructive Designs . . . 60

3.3.1 Progressive designs . . . 61

3.3.2 Instructive designs . . . 62

3.4 Simulation-based evaluation of the expected utility . . . 62

3.5 Second formulation of the problem . . . 65

3.5.1 The model . . . 65

3.5.2 n-node optimal design problem for random graphs . . . 66

3.5.3 Examples . . . 67

4 Optimal Designs for Basic Random Graph Models 70

4.1 Worst case scenarios: indefinitely growing or diminishing vertex configurations . . . 70

4.2 Optimal designs for basic random graphs . . . 73

4.2.1 Two-node design and prior entropy asymptote of the expected utility . . . 73

4.2.2 Progressive and instructive designs: two-node 'black box' design example . . . 76

4.2.3 Three-node star design with two independent edges . . . 80

4.2.4 Proximity graphs . . . 83

4.2.5 Step-like (threshold) probability decay . . . 88

4.2.6 Non-preservation of optimal designs under replication 91

5 Lattice-based Optimal Designs **95**

5.1 Inference and Optimal Design for Percolation Models 96

5.1.1 Nearest-neighbour interaction model and percolation 96

5.1.2 Parameter estimation 101

5.1.3 Bayesian optimal designs and inner-outer plots 120

5.1.4 Implementation of progressive and instructive designs based on inner-outer plots 129

5.2 Lattice designs for inference on random graphs with long-range connections 133

5.2.1 Generalising results from the previous section 134

5.2.2 Square lattice and its deformations 135

6 Grid Approximation of a Finite Set of Points **142**

6.1 Formulation of the problem 142

6.1.1 Basic examples 142

6.1.2 Formulation of the problem and motivation 145

6.2 Finding ϵ-optimal approximation grids 147

6.2.1 Brucker–Meyer approximation in $\mathbb{R}$ and $\mathbb{R}^n$ 147

6.2.2 Approximation by grid nodes 154

6.2.3 Applications 160

7 Conclusions **161**

7.1 Summary 161

7.2 Contributions of the thesis 164

7.3 Directions for future work 165

Bibliography **170**

A Solving $a^{bx+c} = dx + e$ and maximising $x^2/(e^x - 1)$ **185**

A.1 Equation $a^{bx+c} = dx + e$ 185

A.2 Maximisation of $x^2/\left(e^{\theta x} - 1\right)$ 186

B Dirac delta function **187**

C Integration of polylogarithms **189**

D Realisation of 6 distances in $\mathbb{R}^3$ **192**

E Gamma distribution, infectious times and site percolation **194**

List of Figures

1.1 Examples of different types of regular discrete graph topologies. . . 3

1.2 Arrangement of n objects within the set $\mathcal{D}$: there is a link between each pair (u, v) of them with probability $p(r(u, v), \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. In **(a)** $\mathcal{D}$ is a bounded region in $\mathbb{R}^2$, in **(b)** $\mathcal{D} \subseteq \mathbb{Z}^2$, and in **(c)** $\mathcal{D}$ is a subset of nodes of a hexagonal grid—only neighbouring nodes can be connected realising the so called nearest-neighbour interaction. . 5

1.3 The growth of the mycelial colonies as a percolation process studied by Bailey and Gilligan (1997) and Bailey et al (2000). The edge-probability decay may be 'combined' from simpler decays: e.g. the progress of disease in a population of radish plants exposed to primary infection by *R. solani* in the presense/absence of *T. viride* was studied in Bailey and Gilligan (1997) using the following form for the probability of infection: $p(r, \theta) = (\theta_1 + \theta_2 r)e^{-\theta_3 r}$. 9

2.1 Oriented (left) and unoriented (right) multigraph on the same set of vertices. 17

2.2 A subgraph induced by the vertices of the graph from left of degrees distinct from 4 is a cycle from right. 19

2.3 Example of a graph G and its complement $\bar{G}$. 20

3.1 A graph on three nodes with edges of weights r_1, r_2, r_3. 42

3.2 Observation times diagram: solid line is the time axis, and the dotted lines are possible edges of the graph. 43

3.3 Left: A random graph on four nodes with edges of weights $r_1, r_2, r_3, r_4, r_5, r_6$. Right: The optimal random graph on four nodes in plane is a square. 45

3.4 Example of a geometric (proximity) graph on eight nodes with a threshold parameter $\theta \in \Theta = \mathbb{R}_+$. 68
3.5 Five of many more possible arrangements of 17 vertices on the two-dimensional integer grid $\mathbb{Z}^2$. 69

4.1 Function $W_\alpha(d)$ defined in (4.5) when $\alpha = 1$. This function attains its maximum at the point $d \approx 2.52$. 74
4.2 Two-node random multigraph (with no loops) in a black box: n multiple edges connect two sites u and v, each being open with probability p independently of the status of any other edge; it can be only observed whether the nodes are connected or not but not the total number of open edges. 76
4.3 Expected utility (expected KL divergence) of the experimenter A holding a beta prior for p, Beta(α, α), minus the entropy of the prior distribution. 78
4.4 (a) Expected utility (expected KL divergence) of the experimenter A holding a beta prior for p, Beta(α, β) (various sets of values for α and β), minus the entropy of the prior distribution; (b) Expected utility plots for the prior distributions considered in the left plot. . . 78
4.5 Optimal values of n, n^*, derived by maximising the expected KL divergence calculated by B (who knows the exact value of the parameter p, p^*) for the experimenter A holding a uniform prior for p. 79
4.6 Expected utility minus prior entropy surface for the Cauchy edge-probability function (see § 1.2.2); here θ is assumed to take values 1, 2, and 5 with probabilities 0.1, 0.5, and 0.4, respectively. Note that $-\mathrm{Ent}\{\pi(\theta)\} = 0.1 \log 0.1 + 0.5 \log 0.5 + 0.4 \log 0.4 \approx -0.94$, and this is in agreement with the plot (which in turn reflects the statement of Corollary 4.1.2). Horizontal axes correspond to the lengths of the edges, d_1 and d_2. 80

4.7 Expected utility minus prior entropy plots for two independent random edges and **(a)** KL divergence and power-law decay (exponential and Cauchy decays give similar unimodal surfaces); **(b)** negative squared error loss utility and logistic decay; **(c)** KL divergence and logistic decay, **(d)** KL divergence and a 'linear' decay $p(r,\theta) = (1-\theta r)\mathbb{1}_{\{r\leq\theta^{-1}\}}$ with a discrete distribution for θ over a finite set of points; **(e)** KL divergence and $p(r,\theta) = 1-\left(1+e^{(10-r)/\theta}\right)^{-1}$ with a discrete distribution for θ over a finite set; **(f)** KL divergence and $p(r,\theta) = 1-\left(1+e^{(10\theta-r)/0.3}\right)^{-1}$ with the same prior for θ as in (e). 82

4.8 Solution to the optimal design problem for proximity graph with and without metric constrains (six edges, see Example 4.2.3). 88

4.9 Optimal designs as functions of α in the model with threshold edge-probability function: (a) one edge and (b) two independent edges. . 90

5.1 Open clusters emerged as a result of bond percolation on $\mathbb{L}^2$ for different values of p: (a) $p = 0.2$, (b) $p = 0.4$, (c) $p = 0.5$, (d) $p = 0.6$, (e) $p = 0.75$, and (f) $p = 0.9$. The origin of $\mathbb{Z}^2$ is denoted by a circle in the centre of each plot. 100

5.2 An open cluster (black solid dots) containing the origin (a black dot in a circle) as a result of percolation simulation on $\mathbb{L}^2$. Here the bond percolation probability p was taken to be 0.478; the solid bonds represent open bonds. The open cluster can be seen as a finite outbreak of an epidemic with constant infectious periods and infection intensity spread rate $\lambda \approx 2.6$ evolving on $\Pi = \mathbb{Z}^2$ (since $0.478 = 1 - e^{-2.6/4}$). The dotted lines depict directions along which infection did not spread (from black to grey dots); thus, grey dots depict individuals which remain healthy and the dotted lines represent those bonds that must be absent given knowledge of the cluster set. 103

5.3 Solid line corresponds to the likelihood function evaluated for the complete information (both the site and edge configurations are known) on the cluster $\mathcal{C}$ from Figure 5.2. The histogram is based on a sample drawn from the MCMC applied to the site configuration $\mathcal{C}$ (nodes only). 106

5.4 Trace plot for MCMC sampling resulted in the histogram from Figure 5.3 for the cluster $\mathcal{C}$ from Figure 5.2. The trace plot indicates that the mixing properties of the chain are rather satisfactory. It took 28 seconds on Intel(R) Core(TM)2 Duo CPU 2.26GHz to obtain a series of chain updates of the length 10^4. This time could be further reduced by using dynamic graph update algorithms, see the footnote on the p. 129. 108

5.5 Inference on the percolation parameter using MCMC described in Algorithm 2: histograms of obtained samples and trace plots for (a,b) $n = 10$; (c,d) $n = 35$; (e,f) $n = 50$; (g,h) $n = 70$. 113

5.6 Likelihood functions $\mathcal{L}_n(p)$ ($n = 25, 50, 70$) obtained using the MCMC from Algorithm 2 and MCMC sample histogram of $\mathcal{L}_n(p)$ for $n = 70$. 117

5.7 Example of an inner-outer (m, r)-plot in $\mathbb{L}^2$: here $m = 9$ and $r = 3$. The plot is bounded by an $N \times N$ square, where N, according to (5.13), equals 21. 123

5.8 Left: An open cluster $\mathcal{C}$ simulated on the inner-outer plot $\Pi^{(2)}(9, 2)$ in $\mathbb{L}^2$ with $p = 0.52$; the central node (an initially inoculated site) is denoted by a circle. Right: The fully saturated graph derived from $\mathcal{C}$ with respect to the vertex set $\Pi^{(2)}(9, 2)$ and nearest-neighbour interaction. 124

5.9 (a) An open cluster $\mathcal{C}$ simulated on the inner-outer plot $\Pi^{(2)}(23, 4)$ in $\mathbb{L}^2$ with $p = 0.86$; the central node (an initially inoculated site) is denoted by a circle. (b) The fully saturated graph derived from $\mathcal{C}$ with respect to the vertex set $\Pi^{(2)}(23, 4)$ and nearest-neighbour interaction. 125

5.10 Inference on the percolation parameter for the configuration from the left plot in Figure 5.8. Left: Sample histogram obtained by running MCMC for this configuration. Right: MCMC trace plot of updates for p. The value of p for which the configuration in Figure 5.8 was obtained is 0.52. 127

5.11 Left: open cluster from Figure 5.9(a) obtained on the inner-outer $(23, 4)$-plot using $p = 0.86$. Right: MCMC sample histogram for p assuming the uniform prior $\mathrm{U}(0, 1)$ for this parameter. 127

5.12 Left: simulated open cluster obtained on the inner-outer $(13, 2)$-plot using $p = 0.9$. Right: MCMC sample histogram for p assuming the uniform prior $\mathrm{U}(0, 1)$ for this parameter. 128

5.13 Left: MCMC trace plot of updates in p for the site configuration from Figure 5.12. Right: part of the burn-in period of the MCMC trace plot of updates in p for the site configuration from Figure 5.11; this part of the update trace was not used in producing the histogram in Figure 5.11. 128

5.14 *Inner-outer* design plots A, B, and C form the design space $\mathcal{D} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. 131

5.15 Left: sample histogram for the marginal of $h(d, p, y)$ in d, $d \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$, under progressive design and $\pi(p) \sim \mathrm{U}(0, 1)$. Right: evaluated expected utility under instructive design with $\pi^*(p) \equiv \delta(p - 0.9)$ and 95% credibility intervals ($M = 1500$) for the plots A, B, and C, under instructive design. 132

5.16 Updating connected component: graphical representation of Metropolis-Hastings step of Algorithm 2 for long-range interaction locally finite graph models. 134

5.17 Modification of the planar square lattice. The modification parameters are as follows: d_x, the spacing between nodes in the horizontal direction; d_y, the spacing in the vertical direction; and δ_x, a displacement of every second row in the horizontal direction. All nodes of every second row are shifted to the right if $\delta_x > 0$, and to the left if $\delta_x < 0$. 136

5.18 Examples of modified planar square lattices: (a) unchanged square lattice ($d_x = d_y$, $\delta_x = 0$); (b) hexagonal lattice ($d_y = \sqrt{3}d_x/2$); (c) square lattice ($d_y = \delta_x = d_x/2$). The number of nodes is the same in all three plots. 137

5.19 Left: long-range connections with exponential decay $p = e^{-\theta d}$ ($\theta = 1.9$) on a triangular 13×13 lattice plot ($d_x = d_y = 1$) with displacement $\delta_x = 1/2$. Right: the connected component of the graph from the left panel which contains the central node (in circle). 138

5.20 Spline approximation of the expected utility (minus entropy of the prior distribution) for the long-range percolation model with exponential edge-probability function and the following 5×5 lattices: (a) triangular ($d_y = d_x$, $\delta_x = d_x/2$), (c) square ($d_y = d_x$, $\delta_x = 0$), and (e) hexagonal ($d_y = \sqrt{3}d_x/2$). The plots (b), (d), (f) in the right panel depict the first derivative of the corresponding approximation spline. The edge profile decay is of the form $p(d) = e^{-\theta d}$ and the prior distribution for θ was taken to be Gamma$(10, 0.2)$. 140

5.21 Spline approximation of the expected utility (minus entropy of the prior distribution) for the long-range percolation model with Cauchy edge-probability function and the following 5×5 lattices: (a) triangular ($d_y = d_x$, $\delta_x = d_x/2$), (c) square ($d_y = d_x$, $\delta_x = 0$), and (e) hexagonal ($d_y = \sqrt{3}d_x/2$). The plots (b), (d), (f) in the right panel depict the first derivative of the corresponding approximation spline. The edge profile decay is of the form $p(d) = (1 + \theta d^2)^{-1}$ and the prior distribution for θ was taken to be Gamma$(10, 0.2)$. 141

6.1 Typical dependence of minimal $d_{\max}$ on the spacing of the grid for a set X from $\mathbb{R}$ containing 3 or 4 points. In this particular example $X = \{11.8998, 34.0386, 49.8364, 95.9744\}$. Notice, that what is shown is a single graph of such a dependence; this graph exhibits discontinuities at many values of the grid spacing. 143

6.2 Initial configuration X of six points in plane and its approximation by the vertices of the coordinate grid $\mathbb{Z}^2$. The largest 'approximating' distance is 0.5276. Arrows indicate the vertices of the grid which approximate elements of X. 145

6.3 Configuration X of six points in plane from Figure 6.2: the square grid spacing h is 0.2. The largest 'approximating' distance by this grid is 0.1074. 146

6.4 Configuration X from Figure 6.2 in new axes after rotating the coordinate system clockwise at the angle $\theta = 0.7$. The maximal approximating distance is 0.0737. 146

6.5 Representation of X and G on a circle in the Brucker–Meyer univariate approximation problem and translation of the grid (polygon with solid edges) realised via translation of the points from X (polygons with dotted edges). 149

6.6 Function $f(\cdot)$ on the interval $[t, t + \Delta t]$ when $\Delta t > [r(t) - l(t)]/2$. . 151

6.7 Application of the Brucker–Meyer algorithm in the one-dimensional case: initial and optimal grid with spacing $\alpha = 5$ for the set X of 5 points drawn uniformly and independently on [1,100]. 154

6.8 Approximation of a finite set of points by the nodes of a square grid with the spacing α. 155

6.9 Approximation of a finite set of points X in plane by a uniform grid from Example 6.2.2. 158

A.1 Intersection of the graphs of functions $e^{-x} = 1 - x/2$ and $1 - x/2$, $x > 0$. 186

C.1 Plots of the function $I_\alpha(\alpha/\kappa)$ when α is fixed, $\kappa \in [0, 14]$ ($\alpha = 0.5, 0.6, \ldots, 3$). The plots have been obtain both by using numerical evaluation of integrals in (C.1) and representation (C.3). 190

D.1 Realisation of 6 distances in $\mathbb{R}^3$: a working scheme. 193

List of Tables

4.1 Optimal designs for the model with threshold edge-probability function as functions of the threshold α when $n = 4$. 92

5.1 Table comprising some values of m and r (up to 25 for m and 5 for r) as well as corresponding values of N and T. The possible values of r (italicised) are located in the first row of the table, whereas the possible values of m (italicised) are to be found in the second column of it (these values also coincide with N since they correspond to $r = 0$). The values of N can be found at the intersection of a row and a column corresponding to the values of m and r. The total number of nodes T in an (m, r)-plot can be found to the right of the value of $N(m, r)$ in the same row (these numbers are in bold). The values of N and T were calculated using (5.13) and (5.12) respectively. 122

6.1 Arguments of the MATLAB function `optimal_plot` (p. 157) and the input arguments of Algorithm 4 (p. 156). 157

Author editorial notes

Colours in figures

Some figures in this thesis contain colours and this serves the purpose of a more illustrative graphical representation. However, the content in all figures throughout the thesis is independent of any colours used. This means that no information will be lost or become intractable, should one wish to make black and white copy of the thesis or any of its parts.

Description of algorithms

`End`'s in algorithmic structures, such as `FOR`, `IF`, `WHILE`, were all omitted in the description of the algorithms. This, however, does not make their description ambiguous or intractable.

References

All web links given were correct and working at the time of writing of the thesis. The list of authors was shortened to the first author with adding *et al* and the year of publication whenever the number of authors exceeded two.

Software and permission acknowledgment

This thesis was typeset in LaTeX using MikTeX and WinEdt. The figures were either produced in MATLAB or made using the vector graphics editor Adobe Illustrator. Figure 1.3 was reproduced from Bailey et al (2000) from permission of the authors.

Referring to sections and subsections

The sign § is used to refer to a (sub)section within the thesis.

Any errors and misprints that might persist are all my own.

Notation and abbreviations

1. $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{R}$ are used to denote the natural numbers $(1, 2, \ldots)$, integers, and real number, respectively.

2. Bold face and a roman style are generally used to denote a vector, e.g. $\boldsymbol{\theta} \in \mathbb{R}^n$, unless otherwise noted, in contrast to an italic style for scalars, e.g. $\theta \in \mathbb{R}$.

3. Random variables, but not only, are denoted by italicised capitals. The probability of an event A is denoted by $\mathbb{P}(A)$, e.g. $\mathbb{P}(X \in \mathbb{R}_+)$ denotes the probability of the event "the random variable X takes a non-negative real value". The expected value of X is denoted by $\mathbb{E}[X]$ and its variance by $\operatorname{var} X$.

4. The *support* of a distribution is the smallest, with respect to inclusion, closed set whose complement has probability zero. The support of a random variable (object) X is denoted by $\operatorname{supp} X$.

5. It is followed the very convenient, albeit theoretically incorrect practice of using the term density both for continuous random variables and for the probability mass function of discrete random variables. In line with this convention in what follows integrals are to be interpreted as sums when necessary. Thus, if X is a discrete random variable, then
$$\mathbb{E}[X] = \int_{\operatorname{supp} X} x f_X(x) \mathrm{d}x \equiv \sum_{x \in \operatorname{supp} X} x f_X(x).$$

6. The entropy of a distribution with density (or probability mass function) $f_X(x)$ is denoted by $\operatorname{Ent}\{f_X(x)\}$:
$$\operatorname{Ent}\{f_X(x)\} := -\int_{\operatorname{supp} X} f_X(x) \log f_X(x) \mathrm{d}x.$$

7. The following notation is used for the indicator function:

$$\mathbb{1}_{\{A\}} = \begin{cases} 1, & \text{if the event } A \text{ takes place,} \\ 0, & \text{otherwise.} \end{cases}$$

8. The use of vertical bars, $|\cdot|$, when applied to a discrete set, indicates the number of elements in the set, e.g. $|\mathcal{C}| = n$.

9. Notation $\arg\max_x f(x)$ is used to denote the set $\{x|\, f(y) \leq f(x)\, \forall y\}$, that is the value of the argument for which the value of the given expression attains its maximum value.

10. The 'big Oh' notation is used in this thesis to describe the efficiency of the algorithms. For example, the writing $O(n^2)$ means that the time complexity $T(n)$ of a corresponding algorithm is of order n^2 in the following sense:

$$\exists M \in \mathbb{R}_+\, \exists N \in \mathbb{N} : T(n) \leq Mn^2\ \forall n > N.$$

    O-notation is an upper bound asymptotic order notation and is not to be abused by assuming that it gives an exact order of growth: a running time $O(n^2)$ does not imply that the running time is not also $O(n)$ (Graham, Knuth and Patashnik (1990, p. 429–229)).

11. The sign ':=' is used to denote "is defined as" or "equal by definition".

12. It is used the standard notation e to refer to the base of natural logarithms:

$$e := \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n,$$

    and the following notation for binomial coefficients:

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}.$$

13. Natural logarithms are denoted by '$\log x$':

$$\log e = 1;$$

14. The diagonal of an n-ary relation $X^n := X \times X \times \ldots \times X$ (n times) on a set X is denoted by $\operatorname{diag} X^n$:

$$\operatorname{diag} X^n := \{(x_1, x_2, \ldots, x_n) \in X^n |\, x_i = x_j,\ 1 \leq i < j \leq n\}.$$

Chapter 1

Introduction

Waiho i te toipoto, kaua i te toiroa.

(Māori proverb)

Let us keep close together, not far apart.

1.1 Why inference and optimal design on random graphs?

A graph is a mathematical structure which is used to model pairwise relations within a set of objects, often of the same nature. Describing the structure of the interconnection pattern of a network of interacting objects, graphs represent convenient mathematical objects allowing one to capture, analyse, and interpret such interactions and their development.

Graphs are convenient because they are abstract—one can study them regardless of the nature of the set of the interacting objects. However, depending on what these objects actually represent, the corresponding graphs, or their dynamics, may reflect development of the processes observed by biologists, epidemiologists, physicists, engineers, sociologists and ecologists, who often see the same interesting features and phenomena in the network structures that appear in their interdisciplinary studies. Discovery of *small-world networks* and the parallels between the spread of an infectious disease in plant epidemiology or forest fires on the one hand, and percolation processes on discrete and continuum structures on the other

hand, are just two of the numerous possible examples. Not surprisingly, the interest in network science that arose in the early 1990's, and has increased ever since, produced interesting applications in mathematical epidemiology, social networks and computer networks theory[1].

A graph that is generated by some random process is called a *random graph*. Strictly speaking, a random graph as a mathematical object can be regarded as a random element on a certain probability space taking values in a set of graphs, but there may be, and indeed this is often the case, a rule according to which a realisation of such random element can be obtained. In some situations it is reasonable to assume that the vertices of the considered random graph are fixed, while edges occur randomly and the probability that an edge is present between a given pair of vertices obeys a parametric law that depends on the degree to which the corresponding objects are susceptible to an interaction.

A fairly realistic example is the following: a researcher dealing with a phenomenon of signal propagation establishes that the strength of the signal, and hence the chances for its successful reception, decays according to a power law with distance regardless of the physical characteristics[2] of the medium in which the signal propagation evolves. However, there is a correspondence between physical conditions and the exponent of the power law describing the signal strength decay, and the researcher wants to know this correspondence. Taking measurements of the signal strength in a particular medium will give information on the scaling exponent. However, if the researcher is only equipped with signal detectors that can measure the signal's presence or absence with some uncertainty related to the signal's strength and the number of such detectors is fixed, some of their allocations will be more informative and some will be less informative. What choice of the detectors' positions is the most optimal?

Generally speaking, there are three key factors that influence the answer to this

[1]In the author's opinion, the postponement of widespread progress on the dynamics of large-scale networks until the 1990's was, to some extent, due to the lack of sufficient computing power to simulate the behaviour of large complex networks prior to that time.

[2]The fundamental law that the researcher establishes might only hold within some range of values of the medium's characteristics.

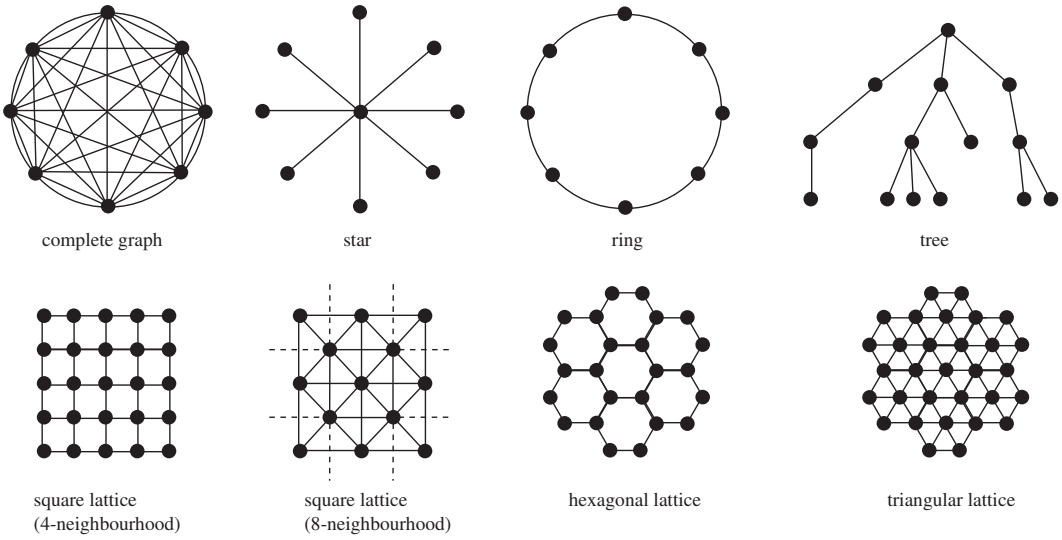


Figure 1.1: Examples of different types of regular discrete graph topologies.

question:

1. the form of the decay function and the probabilistic nature of the signal detection;

2. the local topology of the space within which the signal propagates;

3. the way in which the information derived from the detectors is quantified.

The form of the decay function affects the optimal choice in an obvious way: the higher the chances are for the signal to travel longer distances—the lower the chances are that a clever experimenter will put all the detectors close to the emitter(s) of the signal. The local topology of the space within which the signal propagates describes all permitted directions of the travelling signal to propagate along once it is sent by the emitters; this information should be described for any possible position of an emitter within the considered space. We will refer to this information as the *topology of interactions* or *contact network*. The topology of interactions can be represented by a graph, either discrete or continuum. Figure 1.1 depicts basic examples of different types of regular discrete graph topologies.

Finally, different measures of quantifying information delivered by the detectors will lead to different optimal arrangements of them. Generally, the value of the information carried by data depends on what exactly one intends to do with the data when they are collected.

In the next section we give a general description of the model and problem under study as well as provide further motivation details.

1.2 General model description and further motivation

1.2.1 Model

Consider an arrangement of n objects $x_1, x_2, \ldots, x_n$ within a subset $\mathcal{D}$ of some larger set X, possibly a metric space. There is an unoriented link between each pair x_i and x_j, independently of the positions of the other objects and links between them, with some probability $p_{ij} = p_{ji}$ which depends on the non-negative weight r_{ij} attributed to (x_i, x_j) (in the case of a metric structure these weights will be distances between objects), i.e.

$$p_{ij} := \mathbb{P}(x_i \text{ and } x_j \text{ are connected}) = p(r_{ij}, \boldsymbol{\theta}), \tag{1.1}$$

where $\boldsymbol{\theta}$ is an unknown parameter, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$, and function $p(\cdot, \cdot)$ acts as follows:

$$p : \mathbb{R}_+ \times \Theta \to [0, 1]. \tag{1.2}$$

One may additionally require the following two assumptions to hold, particularly when r_{ij} are distances:

Assumption 1.2.1. *The function $p(r, \boldsymbol{\theta})$ is non-increasing in r for each value of $\boldsymbol{\theta}$.*

Assumption 1.2.2. *The function $p(r, \boldsymbol{\theta})$ tends to zero as r tends to infinity, and it tends to unity as r tends to zero for each value of $\boldsymbol{\theta}$:*

$$\lim_{r \to \infty} p(r, \boldsymbol{\theta}) = 0, \tag{1.3}$$

$$\lim_{r \to 0} p(r, \boldsymbol{\theta}) = 1. \tag{1.4}$$

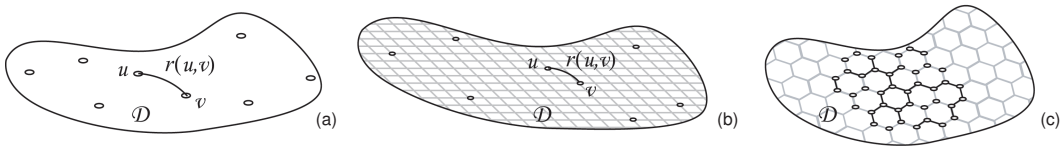


Figure 1.2: Arrangement of n objects within the set $\mathcal{D}$: there is a link between each pair (u, v) of them with probability $p(r(u, v), \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. In **(a)** $\mathcal{D}$ is a bounded region in $\mathbb{R}^2$, in **(b)** $\mathcal{D} \subseteq \mathbb{Z}^2$, and in **(c)** $\mathcal{D}$ is a subset of nodes of a hexagonal grid—only neighbouring nodes can be connected realising the so called nearest-neighbour interaction.

Depending on the context the following names are commonly used to refer to the function $p(r, \boldsymbol{\theta})$:

- *edge-probability function* or *edge-probability profile*;
- *connectivity kernel* or *connection kernel*.

The described procedure of establishing connections between a finite number of objects taken in the set $\mathcal{D}$ results in a finite random graph on these objects as nodes. Some examples of different types of the set $\mathcal{D}$ are shown in Figure 1.2, where long-range connections are possible within $\mathcal{D}$ in (a) and (b), and only connections between adjacent nodes of the hexagonal lattice are allowed in (c) leading to nearest-neighbour interaction.

The **statistical interest** in considering the described model is to make inference on its parameter $\boldsymbol{\theta}$. This should be done after observing a random graph on n nodes, formation of which is governed by the edge-probability function $p(r, \boldsymbol{\theta})$. The **optimal design problem** consists in finding an *optimal arrangement* of these n nodes in order to extract as much information about θ as possible—this should be done before looking at an observation of the random graph, but certainly taking into account all possible outcomes. Information provided by each of these outcomes for a given arrangement should be carefully quantified, so that different arrangements can be compared in terms of their usefulness for solving the problem of parameter estimation.

1.2.2 Motivation: theoretical positions and practical aspects

Theoretical aspects

The random graph model described in § 1.2.1 can be viewed as an extension of the Erdős–Rényi random graph in which each pair of vertices is connected by an edge with probability p. More formally, the Erdős–Rényi random graph $G_{n,p}$ is constructed in the following way. Let $V = \{1, 2, \ldots, n\}$, and let $(X_{ij} : 1 \leq i < j \leq n)$ be independent Bernoulli random variables with parameter p. For each pair $i < j$ an undirected edge (i, j) is placed between vertices i and j if and only if $X_{ij} = 1$. The resulting graph is named after the two prominent Hungarian mathematicians Paul Erdős and Alfréd Rényi (1959, 1960), although historically it appears to have been introduced first by Edgar N. Gilbert (1959).

Being a truly elegant model, the Erdős–Rényi random graph model was initially introduced and studied in order to understand the properties of 'typical' graphs. The random graph $G_{n,p}$ has received an enormous deal of attention, predominantly within the community working on probabilistic combinatorics (Grimmett (2008)).

The Erdős–Rényi random graph on n vertices can be seen as a bond percolation model on the complete graph K_n with the bond percolation probability p (in this percolation model the random graph is obtained by deleting edges of K_n, each with probability p and independently of each other). On the one hand, as noticed by Grimmett (2008), "the parallel with percolation is weak in the sense that the theory of $G_{n,p}$ is largely combinatorial rather than geometrical". On the other hand, we find it useful to indicate an underlying graph on which percolation is considered, and thus to identify the topology of interactions (in the case of the Erdős–Rényi model it is the complete graph K_n since any two nodes can be connected with probability p). This view is formally represented in the next chapter. Some of the results obtained in this thesis refer to classical percolation on $\mathbb{Z}^d$. We believe that these results can further be generalised to percolation models on other lattices or, even more generally, irregular infinite (but locally finite) graphs.

The two fundamental assumptions of the classical $G_{n,p}$ model are that (i) edges are independent of each other, and (ii) edges are equiprobable. Clearly, either of these assumptions may often be inappropriate for modelling real-life

phenomena. While preserving the former assumption, the model introduced in § 1.2.1 improves upon the latter one. For other alternatives see the popular Watts and Strogatz model, which produces graphs that are homogeneous in degree (see Milgram (1967), Travers and Milgram (1969), Watts and Strogatz (1998) and Watts (2003)) and the Barabási-Albert model of preferential attachment (see Albert and Barabási (2002)), which produces graphs with scale-free degree distribution.

Practical aspects and the problem of incomplete observations

Many real-world phenomena can be modelled by random graphs, or more generally, by dynamically changing random graphs. Specifically, host-pathogen biological systems that may combine primary and nearest-neighbour or long-range secondary infection processes can be efficiently described by spatio-temporal models based on random graphs evolving in time (Gibson et al (2006)).

Although a continuous observation of an epidemic is not always possible, a spatial 'snapshot' may provide one with some, albeit highly incomplete, knowledge about the epidemic. In terms of the model this knowledge results in a random graph realised in some metric space. Moreover, under certain experimental circumstances it is not possible to observe some or even all of the edges of such a random graph—all one would know then are the vertices which correspond to the infected sites, that is to those sites which interacted as a result of the evolution of the process under consideration.

One particular application refers to the colonisation of susceptible sites, such as seeds or plants grown on a lattice, by virus, fungal, or bacterial pathogens with limited dispersal abilities. A typical example is the spread of infections through populations of seedlings by the fungal pathogen, *Rhizoctonia solani* Kühn. This economically-important pathogen is wide spread with a remarkably wide host range (Chase (1998)). In addition to its intrinsic economic importance, it has been extensively used as an experimental model system to test epidemiological hypotheses in replicated microcosms (Gibson et al (2004) and Otten et al (2004)) and to study biological control of pathozone behaviour by an antagonistic fungus and disease

dynamics (Bailey and Gilligan (1997)). Transmission of infection between plants occurs by mycelial growth from an infected host, with preferential spread along soil surfaces—hence the missing information about the structure of interactions.

The spread of infections with limited dispersal abilities among plants can be viewed as a spatial *SIR* epidemic with nearest-neighbour secondary infections and removals, and can be related to percolation processes on regular lattices. An illustrative example, classical now (Grimmett (1999), Trapman (2006)), of a problem arising in botanical epidemiology which can be related to percolation is that of an orchard with trees planted at regular distances in such a way that their positions can be seen as vertices of the square lattice. Assume that one of the trees (the central tree, for instance) is infected by a disease. The infection process is such that exactly one time unit after being infected a tree will die[3]. After becoming infected a tree becomes infectious and remains so until its death. While infectious it spreads infectious material to its nearest neighbours, each of which might become infected (if they were not already so) with some probability p. It is also assumed that all infections occur independently of each other.

Bayesian estimation for percolation models of disease spread in plant populations in the context of the spread of *Rhizoctonia solani* has been presented by Gibson et al (2006). Bailey et al (2000) studied the spread of this soil-borne fungal plant pathogen among discrete sites of nutrient resource using simple concepts of percolation theory; a distinction was made between invasive and non-invasive saprotrophic spread (see Figure 1.3). The authors of these papers formulated statistical methods for fitting and testing percolation-based spatio-temporal models that are generally applicable to biological or physical processes that evolve in time in spatially structured populations. Estimation of spatial parameters from a single snapshot of an epidemic evolving on a discretised grid under the assumption that fundamental spatial statistics are near equilibrium was studied in Keeling et al (2004).

The difficulties in performing inference for these models in the presence of ob-

[3]Of course, this is a highly idealised assumption, but we often have to make simplifications in model assumptions and quite often the analysis based on such simplifications rewards us with a valuable insight into the problem under study!

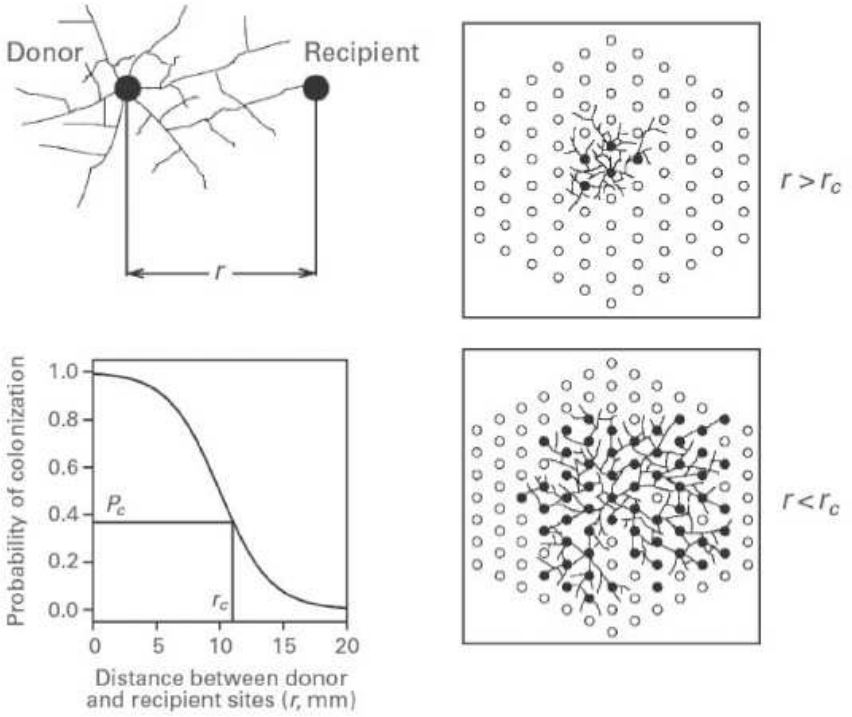


Figure 1.3: The growth of the mycelial colonies as a percolation process studied by Bailey and Gilligan (1997) and Bailey et al (2000). The edge-probability decay may be 'combined' from simpler decays: e.g. the progress of disease in a population of radish plants exposed to primary infection by *R. solani* in the presense/absence of *T. viride* was studied in Bailey and Gilligan (1997) using the following form for the probability of infection: $p(r, \theta) = (\theta_1 + \theta_2 r)e^{-\theta_3 r}$.

servational uncertainty or incomplete observations can be overcome to an extent by employing a Bayesian approach and modern powerful computational techniques—mainly Markov Chain Monte Carlo (for instance, see Gibson (1997)). Markov Chain Monte Carlo methods often offer important advantages over existing methods of analysis. In particular, they allow a much greater degree of modelling flexibility, although the implementation of these methods can be problematic because of convergence and mixing difficulties which arise due to the amount and nature of missing data.

An aspect which has received little attention in the context of the described models is that of experimental design. Statisticians have investigated the question of experimental design in the Bayesian framework (see Chaloner and Verdinelli (1995) for a review). The work of Müller and others (e.g. Müller (1999), Verdinelli (1992)) examined the ways of identifying designs that maximise the expectation of a utility function.

In this thesis we study the problem of optimal design for random graph models within the utility-based Bayesian framework and discuss generic issues that arise in this context. Realisations of random graph can be seen as a final snapshot of nearest-neighbour or long-range disease spread spatio-temporal dynamics or as a result of the percolation process on a node network (see Read and Keeling (2003) and Bailey et al (2000)).

The purpose of the 'optimal design', as presented in this thesis, is not as much relevant to epidemics in large human populations where one employ mean-field considerations, as to networks with more distinctive topological structure. On the other hand, disease evolution on networks and plant epidemiology are not the only possible practical contexts within which the problem of optimal design for random graphs can be studied. The following are just some examples of areas within which random graph and network models have recently been rapidly developed, and which keep creating a demand and open new opportunities for studying non-linear experimental design problems in the context of random graphs:

- radio networks, e.g. random mobile graphs introduced in Tyrakowski and Palka (2005) for analysis of distributed algorithms requiring synchronous

communication in radio networks;

- geophysics: determining locations of seismometers to locate earthquakes with minimum uncertainty, locating receivers optimally within a well to locate induced microseismicity during production, designing source/receiver geometries for acoustic tomography that optimally detects underwater velocity anomalies; see Curtis (2004 a,b) and references therein;

- general temporal stochastic ageing and fatigue processes, e.g. Ryan (2003);

- psychological experiments, e.g. Kueck et al (2009) and neurophysiological experiments, e.g. Paninski (2005).

We conclude this section by listing a few examples of edge-probability decays for the model introduced in § 1.2.1:

$$\textit{threshold decay: } p(r,\theta) = \mathbb{1}_{\{r\leq\theta\}} + \alpha\mathbb{1}_{\{r>\theta\}},\ \alpha\in[0,1), \theta\in\Theta\equiv\mathbb{R}_+;$$

$$\textit{exponential law decay: } p(r,\theta) = e^{-\theta r},\ \theta\in\Theta\equiv\mathbb{R}_+;$$

$$\textit{power-law decay: } p(r,\boldsymbol{\theta}) = (1+\theta_1 r)^{-\theta_2},\ \boldsymbol{\theta}=(\theta_1,\theta_2)\in\Theta\equiv\mathbb{R}_+^2;$$

$$\textit{Cauchy decay: } p(r,\theta) = (1+\theta r^2)^{-1},\ \theta\in\Theta\equiv\mathbb{R}_+;$$

$$\textit{logistic function: } p(r,\boldsymbol{\theta}) = \theta_1/\exp\{\theta_2(r-\theta_3)\},\ \boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3)\in\Theta\equiv\mathbb{R}_+\times\mathbb{R}_+.$$

1.3 Related work on inference and experimental design problems for stochastic interaction and spatial response models

In spatial response models (e.g. image analysis or geostatistics) and stochastic interaction models (e.g. epidemic models) one studies responses as functions on either spatial locations or their interactions as well as possibly time. We give a brief review of existing related work on inference and design problems for stochastic

interaction and spatial response models in this section. This review is inevitably selective and ultimately incomplete.

1.3.1 Spatial response models

Müller (2007) is an excellent account on applications of experimental design theory to spatial statistical response models. The monograph discusses exploratory designs, designs for spatial trend estimation and multipurpose designs for these models and contains many useful references in the field.

Fuentes et al (2007) develop a fully Bayesian spatial statistical methodology to design air pollution monitoring network with good predictive capabilities and minimised costs of monitoring. In order to estimate the associate model parameters the authors use the technique of Reversible Jump Markov Chain Monte Carlo. The design problem was solved using a specific utility function which also took into account monitoring costs.

A different approach is taken by Papadimitriou et al (2005). These authors solve the problem of optimising the location and number of sensors for the purpose of most accurately predicting the response of randomly vibrating structures at unmeasured locations by minimising the errors in the response predictions obtained by the kriging method at unmeasured locations.

Gatrell et al (1996) give a review of a number of methods for the exploration and modelling of spatial point patterns with particular reference to geographical epidemiology (geographical incidence of dicease). Gaudard et al (1999) present a complete Bayesian methodology for analysing spatial data and estimating structural covariance parameters modelling the spatial covariance structure assuming Gaussian random fields.

Besag and Green (1993) present a thorough review on Baysian computation in spatial statistics, tracing the early development of Markov Chain Monte Carlo methods in Bayesian inference for statistical physics and making a particular emphasis on the Bayesian analysis of agricultural experiments.

1.3.2 Stochastic interaction models

Epidemic models represent a very good example of stochastic interaction models. A discussion on the nature of infectious disease data, its modelling aspects as well as an extended review on previous work on epidemic modelling (including statistical analysis of epidemics in homogeneous and structured populations) is presented in the PhD thesis of Demiris (2004) together with an extended bibliographical review on the topics mentioned. Using the Bayesian paradigm the author develops suitably tailored Markov Chain Monte Carlo algorithms in order to perform statistical inference for an epidemic model with two levels of mixing as well as a generalised *SIR* stochastic epidemic with an underlying contact structure using random graphs given the final size(s) of the epidemic outcome.

Keeling (1999) addresses the effects of local spatial structure on epidemiological invasions and determines invasions thresholds by modelling the behaviour of individuals in a fixed network and the spread of a disease through a structured network. The role and implications of network structure for epidemic dynamics are studied in Parham and Ferguson (2006) and Keeling (2005). Estimation of important dispersal and spatial parameters of a spatial epidemic from a single snapshot is presented in Keeling et al (2004).

Neal (2003) considers a generalized stochastic epidemic on a Bernoulli random graph. By constructing the epidemic and graph in unison, the epidemic is shown to be a randomized Reed-Frost epidemic. Exact final-size distribution and extensive asymptotic results are also derived. Ball and Neal (2008) study *SIR* epidemics on social networks involving two levels of mixing: one is due to the network structure and another is independent of it representing casual contacts. The authors derive a deterministic model that approximates the spread of an epidemic that becomes established in a large population.

Glickman and Jensen (2005) study the problem of paired comparison experiments and formalise it as a Bayesian optimal design problem. The authors develop a pairing method that maximises the expected gain in the Kullback–Leibler divergence from the prior to the posterior of the individual's strengths. By changing the utility function Glickman (2008) derives Bayesian locally-optimal design of

knockout (paired comparison) tournaments when the goal is to identify the overall best player.

Curtis (2004 a,b) and references therein represent an account on the theory and practice of experimental design in geophysical problems: locating receivers or sensors optimally in a heterogeneous environment in order to collect data most efficiently is the main topic of this area of research in optimal experimental design theory. Sensor placement applications and the problem of finding optimal sensor locations is the main motivation of the recent paper of Ren et al (2008) in which an adaptive evolutionary Monte Carlo algorithm is used in order to optimise certain complicated "black-box" objective function.

Analysis of the dynamics of spatiotemporal epidemics from time-series data has been done by Filipe et al (2003) and Gibson et al (2006) via semi-spatial modelling and moment-closure approximation approach. This work has seen further development in Cook et al (2008) where the authors studied the problem of optimal observation times in experimental epidemic processes distinguishing between the so called *progressive* and *pedagogic* design scenarios.

1.4 Outline of the thesis

This thesis is organised as follows. Chapter 2 contains a review of the standard mathematical notions, tools and techniques that are used throughout the thesis. By recalling basic notions from the graph theory, the techniques of Bayesian statistical inference and Monte Carlo and Markov Chain Monte Carlo methods we set up a framework for further development.

Chapter 3 argues for the choice of utility-based Bayesian optimal experimentation as a criterial framework of our further considerations. We give a rigorous review of the Shannon entropy, the Lindley information measure and the Kullback–Leibler divergence as measures of informativeness of experiments and discuss simulation-based methods of evaluation of the expected utility and thus identifying optimal designs. We also introduce two different experimental scenarios in this chapter: *progressive* and *instructive* designs. Using graphs as an underlying

interaction topology as well as model objects and utility-based Bayesian framework we give a second, more specific formulation of the model and design problem (n-node optimal arrangement problem for random graphs), concluding the chapter with some examples.

Chapter 4 contains theoretic results for some basic random graph models. We first prove a general worst case scenario result for indefinitely growing or diminishing configurations. We then study two-node and three-node designs through a number of examples which identify important features of expected utility surfaces. We continue with studying proximity (geometric) graphs and graphs with threshold edge-probability decay and obtain an explicit solution to the optimal design problem for proximity graphs on star interaction topologies considered in metric spaces. We also show that the case of a threshold edge-probability decay can be treated numerically. The chapter is concluded by a discussion on how the obtained theoretic result for proximity graphs can be used to easily show non-preservation of optimal designs under replication (in non-linear models).

Chapter 5 concerns inference and experimental design problems for finite clusters from percolation on the integer lattice $\mathbb{Z}^d$, $d \in \mathbb{N}$. We introduce *inner-outer plots* as a design class and show that in presence of incomplete observations for percolation models the most populated design is not necessarily the most optimal design. This chapter contains both theoretical and practical results for such nearest-neighbour interaction models under both *progressive* and *instructive* design scenarios. The chapter concludes with a discussion on what the generalisations of the obtained results and methods might look like for long-range interaction models. We also discuss the potential of deformations of the square lattice as a way towards identifying a whole class of lattice designs that keep the dimensionality and cardinality of the design space low.

Chapter 6 can be regarded as independent of the rest of the thesis. This chapter deals with a problem of grid approximation of a finite set of points—a design problem in its own way. The last chapter, Chapter 7 concludes the thesis by identifying contributions of the thesis and potential directions for future work.

Chapter 2

Tools and methodology

The standard tools and techniques that are used in the rest of the thesis are reviewed in this chapter. These tools include basic notions from the graph theory, the techniques of Bayesian statistical inference and Monte Carlo methods.

2.1 Basic notions from the graph theory

Loosely speaking, a graph consists of vertices connected by edges. Generally, there may be more than one edge connecting the same pair of (not necessarily distinct) vertices and the direction of their connection may also be important. To proceed with formal definitions let us agree on the following notation: define the set of all subsets of two elements of a given set V by $V \otimes V$, that is[1]

$$V \otimes V := \{\{u, v\} : u, v \in V, u \neq v\}. \tag{2.1}$$

As traditionally, the Cartesian product of the set V with itself is denoted by $V \times V \equiv V^2$.

An *oriented (unoriented) multigraph* $G = (V, E, \psi)$ consists of a set of vertices,

[1] Although, conventionally, the sign $\otimes$ is used to denote the tensor product, it is used in this thesis only to denote the set of all two-element subsets of a given set. This should not cause any confusion, since we are far from any context involving tensors.

V, the set of edges E, and a map

$$\psi : E \to V \times V \tag{2.2}$$

$$(\psi : E \to V \otimes V \cup \{\{u\} : u \in V\}) \tag{2.3}$$

that assigns a pair of vertices to each edge $e \in E$:

$$\psi(e) = (u, v) \in V^2, \ \ u, v \in V$$

$$(\psi(e) \in V \otimes V \cup \{\{u\} : u \in V\}).$$

Figure 2.1 shows an example of an oriented and unoriented multigraph on the same vertex set $V = \{u_1, u_2, u_3, u_4\}$. Although the edge sets of both graphs coincide: $E = \{e_1, e_2, e_3, e_4, e_5\}$, the maps ψ are different:

$$\psi(e_1) = (u_1, u_1), \psi(e_2) = (u_1, u_2), \psi(e_3) = (u_3, u_2), \psi(e_4) = (u_2, u_3), \psi(e_5) = (u_4, u_4),$$

for the oriented graph (Figure 2.1, left), and

$$\psi(e_1) = \{u_1\}, \psi(e_2) = \{u_1, u_2\}, \psi(e_3) = \{u_2, u_3\}, \psi(e_4) = \{u_2, u_3\}, \psi(e_5) = \{u_4\},$$

for the unoriented graph (Figure 2.1, right).

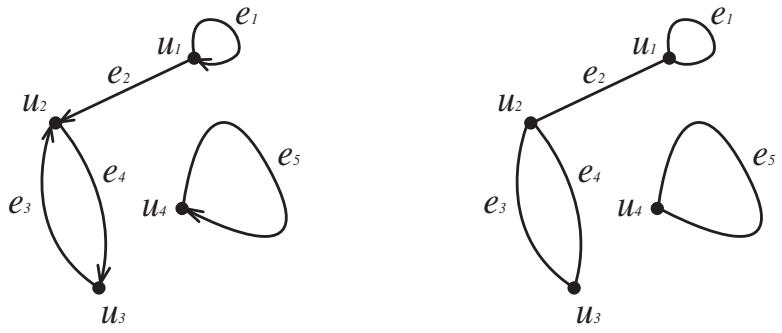


Figure 2.1: Oriented (left) and unoriented (right) multigraph on the same set of vertices.

Vertices of a multigraph are also called nodes. The *order* of a multigraph G is the cardinality of its vertex set $|V|$. The *size* of a multigraph is the cardinality of its edge set $|E|$. Directed multigraphs are also called oriented multigraphs—the orientation of at least some of their edges may be important[2].

[2]In graph theory literature the oriented graphs are often abbreviated as *orgraphs*.

A *simple* graph is a multigraph which has no *loops*[3] and no multiple edges (these connect the same vertices). Thus, a multigraph $G = (V, E, \psi)$ (oriented or unoriented) is simple if and only if the map ψ is injective, that is

$$\psi(e_1) = \psi(e_2) \Rightarrow e_1 = e_2 \; \forall e_1, e_2 \in E,$$

and the image $\psi(E)$ of the map ψ defined either by (2.2) or (2.3), depending on whether G is oriented or unoriented, is a subset of the following set:

$$\psi(E) \subseteq \begin{cases} V \times V \setminus \operatorname{diag} V, & \text{if the multigraph } G = (V, E, \psi) \text{ is oriented} \\ V \otimes V, & \text{otherwise.} \end{cases}$$

If a multigraph G is simple then the map ψ can be considered to be a simple inclusion and depending on whether G is oriented or unoriented it is enough to assume that E is a subset of $V^2 \setminus \operatorname{diag} V$ or $V \otimes V$, correspondingly, to fully define G. In what follows, unless stated otherwise, the term *graph* refers to a simple graph. Moreover, let us refer to the elements of the edge set E of a simple graph $G = (V, E)$ as pairs (u, v) regardless of whether G is oriented or not, keeping in mind that a pair $(u, v) \in E$ is an oriented pair, should G be oriented, and that it is an unoriented pair otherwise.

Every vertex u of an oriented graph has an *out-degree* and an *in-degree*, the former being the number of edges that originate at u, and the latter being the number of edges that have u as a second *end vertex*. Denoting the in-degree of a vertex u by $\deg_{in}(u)$ and its out-degree by $\deg_{out}(u)$, one can formally write:

$$\deg_{in}(u) := |\{v \in V \,|\, (v, u) \in E\}|,$$

$$\deg_{out}(u) := |\{v \in V \,|\, (u, v) \in E\}|.$$

The notions of in-degree and out-degree are no longer applicable in the case of an unoriented graph. Instead, one considers the number of all neighbours of a vertex: *the degree of a vertex* u of an unoriented graph is denoted by $\deg(u)$ and it is (by definition) equal to the cardinality of its *neighbourhood* $N(u) := \{v \in V \,|\, (u, v) \in E\}$. If this is finite for each vertex, we call the graph *locally finite*. Edges of an undirected graph are also called links.

[3] A *loop* is an edge connecting a vertex to itself.

A graph $G' = (V', E')$ is a *subgraph* of a graph $G = (V, E)$ if and only if

1. $V' \subseteq V$,

2. $E' \subseteq E$ and $(u, v) \in E' \Rightarrow u, v \in V'$.

In general, a subgraph need not have all possible edges. If a subgraph inherits every edge with end points belonging to V' from the original graph G, it is a *node-induced subgraph*. In contrast, an *edge-induced subgraph* is a subset of the edges of a graph G together with any vertices that are their endpoints. Any node-induced subgraph will be referred to simply as an *induced subgraph*. An example of a graph and its induced subgraph is given in Figure 2.2.

Figure 2.2: A subgraph induced by the vertices of the graph from left of degrees distinct from 4 is a cycle from right.

A *path* of a graph is a sequence of some of its vertices $u_0, u_1, \ldots, u_{m+1}, \ldots$, such that $(u_{i-1}, u_i) \in E$. A *simple path* is a path in which no vertex occurs more than once. A finite path $u_0, \ldots, u_m$ is closed if $u_0 = u_m$. A finite closed path is called a *cycle*. A finite closed simple path is called a *simple cycle*. A graph is called *connected*, if there exists a path between any two of its vertices. The set of vertices of any graph naturally splits into subsets of vertices which are connected to each other. Graphs induced by these subsets are called *connected components*. A graph is *connected* if and only if it consists of a single connected component.

The *complement* $\bar{G}$ of an undirected graph $G = (V, E)$ is a graph $(V, V \otimes V \setminus E)$. A *complete* graph K_n of order n is a graph with n vertices in which every vertex is *adjacent* to every other[4]. An example of a graph and its complement is shown in Figure 2.3. The graph G has the only cycle consisting of the vertices u_2, u_3, and u_4; the vertex u_1 is connected to this cycle. The vertex u_5 is an isolated

[4]For example, the top-left graph in Figure 1.1 is a complete graph K_8.

vertex, and therefore G has two connected components. Its complement $\bar{G}$ has exactly three cycles and represents a single connected component. By the *union* of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ we will understand the graph $G_1 \cup G_2 := (V_1 \cup V_2, E_1 \cup E_2)$. The union of the graph G and its complement $\bar{G}$ from Figure 2.3 is a complete graph K_5.

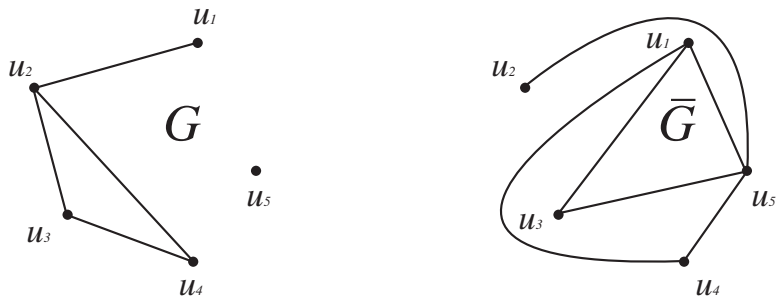


Figure 2.3: Example of a graph G and its complement $\bar{G}$.

A convenient way to represent a graph is to indicate its adjacency structure. When V is finite this can be done in the form of a matrix. If the cardinality of the vertex set V is n_V, then the *adjacency matrix* $A = (a_{ij})$ of this graph is an $n_V \times n_V$ matrix in which entry a_{ij} is equal to 1 if and only if $(i, j) \in E$, and is equal to 0 otherwise. Conventionally, the adjacency matrix of an undirected graph is always symmetric. The adjacency matrices of the graph G and its complement $\bar{G}$ from Figure 2.3 are given below:

$$A_G = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad A_{\bar{G}} = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

Often, it is useful to distinguish between a strong and weak connection between vertices of a graph. This naturally leads to a notion of *weighted graphs* in which each edge $(u, v) \in E$ receives a weight $r(u, v)$. Weights are usually non-negative real numbers: $r(u, v) \in \mathbb{R}_+ \ \forall u, v \in V$. One can extend the $0 - 1$ graph's adjacency representation of weighted graphs by allowing the ij^{th} entry of the adjacency matrix A to take the value of the weight of the edge connecting the i^{th} and the

j^{th} vertices of the graph G. If V is uncountable then a matrix representation of the adjacency structure is not possible, but one can still consider a non-negative *weight function* $r(\cdot,\cdot)$ defined on E:

$$r : E \to \mathbb{R}_+.$$

For convenience we extend this function to V^2 as follows:

WF.1 $r(u,u) = 0 \ \ \forall u \in V$,

WF.2 $r(u,v) = r(v,u)\ \forall u, v \in V$,

WF.3 $r(u,v) < +\infty\ \forall (u,v) \in E$,

WF.4 $r(u,v) = +\infty\ \forall (u,v) \in V^2 \setminus E \setminus \operatorname{diag} V^2$.

Possessing the properties WF.1-4, the weight function $r(\cdot,\cdot)$ contains complete information about the adjacency structure of a simple weighted graph. Let us agree therefore to refer to $r(\cdot,\cdot)$ as R, regardless whether V is countable or uncountable[5], and write $G = (V, R)$ to denote a simple weighted graph with the vertex set V and weight structure R.

The notion of an induced graph can also be naturally generalised to weighted graphs.

Example 2.1.1. *Let us consider the graph G from Figure 2.3 and assume that the edge weights are equal to Euclidean distances between corresponding nodes (in some conditional units of distance measurements). Denote the adjacency matrix representing the corresponding weighted graph by R. Then R is as follows (in some units of length measurement):*

$$R = \big(r(u_i,u_j)\big)_{1\le i,j\le 5} = \begin{pmatrix} 0 & 13.914 & +\infty & +\infty & +\infty \\ 13.914 & 0 & 10.637 & 19.316 & +\infty \\ +\infty & 10.637 & 0 & 10.986 & +\infty \\ +\infty & 19.316 & 10.986 & 0 & +\infty \\ +\infty & +\infty & +\infty & +\infty & 0 \end{pmatrix}.$$

[5]Whenever V is countable R will denote the weight matrix $R = (r(i,j))_{i,j\in V}$.

Example 2.1.2. *Let $G = (V, R)$ be a weighted graph, where $V = \mathbb{R} \times \mathbb{R}$ and R represents Euclidean distances between each two points of the plane $\mathbb{R}^2$. Let N be a natural number and let V' be defined as follows:*

$$V' = \{(x, y) \in V \mid \max\{x, y\} \leq N\} \cap \mathbb{Z}^2.$$

The induced graph $G' = (V', R|_{V' \times V'})$ is then a complete graph representing a $(2N + 1) \times (2N + 1)$ square consisting of the nodes of the integer lattice $\mathbb{Z}^2$ with the origin as a central node. The weight of the edge between any two nodes of this graph is equal to the Euclidean distance between them. Here by $R|_{V' \times V'}$ we denoted the weight structure of G' coinciding with R on the vertex set V', that is to say the restriction of R to $V' \times V'$.

Example 2.1.3. *Let $G = (\mathbb{Z}^2, R)$ be a weighted graph, where R is defined as follows:*

$$r\left(u(x_1, y_1), v(x_2, y_2)\right) := \begin{cases} 1 & \text{if } \|u - v\|_1 := |x_2 - x_1| + |y_2 - y_1| = 1, \\ +\infty & \text{otherwise,} \end{cases} \quad \forall u, v \in \mathbb{Z}^2.$$

Let $N = 3$ and V' be defined as in Example 2.1.2. Then the graph $G' = (V', R|_{V' \times V'})$ induced by V' can be graphically represented as the left graph in Figure 2.2. Each depicted edge of this induced subgraph has weight 1. As in the previous example $R|_{V' \times V'}$ is a weight structure which agrees with R on V'.

2.2 Likelihood and Bayesian statistical inference

The fundamental problem of statistical science is that of *inference*. In order to design as effective an experiment as possible for making inference from consequently observed data, we need to describe the methodology within which inference and experimental design will be made. In this section the fundamental aspects of the likelihood-based statistical inference and inference made within a Bayesian framework are reviewed. A discussion on the measures of informativeness of experiments, when the purpose is inference on the model parameter(s), within each of these two choices is presented in Chapter 3.

2.2.1 Data, likelihood and Fisher information

Data and the likelihood function

Once the model for a studied process is formulated and, typically, parameterised, we need to determine the parameter values in order to be able to use the model and characterise the data obtained. The classical way to do this is via the the likelihood function.

Let $Y_1, \ldots, Y_n$ be n independent random variables with probability density functions $f_1(y_1; \boldsymbol{\theta}), \ldots, f_n(y_n; \boldsymbol{\theta})$ depending on a statistical parameter $\boldsymbol{\theta}$ taking values from some set Θ, possibly a subset of $\mathbb{R}^p$. In the case when Y_i is a discrete random variable $f_i(y_i; \boldsymbol{\theta})$ is a function defining the probabilities for Y_i to take value y_i:

$$f_i(y_i; \boldsymbol{\theta}) = \mathbb{P}(Y_i = y_i \,|\, \boldsymbol{\theta}).$$

The joint density of n independent observations $\mathbf{y} = (y_1, \ldots, y_n)$ of the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta}).$$

The *likelihood function* of $\boldsymbol{\theta}$, associated with a vector of random variables $\mathbf{Y}$ is defined up to a positive factor of proportionality as

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) \propto f(\mathbf{y}; \boldsymbol{\theta}) \tag{2.4}$$

The factor of proportionality in (2.4) may depend on $\mathbf{y}$ but not on $\boldsymbol{\theta}$. Thus, the likelihood function is obtained from the joint density $f(\mathbf{y}; \boldsymbol{\theta})$ by viewing it as a function of the unknown parameter $\boldsymbol{\theta}$, for fixed data $\mathbf{y}$. Let us write simply $\mathcal{L}(\boldsymbol{\theta})$ for the likelihood of $\boldsymbol{\theta}$ whenever the context makes clear what data $\mathbf{y}$ is assumed to be available.

In this setting the random variables $Y_1, \ldots, Y_n$ represent a formalisation of the phenomenon which is studied. Their distributions $f_1, \ldots, f_n$, representing parametrised families of distributions, constitute the model which, as we believe, adequately describes the process. A particular value of the parameter further specifies the model. Finding the value $\hat{\boldsymbol{\theta}}^*(y_1, \ldots, y_n)$ of the parameter $\boldsymbol{\theta}$ that maximises the probability of observing the actual data $(y_1, \ldots, y_n)$ given the model and the

parameters forms the basis of the maximum likelihood approach (Edwards (1972)). The value $\hat{\boldsymbol{\theta}}^*(y_1, \ldots, y_n)$ is seen as a realisation of the following statistic

$$\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}),$$

widely known as the *maximum likelihood estimator*.

The maximum likelihood approach was first considered, but not named as we know it now, by Fisher (1921)[6]. This approach is based on the *likelihood principle*, which asserts that all information about the parameter in a sample is contained in the likelihood function, and on the intuitive reasoning that a $\boldsymbol{\theta}_1$ for which $f(\mathbf{y} \,|\, \boldsymbol{\theta}_1)$ is larger than $f(\mathbf{y} \,|\, \boldsymbol{\theta}_2)$ for some $\boldsymbol{\theta}_2$ is more 'likely' to be the true value of the parameter $\boldsymbol{\theta}$.

Persuasive arguments and theoretical development of this approach, including axiomatic construction, were given consequently by Allan Birnbaum (1962). Birnbaum proved that the likelihood principle follows from two simpler and seemingly reasonable principles, the *conditionality principle* and the *sufficiency principle*. To describe these briefly, recall that the following statements are equivalent definitions of the notion of a *sufficient statistic*[7] $T(\mathbf{Y})$ for $\boldsymbol{\theta}$:

1 $f(\mathbf{y} \,|\, T(\mathbf{y}) = t, \boldsymbol{\theta}) = f(\mathbf{y} \,|\, T(\mathbf{Y}) = t)\ \forall \boldsymbol{\theta} \in \Theta\ \forall t \in \operatorname{supp} T$.

2 $f(\boldsymbol{\theta} \,|\, \mathbf{y}, T(\mathbf{y}) = t) = f(\boldsymbol{\theta} \,|\, T(\mathbf{y}) = t)\ \forall t \in \operatorname{supp} T$ (this definition, however, requires a Bayesian framework which is introduced in § 2.2.2).

3 $f(\mathbf{y}; \boldsymbol{\theta}) = h(\mathbf{y}) g(T(\mathbf{y}), \boldsymbol{\theta})$, i.e. the density of $\mathbf{Y}$ can be factorised into a product of a function depending on $\mathbf{y}$ only and a function depending on $\boldsymbol{\theta}$ and $\mathbf{y}$ only through $T(\mathbf{y})$ (Fisher–Neyman factorisation theorem).

The conditionality principle says that if an experiment is chosen by a random process independent of the true value of $\boldsymbol{\theta}$, then only the experiment actually performed is relevant to inferences about $\boldsymbol{\theta}$.

[6]The historical account on the concept of *likelihood* is given in Edwards (1974). See also Lauritzen (1999) for earlier insights on the concept of likelihood in work of the Danish astronomer, actuary, and mathematician T. N. Thiele.

[7]A sufficient statistic can be a vector valued statistic.

The sufficiency principle says that if $T(\mathbf{Y})$ is a *sufficient statistic* for $\boldsymbol{\theta}$, and if in two experiments with outcomes $\mathbf{y}_1$ and $\mathbf{y}_2$ we have $T(\mathbf{y}_1) = T(\mathbf{y}_2)$, then the evidence about $\boldsymbol{\theta}$ given by the two experiments is the same.

Example 2.2.1. *Let $Y_1, \ldots, Y_n$ be independent Bernoulli random variables with parameter $p \in [0, 1]$, i.e.*

$$\mathbb{P}(Y_i = 1) = 1 - \mathbb{P}(Y_i = 0) = p,\ i = 1, \ldots, n.$$

By the Fisher–Neyman factorisation theorem the random variable $T(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is sufficient for p:

$$f(\mathbf{y}; p) = \prod_{1 \le i \le n} p^{y_i}(1-p)^{1-y_i} = p^{T(\mathbf{y})}(1-p)^{n-T(\mathbf{y})},\ \mathbf{y} \in \{0,1\}^n.$$

The statistic $T(\mathbf{Y})$ is no longer sufficient if at least two distributions among those of $Y_1, \ldots, Y_n$ have different parameters.

Obtaining the likelihood function can be complicated when the observations of data are incomplete, i.e. when one observes a possibly vector-valued data summary $T(y_1, \ldots, y_n)$ of the actual data, and $T(Y_1, \ldots, Y_n)$ is not a sufficient statistic.

The Fisher information

Consider a parametric family of distributions with densities $f(\mathbf{y}; \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Let $\hat{\theta}$ be an unbiased estimator for θ. The result known as the *Cramér–Rao lower bound* gives us the minimum variance that can be expected from $\hat{\theta}$, that is to say the maximum precision on estimating θ when using $\hat{\theta}$:

$$\operatorname{var}\hat{\theta} \ge 1/I(\theta),$$

where

$$I(\theta) = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(\mathbf{Y}; \theta)\right)^2 \,\middle|\, \theta\right] \tag{2.5}$$

is the so called *Fisher information function*. The corresponding theorem, in a more general form, was first proven by Fréchet (1943) and then by Rao (1945) and Cramér (1946). An informal derivation of the Fisher information function can be found in Frieden (2004).

Example 2.2.2. *Let $Y \sim Bin(n, e^{-\theta r})$, where n and r are known and fixed, and $0 < \theta < 1$. Introducing $p = e^{-\theta r}$ and considering $Y \sim Bin(n, p)$ results in the following:*

$$f(y; p) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$\frac{\partial}{\partial \theta} \log f(y; p) = \frac{y}{p} - \frac{n-y}{1-p},\ 0 < \theta < 1.$$

The Fisher information (the amount of information on p) is as follows[8]*:*

$$I(p) = \mathbb{E}\left[\frac{Y}{p} - \frac{n-Y}{1-p}\right] = \frac{n}{p(1-p)}.$$

The amount of information on θ can be obtained by dividing $I(p)$ by $\left(\frac{d\theta}{dp}\right)^2$, where $\theta = -\frac{1}{y}\log p$:

$$I(\theta) = ny^2 \frac{e^{-\theta y}}{1 - e^{-\theta y}}.$$

Generally, certain regularity conditions should be met in order to define the Fisher information function (Zacks (1981, pp. 103, 237)). Particularly, if the following regularity condition is met

$$\int \frac{\mathrm{d}}{\mathrm{d}\theta} f(\mathbf{y}; \theta)\, \mathrm{d}\mathbf{y} = 0,$$

then the Fisher information (2.5) may also be written as follows:

$$I(\theta) = -\mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(\mathbf{Y}; \theta) \,\middle|\, \theta\right]. \tag{2.6}$$

Thus, being the expected value of the second derivative of the log-likelihood function $\log f$, the Fisher information may be seen as a measure of the 'sharpness' of this (random!) function at a given point θ.

The Fisher information is additive: the information yielded by two independent experiments $\mathbf{X}$ and $\mathbf{Y}$ is the sum of the information from each experiment separately:

$$I_{\mathbf{X},\mathbf{Y}}(\theta) = I_{\mathbf{X}}(\theta) + I_{\mathbf{Y}}(\theta). \tag{2.7}$$

[8]Notice also that $I(p) = 1/\mathrm{Var}\,[Y/n]$.

When the model parameter is a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$, then the Fisher information takes the form of an $m \times m$ matrix $I(\boldsymbol{\theta})$ with the ij-element being

$$I_{ij}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\mathbf{Y}; \boldsymbol{\theta}) \,\middle|\, \boldsymbol{\theta} \right]. \tag{2.8}$$

2.2.2 Bayesian concept

In a Bayesian framework we quantify our beliefs about relative likelihood of different parameter values using a prior probability distribution on the parameter space. The data, and specifically the likelihood function, are then used to update the prior distribution to a posterior distribution using the Bayes theorem.

The Bayesian method can be briefly represented as comprising the following principal steps (O'Hagan (1994)):

1 *Likelihood*. Obtain the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$. This step describes the process giving rise to the data $\mathbf{y}$ in terms of the unknown parameter $\boldsymbol{\theta}$.

2 *Prior*. Formulate the prior density $\pi(\boldsymbol{\theta})$. The prior distribution expresses what is known or believed to be known about $\boldsymbol{\theta}$ prior to observing the new data $\mathbf{y}$.

3 *Posterior*. Apply Bayes' theorem to derive the posterior density $\pi(\boldsymbol{\theta} \,|\, \mathbf{y})$. This will now express what is known about the model parameter $\boldsymbol{\theta}$ after observing the data $\mathbf{y}$.

4 *Inference*. Derive appropriate inference statements from the posterior distribution. These statements may include specific inferences such as point estimates, interval estimates, probability of hypotheses or assessment of how different the posterior distribution is from the prior distribution.

Bayes' Theorem

Inference concerning $\boldsymbol{\theta}$ is based on its posterior distribution, given by Bayes' Theorem:

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta})}{\int\limits_{\Theta} \mathcal{L}(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta}} \propto \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta}). \tag{2.9}$$

The integral in the denominator

$$\Phi(\mathbf{y}) := \int_{\Theta} \mathcal{L}(\mathbf{y}; \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{2.10}$$

is the marginal distribution of $\mathbf{y}$ derived from the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$. This distribution is called the *prior predictive distribution for y* (Leonard and Hsu (1999)), but is also known as the *evidence* or *marginal likelihood* (Zacks (1981)) in cases when the likelihood is integrated over some of the model parameters:

$$\Phi(\mathbf{y} \mid \boldsymbol{\delta}) = \int_{\Theta} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\delta}; \mathbf{y}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\theta}.$$

The right-hand side of (2.9) indicates that $\Phi(\mathbf{y})$ is essentially a normalising constant in evaluating $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. It is the calculation of this function that traditionally represents a severe obstacle while performing the Bayesian analysis. However, the calculation of the normalising constant can be often avoided using Markov Chain Monte Carlo (MCMC) methods which permit sampling from the posterior without evaluating this marginal distribution (Section 2.3).

Notice also that if the family of distributions from which $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ stems admits a sufficient statistic $T(\mathbf{Y})$ then for any prior distribution $\pi(\boldsymbol{\theta})$, the posterior distribution is a function of $T(\mathbf{Y})$, and can be determined from the distribution of $T(\mathbf{Y})$ under $\boldsymbol{\theta}$. Indeed, if $T(\mathbf{Y})$ is sufficient for $\boldsymbol{\theta}$ under the model $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ then by the Neyman–Fisher factorisation theorem $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) = h(\mathbf{y}) g(T(\mathbf{y}), \boldsymbol{\theta})$, so that the posterior density

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{g(T(\mathbf{y}), \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\int_{\Theta} g(T(\mathbf{y}), \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}$$

is a function of $T(\mathbf{Y})$. It follows that the conditional density of $\boldsymbol{\theta}$ given $\{T(\mathbf{Y}) = t\}$ coincides with $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ on the sets $\{\mathbf{y} : T(\mathbf{y}) = t\}$ for all $t \in \operatorname{supp} T$.

Bayes' Theorem can be applied sequentially, providing the basis for a Bayesian analysis under sequential experimentation. For instance, suppose that we have

observed two independent data samples $\mathbf{y}_1$ and $\mathbf{y}_2$. Then

$$\begin{aligned}\pi(\boldsymbol{\theta} \,|\, \mathbf{y}_1, \mathbf{y}_2) &\propto f(\mathbf{y}_1, \mathbf{y}_2 \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_2)\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_1)\pi(\boldsymbol{\theta}) \\ &\propto \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_2)\pi(\boldsymbol{\theta} \,|\, \mathbf{y}_1),\end{aligned}$$

that is, in order to obtain the posterior for the full data set $(\mathbf{y}_1, \mathbf{y}_2)$, one can first evaluate $\pi(\boldsymbol{\theta} \,|\, \mathbf{y}_1)$ and then use it as the prior for $\mathbf{y}_2$. This forms a natural setting for performing a *sequential Bayesian analysis*. If the data are incomplete, this will only reflect on evaluation of the likelihood, and the whole construction will be similar.

Prior distribution

The prior distribution is absent from classical methods, but it is an integral part of Bayesian statistics. It represents the knowledge of an investigator about the model parameter $\boldsymbol{\theta}$ before seeing the data. This knowledge, however, takes into account previous experience the investigator might have had, applying the model for another data set or, when no reliable prior concerning the model parameter exists, results in specifying a *non-informative* prior for $\boldsymbol{\theta}$.

Non-informative priors

In the case when the parameter space Θ is of finite measure (length, area, volume), one might take a uniform distribution over Θ to serve as a *'non-informative'* prior—such distribution will contain no information about $\boldsymbol{\theta}$ except its range of values in the sense that it does not favour one value of $\boldsymbol{\theta}$ to another.

For unbounded parameter spaces things are not that straightforward. For instance, when $\Theta \equiv \mathbb{R}_+$ a distribution $\pi(\theta) = c \in \mathbb{R}_+$ is clearly improper (it is not a probability distribution). However, Bayesian analysis is still possible whenever the prior predictive distribution is proper, i.e. if $\int_\Theta \mathcal{L}(\theta; \mathbf{y})\, \mathrm{d}\theta < \infty$.

The problem with uniform distributions as 'non-informative' priors is that a uniform prior is not invariant under reparametrisation of the model, that is to say

a uniform prior will be converted to a non-uniform one, and hence informative, by reparametrising the model in hand. One approach that overcomes this difficulty is the so called *Jeffreys prior*: Jeffreys (1961) justified the use of the following prior

$$\pi(\theta) \propto \left| I(\theta)^{1/2} \right|$$

on grounds of its invariance properties. Here $I(\theta)$ is the Fisher information and the prior $\pi(\theta)$ is such that if $\omega = \phi(\theta)$ is a one-to-one transformation, then $\pi(\theta \,|\, \mathbf{y}) = \pi(\omega \,|\, \mathbf{y})$ for every $\mathbf{y}$. This prior is often improper, since the square root of $|I(\theta)|$ is not always an integrable function. Lindley (1961) showed that $\pi(\theta) \propto \left|I(\theta)^{1/2}\right|$ leads to the maximum expected information gain using entropy-based measure—this being the very reason why the Jeffreys prior is called *non-informative* and why uniform priors are better to refer to simply as *flat* priors (see Irony and Singpurwalla (1997) for an interesting discussion with J. Bernardo on the topic and Berger, Bernardo and Mendoza (1989) for mathematical foundation of deriving non-informative priors for Bayesian inference via maximisation of information measures).

Undoubtedly, the choice of a prior distribution is a critical step of Bayesian procedures. However, the difficulty in selecting the prior distribution is not only in choosing the way in which it represents the prior knowledge on the model parameter(s), but also in the fact that when choosing it one might also need to find a balance between an improvement in the subsequent analytical treatment of the problem and the subjective determination of the prior distribution (and hence to ignore part of the prior information). The reader is referred to Robert (2007, Chapter 3) for a further discussion on the choice of prior distributions.

Inference from posterior

Having obtained the posterior distribution, one can use the following standard tools in order to summarise the results:

1 plot of the density function: this will visualise the current state of our knowledge;

2 numerical summaries of the posterior and point estimation: mean, median, mode and variance; in the case of a flat prior the mode of the posterior distribution will coincide with the maximum likelihood estimate;

3 interval estimation: this involves determination of various sorts of credibility intervals or sets.

For an excellent recent review of the methodology of Bayesian statistics the reader is invited to refer to Bernardo (2003). A range of arguments for Bayesian implementation and use of the likelihood function through Bayesian analysis is presented in Berger and Wolpert (1988, Chapter 3, § 5.3).

2.3 Monte Carlo methods and Markov Chain Monte Carlo

Monte Carlo methods have become standard techniques and an integral part of the arsenal of researchers and practitioners whose interests belong to many different areas of study. Applications of Monte Carlo methods can be found in various fields: operational research (including queueing and network systems analysis and numerical analysis), reliability theory, statistics, finance, to name just a few mathematical areas. Allowing one to model complex nondeterministic time-space evolution, epidemics and social phenomena, these methods have also found wide applications in biological and social sciences.

2.3.1 Monte Carlo methods

Monte Carlo methods are experimental modelling methods. Madras (2002) categorised Monte Carlo experiments into the following two broad classes:

1 direct simulation of a naturally random system or object;

2 addition of artificial randomness to a system of study, followed by simulation of the new system.

Monte Carlo methods are used in this thesis for purposes falling into each of these groups. Estimation of parametric integrals of the form

$$I(d) = \int u(x, d)p(x)\,\mathrm{d}x, \tag{2.11}$$

where $p(x)$ is a probability distribution, will be made by sampling from this distribution and approximating the integral by the corresponding *ergodic average*:

$$I^{(M)}(d) := \frac{1}{M}\sum_{i=1}^{M} u(x_i, d),\ x_i \sim p(x). \tag{2.12}$$

The estimator that gives birth to this point estimate of $I(d)$ is unbiased and almost surely converges to $I(d)$, as $M \to \infty$, by the strong law of large numbers.

By the Central Limit Theorem, an approximate 95% confidence interval for $I(d)$ (for any fixed d) is

$$\left[I^{(M)}(d) - 1.96\frac{\sigma}{\sqrt{M}}, I^{(M)}(d) + 1.96\frac{\sigma}{\sqrt{M}}\right],$$

where σ is the standard deviation of the random variable $u(X, d)$ with X having the density p. The standard deviation σ is often unknown or difficult to calculate, but it can be approximated by the sample standard deviation:

$$s_M = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(u(x_i, d) - I^{(M)}(d)\right)^2}.$$

It is due to the Central Limit Theorem that the accuracy of estimates calculated by Monte Carlo simulation is proportional to $M^{-1/2}$, where M is the size of the sample used. In general, it is true for all Monte Carlo methods that the absolute error of the calculation is inversely proportional to the square root of the computational effort spent. This means, that in order to increase the precision of calculations by a factor of 10, one needs to increase the computational effort (sample size) by a factor of 100. This, in turn, means that Monte Carlo is perhaps not the best choice when one wants to achieve high precision of estimation. It might, however, be one of a very few working methods to tackle the problem in hand, if not the only one. This is particularly true for high-dimensional problems.

Finally, Monte Carlo simulation is also used in this thesis in order to obtain realisations of random graphs. The simulation technique takes its simplest form

in this context: for example, in order to obtain a realisation $G = (\mathcal{D}, E)$ of an unoriented random graph on the fixed set $\mathcal{D}$ of n given nodes $x_1, \ldots, x_n$ with an edge-probability function $p(r, \boldsymbol{\theta})$, as defined in § 1.2.1 by (1.1-1.2), one obtains first the realisations $\{\tilde{U}_{ij}\}_{1 \le i < j \le n}$ of $n(n-1)/2$ independent standard uniform random variables. The edge set E is formed then by all such unoriented pairs (x_i, x_j) for which it is true that $\tilde{U}_{ij} \le p(r(x_i, x_j), \boldsymbol{\theta})$, $1 \le i < j \le n$. Note that the value of $\boldsymbol{\theta}$ is assumed to be fixed prior to obtaining any realisation(s) of such a random graph.

2.3.2 Markov Chain Monte Carlo

The main idea behind Markov Chain Monte Carlo method is to construct a Markov chain, whose unique limiting distribution will coincide with the distribution of interest. If one succeeds in doing so, one can run the corresponding Markov chain for a sufficiently long period of time and then take a sequence of its consequent states, this being an approximate sample from the *target* distribution, which can be a very complex distribution. Constructing Markov chains is particularly helpful while performing a Bayesian analysis, in which case the target distribution is the posterior density.

Discrete time irreducible and aperiodic Markov chains

Markov chains are discrete-time stochastic processes with the *Markov property*: given the present state, the future and past states are independent. Let us consider first a Markov chain $X_0, X_1, X_2, \ldots$, where each X_i takes values in a countable state space $\mathcal{S}$. The k-step transition probabilities are

$$p_{i,j}^{(k)} = \mathbb{P}\left(X_{t+k} = j \,|\, X_t = i\right) \ (i, j \in \mathcal{S},\, k = 0, 1, 2, \ldots.)$$

The transition probability matrix is $P = (p_{ij} \equiv p_{ij}^{(1)})$; it is a basic fact that $p_{ij}^{(k)}$ is the ij^{th} entry of P^k.

Two states i and j are said to be *communicating* if there exists n such that $p_{ij}^{(n)} > 0$. Thus, the state space $\mathcal{S}$ of a Markov chain splits into subsets (communicating classes) that contain communicating states. A Markov chain is said to be

irreducible if the chain can eventually get from each state to every other state, i.e., for every $i, j \in \mathcal{S}$ there exists $k_{ij} \geq 0$ such that $p_{ij}^{(k_{ij})} > 0$. The state space $\mathcal{S}$ of an irreducible Markov chain is a single communicating class.

A state i has period D_i if any return to the state i occurs in a multiple of D_i time steps:

$$D_i := \gcd\{n : p_{ii}^{(n)} > 0\}.$$

If $D_i = 1$, then the state i is said to be aperiodic. Otherwise, the state i is said to be periodic with period D_i. It can be shown that every state in a communicating class has the same period.

An irreducible chain is said to be aperiodic if the period of one of its states (equivalently, all of its states) is unity.

Limiting behaviour of Markov chains with countable state spaces

By one of the fundamental theorems about the long-run behaviour of Markov chains, an aperiodic irreducible Markov chain exhibits stochastically predictable limiting behaviour, which does not depend on the initial state of the chain (Madras (2002, p. 54)). In order to be more precise let us formulate the corresponding well-known theorem.

Theorem 2.3.1. *(Theorem 4.2 in Madras (2002)) Consider an aperiodic irreducible Markov chain with state space $\mathcal{S}$. For every $i, j \in \mathcal{S}$, the limit $\pi_j := \lim_{k\to\infty} p_{i,j}^{(k)}$ exists and is independent of i. Furthermore:*

1 If $\mathcal{S}$ is finite, then

$$\sum_{j\in\mathcal{S}} \pi_j = 1 \quad \textit{and} \quad \sum_{i\in\mathcal{S}} \pi_i p_{i,j} = \pi_j$$

for every $j \in \mathcal{S}$. That is, if we write $\boldsymbol{\pi}$ to denote the row vector whose entries are π_i, then $\boldsymbol{\pi}P = \boldsymbol{\pi}$. Moreover, the only solution of the following system of equations

$$\begin{cases} \mathbf{v}P = \mathbf{v} \\ \sum_{i\in\mathcal{S}} v_i = 1 \\ v_i \geq 0, i \in \mathcal{S} \end{cases} \tag{2.13}$$

is $\mathbf{v} = \boldsymbol{\pi}$.

2 *If* $\mathcal{S}$ *is countably infinite, then there are two possibilities: either (i)* $\pi_j = 0$ *for every* j *and (2.13) has no solutions or (ii)* $\boldsymbol{\pi}$ *satisfies (2.13) and it is the only solution of (2.13).*

Thus, with the exception that is described by the case 2(i) of this theorem, there exists such a distribution $\boldsymbol{\pi}$ that, provided the initial chain state X_0 has distribution $\boldsymbol{\pi}$, the distribution of X_k is exactly $\boldsymbol{\pi}$ for every time step $k = 1, 2, \ldots$. Moreover, the following interpretations can be attributed to $\boldsymbol{\pi}$:

- $\pi_i \approx \mathbb{P}(X_k = i)$ for large k, independent of the distribution of X_0.
- π_i is the long-run fraction of time the system spends in state i:

$$\mathbb{P}\left(\pi_i = \lim_{k\to\infty} \#\{n : X_n = i, n = 1, \ldots, k\}/k\right) = 1.$$

The limiting distribution $\boldsymbol{\pi}$ of a Markov chain is referred to as the *equilibrium* or *invariant* or *steady-state* or *stationary* distribution.

Among Markov chains exhibiting stationary behaviour there is an important class of chains that show the so called time-reversibility: a stationary Markov chain is said to be *reversible* if its transition matrix $P = (p_{ij})$ and stationary distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots,)$ satisfy the *detailed balance equations*:

$$\pi_i p_{i,j} = \pi_j p_{j,i}, \ \forall i, j \in \mathcal{S}. \tag{2.14}$$

Conversely, it is easy to check that any irreducible Markov chain satisfying the detailed balance equations for some $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$, that proves to be a probability distribution over $\mathcal{S}$, is stationary and the equilibrium of such Markov chain is described by $\boldsymbol{\pi}$.

Markov chains with continuous state spaces

In the case when a Markov chain has a continuous state space $\mathcal{X}$, its transitions from a current state X_n to a new state X_{n+1}, both from $\mathcal{X}$, are described by using a transition density $\mathcal{K}(X \mid X_n)$ that is only dependent upon X_n. Density $\mathcal{K}$, uniquely

describing the dynamics of the chain given its initial state, is also referred to as the *transition kernel* of the chain.

Under conditions that are fairly similar to those for Markov chains with discrete state spaces, aperiodic and irreducible Markov chains with continuous state spaces exhibit convergence to a stationary distribution $\pi(\cdot)$:

$$\mathbb{P}\left(X_n \in A\right) \to \int\limits_A \pi(x)dx \; \forall A \in \mathcal{X}, \text{ as } n \to \infty.$$

The stationary distribution is unique when the entire state space can be reasonably explored, that is to say, when any set of states can be reached from any other set of states within a finite number of transitions.

When we wish to construct a Markov chain with a given stationary distribution $\pi^*(\cdot)$, known may be only up to a constant, one way to achieve it is to find a transition kernel that would satisfy the detailed balance equations with respect to $\pi^*(\cdot)$:

$$\pi^*(X)\mathcal{K}(X' \,|\, X) = \pi^*(X')\mathcal{K}(X \,|\, X'), \; \forall X, X' \in \mathcal{X}. \tag{2.15}$$

The two, perhaps most famous and common methods of constructing time reversible Markov chains whose stationary distributions match the target distribution are described here: these are the Metropolis–Hastings algorithm and the Gibbs sampler.

The Metropolis–Hastings algorithm

This algorithm, described first in a special case by Metropolis et al (1953) and generalised later by Hastings (1970), generates approximate samples from a probability density $g(x)$ known up to a constant. Given a conditional density $q(\cdot \,|\, x)$ the algorithm generates a Markov chain $(X_n)_{n=0}^{\infty}$ with an arbitrary initial state X_0 and stationary distribution coinciding with $g(x)$ by updating the current state from X_n to X_{n+1} $(n = 1, 2, \ldots)$ via the following steps:

1 Generate $\xi \sim q(\xi \,|\, X_n)$.

2 Evaluate $\alpha(X_n \to \xi) = \min\left\{1, \frac{g(\xi)}{g(X_n)} \frac{q(X_n \,|\, \xi)}{q(\xi \,|\, X_n)}\right\}$.

3 Set

$$X_{n+1} = \begin{cases} \xi & \text{with probability } \alpha(X_n \to \xi), \\ X_n & \text{with probability} 1 - \alpha(X_n \to \xi). \end{cases}$$

Here the target distribution is $g(x)$, the quantity α is called the *acceptance probability* or *acceptance ratio*, whereas the distribution $q(\cdot \mid x)$, which is used to propose updates of the chain's state, is called the *proposal distribution*. When the support of the proposal distribution $q(\cdot \mid \cdot)$ includes the chain's state space $\mathcal{X}$, the transition kernel is as follows:

$$\mathcal{K}(X \mid X_n) = \alpha(X_n \to X) q(X \mid X_n) + [1 - \zeta(X_n)]\delta(X - X_n), \tag{2.16}$$

where $\zeta(X_n) := \int_{\mathcal{X}} \alpha(X_n \to x) q(x \mid X_n) \, \mathrm{d}x$ is the expected probability of accepting a new point while being in the state X_n, and $\delta(X - X_n)$ is the Dirac delta function that assigns a unit mass to the state X_n (see Appendix B).

The described algorithm ensures the correct stationary distribution for the corresponding Markov chain as long as this chain is irreducible and aperiodic: it is straightforward to check that the chain (2.16) satisfies the detailed balance equations with respect to $g(x)$.

The rate of convergence of the chain to its stationary distribution depends on the choice of the proposal distribution. Among the most basic, but somewhat 'universal' types of proposals, there are the following two types:

1 *independent proposals*

This family consists of the proposal distributions $q(\tilde{x} \mid x)$ which do not depend on x: $q(\tilde{x} \mid x) = f(\tilde{x})$.

2 *symmetric random walk proposals*

This family comprises the proposal distributions $q(\tilde{x} \mid x)$ which are symmetric about θ: $q(\tilde{x} \mid x) = f(|\tilde{x} - x|)$. For such proposals the acceptance probability simply becomes

$$\alpha(X_n \to \xi) = \min \left\{ 1, \frac{g(\xi)}{g(X_n)} \right\},$$

and, clearly, the corresponding Markov chain will tend to remain longer in the points with higher values of the target distribution, while the points with

lower probability will be visited less often. Markov chains with a symmetric proposal are known as *Metropolis random walks*.

Those authors who believe that main ideas deserve short names refer to Metropolis–Hastings algorithm simply as Metropolis algorithm (e.g. MacKay (2003, p. 366)).

The Gibbs Sampler

The *Gibbs sampler* is a special case of the Metropolis–Hastings algorithm when every proposing state is always accepted ($\alpha \equiv 1$). It originated in the seminal work of Geman and Geman (1984). The idea behind this sampling method is in updating the states of the chain in an element-wise way, when the states are some multidimensional objects $\mathbf{X}_0, \mathbf{X}_1, \ldots$, i.e.

$$\mathbf{X}_i = (X_1^{(i)}, \ldots, X_k^{(i)}).$$

Thus, if one needs to sample from a multivariate distribution $g_{\mathbf{X}}(x_1, \ldots, x_k)$, one can use the corresponding one-dimensional full conditional distributions

$$g_1(x_1 \,|\, \cdot), g_2(x_2 \,|\, \cdot), \ldots, g_k(x_k \,|\, \cdot)$$

as follows: given the current state of the chain $\mathbf{X}_n = (X_1^{(n)}, \ldots, X_k^{(n)})$ the next state of the chain is simulated by sampling

$$X_i^{(n+1)} \sim g_i(x_i) \equiv g(x_i \,|\, X_1^{(n)}, X_2^{(n)}, \ldots, X_{i-1}^{(n)}, X_{i+1}^{(n)}, \ldots, X_k^{(n)}),\ i = 1, \ldots, k,$$

and letting $\mathbf{X}_{n+1} := (X_1^{(n)}, X_2^{(n)}, \ldots, X_{i-1}^{(n)}, X_i^{(n+1)}, X_{i+1}^{(n)}, \ldots, X_k^{(n)})$.

There are variations of the Gibbs sampler in which the order of the components' updates is either systematic or random. Moreover, the full conditional distributions need not be one-dimensional and some updates in the Gibbs sampler can be replaced by Metropolis–Hastings steps.

A comprehensive up-to-date review on the topic of Monte Carlo and MCMC methods is provided by Murray (2007). A brief discussion on practical issues related to the output of a chain and its statistical analysis follows.

Implementation

A correctly constructed Markov chain will have as its limiting (stationary) distribution the desired target distribution. However, one should bear in mind that the method is based on asymptotic results, in general the target distribution being achieved only in the limit. The output of a chain should be dealt with carefully therefore. The following are important questions to be asked about the behaviour of a stationary Markov chain:

1 Starting from which step of the chain states' updates may one consider the subsequent updates to form an approximate sample from the target distribution? In other words, how many initial observations shall we discard before starting the sampling itself? The answer to this question obviously relates to the rate of convergence of a particular chain.

2 What to do when the output of a sampler has a complicated dependence structure, and particularly when adjacent steps are highly correlated?

The former question naturally gave rise to the notion of *burn-in* period[9], this being the number of steps one should discard before obtaining an approximate sample from the target distribution. The most simple recipe for the latter question is to keep one sample of chain out of t iterations, and thus *'thinning'* the output of the chain[10]. Not surprisingly, these qualitative solutions are based on empirical evidence: the burn-in period can be estimated from the plot of the sampled values for each variable in the chain *versus* the number of iterations (*trace plot*), and t for thinning depends on the dependence structure and level of correlation in assessment of which, for example, a covariogram or correlogram may be helpful. Finally, in making decisions on how to tackle these problems in a particular situation, the cost of sampling should also be taken into account.

The reader is referred to Levin, Peres and Wilmer (2009) for the most recent and comprehensive account on the subject: the authors of this textbook develop the

[9]Or *warm-up* period, or *mixing time*.

[10]The correlation between adjacent steps should be assessed, since an unnecessary thinning might only make the variance of the output worse (see Murray (2007) and Geyer (1992)).

results on the rate of convergence of a Markov chain to the stationary distribution as a function of the size and geometry of the state space.

Chapter 3

Utility-Based Optimal Designs within the Bayesian Framework

3.1 Introduction: from locally D-optimum to utility-based Bayesian designs

3.1.1 Toy examples: three and four nodes

Consider a graph on three vertices with edges of non-negative weights r_1, r_2 and r_3 as in Figure 3.1 and form a random graph on these vertices in which each of the edges is present independently of any other edge with probability $p(r_k, \theta) = e^{-\theta r_k}$, $\theta \in \mathbb{R}_+$, $k = 1, 2, 3$. The larger the weights r_1, r_2, r_3 are, the larger the chances are to observe no edges at all in a realisation of this random graph. Likewise, the smaller these weights are, the larger the chances are to see all three edges present in a realisation of the random graph. Suppose θ is unknown, and we want to make inference on this model parameter. What are the optimal values for r_1, r_2 and r_3 then?

One approach would be to maximise the Fisher information function (see Example 2.2.2)

$$I(\theta; \boldsymbol{r}) = \sum_{k=1}^{3} r_k^2 \frac{e^{-\theta r_k}}{1 - e^{-\theta r_k}} \tag{3.1}$$

with respect to $\mathbf{r} = (r_1, r_2, r_3) \in \mathbb{R}_+^3$, since it is the Fisher information that, in a

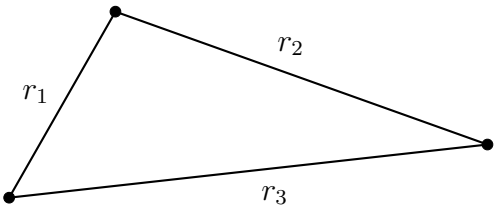


Figure 3.1: A graph on three nodes with edges of weights r_1, r_2, r_3.

sense, is measuring the amount of information that our three-node random graph carries about the model parameter θ upon which the likelihood function depends (in the sense discussed in § 2.2.1). Maximising (3.1) we find[1] that the optimal choice of the weights is as follows:

$$r_1^* = r_2^* = r_3^* \approx 1.6/\theta. \tag{3.2}$$

Indeed, each of the three terms in (3.1) is independent of the two others, and is maximised at the point equal approximately to $1.6/\theta$ (see Appendix A for details). Hence, any other triple of values of r_1, r_2 and r_3 than $\mathbf{r}^* = (r_1^*, r_2^*, r_3^*)$ will only decrease the Fisher information.

This situation can be easily generalised for the case of n independent pairs of vertices or star topology. Consider the following example.

Example 3.1.1. *(based on Example 3.11 in Zacks (1981)) Suppose that n systems $S_1, \ldots, S_n$ operate in parallel and independently. The lifetime T_i of the system S_i is exponentially distributed, $T_i \sim Exp(\theta)$, and assume that $T_1, \ldots, T_n$ are independent random variables. We can check the status of the system S_i at the time instance r_i, $i = 1, \ldots, n$. What is the optimal set of times $r_1, \ldots, r_n$ at which the systems should be approached and examined in order to maximise the amount of information on the 'ageing rate' θ? Modelling the status ('operating' or 'broken') of the system S_i at the time r_i by a Bernoulli random variable with parameter $e^{-\theta r_i}$, and maximising the Fisher information for this model*

$$I(\theta; \boldsymbol{r}) = \sum_{k=1}^{n} r_k^2 \frac{e^{-\theta r_k}}{1 - e^{-\theta r_k}}, \tag{3.3}$$

by maximising its summands separately we find that the optimal set of observation

[1] See Appendix A.2

times is $\mathbf{r}^* = (r_1^*, \ldots, r_n^*)$*, where*

$$r_1^* = \ldots = r_n^* \approx 1.6/\theta.$$

Thus, using the Fisher information function as an information measure and allowing for observation times to be chosen individually for each of the considered systems (Figure 3.2), we found that the optimal times should all be equal and no different from the optimal time for the case when a single observation is only allowed and the number of the broken devices is observed (Example 2.2.2).

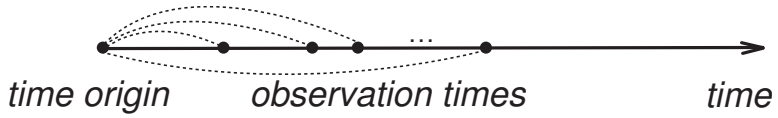


Figure 3.2: Observation times diagram: solid line is the time axis, and the dotted lines are possible edges of the graph.

Notice also that the corresponding random graph presented in Figure 3.2 is topologically equivalent to the star configuration (Figure 1.1, star), with the central node corresponding to the time origin.

If the vertices of the three-node graph are considered to be elements of a metric space, and hence r_1, r_2, and r_3 are distances, then 'the optimal arrangement' is equilateral and it coincides with the one given by (3.2). Indeed, when r_1, r_2, and r_3 are distances, the maximisation of $I(\theta; \mathbf{r})$, $\mathbf{r} = (r_1, r_2, r_3) \in \mathbb{R}^3_+$, should be made in conjunction with the *triangle inequality* constraints:

$$\begin{cases} r_1 \leq r_2 + r_3, \\ r_2 \leq r_1 + r_3, \\ r_3 \leq r_1 + r_2. \end{cases} \tag{3.4}$$

However, since the solution $\mathbf{r}^*$ of the corresponding optimisation problem without the triangle constraints satisfies them, it is also the solution of the optimisation problem under the constraints (3.4).

The optimality based on the maximisation of the Fisher information function $I(\theta; \mathbf{r})$, or, more generally, of the determinant of the Fisher information matrix[2]

[2]whose elements are defined by (2.8), p. 27.

$\det I(\boldsymbol{\theta};\mathbf{r})$ is widely known as *D-optimality*. D-optimality is a particular case of a more general optimality criterion based on the maximisation of a suitable scalar functional $\Psi(I(\boldsymbol{\theta};\mathbf{r}))$ of the Fisher information matrix that in particular permits to arrive at a complete ordering of candidate designs. (For D-optimal designs the functional Ψ is a logarithmic transformation: $\Psi(I(\boldsymbol{\theta};\mathbf{r})) = \log\det I(\boldsymbol{\theta};\mathbf{r})$). The choice of the functional Ψ gives a great variety of design criteria distinguishing within an alphabetical nomenclature (e.g. A-, D-, E-, G, I-, L-optimality) that originated in work of Kiefer (1959) followed by an important paper of Kiefer and Wolfowitz (1960) containing the first equivalence theorem (more on this in Atkinson and Donev (1992) and Ryan (2007)).

The classical interpretation of D-optimum designs is simple: they minimise the volume of the confidence ellipsoid (or the length of the confidence interval in the case of a univariate parameter), and hence are relevant to the inference problem. However, D-optimal designs have serious drawbacks which have been intensively discussed in the literature. The toy examples considered above clearly exhibit some of them causing the following concerns:

1 **The design is a function of the model parameter estimate.**

Indeed, the optimal edge weights (3.2) depend on the true value of the model parameter. Although estimates of the parameter(s) can be obtained, it is still difficult to accept the fact that the design that has to be chosen prior to performing an experiment in order to make inference on the model parameter(s) is strongly dependent upon the knowledge (or a good guess!) of its true value.

Müller (2007) refers to this problem as a 'circular problem': "the information matrix (function) depends upon the true values of the model parameter and not only upon the design variable, which evidently leads to a circular problem: for finding a design that estimates the model parameter efficiently it is required to know its value in advance". Khuri (1984) attributes the following words of irony to William G. Cochran:

> *"You tell me the value of θ and I promise to design the best experiment for estimating θ".*

It is difficult therefore to adopt the D-optimal design as a *bona fide* practical design for the purpose of making inference—what such a design would tell us, for instance, in the context of the toy three-node weighted random graph example, is that were we to set the edge weights too different from $1.6/\theta^*$, where θ^* is the true value of θ, we would lose a considerable amount of information.

2 **Symmetry in the optimal design.**

The solutions to both constrained (planarity conditions) and unconstrained three-node optimal random graph problems considered above suggest that all the edges should be of equal weights. It is not clear, however, why this should be the case: one might intuitively expect the optimal weights to be different as long as one has such a freedom in choosing the edge weights of the random graph in order to maximally increase the information gain on the model parameter (note that this observation is not specific to the choice of the edge function in the considered examples).

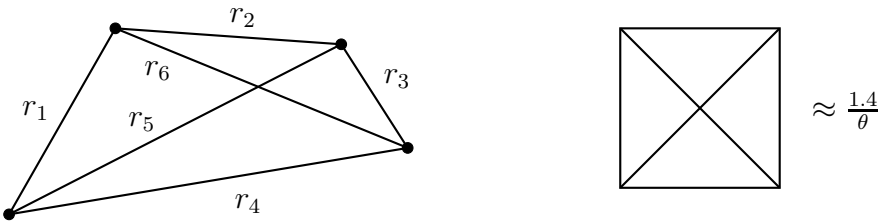


Figure 3.3: Left: A random graph on four nodes with edges of weights $r_1, r_2, r_3, r_4, r_5, r_6$. Right: The optimal random graph on four nodes in plane is a square.

In the case of four nodes the optimal weights are all equal for the unconstrained problem (Example 3.1.1), and among all planar configurations, as in Figure 3.3 (left), the optimal design is an arrangement of vertices of a square with the side's length approximately equal $1.4/\theta$, as in Figure 3.3 (right). The author does not have analytical proof for the latter result: the claim is based on the numerical maximisation of the corresponding Fisher information function. This function was evaluated on the set of four vertex

configurations, with one vertex fixed and three other vertices placed at the nodes of a square grid of small spacing.

3 **D-optimal designs are not invariant under reparametrisation of the model parameter** Although scale invariant, D-optimal designs are not invariant under general model reparametrisation. This has always been considered as one of the most serious drawbacks of D-optimality (e.g. Firth and Hinde (1997b)).

4 **Is there place for using a prior knowledge?** The D-optimum designs rely on a single prior point estimate of the parameter. Can, however, the Bayesian approach be integrated into the mentioned alphabetical optimal design hierarchy? The answer is yes, and an alternative that involves a prior knowledge to the optimality based on $I(\boldsymbol{\theta}; \mathbf{r})$, is simply to maximise the average of a monotone function of a determinant of the Fisher matrix with respect to the prior distribution:

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} \int_{\Theta} \Psi(I(\boldsymbol{\theta}; \mathbf{r}))\pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}. \tag{3.5}$$

(Atkinson and Donev (1992), Atkinson et al (1993), Chaloner and Verdinelli (1995)). Such an approach can also solve two more problems already mentioned above: (i) whatever the choice of Ψ, the optimal design is independent on the model parameter, and (ii) if $\Psi(\cdot) = \log\det\cdot$, then the optimal design is 'parameter neutral', that is reparametrisation invariant.

A similar, though prior-free approach is considered by Firth and Hinde (1997a, 1997b) who suggested to maximise

$$J(\mathbf{r}) = \int_{\Theta} (\det I(\boldsymbol{\theta}; \mathbf{r}))^{1/2}\, \mathrm{d}\boldsymbol{\theta}, \tag{3.6}$$

and thus to avoid dependence on $\boldsymbol{\theta}$ and also to achieve invariance to the choice of parametrisation used to represent the model. These authors also noticed that designs maximising (3.6) are actually 'pseudo-Bayesian' since the information quantity used does not involve a proper prior.

Thus, we have listed enough reasons to turn to a more suitable utility-based Bayesian experimentation paradigm.

3.1.2 Utility-based Bayesian optimal designs

The experimental design problem can be conveniently approached within the Bayesian framework (Chaloner and Verdinelli (1995)). Suppose we study a stochastic process for which we formulate a model M, characterised by a model parameter θ (a variable or a vector). The model M is described by a probability distribution $f(y \mid \theta, d)$ of the outcome y of the studied process under experimental conditions described by the design parameter d given the value of the model parameter θ. Our knowledge about θ is described by a prior distribution $\pi(\theta)$. Whenever the choice of d is under our control there appears a question of choosing the optimal d under which one should observe the stochastic process. Such prescribed experimental conditions are referred to as a *design*, and the optimal design is found under optimality criteria which are specifically formulated depending on the context and the purpose of the experiment.

By employing a utility function $u(d, y, \theta)$ one can specify the purpose of the experiment and measure the value of its outcome y accordingly. The methodology of posing and solving utility-based optimal design problems within the Bayesian paradigm has become somewhat standard (Müller (1999), Cook et al (2008)). The design has to be chosen before performing an experiment and one may choose to maximise the expectation of the utility function $u(d, y, \theta)$ with respect to θ and y (Müller (1999)):

$$d_{\max} = \arg\max_{d \in \mathcal{D}} U(d), \tag{3.7}$$

where

$$U(d) = \int_{\Theta} \int_{\mathcal{Y}} u(d, y, \theta) f(y \mid \theta, d) \pi(\theta) \, \mathrm{d}\theta \, \mathrm{d}y. \tag{3.8}$$

Here $\mathcal{D}$ is the set of possible designs. The set of possible outcomes y of the experiment is denoted by $\mathcal{Y}$. The experiment is defined by a model $f(y \mid \theta, d)$, that is to say by the distribution of y conditional on θ for a given design d.

The utility function $u(d, y, \theta)$ is one of the key elements in this methodology. As its choice reflects the very purpose of experimentation, the utility function may well be contextually specific. For instance, in the context of random graph models (y is a realisation of a random graph) the utility function u might be linked to a

particular property of the resulting graphs, for example, counting the total number of edges or the total number of triangles in the graph.

This, however, should not always be the case: contextually different experiments may still be designed using the same 'context-free' utility functions, especially when the purpose of the experimentation is to make inference on the model parameter θ. Examples include, but are not limited to, the following most common utility functions :

- the *negative squared error loss*:

$$u(d, y, \theta) = -\{\theta - \mathbb{E}\left[\theta \,|\, \mathbf{y}, d\right]\}^2; \tag{3.9}$$

- the inverse of the posterior variance:

$$u(d, y, \theta) = \left[\mathbb{V}(\theta \,|\, y, d)\right]^{-1} \tag{3.10}$$

  (this quantity can be regarded as the precision in the Bayesian sense);

- logarithmic ratio of the posterior distribution to the prior distribution:

$$u(d, y, \theta) = \log \frac{\pi(\theta \,|\, y, d)}{\pi(\theta)}. \tag{3.11}$$

The mathematical expressions for the negative squared error loss and the inverse of the posterior variance are self-explanatory. Although simple, designed to decrease the posterior uncertainty about the parameter θ these utility functions have serious drawbacks. The utility function $u(d, y, \theta)$ from (3.11) overcomes the following two most important of them: (i) its expected value $U(d)$ defined by (3.8) represents the average gain in information about θ rather than decreases the posterior uncertainty about this parameter while performing the experiment under design d, and (ii) $U(d)$ is invariant under a change of parameter, that is to say under the model reparametrisation. These two features are discussed in greater detail in the next section.

It is worth mentioning that a utility function can also take forms that describe more than a single purpose while designing an experiment. For example, the cost of the experimental units used might also be taken into account whilst trying to

achieve the primary goal(s) of the experiment. In such cases context-free and context-specific utility functions can be combined to obtain a more complicated compound utility measure incorporating more than one design criterion. The reader is referred to the monograph of Müller (2007, Chapter 7) references therein for more information on multipurpose designs, and to Parmigiani and Berry (1994), Müller (1999), Chaloner and Verdinelli (1995), Clyde (2004), Fuentes et al (2007) for examples of use of utility functions related to prediction, hypothesis testing, model discrimination, and for applications of compound utility functions involving costs.

When the sole purpose of experimentation is to increase the knowledge about the model parameter there are strong arguments for using the logarithmic ratio $\log \frac{\pi(\theta \mid y,d)}{\pi(\theta)}$ of the posterior to prior as a utility function. The corresponding optimisation problem (3.7-3.8) is directly related to the well-known Lindley information measure and the Kullback–Leibler divergence in this case. We explore these information measures in detail in the next section.

3.2 Shannon entropy, Lindley information measure and Kullback–Leibler divergence

3.2.1 Bits of history

Lindley (1956) in his seminal paper has mentioned that it was Claude Shannon who introduced the following two important ideas into the theory of information in communications engineering:

1 information is a statistical concept—the statistical frequency distribution of the symbols that a message consists of must be considered before the notion can be discussed adequately;

2 there is essentially a unique function of the symbol frequency distribution which measures the amount of the information.

Kullback and Leibler (1951) and subsequently Kullback (1952, 1954) applied the former of these ideas to statistical theory. Lindley (1956) further developed the

theory applying these two ideas and discussing the notion of information carried by an experiment in a general context, rather than specific to communication engineering.

As well as papers of Kullback and Leibler[3], there were works of further authors preceding the paper of Dennis Lindley (1956) (and, indeed, acknowledged by him) discussing and applying similar ideas in various contexts: McMillan (1953) gave the interpretation of Shannon's ideas in the statistical theory; Cronbach (1953) applied Shannon's theory in psychometric problems and essentially gave a definition of the average amount of information provided by an experiment. Methods of comparing experiments (as sampling procedures) involving the decision-theoretic paradigm and consideration of losses have been suggested by Bohnenblust, Shapley and Sherman (in private communication to Blackwell) and by Blackwell (1951).

Subsequently, DeGroot (1962) is concerned with a general experimental methodology when the purpose is to decrease uncertainty in knowledge about the model parameter (or *"... about the true state of nature"*) within the Bayesian context. From a more general position the prior and the posterior knowledge are viewed as uncertainties, and, assuming that these uncertainties can be measured[4], the information in an experiment Y is defined as the difference between the uncertainty in θ prior to observing Y and the expected uncertainty after having observed Y. In the later paper DeGroot (1984) studies the relationship between information measures that are based on both the prior knowledge for θ and the utility function of the experimenter, and measures that are based only on the experimenters prior belief about θ.

The reader is referred to Ginebra (2007), and references therein, as an excellent account on the topic of how to measure information in a statistical experiment. The author focused on a characterisation of the measure of the information in an exper-

[3]As pointed out by MacKay (2003), the diphtong *'ei'* in *'Leibler'* should be pronounced the same as in the word *'heist'*, that is according to German language pronunciation rules.

[4]Essentially, by introducing a functional on the space of all possible prior distributions; the Shannon entropy taken with the negative sign would then be just one of many other possible choices (see Venegas-Martínez (2004) for an account on a general family of information functionals in the context of producing informative and non-informative priors).

iment that encompasses as special cases the measures of information considered by Lindley (1956), Kiefer (1959), Raiffa and Schlaiffer (1961), DeGroot (1962, 1984), Csiszár (1967).

3.2.2 Lindley information

Suggesting a measure of information provided by an experiment whose objective is not to reach decisions but rather to gain knowledge about the model parameter θ, Lindley (1956) exploited Shannon's information measure. Following Dennis Lindley, but slightly reducing the level of rigour (the fact that we will be dealing with random graphs on finite vertex sets allows us to do so), we start with the general definition of an experiment.

Definition 3.2.1. *An experiment E is the ordered triple $E = (\mathcal{Y}, \Theta, \Upsilon)$, where $\Upsilon = \{f(\cdot \,|\, \theta)\}_{\theta \in \Theta}$ is a parametrised family of probability densities (probability mass functions) describing a random object $Y \in \mathcal{Y}$.*

The following is the definition of the *Lindley information measure* given for a prior distribution π.

Definition 3.2.2. *For a prior distribution $\pi(\cdot)$ of θ, the amount of information $\mathcal{I}_0$ contained in this distribution is defined to be minus the Shannon entropy:*

$$\mathcal{I}_0 := -\mathrm{Ent}\{\pi(\theta)\} = \int_\Theta \pi(\theta) \log \pi(\theta)\, d\theta =: \mathbb{E}_\theta[\log \pi(\theta)]. \tag{3.12}$$

Taking into account that $x \log x \to 0$, as $x \to 0$, define $\pi(\theta) \log \pi(\theta) := 0$ for any θ such that $\pi(\theta) = 0$.

The more the function π is concentrated on a single value of θ, the greater the amount of information $\mathcal{I}_0$ is. On the other hand, the more this function is spread over Θ, the smaller this information measure is. Notice, however, that $\mathcal{I}_0$ is not invariant under reparametrisations.

After the experiment has been performed and the observation y of Y obtained, the posterior distribution of θ is $\pi(\cdot \,|\, y)$, given by (2.9). Thus the amount of information associated with $\pi(\cdot \,|\, y)$ is as follows (by analogy with Definition 3.2.2):

$$\mathcal{I}_1(y) := \int \pi(\theta \,|\, y) \log \pi(\theta \,|\, y)\, \mathrm{d}\theta. \tag{3.13}$$

The increase in information provided by the experiment E when the observation y was obtained can be expressed as the difference between $\mathcal{I}_1(y)$ and $\mathcal{I}_0$:

$$\mathcal{I}(E, \pi, y) := \mathcal{I}_1(y) - \mathcal{I}_0.$$

Clearly, some observations are more informative than others (for a given prior information π). Lindley (1956) defined the average amount of information provided by the experiment E by averaging the increase in information provided by the experiment E over all its possible outcomes.

Definition 3.2.3. *The average amount of information provided by the experiment E, with prior knowledge $\pi(\theta)$, is*

$$\mathcal{I}(E, \pi) := \mathbb{E}_Y[\mathcal{I}(E, \pi, y)] = \int_{\mathcal{Y}} (\mathcal{I}_1(y) - \mathcal{I}_0)\Phi(y)\, dy, \tag{3.14}$$

where $\Phi(y)$ is the marginal likelihood:

$$\Phi(y) = \int_{\Theta} f(y \,|\, \theta)\pi(\theta)\, d\theta.$$

Since

$$\int_{\mathcal{Y}} \mathcal{I}_1(y)\Phi(y)\, \mathrm{d}y = \int_{\Theta} \int_{\mathcal{Y}} \log \pi(\theta \,|\, y) f(y \,|\, \theta)\pi(\theta)\, \mathrm{d}\theta \mathrm{d}y$$

and

$$\int_{\mathcal{Y}} \mathcal{I}_0\Phi(y)\, \mathrm{d}y = \mathcal{I}_0 = \int_{\Theta} \int_{\mathcal{Y}} \log \pi(\theta) f(y \,|\, \theta)\pi(\theta)\, \mathrm{d}\theta \mathrm{d}y,$$

it follows immediately from Definition 3.2.3 that

$$\mathcal{I}(E, \pi) = \mathbb{E}_\theta \mathbb{E}_{Y \,|\, \theta} \left[\log \frac{\pi(\theta \,|\, y)}{\pi(\theta)} \right] = \mathbb{E}_Y \mathbb{E}_{\theta \,|\, Y} \left[\log \frac{\pi(\theta \,|\, y)}{\pi(\theta)} \right], \tag{3.15}$$

and from the Bayes theorem that

$$\mathcal{I}(E, \pi) = \mathbb{E}_\theta \mathbb{E}_{Y \,|\, \theta} \left[\log \frac{f(y \,|\, \theta)}{\Phi(y)} \right] = \mathbb{E}_Y \mathbb{E}_{\theta \,|\, Y} \left[\log \frac{f(y \,|\, \theta)}{\Phi(y)} \right]. \tag{3.16}$$

The two representations (3.15) and (3.16) suggest the symmetry between θ and y, and indeed, the third alternative form for $\mathcal{I}(E, \pi)$, that best expresses this symmetry, can also be easily derived:

$$\mathcal{I}(E,\pi) = \int\limits_{\Theta}\int\limits_{\mathcal{Y}} p(y,\theta)\log\frac{p(y,\theta)}{\Phi(y)\pi(\theta)}\,\mathrm{d}\theta\mathrm{d}y, \tag{3.17}$$

where $\Phi(\cdot)$ is, as before, the prior predictive distribution of Y, and $p(y,\theta)$ is the joint distribution of Y and θ.

One should notice that in contrast to $\mathcal{I}_0$, $\mathcal{I}_1(y)$, and $\mathcal{I}(E,\pi,y)$, the expected gain in information $\mathcal{I}(E,\pi)$ prior to performing the experiment E is invariant under one-to-one transformations of the parameter space Θ.

The informativeness of experiments can be measured using the expected Lindley information gain: if E_1 and E_2 are two experiments such that

$$\mathcal{I}(E_1,\pi(\theta)) \leq \mathcal{I}(E_2,\pi(\theta)),$$

then we are saying that E_2 is not less informative than E_1.

3.2.3 Comparing informativeness of experiments: expected Kullback–Leibler divergence and expected Lindley information gain as expected utility and their properties

The average amount of information that will be obtained after performing an experiment (and calculated prior to performing it) is directly related to the Kullback–Leibler divergence—a well-known functional that measures the difference between two probability distributions.

Kullback–Leibler divergence and its basic properties

Definition 3.2.4. *The Kullback–Leibler divergence of the probability density $g(t)$ from the probability density $h(t)$ is defined as*

$$D_{KL}\{h(t) \parallel g(t)\} := \int\limits_{\mathbb{R}} h(t)\log\frac{h(t)}{g(t)}\,dt.$$

Here the probability densities become probability mass functions whenever the supports of the distributions involved are countable sets—integration should be replaced by summation then.

This measure of difference between two distributions was originally introduced by Kullback and Leibler (1951) and considered as a "directed divergence". The Kullback–Leibler (KL) divergence cannot be considered a true distance, as, although it is a positive quantity, it is not symmetric. Neither does this divergence measure satisfy the triangle inequality.

The basic properties of the Kullback–Leibler divergence follow.

KL.1 (positiveness) $D_{KL}\{h(t) \parallel g(t)\} \geq 0$, for any distributions h and g, with equality if, and only if, $h(t) = g(t)$ almost everywhere on $\mathbb{R}$.

Proof. To verify this we recall that the logarithm $\log(\cdot)$ is a concave function and by the Jensen inequality

$$\int_{-\infty}^{\infty} \log r(x) f(x)\,\mathrm{d}x \leq \log \int_{-\infty}^{\infty} r(x) f(x)\,\mathrm{d}x$$

for any real-valued measurable function r and density f with equality when $r(x)$ is a constant almost everywhere. Hence,

$$-D_{KL}\{h(t) \parallel g(t)\} = \int_{\Theta} \log \frac{g(t)}{h(t)} h(t)\,\mathrm{d}\theta \leq \log \int_{\Theta} \frac{g(t)}{h(t)} h(t)\,\mathrm{d}\theta \equiv 0,$$

and equality holds if, and only if, $r(x) := g(t)/h(t) \equiv \text{const}$, which is only possible when $g(t) = h(t)$, since these two functions are probability densities.

□

KL.2 (asymmetry) There exist probability densities h and g such that

$$D_{KL}\{h(t) \parallel g(t)\} \neq D_{KL}\{g(t) \parallel h(t)\}.$$

KL.3 (triangle inequality breakdown) There exist such probability densities f, h, and g, that

$$D_{KL}\{h \parallel f\} + D_{KL}\{f \parallel g\} < D_{KL}\{h \parallel g\}.$$

KL.4 The expected Lindley information gain prior to performing an experiment E with a prior distribution $\pi(\theta)$ coincides with the expected KL divergence of $\pi(\theta)$ from the corresponding posterior $\pi(\theta \,|\, y)$:

$$\mathcal{I}(E, \pi) = \mathbb{E}_Y[D_{KL}\{\pi(\theta \,|\, y) \parallel \pi(\theta)\}].$$

Proof. The proof follows immediately from the definition of the Kullback–Leibler divergence and the form (3.15) for the expected Lindley information gain $\mathcal{I}(E, \pi)$. □

The Kullback–Leibler divergence can be viewed as a particular case of a more general measure of divergence between two distribution—the α-divergence (see Paquet (2008), Amari (1985, 2005), Minka (2005). This viewpoint is especially important from the position of information geometry (see Amari and Nagaoka (2000)).

Comparing informativeness of experiments

Often experiments can be controlled, and then they can be distinguished by different values of the control variables. Generalising Definition 3.2.1 consider a family of experiments $E_d = (\mathcal{Y}_d, \Theta_d, \Upsilon_d)$ that are labelled by some control variable d, so that $\Upsilon_d = \{f(\cdot \,|\, \theta, d)\}_{\theta \in \Theta_d}$ for $d \in \mathcal{D}$. This is a fairly general set up. However, it is natural to assume that the parameter spaces Θ_d do not depend on the control variable d: $\Theta_d \equiv \Theta \ \forall d \in \mathcal{D}$. In view of the design problem discussed in § 3.1.2, we relate to the control variable d as a design. We also consider that the prior knowledge $\pi(\theta)$ does not depend on d.

Since the expected KL divergence of the prior $\pi(\theta)$ from the posterior $\pi(\theta \,|\, y)$ coincides with the expected Lindley information gain (by KL.4), and the latter coincides with the expected utility $U(d)$ defined by (3.8) with the utility function $u(d, y, \theta) = \log \frac{\pi(\theta \,|\, y, d)}{\pi(\theta)}$, one can write the following:

$$U_{\mathrm{KL}}(d) := \mathbb{E}_{y,\theta}\left[\log \frac{\pi(\theta \,|\, y, d)}{\pi(\theta)}\right] \equiv \mathbb{E}_y\left[D_{KL}\{\pi(\theta \,|\, y, d) \parallel \pi(\theta)\}\right] \equiv \mathcal{I}(E_d, \pi), \quad (3.18)$$

denoting the expected utility based on the Kullback–Leibler divergence[5] by $U_{\mathrm{KL}}(d)$. This combines together the notions of the Lindley information gain and the KL divergence, and fits them into the utility based Bayesian framework presented in § 3.1.2.

We list (without proofs) the most important properties of the expected Lindley information measure with a view of comparing experiments. More properties are

[5] or on the utility (3.11).

given in Lindley (1956) with proofs. We omit writing $\pi(\theta)$ as long as it remains unchanged while comparing different experiments with the design parameter d:

$$\mathcal{I}(E_d, \pi) = \mathcal{I}(E_d).$$

By a sum of two experiments $E_{d_1} = (\mathcal{Y}_{d_1}, \Theta, \Upsilon_{d_1})$ and $E_{d_2} = (\mathcal{Y}_{d_2}, \Theta, \Upsilon_{d_2})$ we understand an experiment E_{d_1,d_2} which consists in observing an unordered pair (y_{d_1}, y_{d_2}), $d_1, d_2 \in \mathcal{D}$.

LIG.1 Any experiment is informative on the average, unless the density of Y does not depend on θ. That is,

$$\mathcal{I}(E_d) \geq 0,$$

with equality if, and only if, $f(y \mid \theta)$ does not depend on θ, except possibly on a set of zero Lebesgue measure.

LIG.2 The sum of two experiments is conditionally additive:

$$\mathcal{I}(E_{d_1,d_2}) = \mathcal{I}(E_{d_1}) + \mathcal{I}(E_{d_2} \mid E_{d_1}),$$

where $\mathcal{I}(E_{d_2} \mid E_{d_1})$ is the average Lindley information gain prior to performing the experiment E_{d_2} with the prior knowledge $\pi(\theta \mid y_{d_1})$.

LIG.3 If y_{d_1} is sufficient for $y_{d_1,d_2} = (y_{d_1}, y_{d_2})$ in the Neyman–Fisher sense (p. 24), then

$$\mathcal{I}(E_{d_1,d_2}) = \mathcal{I}(E_{d_1}).$$

LIG.4 If two experiments E_{d_1} and E_{d_2} are independent, that is to say if

$$f(y_{d_1}, y_{d_2} \mid \theta) = f(y_{d_1} \mid \theta) f(y_{d_2} \mid \theta) \; \forall \theta \in \Theta,$$

then

$$\mathcal{I}(E_{d_2} \mid E_{d_1}) \leq \mathcal{I}(E_{d_2}),$$

with equality if, and only if, y_{d_1} and y_{d_2} are independent (their joint prior predictive distribution factorises into its marginals).

LIG.5 The Lindley information gain is *subadditive*: if E_{d_1} and E_{d_2} are independent experiments, then

$$\mathcal{I}(E_{d_1}) + \mathcal{I}(E_{d_2}) \geq \mathcal{I}(E_{d_1,d_2}),$$

with equality if, and only if, y_{d_1} and y_{d_2} are (unconditionally) independent. Note that the unconditional independence here means the same as in LIG.4, and thus LIG.5 is implied by the properties LIG.2 and LIG.4.

An alternative form for the expected KL divergence and first-order conditions for the expected utility

The following useful representation appears in Lindley (1956) without a proof. This representation complements the ones presented in (3.15-3.17). We use this representation to derive first-order conditions for the expected utility based on the KL divergence (Theorem 3.2.7) and to prove the worst case scenario result for indefinitely growing or diminishing vertex configurations (Theorem 4.1.1).

Lemma 3.2.5. *The expected utility $U_{KL}(d)$ can be represented in the following form:*

$$U_{KL}(d) = Ent\{\Phi(y \,|\, d)\} - \mathbb{E}_\theta \left[Ent\{f(y \,|\, \theta, d)\} \right], \tag{3.19}$$

where $\Phi(y \,|\, d)$ is the marginal of the joint distribution of y and θ, i.e. the prior predictive distribution

$$\Phi(y \,|\, d) := \int_\Theta f(y \,|\, \theta, d)\pi(\theta) \, \mathrm{d}\theta.$$

Proof. If $f(y \,|\, \theta, d)$ is the model distribution and $\pi(\theta)$ is the prior distribution for θ, then, as follows from Bayes' theorem,

$$\frac{\pi(\theta \,|\, y, d)}{\pi(\theta)} = \frac{f(y \,|\, \theta, d)}{\Phi(y \,|\, d)}. \tag{3.20}$$

Therefore

$$
\begin{aligned}
U_{\mathrm{KL}}(d) &= \int_{\Theta}\int_{Y} \log \frac{f(y \mid \theta, d)}{\Phi(y \mid d)} f(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\mathrm{d}y \\
&= \int_{\Theta}\int_{Y} \log f(y \mid \theta, d) f(y \mid \theta, d)\,\mathrm{d}y\,\pi(\theta)\,\mathrm{d}\theta - \int_{\Theta}\int_{Y} \log \Phi(y \mid d) f(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}y \\
&= -\int_{\Theta} \mathrm{Ent}\{f(y \mid \theta, d)\}\pi(\theta)\,\mathrm{d}\theta - \int_{Y} \log \Phi(y \mid d) \int_{\Theta} f(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}y \\
&= -\int_{\Theta} \mathrm{Ent}\{f(y \mid \theta, d)\}\pi(\theta)\,\mathrm{d}\theta - \int_{Y} \log \Phi(y \mid d)\Phi(y \mid d)\,\mathrm{d}y \\
&= \mathrm{Ent}\{\Phi(y \mid d)\} - \int_{\Theta} \mathrm{Ent}\{f(y \mid \theta, d)\}\pi(\theta)\,\mathrm{d}\theta,
\end{aligned}
$$

and the lemma is proven. □

Lemma 3.2.6. *Let* $\{f(x \mid t),\, x \in \mathbb{R}\}_{t \in \mathcal{T} \subseteq \mathbb{R}}$ *be a parametrised family of univariate probability density functions (or probability mass functions) with parameter* t*. Then*

$$
\frac{\mathrm{d}}{\mathrm{d}t} Ent\{f(x \mid t)\} = -\int_{\mathbb{R}} \frac{\partial f(x \mid t)}{\partial t} \log f(x \mid t)\,\mathrm{d}x, \tag{3.21}
$$

provided the function $f(x \mid t)$ *is such that differentiation and integration are interchangeable.*

Proof. Differentiation of the entropy of $f(x \mid t)$ can be done using the Leibniz integration rule when $f(x \mid t)$ is a density provided that this function is continuous with respect to x and its partial derivative with respect to t exists and is also continuous within the interval of differentiation:

$$
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Ent}\{f(x \mid t)\} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}} f(x \mid t) \log f(x \mid t)\,\mathrm{d}x = -\int_{\mathbb{R}} \frac{\partial}{\partial t}\left(f(x \mid t) \log f(x \mid t)\right)\,\mathrm{d}x. \tag{3.22}
$$

Since

$$
\frac{\partial}{\partial t}\left(f(x \mid t) \log f(x \mid t)\right) = \frac{\partial}{\partial t} f(x \mid t) + \left(\frac{\partial}{\partial t} f(x \mid t)\right) \log f(x \mid t)
$$

and $\int_{\mathbb{R}} \frac{\partial}{\partial t} f(x \mid t)\,\mathrm{d}x = 0$, as a derivative of unity, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Ent}\{f(x \mid t)\} = -\int_{\mathbb{R}} \frac{\partial f(x \mid t)}{\partial t} \log f(x \mid t)\,\mathrm{d}x.$$

If $f(x \mid t)$ is a probability mass function of a random variable X taking values in a finite set, then the differentiation can be directly applied to the corresponding finite sum, providing $f(x \mid t)$ is differentiable in t for each $x \in \operatorname{supp} X$. If X takes values in an infinite countable set, then a sufficient condition for (3.21) to hold would be uniform convergence of the sum of partial derivatives of $f(x \mid t)$ with respect to t over $x \in \operatorname{supp} X$. □

The following first-order conditions result was first established by Parmigiani and Berry (1994). Our proof is based on Lemmas 3.2.5 and 3.2.6.

Theorem 3.2.7. *(First-order conditions) The derivative of the expected utility $U_{KL}(d)$ with respect to a continuous design variable $d \in \mathcal{D} \subseteq \mathbb{R}$ can be calculated as follows:*

$$U'_{KL}(d) = \int_{\Theta}\int_{\mathcal{Y}} \log \frac{\pi(\theta \mid y, d)}{\pi(\theta)} f'_d(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\mathrm{d}y, \tag{3.23}$$

provided the functions $f(\cdot \mid \cdot, \cdot)$, $\pi(\cdot)$ are such that differentiation of $U_{KL}(d)$ and the corresponding integration are interchangeable.

Proof. Taking U_{KL} in the form (3.19) derived in Lemma 3.2.5 and applying Lemma 3.2.6 we obtain

$$U'_{\mathrm{KL}}(d) = \left(\mathrm{Ent}\{\Phi(y \mid d)\} - \int_{\Theta} \mathrm{Ent}\{f(y \mid \theta, d)\}\pi(\theta)\,\mathrm{d}\theta\right)'_d \tag{3.24}$$

$$= -\int_{\mathcal{Y}} [\Phi(y \mid d)]'_d \log \Phi(y \mid d)\,\mathrm{d}y + \int_{\Theta}\int_{\mathcal{Y}} f'_d(y \mid \theta, d) \log f(y \mid \theta, d)\pi(\theta)\,\mathrm{d}y\,\mathrm{d}\theta. \tag{3.25}$$

On the other hand, denoting the right-hand side of the hypothetical identity (3.23) by $J(d)$ and using (3.20), we obtain

$$J(d) = \int_{\Theta}\int_{\mathcal{Y}} \log \frac{f(y \mid \theta, d)}{\Phi(y \mid d)} f'_d(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}y = \int_{\Theta}\int_{\mathcal{Y}} \log f(y \mid \theta, d) f'_d(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}y$$
$$- \int_{\Theta}\int_{\mathcal{Y}} \log \Phi(y \mid d) f'_d(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}y,$$

so that

$$U'_{\mathrm{KL}}(d) - J(d) = \int_{\Theta}\int_{\mathcal{Y}} \log \Phi(y \mid d) f'_d(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}y - \int_{\mathcal{Y}} [\Phi(y \mid d)]'_d \log \Phi(y \mid d)\,\mathrm{d}y.$$

Differentiation of (3.20) with respect to d produces

$$f'_d(y \mid \theta, d) = \frac{1}{\pi(\theta)} \left[\pi'_d(\theta \mid y, d)\Phi(y \mid d) + \Phi'_d(y \mid d)\pi(\theta \mid y, d)\right],$$

consequently resulting in the following:

$$\int_{\Theta}\int_{\mathcal{Y}} \log \Phi(y \mid d) f'_d(y \mid \theta, d)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}y = \int_{\mathcal{Y}} \log \Phi(y \mid d)\Phi(y \mid d) \int_{\Theta} \pi'_d(\theta \mid y, d)\,\mathrm{d}\theta\,\mathrm{d}y$$
$$+ \int_{\mathcal{Y}} \log \Phi(y \mid d)\Phi'_d(y \mid d) \int_{\Theta} \pi(\theta \mid y, d)\,\mathrm{d}\theta\,\mathrm{d}y.$$

Since $\int_{\Theta} \pi(\theta \mid y, d)\,\mathrm{d}\theta \equiv 1$ and $\int_{\Theta} \pi'_d(\theta \mid y, d)\,\mathrm{d}\theta \equiv 0$, $d \in \mathcal{D}$, it follows that

$$U'_{\mathrm{KL}}(d) - J(d) \equiv 0,\ d \in \mathcal{D}.$$

The proof is complete. □

3.3 Progressive and Instructive Designs

When the experimental motivation consists in increasing one's knowledge about the model parameter the expected KL divergence of prior from posterior coincides with the expected utility $U(d)$ (see (3.8)) based on $u(d, y, \theta) = \log \frac{\pi(\theta \mid y, d)}{\pi(\theta)}$ (recall that we denoted such expected utility by $U_{\mathrm{KL}}(d)$). If, however, the purpose of the experiment is to instruct someone holding the prior $\pi(\theta)$ using one's superior

knowledge $\pi^*(\theta)$ of the system under study, then the expected KL divergence should be calculated in the form

$$U^*_{KL}(d) = \mathbb{E}_y \left[D_{KL}\{\pi(\theta \,|\, y, d) \parallel \pi(\theta)\} \right], \tag{3.26}$$

where the integration over the space of observables $\mathcal{Y}$ is to be carried out using one's superior knowledge $\pi^*(\theta)$. We refer to these two different experimental motivations as to *progressive design* and *instructive design*[6], correspondingly, and discuss them next in more detail.

3.3.1 Progressive designs

In this setting there is an experimenter A who holds a prior knowledge about the model parameter in the form of its prior distribution $\pi(\theta)$ (which could have been obtained on earlier stages of analysing the process or experimenting with it) and whose goal is to design an optimal experiment in order to increase this knowledge. With the KL divergence as an information gain measure in hand, this experimenter maximises the expected KL divergence (3.26):

$$U_{\mathrm{KL}}(d) = \mathbb{E}_Y \mathbb{E}_\Theta \left[\log \frac{\pi(\theta \,|\, y, d)}{\pi(\theta)} \,\middle|\, y \right] = \int\limits_\Theta \int\limits_{\mathcal{Y}} \log \frac{\pi(\theta \,|\, y, d)}{\pi(\theta)} f(y, \theta \,|\, d) \, \mathrm{d}\theta \mathrm{d}y,$$

where $f(y, \theta \,|\, d)$ is the joint density of y and θ. Thus, as discussed in § 3.2.3, $U_{\mathrm{KL}}(d)$ coincides with (3.8) where $u(d, y, \theta) = \log \frac{\pi(\theta \,|\, y)}{\pi(\theta)}$. Notice, also, that since the prior distribution $\pi(\theta)$ does not depend on the chosen design d, the expected information gain $U_{\mathrm{KL}}(d)$ can be written as follows:

$$U_{\mathrm{KL}}(d) = \int\limits_{\mathcal{Y}} \int\limits_\Theta \log \pi(\theta \,|\, y, d) f(y \,|\, \theta, d) \pi(\theta) \, \mathrm{d}\theta \mathrm{d}y + \mathrm{Ent}\{\pi(\theta)\}, \tag{3.27}$$

with the second term being independent of d, so that it suffices to maximise the first term only.

[6] In Cook et al (2008) these experimental scenarios are called *progressive* and *pedagogic* designs.

3.3.2 Instructive designs

In contrast to the progressive design scenario, in the instructive case there is an experimenter A, holding a prior $\pi(\theta)$, and a better informed *trainer* [7] B whose knowledge about the model parameter is summarised in a distribution $\pi^*(\theta)$. The purpose here is to help maximising the change in experimenter's information from $\pi(\theta)$ to $\pi(\theta \,|\, y)$ by designing an experiment using the existing superior knowledge $\pi^*(\theta)$.

By analogy with (3.7-3.8) such an optimisation problem can be formulated in the following way:

$$d^*_{\max} = \arg\max_{d \in D} U^*_{\mathrm{KL}}(d), \tag{3.28}$$

$$U^*_{\mathrm{KL}}(d) = \int_{\mathcal{Y}} D_{\mathrm{KL}}\{\pi(\theta \,|\, y, d) \parallel \pi(\theta)\}\Phi^*(y)\,\mathrm{d}y, \tag{3.29}$$

where, as before, $\pi(\theta \,|\, y)$ is the posterior of the experimenter A, and $\Phi^*(y)$ is the model as it is understood (known) by the instructor B:

$$\Phi^*(y) = \int_{\Theta} f(y \,|\, \theta, d)\pi^*(\theta)\,\mathrm{d}\theta,$$

that is to say, $\Phi^*(y)$ is the prior predictive distribution of y under the prior $\pi^*(\theta)$.

In particular, if the instructor B knows the exact value of θ, θ^*, and hence $\pi^*(\theta)$ is the Dirac function $\delta(\theta - \theta^*)$ (see Appendix B), then $\Phi^*(y) = f(y \,|\, \theta^*, d)$, so that

$$U^*_{\mathrm{KL}}(d) = \int_{\mathcal{Y}} D_{\mathrm{KL}}\{\pi(\theta \,|\, y, d) \parallel \pi(\theta)\} f(y \,|\, \theta^*, d)\,\mathrm{d}y.$$

3.4 Simulation-based evaluation of the expected utility

Generally, the solution to the optimal design problems (3.7-3.8) and (3.28-3.29) cannot be obtained analytically. This is mainly due to the following three reasons. First, the design space $\mathcal{D}$ may be complicated, with many design variables, some of them having a continuum range of values. Second, even if the design space has a

[7] *instructor* or *tutor*

simple structure, the utility function may not be easy to evaluate, and the expected utility $U(d)$ cannot be obtained explicitly. Finally, for incomplete observations of highly non-linear stochastic processes such as epidemic models, the likelihood is not usually available in a closed form, and this results in computationally intensive evaluations or high cost sampling procedures.

A review of analytical and approximate numerical solutions to Bayesian optimal design problems for the traditional experimental design involving linear and non-linear models can be found in Verdinelli (1992) and Chaloner and Verdinelli (1995).

Müller (1999) reviews simulation-based methods for optimal design problems where the expected utility $U(d)$ is evaluated by Monte-Carlo simulation. In its simplest form an estimate $\widehat{U}$ of U for any given design d in the progressive case is as follows (§ 2.3.1):

$$\widehat{U}(d) = \frac{1}{M} \sum_{i=1}^{M} u(d, \theta_i, y_i), \tag{3.30}$$

where $\{(\theta_i, y_i), i = 1, \ldots, M\}$ is a Monte-Carlo sample generated values:

$$\theta_i \sim \pi(\theta),\ y_i \sim f(y \,|\, \theta, d). \tag{3.31}$$

The expected utility U may, in particular, be based on the KL divergence.

Analogously, the expected utility U^*_{KL} under instructive scenario, when the instructor knows the true value of the model parameter, θ^*, can be evaluated using the following scheme:

$$\widehat{U}^*_{\mathrm{KL}}(d) = \frac{1}{M} \sum_{i=1}^{M} \widehat{KL}(y^*_i, d), \tag{3.32}$$

where $y^*_i \sim f(y \,|\, \theta^*, d)$, $i = 1, \ldots, M$, and $\widehat{KL}(y_i, d)$ is an estimate of the KL divergence $D_{\mathrm{KL}}\{\pi(\theta \,|\, y^*, d) \parallel \pi(\theta)\}$ that can be obtained via numerical integration of $\log \frac{\pi(\theta \,|\, y, d)}{\pi(\theta)}$ with respect to the posterior $\pi(\theta \,|\, y, d)$. The former function, in turn, may need to be evaluated through simulation methods—perhaps using an MCMC scheme.

Evaluation of a continuous expected utility surface $U(d)$ by (3.30) or (3.32) can be done by computing its values on a discretised grid of points and further smoothing of the obtained set of values in order to approximate the expected utility

landscape. This, however, may be problematic when $d \in \mathcal{D}$ is a multidimensional design parameter. An alternative to this method might be the augmented probability simulation approach which is studied in Clyde, Müller and Parmigiani (1995), Bielza, Müller and Rios-Insua (1999), and reviewed in Müller (1999).

The augmented probability simulation approach assumes that $u(d, \theta, y)$ is a non-negative bounded function. (This condition, although not always automatically satisfied, can be easily achieved by correspondingly modifying the utility function.) An artificial distribution, proportional to $u(d, \theta, y) f(y \,|\, \theta, d)\pi(\theta)$ can be defined then on d, θ, y (Müller (1999)):

$$h(d, \theta, y) \propto u(d, \theta, y) f(y \,|\, \theta, d)\pi(\theta), \tag{3.33}$$

so that its marginal in d is proportional to the expected utility $U(d)$. Sampling from $h(\cdot, \cdot, \cdot)$ can be used to obtain a sample from its marginal using the following Metropolis–Hastings MCMC scheme described by Müller (1999):

1 Start with a design $d^{(0)}$. Simulate $(\theta^{(0)}, y^{(0)}) \sim f(y \,|\, \theta, d^{(0)})\pi(\theta)$. Evaluate $u^{(0)} = u(d^{(0)}, \theta^{(0)}, y^{(0)})$.

2 Set $k := 1$.

3 Generate a 'candidate' $\tilde{d}$ from a probing distribution $g(\tilde{d} \,|\, d^{(k-1)})$.

4 Simulate $(\tilde{\theta}, \tilde{y}) \sim f(\tilde{y} \,|\, \tilde{\theta}, \tilde{d})\pi(\tilde{\theta})$. Evaluate $\tilde{u} = u(\tilde{d}, \tilde{\theta}, \tilde{y})$.

5 Compute the acceptance probability

$$\begin{aligned} \alpha &= \min\left\{1, \frac{h(\tilde{d}, \tilde{\theta}, \tilde{y})}{h(d^{(k-1)}, \theta^{(k-1)}, y^{(k-1)})} \frac{g(d^{(k-1)} \,|\, \tilde{d})}{g(\tilde{d} \,|\, d^{(k-1)})} \frac{f(\theta^{(k-1)} \,|\, y^{(k-1)}, d^{(k-1)})}{f(\tilde{\theta} \,|\, \tilde{y}, \tilde{d})}\right\} \\ &= \min\left\{1, \frac{\tilde{u} g(d^{(k-1)} \,|\, \tilde{d})}{u^{(k-1)} g(\tilde{d} \,|\, d^{(k-1)})}\right\}. \end{aligned}$$

6 Set $(d^{(k)}, u^{(k)}) := \begin{cases} (\tilde{d}, \tilde{u}) & \text{with probability } \alpha, \\ (d^{(k-1)}, u^{(k-1)}) & \text{with probability } 1 - \alpha. \end{cases}$

7 Set $k := k + 1$ and repeat steps 3 through 6 until the chain is judged to have practically converged (see discussion at the end of § 2.3.2).

When using the augmented probability modelling approach one approximates the optimal design by an empirical mode of the marginal of the artificial distribution h, that is by the mode of the first component of the obtained sample (i.e. taking the mode of $\{(d^{(k)}, u^{(k)})\}$ and discarding the values $u^{(1)}, u^{(2)} \ldots$). Notice that the evaluation of $u(d, \theta, y)$ at each step may generally involve MCMC sampling from the posterior $\pi(\theta \,|\, y, d)$.

Finally, we would like to mention the work of Paquet (2008) which examined both deterministic and stochastic methods of treating integrals that involve intractable posterior distributions, in particular for the models where the parameter space can be extended with additional latent variables in order to obtain distributions that are easier to handle algorithmically. Ryan (2003) discussed some properties of estimators of the Kullback–Leibler expected information based on Laplace approximation of the prior predictive distribution in the context of optimal design problems with application to the random fatigue-limit model.

3.5 Second formulation of the problem

The weighted random graph model and the optimal random graph problem were introduced in § 1.2.1. The purpose of designing an experiment using that model was to make inference on the model parameter(s). Here we describe our weighted random graph model using the notion of a node-induced weighted random subgraph (Section 2.1), p.19, and formulate the optimal design problem for such graphs using the utility based Bayesian methodology discussed in § 3.1.2 and Sections 3.2-3.4.

3.5.1 The model

Let $G = (V, R)$ be a simple weighted graph with a possibly uncountable vertex set V and a weight structure R. That is, R represents a non-negative symmetric function $r(\cdot, \cdot)$ that can take value $+\infty$, and it defines completely the adjacency structure of the graph G (see properties WF.1-4 on p. 21).

For any given fixed θ, an edge-probability function $p(r, \theta)$ defined by (1.2) and satisfying Assumptions 1.2.1-1.2.2, and a subset $V' \subseteq V$, we define a random

graph $\mathcal{G}_{V'} := (V', R|_{V'\times V'}, p(r,\theta))$ by deleting each edge (u,v) of the node-induced subgraph $G' = (V', R|_{V'\times V'})$ of the graph G with probability $1 - p(r(u,v),\theta)$ independently of the status of other edges.

Let us call any countable subset V' of the vertex set V a *node arrangement design*. A set of subsets, $\mathcal{D}$, of the vertex set V, $\mathcal{D} := \{V' : V' \subseteq V\}$, is called a *node arrangement design space*. Any finite node arrangement design containing n nodes is called to be an *n-node configuration design*. For any given vertex set V denote the set of all n-node configuration designs by $\mathcal{D}^{(n)}$:

$$\mathcal{D}^{(n)} := \{V' \subseteq V : |V'| = n\}.$$

Any realisation of the described random graph $\mathcal{G}_{V'}$ defines a $0-1$ map y on $V' \times V'$ as follows:

$$y : V' \times V' \to \{0,1\}, \tag{3.34}$$

$$y(u,v) = \begin{cases} 0 & \text{if the edge } (u,v) \text{ was deleted,} \\ 1 & \text{otherwise.} \end{cases} \tag{3.35}$$

If θ is unknown and assumed to be random, given a realisation y of a random graph $\mathcal{G}_{V'}$ the likelihood function for θ is as follows:

$$\mathcal{L}_{V'}(\theta; y) = f(y \,|\, \theta, V') = \prod_{\{u,v\}\in V'\otimes V'} [p(r(u,v),\theta)]^{y(u,v)} \, [1 - p(r(u,v),\theta)]^{1-y(u,v)}, \tag{3.36}$$

where $V' \otimes V'$ is the set of all two element subsets of V', as in (2.1).

Finally, denote the set of all possible observations y by $\mathcal{Y}$.

3.5.2 n-node optimal design problem for random graphs

The notations of the previous paragraph make it possible to directly translate formulation of the optimal design problem to weighted random graphs. In the Bayesian context, the n-node optimal design problem consists in finding an n-node configuration design that maximises the expected Kullback–Leibler divergence (§ 3.3). A utility $u(y,\theta,d)$ corresponding to an observation $y \in \mathcal{Y}$, parameter $\theta \in \Theta$, and design $d \in \mathcal{D}^{(n)}$ can be other than $\log \frac{\pi(\theta \,|\, y,d)}{\pi(\theta)}$; in this case the whole

construction of § 3.1.2 remains the same but the choice of the utility function u. The notions of the progressive and instructive designs from § 3.3, can also be analogously formulated for utilities other than the logarithmic ratio of the posterior to prior distribution (corresponding to the KL divergence).

3.5.3 Examples

Example 3.5.1. *Let $G = (V, R)$, where $V = \mathbb{R}_+$ and*

$$R:\ r(u,v) = \begin{cases} v, & u = 0, \\ u, & v = 0, \\ +\infty, & \textit{otherwise.} \end{cases}$$

Whatever the optimality condition and the edge-probability function are chosen, the optimal design $d^* = V' \in \mathcal{D}^{(n)}$ will clearly contain the origin as a vertex, otherwise all vertices of the resulting random graph will be isolated. It is reasonable to assume that $n \geq 2$.

In this example a design can be interpreted as a set of observation moments in time (Example 3.1.1, Figure 3.2), that is to say is topologically equivalent to a star (Figure 1.1, star).

Example 3.5.2. *Let $G = (V, R)$, where $V = \mathbb{R}^m$ and R is the Euclidean metric, that is (V, R) is an m-dimensional Euclidean space (it is the Euclidean plane when $m = 2$).*

A design $d \in \mathcal{D}^{(n)}$ consists of n points chosen in the space $\mathbb{R}^m$. Notice, however, that since translations and rotations do not change the edge structure of the corresponding random graph, all n-point configurations with a given set of $n(n-1)/2$ edge lengths will be equally informative. Thus, there is an 'equivalent' analogue for $\mathcal{D}^{(n)}$, and namely:

$$\tilde{\mathcal{D}}^{(n)} = \{\boldsymbol{r} = (r_{12}, r_{13}, \ldots, r_{n-1,n}) \in \mathbb{R}^{n(n-1)/2} : \textit{there exists a planar graph}$$

$$\textit{on } n \textit{ nodes and edge lengths from } \boldsymbol{r}\}.$$

Of course, any permutation of the components of a design $\boldsymbol{r} \in \tilde{\mathcal{D}}^{(n)}$ will be a design of the same informativeness, so that the corresponding expected utility

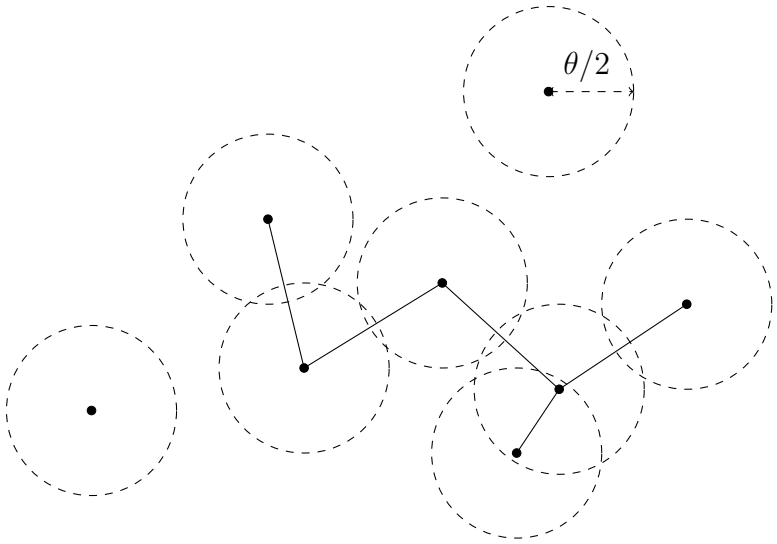


Figure 3.4: Example of a geometric (proximity) graph on eight nodes with a threshold parameter $\theta \in \Theta = \mathbb{R}_+$.

function $U(\boldsymbol{r})$ is a symmetric function. It might, however, be of interest to ask whether $U(\boldsymbol{r})$ is unimodal or multimodal for any given choice of the edge-probability function p, the utility function u, the number of nodes n, and the dimension of space, m.

Example 3.5.3. *Let $G = (V, R)$, where $V = \mathbb{R}^2$ and R is the Euclidean metric, that is (V, R) is the Euclidean plane. Let the edge-probability function $p(\cdot, \cdot)$ be a step function as follows:*

$$p(r, \theta) = \begin{cases} 1, & r \leq \theta, \\ 0, & \textit{otherwise.} \end{cases}$$

Thus, the model parameter θ can be interpreted as a threshold—there is an edge between two nodes if, and only if, the distance between the nodes does not exceed θ (or, equivalently, if and only if the intersection of the two discs of of the radius $\theta/2$ with centres in these nodes is not empty).

The described graphs are known as geometric graphs or proximity graphs (see Penrose (2003)). An example of a geometric graph in plane is presented in Figure 3.4.

Example 3.5.4. *Let $G = (V, R)$, where $V \equiv \mathbb{Z}^2$ and R, as in Example 2.1.3, is*

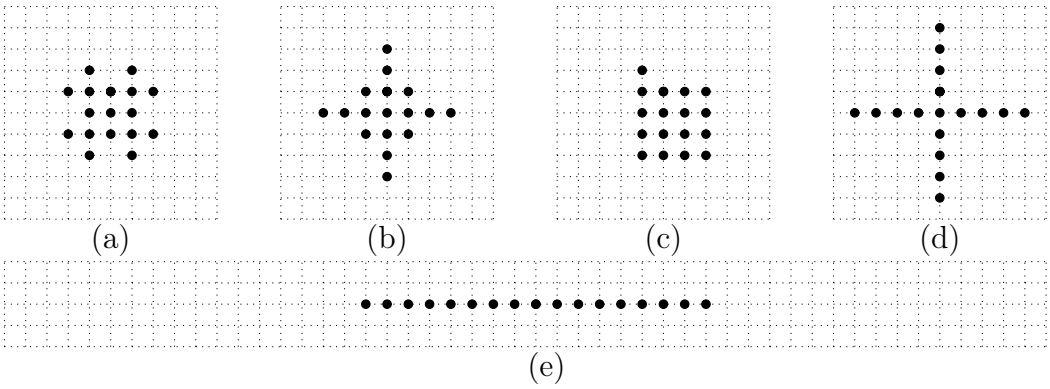


Figure 3.5: Five of many more possible arrangements of 17 vertices on the two-dimensional integer grid $\mathbb{Z}^2$.

the Euclidean metric restricted to a four-neighbourhood as follows:

$$r\left(u(x_1,y_1),v(x_2,y_2)\right) := \begin{cases} 1 & \text{if } \|u-v\|_1 = |x_2-x_1|+|y_2-y_1| = 1, \\ +\infty & \text{otherwise,} \end{cases} \quad \forall u,v \in V.$$

Figure 3.5 depicts examples of five possible node arrangement designs when $n = 17$.

It is easy to see that since any two edges of the considered random graph are present in its realisation independently of each other, an optimal n-node configuration design based on the KL divergence (or, equivalently, on the Lindley information measure) is the one that has the greatest possible number of potential edges[8]*. For instance, the node arrangement (c) in Figure 3.5 will lead to at most 25 edges in a realisation of the corresponding random graph on this nodes, whereas the node arrangement (e), Figure 3.5, may not give more than 16 edges. The node configurations from (a) and (b), both in Figure 3.5, inducing a graph with the same number of possible edges (twenty edges), are equally informative.*

The problem is not that trivial when there is missing information in observations. For example, one might only be able to see the endpoints of the present edges in a realisation of the random graph under consideration, but not the edges themselves. It is not straightforward to answer the question whether the most 'packed' configuration of nodes is the most informative then. In fact, this is not the case, as we shall learn in §§ 5.1.3, 5.1.4, Chapter 5 after introducing inner-outer design plots for percolation model.

[8]This is due to the property LIG.5 of the expected Lindley information gain, p.57.

Chapter 4

Optimal Designs for Basic Random Graph Models

4.1 Worst case scenarios: indefinitely growing or diminishing vertex configurations

When the purpose of experimentation is to make inference on the model parameter and the edge-probability function decreases with the edge weight, an experiment when all the edges of an experimental random graph have very large or very small weights is intuitively fairly uninformative. Here we give the exact meaning to this assertion using the expected Kullback–Leibler divergence.

Let $G = (V, R)$ be a simple weighted graph with a possibly uncountable vertex set V and the weight structure R. Let V' be a finite-order node arrangement design, that is $V' \subseteq V$ and $|V'| < \infty$. Let $\delta(V')$ be the smallest edge weight among the weights of the corresponding node induced graph $(V', R|_{V' \times V'})$ and $\Delta(V')$ be its largest edge weight:

$$\delta(V') := \min\{r(u, v) \,|\, (u, v) \in V' \times V'\},$$

$$\Delta(V') := \max\{r(u, v) \,|\, (u, v) \in V' \times V'\}.$$

Let us assume, as we did before, that the edge-probability function $p(\cdot, \cdot)$ satis-

fies Assumptions 1.2.1-1.2.2. That is

$$p(r,\theta) \to 0, \text{ as } r \to \infty \text{ for any fixed } \theta \in \Theta, \tag{4.1}$$

$$p(r,\theta) \to 1, \text{ as } r \to 0 \text{ for any fixed } \theta \in \Theta. \tag{4.2}$$

As previously, the set of all n-node configuration designs is denoted by $\mathcal{D}^{(n)}$.

Theorem 4.1.1. *If a sequence of random graph designs $d_k = V_k \in \mathcal{D}^{(n_k)}$, $k = 1, 2, \ldots$, is such that (i) the sequence $\{n_k \in \mathbb{N}\}_{k=1}^{\infty}$ is bounded, and (ii) either $\delta(V_k) \to \infty$ or $\Delta(V_k) \to 0$ when $k \to \infty$, then $U_{KL}(d_k) \to 0$ (for any proper prior distribution for θ on Θ).*

Proof. We will first prove the theorem for the case when $\{n_k\}_{k=1}^{\infty}$ is a constant sequence, that is when $n_k \equiv n$ for all $k \in \mathbb{N}$.

The joint edge distribution corresponding to the design $d_k = V_k$ is as follows:

$$f(y \,|\, \theta, d_k) = \prod_{\{u,v\} \in V_k \otimes V_k} [p(r(u,v),\theta)]^{y(u,v)} \, [1 - p(r(u,v),\theta)]^{1-y(u,v)}, \tag{4.3}$$

where y is as defined by (3.34) and (3.35), and the product is taken over all unordered pairs of vertices from V_k. Under condition (ii) this distribution converges, due to (4.1) and (4.2), to a one-point mass distribution. Moreover,

- if $\delta(V_k) \to \infty$, then the limiting one-point distribution is concentrated at $y_\delta = (0, 0, \ldots, 0) \in \mathbb{R}^{n(n+1)/2}$ (all $n(n+1)/2$ edges are absent);
- if $\Delta(V_k) \to 0$, then the limiting one-point mass is concentrated at $y_\Delta = (1, 1, \ldots, 1) \in \mathbb{R}^{n(n+1)/2}$ (all $n(n+1)/2$ edges are present).

Thus, under condition (ii) the sequence $\{f(y \,|\, \theta, d_k)\}_{k=1}^{\infty}$ converges to a one-point distribution, and hence the sequence of corresponding entropies converges to zero:

$$\mathrm{Ent}\{f(y \,|\, \theta, d_k)\} \to 0, \; k \to \infty \; \forall \theta \in \Theta.$$

Analogously, it is due to (3.34) and (3.35) that the sequence of prior predictive distributions (marginal distributions of y), $\{\Phi(y \,|\, d_k)\}_{k=1}^{\infty}$, corresponding to $\{f(y \,|\, \theta, d_k)\}_{k=1}^{\infty}$ (under any prior distribution $\pi(\theta)$) converges to a one-point mass

distribution (concentrated at y_δ if $\delta(d_k) \to \infty$ or at y_Δ if $\Delta(d_k) \to 0$). Consequently, the sequence of the entropies of the prior predictive distributions vanishes when k grows:

$$\mathrm{Ent}\{\Phi(y \,|\, d_k)\} \to 0,\ k \to \infty.$$

One can apply now Lemma 3.2.5 (representation (3.19) of the expected utility based on the KL divergence) to deduce that $U_{\mathrm{KL}}(d_k) \to 0$ as follows:

$$\lim_{k\to\infty} U_{\mathrm{KL}}(d_k) = \lim_{k\to\infty} \left(\mathrm{Ent}\{\Phi(y \,|\, d_k)\} - \mathbb{E}_\theta \left[\mathrm{Ent}\{f(y \,|\, \theta, d_k)\}\right]\right) = 0 - 0 = 0.$$

We consider now the case when the sequence of designs $d_k = V_k$ is such that the sequence of their orders, $\{n_k\}_{k=1}^{\infty}$, is bounded:

$$\exists n \in \mathbb{N} :\ |n_k| \leq n\, \forall k \in \mathbb{N}.$$

We construct a new sequence of designs $\tilde{d}_k = \{\tilde{V}_k\}$ and 'modify' correspondingly the weight function r as follows:

$$\tilde{V}_k := V_k \cup \{u_k^{(1)}, u_k^{(2)}, \ldots, u_k^{(n-n_k)}\},$$

where the added nodes $u_k^{(1)}, u_k^{(2)}, \ldots, u_k^{(n-n_k)}$ are some formal 'fictitious' nodes such that

- $r(u_k^{(i)}, u) = +\infty$, $i = 1, \ldots, n - n_k\ \forall u \in V_k\, \forall k \in \mathbb{N}$, if $\delta(V_k) \to \infty$;
- $r(u_k^{(i)}, u) = 0$, $i = 1, \ldots, n - n_k\ \forall u \in V_k\, \forall k \in \mathbb{N}$, if $\Delta(V_k) \to 0$.

It is clear that all the designs in the sequence $\tilde{d}_k = \tilde{V}_k$ have the same order, n, and that $U_{\mathrm{KL}}(\tilde{d}_k) = U_{\mathrm{KL}}(d_k)$ for any $k \in \mathbb{N}$. Applying the proved result to the sequence of same order designs $\tilde{d}_k$ we complete the proof of the theorem in its general form (that is when $\{n_k\}$ is an arbitrary bounded sequence of design orders):

$$U_{\mathrm{KL}}(d_k) = U_{\mathrm{KL}}(\tilde{d}_k) \to 0,\ k \to \infty.$$

Alternatively, one can consider all **infinite** index sequences $\{k_i(m)\}$ characterised as follows:

$$n_{k_i(m)} = m,\ i = 1, \ldots \infty,\ m \in \{1, \ldots, n\}.$$

By the first part of the theorem

$$U_{\mathrm{KL}}(d_{k_i(m)}) \to 0,\ i \to \infty,$$

for each of these index sequences, and hence the same is true for the parent sequence $\{n_k\}$:

$$U_{\mathrm{KL}}(d_k) \to 0,\ k \to \infty.$$

The proof is complete. □

Since the Kullback–Leibler divergence, and hence its expectation, is a non-negative information quantity, the two scenarios under which the size of the design graph is either indefinitely growing or diminishing in the limit are in a certain sense the worst case design scenarios.

Corollary 4.1.2. *The function* $W(d) = U_{KL}(d) - \mathrm{Ent}\{\pi(\theta)\}$ *has an asymptote* $W = -\mathrm{Ent}\{\pi(\theta)\}$ *when either* $\delta(d) \to \infty$ *or* $\Delta(d) \to 0$.

4.2 Optimal designs for basic random graphs

The purpose of this section is to illustrate some aspects and main difficulties related to the evaluation of the expected utility. In addition, we investigate the simplest models of random graphs for which it is possible to solve the utility based optimal design problem either analytically/numerically, or by combining both analytical and numerical techniques.

4.2.1 Two-node design and prior entropy asymptote of the expected utility

Let us consider two points u and v taken in the Euclidean plane $\mathbb{R}^2$ and form a random graph on these points as nodes as follows: there is a link between the nodes u and v with probability $e^{-\theta d(u,v)}$, where $d(u,v)$ is the distance between them and $\theta \in \mathbb{R}_+$ is the rate of the exponential edge-probability decay.

Given a prior distribution $\pi(\theta)$ the optimal design problem, namely to find such points u^* and v^* that the expected KL divergence between the posterior and the

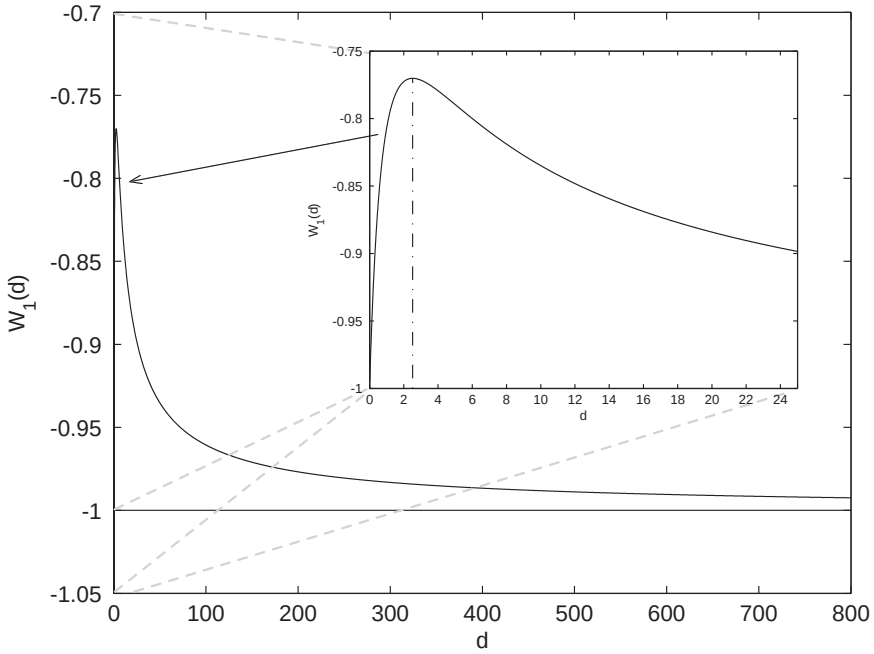


Figure 4.1: Function $W_\alpha(d)$ defined in (4.5) when $\alpha = 1$. This function attains its maximum at the point $d \approx 2.52$.

prior is maximised, can be posed using the formalism presented in §§ 3.5.1, 3.5.2 and the model discussed in Example 3.5.1 (when $n = 2$). We identify one of the vertices of our random graph, say u, with the origin and allow the other, v, to be chosen on $\mathbb{R}_+$. Consequently, the optimal progressive design d^* in this case is the distance d of v from the origin at which the corresponding expected KL divergence

$$U_{\mathrm{KL}}(d) = \int\limits_{\mathbb{R}_+} \log \frac{\pi(\theta \,|\, d, y=1)}{\pi(\theta)} e^{-\theta d} \pi(\theta)\, \mathrm{d}\theta + \int\limits_{\mathbb{R}_+} \log \frac{\pi(\theta \,|\, d, y=0)}{\pi(\theta)} [1 - e^{-\theta d}] \pi(\theta)\, \mathrm{d}\theta$$

is maximised.

It is due to the identity (3.27) that

$$\begin{aligned} U_{\mathrm{KL}}(d) - \mathrm{Ent}\{\pi(\theta)\} = &\int\limits_{\mathbb{R}_+} \log \pi(\theta \,|\, d, y=1) e^{-\theta d} \pi(\theta)\, \mathrm{d}\theta \\ &+ \int\limits_{\mathbb{R}_+} \log \pi(\theta \,|\, d, y=0) [1 - e^{-\theta d}] \pi(\theta)\, \mathrm{d}\theta. \end{aligned} \tag{4.4}$$

Let us assume that the prior distribution of θ is exponential with parameter $\alpha > 0$: $\pi \sim \mathrm{Exp}(\alpha)$. After introducing the following notation

$$W_\alpha(d) := U_{\mathrm{KL}}(d; \alpha) - \mathrm{Ent}\{\mathrm{Exp}(\alpha)\} \tag{4.5}$$

it is straightforward to obtain the following identity:

$$\frac{1}{\alpha}W_\alpha(d) = \frac{1}{\alpha+d}\log(\alpha+d) - \frac{1}{\alpha+d} + \frac{d}{\alpha}\frac{1}{\alpha+d}[\log\alpha + \log(1+\frac{\alpha}{d})] +$$

$$+ \int_0^\infty \log[1-e^{-\theta d}](1-e^{-\theta d})e^{-\alpha\theta}\mathrm{d}\theta - \frac{d}{\alpha}\frac{d+2\alpha}{(\alpha+d)^2}. \tag{4.6}$$

It is also easy to verify that

$$W_{t\alpha}(td) = W_\alpha(d) + \log t, \tag{4.7}$$

and therefore

$$\arg\max_{d\geq 0} W_{t\alpha}(d) = t \cdot \arg\max_{d\geq 0} W_\alpha(d). \tag{4.8}$$

Thus, it suffices to maximise $W_1(d)$ in order to find maximum of W_α for any $\alpha > 0$.

The integral in (4.6) can be calculated analytically when $\alpha/d \in \mathbb{N}$ as an integral of polylogarithms and it can be further reduced to the digamma function for any non-negative α and d (see Appendix C). The numerical evaluation of this integral suggests that $W_1(d)$ is maximised at $d \approx 2.51895$ (see Figure 4.1). One can also notice that the graph of W_α, depicted in Figure 4.1, closely approaches the horizontal line at the level $-\mathrm{Ent}\{\mathrm{Exp}(1)\} = -1$ when d increases—a fact that is indeed expected in view of Corollary 4.1.2. This kind of verification can be suggested as a validation tool of checking whether numerical evaluation of involved integrals has been implemented correctly.

The particular model we have just considered represents an example combining a very simple random graph design model and a simple prior distribution—and yet, one should acknowledge, the fully analytical evaluation of the expected utility was not a very elementary task. It is clear that in more general models with more complex prior distributions and edge-probability functions the analytical treatment of the expected utility would generally not be possible.

Finally notice that for any given α the expected utility $U_{\mathrm{KL}}(d;\alpha)$ is a unimodal function for the function $W_\alpha(d)$ is so. It is interesting to ask in this regard whether there exist examples of multimodal expected utilities when d is a one-dimensional (multidimensional) design parameter? If yes, can it be derived a characterisation of cases when the expected utility is a globally unimodal function? The advantage of such characterisation in the multidimensional case, for example, would be

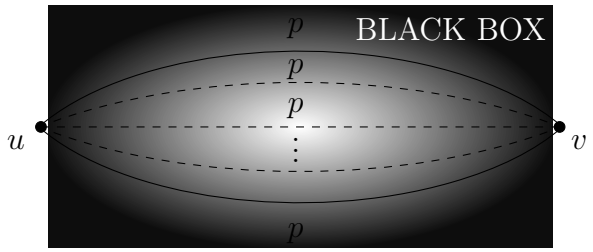


Figure 4.2: Two-node random multigraph (with no loops) in a black box: n multiple edges connect two sites u and v, each being open with probability p independently of the status of any other edge; it can be only observed whether the nodes are connected or not but not the total number of open edges.

obvious—since the expected utility is a symmetric function with respect to any permutation of its arguments (elements of the design parameter $d = (d_1, \ldots, d_{|\mathcal{D}|})$), unimodality will imply that all the components of the optimal design d^* should be taken equal: $d_1^* = \ldots = d_{|\mathcal{D}|}^*$. This property could greatly simplify the search for the optimal design.

4.2.2 Progressive and instructive designs: two-node 'black box' design example

In order to illustrate the concepts of the progressive and instructive designs introduced in Section 3.3 we consider a **multigraph** on two nodes without loops. Assume that there are n edges between two vertices u and v and each edge is open with probability p independently of the status of any other edge. The vertices are considered to be connected if, and only if, there is at least one open edge connecting them. The number of open edges, however, cannot be observed, thus representing a kind of a black box (Figure 4.2). That is, all that can be observed for a given (and known!) n is whether u and v are connected or not.

Suppose that there is an experimenter A whose prior knowledge about p is expressed in a prior distribution $\pi(p)$. The experimenter A, however, wishes to find the optimal number of links, n, to equip the black box with in order to maximise the increase in his or her knowledge about p after observing whether u and v are connected or not. Thus, in this example n is the design parameter and p

is the model parameter that the experimenter A would like to make inference on.

Clearly, the sites u and v are connected with probability $1-(1-p)^n$, so that, in line with the settings of § 3.3.1, the experimenter maximises the corresponding expected utility:

$$U_{\mathrm{KL}}(n) = \int_0^1 \log \frac{\pi(p \,|\, y=1,n)}{\pi(p)} [1-(1-p)^n] \pi(p)\, \mathrm{d}p$$

$$+ \int_0^1 \log \frac{\pi(p \,|\, y=0,n)}{\pi(p)} (1-p)^n \pi(p)\, \mathrm{d}p$$

$$= \int_0^1 \log \pi(p \,|\, y=1,n) [1-(1-p)^n] \pi(p)\, \mathrm{d}p \tag{4.9}$$

$$+ \int_0^1 \log \pi(p \,|\, y=0,n) (1-p)^n \pi(p)\, \mathrm{d}p + \mathrm{Ent}\{\pi(p)\}. \tag{4.10}$$

where

$$y = \begin{cases} 1, & \text{if } u \text{ and } v \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases}$$

Assume that the prior $\pi(p)$ is modelled by a beta distribution: $\pi(p) \sim \mathrm{Beta}(\alpha, \beta)$. In this case its entropy can be calculated as follows:

$$\mathrm{Ent}\{\pi(p)\} = \log B(\alpha,\beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta),$$

where ψ is the *digamma function*, $\psi(z) = \Gamma'(z)/\Gamma(z)$, and the integration in (4.9-4.10) with respect to the prior distribution can be performed numerically.

Figure 4.3 depicts plots of the function $U_{\mathrm{KL}}(n) - \mathrm{Ent}\{\pi(p)\}$ when $\pi(p) \sim \mathrm{Beta}(\alpha,\alpha)$, $\alpha = 1,2,3,4$. It is important to note that the plots were produced after numerically evaluating the integrals in (4.9-4.10) and that the horizontal asymptotic behaviour, when $n \to \infty$, is the expected behaviour which can be validated by plotting the horizontal line that corresponds to $\mathrm{Ent}\{\pi(p)\}$. This observation suggests the value of plotting the result of integration and expected prior entropy asymptote as a basic tool for checking whether the integration was carried out correctly or not whenever the prior entropy can be easily calculated.

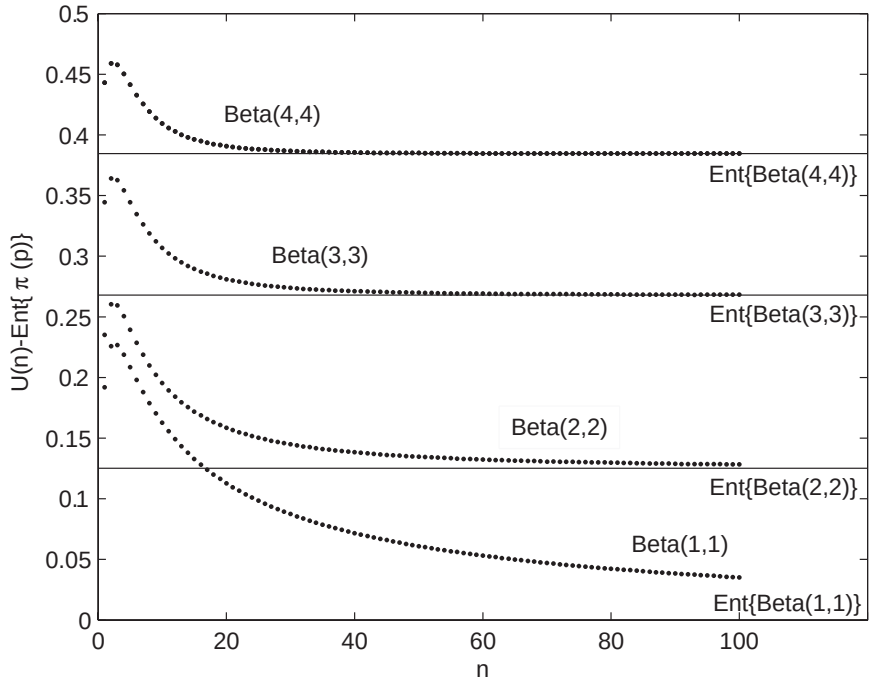


Figure 4.3: Expected utility (expected KL divergence) of the experimenter A holding a beta prior for p, Beta(α, α), minus the entropy of the prior distribution.

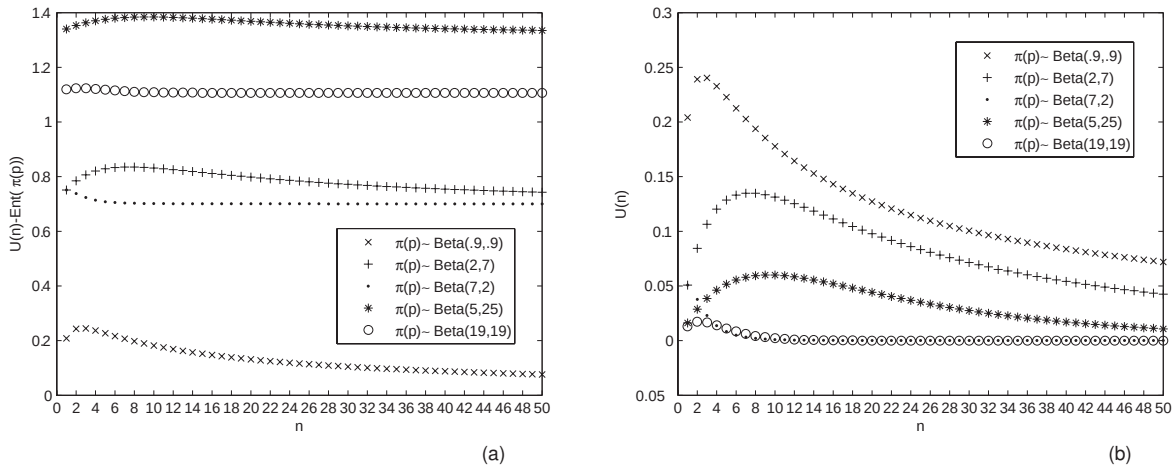


Figure 4.4: (a) Expected utility (expected KL divergence) of the experimenter A holding a beta prior for p, Beta(α, β) (various sets of values for α and β), minus the entropy of the prior distribution; (b) Expected utility plots for the prior distributions considered in the left plot.

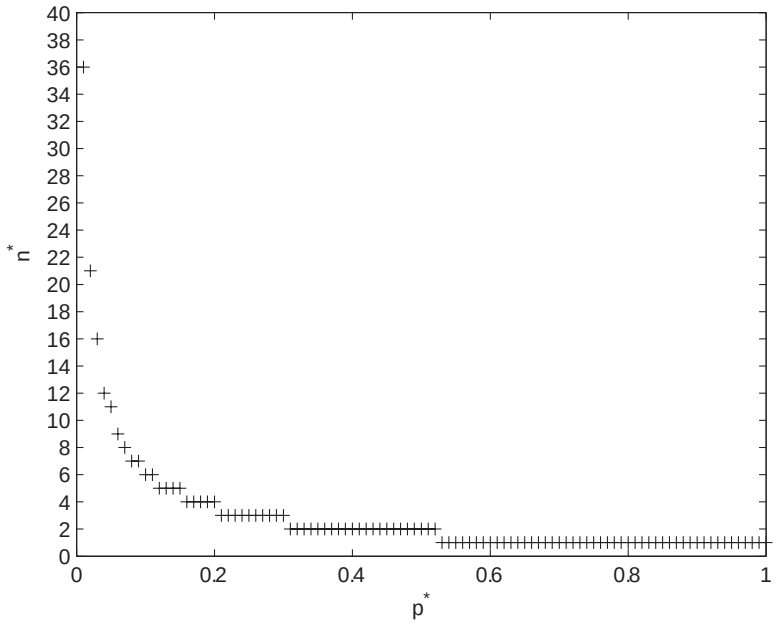


Figure 4.5: Optimal values of n, n^*, derived by maximising the expected KL divergence calculated by B (who knows the exact value of the parameter p, p^*) for the experimenter A holding a uniform prior for p.

However, one should interpret these plots with care. Such combined plots for different prior distributions allow one to read the optimal design parameter's values and compare them for the same prior distribution. However, it is not feasible to compare different designs corresponding to different prior distributions using such plots. Such a comparison can be done by plotting graphs of the expected utility $U_{\mathrm{KL}}(n)$ corresponding to the prior distributions of interest. Examples provided in Figure 4.4 illustrate this idea.

Now let us assume that there is an instructor B who knows the exact value of p, p^*. Acting in accordance with (3.28-3.29), the instructor, in order to find the best 'convincing' design for the experimenter A, should maximise the expected utility

$$U^*_{\mathrm{KL}}(n) = [1-(1-p^*)^n] \int_0^1 \log \frac{\pi(p \,|\, y=1,n)}{\pi(p)} \pi(p \,|\, y=1,n)\, \mathrm{d}p$$

$$+ (1-p^*)^n \int_0^1 \log \frac{\pi(p \,|\, y=0,n)}{\pi(p)} \pi(p \,|\, y=0,n)\, \mathrm{d}p.$$

Figure 4.5 shows the dependence of the optimal number of edges, n, on the true value of p known to B when the experimenter A is ignorant about p and chooses

to represent this using a uniform prior (and thus $\pi(p)$ is a uniform distribution).

Finally notice that the discussed experiment can be easily modelled using simple weighted graphs within the formalisation framework presented in § 3.5.1. For instance, if $V \equiv \mathbb{Z}$, $R = (r(i,j))_{i,j \in V}$ is such that $r(i,j) := |i - j|$, and $p(r(i,j), \theta) := (1-\theta)^{r(i,j)}$, $\theta \in \Theta \equiv [0,1]$, then the solution to the optimal design problem with two-node configuration design space $\mathcal{D}^{(2)}$, $d^* = (i^*, j^*)$, relates to the optimal number of edges from the black box, n^*, as follows: $n^* = |i^* - j^*|$.

4.2.3 Three-node star design with two independent edges

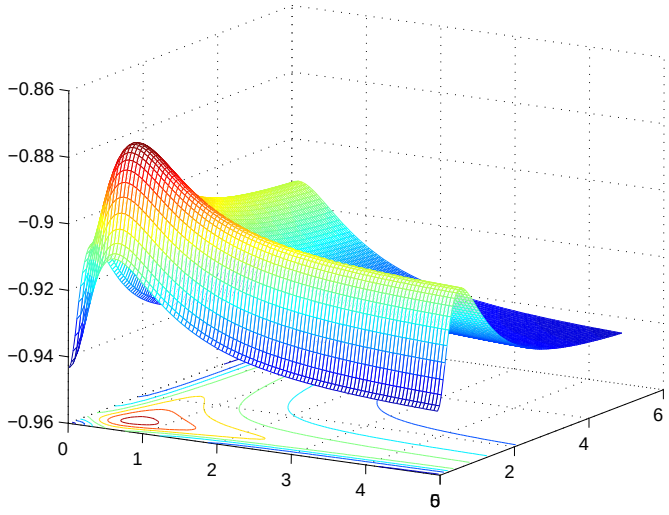


Figure 4.6: Expected utility minus prior entropy surface for the Cauchy edge-probability function (see § 1.2.2); here θ is assumed to take values 1, 2, and 5 with probabilities 0.1, 0.5, and 0.4, respectively. Note that $-\mathrm{Ent}\{\pi(\theta)\} = 0.1 \log 0.1 + 0.5 \log 0.5 + 0.4 \log 0.4 \approx -0.94$, and this is in agreement with the plot (which in turn reflects the statement of Corollary 4.1.2). Horizontal axes correspond to the lengths of the edges, d_1 and d_2.

So far we encountered only unimodal expected utilities: in particular, the expected utility function in the univariate design problem considered in § 4.2.1 was unimodal. Figure 4.6 shows the landscape of the expected utility for the problem with two independent random edges of lengths d_1 and d_2 (three nodes and a

star interaction topology) and the Cauchy edge-probability function—the expected utility surface is again unimodal.

Are there multimodal expected utilities? Here we answer positively this question by providing simple examples of two-dimensional expected utility surfaces which are multimodal (not only for a utility function based on KL divergence). In fact the character of the multimodality can be quite diverse. Figure 4.7 depicts plots of two-dimensional expected utility surfaces (as functions of d_1 and d_2) as well as projections of isolines onto the design space for various prior distributions and edge-probability functions.

Figure 4.7(a) shows a unimodal expected utility surface corresponding to a power-law edge-probability decay and an exponential prior distribution for θ. The plot is similar to that shown in Figure 4.6 and, indeed, our experience is that expected utility landscapes for exponential and Cauchy decays have similar shapes. The plot of the expected negative squared error loss (p. 48) under logistic decay (p. 11) presented in Figure 4.7(b) exhibits a global mode as well as two local maxima. In addition, and it is easily noticeable, the surface is fairly flat around the modes. This is also often the case with expected Kullback–Leibler divergence surfaces. The plot of the expected KL divergence from Figure 4.7(c) is not as flat around its mode as that of Figure 4.7(b) but similar otherwise; it corresponds to the logistic decay (p. 11).

The plot of the expected KL divergence under a 'linear' edge-probability decay

$$p(r, \theta) = (1 - \theta r)\mathbb{1}_{\{r \leq 1/\theta\}}$$

and a discrete prior for θ presented in Figure 4.7(d) suggests that there are just two modes in total. As a consequence of this (and the fact that the function of interest is symmetric) the modes are global and any designs on the line $d_1 = d_2$ are far from being 'good' in this case. This is not the case in Figure 4.7(e) which depicts a plot of the expected KL divergence with the edge-probability decay

$$p(r, \theta) = 1 - \left(1 + e^{(10-r)/\theta}\right)^{-1}$$

and discrete prior distribution for θ—there are four equally significant modes in this case and two of them are achieved at points from the line $d_1 = d_2$.

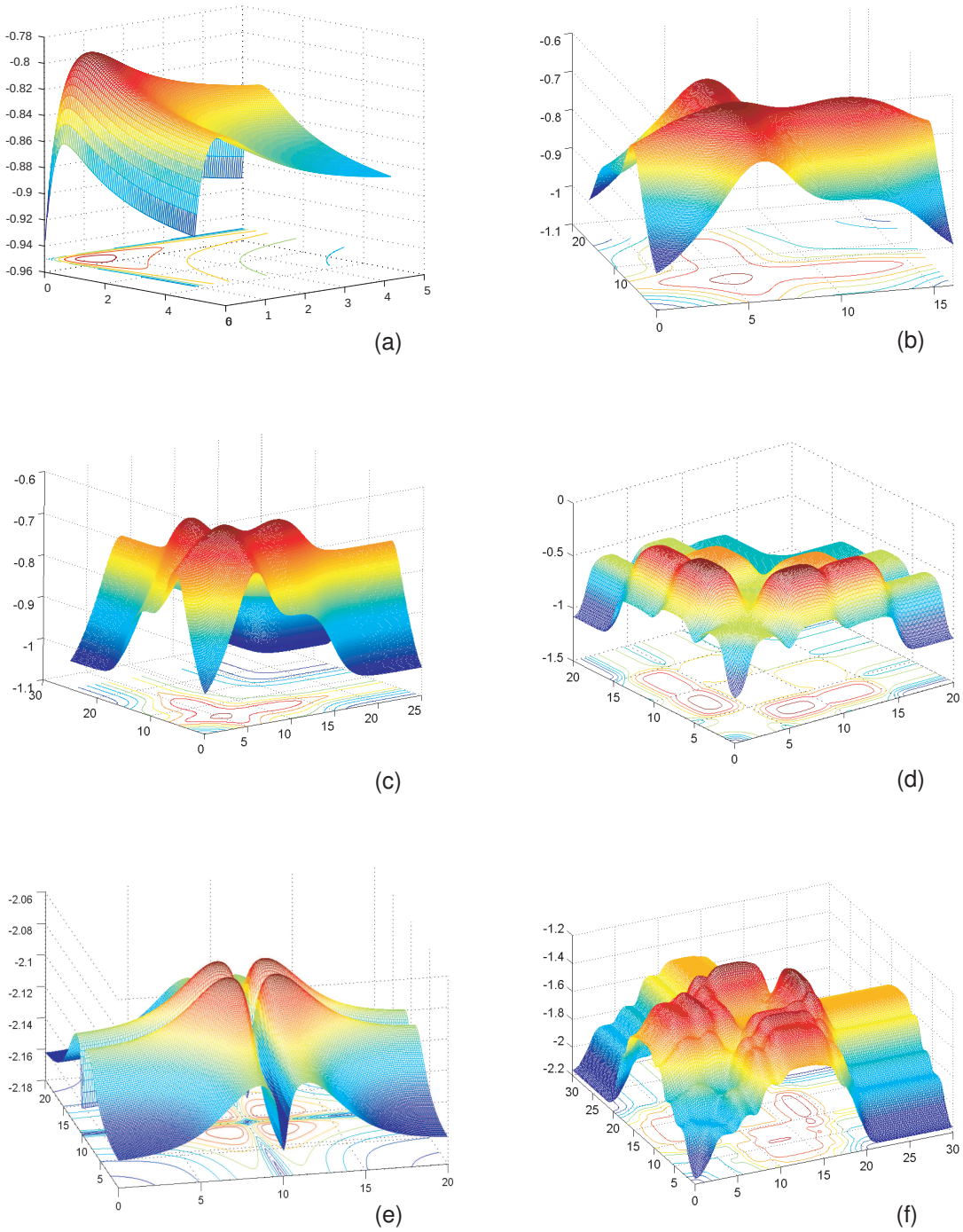


Figure 4.7: Expected utility minus prior entropy plots for two independent random edges and **(a)** KL divergence and power-law decay (exponential and Cauchy decays give similar unimodal surfaces); **(b)** negative squared error loss utility and logistic decay; **(c)** KL divergence and logistic decay, **(d)** KL divergence and a 'linear' decay $p(r,\theta) = (1-\theta r)\mathbb{1}_{\{r \leq \theta^{-1}\}}$ with a discrete distribution for θ over a finite set of points; **(e)** KL divergence and $p(r,\theta) = 1 - \left(1 + e^{(10-r)/\theta}\right)^{-1}$ with a discrete distribution for θ over a finite set; **(f)** KL divergence and $p(r,\theta) = 1 - \left(1 + e^{(10\theta - r)/0.3}\right)^{-1}$ with the same prior for θ as in (e).

Finally, the plot from Figure 4.7(f) corresponding to the expected KL divergence for the edge-probability decay

$$p(r, \theta) = 1 - \left(1 + e^{(10\theta - r)/0.3}\right)^{-1}$$

and the same discrete prior as in previous example differs dramatically from its counterpart in Figure 4.7(e). The 'hills' containing the two global maxima (at points different from any pairs $d_1 = d_2$) are very thin in one direction and very flat along another direction (these directions are different, of course, for each of these modes).

The examples considered above show wide versatility in the shapes of the expected utility surface as well as the geometry of its maxima. With increasing number of the experimental graph vertices the complexity of the optimisation problem will only grow: potential analytical intractability suggests that numerical or simulation techniques of the expected utility evaluation may be of greater use; however considerable 'flatness' of the expected utility function (in high dimensions) may diminish the power of these techniques. Reducing the design space by decreasing its dimensionality is one solution to overcome this difficulty. Deferring a more detailed discussion on this until next chapter we move next to the study of proximity random graph models and closely related models, for which the expected utility has multiple global modes but is analytically tractable—hence no reduction in dimensionality is needed.

4.2.4 Proximity graphs

Proximity graphs or *geometric graphs* are graphs consisting of nodes placed in n-dimensional metric space, with edges connecting only those pairs of nodes which are in some sense close to each other (Penrose (2003)). Given a finite subset V of a metric space X with a metric r and a distance threshold θ, the unoriented graph with vertex set V and undirected edges connecting those pairs $\{u, v\} \subseteq V$ for which it is true that $d(u, v) \leq \theta$ is a geometric graph. The set V may in particular be a subset of $\mathbb{R}^m$. When $m = 1$ the resulting graphs are **related** to what is known as *interval graphs* (Golumbic (2004)); when $m = 2$ proximity graphs are known

as *disk* graphs (Penrose (2003)). Since proximity graphs are essentially defined by looking at intersections of neighbourhoods of their vertices, these graphs can be seen as a particular case of *set intersection graphs* (Fulkerson and Gross (1965)).

Let us consider an optimal design problem for graphs with star topology (as in Example 3.5.1) and the 0-1 step edge-probability function:

$$p(d(u,v),\theta) = \mathbb{1}_{\{d(u,v)\leq\theta\}},\ \theta\in\Theta\subseteq\mathbb{R}_+, \tag{4.11}$$

where $d(u,v)$ is a weight attributed to the pair $\{u,v\}$. Here, as before, the function $p(d,\theta)$ denotes the probability for any pair of two vertices with weight d to be connected given the value of the model parameter θ:

$$p(d(u,v),\theta) := \mathbb{P}(u \text{ and } v \text{ are connected} \,|\, \theta).$$

Note that for a fixed value of θ the resulting graph is not random.

Recall that the optimal arrangement $d^*\in\mathcal{D}^{(n)}$ for the n-node optimal design problem in the case of a star topology always contains the origin as one of the nodes (Example 3.5.1), that is we may assume that the design parameter d is as follows:

$$d=\{0,d_1,\ldots,d_{n-1}\},\ d_i\in\mathbb{R}_+\ \forall i=1,\ldots,n-1.$$

Let y_i be a binary variable which takes value 1 whenever there is an edge between a node u_i placed at the distance d_i from the origin (as a centre of the star), and it takes value 0 otherwise. The likelihood function of the model parameter θ given an (essential) observation $y=(y_1,\ldots,y_{n-1})$ is as follows then:

$$f(y\,|\,\theta,d) = \prod_{i=1}^{n-1}\left(y_i\mathbb{1}_{\{d_i\leq\theta\}} + (1-y_i)\mathbb{1}_{\{d_i>\theta\}}\right). \tag{4.12}$$

Example 4.2.1. *Let $n=3$ and let us assume that the prior for θ is exponential distribution with parameter λ:*

$$\pi(\theta) = \lambda\mathbb{1}_{\{\theta\geq 0\}}e^{-\lambda\theta}.$$

In this case

$$W(d_1,d_2) = U_{KL}(d_1,d_2) - \mathrm{Ent}\{\pi(\theta)\} = \sum_{y\in\{0,1\}^2:\,y\neq(0,1)}\int\limits_{\Theta_y}\log\frac{\pi(\theta)}{\int\limits_{\Theta_y}\pi(\phi)\,d\phi}\pi(\theta)\,d\theta, \tag{4.13}$$

where

$$\Theta_y = \begin{cases} [0, d_1), & \text{if } y = (0,0), \\ [d_1, d_2), & \text{if } y = (1,0), \\ [d_2, \infty), & \text{if } y = (1,1), \end{cases}$$

and $d_1 \leq d_2$. *Notice that the outcome* $y = (0,1)$ *is an impossible event.*

Thus,

$$W(d_1, d_2) = \log\{1 - e^{-\lambda d1}\}[e^{-\lambda d_1} - 1] + \log\{e^{-\lambda d_1} - e^{-\lambda d_2}\}[e^{-\lambda d_2} - e^{-\lambda d_1}] + \lambda d_2 e^{-\lambda d_2},$$

and the partial derivatives of W *are as follows:*

$$\frac{\partial}{\partial d_1} W = \lambda e^{-\lambda d_1} \log \frac{e^{-\lambda d_1} - e^{-\lambda d_2}}{1 - e^{-\lambda d_1}}$$

$$\frac{\partial}{\partial d_2} W = -\lambda e^{-\lambda d_2} [\log\{e^{-\lambda d_1} - e^{-\lambda d_2}\} + \lambda d_2], \tag{4.14}$$

so that the stationary point is $d_{st} = \left(\frac{\log 3/2}{\lambda}, \frac{\log 3}{\lambda}\right)$. *This is the point of maximum of* W, *and hence of the expected utility* U_{KL}.

Recall that by a *quantile of order* p, $p \in (0,1)$, of a random variable X with distribution function $F_X(x)$ one understands any number q_p such that $F_X(q_p) \leq p$ and $F_X(q_p+) \geq p$. One can easily recognise quantiles of the orders $1/3$ and $2/3$ of the exponential distribution with parameter λ in the values of the optimal distances found in Example 4.2.1. As the following theorem shows this fact is not a coincidence.

Let us assume that the prior distribution of θ is given by a probability density function $\pi(\theta)$, and that θ has a non-negative support $\operatorname{supp}\theta = \Theta \subseteq \mathbb{R}_+$.

Theorem 4.2.2. *Solution to the* n*-node optimal design problem for a proximity graph under star topology is given by* $n-1$ *quantiles of the prior* $\pi(\theta)$ *dividing this distribution into* n *even parts.*

Proof. Without loss of generality we assume that the distances are ordered as follows:

$$d_0 := 0 \leq d_1 \leq d_2 \ldots \leq d_{n-1} \leq d_n := +\infty.$$

The design vector d naturally defines the following discrete probability distribution $P_{d\,|\,\pi} := \{p_1, \ldots, p_n\}$:

$$p_i := \mathbb{P}(d_{i-1} \leq \theta < d_i) = \int_{d_{i-1}}^{d_i} \pi(\theta)\,\mathrm{d}\theta,\ i = 1, \ldots, n,$$

where some p_i's are possibly zeros (this will happen if some d_i's lie outside Θ).

The space of observables $\mathcal{Y}$ can be described as the following set of n zero-one $(n-1)$-tuples:

$$\mathcal{Y} = \{(0, \ldots, 0), (1, 0, \ldots, 0), (1, 1, 0, \ldots, 0), \ldots, (1, \ldots, 1, 0, \ldots, 0) \ldots, (1, \ldots, 1)\}.$$

Denoting by Δ_y the interval $[d_{i-1}, d_i]$, where i is the place of the first zero in $y \in \mathcal{Y}$, one obtains:

$$\pi(\theta\,|\,y) = \mathbb{1}_{\{\theta \in \Delta_y\}}\pi(\theta)/p_i,\ i = 1, \ldots, n,$$

so that

$$U_{\mathrm{KL}}(d) = \sum_{y \in \mathcal{Y}} \int_{\Delta_y} \log \frac{\pi(\theta\,|\,y)}{\pi(\theta)} \pi(\theta)\,\mathrm{d}\theta = \sum_{i=1}^{n} \int_{d_{i-1}}^{d_i} \log p_i^{-1} \pi(\theta)\,\mathrm{d}\theta = \mathrm{Ent}\{P_{d\,|\,\pi}\}. \quad (4.15)$$

The entropy of the discrete distribution $P_{d\,|\,\pi} := \{p_1, \ldots, p_n\}$ is maximised when $p_1 = \ldots = p_n$. This means that the points $d_1, \ldots, d_{n-1}$ divide $\pi(\theta)$ into n even parts, that is $d_1, \ldots, d_{n-1}$ are quantiles of the prior distribution π of the orders $1/n, \ldots, (n-1)/n$ respectively. The theorem is proved. □

This theorem gives a clear recipe for solving the problem of n-node optimal arrangement design in the case of underlying star topology (or, equivalently, in the case of n independent pair of vertices). In the case when the design space assumes, for example, embedding in a metric space, one can be suggested to maximise the function $\mathrm{Ent}\{P_{d\,|\,\pi}\}$ under triangle inequalities imposed on correspondingly related distances, that is to solve the following optimisation problem with linear

constraints[1]:

$$\textit{Maximise } \mathrm{Ent}\{P_{d\,|\,\pi}\},$$

$$\textit{subject to } 0 \leq d_{ij} \leq d_{ik} + d_{kj},$$

$$1 \leq i, j \leq n,\ k = 1, \ldots, n \,\&\, k \neq i \neq j \neq k.$$

Example 4.2.3. *Let $n = 4$ and $\pi \sim \Gamma(3.4, 0.5)$. The solution to the following optimisation problem*

$$\textit{Maximise } \mathrm{Ent}\{P_{d\,|\,\pi}\},$$

$$\textit{subject to } 0 \leq d_{ij} \leq d_{ik} + d_{kj},$$

$$1 \leq i, j \leq 4,\ k = 1, 2, 3, 4 \,\&\, k \neq i \neq j \neq k,$$

obtained using the MATLAB function `fmincon` *(Optimization Toolbox) is as follows:*

$$d^* = (d^*_{34}, d^*_{24}, d^*_{23}, d^*_{14}, d^*_{13}, d^*_{12}) \approx (0.99, 1.47, 1.86, 2.40, 2.85, 3.87).$$

For comparison, the vector of quantiles, and hence the solution to the optimal design problem with star topology and $n = 7$, of orders $1/7, 2/7, \ldots, 6/7$ for the gamma distribution with parameters 3.4 and 0.5 is as follows

$$q = (q_{1/7}, q_{2/7}, q_{3/7}, q_{4/7}, q_{5/7}, q_{6/7}) \approx (1.095, 1.384, 1.696, 2.080, 2.656).$$

These values were also obtained numerically. Figure 4.8 shows the plot of the prior density function $\pi(\theta)$ and the solutions obtained.

In Example 4.2.3 the solution d^* satisfying the metric inequalities is such that

$$d^*_{14} + d^*_{24} = d^*_{12}.$$

Basic geometric considerations show that there is no four-vertex configuration in $\mathbb{R}^3$ with lengths from d^*. A simple procedure of detecting whether six non-negative numbers $d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}$ can be pairwise distances between some four points U_1, U_2, U_3 and U_4 in the Euclidean space $\mathbb{R}^3$ is described in Appendix D.

[1] There will be $3\binom{n}{3}$ triangle inequalities and $\binom{n}{2}$ non-negativeness inequalities—in total $n(n-1)^2/2$ linear constraints.

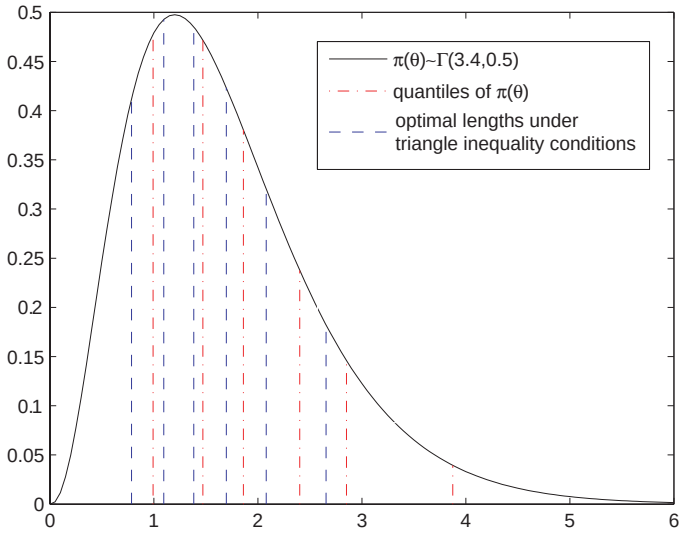


Figure 4.8: Solution to the optimal design problem for proximity graph with and without metric constrains (six edges, see Example 4.2.3).

Further investigation is generally needed to decide whether the obtained solution (represented by a set of pairwise distances) corresponds to any realisable point configuration in case when the metric space of interest is also Euclidean space. Some advanced methods and techniques can be used in answering questions similar to finding the dimension of the Euclidean space in which the obtained design is realisable (e.g. see Vempala (2006)) or, when the distances cannot be preserved exactly, finding an embedding which would preserve distances as much as possible minimising the *distortion* (a measure of preserving distances by a transformation) of the embedding (see Shavitt and Tankel (2004), Gupta (1999) and references therein).

4.2.5 Step-like (threshold) probability decay

Undoubtedly, the geometric graph model considered above can be generalised in a number of various ways. One of them is particularly interesting: it is simple and it permits, similarly to geometric graph case, analytical treatment of the optimal design problem combined with numerical optimisation techniques.

Let α be a non-negative real number which is less than unity: $0 \leq \alpha < 1$.

Consider the following generalisation of a $0-1$ edge-probability function:

$$p(d,\theta) = \mathbb{1}_{\{d\le\theta\}} + \alpha\mathbb{1}_{\{d>\theta\}},\ \theta\in\Theta\subseteq\mathbb{R}_+. \tag{4.16}$$

Notice, that in contrast with proximity graphs, graphs whose links are declared present or absent in accordance with (4.16) with at least one link of length exceeding θ are random *per se*[2].

Because of the invariance of the expected Kullback–Leibler divergence under the change of the model parameter (see p. 53), one can assume that $\theta\sim \mathrm{U}_{[0,1]}$.

Given n nodes and a star interaction topology, that is $n-1$ independent random edges, we order their lengths, just as we did before, adding two fictitious elements:

$$0=:d_0\le d_1\le d_2\le\ldots\le d_{n-1}\le d_n:=1.$$

What we observe is an $(n-1)$-tuple $\mathbf{y}$ of zeros and ones. In addition, we introduce into consideration the following statistic:

$$I(\mathbf{y}) = \begin{cases} k, & \text{where } k \text{ is the place of the first zero in } y,\\ n, & \text{if there are no zeros in } y. \end{cases}$$

The statistic I is well defined (in the sense that it is assigned a value for any possible outcome of $\mathbf{y}$) and sufficient for θ. Note also that the support of the posterior distribution of θ is located to the left of d_k if $I=k$. The posterior distribution remains equivalent to the prior distribution if $I=n$ (the observation 'all edges are present' is a non-informative one).

Let $p_i := d_i - d_{i-1}$, $i=1,...,n$. The posterior distribution of θ is constant in any interval $\Delta_i=(d_{i-1},d_i]$. Therefore, one is interested in the following posterior probabilities:

$$\pi(\theta\in\Delta_s\,|\,I=k) = \frac{\mathbb{P}(I=k\,|\,\theta\in\Delta_s)p_s}{\int\limits_0^1 \mathbb{P}(I=k\,|\,\theta)d\theta}$$

$$= \frac{\mathbb{P}(I=k\,|\,\theta\in\Delta_s)p_s}{\sum\limits_{i=1}^{k}\mathbb{P}(I=k\,|\,\theta\in\Delta_i)p_i},\ s=1,...,k,\ k=1,\ldots,n,$$

[2]For a fixed threshold θ proximity graphs are not random—all the randomness in the design problem comes from the assumption that θ is unknown and distributed according to experimenter's prior belief!

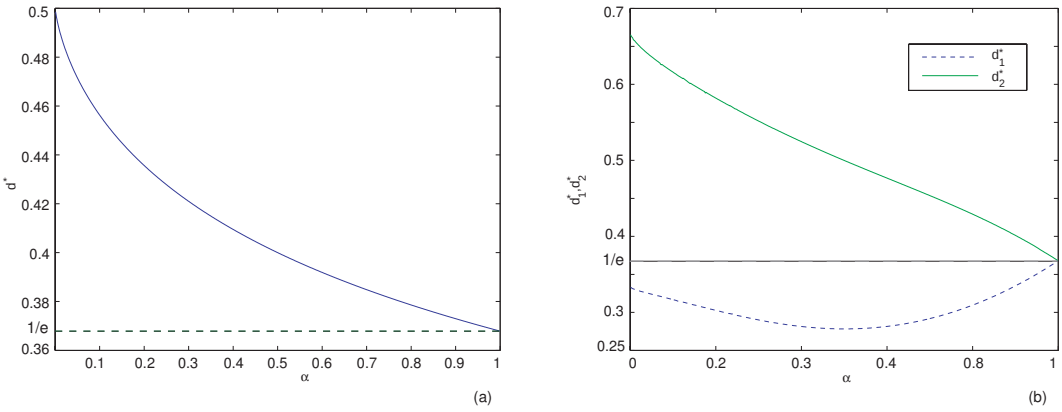


Figure 4.9: Optimal designs as functions of α in the model with threshold edge-probability function: (a) one edge and (b) two independent edges.

where $\mathbb{P}(I = k \,|\, \theta \in \Delta_s) = \alpha^{k-s}(1-\alpha)$.

Since I is a sufficient statistic for θ, one can express the utility function based on the Kullback–Leibler divergence as follows (LIG.3, p. 56):

$$U_{\mathrm{KL}} = \sum_{k=1}^{n} \int_0^1 \log \frac{\pi(\theta \,|\, I = k)}{\pi(\theta)} \mathbb{P}(I = k \,|\, \theta)\pi(\theta)d\theta$$

$$= \sum_{k=1}^{n} \sum_{s=1}^{k} \log \frac{\mathbb{P}(I = k \,|\, \theta \in \Delta_s)}{\sum_{i=1}^{k} \mathbb{P}(I = k \,|\, \theta \in \Delta_i)p_i} \mathbb{P}(I = k \,|\, \theta \in \Delta_s)p_s.$$

Finally, one obtains:

$$U_{\mathrm{KL}}(d) = U_{\mathrm{KL}}(p_1, p_2, \ldots, p_n) = (1-\alpha) \sum_{k=1}^{n-1} \sum_{s=1}^{k} \alpha^{k-s} p_s \log \frac{\alpha^{k-s}}{\sum_{i=1}^{k} \alpha^{k-i} p_i}$$

$$+ \sum_{j=1}^{n} \alpha^{n-j} p_j \log \frac{\alpha^{n-j}}{\sum_{i=1}^{n} \alpha^{n-i} p_i}. \tag{4.17}$$

In particular, when $n = 2$ (a pair of vertices) then

$$U_{\mathrm{KL}}(d) = (\alpha - 1)d \log d + \alpha \ln \alpha - [1 + (\alpha - 1)d] \ln(1 + (\alpha - 1)d). \tag{4.18}$$

Under restriction $d \in [0, 1]$ the expected utility $U_{\mathrm{KL}}(d)$ is maximised at

$$d^*(\alpha) = \frac{\alpha^{\frac{\alpha}{1-\alpha}}}{1 - (\alpha - 1)\alpha^{\frac{\alpha}{1-\alpha}}}. \tag{4.19}$$

Figure 4.9(a) shows the plot of $d^*(\alpha)$. Interestingly enough to notice that

$$\lim_{\alpha \to 1} d^*(\alpha) = 1/e. \tag{4.20}$$

For general n and α the problem of optimal design for the model with threshold edge-probability function can be solved by numerically solving the following optimisation problem:

$$\textit{Maximise } U_{\mathrm{KL}}(p_1, p_2, \ldots, p_n), \tag{4.21}$$

$$\textit{subject to } \sum_{i=1}^{n} p_i = 1, \tag{4.22}$$

$$p_i \geq 0,\ i = 1, \ldots, n, \tag{4.23}$$

where $U_{\mathrm{KL}}(p_1, p_2, \ldots, p_n)$ is taken in the form (4.17).

When $n = 3$ there are two independent edges of lengths d_1 and d_2. Figure 4.9(b) represents 2-edge optimal designs d_1^* and d_2^* as functions of α.

Table 4.1 contains optimal designs which correspond to different values of α from the interval [0,1) in the case of 3 independent edges (n=4). Notably, the optimal designs seem to be planar for any α in this case (but one should keep in mind that θ was taken to be uniformly distributed, and if we reparametrise the model by accordingly transforming θ and its support, the optimal design can be obtained as corresponding quantiles of the new prior distribution of θ and the planarity may be easily 'violated' by such procedure.).

Figure 4.9 and Table 4.1 were obtained by solving numerically the optimisation problem (4.21) with linear constraints (4.22) and (4.23).

Finally notice that it looks very convincing from Figure 4.9 and Table 4.1 that optimal edge lengths $d_i^*(\alpha)$, $i = 1, \ldots, n$, tend to $1/e \approx 0.368$ each, as α goes to 1. We currently do not have analytic proof of this for general values of n.

4.2.6 Non-preservation of optimal designs under replication

Although optimal designs are often maintained under replication in the case of linear (or linearisable) models with normal errors, the following trivial example

α	d_1	d_2	d_3
0.0	0.25	0.5	0.75
0.1	0.2499	0.4759	0.7132
0.2	0.2481	0.4526	0.6872
0.3	0.2435	0.4278	0.6604
0.4	0.2378	0.4040	0.6298
0.5	0.2344	0.3840	0.5967
0.6	0.2365	0.3702	0.5615
0.7	0.2468	0.3633	0.5238
0.8	0.2691	0.3625	0.4819
0.9	0.3077	0.3654	0.4319

Table 4.1: Optimal designs for the model with threshold edge-probability function as functions of the threshold α when $n = 4$.

shows that this is not generally true. A related point that this example shows is that the sequential optimal design of replicated experiments need not be the same as the optimal design of simultaneous replicated experiments.

The following elementary proof is given in Cook et al. (2008) and appeared in discussion with Alex Cook. It uses the results for optimal designs for geometric random graphs discussed in § 4.2.4.

Imagine the following situation. There are two replicate populations of n individuals each. Individuals pass from state S to state I after a constant period of time μ. Replicate A is observed once at time τ_A and replicate B once at time τ_B. Without loss of generality, $\tau_A \leq \tau_B$. Let $I_i(t)$ be the number of individuals in replicate i in the state I at time t. Clearly, $I_i(t) = 0$ if $t < \mu$ and $I_i(t) = n$ if $t \geq \mu$.

Assume that the prior knowledge for μ is vague and expressed via the following prior distribution:

$$\pi(\mu) = \mathbb{1}_{\{\mu \in (0,1)\}}.$$

Let us restrict our attention to the designs such that $\tau_i \in [0, 1]$, since any other design would yield no more information.

The uniform prior $\pi(\mu)$ translates to the following priors for $\{I_A(\tau_A), I_B(\tau_B)\}$:

$$\mathbb{P}\left(\{I_A(\tau_A), I_B(\tau_B)\} = (0,0)\right) = 1 - \tau_B \tag{4.24}$$

$$\mathbb{P}\left(\{I_A(\tau_A), I_B(\tau_B)\} = (n,0)\right) = \tau_B - \tau_A \tag{4.25}$$

$$\mathbb{P}\left(\{I_A(\tau_A), I_B(\tau_B)\} = (n,n)\right) = \tau_A, \tag{4.26}$$

with the outcome in (4.25) having probability 0 if an identical choice of design in the two replicates is made, i.e. if $\tau_A = \tau_B$.

It follows that the posterior for μ is:

$$\pi(\mu \,|\, \{I_A(\tau_A), I_B(\tau_B)\} = (0,0)) = \frac{1}{1-\tau_B} \mathbb{1}_{\{\mu \in (\tau_B, 1)\}} \tag{4.27}$$

$$\pi(\mu \,|\, \{I_A(\tau_A), I_B(\tau_B)\} = (n,0)) = \frac{1}{\tau_B - \tau_A} \mathbb{1}_{\{\mu \in (\tau_A, \tau_B)\}} \tag{4.28}$$

$$\pi(\mu \,|\, \{I_A(\tau_A), I_B(\tau_B)\} = (n,n)) = \frac{1}{\tau_B - \tau_A} \mathbb{1}_{\{\mu \in (\tau_0, \tau_A)\}}. \tag{4.29}$$

If $\tau_A < \tau_B$, the expected utility, based on the Kullback–Leibler divergence, is

$$\mathbb{E}\left[U(\tau_A, \tau_B)\right] = (1-\tau_B) \log \frac{1}{1-\tau_B} + (\tau_B - \tau_A) \log \frac{1}{\tau_B - \tau_A} + \tau_A \log \tau_A^{-1}, \tag{4.30}$$

which is maximised (Theorem 4.2.2) by $(\tau_A, \tau_B) = (1/3, 2/3)$, with the expected information yield $U(1/3, 2/3) = \log 3$.

If, on the other hand, $\tau_A = \tau_B = \tau$, the expected utility becomes

$$\mathbb{E}\left[U(\tau, \tau)\right] = (1-\tau) \log \frac{1}{1-\tau} - \tau \log \tau, \tag{4.31}$$

which is maximised by $\tau = 1/2$, giving utility $U(1/2, 1/2) = \log 2 < U(1/3, 2/3)$.

In fact, it can readily be seen that taking the same design in both replicates yields no more information than having a single replicate with that design. It can also be seen that replicates containing a single individual yield the same information as those containing more than one individual. It seems to be intuitively obvious that if the lifetimes are random and their variance is much smaller than the variance of the prior for the mean, a similar result will hold.

A related point that this example shows is that the sequential optimal design of replicated experiments need not be the same as the optimal design of simultaneous replicated experiments. If we ran the above experiment simultaneously, the best design, as found above, is $\tau_A = 1/3$ and $\tau_B = 2/3$ with the utility equal $\log 3$. If, however, we allowed the inference which results from replicate A to be used for designing an experiment B at some later time, an argument similar to that above shows that the optimal design is to take $\tau_A = 1/2$, followed by $\tau = 1/4$ if $I_A(1/2) = n$, and $\tau_B = 3/4$ if $I_A(1/2) = 0$. This sequential design has utility $\log 4$ and thus is more informative than the simultaneous design.

Chapter 5

Lattice-based Optimal Designs

In the first section of this chapter we study inference and optimal design problems for finite clusters from percolation on the integer lattice $\mathbb{Z}^d$ or, equivalently, for *SIR* epidemics evolving on a bounded subset of $\mathbb{Z}^d$ with constant infectious times. The corresponding percolation probability p is considered to be unknown, possibly depending, through the experimental design, on other parameters. We consider inference under each of the following two scenarios:

(i) The observations consist of the set of sites which are ever infected, so that the routes by which infections travel are not observed (in terms of the bond percolation process, this corresponds to a knowledge of the connected component containing the initially infected site—the location of this site within the component not being relevant to inference for p).

(ii) All that is observed is the *size* of the set of sites which are ever infected. By the set size we mean cardinality here.

We discuss practical aspects of Bayesian utility-based optimal designs for the former scenario and prove that the sequence of maximum likelihood estimates for p converges to the critical percolation probability p_c under the latter scenario (when the size of the finite cluster grows infinitely).

In the second section we outline how the results for nearest-neighbour graph models can be generalised to the case of long-range connections.

5.1 Inference and Optimal Design for Percolation Models

5.1.1 Nearest-neighbour interaction model and percolation

Brief historical account on percolation

The concept of percolation has received enormous interest among physicists since it was introduced by Broadbent and Hammersley (1957). One reason for that, perhaps, is that it provides a clear and intuitively appealing model of the geometry which appears in disordered systems. Percolation has been used to model and analyse the spreading of oil in water and transport phenomena in porous media and materials (Yanuka (1992), Stauffer and Aharony (1992), de Gennes and Guyon (1978), Larson et al (1981), Sahimi (1994), Odagaki and Toyufuku (1998), Tobochnik (1999), De Bondt et al (1992), Bunde et al (1995), Bentz and Garboczi (1992), Machta (1991), Moon and Girvin (1995)), to model the spread of infections and forest fires via nearest and finite range percolation (Zhang (1993), Cox and Durrett (1988), Gibson et al (2006)) and via continuum percolation (Meester and Roy (1996)). It has also been used in studying failures of electronic devices and integrated circuits (Gingl et al (1996)), in modelling random resistor networks (Pennetta et al (2002)), and in studying transport and electrical properties of percolating networks (Adam and Delsanti (1989)). Percolation models have also been used outside physics to model ecological disturbances (With and Crist (1992)), robustness of the Internet and other networks (Cohen et al (2000), Callaway et al (2000)), biological evolution (Ray and Jan (1994)), and social influence (Solomon et al (2000)). Percolation is one of the simplest models which exhibits phase transition, and the occurrence of critical phenomena is central to the appeal of percolation. The reader is referred to Chapter 1 of Grimmett (1999) for further details on modelling a random medium using percolation models.

Classical *SIR* epidemic model and percolation

Disease spread as a result of (typically) short-range contact between, for example, plants can be modelled as a transmission process on an undirected graph. Nodes, or vertices, of the graph correspond to possible locations of plants, and edges of the graph link locations which are considered to be *neighbours*. In a classical *SIR* model each node, or vertex, of the graph is in one of three states: either it is occupied by a *healthy*, but susceptible, plant (state S), or it is occupied by an *infected* and infectious plant (state I), or finally it is *empty*, any plant at that location having died and thus being considered removed (state R). A plant at node i, once infected (or from time 0 if initially infected), remains in the infected (and infectious) state I for some random time τ_i after which it dies, so that node i then remains in the empty state R ever thereafter. During its infectious time the plant at node i sends further infections to each of its neighbouring nodes j as a Poisson process with rate λ_{ij} (so that the probability that an infection travels from i to j in any small time interval of length h is $\lambda_{ij}h + o(h)$ as $h \to 0$ while the probability that two or more infections travel in the same interval is $o(h)$ as $h \to 0$); any infection arriving at node j changes the state of any *healthy* plant there to *infected*, and otherwise has no effect. All infectious periods and infection processes are considered to be independent of each other. The initial state of the system is typically defined by one or more nodes being occupied by infected plants, the remaining nodes being occupied by healthy plants. The epidemic may *die out* at some finite time at which the set of infected nodes first becomes empty, or, on an infinite graph only, it is possible that it may continue forever.

Thus, for any infected node i, the event E_{ij} that any neighbouring node j receives at least one infection from node i has probability $p_{ij} = 1 - \mathbb{E}[\exp(-\lambda_{ij}\tau_i)]$ (here, as previously, $\mathbb{E}$ denotes expectation). Note that, for any given node i, even though the infection processes are independent, the events E_{ij} are themselves independent if and only if the random infectious period τ_i is a constant. We now suppose that this is the case and that furthermore, for all ordered pairs (i, j) of neighbours, we have $p_{ij} = p$ for some probability p. Suppose further that it is possible to observe neither the time evolution of the epidemic nor the edges of

the graph by which infections travel, but only the initially infected set of nodes and the set of nodes which are at some time infected and thus ultimately in the empty state R. It is then not difficult to see, and is indeed well known (e.g. Kulasmaa and Zachary (1984)), that the epidemic may be probabilistically realised as an unoriented *bond percolation* process on the graph in which each edge is independently *open* with probability p, and in which the set of nodes which are ever infected consists of those nodes reachable along open *paths* (chains of open edges) from those initially infected. (Note that the ability to use an *unoriented* bond percolation process requires both the assumptions that the above events E_{ij} are independent and that $p_{ij} = p_{ji}$ for all i, j; in the absence of *either* of these assumptions one would in general need to consider an oriented process with the appropriate dependence structure [1].)

Further we consider the epidemic to take place on some subset Π of the two-dimensional integer lattice $\mathbb{Z}^d$, where we allow $\Pi = \mathbb{Z}^d$ as a possibility. Two sites (nodes) are considered *neighbours* if and only if they are distance 1 apart. Thus in the case $\Pi = \mathbb{Z}^2$ each node has 4 neighbours. This may be considered as a model for *nearest-neighbour* interaction. We assume furthermore that initially there is a single infected site, and that all other sites in Π are occupied by healthy individuals.

Bond percolation in graph-theoretic terms

We now establish the basic definitions and notation for bond and site percolation on the integer lattice. As usual, we write $\mathbb{Z}^d$ for the set of all vectors $x = (x_1, x_2, \ldots, x_d)$ with integer coordinates. The norm $\| \cdot \|_1$ defines a distance between each two elements of $\mathbb{Z}^d$, x and y, as follows:

$$\delta(x, y) := \| x - y \|_1 = \sum_{i=1}^{d} |x_i - y_i|.$$

[1] Non-constant infectious period distributions τ_i can similarly lead to other interesting percolation processes. For example, *site percolation* may be approximated arbitrarily closely by an infectious period distribution which with some sufficiently small probability takes some sufficiently large value, and which otherwise takes the value zero (see Appendix E).

The set $\mathbb{Z}^d$ may be turned into a graph using the '4-neighbourhood relationship' as follows: two elements x and y are declared to be neighbours (or *adjacent*) if and only if $\delta(x, y) = 1$. If x and y are adjacent, then we write $x \sim y$. The set of edges obtained in this way is denoted by $\mathbb{E}^d$ and the corresponding graph $(\mathbb{Z}^d, \mathbb{E}^d)$ is called the *d-dimensional cubic lattice* (Grimmett (1999)). We denote this lattice by $\mathbb{L}^d$ and the origin of $\mathbb{Z}^d$ by 0.

The following describes the percolation process on $\mathbb{L}^d$. Let p be a real number between zero and one: $0 \leq p \leq 1$. We declare each edge of the lattice $\mathbb{L}^d$ to be *open* with probability p and closed otherwise, independently of the status of any other edge. The random subgraph of $\mathbb{L}^d$ formed in this way contains the vertex set $\mathbb{Z}^d$ and the open edges only. The connected components of this graph are called *open clusters*. The open cluster containing the vertex x is denoted by $\mathcal{C}(x)$. It is clear that the distribution of $\mathcal{C}(x)$ is independent of the choice of x. The open cluster $\mathcal{C} := \mathcal{C}(0)$ containing the origin is typical in this sense. Figure 5.1 depicts examples of open clusters of a percolation process on $\mathbb{L}^2$ (for different values of p) restricted to the bounding box $[-31, 31] \times [-31, 31]$.

A central quantity of interest in percolation theory is that of *percolation probability* $\theta(p)$, this being the probability that the origin (or any other given vertex) belongs to an infinite open cluster:

$$\theta(p) := \mathbf{P}_p(|\mathcal{C} = \infty|) = 1 - \sum_{n=1}^{\infty} \mathbb{P}_p(|\mathcal{C}| = n).$$

The following critical phenomenon results are of fundamental importance in percolation theory (Grimmett (1999)):

- The function θ is a non-decreasing function of p:

$$0 = \theta(0) \leq \theta(p') \leq \theta(p''), \ \forall p', p'' : \ 0 \leq p' \leq p'' \leq 1.$$

- There exists a critical value $p_c(d)$ of p such that $\theta(p) = 0$ for any $p < p_c(d)$ and $\theta(p) > 0$ for any $p > p_c(d)$. The value $p_c(d)$ is called the *critical probability* and can formally be defined as follows:

$$p_c(d) := \sup\{p : \ \theta(p) = 0\}.$$

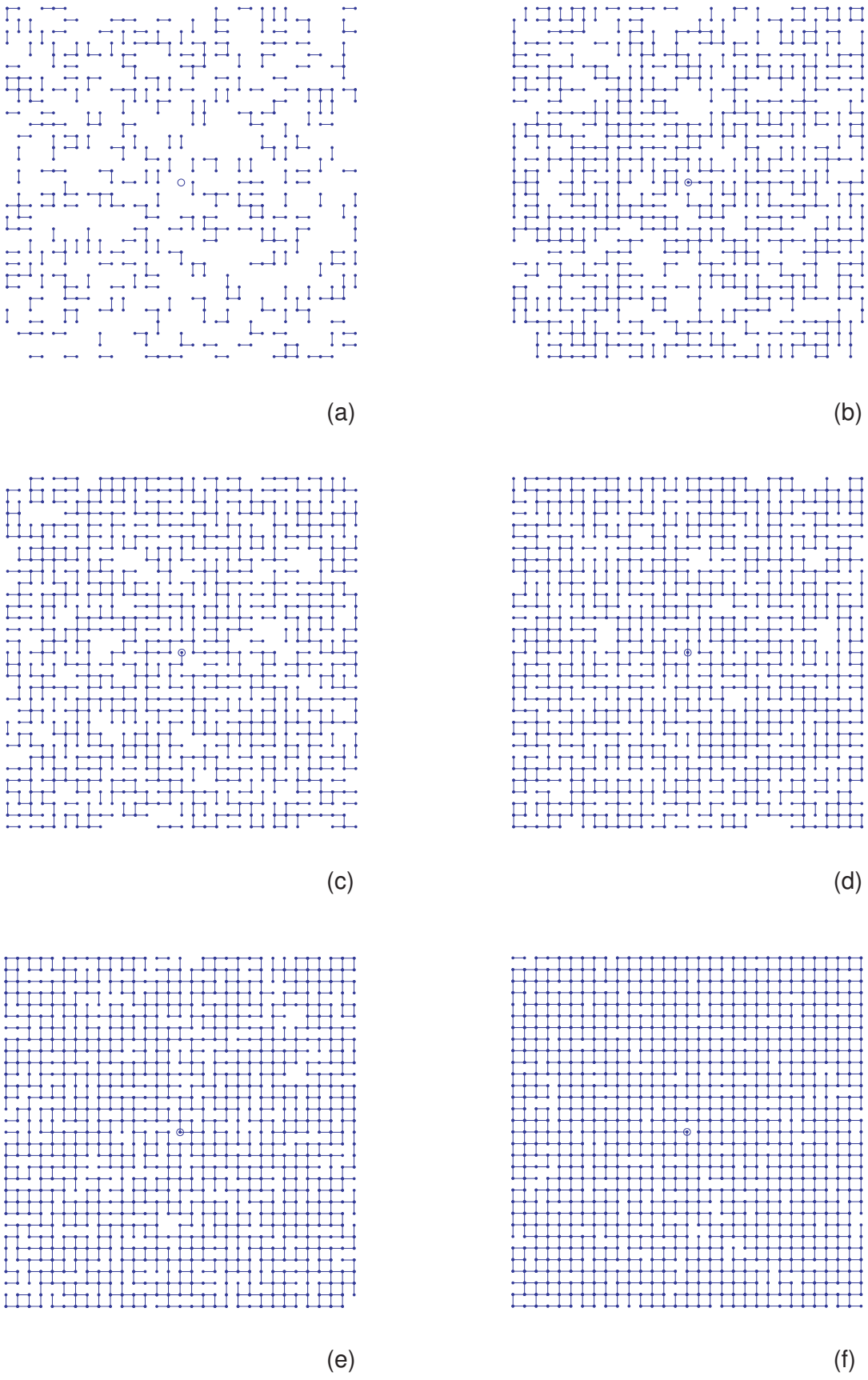


Figure 5.1: Open clusters emerged as a result of bond percolation on $\mathbb{L}^2$ for different values of p: (a) $p = 0.2$, (b) $p = 0.4$, (c) $p = 0.5$, (d) $p = 0.6$, (e) $p = 0.75$, and (f) $p = 0.9$. The origin of $\mathbb{Z}^2$ is denoted by a circle in the centre of each plot.

- The critical probability is unity in the one-dimensional case: $p_c(1) = 1$.
- The critical probability exists and is strictly between zero and one on the lattice $\mathbb{L}^d$, $d \geq 2$:

$$0 < p_c(d) < 1, \quad \text{for any } d \geq 2.$$

- The critical probability is a monotonically decreasing function in d:

$$p_c(d+1) < p_c(d), \quad \text{for } d \geq 1.$$

Incomplete observations

The probability p introduced above is considered to be unknown, but may depend on other parameters. For instance, this probability may depend on the distance between plants (lattice vertices) or, if $\Pi = \mathbb{Z}^2$ and the Poisson process of emitting germs by infectious plants is isotropic, it may be related to its intensity $\lambda = 4\lambda_{ij}$ (each site has four neighbours in $\mathbb{L}^2$). In the latter case p may be taken to be of the form $p = 1 - e^{-\lambda/4}$ and it is λ that would be an object of interest for plant epidemiologists.

We consider inference under each of the following two scenarios:

(i) the observations consist of the set of sites which are ever infected, so that the routes by which infections travel are not observed; note that, in terms of the bond percolation process, this corresponds to knowledge of the connected component containing the initially infected site—the location of this site within the component not being relevant to inference for p (see below);

(ii) all that is observed is the *size* of the set of sites which are ever infected.

We denote and refer further to the former of these two scenarios as $\mathcal{S}1$ and to the latter scenario as $\mathcal{S}2$.

5.1.2 Parameter estimation

Distribution of ever-infected sites

Consider our *SIR* constant infectious period epidemic on a locally finite graph in which the probability that any individual i sends at least one infection to any given

neighbour j is p. By the definition of the epidemic these events are independent.

The following basic result is well-known. However, the author was unable to find a reference to the formulated and rigorously proven result—this theorem can be well regarded as a part of mathematical folklore of the sort "It is easy to see that..." (e.g., see Grassberger (1983)).

Theorem 5.1.1. *For any given set of initially infected sites, the distribution of the set of ever-infected sites is the same as for the corresponding unoriented bond percolation process (with the same initial set).*

Proof. Given the realisation of the epidemic we construct a realisation of the unoriented bond percolation process as follows. For each unordered pair of neighbours $\{i, j\}$, if i, say, becomes infected before j then we construct an open link between i and j if and only if, in the epidemic, i sends at least one infection to j; if either i and j are both initially infected or i and j are both never infected, then we construct an open link between i and j with probability p independent of all else. Since the probability for two vertices to become infected at exactly the same time is 0, it is clear from consideration of the temporal evolution of the epidemic that all edges are open with probability p independently of each other. Furthermore, the set of ever-infected sites in the epidemic is the same as the set of 'wetted' sites (sites linked by open edges to the initial wet set) in the bond percolation process. □

It follows that, for inference, if all that is observed is the set of ever-infected sites, then we may calculate the likelihood function using the unoriented bond percolation model. However, one cannot think of any scenario in which we also obtain any information about the links used to spread the epidemic for which a similar conclusion holds. Here are two possible scenarios with counter-examples.

- For at least some *unordered* pairs $\{i, j\}$ of neighbours, we observe whether or not an infection passed between i and j (even if both were already infected). Consider the graph with 2 vertices and one edge, and suppose we observe the edge to have been used; then the likelihood for the epidemic model is $2p - p^2$, while that for the unoriented bond percolation model is p.

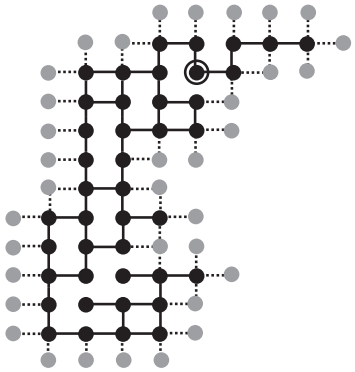


Figure 5.2: An open cluster (black solid dots) containing the origin (a black dot in a circle) as a result of percolation simulation on $\mathbb{L}^2$. Here the bond percolation probability p was taken to be 0.478; the solid bonds represent open bonds. The open cluster can be seen as a finite outbreak of an epidemic with constant infectious periods and infection intensity spread rate $\lambda \approx 2.6$ evolving on $\Pi = \mathbb{Z}^2$ (since $0.478 = 1 - e^{-2.6/4}$). The dotted lines depict directions along which infection did not spread (from black to grey dots); thus, grey dots depict individuals which remain healthy and the dotted lines represent those bonds that must be absent given knowledge of the cluster set.

- For at least some *ordered* pairs (i, j) of neighbours, we observe whether or not an infection passed from i to j (even if j was already infected). Consider the graph with 3 vertices and 3 edges, and suppose (with vertex 1 initially infected) we observe infections to have passed from 1 to 2 and from 1 to 3 and also that no other infections have passed; then the likelihood for the epidemic model is $p^2(1-p)^4$, while that for the unoriented bond percolation model is $p^2(1-p)$.

In the first of the above scenarios, if we made the observation for every unordered pair of neighbours, then, for inference, we could pass to the unoriented bond percolation model with parameter $p' = 2p - p^2$.

The result proved in Theorem 5.1.1 means that a final snapshot of an *SIR* epidemic with nearest-neighbour interaction and constant infectious periods evolving on $\mathbb{Z}^2$ can be seen as an open cluster of the corresponding percolation process on $\mathbb{L}^2 = (\mathbb{Z}^2, \mathbb{E}^2)$, had the infection process started with a single initially inoculated

site (placed at the origin of the lattice, for example). Figure 5.2 shows an open cluster obtained by simulation of percolation process on the integer lattice in plane when $p = 0.478$. This connected component containing the origin can be seen as a final (and finite) outbreak of an *SIR* epidemic process of the kind discussed above. The origin (or, indeed, any other vertex of the open cluster) may be considered to be the site where the initially inoculated individual has been placed. Clearly, the realised bond structure is not the only possible way resulting in the site configuration seen in Figure 5.2. However, the distribution of this site configuration as an extinct *SIR* epidemic coincides with that of the corresponding unoriented bond percolation process.

Scenario $\mathcal{S}_1$: hidden bond structure

Let Π be a (proper or improper) subgraph of $\mathbb{L}^d = (\mathbb{Z}^d, \mathbb{E}^d)$ containing the origin and let $\mathcal{C}$ be an open cluster of a percolation process on the graph Π containing the origin. The set of nodes $\mathcal{C}$ represents a snapshot of an extinct outbreak of our spatial *SIR* epidemic evolving on $\Pi \subseteq \mathbb{L}^2$.

Let us introduce some additional notions. Let $G = (V, E)$ be a locally finite graph and let $G' = (V', E')$ be a subgraph of G. By the *saturation* of the graph G' with respect to G we understand the graph $\tilde{G} = (\tilde{V}, \tilde{E})$ such that

$$\tilde{V} = V' \text{ and } \tilde{E} = \{(x, y) \,|\, x, y \in V' \,\&\, (x, y) \in E\}.$$

Thus, in order to obtain the saturation of a subgraph G' of a given graph G one needs to add to G' all possible edges from G with endpoints from G', and hence 'saturate' it.

We denote the saturation of G' with respect to G by $\mathrm{Satur}_G\, G'$ or, in cases when it is clear from the context with respect to what graph the saturation takes place, by $\mathrm{Satur}\, G'$. A graph G' whose saturation (with respect to some graph G) coincides with itself is called a *fully saturated graph*. For example, the fully saturated graph (with respect to $\mathbb{L}^2$) is obtained from the graph depicted in Figure 5.2 by connecting all pairs of neighbouring black sites (according to the 4-neighbourhood relationship). Note that the operation of saturation may also be applied solely to a subset of vertices of the original graph, since it does not make use of the edges

of the subgraph-operand (alternatively, one may think about the subset of the original graph vertex set as a subgraph with an empty edge set).

In order to distinguish between the boundary points of a graph and their neighbours, which are not in the graph, we introduce the notions of the *surface* and the *frontier* of the graph (again, with respect to another graph). Let us denote by ∂G the surface of G in Π, $G \subseteq \Pi$, that is to say the set

$$\partial G := \{x \in G : \exists y \in \Pi \backslash G \text{ such that } x \text{ and } y \text{ are neighbours in } \Pi\},$$

and by Γ_G the frontier of G in Π, i.e. the set $\partial(\Pi \backslash G)$.

In order to identify the likelihood function we introduce the set $\mathcal{G}(\mathcal{C})$ of all *connected* subgraphs of Π with $\mathcal{C}$ as a vertex set. Note that the set $\mathcal{G}(\mathcal{C})$ is necessarily nonempty. For each $G \in \mathcal{G}(\mathcal{C})$ the number of edges between the vertices of the graph G and the elements of its frontier Γ_G is the same—we denote it by $w_{\mathcal{C}}$. Finally, we denote the total number of edges present in G by $e(G)$.

The probability that $\mathcal{C}$ represents the set of ever-infected sites and that the edges of G correspond to those routes along which the infection travelled is

$$\mathbb{P}_p(G) = p^{e(G)}(1-p)^{e(\operatorname{Satur}\mathcal{C})-e(G)+w_{\mathcal{C}}},$$

and the likelihood function associated with the observed set $\mathcal{C}$ of ever-infected sites is given by

$$\mathcal{L}(p) = \mathbb{P}_p(\mathcal{C}) = \sum_{G \in \mathcal{G}(\mathcal{C})} \mathbb{P}_p(G).$$

Hence, under assumption of a uniform prior for p, its posterior distribution $\pi(p \,|\, \mathcal{C})$ is a mixture of beta distributions:

$$\pi(p \,|\, \mathcal{C}) \propto \sum_k r(k) \operatorname{Beta}\left(k+1, e(\operatorname{Satur}\mathcal{C}) - k + w_{\mathcal{C}} + 1\right),$$

where

$$r(k) := \#\{G \in \mathcal{G}(\mathcal{C}) \,|\, e(G) = k\}.$$

It is not feasible to calculate $\pi(p \,|\, \mathcal{C})$ in the above form for reasons of difficulty in calculating efficiently the coefficients $r(k)$, since it is hard to enumerate all corresponding graphs. We describe therefore an MCMC algorithm that allows

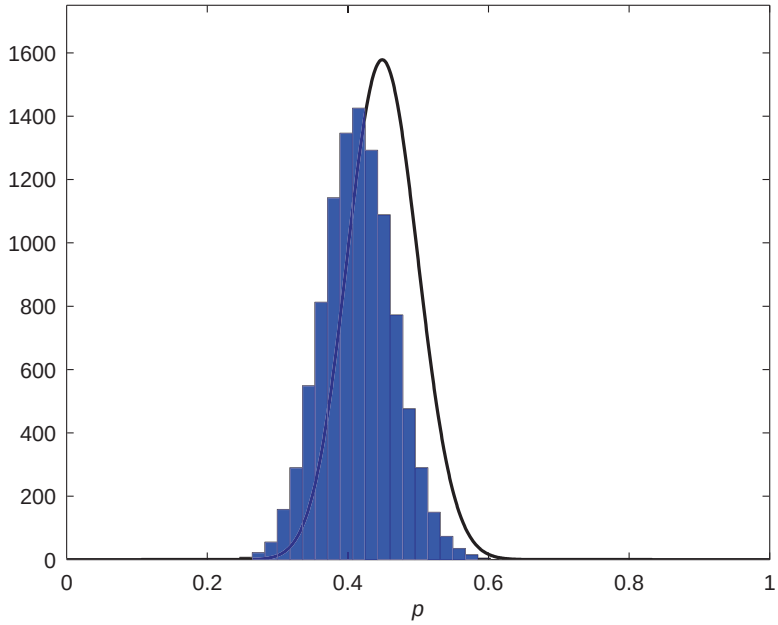


Figure 5.3: Solid line corresponds to the likelihood function evaluated for the complete information (both the site and edge configurations are known) on the cluster $\mathcal{C}$ from Figure 5.2. The histogram is based on a sample drawn from the MCMC applied to the site configuration $\mathcal{C}$ (nodes only).

one to sample from the distribution $\pi(p \,|\, \mathcal{C})$ under the uniform prior on p, that is, effectively, to evaluate the likelihood function of p.

Our Markov chain explores the joint space of values for p and graphs from $\mathcal{G}(\mathcal{C})$, that is to say the set $[0,1] \times \mathcal{G}(\mathcal{C})$. The stationary distribution of the chain is the joint posterior distribution of p and $G \in \mathcal{G}(\mathcal{C})$. The description of the chain is given in Algorithm 1. This Markov chain explores the set of all connected graphs $\mathcal{G}(\mathcal{C})$ by simply deleting or adding an edge from the current graph preserving the connectivity of the given site configuration $\mathcal{C}$.

The proposed MCMC is irreducible by construction: there is a positive probability for the chain to switch between any two connected graphs from $\mathcal{G}(\mathcal{C})$ since any two such graphs have the same vertex set and differ by a finite number of edges only.

Example 5.1.2. *We apply Algorithm 1 to the site configuration $\mathcal{C}$ from Figure 5.2 (black dots only). This open cluster at the origin was obtained by simulating the percolation process in $\mathbb{Z}^2$ using the value of the percolation parameter $p = 0.478$.*

Algorithm 1 Markov Chain Monte Carlo: scenario $\mathcal{S}1$

Require: an open cluster $\mathcal{C}$;

take an initial value p_0 arbitrary from $(0,1)$;
$t := 0$ $X_t := (p_t, \operatorname{Satur}\mathcal{C})$;
repeat
  Gibbs sampler steps:
  $p_{2t+1} \sim \operatorname{Beta}\left(e(G_{2t}) + 1, e(\operatorname{Satur}\mathcal{C}) - e(G_{2t}) + w_{\mathcal{C}} + 1\right)$;
  $X_{2t+1} := (p_{2t+1}, G_{2t})$;
  Metropolis–Hastings sampler steps:
  choose an edge e uniformly at random from $\operatorname{Satur}\mathcal{C}$;
  if $e \in G_{2t}$ **then**
    $U \sim \operatorname{Uniform}[0,1]$;
    if $U \leq \min(1, \frac{1-p_{2t+1}}{p_{2t+1}} \mathbb{1}_{\{G_{2t+1}\setminus\{e\} \text{ is connected}\}})$ **then**
      $X_{2t+2} := (p_{2t+1}, G_{2t+1} \setminus \{e\})$;
    else
      $X_{2t+2} := (p_{2t+1}, G_{2t+1})$;
  else
    $U \sim \operatorname{Uniform}[0,1]$;
    if $U \leq \min(1, \frac{p_{2t+1}}{1-p_{2t+1}})$ **then**
      $X_{2t+2} := (p_{2t+1}, G_{2t+1} \cup \{e\})$;
    else
      $X_{2t+2} := (p_{2t+1}, G_{2t+1})$;
  $t := t + 2$;
until we judge that the chain has converged and a sample of sufficient size is recorded

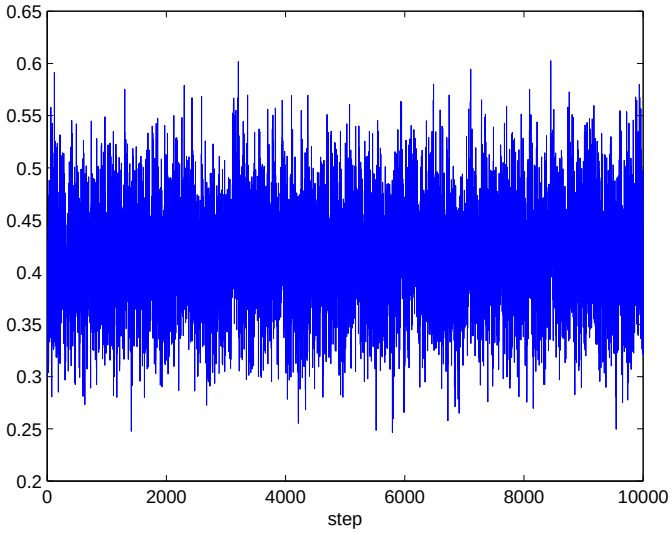


Figure 5.4: Trace plot for MCMC sampling resulted in the histogram from Figure 5.3 for the cluster $\mathcal{C}$ from Figure 5.2. The trace plot indicates that the mixing properties of the chain are rather satisfactory. It took 28 seconds on Intel(R) Core(TM)2 Duo CPU 2.26GHz to obtain a series of chain updates of the length 10^4. This time could be further reduced by using dynamic graph update algorithms, see the footnote on the p. 129.

Figure 5.3 shows the likelihood function of the model parameter for the complete observation (i.e. nodes and edges of the cluster) and a histogram of a sample from the posterior distribution $\pi(p \,|\, \mathcal{C})$ *obtained by running the MCMC described in Algorithm 1 when the prior distribution is uniform on the interval* $(0, 1)$.

An animated example of using Algorithm 1 can be found at the WEB address `http://www.cl.cam.ac.uk/~aib29/HWThesis/Video/`. We defer the discussion on the mixing properties of the suggested chain until § 5.1.4.

Scenario $\mathcal{S}_2$: unknown site configuration

Under this scenario only the size n of the outbreak of our *SIR* epidemic evolving on $\Pi = \mathbb{L}^d$ is given.

Let $\mathcal{G}_n$ be the set of all possible connected graphs on n vertices including the origin. These graphs represent the outbreaks of the size n and we distinguish all isomorphic graphs which have different orientation.

We denote the number of edges of Π between the vertices of the graph $G \in \mathcal{G}_n$ and the vertices of its frontier Γ_G by $w(G)$.

Given the epidemic of size n, the inference on p involves evaluation of the likelihood function $\mathcal{L}_n(p) := \mathbb{P}_p(|\mathcal{C}| = n)$ which can be represented as follows:

$$\mathcal{L}_n(p) = \sum_{G \in \mathcal{G}_n} \mathbb{P}_p(G).$$

As previously, under assumption of a uniform prior for p its posterior distribution $\pi(p \,|\, |\mathcal{C}| = n)$ is a mixture of beta distributions:

$$\pi(p \,|\, |\mathcal{C}| = n) \propto \sum_{s,k,l} q(s,k,l)\, \mathrm{Beta}\,(k+1, s-k+l+1)\,, \tag{5.1}$$

where

$$q(s,k,l) := \#\{G \in \mathcal{G}_n \,|\, e(\mathrm{Satur}\, G) = s, e(G) = k, w(G) = l\}.$$

This, again, represents a hard enumeration problem. However, inference on p can be made using the MCMC technique. Algorithm 2 contains a description of a Markov chain which serves the purpose of sampling from the posterior distribution $\pi(p \,|\, |\mathcal{C}| = n)$, given the prior distribution of p is uniform on [0,1]. The chain explores the joint space of all possible connected graphs on n nodes and possible values for the percolation parameter p by deleting a vertex from the current graph and adding a vertex from its frontier.

The presented Markov Chain Monte Carlo algorithm, similarly to the one suggested for the previously considered scenario $\mathcal{S}1$, is a combination of Gibbs and Metropolis–Hastings steps. The marginal of the chain limiting distribution $f(p, G)$ in p coincides with the posterior distribution $\pi(p \,|\, |\mathcal{C}| = n)$.

We give now explicit expressions for the proposal probabilities used in the Metropolis–Hastings part of this algorithm. Assume that the current graph within the Metropolis–Hastings step is G and a graph $\tilde{G}$ is proposed, the latter being obtained from the former by deleting a vertex u with all edges adjoining it and inserting a vertex v with every possible edge, each independently with probability p, which was determined by the preceding Gibbs step. We assume that at least one such edge is inserted and denote the number of deleted and added edges by

Algorithm 2 Markov Chain Monte Carlo: scenario $\mathcal{S}2$

Require: the value of n.

take a value p_0 arbitrary from $(0,1)$ and a graph G_0 arbitrary from $\mathcal{G}_n$;
$t := 0$ $X_t := (p_t, G_t)$;
repeat
  `move_on`:=1;
  Gibbs sampler steps:
  $p_{2t+1} \sim \mathrm{Beta}\,(e(G_{2t}) + 1, e(\mathrm{Satur}\, G_{2t} - e(G_{2t}) + w(G_{2t}) + 1))$;
  $X_{2t+1} := (p_2 t + 1, G_{2t})$;
  Metropolis–Hastings sampler steps:
  choose a vertex u uniformly at random from G_{2t+1} and choose a vertex v uniformly at random from $\Gamma_{G_{2t+1}}$. Derive a graph $\tilde{G}$ from G_{2t+1} by deleting all edges which adjoin u (in G_{2t+1}) and adding the edges that connect v with vertices of the graph $G_{2t+1} \setminus \{u\}$ in Π, each independently with probability p_{2t+1} (conditioning on the event that at least one edge is added).
  if $\tilde{G}$ is disconnected **then**
    $X_{2t+2} := (p_{2t+1}, G_{2t+1})$; `move_on`:=0
  if `move_on` **then**
    $U \sim \mathrm{Uniform}(0,1)$;
    $\tilde{d}(v) := \#\{e \,|\, e = (v,z)\, \exists z \in G_{2t+1} \setminus \{u\}\}$;
    $\tilde{d}(u) := \#\{e \,|\, e = (u,z)\, \exists z \in \tilde{G} \setminus \{v\}\}$;
    $\nu(u) := \#\{x \,|\, x \in \Gamma_{G_{2t+1}} \,\&\, (u,x) \in \Pi\}$;
    $\nu(v) := \#\{x \,|\, x \in \Gamma_{\tilde{G}} \,\&\, (v,x) \in \Pi\}$;
    $\kappa := \tilde{d}(u) - \tilde{d}(v) + \nu(v) - \nu(u)$; $U \sim \mathrm{Uniform}(0,1)$;
    if $U \le \min\left(1, \frac{|\Gamma_{G_{2t+1}}|}{|\Gamma_{\tilde{G}}|} \frac{1-(1-p_{2t+1})^{\tilde{d}(v)}}{1-(1-p_{2t+1})^{\tilde{d}(u)}} (1-p_{2t+1})^{\kappa}\right)$ **then**
      $X_{2t+2} := (p_{2t+1}, \tilde{G})$;
    else
      $X_{2t+2} := (p_{2t+1}, G_{2t+1})$;
  $t := t + 2$;
until we judge that the chain has converged and a sample of sufficient size is recorded

$d(u)$ and $d(v)$ respectively. Then,

$$q(G,\tilde{G}) = \frac{1}{n}\frac{1}{|\Gamma_G|}\frac{p^{d(v)}(1-p)^{\tilde{d}(v)-d(v)}}{1-(1-p)^{\tilde{d}(v)}}\mathbb{1}_{\{\tilde{G} \text{ is connected}\}}, \tag{5.2}$$

and similarly,

$$q(\tilde{G},G) = \frac{1}{n}\frac{1}{|\Gamma_{\tilde{G}}|}\frac{p^{d(u)}(1-p)^{\tilde{d}(u)-d(u)}}{1-(1-p)^{\tilde{d}(u)}}\mathbb{1}_{\{G \text{ is connected}\}}. \tag{5.3}$$

Clearly,

$$\mathbb{P}_p(G) \propto p^{d(u)}(1-p)^{\tilde{d}(v)}(1-p)^{\tilde{d}(u)-d(u)+\nu(u)}$$

$$\mathbb{P}_p(\tilde{G}) \propto p^{d(v)}(1-p)^{\tilde{d}(v)}(1-p)^{\tilde{d}(v)-d(v)+\nu(v)},$$

so that the acceptance probability at the Metropolis–Hastings step, α, is as follows:

$$\begin{aligned}
\alpha &= \min\left(1, \frac{q(\tilde{G},G)}{q(G,\tilde{G})}\frac{\mathbb{P}_p(\tilde{G})}{\mathbb{P}_p(G)}\right) \\
&= \min\left(1, \frac{|\Gamma_G|}{|\Gamma_{\tilde{G}}|}\frac{(1-p)^{\tilde{d}(u)+\nu(v)}}{(1-p)^{\tilde{d}(v)+\nu(u)}}\frac{1-(1-p)^{\tilde{d}(v)}}{1-(1-p)^{\tilde{d}(u)}}\right) \\
&= \min,\left(1, \frac{|\Gamma_G|}{|\Gamma_{\tilde{G}}|}\frac{1-(1-p)^{\tilde{d}(v)}}{1-(1-p)^{\tilde{d}(u)}}(1-p)^{\kappa}\right),
\end{aligned}$$

where, as it was introduced in the description of Algorithm 2,

$$\kappa := \tilde{d}(u) - \tilde{d}(v) + \nu(v) - \nu(u).$$

We claim that the constructed chain is irreducible, that is to say this chain can get from each state to any other state. In graph theory notions this means that the chain can get from each connected graph on n vertices including the origin to any other connected graph on n vertices (also including the origin) on the considered lattice. We show the irreducibility of the proposed MCMC by constructing a sequence of steps in which any graph of $\mathcal{G}_n$ is transformed to a so called *line-skeleton* graph on n vertices. By such a graph we mean any tree (a graph with no cycles) containing the origin of the lattice and having n vertices so that only two of these vertices have degree one. Choose and fix one of such line-skeletons denoting it by S.

Consider a graph G from $\mathcal{G}_n$ and denote the length of the shortest path from $x \in G$ to S by $\delta(x, S)$ (δ here is the distance introduced in § 5.1.1). Each vertex x from G receives a well defined finite weight $\delta(x, S)$, since the graph G is connected and finite. By using the description of our Markov chain we can delete any vertex from our current graph for which $\delta(x, S)$ is maximal and add to this graph a vertex from the chosen line skeleton S without making the graph disconnected or containing cycles until the maximum value of $d(x, S)$ is zero—in this case all vertices are forming the line skeleton. Since this procedure can be reversed it follows that $\mathcal{G}_n$ in the described Markov chain is indeed a communicating class, and hence the chain is irreducible.

An animated example of using Algorithm 2 when $n = 25$ can be found at the WEB address `http://www.cl.cam.ac.uk/~aib29/HWThesis/Video/`. Figure 5.5 shows histograms of samples from the distribution $\pi(p \,|\, |\mathcal{C}| = n)$ obtained using the proposed MCMC for the scenario $\mathcal{S}2$ for the cluster size values $p = 10, 35, 50, 70$ and corresponding trace plots.

In realisations of either of the described algorithms (Algorithm 1 and Algorithm 2) we used the MATLAB library *MatlabBGL* (`http://www.stanford.edu/~dgleich/programs/matlab_bgl/`) for checking connectivity of the proposal graphs.

Convergence of inferences for $\mathcal{S}2$ with increasing cluster size

Percolation exhibits a phenomenon of *criticality*, this being central in the percolation theory: as p increases, the sizes of open clusters (connected components) also increase, and there is a critical value of p_c at which there appears a cluster which dominates the rest of the pattern. Loosely speaking, as more and more edges are assigned to be open, there comes a moment when large-scale connections are formed across the lattice. If $p < p_c$, then with probability one all open clusters are finite, but there is a single infinite open cluster when $p > p_c$ almost surely. The bond percolation on the square lattice seems to be most studied to date of all percolation processes. The critical probability p_c in the case of a square lattice is $\frac{1}{2}$. What follows, however, holds for any lattice $\mathbb{Z}^d$, $d \geq 2$.

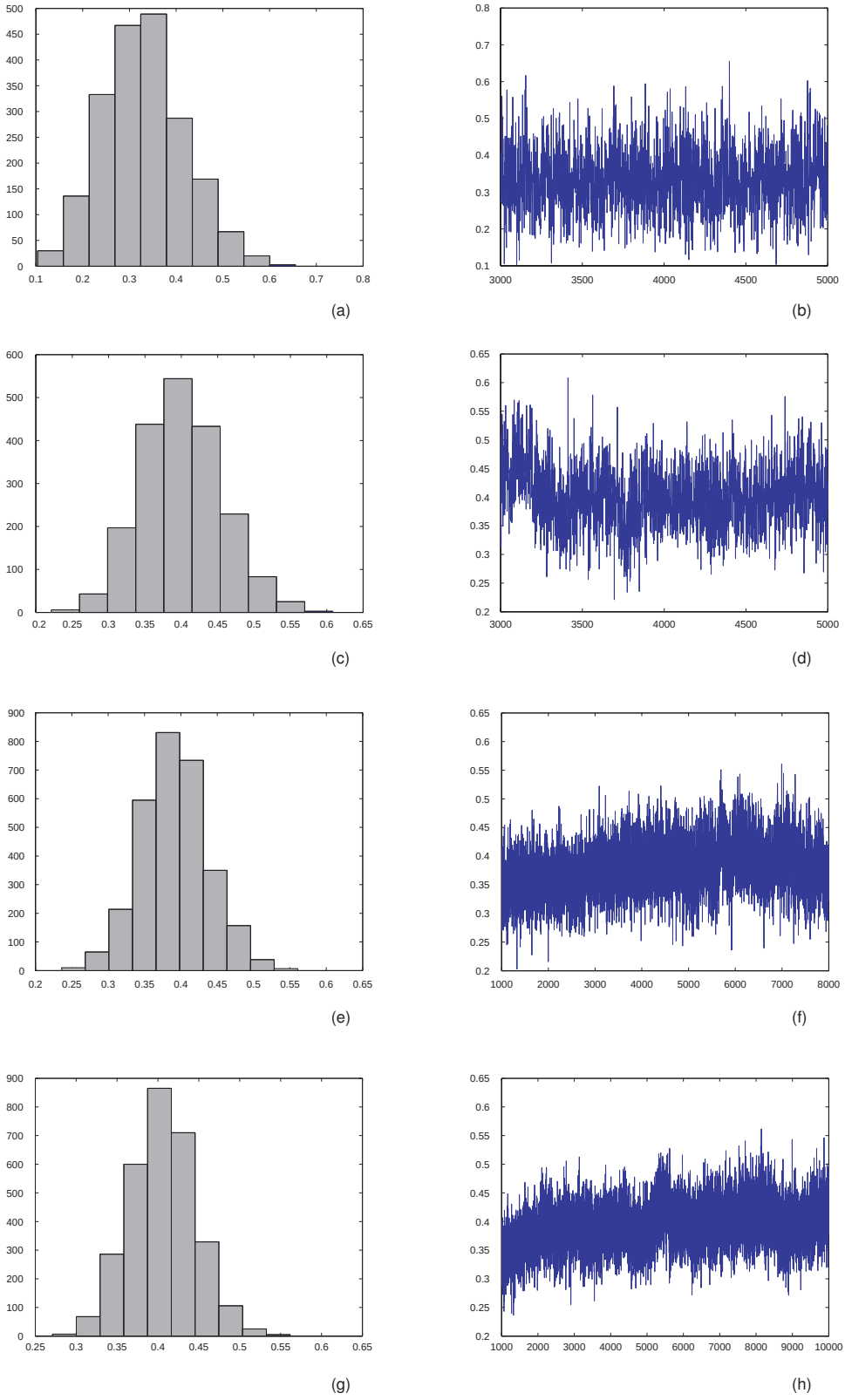


Figure 5.5: Inference on the percolation parameter using MCMC described in Algorithm 2: histograms of obtained samples and trace plots for (a,b) $n = 10$; (c,d) $n = 35$; (e,f) $n = 50$; (g,h) $n = 70$.

As before, denote by $\mathcal{C}(x)$ the open cluster (connected component) which contains the vertex x. Let us write $\chi(p) = \mathbb{E}_p|\mathcal{C}|$ for the mean number of vertices in the open cluster $\mathcal{C} := \mathcal{C}(0)$ at the origin. Using the translation invariance of the process on $\mathbb{Z}^d$, we have $\chi(p) = \mathbb{E}_p|\mathcal{C}(x)|$ for all vertices x. The percolation theory tells us that if $p < p_c$, then $\chi(p) < \infty$ (Grimmett (1999, p. 20)). When $p > p_c$, then $\chi(p) = \infty$ and the function χ is not of a much interest in this case. Instead, one studies the function $\chi^f(p) = \mathbb{E}_p[|\mathcal{C}| : |\mathcal{C}| < \infty]$. The function $\chi(p)$ ($\chi^f(p)$) monotonically increases as $p \uparrow p_c$ ($p \downarrow p_c$), having $p = p_c$ as its asymptote. It is known that there is no infinite open cluster when $p = p_c$ for percolation on the square lattice. How likely it is to observe an open cluster of size n when n is very large? What value of p should one suggest if one happened to observe a large epidemic of size n?

It is intuitively unlikely that if p is much smaller than p_c ($p \ll p_c$) or much larger than this value ($p \gg p_c$) that, having attained a sufficiently large size n, the epidemic would have burned out. Intuition suggests therefore that the likelihood function for p, given that the size n of the connected component containing the origin is increasing, should be increasingly concentrated around p_c.

Let $\mathbb{P}_p(|\mathcal{C}| = n)$ be the probability that an open cluster is of size n in percolation with the edge density p. Assume that we observe a spread of infection on $\mathbb{Z}^d$ through nearest-neighbour interactions and assume that its final size is n. The likelihood function $\mathcal{L}_n(p)$ for the percolation probability p is nothing but the probability $\mathbb{P}_p(|\mathcal{C}| = n)$ considered as a function of p:

$$\mathcal{L}_n(p) = \mathbb{P}_p(|\mathcal{C}| = n).$$

Let $\hat{p}_n$ be the maximum likelihood estimate for p derived from $\mathcal{L}_n(p)$. Then the following theorem holds.

Theorem 5.1.3. *The sequence of maximum likelihood estimates $\hat{p}_n$ for p converges to the critical probability p_c.*

Note that in the formulation of the theorem p_c stands for $p_c(d)$. The proof of the theorem is based on the following lemma.

Lemma 5.1.4. *For any $p \in (0,1)$ different from p_c the following holds:*

$$L = \lim_{n\to\infty} \frac{\mathcal{L}_n(p)}{\mathcal{L}_n(p_c)} = 0.$$

Moreover, this convergence is uniform for any closed interval which does not contain p_c.

Proof. (Lemma 5.1.4) Since the mean number $\chi(p)$ of vertices in the open cluster at the origin is infinite when $p = p_c$, the cluster size distribution $P_{p_c}(|\mathcal{C}| = n)$ cannot decay faster than any sub-exponential function. (It is strongly believed that $P_{p_c}(|\mathcal{C}| = n) \approx n^{-1-1/\delta}$, where "$\approx$" is a logarithmic equivalence, that is $\lim\limits_{n\to\infty} -\frac{\log P_{pc}}{(1+1/\delta)\log n} = 1$, but no rigorous proof of this is known, see Chapter 9 in Grimmett (1999)).

We shall further distinguish two cases:

1 *Subcritical case $p < p_c$.* In this case the cluster size distribution decays exponentially (Grimmett (1999, p. 132)), i.e.

$$\exists\, \lambda(p) > 0 :\ P_p(|\mathcal{C}| = n) \le e^{-n\lambda(p)}\ \forall n \ge 1.$$

Therefore,

$$L \le \lim_{n\to\infty} \left(e^{-n\lambda(p)} / P_{p_c}(|\mathcal{C}| = n) \right) = 0. \tag{5.4}$$

2 *Supercritical case $p > p_c$.* In this case the decay is sub-exponential (Grimmett (1999, p. 216)):

$$\exists\, \eta(p) > 0 :\ \ P_p(|\mathcal{C}| = n) \le e^{-\eta(p) n^{(d-1)/d}}\ \forall n \ge 1,$$

and therefore

$$L \le \lim_{n\to\infty} \left(e^{-n^{(d-1)/d}\eta(p)} / P_{p_c}(|\mathcal{C}| = n) \right) = 0. \tag{5.5}$$

Since L is non-negative, it follows from (5.4) and (5.5) that $L = 0$, $p \neq p_c$.

In fact, the convergence here is uniform, as both $\lambda(p)$ and $\eta(p)$ can be separated from zero uniformly for all values of p from any interval Δ of the following form:

$$\Delta = [\alpha, p_c - \gamma] \cup [p_c + \gamma, \beta] \subset (0,1).$$

Thus, if

$$L_n(p) := \frac{\mathcal{L}_n(p)}{\mathcal{L}_n(p_c)},$$

then

$$L_n(p) \to \mathbb{1}_{\{p=p_c\}} = \begin{cases} 0 & p \neq p_c \\ 1 & p = p_c \end{cases},$$

pointwise for all $p \in (0,1)$, and the convergence

$$L_n(p) \to 0$$

is uniform on any interval

$$\Delta = [\alpha, p_c - \gamma] \cup [p_c + \gamma, \beta],\ \Delta \subset (0,1).$$

□

We continue with the proof of Theorem 5.1.3.

Proof. (Theorem 5.1.3) Consider the likelihood function $\mathcal{L}_n(p)$. The result of Lemma 5.1.4 being reformulated in ε-terms would mean that $\forall \varepsilon > 0\ \exists\, N(\varepsilon, \gamma) > 0$, such that

$$\mathcal{L}_n(p) < \varepsilon \mathcal{L}_n(p_c),\ \forall n > N(\varepsilon, \gamma)\, \forall p \in \Delta = [\alpha, p_c - \gamma] \cup [p_c + \gamma, \beta), \tag{5.6}$$

for any α and β, such that $\Delta \subset (0,1)$.

The quantity $\hat{p}_n$ being the maximum likelihood estimate for p is the mode of $\mathcal{L}_n(p)$:

$$\hat{p}_n := \arg \max_{p \in (0,1)} \mathcal{L}_n(p),$$

i.e.

$$\mathcal{L}_n(\hat{p}_n) \geq \mathcal{L}_n(p)\ \forall p \in (0,1). \tag{5.7}$$

Consider the sequence of maximum likelihood estimates $\{\hat{p}_n\}_{n=1}^{\infty}$. We will prove now that this sequence converges to p_c. Suppose, conversely, this is not the case. That would in particular mean that

$$\exists\, \zeta \in (0,1) :\ \forall M > 0\, \exists\, n > M :\ |\hat{p}_n - p_c| > \zeta.$$

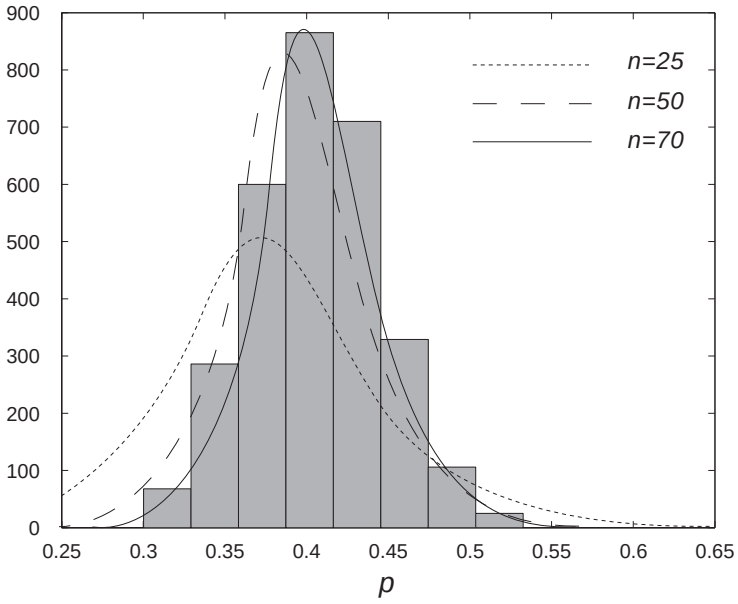


Figure 5.6: Likelihood functions $\mathcal{L}_n(p)$ $(n = 25, 50, 70)$ obtained using the MCMC from Algorithm 2 and MCMC sample histogram of $\mathcal{L}_n(p)$ for $n = 70$.

Take $M(\zeta) = N(\zeta, \zeta/2)$, then $\exists\, n > M(\zeta) : |\hat{p}_n - p_c| > \zeta$, i.e. $p_c \neq \hat{p}_n$. At the same time (when $\varepsilon = \zeta$) the following holds by (5.6):

$$\mathcal{L}_n(\hat{p}_n) < \zeta \mathcal{L}_n(p_c) < \mathcal{L}_n(p_c),$$

which is in contradiction with (5.7). Hence, $\hat{p}_n \to p_c$, $n \to \infty$. □

Figure 5.6 depicts the plots of likelihood function $\mathcal{L}_n$ when $n = 25, 50, 70$. These plots were obtained by smoothing the histograms of samples generated by the MCMC described in Algorithm 2. It is noticeable, and indeed intuitively expected, that the maximums of $\mathcal{L}_{25}(p)$, $\mathcal{L}_{50}(p)$ and $\mathcal{L}_{70}(p)$ are increasing. This observation, together with intuitive expectation, gives rise to the following conjecture.

Conjecture 5.1.5. *The sequence* $\{\hat{p}_n\}$ *converges to* p_c *monotonically from the left.*

Recall that the posterior distribution $\pi(p \,|\, n) := \pi(p \,||C| = n)$ is a density function proportional to both the likelihood and the prior distribution $\pi(p)$:

$$\pi(p \,|\, n) \propto \mathcal{L}_n(p)\pi(p).$$

We use the following heuristical argument to formulate a conjecture regarding the asymptotic form of the posterior function. If n is very large, then it is both very unlikely that the true value of p is either less or greater than p_c. We believe therefore that the likelihood function is increasingly concentrated around $p = p_c$ in such a way that it has certain implications on the posterior distribution of p, should p_c be not ignored by the prior $\pi(p)$. We formulate the corresponding conjecture using the notion of a delta sequence (see Appendix B).

Conjecture 5.1.6. *Provided $p \in \operatorname{supp}\pi(\cdot)$ the functional sequence $\{\pi(p \mid n)\}_{n=1}^{\infty}$ is a delta sequence which generates the delta function $\delta(p - p_c)$.*

Thus, we believe that the limiting posterior distribution of the percolation parameter is a one-point mass distribution at $p = p_c$, or the Dirac delta function $\delta(p - p_c)$ (Appendix B).

Theorem 5.1.3 and Conjectures 5.1.5, 5.1.6 together with MCMC described in Algorithm 2 give a tool of approximate estimation of p_c. It follows from Theorem 5.1.3 that maximums of the likelihoods tend to p_c as the size n of the cluster $\mathcal{C}$ increases, and this convergence, if Conjecture 5.1.6 is true, is monotonic. These maximums, however, can be approximated by MCMC sampling using Algorithm 2 and taking the uniform prior for p. By virtue of Conjecture 5.1.6—should this conjecture hold—the error of such approximation should diminish as n increases. Knowledge of the rate at which this error is (hypothetically) decreasing could help to better understand the limits of this method of estimation of p_c; this in particular includes scenarios of other sorts of lattices (perhaps, locally finite lattices) for which the exact values of p_c are unknown, but results similar to Theorem 5.1.3 and Conjectures 5.1.5, 5.1.6 hold true.

Combinatorial characterisation of large percolation clusters on $\mathbb{L}^d$

The theoretical results obtained and conjectured previously for inference under scenario $\mathcal{S}2$ can be used to derive their combinatorial analogues regarding the relative number of realisations of the process with the cluster size n. Under scenario $\mathcal{S}2$ the posterior distribution $\pi(p \mid |\mathcal{C}| = n)$ can be seen as a mixture of beta distributions, as in (5.1). Conjecture 5.1.6 implies that the number $q(s, k, l)$ of

graphs G corresponding to open clusters $\mathcal{C}$ which could emerge as a result of the percolation process with parameter p_c and for which it holds that

$$\frac{k+1}{s+l+2} \approx p_c, \tag{5.8}$$

where k is the number of edges in G, s is the number of edges in the saturation of G, and l is the number of edges of Π between the surface and the frontier of G, is far greater than the number of all other graphs. This is so, since the sequence of beta distributions $\text{Beta}(\alpha_n, \beta_n)$ is a delta sequence generating the delta function at p_c if and only if $\alpha_n/(\alpha_n + \beta_n) \to p_c$.

Thus, in percolation processes on $\mathbb{L}^d$ the number of finite graphs corresponding to open clusters of size n (where n is a very large number) that satisfy the condition

$$\frac{e(G)+1}{e(\text{Satur}\, G) + w(G) + 2} \approx p_c \tag{5.9}$$

largely exceeds the number of all other connected components on n nodes. In other words, a typical graph corresponding to an open cluster of a large size in percolation process on $\mathbb{L}^2$ is such a graph for which (5.9) holds. In particular, when $d = 2$:

$$e(G) - w(G) \approx e(\text{Satur}\, G) - e(G); \tag{5.10}$$

that is the number of present edges in G approximately equates to the total number of 'absent' edges[2] and edges between G and its frontier.

Large percolation clusters as rare events

When n is large, the appearance of finite open clusters of size n is highly unlikely: the distribution of the cluster size (hypothetically) decays as $n^{-1-1/\delta}$, $\delta > 0$, when $p = p_c$, and the decay is exponential (sub-exponential) when $p < p_c$ ($p > p_c$). Large finite percolation clusters can therefore be viewed as rare events. Since the state space of the MCMC proposed for inference on the percolation parameter p under scenario $\mathcal{S}2$ and described in Algorithm 2 involves the set of all open clusters on n nodes, this algorithm can be readily used in order to obtain realisations of these rare events.

[2]By 'absent' edges of G we mean edges of Π with both endpoints from G.

5.1.3 Bayesian optimal designs and inner-outer plots

In § 5.1.2 we introduced two scenarios of incomplete observations for percolation processes on the square lattices and considered the problem of inference on p under each of those scenarios. We turn to the question of optimal experimentation within the utility-based Bayesian framework in the context of percolation model now.

We would like to make a few observations before moving on. First, we would like to mention that the percolation model that we consider in this section can be seen as a particular case of the pairwise interaction model introduced earlier in § 1.2.1 and further specified in graph-theoretic terms in § 3.5.1 (see Examples 2.1.3, 3.5.4). Secondly, as it was noted in Example 3.5.4, the most 'packed' node configuration (given the configuration's size is fixed) is the most optimal design for random graph with nearest-neighbour links when one has access to complete information about graph realisations. Our final comment relates therefore to the problem of n-node optimal design for random graphs which was formulated in § 3.5.2. In this problem the design space was restricted to consist of sets of nodes of a fixed size (cardinality). In what follows we are going to relax this restriction.

We now adopt a Bayesian approach to design optimal experiments for our particular percolation model with incomplete observations of the kind described by scenario $\mathcal{S}1$ by identifying a class of designs (node configurations) which we call *'inner-outer' plots*. These plots of limited size are obtained by removing some sites from mostly 'packed' configurations, or, equivalently, by removing some nodes of the underlying grid. We refer to this process as *'sparsification of the grid'*. A typical example of an inner-outer plot is given in Figure 5.7. We will focus our attention on inner-outer designs which we now formally describe.

Let us assume that m is an odd positive integer and $r \in \mathbb{N} \cup \{0\}$. An *inner-outer* (m, r)*-plot* $\Pi_0^{(d)}(m, r)$ *in* $\mathbb{L}^d$ *with centre at the origin* is a d-dimensional box $B_N^{(d)}$ with side-length $N = m + 4r$ and some vertices removed as follows:

$$\Pi_0(m,r) := \begin{cases} B_N^{(d)}, & r = 0 \\ B_N^{(d)} \setminus \{x \in B_N^{(d)} : \| x \|_\infty = m + 2j + 1, j = 0, \ldots, r-1 \,\& \\ \qquad \& \| x \|_1 \equiv 0 \bmod 2\}, & r > 0, \end{cases} \tag{5.11}$$

where $\| x \|_\infty = \max(|x_1|, \ldots, |x_d|)$ and $\| x \|_1 = \sum_{i=1}^{d} |x_i|$ for any $x = (x_1, \ldots, x_d) \in \mathbb{Z}^2$ and the box $B_N^{(d)}$ is defined as follows[3]:

$$B_N^{(d)} := [-(N-1)/2, (N-1)/2]^d = \{x \in \mathbb{Z}^d : \| x \|_\infty \leq (N-1)/2\}.$$

We call any plot that can be obtained by shifting the plot $\Pi_0^{(d)}(m, r)$ in $\mathbb{L}^d$ an *inner-outer* (m, r)-plot, or simply an inner-outer plot, and denote it by $\Pi^{(d)}(m, r)$.

The total number of nodes contained in an (m, r)-plot can be calculated by subtracting the total number of 'sparsifying' (removed) nodes from the outer plot (as prescribed by (5.11)) as follows:

$$\begin{aligned} T &= N(m,r)^2 - 4\sum_{i=1}^{r}\left(\frac{m-1}{2} + 2i - 1\right) \\ &= N(m,r)^2 - 2mr - 4r(r+1) + 6r \\ &= N(m,r)^2 - 2r(m + 2r - 1), \end{aligned} \tag{5.12}$$

where

$$N(m, r) = m + 4r. \tag{5.13}$$

The inner-outer (m, r)-plot presented in Figure 5.7 is from $\mathbb{L}^2$. In this example the size of the inner plot is $m \times m$, where $m = 9$, and there are $r = 3$ 'circles' (with respect to the metric $\| \cdot \|_\infty$) in the outer plot from which every second site is removed. The size of the bounding box is $N \times N$, where $N = m + 4r = 21$. The total number T of nodes that this configuration contains is 357 (according to (5.12)).

Table 5.1 contains some values for m and r (up to 25 for m and 5 for r) as well as corresponding values of N and T. The possible values of r are located in the first row of the table, whereas the possible values of m are to be found in the second column of it (these values also coincide with N since they correspond to $r = 0$). The values of N can be found at the intersection of a row and a column corresponding to the values of m and r. The total number of nodes in an (m, r)-plot can be found to the right of the value of $N(m, r)$ in the same row.

[3]Note that N is a positive integer number since m is odd.

$r =$	*0*	*0*	*1*	*1*	*2*	*2*	*3*	*3*	*4*	*4*	*5*	*5*
	N	T	N	T	N	T	N	T	N	T	N	T
$m =$	*3*	**9**	7	**41**	11	**97**	15	**177**	19	**281**	23	**409**
	5	**25**	9	**69**	13	**137**	17	**229**	21	**345**	25	**485**
	7	**49**	11	**105**	15	**185**	19	**289**	23	**417**	27	**569**
	9	**81**	13	**149**	17	**241**	21	**357**	25	**497**	29	**661**
	11	**121**	15	**201**	19	**305**	23	**433**	27	**585**	31	**761**
	13	**169**	17	**261**	21	**377**	25	**517**	29	**681**	33	**869**
	15	**225**	19	**329**	23	**457**	27	**609**	31	**785**	35	**985**
	17	**289**	21	**405**	25	**545**	29	**709**	33	**897**	37	**1109**
	19	**361**	23	**489**	27	**641**	31	**817**	35	**1017**	39	**1241**
	21	**441**	25	**581**	29	**745**	33	**933**	37	**1145**	41	**1381**
	23	**529**	27	**681**	31	**857**	35	**1057**	39	**1281**	43	**1529**
	25	**625**	29	**789**	33	**977**	37	**1189**	41	**1425**	45	**1685**

Table 5.1: Table comprising some values of m and r (up to 25 for m and 5 for r) as well as corresponding values of N and T. The possible values of r (italicised) are located in the first row of the table, whereas the possible values of m (italicised) are to be found in the second column of it (these values also coincide with N since they correspond to $r = 0$). The values of N can be found at the intersection of a row and a column corresponding to the values of m and r. The total number of nodes T in an (m, r)-plot can be found to the right of the value of $N(m, r)$ in the same row (these numbers are in bold). The values of N and T were calculated using (5.13) and (5.12) respectively.

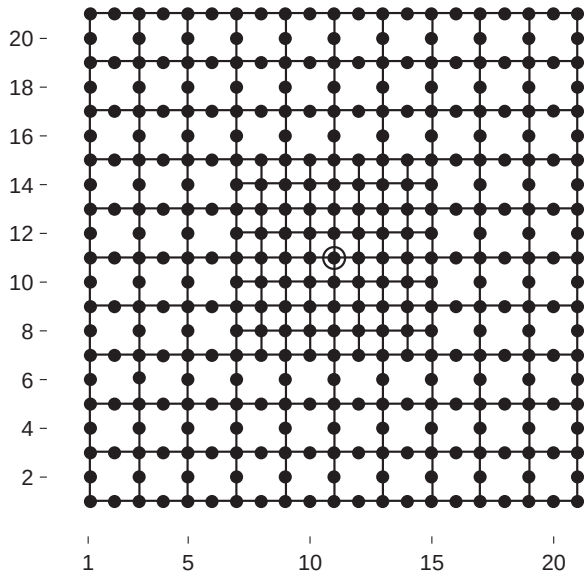


Figure 5.7: Example of an inner-outer (m, r)-plot in $\mathbb{L}^2$: here $m = 9$ and $r = 3$. The plot is bounded by an $N \times N$ square, where N, according to (5.13), equals 21.

Optimal design problem: the model and design space

The model that we consider now is that of an *SIR* epidemic with constant infectious periods taking off from the central site of an inner-outer plot in $\mathbb{L}^d$ and evolving on that plot according to nearest-neighbour interaction rule. This model is equivalent to the model discussed in § 5.1.1 when Π is an inner-outer plot $\Pi^{(d)}(m, r)$. We consider this model in the context of scenario $\mathcal{S}1$, that is when the only information available about the outcome of the epidemic is its site configuration.

Since our model is essentially the same percolation model considered previously, all the terminology from § 5.1.2 remains the same. Figure 5.8 depicts a simulated connected component (from left) emerged as a result of the percolation process with parameter $p = 0.52$ on the inner-outer $(9, 2)$ plot in $\mathbb{L}^2$. This connected component contains the central node (denoted by a circle) of the underlying inner-outer plot. The plot from right in the same figure depicts the saturation graph of the site configuration from left with respect to the underlying inner-outer plot. Similarly, Figure 5.9 shows a simulated connected component on an inner-outer plot of the larger size ($m = 23$, $r = 4$) and higher value of the percolation parame-

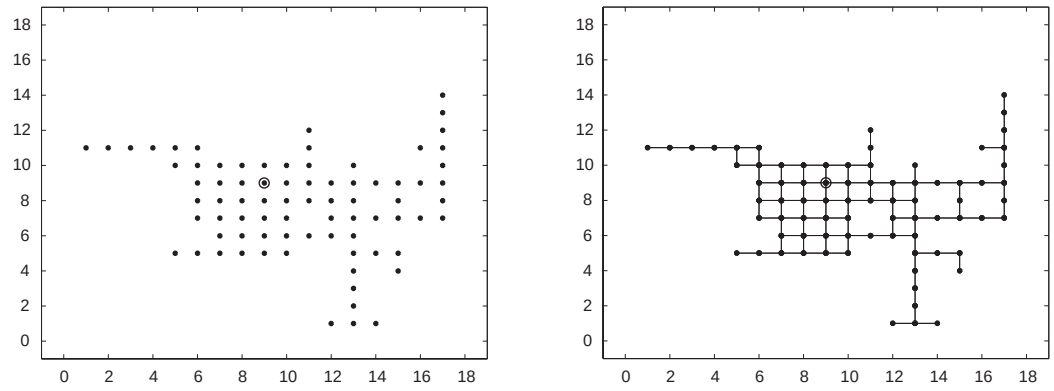


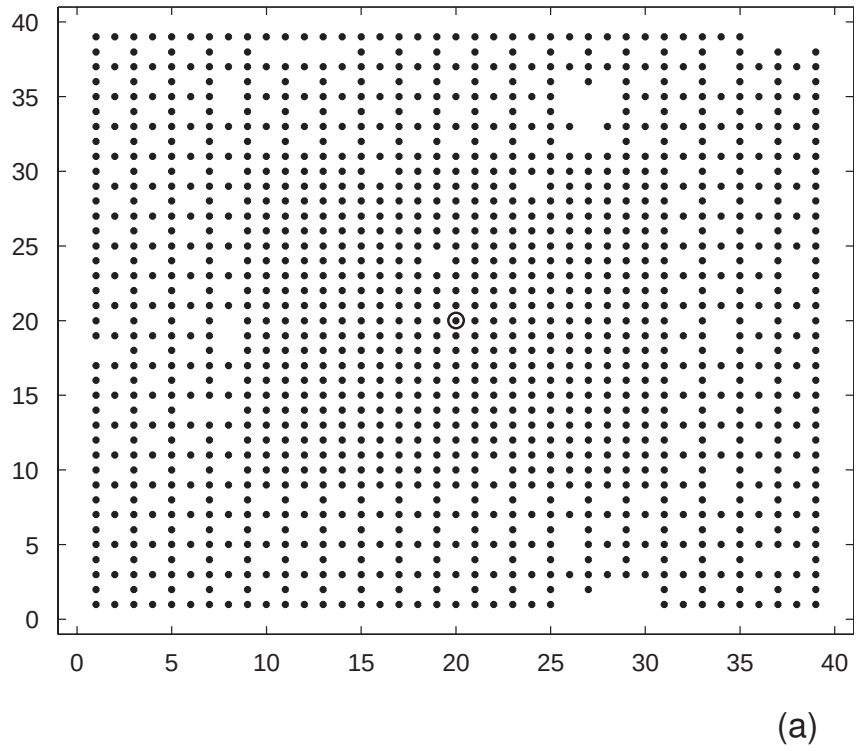
Figure 5.8: Left: An open cluster $\mathcal{C}$ simulated on the inner-outer plot $\Pi^{(2)}(9,2)$ in $\mathbb{L}^2$ with $p = 0.52$; the central node (an initially inoculated site) is denoted by a circle. Right: The fully saturated graph derived from $\mathcal{C}$ with respect to the vertex set $\Pi^{(2)}(9,2)$ and nearest-neighbour interaction.

ter, $p = 0.86$ as well as the fully saturated graph induced by this site configuration, the underlying inner-outer plot $\Pi^{(2)}(23,4)$ and edge set $\mathbb{E}^2$.

The optimal design problem which we formulate for percolation processes evolving on inner-outer plots is based entirely on the Bayesian approach (involving a utility function) presented in §3.1.2. Since our ultimate goal in designing and performing an experiment is to learn as much information about the model parameter (percolation probability parameter p in this case) as possible, the utility function $u(d, y, p)$ is the logarithmic ratio of the posterior distribution $\pi(p \,|\, y, d)$ and the prior distribution $\pi(p)$ of p given a realisation y of the percolation process on an inner outer plot d: $u(d, y, p) = \log \frac{\pi(p \,|\, y, d)}{\pi(p)}$. Thus, the design space $\mathcal{D}$ is a set of inner-outer plots of certain type (which is to be further specified) and the object of interest is the expected utility

$$U(d) = \int_0^1 \int_{\mathcal{Y}} \log \frac{\pi(p \,|\, y, d)}{\pi(p)} f(y \,|\, p, d) \pi(\theta) \, \mathrm{d}p \, \mathrm{d}y \tag{5.14}$$

which has to be maximised in order to find the optimal inner-outer plot $d = \Pi^{(n)}(m, r) \in \mathcal{D}$, $n = 2, 3, \ldots$. Here $\mathcal{Y}$ denotes the set of all connected components on d containing the central node. We use the methods discussed in §§ 3.3, 3.4 in order to solve the optimal design problem in hand. This will be done for both the progressive and instructive design scenarios.



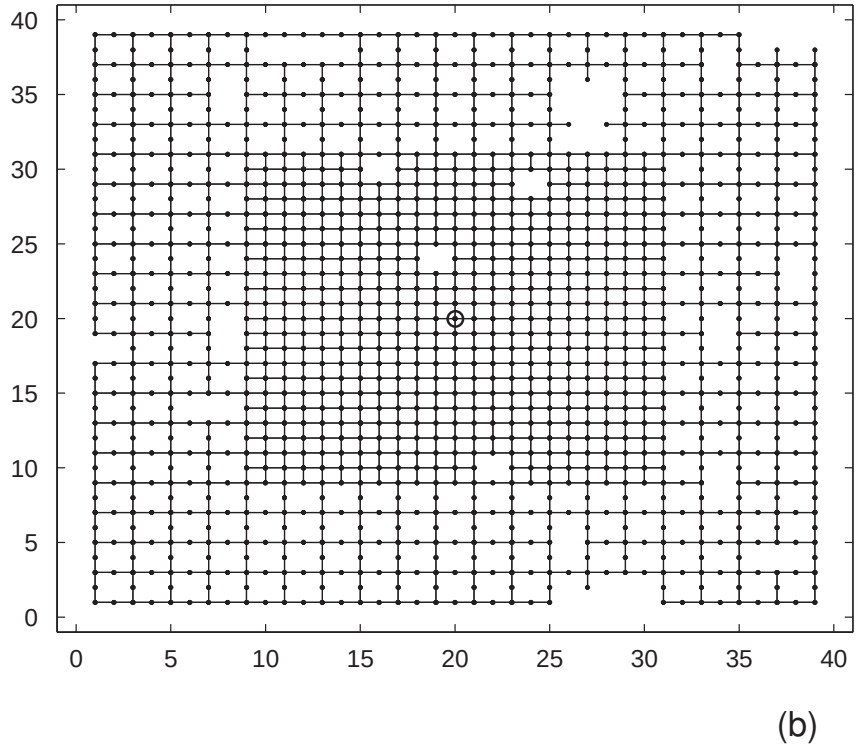


Figure 5.9: (a) An open cluster $\mathcal{C}$ simulated on the inner-outer plot $\Pi^{(2)}(23, 4)$ in $\mathbb{L}^2$ with $p = 0.86$; the central node (an initially inoculated site) is denoted by a circle. (b) The fully saturated graph derived from $\mathcal{C}$ with respect to the vertex set $\Pi^{(2)}(23, 4)$ and nearest-neighbour interaction.

The choice of the design space $\mathcal{D}$ can be made in a number of ways, possibly reflecting such restrictions as a limited number of experimental units or limited size of the experimental plot. In the context of inner-outer plots the former restriction would mean that T from (5.12) is bounded from above, whereas the latter condition is equivalent to bounding the quantity $N(m, r)$. A combination of these conditions or some other information can be also taken into account when identifying the design space.

In our further practical examples we will assume that the design space is of the form

$$\mathcal{D} = \{d \,|\, d = \Pi^{(2)}(m, r) \;\&\; m + 4r = N\}$$

given the side length N of the experimental plot (N should be an odd number). For example, if $N = 19$ then, as it can easily be read from Table 5.1

$$\mathcal{D} = \{\Pi^{(2)}(19, 0), \Pi^{(2)}(15, 1), \Pi^{(2)}(11, 2), \Pi^{(2)}(7, 3), \Pi^{(2)}(3, 4)\}.$$

Inference for percolation process on inner-outer plots

Finding the optimal design under both the 'progressive' and 'instructive' experimentation scenarios involves evaluation of the likelihood function of the model parameter for incomplete observations (under scenario $\mathcal{S}1$). The likelihood can be evaluated using the MCMC presented in § 5.1.2 and described in Algorithm 1.

Let us consider a few examples with different choice of the inner-outer plot and values of the percolation parameter.

Figure 5.10 depicts a histogram of a sample from the posterior distribution (assuming a uniform prior $\mathrm{U}(0, 1)$) obtained using the MCMC from Algorithm 1 for the site configuration from Figure 5.8 on the $(9, 2)$-plot embedded in the lattice $\mathbb{L}^2$ and the corresponding trace plot for the first $15 \cdot 10^3$ steps of the chain.

Figure 5.11 contains plots of a simulated open cluster obtained on the inner-outer plot $\Pi^{(2)}(23, 4)$ using $p = 0.86$ (left plot) and the likelihood sample histogram obtained after running our Markov chain from Algorithm 1 with this open cluster as the input data. Figure 5.12 shows the same kind of plots for a realisation of the percolation process on $\Pi^{(2)}(13, 2)$ when the parameter p was taken to be equal to 0.9.

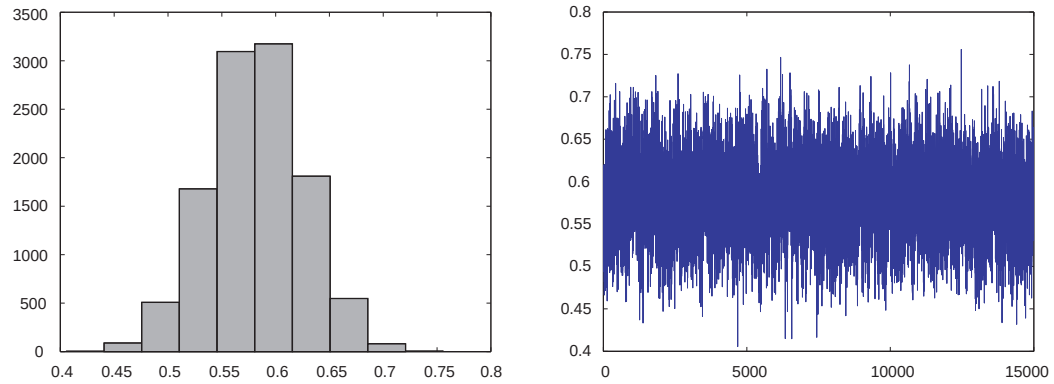


Figure 5.10: Inference on the percolation parameter for the configuration from the left plot in Figure 5.8. Left: Sample histogram obtained by running MCMC for this configuration. Right: MCMC trace plot of updates for p. The value of p for which the configuration in Figure 5.8 was obtained is 0.52.

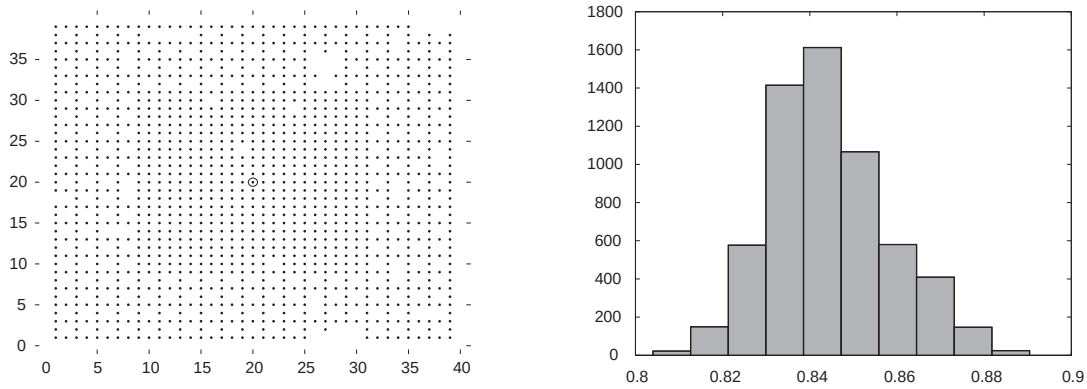


Figure 5.11: Left: open cluster from Figure 5.9(a) obtained on the inner-outer $(23, 4)$-plot using $p = 0.86$. Right: MCMC sample histogram for p assuming the uniform prior $\mathrm{U}(0, 1)$ for this parameter.

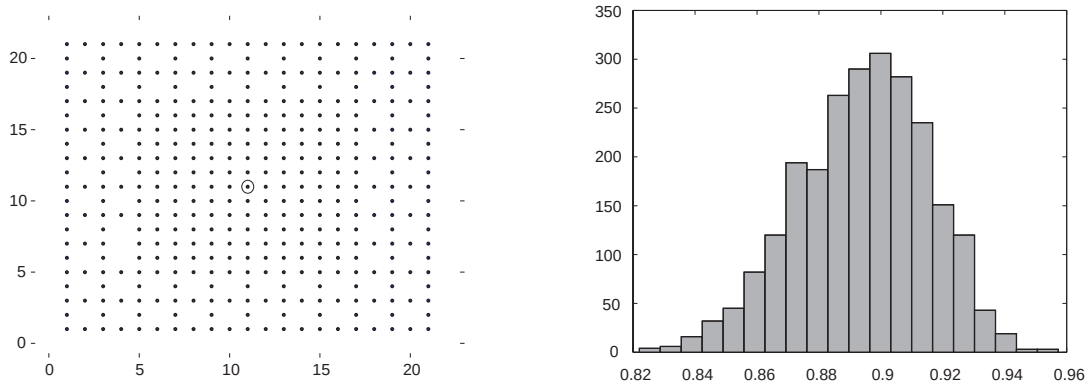


Figure 5.12: Left: simulated open cluster obtained on the inner-outer $(13, 2)$-plot using $p = 0.9$. Right: MCMC sample histogram for p assuming the uniform prior $\mathrm{U}(0, 1)$ for this parameter.

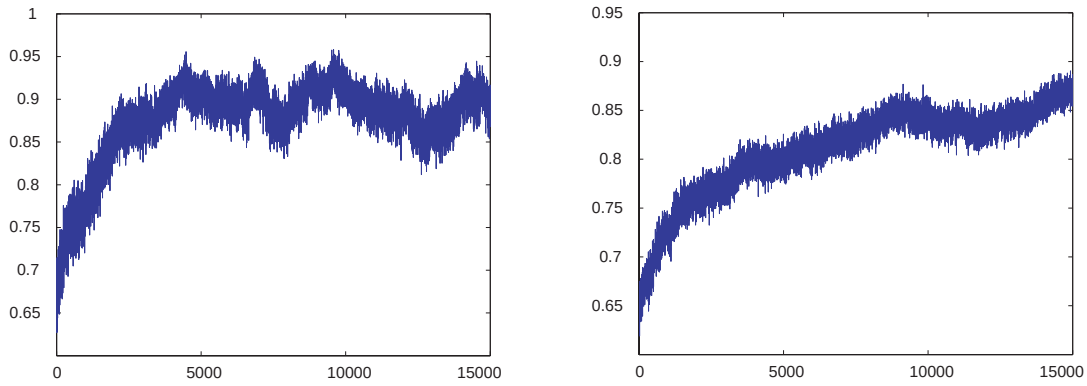


Figure 5.13: Left: MCMC trace plot of updates in p for the site configuration from Figure 5.12. Right: part of the burn-in period of the MCMC trace plot of updates in p for the site configuration from Figure 5.11; this part of the update trace was not used in producing the histogram in Figure 5.11.

Figure 5.10 (right) and Figure 5.13 display early stages of updates in p (trace plots) for three different site configurations. These figures show that the mixing time of the chain may vary considerably depending on the size of the site configuration and its connectivity properties. In practice, it never takes too long to see the chain described by Algorithm 1 converging: both the connected component updates and the percolation parameter updates can be efficiently realised[4], allowing one to perform the sampling from the posterior distribution $\pi(p \,|\, \mathcal{C})$ reasonably fast. However, it is important to note that Figure 5.10 and Figure 5.13 suggest that chains from Algorithm 1 corresponding to larger site configurations have rather poor mixing properties. This is in contrast to trace plots from Figures 5.4 and 5.10 which correspond to smaller configurations. One solution to this problem would be to update p not as often as the connected component is updated. This certainly requires further experimentation and exploration in order to find recipes of better chain mixing.

5.1.4 Implementation of progressive and instructive designs based on inner-outer plots

Given a finite design space $\mathcal{D}$ it is fairly straightforward to solve the optimal design problem for percolation model on inner-outer plots using the tools we developed previously. Let us make a few comments about solving the problem under each of the two design scenarios.

Progressive design: expected utility evaluation through augmented modelling

Since $\mathcal{D}$ is finite we choose to identify the design that maximises the expected utility function in the 'progressive' case using augmented modelling which was described in Section 3.4. Recall that this is based on an artificial distribution $h(d, p, y) \propto u(d, p, y) f(y \,|\, p, d) \pi(p)$ from which samples are taken via a Metropolis–

[4]Updates of the connected component, that is deletion and insertion of edges while preserving the *connectedness* of the underlying graph, can be greatly improved using dynamic graph algorithms (e.g. see Zaroliagis (2002)).

Hastings sampler. The optimal design d^* is identified then as a value of d at which the marginal of h is maximised.

Instructive design: Monte Carlo evaluation of the expected utility

We treat the 'instructive' case differently from that of the 'progressive' one because of the form of the expected utility under this scenario. Recall that in the 'instructive' case whenever the 'instructor' knows the true value p^* of the model parameter p one can write the expected utility based on the Kullback–Leibler divergence as follows:

$$U^*_{\mathrm{KL}}(d) = \int\limits_{\mathcal{Y}} \int\limits_{0}^{1} \log \frac{\pi(p \,|\, y, d)}{\pi(p)} \pi(p \,|\, y, d) \,\mathrm{d}p \, f(y \,|\, p^*, d) \,\mathrm{d}y,$$

where $f(y \,|\, p^*, d)$ is the likelihood function evaluated at the true value of the model parameter p^* given the open cluster y and design d.

One can see that if we choose to evaluate the expected utility $U^*_{\mathrm{KL}}(d)$ via standard Monte Carlo simulation, then a Markov chain has to be run each time we sample a new observation (an open cluster) y and also the potentially time-consuming integration has to be done with respect to the model parameter p. This integration can be implemented in the following way. Since we can sample from the posterior $\pi(p \,|\, y, d)$ via MCMC technique (Algorithm 1) for any given open cluster y, we do so and then fit the beta distribution (or some other distribution) to the MCMC sample obtained in order to perform integration numerically in a more efficient way.

Thus, the expected utility evaluation scheme for inner-outer plots in the 'instructive' case and scenario $\mathcal{S}1$ can be described as follows.

For each inner-outer plot $d \in \mathcal{D}$ do the following:

1 generate a random sample of M independent connected clusters $\{y_i\}_{i=1}^{M}$ on $\Pi^{(2)}(m, r)$: $y_i \sim f(y \,|\, p^*, d)$;

2 perform M MCMC's in order to obtain M independent samples for the posterior distribution $\pi(p \,|\, y, d)$;

3 fit beta distribution to each of the obtained samples; refer to the fitted distributions as $\pi(p \,|\, y_i, d)$;

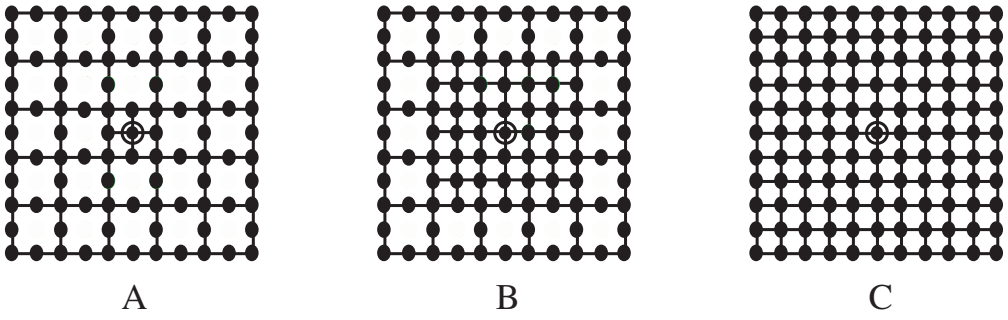


Figure 5.14: *Inner-outer* design plots A, B, and C form the design space $\mathcal{D} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$.

4 evaluate numerically the integrals $I_i := \int_0^1 \log \frac{\pi(p \,|\, y_i, d)}{\pi(p)} \pi(p \,|\, y_i, d)\, \mathrm{d}p$;

5 evaluate the expected utility: $\bar{U}_M = \frac{1}{M} \sum_{i=1}^{M} I_i$.

One may wonder how well the true posteriors can be fitted by the Beta distribution family (step 3 of the above scheme). The author's experience suggests that such a fit never affects the outcome of the analysis on the qualitative level unless the prior distribution has more than one local mode or its support is smaller than the entire interval $(0, 1)$. Recall that the purpose of this step is to make evaluation of the integrals I_i easier and faster, and hence the family of beta distributions is only one of possible choices. For instance, if the prior distribution $\pi(p)$ is uniform on $(0, 1)$, then the integrals I_i represent the entropies of the fitted distributions. In the case when the fitting is done by beta distributions, each of these integrals can be quickly calculated using the following analytical formula for the entropy of the beta distribution $\mathrm{Beta}(\alpha, \beta)$:

$$\mathrm{Ent}\{\mathrm{Beta}(\alpha, \beta)\} = \log B(\alpha, \beta) - (\alpha - 1)\psi(\alpha) - (\beta - 1)\psi(\beta) + (\alpha + \beta - 2)\psi(\alpha + \beta),$$

where ψ is the *digamma function*, $\psi(z) = \Gamma'(z)/\Gamma(z)$. Other forms of the prior distribution may condition the choice of the family of fitting distributions when trying to facilitate the calculation of the integrals I_i, $i = 1, \ldots, M$; methodologically, discretising both the prior and posterior is also an option for this stage of the solution to the optimisation problem.

Example 5.1.7. *In our example we consider all inner-outer plots in* $\mathbb{L}^2$ *whose sizes do not exceed* $N = 11$. *There are only three such plots:* $\Pi^{(2)}(3, 2)$, $\Pi^{(2)}(7, 1)$,

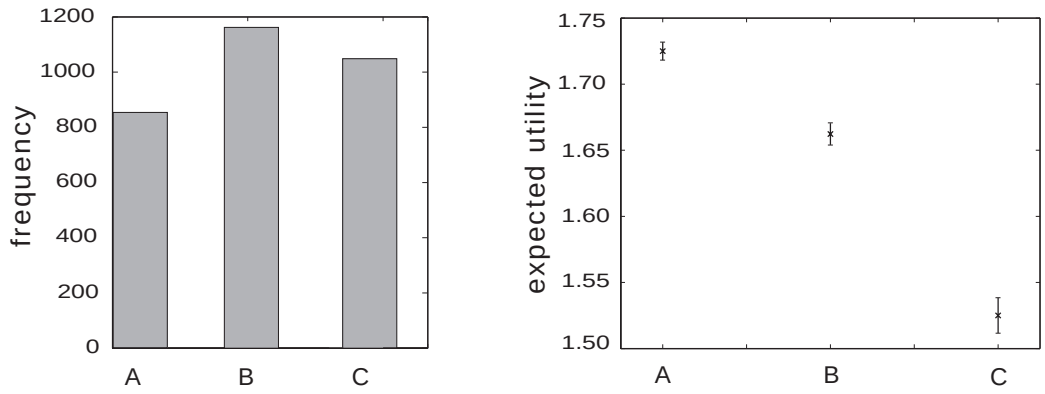


Figure 5.15: Left: sample histogram for the marginal of $h(d, p, y)$ in d, $d \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$, under progressive design and $\pi(p) \sim \mathrm{U}(0, 1)$. Right: evaluated expected utility under instructive design with $\pi^*(p) \equiv \delta(p-0.9)$ and 95% credibility intervals ($M = 1500$) for the plots A, B, and C, under instructive design.

and $\Pi^{(2)}(11, 0)$*. For ease of reference we mark them A, B, and C respectively (as depicted in Figure 5.14). Thus, the design space* $\mathcal{D} = \{A, B, C\}$ *consists of three designs, among which A is the mostly sparsified plot whereas no nodes are removed from C at all.*

Figure 5.15 represents graphically the results of the comparison of designs from $\mathcal{D}$ *under both 'progressive' and 'instructive' case when the prior distribution* $\pi(p)$ *is uniform on the interval* $(0, 1)$*. The left panel of the figure corresponds to the former scenario and depicts a histogram of a sample corresponding to the marginal of the artificial augmenting distribution* $h(d, p, y) \propto u(d, p, y) f(y \mid p, d)\pi(p)$ *in* $d \in \mathcal{D}$*. The right panel corresponds to the latter scenario and shows the Monte Carlo estimated values of the expected utilities and 95% credibility intervals for each of the three considered designs (*$M = 1500$*, see (3.30) in Section 3.4) assuming that the instructor's knowledge* $\pi^*(p)$ *about the model parameter is exact,* $\pi^*(p) = \delta(p-0.9)$*.*

The plots from Figure 5.15 indicate that the solutions to the optimal design problem under the two scenarios are different from each other. The 'moderately sparsified' plot B maximises the expected utility in the progressive case, that is in the case when there is just a single experimenter designing an experiment for himself. If, however, it is the instructor who knows the true value of the model parameter ($p = 0.9$*) and wants to choose the best convincing inner-outer plot from*

the set $\mathcal{D}$ for the experimenter to use it (instructive scenario), then the optimal plot is the 'mostly sparsified' inner-outer plot A. Notably, the 'mostly dense' plot C would be the worst choice in the instructive case, whereas it outperforms the 'mostly sparsified' plot A in the progressive case, but is worse than the 'moderately sparsified' plot B.

Although the inner-outer design plots introduced above represent a limited range of designs which can be defined using a lattice structure, the advantage of their use is that the dimension of the design space is reduced to one (recall that the design space is completely determined by the value of the inner-outer plot's side length N). Low dimensionality of the design space and its more complex structure and richness can still be achieved by considering less restrictive sparsification of the lattice-based plots—for example, by considering all connected components containing the origin within a set of nodes contained in a square or rectangle of a fixed size to be designs. The optimisation techniques employing MCMC sampling based on exploration of the connected components induced by these designs and augmented modelling remain, however, the same. These, together with more detailed study of dependence of optimal designs on the experimenter's prior $pi(p)$ and instructor's prior knowledge $\pi^*(p)$, will be investigated in future studies.

5.2 Lattice designs for inference on random graphs with long-range connections

Throughout the whole previous section it was assumed that we deal with a square lattice-based random graph model with nearest-neighbour connections. In this section we briefly discuss the potential and possibilities of working with greater variety of lattices, while keeping the dimension of the design space low, and also allowing long-range connections between graph nodes.

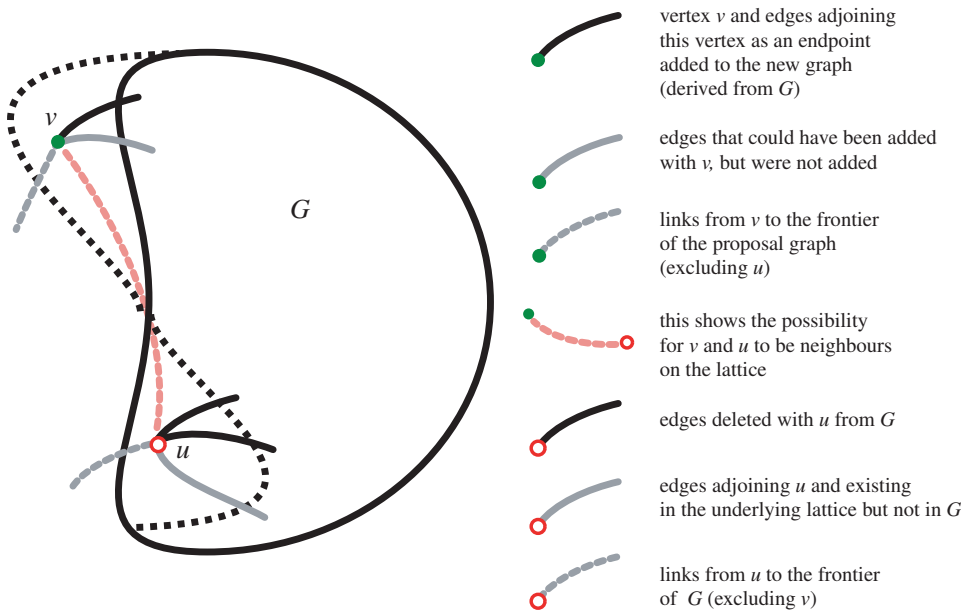


Figure 5.16: Updating connected component: graphical representation of Metropolis-Hastings step of Algorithm 2 for long-range interaction locally finite graph models.

5.2.1 Generalising results from the previous section

The results presented in the previous section with regard to making inference under scenarios $\mathcal{S}1$ and $\mathcal{S}2$ and looking for optimal node arrangements under scenario $\mathcal{S}1$ can be easily extended to the case of long-range connections. In fact, Algorithms 1 and 2 are already described in such a way that they can immediately be used for any locally finite graph as an underlying interaction topology. We will illustrate this using a schematic description of the main procedures that the mentioned algorithms involve: insertion and deletion of vertices and edges.

For example, in Algorithm 2 at each step of updating the current connected component G a vertex u is deleted at random from G and a vertex v, taken from the frontier Γ_G of the graph G, is added to G, thus forming a proposal graph $\tilde{G}$. Figure 5.16 graphically depicts this process: the vertex u is chosen randomly from G and will be deleted from G with all the edges which contain this vertex. The vertex v is chosen randomly from the frontier of G, Γ_G, and is added to the graph G with every possible edge, each independently with corresponding probability (see Algorithm 2).

The number of vertices of the resulting proposal graph $\tilde{G}$ remains unchanged, whereas the number of present and absent edges as well as the number of edges between $\tilde{G}$ and its frontier $\Gamma_{\tilde{G}}$ may be changed as a result of these operations, but can easily be maintained. Then the acceptance probability is calculated after it is detected that $\tilde{G}$ is connected. The latter check can be efficiently done using the classical *depth-first* or *breadth-first search* algorithms (see Gibbons (1985)), by starting traversing the graph from a single node and counting all nodes reached. Since every node and every edge will be explored in the worst case, (undirected) graph connectivity can be diagnosed in $O\left(n + e(\tilde{G})\right)$ steps[5], where n is the number of nodes in the graphs G and $\tilde{G}$ and $e(\tilde{G})$ is the number of edges in the proposal graph $\tilde{G}$.

In Theorem 5.1.3 it was shown that under Scenario $\mathcal{S}2$ the sequence of maximum likelihood estimates for the percolation parameter p converges to the critical percolation probability $p_c(d)$ of the square integer lattice $\mathbb{L}^{(d)}$, $d \geq 2$. The author of this thesis conjectures that a similar result holds for any long-range percolation model on infinite locally finite graphs.

5.2.2 Square lattice and its deformations

In Section 3.5 we formulated the n-node optimal design problem for random graphs. This problem consists in finding an n-node configuration design that maximises the expected utility function (the expected Kullback–Leibler divergence). The design parameters in this problem are either locations of the nodes or distances between them (or weights defined on the node binary relationship). If the design n nodes are to be taken from a region of cardinality of the continuum, then the cardinality of the design space would also become continuum. This would make the search for the optimal design excessively time-consuming. Identification of the optimum would also be difficult, since potential symmetries in the node arrangements would inevitably necessitate complex shaped constraints.

For example, consider three nodes arranged at the points d_1, d_2 and d_3 in $\mathbb{R}^3$. Clearly, the design $d = (d_1, d_2, d_3)$ has the same expected utility as any translation,

[5]That is in $O\left(n^2\right)$ steps in the worst case, when all or 'almost all' egdes are present.

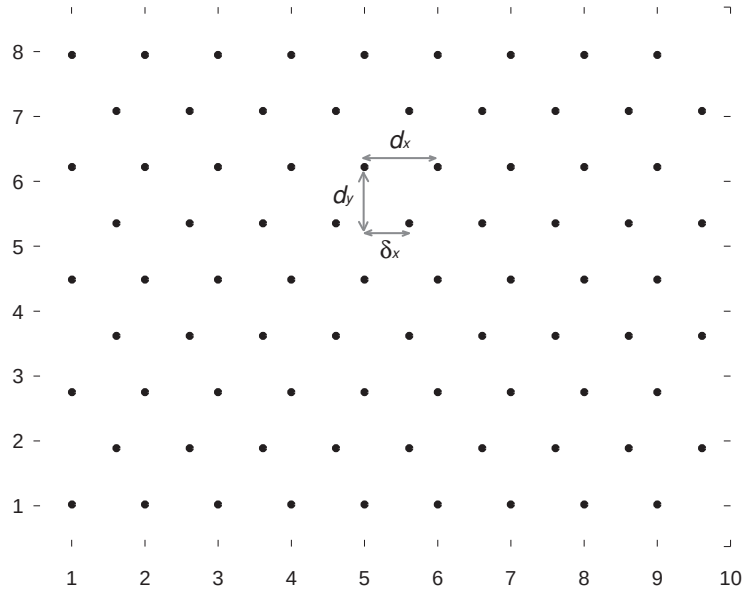
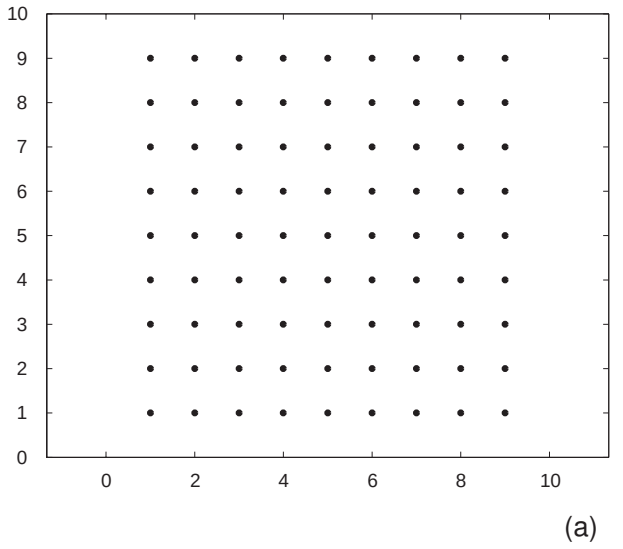


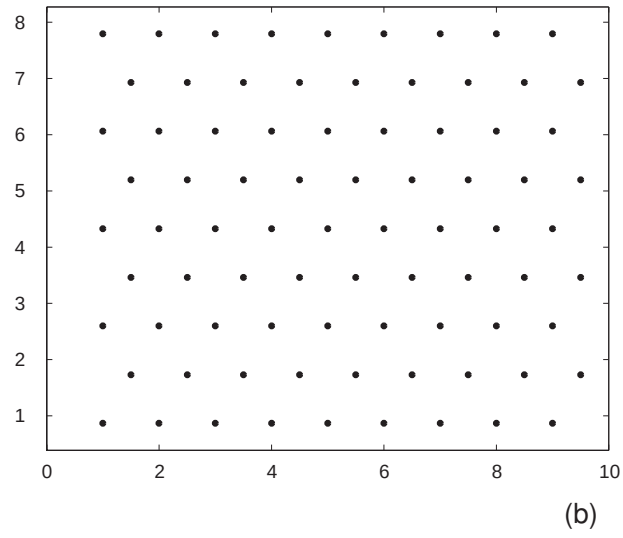
Figure 5.17: Modification of the planar square lattice. The modification parameters are as follows: d_x, the spacing between nodes in the horizontal direction; d_y, the spacing in the vertical direction; and δ_x, a displacement of every second row in the horizontal direction. All nodes of every second row are shifted to the right if $\delta_x > 0$, and to the left if $\delta_x < 0$.

rotation or reflection of it, and so the optimal design as well as any other design has an infinite number of superficial variants. Searching for the optimum requires (i) imposing constraints on the design space and, even if that is done, (ii) exploration of arrangements from a continuum design space.

The approach one might wish to take (and it is partly what we did in the previous section) is to impose a lattice structure on the points, thereby simplifying the design space and reducing its size considerably. More specifically, for planar designs we consider deformations of a square lattice with three design parameters that control the spacing and structure of the lattice: d_x, the spacing between nodes in the horizontal direction; d_y, the spacing in the vertical direction; and δ_x, a displacement of every second row in the horizontal direction. By varying these distances one can obtain the following lattices among others:

- *square* lattices ($d_x = d_y$, $\delta_x = 0$ or $d_y = \delta_x = d_x/2$), as in Figure 5.17 (a,c);
- *rectangular* lattices ($\delta_x = 0$);





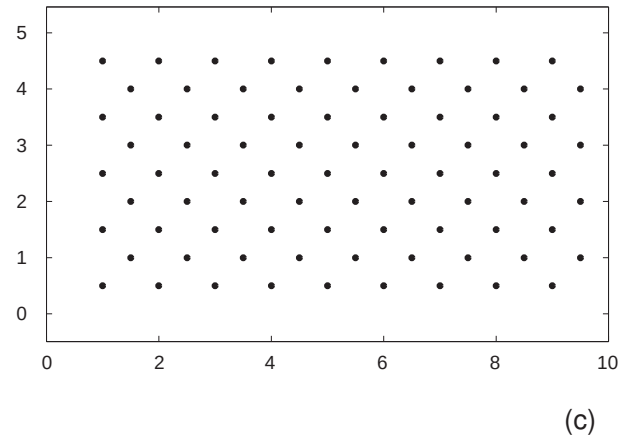


Figure 5.18: Examples of modified planar square lattices: (a) unchanged square lattice ($d_x = d_y$, $\delta_x = 0$); (b) hexagonal lattice ($d_y = \sqrt{3}d_x/2$); (c) square lattice ($d_y = \delta_x = d_x/2$). The number of nodes is the same in all three plots.

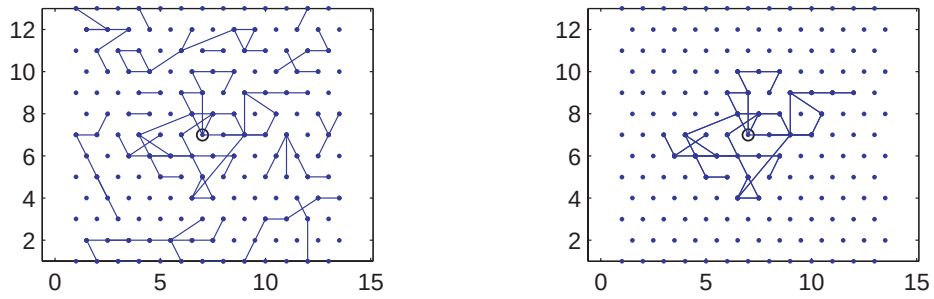


Figure 5.19: Left: long-range connections with exponential decay $p = e^{-\theta d}$ ($\theta = 1.9$) on a triangular 13×13 lattice plot ($d_x = d_y = 1$) with displacement $\delta_x = 1/2$. Right: the connected component of the graph from the left panel which contains the central node (in circle).

- *hexagonal* lattices ($d_y = \sqrt{3}d_x/2$, $\delta_x = d_x/2$), as in Figure 5.17 (b); note that this terminology may seem to be slightly ambiguous—the lattice for which it holds that $d_y = \sqrt{3}d_x/2$ and from which every second node of every second row is removed could also be justly called *hexagonal*;
- *triangular* lattices (e.g. $d_x = d_y = 2\delta_x$, as in Figure 5.19).

Not only the described lattices represent the underlying interaction topologies on which random graphs are considered, but they can also represent designs. For instance, when the size $n \times m$ of the design lattice-based plot is fixed, one can identify the design space by discretising one or more lattice parameters (d_x, d_y, δ_x). For example, if N is a fixed natural number, ε is a non-negative real number, $\varepsilon \in \mathbb{R}_+$, and H_ε denotes the set $\{h\varepsilon \,|\, h = 1, \ldots, N\}$, the design space $\mathcal{D}$ may be defined as a set of

- square lattices of the size $n \times m$ with $d_x \in H_\varepsilon$;
- rectangular lattices of the size $n \times m$ with $(d_x, d_y) \in H_\varepsilon \times H_\varepsilon$;
- hexagonal lattices of the size $n \times m$ with $d_x \in H_\varepsilon$.

The design space $\mathcal{D}$ can also contain lattices of different types. This may be useful when one wants to compare them and choose the type of the most optimal lattice in the random graph design problem.

Example 5.2.1. *In this example the model is a long-range percolation on a lattice with origin. For example, consider a triangular* 13×13 *lattice* ($d_x = d_y = 2\delta_x$). *The left plot in Figure 5.19 depicts a realisation of a long-range percolation process on such lattice when the connections are made according to exponential probability decay* $p(d) = e^{-\theta d}$ *with* $\theta = 1.9$. *The right plot in Figure 5.19 depicts the connected component containing the origin (the central node in a circle) only. We compare three different lattice types and two edge-probability decays when the graph is the connected component of the long-range percolation process at the origin.*

Let us consider lattices of the size 5×5 *nodes. In order to keep the dimension of the design space low we consider the triangular, square and hexagonal lattices for which the horizontal and vertical spacings are parametrised identically, that is we allow* d_x *to take values from a finite set of non-negative reals, and*

- *in the case of the triangular or square lattice we set* $d_y = d_x$;
- *in the case of the hexagonal lattice we set* $d_y = \sqrt{3}d_x/2$;

the shift δ_x *is varied in the same way for each of these lattices:* $\delta_x = d_x/2$.

Figure 5.20 and Figure 5.21 show the result of the expected utility evaluation (the expected utility minus the prior distribution entropy) for each of the lattices mentioned above and the two edge-probability functions: exponential, $p(d) = e^{-\theta d}$, *and Cauchy,* $p(d) = (1 + \theta d^2)^{-1}$, *respectively. The expected utility was evaluated on a finite set of values for the model parameter* d_x *via standard Monte Carlo sampling and averaging. Cubic spline interpolation, least squares spline approximation and smoothing spline of order 4 were used to represent the expected utility as a continuous curve. One can see the plots of the first derivative of the fitted least squares spline (or smoothing spline) in the right panels of these figures. The prior knowledge about the model parameter* θ *was taken to be* $\Gamma(10, 0.2)$ *in each of these two models.*

It is clear from the plots that the type of the lattice used did not affect essentially the solution of the optimal design problem in either of the these two models—the expected utility is maximised at roughly the same values of the lattice size parameter d_x.

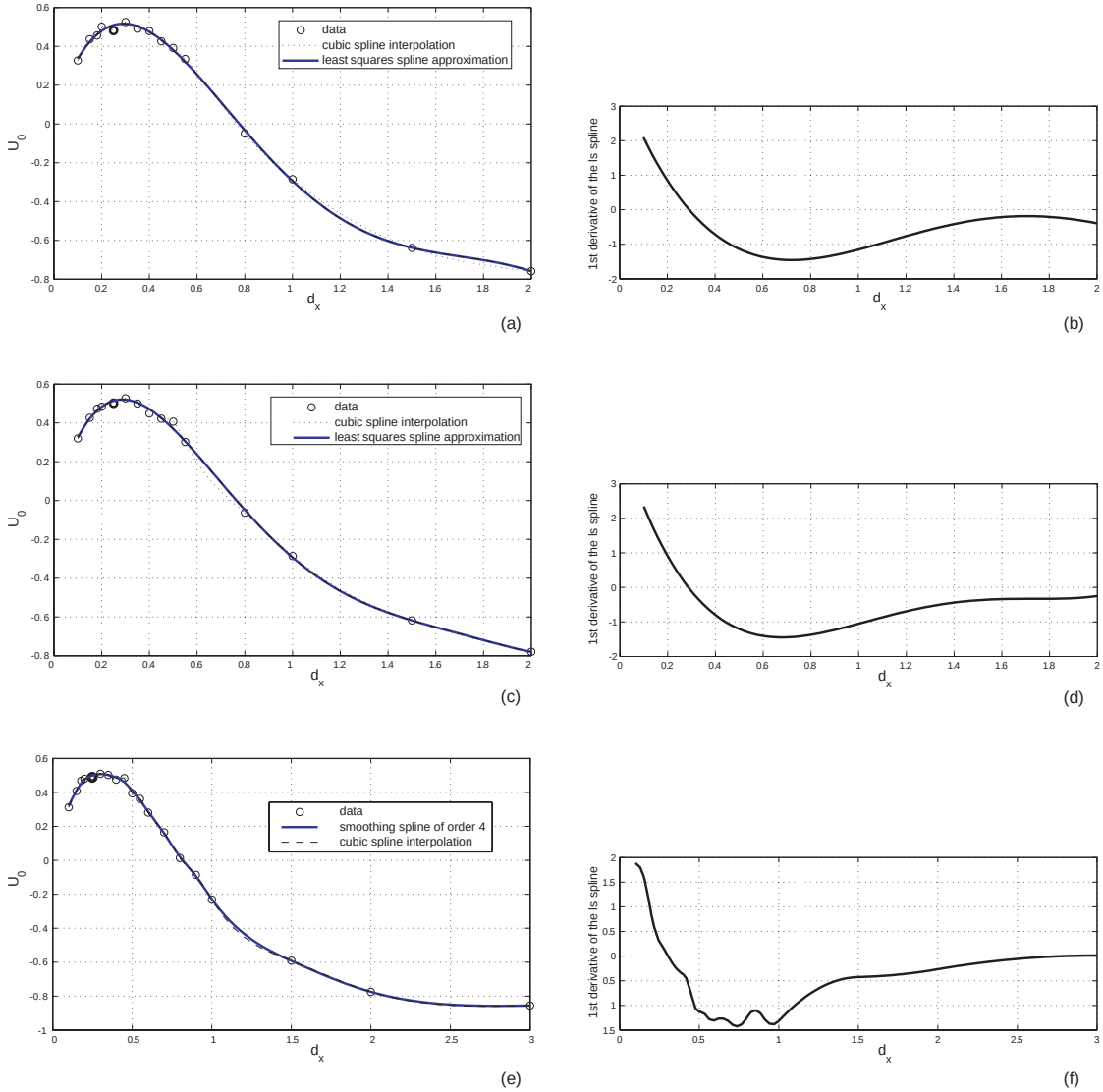


Figure 5.20: Spline approximation of the expected utility (minus entropy of the prior distribution) for the long-range percolation model with exponential edge-probability function and the following 5×5 lattices: (a) triangular ($d_y = d_x$, $\delta_x = d_x/2$), (c) square ($d_y = d_x$, $\delta_x = 0$), and (e) hexagonal ($d_y = \sqrt{3}d_x/2$). The plots (b), (d), (f) in the right panel depict the first derivative of the corresponding approximation spline. The edge profile decay is of the form $p(d) = e^{-\theta d}$ and the prior distribution for θ was taken to be Gamma$(10, 0.2)$.

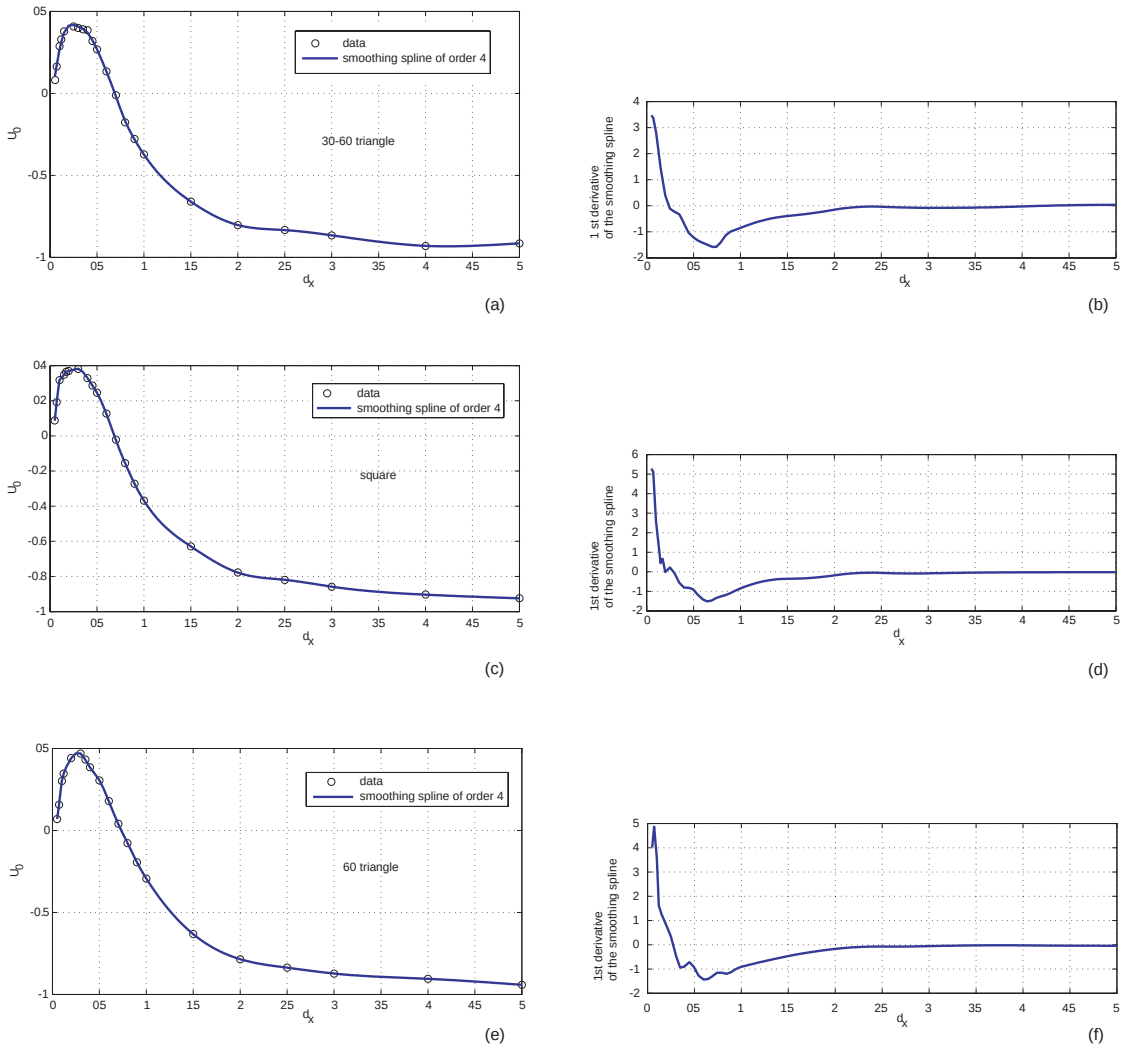


Figure 5.21: Spline approximation of the expected utility (minus entropy of the prior distribution) for the long-range percolation model with Cauchy edge-probability function and the following 5×5 lattices: (a) triangular ($d_y = d_x$, $\delta_x = d_x/2$), (c) square ($d_y = d_x$, $\delta_x = 0$), and (e) hexagonal ($d_y = \sqrt{3}d_x/2$). The plots (b), (d), (f) in the right panel depict the first derivative of the corresponding approximation spline. The edge profile decay is of the form $p(d) = (1 + \theta d^2)^{-1}$ and the prior distribution for θ was taken to be Gamma$(10, 0.2)$.

Chapter 6

Grid Approximation of a Finite Set of Points

6.1 Formulation of the problem

6.1.1 Basic examples

Consider a set of points X on the real line $\mathbb{R}$. Let $d_{\max}$ be the maximum of distances from each point of X to the nearest point of a uniform grid of points from the same axis. For each spacing of such a grid there exists an optimal shift of it which minimises $d_{\max}$. A typical plot showing the dependence of the minimal $d_{\max}$ on the grid spacing in the case when X contains 3 or 4 points is shown in Figure 6.1. The more elements X contains, the less the plot is cluttered and the more points of the plot lie closer to the straight line with the slope $1/2$ and the null intercept[1].

If each point of X is approximated by the closest node of a grid of a certain spacing[2], then there is a certain flexibility in choosing the spacing of the grid: for example, if the minimal $d_{\max}$ should not exceed 0.25, then the grid spacing 2 is as good as 0.5 or any smaller value, or if $d_{\max}$ should not exceed 1, then the grid

[1] No points can lie above this line since the distance from any data point from X to the nearest node of the grid does not exceed half of the grid spacing.

[2] This operation consisting in replacing each point of X by the nearest grid point is sometimes called 'rounding' or 'snapping' in computational geometry.

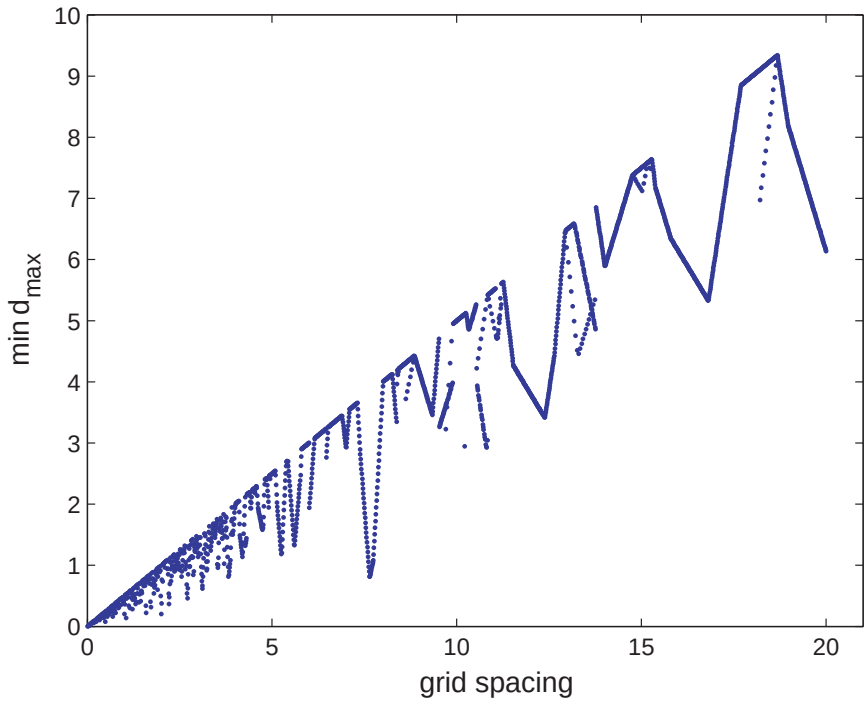


Figure 6.1: Typical dependence of minimal $d_{\max}$ on the spacing of the grid for a set X from $\mathbb{R}$ containing 3 or 4 points. In this particular example $X = \{11.8998, 34.0386, 49.8364, 95.9744\}$. Notice, that what is shown is a single graph of such a dependence; this graph exhibits discontinuities at many values of the grid spacing.

spacings from the range (7.62,7.71) will be as good as any spacing less than 2 (Figure 6.1). In some applications it might be necessary to minimise the number of the grid nodes that fall within the approximation region[3] while satisfying the approximation error, or to minimise the total number of approximating grid nodes (which may well be less than the number of points in X).

Consider another example. Take the following planar configuration of 6 points defined by their Cartesian coordinates

$$X = \{(-0.1553, 6.3511), (-1.4809, 7.9482), (1.2534, 6.2070)$$

$$(0.7213, 8.4480), (0.8821, 9.8563), (4.1944, 7.4268)\}.$$

These points can be approximated by vertices of the coordinate grid $\mathbb{Z}^2$ in which their coordinates are given, for example, by rounding their coordinates to the nearest integer (see Figure 6.2). Such approximation can be characterised by the largest among all deviations of the configuration points from the nearest nodes of the grid. For instance, in the considered example such distance is equal 0.5276. This approximation error may be too large for certain needs and one may want to consider approximations by more refined grids by simply rescaling the initial coordinate grid, i.e. considering the grid $h\mathbb{Z}^2$, $h < 1$, or even $h\mathbb{Z} \times l\mathbb{Z}$, $h, l < 1$. For instance, the largest among all deviations of the configuration points from X from the nearest nodes of the grid $0.2\mathbb{Z}$ is 0.1074, Figure 6.3. Clearly, the approximation error cannot exceed $\frac{1}{2}\sqrt{h^2 + l^2}$ when a grid $h\mathbb{Z} \times l\mathbb{Z}$ is used, and one can always decrease h and l so that this quantity does not exceed the 'target' approximation error.

This is not the best strategy for looking for good approximation grids, especially when there should be a trade-off between the spacing of the grid and approximation error: the latter is often desired to be as small as possible, whereas the former may be required to be not too small—to reduce the number of nodes within the approximation area and to obtain therefore not too dense grid, for example.

Figure 6.4 shows the configuration X placed in the new coordinate system, ob-

[3] i.e., to maximise the grid spacing

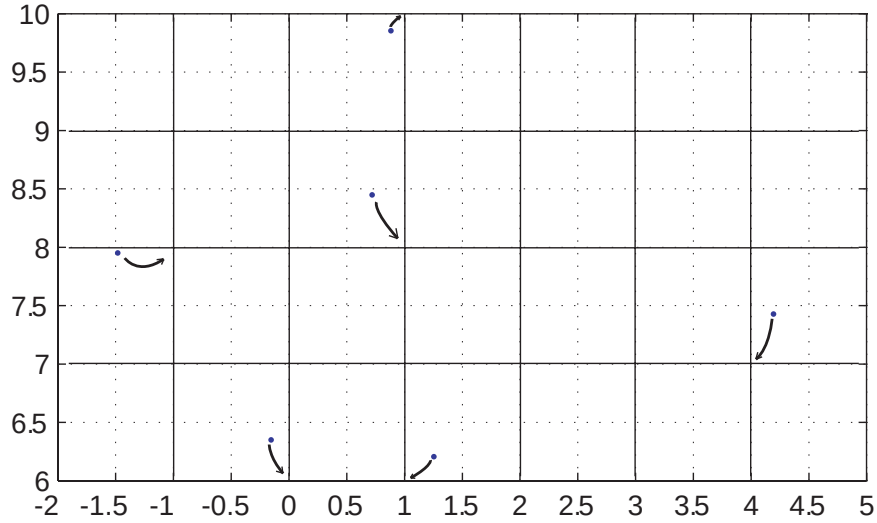


Figure 6.2: Initial configuration X of six points in plane and its approximation by the vertices of the coordinate grid $\mathbb{Z}^2$. The largest 'approximating' distance is 0.5276. Arrows indicate the vertices of the grid which approximate elements of X.

tained after rotating the initial coordinate axes at the angle 0.7. If we approximate the points from X by the nodes of the new grid with (with the unity spacing!) the approximation error becomes 0.0737, and it is smaller than the approximation error resulting from the use of the square grid with the spacing 0.2, coordinated with the axes in which the coordinates of the points from X are initially given.

6.1.2 Formulation of the problem and motivation

We consider the problem of approximating a finite set of points $X \subset \mathbb{R}^n$ by a *uniform grid*

$$G(\alpha, E) = \{\sum_{i=1}^{n} \alpha z_i \mathbf{e}_i \,|\, (z_1, \ldots, z_n) \in \mathbb{Z}^n\}, \tag{6.1}$$

where $E = \{\mathbf{e}_i\}_{i=1}^n$ is an orthonormal basis of $\mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Each point of X is approximated by the nearest node of $G(\alpha, E)$. Let the (Euclidean) distance between $x_k \in X$ and its nearest node be $\delta_k(G)$, $k = 1, \ldots, |X|$, and define the 'distance' from X to G to be

$$\delta(X, G) = \max_{i=1,\ldots,m} \delta_i(G). \tag{6.2}$$

For clarity we will call $\delta(X, G)$ *δ-deviance*.

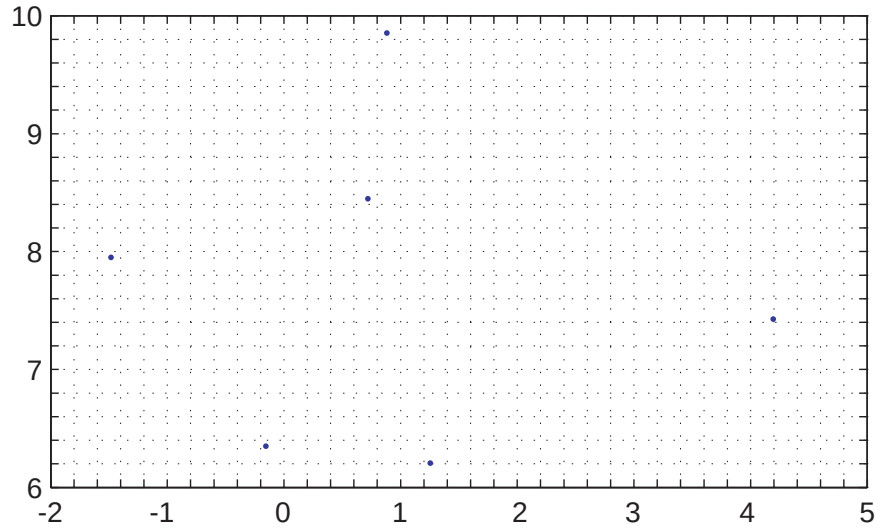


Figure 6.3: Configuration X of six points in plane from Figure 6.2: the square grid spacing h is 0.2. The largest 'approximating' distance by this grid is 0.1074.

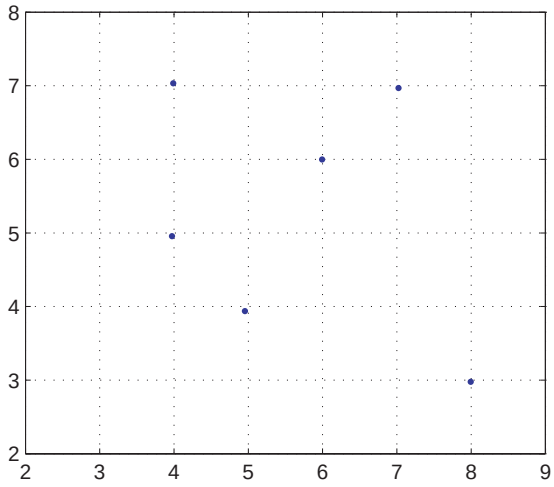


Figure 6.4: Configuration X from Figure 6.2 in new axes after rotating the coordinate system clockwise at the angle $\theta = 0.7$. The maximal approximating distance is 0.0737.

For any given $\epsilon > 0$ and basis E let $\mathcal{A}_\epsilon(E)$ be the set of all α such that

$$\delta(X, G(\alpha, E)) < \epsilon. \tag{6.3}$$

The *ϵ-optimal approximation grid* $G_\epsilon(\alpha^*, E^*)$ is such a grid that

$$\alpha^* = \max_E \max_\alpha \mathcal{A}_\epsilon(E) \tag{6.4}$$

and

$$E^* = \arg\max_E \max_\alpha \mathcal{A}_\epsilon(E). \tag{6.5}$$

In this problem E is the orientation of the grid and α is its spacing. The motivation behind the considered problem is in obtaining approximation grids which are not too dense—once such a grid is obtained a more dense grid can always be constructed, for instance, by consequently halving the spacing of the grid.

In the next section we will review and use the optimisation results of Brucker and Meyer (1987) to provide a numerical recipe of finding a good candidate for the ϵ-optimal approximation grid $G_\epsilon(\alpha^*, E^*)$. Furthermore, the approximation of Brucker and Meyer is adapted to solve the problem of approximation of a finite set of points in $\mathbb{R}^n$ by the nodes of a truly rectangular grid

$$G(\boldsymbol{\alpha}, E) = \{\sum_{i=1}^{n} \alpha_i z_i \mathbf{e}_i \,|\, (z_1, \ldots, z_n) \in \mathbb{Z}^n\}, \tag{6.6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. The complexities of the suggested procedures are also discussed.

6.2 Finding ϵ-optimal approximation grids

6.2.1 Brucker–Meyer approximation in $\mathbb{R}$ and $\mathbb{R}^n$

Brucker–Meyer n-dimensional approximation problem

Brucker and Meyer (1987) in a short communication note proposed a procedure for translating a rectangular grid in such a way that a finite set of points in $\mathbb{R}^n$ is approximated as well as possible by points of translated grid. The closeness between a grid and a set of points in Brucker and Meyer's approximation was

measured by the maximal deviation to the grid lines from the configuration points. Let us review their main results starting with preliminary definitions.

Let $E = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be an orthonormal basis in $\mathbb{R}^n$ and $\alpha_1, \ldots, \alpha_n$ positive real numbers. Consider the following rectangular grid

$$G(\boldsymbol{\alpha}, E) = \{\sum_{i=1}^{n} \alpha_i z_i \mathbf{e}_i \,|\, (z_1, \ldots, z_n) \in \mathbb{Z}^n\}. \tag{6.7}$$

The set $G(\boldsymbol{\alpha}, E) + \boldsymbol{\beta}$ is obtained by adding a vector $\boldsymbol{\beta} \in \mathbb{R}^n$ to all points of the grid G and is called a *translation* of G.

For an arbitrary $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ let

$$d_i(\mathbf{x}, G) = \min\{\|x_i - \alpha_i z\| \;|\; z \in \mathbb{Z}\} \tag{6.8}$$

and

$$d(\mathbf{x}, G) = \max_{i=1,\ldots,n} d_i(\mathbf{x}, G). \tag{6.9}$$

For a finite set of points $X \subset \mathbb{R}^n$ define

$$d(X, G) = \max\{d(x, G) \,|\, \mathbf{x} \in X\}, \tag{6.10}$$

and call this quantity *d-deviance* of X from G (compare to the notion of δ-deviance defined by (6.2) in § 6.1.2).

The following problem was considered by Brucker and Meyer (1987): Given a finite set $X \subset \mathbb{R}^n$ and a grid $G(\boldsymbol{\alpha}, E) \subset \mathbb{R}^n$, find a translation G^* of G such that the d-deviance $d(X, G^*)$ is minimal. We will refer to this problem as the *Brucker–Meyer approximation problem*.

Brucker–Meyer algorithm for the one-dimensional problem

Since the multidimensional Brucker–Meyer problem can be reduced to the one-dimensional case let us first consider the solution of this approximation problem in $\mathbb{R}$ suggested by Brucker and Meyer (1987).

Without loss of generality one can consider that all points of X are non-negative and that $0 \in X$. Let U be the smallest number in the one-dimensional grid $G = \{\alpha z \,|\, z \in \mathbb{Z}\}$ which is greater or equal to all $x \in X$. Then the optimisation problem on the line $\mathbb{R}$ can be reduced to an optimisation problem on a circle of

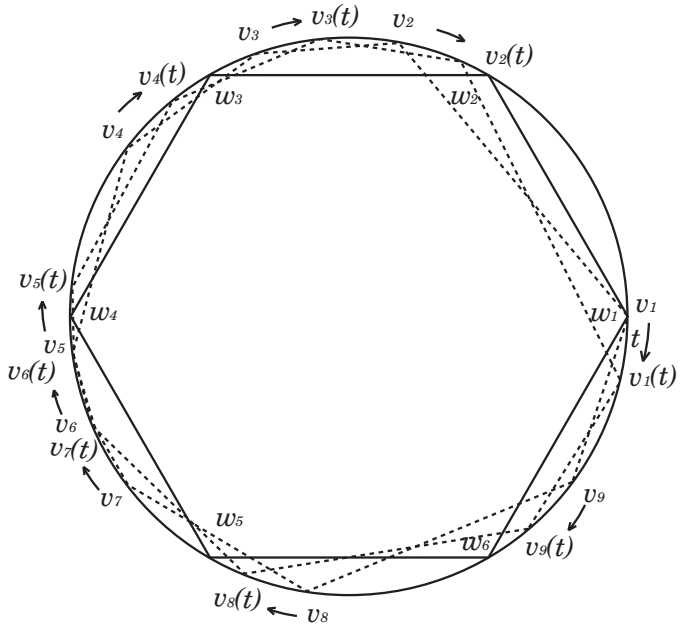


Figure 6.5: Representation of X and G on a circle in the Brucker–Meyer univariate approximation problem and translation of the grid (polygon with solid edges) realised via translation of the points from X (polygons with dotted edges).

circumference U. Consider such a circle C and represent the grid G by the vertices of a regular polygon A with $k = \frac{U}{\alpha}$ vertices $w_1, \ldots, w_n$ on the circle line. The set X will be represented by a (generally irregular) polygon B with $m := |X|$ vertices $v_1, \ldots, v_m$ on the circle. A translation of the grid will correspond then with a movement of the polygon A relative to the polygon B. If we choose a clockwise rotation then a translation can be described by the clockwise distance t from w_1 to v_1 along C. For a fixed t let $d_i(t)$ to be the distance between v_i and its nearest neighbour w_j on the circle line. Then the problem is equivalent to finding such t that

$$f(t) = \max_{i=1,\ldots,m} d_i(t) \tag{6.11}$$

is minimised. Figure 6.5 depicts an example of the described construction.

The advantage of this construction is in the following obvious property of $f(t)$: it is a periodic function with period α. Therefore it suffices to minimise $f(t)$ on the interval $[0, \alpha]$ only.

The idea of Brucker and Meyer for solving this particular problem was in partitioning $[0, \alpha]$ into $2m$ intervals $I_1 = [0, t_1]$, $I_2 = [t_1, t_2]$,..., $I_{2m} = [t_{2m-1}, \alpha]$ such

that the minimum of $f(t)$ in I_j $(j = 1, \ldots, 2m)$ can be calculated easily. Specifically, the interval I_j is chosen so that within this interval no B-vertex passes any A-vertex or the midpoint of two A-vertices w_j and w_{j+1}—clearly, such construction is always possible. The solution of the problem is obtained then by taking the minimum of $f(t)$ on each of these subintervals.

The values $t_1, \ldots, t_{2m-1}$ defining the intervals $I_1, \ldots, I_{2m}$ can be calculated iteratively. To present the corresponding construction Brucker and Meyer used the following notation.

For a fixed t define the following sets

$$R(t) = \{i \,|\, d_i(t) \text{ is a } BA\text{-distance}\},$$

$$L(t) = \{i \,|\, d_i(t) \text{ is a } AB\text{-distance}\},$$

that is if while going clockwise around the circle line the distances $d_i(t)$, $i \in R(t)$ $(i \in L(t))$ are distances from B-vertices (A-vertices) to A-vertices (B-vertices). If, however, v_i is a midpoint of an arc with two A-vertices w_j and w_{j+1} as its endpoints, it will be assumed that i is in $R(t)$. Finally, let

$$r(t) := \max\{d_i(t) \,|\, i \in R(t)\},$$

$$l(t) := \max\{d_i(t) \,|\, i \in L(t)\}.$$

To calculate the t_j values we start with $t = 0$ and let

$$\Delta_1 t := \min\{d_i(t) \,|\, i \in R(t)\},$$

$$\Delta_2 t := \min\{\alpha/2 - d_i(t) \,|\, i \in L(t)\} = \alpha/2 - l(t),$$

$$\Delta t : = \min\{\Delta_1 t, \Delta_2 t\}..$$

Then set $\tilde{t} := t + \Delta t$. Replacing t by $\tilde{t}$ this process is repeated yielding a sequence of the endpoints of the intervals $I_1, \ldots, I_{2m}$.

To calculate the minimum $f(t^*)$ of $f(\cdot)$ in $[t, t + \Delta]$ we have to consider two cases.

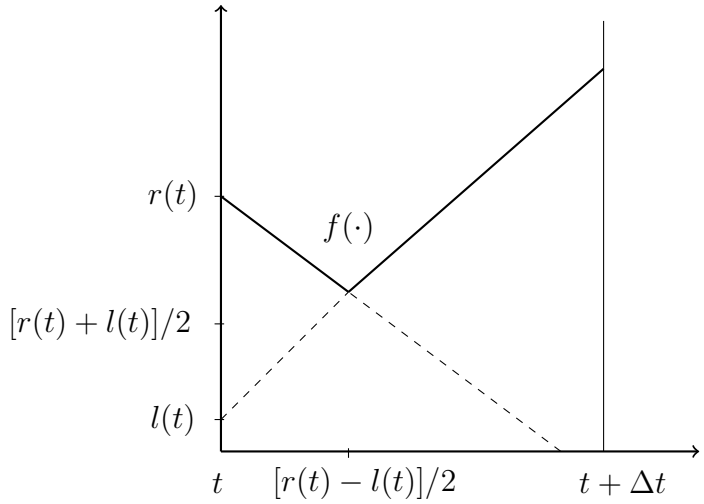


Figure 6.6: Function $f(\cdot)$ on the interval $[t, t+\Delta t]$ when $\Delta t > [r(t) - l(t)]/2$.

1 $r(t) < l(t)$

Then $f(t)$ is strictly increasing in $[t, t+\Delta t]$. Thus, $t^* = t$ and $f(t^*) = l(t)$.

2 $l(t) \leq r(t)$

If $\Delta t > [r(t) - l(t)]/2$ then the graph of $f(\cdot)$ in $[t, t+\Delta t]$ is as shown in Figure 6.6. Thus, $t^* = t + [r(t) - l(t)]/2$. If, however, $\Delta \leq [r(t) - l(t)]/2$ then $t^* = t + \Delta t$. In both cases $f(t^*) = r(t^*)$.

When a B-vertex v_i is passing an A-vertex w_j or the midpoint between two A-vertices w_j and w_{j+1}, we have to change the index sets $R(t)$ and $L(t)$. If v_i passes an A-vertex we have to eliminate the index i from $R(t)$ and to insert i into $L(t)$. If v_i passes the midpoint of two A-vertices we have to eliminate the index i from $L(t)$ and to insert i into $R(t)$.

Algorithm 3 is a slightly corrected and working version of the one suggested by Brucker and Meyer (1987) which finds the solution for the one-dimensional grid approximation problem. The correction has been made in the **while**-loop to check whether the index set $R(t)$ is non-empty so that the original Brucker–Meyer algorithm avoids endless loops and terminates correctly. The algorithm requires the set of deviations $d_1(0), \ldots, d_m(0)$ of the data points from the initial grid, and finds

$$t_{\text{opt}} := \arg\max f(t),$$

Algorithm 3 Brucker–Meyer algorithm

Require: $d_1(0), \ldots, d_m(0)$.

Ensure: t_{opt}, f_{opt}.

$t := 0$; $f_{\text{opt}} := \infty$; $t_{\text{opt}} := 0$;
for $i := 1$ to m **do**
  $d_i := d_i(0)$;
R:=R(0); L:=L(0);
while $t \leq \alpha$ AND $R \neq \emptyset$ **do**
  $r := \max\{d_i \,|\, i \in R\}$; $l := \max\{d_i \,|\, i \in L\}$;
  $\Delta_1 := \min\{d_i \,|\, i \in R\}$; $\Delta_2 := \alpha/2 - l$; $\Delta := \min\{\Delta_1, \Delta_2\}$;
  if $r < l$ **then**
    $t^* := t$; $f(t^*) := l$;
  else
    if $\Delta > (r - l/2)$ **then**
      $\theta := (r - l)/2$
    else
      $\theta := \Delta$; $t^* := t + \theta$; $f(t^*) := r - \theta$
  if $f(t^*) < f_{\text{opt}}$ **then**
    $t_{\text{opt}} := t^*$; $f_{\text{opt}} := f(t^*)$;
  for all $i \in L$ **do**
    $d_i := d_i + \Delta$;
  for all $i \in R$ **do**
    $d_i := d_i - \Delta$;
  $t := t + \Delta$;
  for all $i \in L$ **do**
    if $d_i = \alpha/2$ **then**
      $L := L \setminus \{i\}$; $R := R \cup \{i\}$;
  for all $i \in R$ **do**
    if $d_i = 0$ **then**
      $R := R \setminus \{i\}$; $L := L \cup \{i\}$;
return t_{opt}, f_{opt}

where $f(t)$ is defined by (6.11), and $f_{\text{opt}} := f(t_{\text{opt}})$. The complexity of this algorithm is $O(m^2)$, where $m = |X|$. However, as Brucker and Meyer (1987) note, the complexity can be reduced to $O(m \log m)$ by using appropriate data structure that would permit treatment the sets L and R together with the operations of deletion, insertion, and finding the maximal and minimal elements as priority queues.

Brucker–Meyer approximation in the multidimensional case

Since the objective function in the Brucker and Meyer n-dimensional approximation problem can be written as

$$d(X, G) = \max_{x \in X} \max_{i=1,\ldots,n} d_i(x, G) = \max_{i=1,\ldots,n} \max_{x \in X} d_i(x, G), \tag{6.12}$$

this d-deviance can be maximised by maximising d-deviances along each dimension separately and independently (Brucker and Meyer (1987)). For each dimension $i = 1, \ldots, n$ one can solve the problem of finding a one dimension optimal shift t_i^* of the coordinate system as described above. It follows from (6.12) that the shift given by vector $\boldsymbol{\beta}^* = (t_1^*, \ldots, t_n^*)$ is optimal for the n-dimensional problem.

Assuming the optimal choice of the data representation the complexity of the Brucker–Meyer approximation in the multidimensional situation is $O(nm \log m)$. In general, the optimal solution to the Brucker–Meyer approximation problem is not unique.

Realisation

Algorithm 3 has been implemented by the author in a MATLAB function called `unidimapprox`. This function, allowing one to apply the Brucker–Meyer algorithm in one dimension, has the following syntax:

```
function [d x3 t_opt]=unidimapprox(x,scale,to_plot)
```

Here `d` contains the deviances $d_i(x, G^*)$ from data points `x` to the optimal grid G^* and spacing `scale`. The optimal grid is described by the coordinates `x3` and the optimal shift `t_opt`. One should set `to_plot` to 1 to obtain a graphical output.

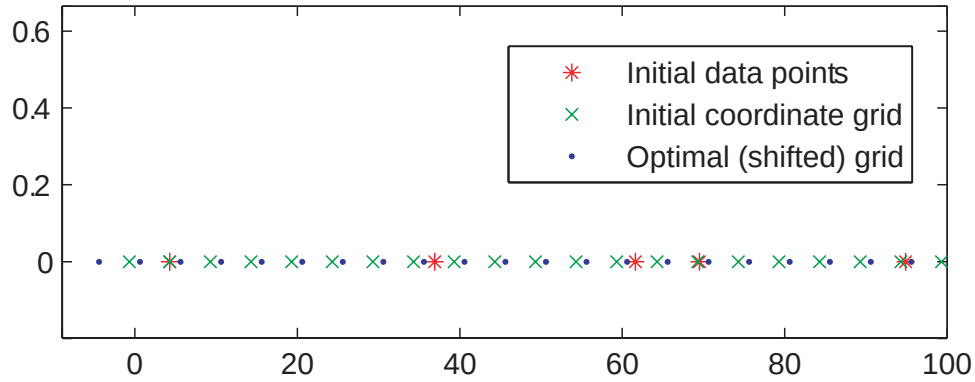


Figure 6.7: Application of the Brucker–Meyer algorithm in the one-dimensional case: initial and optimal grid with spacing $\alpha = 5$ for the set X of 5 points drawn uniformly and independently on [1,100].

Example 6.2.1. *Figure 6.7 illustrates application of the Brucker–Meyer algorithm and contains a graphical output as a result of the execution of the following MATLAB code:*

```
>> x = unifrnd(0,100,1,5);
>> alpha = 5; show_plot = 1;
>> unidimapprox(x,alpha,show_plot);
```

The function `unidimapprox.mat` *can be found on the WWW [1] (folder* `../BM_approximation`*).*

6.2.2 Approximation by grid nodes

Note that the notions of δ-deviance, introduced in § 6.1.2, and that of d-deviance, introduced in § 6.2.1, coincide in the one-dimensional case. In the multidimensional case δ-deviance and d-deviance represent different quantities. However, a generalisation of the Brucker–Meyer algorithm for the multidimensional case can be used to obtain an approximation to the solution to the ϵ-optimal approximation grid problem (6.3)-(6.5).

Consider first a uniform grid $G(\alpha, E)$. Since E is an orthonormal basis in $\mathbb{R}^n$, it can be represented by an angle θ at which a fixed basis[4] E_0 should be rotated

[4]For example, it may be the canonical basis $E_0 = \{(1,0,\ldots,0),\ldots,(0,\ldots,0,1)\}$, if E and E_0, viewed as coordinate systems, have the same orientation.

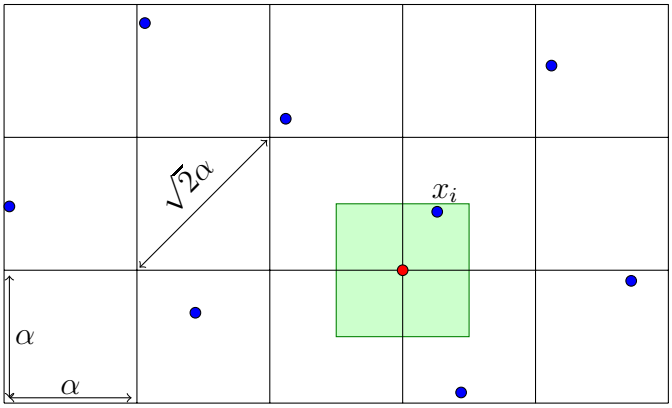


Figure 6.8: Approximation of a finite set of points by the nodes of a square grid with the spacing α.

in order to obtain E. For a given configuration X and 'orientation' θ of the grid G with spacing α the δ-deviances $\delta_i(X, G)$ are bounded from above as follows (Figure 6.8):

$$\delta_i(G) \leq \frac{\sqrt{n}}{2}\alpha, \ i = 1, \ldots, n. \tag{6.13}$$

It follows that

$$\delta(X, G) = \max_{i=1,\ldots,n} \delta_i(X, G) \leq \frac{\sqrt{n}}{2}\alpha,$$

and attributing any non-negative value which does not exceed $\frac{2\epsilon}{\sqrt{n}}$ to α ensures that the corresponding grid $G(\alpha, E)$ is an 'admissible solution' to the ϵ-optimal grid approximation problem.

Algorithm 4 represents a procedure of finding approximations to α^* and E^* (through its orientation angle θ^*) for the ϵ-optimal approximation uniform grid problem (6.3)-(6.5) using the multidimensional Brucker–Meyer algorithm. The idea behind Algorithm 4 is to consequently apply the Brucker–Meyer algorithm to a given point configuration when the grid is rotated by a small incremental angle along a chosen direction and then choose the best grid approximation. The description of Algorithm 4 can be found on p. 156.

Algorithm 4 Approximate solution to the ϵ-optimal approximation grid problem

Require: X (contains m n-dimensional points), ϵ, h_α, h_θ, $\alpha_{\max}$.

Ensure: $\tilde{\alpha}_{\text{opt}}$, $\tilde{\theta}_{\text{opt}}$.

1: $j := 1$; $X_0 := X$; $\tilde{\alpha}_{\text{opt}} := 0$;

2: **while** $(j-1)h_\theta \leq \pi$ **do**

3: $X := \text{rotate}(X_0, (j-1)h_\theta)$; % choose a rotation direction and rotate each element of the initial data matrix at the angle jh_θ along the chosen direction assuming that X has the following form:
$X = \{(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}), \ldots, (x_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)})\}$.

4: $\alpha_0 := \min(h_\alpha, 2\epsilon/\sqrt{n})$; $\alpha := \alpha_0$; $i = 1$;

5: **while** $\alpha \leq \alpha_{\max}$ **do**

6: **for** $k := 1$ to n **do**

7: $[dx_k\ f\ t_{\text{opt}}(k)] := unidimapprox((x_k^{(1)}, \ldots, x_k^{(m)}), \alpha, 0)$;
% the MATLAB function `unidimapprox` implements the Brucker–Meyer algorithm, see Algorithm 3 and the description of this function on p. 153;
% at this moment dx_k is a vector of m d-deviances along the k^{th} coordinate: $dx_k = (dx_k^{(1)}, \ldots, dx_k^{(m)})$;
% the next command has the following syntax (MATLAB version): `[a,b]:=max(x)`, where `a` is the value of the maximal element in `x`, and `b` is the index of this element (if there is more than one such element, than `b` is the smallest index);

8: $[dmax, f] :=$

$$\max\left(\sqrt{\left(dx_1^{(1)}\right)^2 + \ldots + \left(dx_n^{(1)}\right)^2}, \ldots, \sqrt{\left(dx_1^{(m)}\right)^2 + \ldots + \left(dx_n^{(m)}\right)^2}\right);$$

9: **if** $dmax < \epsilon$ AND $i > \tilde{\alpha}_{\text{opt}}/h_\alpha$ **then**

10: $\tilde{\alpha}_{\text{opt}} := ih_\alpha$; $\tilde{\theta}_{\text{opt}} := jh_\theta$;

11: $\alpha := \alpha_0 + ih_\alpha$; $i := i + 1$;

12: $j := j + 1$;

13: **return** $\tilde{\alpha}_{\text{opt}}$, $\tilde{\theta}_{\text{opt}}$;

Realisation

Algorithm 4 has been implemented by the author in a MATLAB function called `optimal_plot`. This function has the following syntax:

```
function [theta_optimal dx_optimal dy_optimal ...
          scale_optimal d_max txopt tyopt]=...
          optimal_grid(points,epsilon,scale_max,...
                       scale_search_step,theta_search_step)
```

The correspondence between the arguments of the function `optimal_plot` and the input arguments of Algorithm 4 is as in Table 6.1.

arguments of `optimal_plot`	arguments of Algorithm 4
`points`	X
`epsilon`	ϵ
`scale_max`	$\alpha_{\max}$
`scale_search_step`	h_α
`theta_search_step`	h_θ

Table 6.1: Arguments of the MATLAB function `optimal_plot` (p. 157) and the input arguments of Algorithm 4 (p. 156).

The function `optimal_grid` can be found on the WWW [1] (folder `../BM_approximation`).

Example 6.2.2. *Consider the following configuration of 10 points*

$$X = \{(6.9523, 1.0399), (0.6842, 7.5864), (1.1252, 7.5654),$$

$$(0.3451, 8.1266), (9.1066, 5.9876), (6.6316, 5.4967),$$

$$(1.0285, 6.6941), (2.0599, 6.6352), (6.0404, 7.5027), (4.7034, 8.3053)\}$$

taken at random from the square $\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 10, 0 \leq y \leq 10\}$ *using the following MATLAB code:*

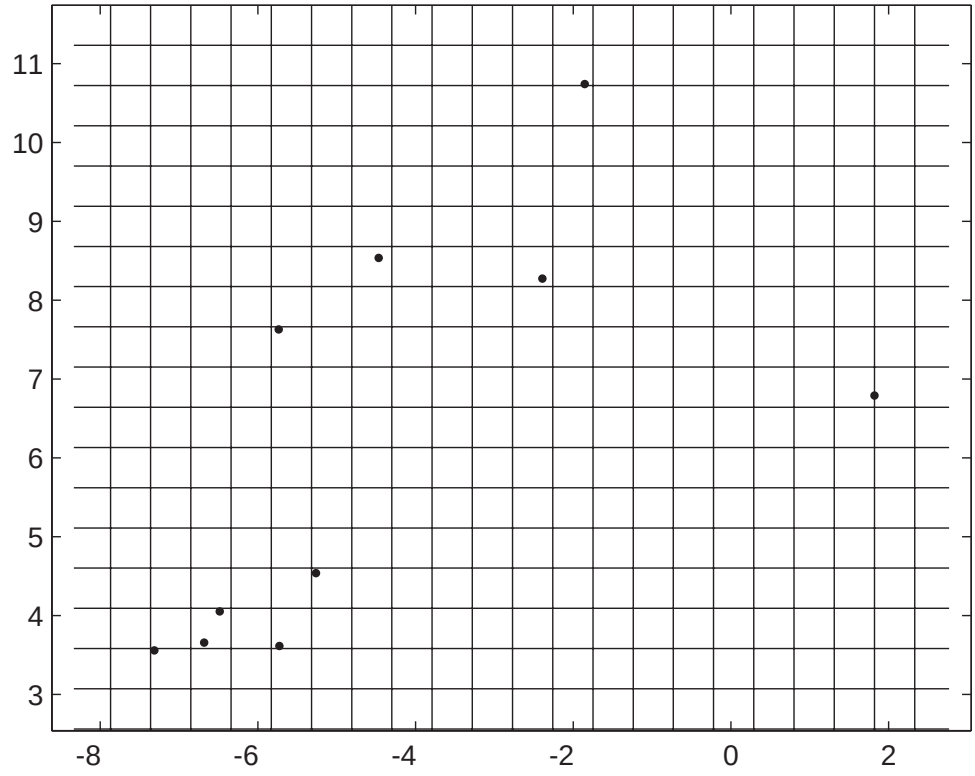


Figure 6.9: Approximation of a finite set of points X in plane by a uniform grid from Example 6.2.2.

```
>> points = zeros(10,2); a = 0; b = 10;
>> points(:,1) = a + (b-a) * rand(10,1);
>> points(:,2) = a + (b-a) * rand(10,1);
```

Algorithm 4 can be applied to the set X by using the described above MATLAB function `optimal_plot`*:*

```
>> scale_max = 2; scale_search_step = 0.01; theta_search_step = 0.01;
>> epsilon = 0.2;
>> [theta_optimal dx_optimal dy_optimal scale_optimal d_max] = ...
   optimal_grid(points,epsion,scale_max,scale_search_step, ...
   theta_search_step)
```

In the new coordinate system, which is obtained by rotation of the initial axes at the angle $\tilde{\theta}_{opt} = 1.16$, the points of X have the following coordinates (Figure 6.9):

$$X = \{(1.8229, 6.7891), (-6.6819, 3.6568), (-6.4867, 4.0528),$$

$$(-7.3127, 3.5617), (-1.8528, 10.7401), (-2.3912, 8.2749),$$

$$(-5.7265, 3.6162), (-5.2606, 4.5383), (-4.4663, 8.5340), (-5.7361, 7.6287)\}.$$

The optimal grid $\tilde{G}^*$ is obtained in this new coordinate system as follows:

$$\tilde{G}^* = \tilde{\alpha}_{opt}(\mathbf{x}_{min} - \mathbb{Z}^2) - \mathbf{t}_{opt},$$

where the rescaling coefficient is $\alpha_{opt} = 0.51$, the optimal shift vector is

$$\mathbf{t}_{opt} = (0.0464, 0.4901),$$

and $\mathbf{x}_{min}$ is calculated as follows:

$$\mathbf{x}_{min} := (\min\{x \,|\, (x, y) \in X\}, \min\{y \,|\, (x, y) \in X\}).$$

In this example the approximation error ϵ was taken to be 0.2 *and the realised δ-deviance is* 0.1911*.*

Additional requirements to the ϵ-optimal grid approximation problem (6.3)-(6.5) can be added. For instance, one may demand that no more than one point of the initial point configuration should be approximated by a node of the optimal grid. Algorithm 4 can still be used in this case with corresponding minor modifications.

It is also straightforward to make corresponding changes in Algorithm 4 in order to obtain approximate solutions to the problem of approximating point configurations by truly rectangular grids $G(\boldsymbol{\alpha}, E)$ defined by (6.7) (see § 6.1.2). In this case, the grids that are checked for approximation optimality should be rescaled independently in each dimension of the vector α.

Our final remarks refer to the complexity of Algorithm 4, its possible modifications and improvement. Since the complexity of the Brucker-Meyer approximation algorithm is $O(nm \log m)$, it follows from the structure of Algorithm 4 that its complexity is $O(\frac{\alpha_{\max}}{h_\alpha} \frac{\pi}{h_\theta} nm \log m)$. The complexity of Algorithm 4 adapted for the

search of a rectangular optimal approximation grid is therefore $O\left(\left(\frac{\alpha_{\max}}{h_\alpha}\right)^n \frac{\pi n m}{h_\theta} \log m\right)$. One could slightly improve on the structure and performance of Algorithm 4 for some particular cases by changing its structure so that the iterations for α for any given θ run from $\alpha_{\max}$ downwards to α_0 and stopping the iterations for θ once a corresponding δ-deviance does not exceed ϵ. Notice, however, that this does not generally reduce the complexity of the algorithm.

6.2.3 Applications

Besides natural applications in interpolation and solution methods for partial differential equations, when calculations are to be done on a grid and the initial values have to be approximated by the nodes of a grid, the discussed algorithms can be used in the analysis of spatial patterns and in optimal design problems for spatial interaction models. $\epsilon-$optimal approximation grids can also be used in approximation of optimal space-filling designs by the grid nodes.

The ϵ-optimal approximation grid problem is essentially a computational geometry problem. An ϵ-optimal grid can be viewed as a *core set* for the original point set X. The notion of a core set is a general notion emerged from papers of Barequet and Har-Peled (2001), Agarwal, Har-Peled and Varadarajan (2003) and others in high dimensional computation geometry and proved to be helpful in obtaining approximations to optimisation problems on point data sets (Chan (2006)). The core set framework can be described in very general lines as follows: suppose that a geometric optimisation problem on a set of points X of size m (such as finding the diameter of a set, for example) is to be solved and no fast algorithm is known for this. One transforms the set X into a set X' in an attempt to achieve the following: (i) the cardinality of X' is small, and (ii) the solution of the same optimisation problem for X' represents a 'good' approximation of the solution to the original problem involving X. If one can find quickly a core set of a small size, then the approximate solution can be obtained faster than the solution to the original problem, even though one may have to apply a 'brute force' algorithm to the core set found. The reader is referred to Chan (2006), Agarwal, Har-Peled and Varadarajan (2005), and Bijay and Vigneron (2005) for further details.

Chapter 7

Conclusions

The problem of optimal experimental design for random graph models has been formulated and studied in this thesis. Here we give a summary of our work, highlight its key contributions, and discuss potential directions for future research.

7.1 Summary

In Chapter 1 we presented motivation behind studying inference and optimal design problems on random graphs. We gave the description of a general random graph model in which nodes are fixed but connections between them are random, establishing according to the so called edge-probability function. This function depends monotonically on the weight of the possible edge, or, indeed, distance between two nodes in the case of a metric space. We also assumed that the edge-probability function is parametrised by a statistical parameter and identified the statistical interest in considering the described model as a wish to be able to make inference on the model parameter(s). The optimal design problem consists then in finding an optimal arrangement of the graph nodes within some predefined region. The chapter contains a selective and brief review of related work on inference and optimal experimental design for, mainly non-linear, spatial response and stochastic interaction models.

Chapter 2 provided a review of some standard mathematical notions from the graph theory and also main elements of statistical likelihood-based and Bayesian

inference techniques. Monte Carlo and Markov Chain Monte Carlo methods were also reviewed in this chapter as the main computational machinery that is subsequently used in the thesis.

Chapter 3 began by considering D-optimal designs for the problem of optimal arrangement of random graph nodes. This was done through some 'toy' examples of graphs with three nodes. We identified the following drawbacks of using informativeness criteria based on the Fisher information in optimal design problem for random graphs: (i) the optimal node arrangement is equidistant; (ii) the design depends on the model parameter's true value; (iii) optimal designs possess symmetries which put a question whether the freedom in choosing the positions of the graph nodes was used efficiently; (iv) the obtained solutions are not invariant under reparametrisation of the model parameter. We therefore turned to a more suitable utility-based Bayesian experimentation paradigm. The Shannon entropy, Kullback–Leibler divergence and Lindley information measure play a particularly important role in this framework. We identified the expected Kullback–Leiber divergence and expected Lindley information gain as expected utilities and thoroughly studied their properties, thus comparing informativeness of experiments. In connection with the problem of expected utility maximisation we gave an alternative proof of the first-order conditions first established by Parmigiani and Berry (1994).

Two different experimental scenarios, progressive design and instructive design, were introduced in Chapter 3. In the former scenario the experimental motivation consists in increasing one's knowledge about the model parameter, whereas in the latter scenario the purpose of the experiment is to instruct someone pursuing an optimal design holding a prior for the model parameter using one's superior knowledge about it. We reviewed simulation-based methods of evaluation of the expected utility based on the Kullback–Leibler divergence for each of these two experimental scenarios in this chapter. Using graphs as an underlying interaction topology as well as model objects and utility-based Bayesian framework we described the studied model in more specific terms and gave a second, more specific formulation of the problem—the n node optimal design problem for random

graphs. The chapter concluded with some illustrative examples.

Chapter 4 delivered theoretic results for some basic random graph models. We first proved a general worst case scenario result for indefinitely growing or diminishing configurations under the progressive design scenario. This intuitively obvious but mathematically not straightforward result says that weights (distances) of the optimal graph edges cannot be all either too large or too small. We were unable to prove a similar theorem in the instructive case, although we strongly believe that it holds in this case as well. We then studied two-node and three-node designs using many examples identifying the following possible crucial features of expected utility surfaces: (i) flatness around modes, and (ii) multimodality. The chapter continued with studying proximity (geometric) graphs and graphs with threshold edge-probability decay. We obtained an explicit solution to the optimal design problem for proximity graphs considered in metric spaces on star topologies—this solution can be represented via quantiles of the prior distribution. We also showed that the case of a threshold edge-probability decay can be relatively easily treated numerically. We concluded the chapter by showing how the obtained theoretic result for proximity graphs can be used to easily show non-preservation of optimal designs under replication (in non-linear models).

Chapter 5 was wholly devoted to inference and experimental design problems for finite clusters from bond percolation on the integer lattice $\mathbb{Z}^d$, $d \in \mathbb{N}$ (and also its modifications), or, equivalently, for SIR epidemics evolving on a bounded or unbounded subset of $\mathbb{Z}^d$ with constant life times[1]. The bond percolation probability p was considered to be unknown. We considered inference under each of the following two scenarios:

- The observations consist of the set of sites which are ever infected, so that the routes along which infections travel are not observed (in terms of the bond percolation process, this corresponds to a knowledge of the connected

[1]Of course, the assumption of constant infectious periods is a very restrictive assumption. However, it is the assumption which allowed us to draw parallels to the unoriented bond percolation process. Generally, the methods used for making inference on the model parameter(s) can be further extended to the Markovian SIR model with variable (or even random) infectious times in the spirit of Demiris (2004) using oriented graphs and connections to oriented percolation

component containing the initially infected site—the location of this site within the component not being relevant to inference for p);

- All that is observed is the size of the set of sites which are ever infected.

We presented MCMC algorithms for making inference on the model parameter in each of these scenarios. We presented a theoretical result stating that the sequence of maximum likelihood estimates for the bond percolation probability converges to the critical probability of the lattice in the case of increasing finite size clusters. We also conjectured that the posterior distribution of the bond percolation parameter 'converges' to the point-mass distribution at the critical probability in this case. These theoretical results have implications of combinatorial nature on the relative number of realisations of the process with a large cluster size. A corresponding combinatorial characterisation is given in the case of the square lattice $\mathbb{Z}^d$.

We introduced inner-outer design plots by 'sparsifying' the underlying lattice and showed that in the case of incomplete observations for percolation models the mostly populated design is not necessarily the most optimal design under either of considered experimental motivations (progressive and instructive scenarios). This has been done using the MCMC algorithms mentioned above. Chapter 5 concluded by a discussion how the obtained results could be generalised to long-range percolation models. We also considered deformations of the square lattice as a way towards identifying a whole class of lattice designs that keep the dimensionality and cardinality of the design space low.

A problem of grid approximation of a finite set of points is formulated in Chapter 6. We introduced ϵ-optimal approximation grids and described a solution to this problem. The practical solution that we suggested here combines fast Brucker–Meyer approximation technique and slow consequent rotations of the grids of optimal scaling. We described the corresponding algorithm and discuss applications of ϵ-optimal approximation grids.

7.2 Contributions of the thesis

We highlight main contributions of this thesis. In this work we

1 formulated the problem of n-node optimal design for weighted random graphs using a utility-based Bayesian statistical framework; the design problem was considered using two different experimentation motivations which we referred to as progressive and instructive designs;

2 gave an alternative proof of the first-order conditions for the expected utility based on the Kullback–Leibler divergence (this result was first proved by Parmigiani and Berry (1994));

3 showed that indefinitely growing or diminishing vertex configurations are asymptotically the worst designs when the Kullback–Leibler divergence is employed as a utility; this was shown for the progressive designs;

4 identified possibility and main features of multimodality of expected utility surfaces in the optimal design problem for random graphs;

5 derived an explicit solution to the problem of optimal design for star-shaped proximity graphs and proximity graphs in metric spaces; showed how to solve the design problem in the case of a threshold edge-probability function numerically;

6 studied inference and optimal design problems for finite clusters from bond percolation on the integer lattice $\mathbb{Z}^2$; introduced inner-outer lattice designs and showed that the mostly populated designs are not necessarily the most optimal designs in the case of incomplete observations under both progressive and instructive design scenarios;

7 formulated the problem of finding ϵ-optimal approximation grids and described a solution to this problem.

7.3 Directions for future work

We identify lacunae of this research and questions to which final (or definite) answers have not yet been found. We also indicate additional areas of potential research on the topic of optimal experimental design for random graphs.

The author believes that the further study of the following questions and problems would complement the results presented in this thesis:

1 ***Characterisation of graphs*** *for which the expected utility is unimodal.* It was shown in § 4.2.3 that the expected utility can be a unimodal function as well as a multimodal function of the design, including the case when the global mode (or modes) $\mathbf{d}^* = (d_1^*, d_2^*, \ldots, d_n^*)$ is not of the form $d_1^* = d_2^* = \ldots = d_n^*$. When does the unimodality appear and what does it depend on? Does the 'steepness' of the edge-probability have an effect? Is it possible to derive a characterisation of graphs (via indicating the type of the probability-edge function for a given utility chosen) for which, for example, the expected utility is a unimodal function? For metric structures (or star topologies) such a result would clearly facilitate the search of optimal designs, as it would allow one to rule out the corresponding type of the optimal design before choosing the optimisation strategy.

2 ***Worst case scenarios*** *for indefinitely growing or diminishing configurations—* ***instructive designs****.* It follows from Theorem 4.1.1 that the weights of edges in the optimal experimental graph cannot be all either too small or too large. This was proved in the case of progressive designs when the Kullback–Leibler divergence is employed as experiment information quantifying measure. We strongly believe that the same result holds in the instructive case as well.

3 ***Solution*** *to the problem of optimal design for* ***proximity graphs*** *in Euclidean spaces.* Theorem 4.2.2 from Chapter 4 states that the solution to the n-node optimal design problem for star-shaped proximity (geometric) graphs is explicitly described by $n-1$ quantiles of the prior distribution of the interaction radius, the threshold θ. If proximity graphs are considered in metric spaces, then the solution to the problem can be obtained numerically as an optimisation problem with linear constraints (Example 4.2.3 in § 4.2.4). Can an efficient algorithm be devised in order to solve the problem for proximity graphs in Euclidean spaces? The same question should also be

answered in the case of the random graph model with a more general form of the edge-probability function. Genetical algorithms in MCMC discussed in Ruiz et al (2007) together with lattice approximation of the graph nodes can be used for further development of experimental design for random graph models in regions of Euclidean spaces.

4 ***Asymptotic behaviour of maximum likelihood estimates*** *for finite percolation clusters* ***on*** *arbitrary* ***locally finite graphs***. Theorem 5.1.3, proved in § 5.1.2, states that for square lattices $\mathbb{L}^d$ the sequence of maximum likelihood estimates $\hat{p}_n$ of the bond percolation probability p stemmed from finite clusters of size n converges to the critical probability $p_c(d)$. Conjecture 5.1.5 states that this convergence is monotone and no estimate from the sequence exceeds the critical probability. Conjecture 5.1.6 is a strong assertion about the corresponding sequence of posterior distributions of the bond percolation probability under the same circumstances—it is stated that the limiting function of this functional sequence is the Dirac delta function, regardless the choice of the prior distribution. Furthermore, we conjecture that results similar to the mentioned ones still hold for percolation on any infinite locally finite graphs. Can such statements be proved or shown to be wrong?

5 ***Inner-outer design plots***. These symmetric designs were constructed for inference on the percolation parameter when edges are not observed (scenario $\mathcal{S}1$) by removing some sites from square plots. It was shown that within a class of inner-outer plots the most dense plot (a square with no removed nodes) may not generally be the optimal design. Are there even more 'sparsified' configurations that outperform inner-outer plots under this observation scenario? How could we identify them?

6 ***ε-Optimal approximation grids***. A weak point of the (approximate) solution to this problem presented in this thesis is that the optimal orientation of the coordinate system (grid or lattice) is chosen by consequently rotating the grid at small angles and applying Brucker–Meyer approximation. Can the number of grid rotations be reduced or avoided at all?

7 *Graphs as* ***snapshots*** *and* ***temporal graphs.*** Although random, the graphs considered in this thesis are static objects. It would be worthwhile to study design issues, similar to those discussed in the thesis, with regard to temporal random graphs, that is with regard to a sequence of graphs, possibly on the same fixed set of nodes with the edge structure changing over time.

8 ***Independence*** *in* ***edge formations.*** In our model we assumed that edges of the random graph appear at random, each being independent of the status of the rest. This is not always a sensible assumption. This assumption can be relaxed using so called *Markov graphs* for which one assumes conditional independence between an edge and all edges which are not adjacent to it. Markov graphs have found further generalisations—for example, in the exponential random graph family. The probability of a given graph from this family is an exponential function of a linear combination of some 'relevant' graph statistics (this kind of generalisation stemmed from a fundamental result for Markov graphs—the Hammersley–Clifford theorem; see Chapter 5 in Zager (2008) for more details and references).

9 *Random graphs with* ***fixed*** *nodes.* In our random graph model the edges are formed randomly but the nodes are fixed and it is their locations that are controlled in the experiment. However, if we relax the condition that the number of experimental sites is fixed or change it by setting an upper bound and allow the nodes to stem from a certain stochastic point process, then the dimensionality of the design space may be considerably reduced. For example, one may ask: "What is the optimal intensity of a homogeneous Poisson point process in $\mathbb{R}^d$ in which each pair of points within distance θ is independently joined with probability p?". As before, the purpose of such an experiment is to make inference on the model parameters (θ and p) while keeping the number of nodes limited. In order to formulate and solve the design problem within the utility-based Bayesian framework multipurpose utilities may have to be used.

10 ***Model discrimination, model robustness*** *and* ***sequential designs*** *in*

optimal design problems for random graphs. These are the issues which have not been addressed in this thesis. Our primary goal in considering optimal design problems for models based on random graphs was to make inference on the model parameter(s). This was done under assumption that the model in hand adequately described the phenomenon under study. Clearly, this is not always a sensible assumption and one may want to design an experiment to actually check its validity by considering a *model discrimination problem*. *Model robustness* is the degree to which the optimal designs depend on the prior distribution, that is how sensible the solution to the design problem is to the supplied prior information on the model parameter(s). (See Chaloner and Verdinelli (1995) and references therein for more details on designs for model discrimination and robustness to the prior distribution.) Finally, although sequential designs were not addressed in this paper, the chosen utility-based Bayesian framework of solving the problem of optimal node arrangement is perfectly suited for developing the methods of finding optimal designs sequentially (e.g. see DeGroot (1962) and Müller et al (2007)).

Bibliography

[1] `http://www.cl.cam.ac.uk/~aib29/HWThesis/Codes/`

[2] Adam, M., Delsanti, M. (1989) Percolation-type growth of branched polymers near gelation threshold. *Contemp. Phys.*, **30**(3), pp. 203–218.

[3] Agarwal, P. K., Har-Peled, S., Varadarajan, K. R. (2005) Geometric approximation via coresets. In *Combinatorial and Computational Geometry,* vol. 52, eds. J. E. Goodman, J. Pach and E. Welzl), Cambridge University Press, MSRI Publications.

[4] Agarwal, P. K., Har-Peled, S., Varadarajan, K. R. (2003) Approximating extent measures of points.
`http://valis.cs.uiuc.edu/~sariel/research/papers/01/fitting/`
Preliminary versions appeared in *Proc. 12th ACM-SIAM Sympos. Discrete Algorithms*, pp. 148–157, 2001 and in *Proc. 42nd IEEE Sympos. Found. Comput. Sci.*, pp. 66–73, 2001.

[5] Albert, R., Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Review of Modern Physics*, **74**(1), pp. 47–97.

[6] Amari, S. (2005) Population Coding, Bayesian Inference and Information Geometry. In *Advances in Neural Networks—ISNN 2005. Second International Symposium on Neural Networks, Chongqing, China, May 30–June 1, 2005, Proceedings, Part II.* Eds.: J. Wang, X. Liao, and Z. Yi, pp. 1–4.

[7] Amari, S., Nagaoko, H. (2000) *Methods of Information Geometry.* Transactions of Mathematical Monographs, vol. 191. American Mathematical Society.

[8] Atkinson, A. C., Chaloner, K., Herzberg, A. M., Juritz, J. (1993) Optimum experimental designs for properties of a compartmental model. *Biometrics*, **49**, pp. 325–337.

[9] Atkinson, A. C., Donev, A. N. (1992) *Optimum Experimental Designs*. Clarendon Press, Oxford.

[10] Ball, F., Neal, P. (2008) Network epidemic models with two levels of mixing. *Mathematical Biosciences,* **212**, p. 69–87.

[11] Barabasi, A. L. (2002) *Linked, The New Science of Networks.* Perseus, Cambridge, MA.

[12] Bailey, D. J., Otten, W., Gilligan, C. A. (2000) Saprotrophic invasion by the soil-borne fungal plant pathogen *Rhizoctonia solani* and percolation thresholds. *New Phytol.*, **146**, pp. 535–544.

[13] Bailey, D. J., Gilligan, C.A. (1997) Biological control of pathozone behaviour and disease dynamics of *Rhizoctonia solani* by *Trichoderma viride*. *New Phytol.*, **136**, pp. 359–367.

[14] Barequet, G., Har-Peled, S. (2001) Efficiently approximating the minimum-volume bounding box of a point set in three dimensions. *J. Algorithms*, **38**, pp. 91–109.

[15] Bejan, A. Iu. (2008) Grid approximation of a finite set of points. Conference *Mathematics & IT: Research and Education 2008*, Chişinău, October, 1-4.

[16] Bejan, A. Iu., Gibson, G. J., Zachary, S. (2008) Inference and experimental design for some random graph models. Workshop *Designed Experiments: Recent Advances in Methods and Applications (DEMA2008)*, Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, 11-15 August 2008.

[17] Bejan, A. (2008) MCMC in modern applied mathematics. *Lecture Notes*, Center for Education and Research in Mathematics and Computer Science, Department of Mathematics and Computer Science, State University of Moldova. `http://www.cl.cam.ac.uk/~aib29/CECMI/MCMC/notes.pdf`

[18] Bentz, D. P., Garboczi, E. J. (1992) Modelling of the leaching of calcium hydroxide from cement paste: Effects on pore space percolation and diffusivity. *Materials and Structures*, **25**(9), pp. 523–533.

[19] Berger, J. O., Bernardo, J. M., Mendoza, M. (1989) On priors that maximize expected information. In *Recent Developments is Statistics and Their Applications,* eds.: J. P. Klein and J. C. Lee.

[20] Berger, J. O., Wolpert, R. L. (1988) *The Likelihood Principle*. Institute of Mathematical Statistics. Lecture Notes–Monograph Series, vol. 6, ed.: Shanti S. Gupta.

[21] Bernardo, J. M. (2003) Bayesian statistics. In *Probability and Statistics* of the *Encyclopedia of Life Support Systems (EOLSS)*, ed. R. Viertl. Oxford, UK: UNESCO.

[22] Besag, J., Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc. B*, **55**(1), pp. 25–37.

[23] Bielza, C., Müller, P., Ríos-Insúa, D. (1999) Monte Carlo methods for decision analysis with applications to influence diagrams. *Manag. Sci.*, **45**(7), pp. 995–1007.

[24] Bijay, K. G., Vigneron,A. (2005) A practical approach to approximating diameter of point-set in low dimensions. In *Proceedings of the 17th Canadian Conference on Computational Geometry (CCCG'05)*, pp. 3–6.

[25] Birnbaum, A. (1962) On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, **57**, pp. 269–306 (with comments by L. J. Savage, G. A. Barnard, J. Cornfield, Irwin Bross, G. E. P. Box, I. J. Good, D. V. Lindley, C. W. Clunies-Ross, J. W. Pratt, H. Levene, T. Goldman, A. P. Dempster, O. Kempthorne and reply by A. Birnbaum).

[26] Blackwell, D. (1951) Comparison of experiments. *Proceedings of the Second Berkely Symposium on Mathematical Statistics and Probability*. University of California Press, pp. 93–102.

[27] Bowman, K. O., Shenton,L. R. (1988) *Properties of Estimators for the Gamma Distribution.* Statistics: Textbooks and Monographs, vol. 89, Marcel Dekker, Inc.

[28] Broadbent, S. R., Hammersley, J. M. (1957) Percolation processes, I and II. *Proc. Camb. Phil. Soc.*, **53**, pp. 629–645.

[29] Brucker, P., Meyer, W. (1987) Approximation of a set of points by points of a grid. *Computing*, **38**, pp. 341–345.

[30] Bunde, A., Havlin, S., Porto, M. (1995) Are branched polymers in the universality class of percolation? *Phys. Rev. Lett.*, **74**, pp. 2714–2716.

[31] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., Watts, D. J. (2000) Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, **85**(25), pp. 5468–5471.

[32] Chaloner, K., Verdinelli, I. (1995) Bayesian experimental design: A review. *Stat. Sci.*, **10**, pp. 273–304.

[33] Chan, T. M. (2006) Faster core-set constructions and data-stream algorithms in fixed dimensions. *Computational Geometry: Theory and Applications*, vol. 35(1), pp. 20–35.

[34] Chase, A. R. (1998) Rhizoctonia diseases on ornamentals. *Western connection. Turf and ornamentals*, **1**(2), pp. 1–4.

[35] Clyde, M.A. (2004) Experimental design: a Bayesian perspective. *Int. Encyc. Social and Behaviour Sciences*, Elsevier Ltd, Editors-in-Chief: Neil J. Smelser and Paul B. Baltes, pp. 5075–5081.

[36] Clyde, M.A., Müller, P., Parmigiani, G. (1995) Inference and design strategies for a hierarchical logistic regression model. *Bayesian Biostatistics*, eds.: A. D. Berry and D. Stangl, New York: Marcel Dekker, pp. 297–320.

[37] Cohen, R., Erez, K., ben-Avraham, D., Havlin, S. (2000) Resilience of the Internet to random breakdowns. *Phys. Rev. Lett.*, **85**, pp. 4626–4628.

[38] Cook, A. R., Gibson, G. J., Gilligan, C. A. (2008) Optimal observation times in experimental epidemic processes. *Biometrics*, **64**(3), pp. 860–868.
Web Appendices available at
`http://www.biometrics.tibs.org/datasets/070104.pdf`

[39] Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., Knuth, D. E. (1996) On the Lambert W function. *Advances in Computational Mathematics*, **5**, pp. 329Ű-359.

[40] Cox, J. T., Durrett, R. (1988) Limit theorems for spread epidemics and forest fires. *Stochastic Process Appl.*, **30**, pp. 171–191.

[41] Cramér, H. (1946) A contribution to the theory of statistical estimation. *Skand. Aktuar.*, **29**, pp. 85–94.

[42] Cronbach, L. J. (1953) A consideration of information theory and utility theory as tools for psychometric problems. *Technical Report No. 1, Contract N6ori-07146, Urbana, Illinois.*

[43] Csiszár, I. (1967) Information-type measures of difference of probability distributions, and indirect observations. *Studia Scientarum Mathematicarum Hungarica*, **2**, pp. 191–213.

[44] Curtis, A. (2004a) Theory of model-based geophysical survey and experimental design. Part I—linear problems. *The Leading Edge*, **23**(10), pp. 997–1004.

[45] Curtis, A. (2004b) Theory of model-based geophysical survey and experimental design. Part II—nonlinear problems. *The Leading Edge*, **23**(11), pp. 1112–1117.

[46] de Gennes, P. G., Guyon, E. (1978) Lois générales pour l'injection d'un fluide un milieu poreux alétoire. *J. de Mécanique*, **3**, p. 403.

[47] De Bondt, S.,Froyen, L., Deruyttere, A. (1992) Electrical conductivity of composites: a percolation approach. *J. Mater. Sci.*, **27**, pp. 1983–1988.

[48] DeGroot, M. H. (1984) Changes in utility as information. *Theory and Decision*, **17**, pp. 287–303.

[49] DeGroot, M. H. (1962) Uncertainty, information, and sequential experiment. *The Annals of Mathematical Statistics*, **33**(2), pp. 404–419.

[50] Demiris, N. (2004) Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo. *PhD thesis*. University of Nottingham.

[51] Devoto, A., Duke, D. W. (1984) Table of integrals and formulae for Feynman diagram calculations. *La Rivista del Nuovo Cimento*, **7**(6), pp. 1–39.

[52] Dirac, P. A. M. (1958) *The principles of quantum dynamics*. 4th ed. Oxford, UK: Clarendon Press.

[53] Dirac, P. A. M. (1927) The physical interpretation of the quantum dynamics. *Proceedings of the Royal Society of London. Series A, Containing Papers of Mathematical or Physical Character*, **113**(765), pp. 621–641.

[54] Dorogovtsev, S. N., Mendes, J. F. F. (2003) *Evolution of Networks, From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford.

[55] Edwards, A. W. F. (1974) The history of likelihood. *International Statistical Review*, **42**, pp. 9–15.

[56] Edwards, A. W. F. (1972) *Likelihood*. Cambridge University Press, Cambridge (expanded edition, 1992, Johns Hopkins University Press, Baltimore).

[57] Erdős, P., Rényi, A. (1960) The evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5**, pp. 17–61.

[58] Erdős, P., Rényi, A. (1959) On random graphs. I. *Publicationes Mathematicae*, **6**, pp. 290–297.

[59] Filipe, J. A. N., Otten, W., Gibson, G., Gilligan, C. A. (2003) Inferring the dynamics of a spatial epidemic from time-series data.

[60] Firth, D., Hinde, J.P. (1997a) On Bayesian D-optimum design criteria and the equivalence theorem in non-linear models. *J. R. Statist. Soc. B*, **59**(4), pp. 793-797.

[61] Firth, D., Hinde, J. P. (1997b) Parameter neutral optimum design for non-linear models. *J. R. Statist. Soc. B*, **59**(4), pp. 799–811.

[62] Fisher, R. A. (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, **1**, pp. 3–32.

[63] Fréchet, M. (1943) Sur l'extension de certain evaluations statistiques au cas des petit echantillons. *Rev. Inst. Stat.*, **11**, pp. 182–205.

[64] Frieden, B. R. (2004) *Science from Fisher Information*. New York: Cambridge University Press.

[65] Fuentes, M., Chaudhuri, A., Holland, D.M. (2007) *Environmental and Ecological Statistics*, **14**(3), pp. 323–340.

[66] Gatrell, A. C., Bailey, T. C., Diggle, P. J., Rowlingson, B. S. (1996) Spatial point pattern analysis and its application in geographical epidemiology. *Trans. Instr. Br. Geogr.*, **21**, pp. 256–274.

[67] Geman, S., Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, **6**, pp. 721–741.

[68] Geyer, C. J. (1992) Practical Markov Chain Monte Carlo. *Statistical Science*, **7**(4), pp. 473–511.

[69] Gibbons, A. M. (1985) *Algorithmic Graph Theory*. Cambridge University Press.

[70] Gibson, G. (1997) Markov Chain Monte Carlo methods for fitting and testing spatio-temporal stochastic models in plant epidemiology. *Appl. Stat.*, **46**(2), pp. 215–233.

[71] Gibson, G. J., Kleczkowski, A., Gilligan, C. A. (2004) Bayesian analysis of botanical epidemics using stochastic compartmental models. *Proc. Nat. Acad. Sci.*, **101**, pp. 12120–12124.

[72] Gibson, G. J., Otten, W., Filipe, J. A. N. Cook, A., Marion, G., Gilligan, C. A. (2006) Bayesian estimation for percolation models of disease spread in plant populations. *Stat. Comput.*, **16**(4), pp. 391–402.

[73] Gilbert, E. (1959) Random graphs. *The Annals of Mathematical Statistics*, **30**, pp. 1141–1144.

[74] Ginebra, J. (2007) On the measure of the information in a statistical experiment. *Bayesian Analysis*, **2**(1), pp. 167–212.

[75] Gingl, Z., Pennetta, C., Kiss, L. B., Reggiani, L. (1996) Biased percolation and abrupt failure of electronic devices. *Semiconductor science and technology*, **11**(12), pp. 1770–1775.

[76] Golumbic, M. C. (2004) Algorithmic graph theory and perfect graphs. Second edition included in *Annals of Discrete Mathematics*, **57**.

[77] Graham, R. L., Knuth, D. E., Patashnik, O. (1999) *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Publishing Company, sixth printing, with corrections.

[78] Grassberger, P. (1983) On the critical behaviour of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, **63**(2), pp. 157–172.

[79] Grimmett, G. R. (2008) *Probability on Graphs*. Lecture Notes for 2008 PIMS-UBC Summer School in Probability, Institut Henri Poincaré.
`http://www.statslab.cam.ac.uk/~grg/books/pgsUS.pdf`

[80] Grimmett, G. (1999) *Percolation*. Springer Verlag, Berlin, 2nd ed.

[81] Gudelj, I., White, K. A. J., Britton, N. F. (2004) The effects of spatial movement and group interactions on disease dynamic of social animals. *Bulletin of Mathematical Biology*, **66**, pp. 91–108.

[82] Gupta, A. (1999) Embedding tree metrics into low dimensional Euclidean spaces. Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing, pp. 694–700.

[83] Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **82**, pp. 711–732.

[84] Irony, T. Z., Singpurwalla, N. D. (1997) Non-informative priors do not exist. A dialogue with José M. Bernardo. *Journal of Statistical Planning and Inference*, **65**, pp. 159–189.

[85] Jammer, M. (1994) *The conceptual Development of Quantum Mechanics*. New York, NY, USA: McGraw-Hill.

[86] Keeling, M. (2005) The implications of network structure for epidemic dynamics. *Theoretical Population Biology*, **67**, pp. 1–8.

[87] Keeling, M. J., Brooks, S. P., Gilligan, C. A. (2004) Using conservation of pattern to estimate parameters from a single snapshot. *Proc. Nat. Adac. Sci.*, **101**, pp. 9155–9160.

[88] Keeling, M. J. (1999) The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond.*, **266**, pp. 859–867.

[89] Kiefer, J. (1959) Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society, Series B*, **21**, pp.271–319.

[90] Khuri, A. I. (1984) A note on D-optimal designs for partially nonlinear regression models. *Technometrics*, **26**(1), pp. 59–61.

[91] Kueck, H., Hoffman, M., Doucet, A., de Freitas, N. (2009) Inference and Learning for Active Sensing, Experimental Design and Control. In *Lecture Notes in Computer Science*, **5524**, pp. 1–10.

[92] Kullback, S. (1968) *Information Theory and Statistics*. Dover Publications, Inc., second revised edition.

[93] Kullback, S. (1954) Certain inequalities in information theory and the Cramér–Rao inequality. *Ann. Math. Stat.*, **25**, pp. 745–751.

[94] Kullback, S. (1952) An application of information theory to multivariate analysis. *Ann. Math. Stat.*, **23**, pp. 88–102.

[95] Kullback, S., Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, pp. 79–86.

[96] Kuulasmaa, K., Zachary, S. (1984) On spatial general epidemics and bond percolation processes. *J. Appl. Prob.*, **21**, pp. 911–914.

[97] Lalley, S. P. (2007) Critical scaling of stochastic epidemic models. *IMS Lecture Notes-Monograph Series. Asymptotics: Particles, Processes and Inverse Problems*, vol. 55, pp. 167–178.

[98] Larson, R. G., Scriven L. E., Davis, H. T. (1981) Percolation theory of two-phase flow in porous media. *Chem. Eng. Sci.*, **15**, pp. 57–73.

[99] Lauritzen, S. L. (1999) Aspects of T. N. Thiele's contributions to statistics. *Bulletin of the International Statistical Institute*, **58**, pp. 27–30.

[100] Levin, D. A., Peres, Y., Wilmer, E. L. (2009) *Markov Chains and Mixing Times*. American Mathematical Society.
`http://www.uoregon.edu/~dlevin/MARKOV/`

[101] Leonard, Th., Hsu, J. S. J. (1999) *Bayesian Methods. An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge University Press.

[102] Lindley, D.V. (1972) Bayesian Statistics. *A Review*. SIAM, Philadelphia.

[103] Lindley, D.V. (1961) The use of prior probability distributions in statistical inference and decisions. *Proc. Fourth Berkely Symp. on Math. Statist. and Prob.*, Vol. I, pp. 453–468.

[104] Lindley, D.V. (1956) On the measure of information provided by an experiment. *Annals of Statistics*, **27**, pp. 986–1005.

[105] Machta, J. (1991) Phase transitions in fractal porous media. *Phys. Rev. Lett.*, **66**, pp. 169–172.

[106] MacKay, D. J. C. (2003) *Information theory, inference, and learning algorithms*. Cambridge University Press.
`http://www.inference.phy.cam.ac.uk/mackay/itilia/book.html`

[107] McMillan, B. (1953) The basic theorems of information theory. *Ann Math. Stat.*, vol. 24, pp. 196–219.

[108] Madras, N. N. (2002) *Lectures on Monte Carlo Methods.* American Mathematical Society.

[109] Meester, R., Roy, R. (1996) *Continuum percolation.* Cambridge University Press.

[110] Metropolis, N., Rosenbluth, A. W., Teller, M. N., Teller, E. (1953) Equation of state calculations by fast computing machinces. *Journal of Chemical Physics,* **21**, pp. 1087–1092.

[111] Milgram, S. (1967) The small world problem. *Psychology Today*, **1**(1), pp. 60–67.

[112] Minka, T. (2005) Divergence measures and message passing. Technical report MSR-TR-2005-173, Microsoft Research, Cambridge, UK.

[113] Moon, K., Girvin, S. M. (1995) Critical behavior of superfluid ^{4}He in aerogel. *Phys. Rev. Lett.*, **75**, pp. 1328–1331.

[114] Müller, W. G. (2007) *Collecting Spatial Data: Optimum Design of Experiments for Random Fields.* Springer, New York.

[115] Müller, P., Berry, D. A., Grieve, A. P., Smith, M., Krams, M. (2007) Simulation-based sequential Bayesian designs. *Journal of Statistical Planning and Inference*, **137**(10), pp. 3140–3150.

[116] Müller, P. (1999) Simulation-based Optimal Design. In *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds.) Oxford University Press, pp. 459–474.

[117] Murray, I. (2007) Advances in Markov chain Monte Carlo methods. *PhD Thesis*, University College of London.

[118] Neal, P. (2003) SIR epidemics on a Bernoulli random graph. *J. Appl. Prob.*, **40**, pp. 779–782.

[119] Odagaki, T., Toyufuku, S. (1998) Properties of percolation clusters in a model granular system in two dimensions. *J. Phys. CM*, **10**, pp.6447–6452.

[120] O'Hagan, A. (1994) *Bayesian Inference*. Kendall's Library of Statistics, Vol. 2B, Edward Arnold.

[121] Otten, W., Bailey, D. J., Gilligan, C. A. (2004) Empirical evidence of spatioal thresholds to control invasion of fungal parasites and saprotrophs. *New Phytol.*, **163**, pp. 125–132.

[122] Paninski, L. (2005) Asymptotic theory of Information-theoretic experimental design. *Neural Computation*, **17**, pp. 1480–1507.

[123] Papadimitriou, C., Haralampidis, Y., Sobczyk, K. (2005) Optimal experimental design in stochastic structural dynamics. *Probabilistic Engineering Mechanics*, **20**, pp. 67–78.

[124] Paquet, U. (2008) Bayesian inference for latent variable models. *Technical Report 724*. Computer Laboratory, University of Cambridge.

[125] Parham, P. E., Ferguson, N. M. (2006) Space and contact networks: capturing the locality of disease transmission. *J. R. Soc. Interface*, **3**(9), pp. 483–493.

[126] Parmigiani, G., Berry, D. A. (1994) Applications of Lindley information measure to the design of clinical experiments. In *Aspects of Uncertainty: a Tribute to D. V. Lindley*. (P. R. Freeman, and A. F. M. Smith, eds.), Chichester: Wiley, pp. 329–348.

[127] Pennetta, C., Reggiani, L., Alfinito, E., Trefan, G. (2002) Stationary regime of random resistor networks under biased percolation. *J. Phys.: Condens Matter*, **14**(9), pp. 2371–2378.

[128] Penrose, M. (2003) *Random Geometric Graphs.* Oxford University Press.

[129] Raiffa, H., Schlaifer, R. O. (1961) *Applied Statistical Decision Theory*. Cambridge: M.I.T. Press.

[130] Rao, C. R. (1945) Information and accuracy attainable in estimation of statistical parameters. *Bull. Cal. Math. Soc.*, **37**, pp. 81–91.

[131] Ray, T. S., Jan, N. (1994) Anomalous approach to the self-organized critical state in a model for life at the edge of chaos. *Phys. Rev. Lett.*, **72**, pp. 4045–4048.

[132] Read, J. M., Keeling, Matt, J. (2003) Disease evolution on networks: the role of contact structure. *Proc. R. Soc. Lond. B*, **270**, pp. 699–708.

[133] Ren, Y., Ding, Yu, Liang, F. (2008) Adaptive evolutionary Monte Carlo algorithm for optimization with applications to sensor placement problems. *Stat. Comput.*, **18**, p. 375–390.

[134] Robert, C. P. (2007) *The Bayesian Choice*. Springer Texts in Statistics. Springer, New York, Second Edition.

[135] Ruiz, F., Ferreira, M. A. R., Schmidt, A. M. (2007) Evolutionary Markov chain Monte Carlo algorithms for optimal monitoring network designs. *Proceedings of the Joint Statistical Meetings*, Section on Bayesian Statistical Science, pp. 1332–1338.

[136] Ryan, Th. P. (2007) *Modern Experimental Design*. Wiley-Interscience.

[137] Ryan, K., J. (2003) Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *Journal of Computational and Graphical Statistics*, **12**(3), pp. 1–19.

[138] Sahimi, M. (1994) Long-range correlated percolation and flow and transport in heterogeneous porous media. *J. Physique I*, **4**, pp. 1263–1268.

[139] Shavitt, Yu., Tankel, T. (2004) Big-bang simulation for embedding network distances in Euclidean space. *IEEE/ACM Transactions on Networking*, **12**(6), pp. 993–1006.

[140] Solomon, S., Weisbuch, G., de Arcengelis, L., Stauffer, N. Jan D. (2000) Social percolation models. *Physica A*, **277**, pp. 239–247.

[141] Sondow, J., Weisstein, E. W. (2009) *Harmonic Number*. From *MathWorld*—A Wolfram Web Resource.

[142] Stauffer, D., Aharony, A. (1992) *Introduction to Percolation Theory*. Taylor and Francis, London (second edition).

[143] Tobochnik, J. (1999) Granular collapse as a percolation transition. *Phys. Rev. E*, **60** pp. 7137–7142.

[144] Trapman, P. (2006) On stochastic models for the spread of infections. *PhD Thesis*.

[145] Travers, J., Milgram, S. (1969) An experimental study of the small world problem. *Sociometry*, **32**(4), pp. 425–443.

[146] Tyrakowski, T., Palka, Z. (2005) A Random Graph Model of Mobile Wireless Networks. In *Electronic Notes in Discrete Mathematics*, **22** pp. 311–314.

[147] Vempala, S. S. (2006) *The Random Projection Method*. Series in Discrete Mathematics and Theoretical Computer Science, vol. 65 American Mathematical Society.

[148] Venegas-Martínez, F. (2004) On information measures and prior distributions: a synthesis. *Morfismos*, **8**(2), pp. 27–50.

[149] Verdinelli, I. (1992) Advances in Bayesian experimental design. *Bayesian Statistics*, **4** (J.M.Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.) Oxford: University Press, pp. 448–467.

[150] Watts, D. J. (2003) *Small Worlds: The Dynamics of Networks between Order and Randomness*. (Princeton Studies in Complexity). Princeton University Press, illustrated edition.

[151] Watts, D. J. (1998) Collective dynamics of small world networks. Nature, **393**, pp. 440–442.

[152] With, K. A., Crist, T. O. (1992) Critical thresholds in species' responses to landscape structures. *Ecology*, **76**(8), pp. 2446–2459.

[153] Yanuka, M (1992) Percolation theory approach to transport phenomena in porous media. *Transport in Porous Media*,**7**(3), pp. 265–282.

[154] Zacks, S. (1981) Parametric Statistical Inference. *International Series in Nonlinear Mathematics: Theory, Methods and Applications*, Editors: V. Lakshmikantham, C.P. Tsokos. **Vol. 4**, pp. 103–104.

[155] Zaroliagis, C. D. (2002) Implementation and experimental studies of dynamic graph algorithms. In *Experimental Algorithmics—From Algorithm Design to Robust and Efficient Software. Lecture Notes of Computer Sciences*, **2547**, pp. 229–278.

[156] Zhang, Yu. (1993) *A shape theorem for epidemics and forest fires with finite range interactions. Ann. Probab.*, **21**(4), pp. 1755–1781.

Appendix A

Solving $a^{bx+c} = dx + e$ and maximising $x^2/(e^x - 1)$

A.1 Equation $a^{bx+c} = dx + e$

The equation of the form

$$a^{bx+c} = dx + e \tag{A.1}$$

is a transcendental equation which can be solved using the Lambert W function (see Corless et al (1996)). This function $W(x)$ satisfies the equation

$$W(x)e^{W(x)} = x,$$

which cannot be solved in elementary functions.

Before solving equation (A.1) we notice that the equation $ta^t = A$, a>0, can be solved in terms of the Lambert function as follows: $t = W(A \log a)/\log a$.

We now solve equation (A.1) using the substitution $t = -b\,(x + e/d)$. Under this transformation the original equation (A.1) becomes

$$ta^t = A = -\frac{b}{d}(a^c - e),$$

and, thus,

$$-bx - \frac{be}{d} = \frac{W(A \log a)}{\log a},$$

producing the solution

$$x = -\frac{W(-\frac{b \log a}{d} a^{(c-be/d)})}{b \log a} - e/d.$$

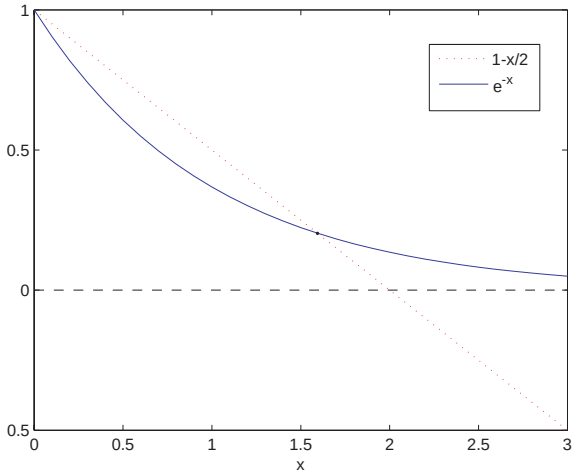


Figure A.1: Intersection of the graphs of functions $e^{-x} = 1 - x/2$ and $1 - x/2$, $x > 0$.

In particular, if $e^{-x} = 1 - x/2$, then $a = e$, $b = -1$, $c = 0$, $d = -1/2$, $e = 1$, and the solution is $x = W(-2e^{-2}) - 2 \approx 1.593624$.

A.2 Maximisation of $x^2/\left(e^{\theta x} - 1\right)$

We start with the observation that in order to maximise the function

$$x \to x^2/\left(e^{\theta x} - 1\right), \; x > 0, \tag{A.2}$$

where the value of the parameter $\theta > 0$ is fixed, it is sufficient to maximise the function $f(x) = x^2/\left(e^x - 1\right)$, since $f(\theta x) = \theta^2 f(x)$; this means that the maximum x^*_θ of (A.2) relates to the maximum x^* of $f(x)$, as follows: $x^*_\theta = x^*/\theta$. We will find x^* now.

The derivative $f'(x)$ of $f(x)$ is as follows:

$$f'(x) = \frac{(2-x)e^x - 2}{\left(e^x - 1\right)^2},$$

and equating this derivative to zero is equivalent to solving the equation $e^{-x} = 1 - x/2$ when $x > 0$. This equation has a single root in $(0, +\infty)$, as is shown in Figure A.1; its value x^* found in Appendix A can be expressed in terms of the Lambert special function $W(x)$ as follows: $x^* = W(-2e^{-2}) - 2 \approx 1.593624$. It follows that $x^*_\theta = (W(-2e^{-2}) - 2)/\theta \approx 1.593624/\theta$.

Appendix B

Dirac delta function

The delta function was introduced by the English physicist Paul Dirac (1927, 1958) in the context of quantum mechanics. This notion relates to previous work by G. Kirchhoff and O. Heaviside (see Jammer (1966)). The delta function is a reflection of Dirac's idea of constructing a strictly localised function on the real numbers: $\delta(x)$ is zero for any x, except for $x = 0$, where it is peaked. The following characteristic property makes this idea more precise:

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_0)dx = f(x_0)$$

for any smooth and absolutely integrable function f. This identity is called the *sifting property of the delta function*.

The delta Dirac function can be defined as a limit of a sequence of functions. A *delta* or *Dirac sequence* of functions $g_n(x)$, $n \in \mathbb{N}$, is a sequence of non-negative strongly peaked functions for which

$$\lim_{n\to\infty} \int_{\infty}^{\infty} g_n(x)f(x)\,dx = f(0)$$

for any smooth and absolutely integrable function $f(x)$.

Equivalently, a delta sequence $\{g_n\}_{n=1}^{\infty}$ satisfies the following conditions:

1 $g_n(x) \geq 0$ for all n and all $x \in \mathbb{R}$.

2 $\int_{\infty}^{\infty} g_n(x)\,dx = 1$ for all n.

3 For every $\gamma > 0$ and $\epsilon > 0$ there is $N \in \mathbb{N}$, such that for all $n > N$

$$\int_{\mathbb{R}\setminus[-\gamma,\gamma]} g_n(x)\,dx < \epsilon.$$

A delta sequence $\{g_n\}_{n=1}^{\infty}$ *'converges'* to, or generates, the Dirac delta function $\delta(x)$. Shifted delta function $\delta(x - x_0)$ can be considered by appropriate shifting in the construction presented above.

Appendix C

Integration of polylogarithms

We consider the following integral:

$$I_\alpha(d) := \int_0^\infty e^{-\alpha\theta}(1-e^{-\theta d})\log(1-e^{-\theta d})\,\mathrm{d}\theta,\ \alpha, d > 0.$$

This integral can be reduced to an integral of a polylogarithm (and hence computed analytically for some particular values of $d = d(\alpha)$) using the following change of variables:

$$x = e^{-\theta d},\ \mathrm{d}\theta = -\frac{1}{xd}\mathrm{d}x.$$

Thus,

$$\begin{aligned} I_\alpha(d) &= -\frac{1}{d}\int_1^0 \frac{x^{\alpha/d}}{x}(1-x)\log(1-x)\,\mathrm{d}x \\ &= \frac{1}{d}\int_0^1 y^{\kappa-1}(1-y)\log(1-y)\,\mathrm{d}x \\ &= \frac{1}{d}\left(\int_0^1 y^{\kappa-1}\log(1-y)\,\mathrm{d}x - \int_0^1 y^{\kappa}\log(1-y)\,\mathrm{d}x\right), \end{aligned} \tag{C.1}$$

where $\kappa := \alpha/d$.

Suppose that $\kappa \in \mathbb{N}$. In this case ((4.2.4) in Devoto and Duke (1984, p.30))

$$\int_0^1 x^\kappa \log(1-x)\,\mathrm{d}x = -\frac{H_\kappa}{\kappa+1},$$

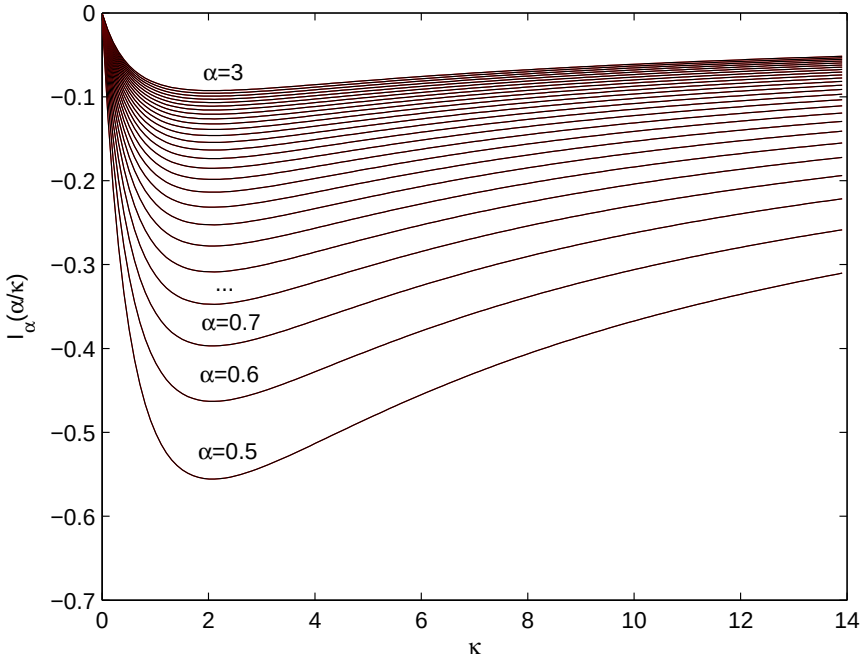


Figure C.1: Plots of the function $I_\alpha(\alpha/\kappa)$ when α is fixed, $\kappa \in [0, 14]$ ($\alpha = 0.5, 0.6, \ldots, 3$). The plots have been obtain both by using numerical evaluation of integrals in (C.1) and representation (C.3).

where H_κ is the κ^{th} harmonic number:

$$H_\kappa := \sum_{i=1}^{\kappa} \frac{1}{i}.$$

Thus we obtain:

$$I_\alpha(\alpha/\kappa) = \alpha^{-1}\left(\frac{\kappa}{\kappa+1}H_{\kappa+1} - H_\kappa\right), \text{ for all } \kappa \in \mathbb{N}. \tag{C.2}$$

It is interesting to note that the representation (C.2) can further be generalised for any $\kappa \in \mathbb{R}_+$ providing an alternative to numerical evaluation of integrals participating in (C.1). This can be done using the fact that a harmonic number H_κ can be expressed analytically as follows (Sondow and Weisstein (2009)):

$$H_\kappa = \gamma + \psi(\kappa + 1), \tag{C.3}$$

where γ is the Euler-Mascheroni constant ($\gamma = 0.577215664901\ldots$) and ψ is the *digamma function*, $\psi(z) = \Gamma'(x)/\Gamma(z)$. Thus, it follows that

$$\alpha I_\alpha(d) = \frac{\alpha\psi(\alpha/d + 1) - \gamma d}{\alpha + d} - \psi(\alpha/d), \tag{C.4}$$

and the right-hand side of (C.4) depends on the ratio α/d only.

Figure C.1 depicts plots of $I_\alpha(\alpha/\kappa)$ (for some fixed values of α) obtained using numerical evaluation of the integrals in (C.1) on the one hand, and (C.3) on the other hand—the plots thus obtained, corresponding to the same values of α, are practically indistinguishable.

Appendix D

Realisation of 6 distances in $\mathbb{R}^3$

Let us assume that we are given a set of non-negative numbers d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34} satisfying the triangle inequalities:

$$d_{ij} + d_{jk} \geq d_{ik},\ i, j, k = 1, 2, 3, 4. \tag{D.1}$$

Are there such points U_1, U_2, U_3 and U_4, all from $\mathbb{R}^3$, that d_{ij} equals the distance between U_i and U_j?

Let us fix any two points from $\mathbb{R}^3$ calling them U_1 and U_2 such that the length of the segment U_1U_2 is d_{12}. Since all our considerations are valid up to an orthonormal transformation, we can assume without loss of generality that U_1 and U_2 are symmetric with respect to the centre O of the coordinate system $Oxyz$ and lie on the coordinate axis Ox: $U_1 = U_1(-d_{12}/2, 0)$ and $U_2 = U_2(d_{12}/2, 0)$.

Consider the plane Oxy. It is straightforward to find points T_1 and T_2 in this plane such that

$$U_1T_1 = d_{13},\ U_2T_1 = d_{23}$$

$$U_1T_2 = d_{14},\ U_2T_2 = d_{24}.$$

Indeed, the point T_1 lies on an ellipse E_1 with U_1 and U_2 as its foci, the semi-major axis $a = (d_{13} + d_{23})/2$, and the semi-minor axis $b = \sqrt{a^2 - U_1U_2^2} = \sqrt{a^2 - d_{12}^2}$. Analogously, the point T_1 lies on an ellipse E_2 with U_1 and U_2 being its foci, the semi-major axis $a = (d_{13} + d_{23})/2$, and the semi-minor axis $b = \sqrt{a^2 - U_1U_2^2} = \sqrt{a^2 - d_{12}^2}$. Figure D.1 illustrates this construction and all subsequent constructions.

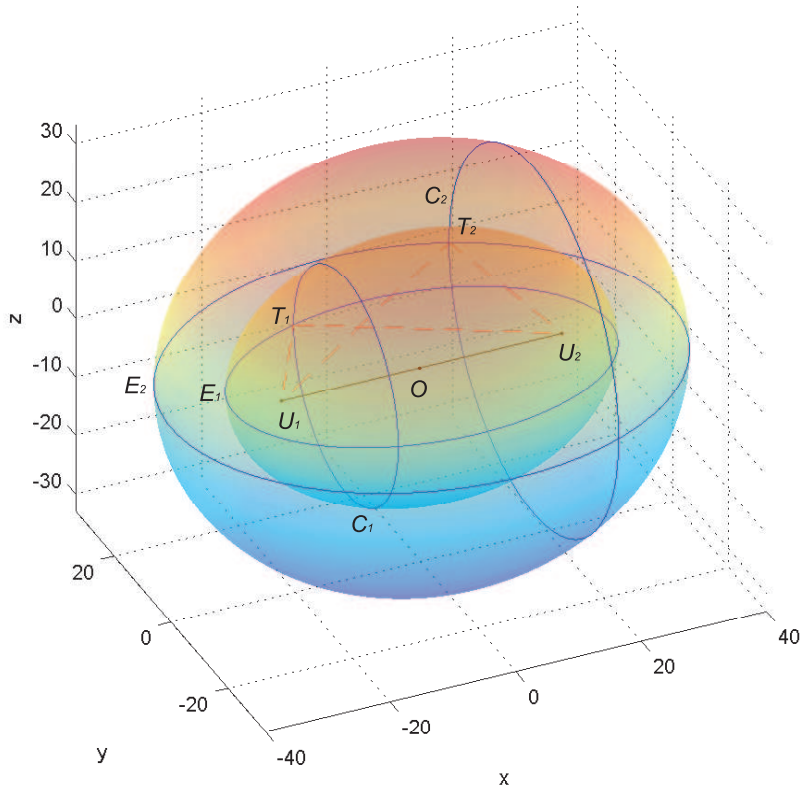


Figure D.1: Realisation of 6 distances in $\mathbb{R}^3$: a working scheme.

The distance between T_1 and T_2 will not generally be d_{34}. To achieve this we consider two circles C_1 and C_2 contained in the planes which are perpendicular to U_1U_2: these circles are such that C_1 contains T_1 and C_2 contains T_2. If $\min(C_1, C_2) \leq d_{34} \leq \max(C_1, C_2)$, where $\min(C_1, C_2)$ is the shortest distance between C_1 and C_2 and $\max(C_1, C_2)$ is the longest distance between C_1 and C_2, then two points U_3 and U_4 can be identified (in a non-unique way!) on C_1 and C_2, respectively, such that the pairwise distances between the points of the configuration $U_1U_2U_3U_4$ equal to the given non-negative numbers d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34} satisfying (D.1).

Appendix E

Gamma distribution, infectious times and site percolation

Site percolation may be approximated arbitrarily closely by an *SIR* epidemic model with infectious time distribution which with some sufficiently small probability takes some sufficiently large value, and which otherwise takes the value zero. We first briefly describe site percolation and then construct a sequence of gamma random variables with the mentioned property.

Site percolation model can be described as follows. One declares each vertex of the grid $\mathbb{L}^d$ open with probability p, independently of the status of the other vertices. The vertices which were not declared open are declared closed. A path on lattice is called open if it consists of open vertices. Similarly to bond percolation, the open cluster $\mathcal{C}(x)$ at x is defined as the set of all vertices 'reachable' by open paths from x (if x is closed then $\mathcal{C}(x)$ is empty). It is easy to see that our *SIR* epidemic model with an improper distribution of infectious times which takes the value $+\infty$ with probability p and the value 0 with probability $q = 1-p$ corresponds to site percolation (in the sense that the distributions of site configurations as epidemic outbreaks and open clusters in percolation coincide).

Consider a sequence of gamma random variables $X_n \sim \Gamma(\kappa_n, \theta_n)$. Here κ_n is a shape parameter of X_n, θ_n is its scale parameter, and the p.d.f. of X_n is as follows:

$$f_n(x; \kappa_n, \theta_n) = \frac{1}{\theta^{\kappa_n}\Gamma(\kappa_n)} x^{\kappa_n - 1} e^{-x/\theta_n}.$$

Fix some $q \in (0, 1)$, and let $\kappa_n \to 0$ as $n \to \infty$. Let also $\theta_n = q^{-1/\kappa_n}$, and note

that both the sequence of means of X_n and of the corresponding variances diverge:

$$\mathbb{E}X_n = \kappa_n\theta_n = \kappa_n q^{-1/\kappa_n} \to \infty,$$

$$\mathbb{V}X_n = \kappa_n\theta_n^2 = \kappa_n q^{-2/\kappa_n} \to \infty.$$

The sequence of random variables X_n converges in distribution to the following discrete random variable X with the support consisting of just two points:

$$\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = +\infty) = q.$$

To show this we take an arbitrary positive number x_0 and calculate

$$\lim_{n\to\infty} \mathbb{P}\left(X_n \in [0, x_0]\right) = \lim_{n\to\infty} \int_0^{x_0} f_n(x;\kappa_n)\, dx, \tag{E.1}$$

where $f_n(x;\kappa_n) = \frac{q}{\Gamma(\kappa_n)} x^{\kappa_n - 1} e^{-xq^{1/\kappa_n}}$. Note, however, that $e^{-xq^{1/\kappa_n}}$ converges uniformly to 1 as $n \to \infty$ on any closed interval $[0, x_0]$, and therefore

$$\lim_{n\to\infty} \mathbb{P}\left(X_n \in [0, x_0]\right) = \lim_{n\to\infty} \frac{q x_0^{\kappa_n}}{\Gamma(\kappa_n + 1)} \to q. \tag{E.2}$$

Hence the limit (E.1) does not depend on x_0. It follows that the limiting cumulative distribution function corresponds to an improper distribution with the mass q at $x = 0$ and mass $1 - q$ at $x = +\infty$.

Alternatively, this result can be proven using the series expansion for the incomplete gamma function $\gamma(x;\kappa) = \int_0^x e^{-t} t^{\kappa-1}\, dt$ (Bowman and Shenton (1988)):

$$\gamma(x;\kappa) = x^\kappa \left(\frac{1}{\kappa} - \frac{x}{1!(\kappa+1)} + \frac{x^2}{2!(\kappa+2)} - \dots \right), \tag{E.3}$$

so that when $\kappa_n \to 0$

$$\mathbb{P}(X_n \le x_0) = \frac{\gamma\left(x_0/\theta_n;\kappa_n\right)}{\Gamma(\kappa_n)} \tag{E.4}$$

behaves as follows:

$$\mathbb{P}(X_n \le x_0) \sim \frac{\left(\frac{x_0}{q^{-1/\kappa_n}}\right)^{\kappa_n}}{\Gamma(\kappa_n)} \frac{1}{\kappa_n} = \frac{q x_0^{\kappa_n}}{\Gamma(\kappa_n + 1)} \to q,\ n \to \infty.$$

The last limit does not depend on the chosen $x_0 > 0$.