SCALABLE ALGORITHMS FOR DISTRIBUTED STATISTICAL INFERENCE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Animashree Anandkumar

August 2009

SCALABLE ALGORITHMS FOR DISTRIBUTED STATISTICAL INFERENCE

Animashree Anandkumar, Ph.D.

Cornell University 2009

The classical framework on distributed inference considers a set of nodes taking measurements and a fusion center making the final decision on the underlying phenomenon, without dealing with the issue of transporting the measurements to the fusion center. Such an approach introduces significant overhead in communication. Communicating all the raw data for inference is not scalable: in this case, the per-node average energy consumption and the total bandwidth requirement become unbounded as the network grows.

We design scalable algorithms for two scenarios with guarantees for inference whose communication requirements and complexity are bounded even as the network grows. This is achieved through distributed computation of a sufficient statistic, which results in reduction of data dimensionality while ensuring no loss in inference accuracy at the fusion center. The first scenario deals with multihop routing and fusion of spatially correlated measurements, incorporated through a Markov random field model. The second scenario deals with design of medium-access control (MAC) with the aim of computing a sufficient statistic for inference over a multiple access channel.

BIOGRAPHICAL SKETCH

Animashree Anandkumar hails from a family of engineers and teachers, and spent her childhood in the city of Mysore in southern India. She received her B.Tech in Electrical Engineering from the Indian Institute of Technology Madras in 2004 with a minor in Theoretical Computer Science. She is a PhD student in Electrical Engineering at Cornell University with a minor in Applied Mathematics. Since Fall 2004, she has been working with the Adaptive Communications and Signal Processing (ACSP) group under the direction of Prof. Lang Tong. She spent two summers at IBM Watson Research, Hawthorne, NY working on transaction monitoring in financial systems. She is currently visiting the Stochastic Systems Group (SSG) at MIT hosted by Prof. Alan Willsky.

Anima received the 2008 IEEE Signal Processing Society (SPS) Young Author award for her paper co-authored with Lang Tong appearing in the IEEE Transactions on Signal Processing. The award is given to the author of a meritorious paper under the age of thirty appearing in the transactions in a three-year window. She is the recipient of the Fran Allen IBM Ph.D fellowship for the year 2008-09, presented annually to one female Ph.D. student in conjunction with the IBM Ph.D. Fellowship Award. She was named a finalist for the Google Anita-Borg Scholarship 2007-08. She received the Student Paper Award at the 2006 International Conference on Acoustic, Speech and Signal Processing (ICASSP). Her research interests are in the area of statistical-signal processing, information theory and networking. Specifically, she has been worked on distributed inference and learning of graphical models, routing and random-access schemes, error exponents and queueing models.

Anima got a chance to explore a variety of hobbies and interests during her stay at Cornell. The diverse student community at Cornell enriched her experi-

ence. She loves to dance and is trained in Bharatnatyam (Indian classical dance), middle eastern and Latin dances. The gorgeous outdoors offered excellent opportunities for hiking, kayaking, snowboarding, ice climbing, and so on. Anima loves to travel and her memorable experiences include the Macchu Pichu Inca trail, fishing for piranhas in the Amazon, torrential downpour in the Sahara desert, and living in a reed island on Lake Titicaca.

This document is dedicated to my dad whose ideas continue to inspire me and to my mom for her unconditional support.

ACKNOWLEDGEMENTS

I want to first and foremost thank my advisor Lang Tong for his continued support and guidance. He has been a great mentor to me and I have learnt a lot on how to formulate, solve and present my research. These skills will no doubt be a tremendous asset to me for years to come. His enthusiastic outlook to research and his boundless energy continues to inspire me. His approach to teaching and mentoring students have become my benchmark as I plan to pursue a career in academia.

I would also like to thank my other committee members Aaron Wagner, Kevin Tang, David Williamson and Ananthram Swami for collaborating with me on different research problems and providing me with valuable research ideas. I want to thank Toby Berger for previously being on my committee and for encouraging me to work on challenging problems. I want to thank Sheila Hemami, Rajit Manohar, Sunil Bhave, Anna Scaglione, Michal Lipson, Cliff Pollock, Jose Martinez, Alyssa Apsel and the late Sergio Servetto for being excellent role models and for pro-actively interacting with students.

I want to thank Joe Yukich, Ananthram Swami and Anthony Ephremides for highly fruitful collaborations. I had a great time interning with the networking technologies group at IBM Watson research over the summers of 2007 and 2008, where I worked with Chatschik Bisdikian, Dakshi Agrawal and Ting He. I want to thank Alan Willsky for hosting me here at MIT since Fall 2008. It has been a pleasure working with him and his group, and I look forward to being here as a post-doctoral fellow. I also want to thank John Fisher, Devavrat Shah, David Gamarnik and Munther Dahleh for on-going highly fruitful research discussions.

Through my graduate life, I have had the privilege of interacting with many fellow students. During my first years, Youngchul, Parv, Vidyut, Ting and Zhiyu helped me learn the ropes. My officemates at various points in time Tae-Eung, Chin Chen, Saswat, Matt, Jin Sub and Brandon were wonderful company. I had the pleasure of interacting with other current and former ACSP members Abhishek, Oliver, Stefan, Meng, Aaron, Amine, Pinto, John and Gael. My friends at Cornell Ania, Meisha, Muthu, LP, Smita, Gerri, Niousha, Richa, Bettina, among others, were great people to hang out with. Lastly, my colleagues here at MIT, Emily, Kush, Matt, Vincent, Jin, Venkat, Mike, Ying, Ishai, Lavanya and Mythili instantly made me feel at home.

I want to thank my dad, mom and my brother Amod for being a constant source of support and for reading numerous drafts of my papers and presentations. I also want to thank my extended family here for their continued support.

This work was supported by the collaborative participation in Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011 and by the Army Research Office under Grant ARO-W911NF-06-1-0346 and by the IBM PhD fellowship 2008-09.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . v
Acknowledgements . . . vi
Table of Contents . . . viii
List of Figures . . . xi

1 Introduction **1**
1.1 Multihop In-network Processing . . . 2
1.2 Medium-Access Control . . . 3
1.3 Related Work on In-network Processing for Inference . . . 5
1.4 Related Work on MAC Design for Statistical Inference . . . 8
1.5 Organization of Thesis . . . 9

2 Spatial Data Correlation Model **10**
2.1 Literature on Spatial Correlation Models . . . 10
2.2 Definition and Properties of MRF . . . 12
2.3 Statistical Inference of Markov Random Fields . . . 19
2.A Proofs . . . 22

3 Minimum Cost In-Network Processing for Optimal Inference **26**
3.1 Network and Cost Model . . . 28
3.2 Formulation of Minimum Cost Fusion . . . 31
3.2.1 Local Processor Assignment . . . 33
3.2.2 0-1 Integer Programming Formulation . . . 35
3.3 Special Case: IID Measurements . . . 37
3.3.1 Data Fusion for Markov Random Fields (DFMRF) Scheme 38
3.4 Steiner Tree Reduction . . . 41
3.4.1 Simplified Integer Program . . . 41
3.4.2 Examples . . . 47
3.5 Simulation Results . . . 48
3.6 Extension to M-ary Hypothesis Testing . . . 50
3.7 Conclusion . . . 54
3.A Overview of Steiner Trees and Approximation Algorithms . . . 54
3.B Proof of Lemma 1 . . . 57

4 Cost-Performance Tradeoff For Inference **59**
4.0.1 Problem Formulation . . . 60
4.0.2 Preliminary Observations & Results . . . 63
4.1 IID Measurements . . . 64
4.2 Correlated Measurements: MRF Model . . . 67
4.2.1 In-network Aggregation of LLR . . . 68
4.2.2 Error Exponent & Penalty Function . . . 69

4.2.3 Special Case of MRF: Disjoint Cliques . . . 70
4.3 Node Selection Heuristics . . . 75
4.4 Numerical Analysis . . . 80
4.4.1 Simulation Environment . . . 80
4.4.2 Results: IID Measurements . . . 81
4.4.3 Results: Correlated Measurements . . . 83
4.A Proofs . . . 85

5 Energy Scaling Laws for Optimal Inference in Random Networks 94
5.1 Scalable data fusion . . . 94
5.2 Summary of Scaling Results . . . 97
5.3 Random Fusion Network Model . . . 99
5.3.1 Stochastic model of sensor locations . . . 99
5.3.2 Random dependency graphs . . . 100
5.3.3 Cost Model . . . 101
5.4 Energy Scaling Laws . . . 101
5.4.1 Energy scaling for optimal fusion: independent case . . . 101
5.4.2 Energy scaling for optimal fusion: MRF case . . . 105
5.5 Numerical Illustrations . . . 114
5.A Functionals on random points sets . . . 119
5.B Proofs . . . 122

6 Inference Accuracy Scaling Laws for Random Networks 124
6.0.1 Additional Assumptions on the Inference Model . . . 125
6.1 Error Exponent as a Graph Functional . . . 126
6.1.1 Acyclic Dependency Graphs . . . 127
6.2 Detection Error Exponent . . . 128
6.2.1 Testing Against Independence . . . 128
6.2.2 General Hypothesis Testing . . . 130
6.3 Gaussian Distribution on Acyclic Graphs . . . 133
6.3.1 1-Nearest Neighbor Dependency . . . 136
6.3.2 Numerical Results . . . 141
6.A Proofs . . . 142

7 Energy-Constrained Inference in Random Networks 146
7.1 Overview of Results and Approach . . . 147
7.2 Optimal Node Density . . . 148
7.2.1 Detection Error Exponent . . . 149
7.2.2 Feasible Node Density Set . . . 152
7.2.3 Infinite Node Density . . . 155
7.2.4 Threshold Behavior of Optimal Density . . . 156
7.2.5 Numerical Analysis . . . 160
7.2.6 Sensitivity Analysis . . . 161
7.3 Energy-density Constraint . . . 163

7.4 Conclusions . . . 167
7.A Proofs . . . 167

8 Medium Access Design for Statistical Inference **170**
8.0.1 Summary of main results . . . 172
8.1 Model and problem formulation . . . 176
8.1.1 TBRA: Transceiver and sufficient statistics . . . 176
8.1.2 Spatio-temporal tradeoff and problem formulation . . . 179
8.2 Optimal type-based random access . . . 180
8.2.1 Optimal non-coherent detector . . . 181
8.2.2 Detection-error exponents . . . 181
8.2.3 Asymptotic distribution and Gaussian approximation . . . 184
8.3 Optimal TBRA for Parameter Estimation . . . 191
8.3.1 Bayesian Cramér-Rao Bound . . . 191
8.3.2 Critical Coherence Index γ_* . . . 192
8.4 Numerical results and simulations . . . 195
8.4.1 Evaluation of error exponents . . . 196
8.4.2 Performance and discussion . . . 198
8.4.3 Numerical Analysis for Parameter Estimation . . . 201
8.5 Conclusions . . . 201
8.A Proofs . . . 202

9 Outlook for the Future **214**
9.1 Scalable Learning & Inference of High-Dimensional Data . . . 215
9.2 Cost-Performance Tradeoff for Inference & Learning . . . 216
9.3 Relevance . . . 216

Bibliography **218**

LIST OF FIGURES

2.1 Linear dependency graph for first-order AR process. 13
2.2 Global Markov Property: $\mathbf{Y}_A \perp\!\!\!\perp \mathbf{Y}_B|\mathbf{Y}_C$. 15
2.3 Fill pattern of potential matrix is same as the dependency graph. 17

3.1 Fusion policy digraph $\mathcal{F}_\Upsilon$: each edge routes one real number. . . 30
3.2 Fusion policy DFMRF for inference at fusion center. 31
3.3 Inputs to the problem of minimum cost fusion for inference. . . . 34
3.4 Cut set S separating set of processors and fusion center. 35
3.5 The optimal fusion graph DMST for independent observations. . 39
3.6 Data fusion for Markov random fields (DFMRF) policy. 40
3.7 δ-approximate fusion policy Υ (AggApprox). 41
3.8 Map-All($\bar{N}_g; \mathcal{E}^{\text{SP}}, C$) adds virtual nodes for each non-trivial clique. 43
3.9 RevMap-All($G; \mathbf{v}_c, \mathbf{v}, C$) maps tree G to fusion scheme Υ. 44
3.10 Example 1: Minimum cost fusion for chain dependency graph. . 45
3.11 Example 2: Minimum cost fusion for given dependency graph. . 47
3.12 Simulation results for k nearest-neighbor dependency graphs. . . 50
3.13 δ-approximate fusion Υ for M-ary hypothesis (AggApprox_M). . 54

4.1 Aggregation of i.i.d. measurements along the PCST. 64
4.2 Computation of the log-likelihood ratio $L(\mathbf{Y}_{\mathbf{v}_s})$ for $\mathbf{v}_s \subset \mathbf{v}$. 67
4.3 Clique selection and fusion via PCST reduction for binary cliques. 71
4.4 Approx_PCDF(Map-Sub,Algo) selects groups $\mathcal{M}_s$ and policy Υ. . 75
4.5 Map-Sub($N_g; \mathcal{M}, \mathcal{E}^{\text{SP}}, \Pi, \gamma$) adds nodes for each non-trivial group. 76
4.6 RevMap-Sub($G'; \mathbf{v}_c, \mathbf{v}, \mathcal{M}$) selects groups $\mathcal{M}_s$ and policy Υ. 77
4.7 Objective Value obj vs. Tradeoff Factor under Uniform Placement 78
4.8 % of Nodes Not Selected vs. Tradeoff Factor, Uniform Placement 79
4.9 Cost vs. Performance for Selected Set $\mathbf{v}_s$ under Uniform Placement 79
4.10 Samples of i.i.d uniform placement and Matern cluster process. . 80
4.11 Group Selection for Gaussian data, 60 runs, $\delta = 1.2, n = 50$ 81
4.12 Component Selection for Gaussian data, 60 runs, $\gamma = 140, n = 200$ 81
4.13 Component Selection for Gaussian data, 60 runs, $\gamma = 140, n = 200$ 82

5.1 Ratio of energy under node pdf τ and uniform pdf. 104
5.2 Avg. energy, k-NNG dependency, $\tau \equiv 1$, 500 runs, $\nu = 2$. 114
5.3 Approx. ratio, k-NNG dependency, $\tau \equiv 1$, 500 runs, $\nu = 2$. 114
5.4 Approx. ratio, k-NNG dependency, $\tau \equiv 1$, 500 runs, $n = 190$. . . . 115
5.5 Disk Dependency graph, $\nu = 2$, uniform ($\tau \equiv 1$). 115
5.6 Disk Dependency graph, radius $\delta = 0$, $n = 190$. 116
5.7 Disk Dependency graph, radius δ, $\nu = 4$. 116
5.8 Sample realization of $n = 190$ points on unit square under pdf τ. . 117
5.9 Avg. energy for shortest-path routing. 500 runs and $n = 190$. . . . 118
5.10 LLN for sum graph edges on uniform point sets ($\tau \equiv 1$). 120

6.1 Directed & undirected versions of nearest-neighbor graph. 137
6.2 Error exponent $\mathcal{D}$ vs. ratio of variances K, node density $\lambda = 1$. . . 138
6.3 Event that the origin is a biroot in the directed 1-NNG. 143

7.1 Optimal node density vs. variance ratio K. See Theorem 21. . . . 147
7.2 Error exponent $\mathcal{D}$ vs. variance ratio K. 153
7.3 Trend under correlation $\rho(R) = M \exp[-aR]$, $M = 0.5, a = 1$. 153
7.4 Optimal density λ_* decided by threshold $K_t(M)$. 158
7.5 Error exponent vs. total energy. $\nu = 2, C_p = c_1(2), M = 0.6$. 166

8.1 Distributed detection over multi-access fading channel. 171
8.2 Performance M_λ under transmission rate λ and coherence index γ. 175
8.3 Optimal transmission rate λ_* under coherence index γ. 192
8.4 Performance metric (SNR= -5dB, $\sigma_H^2 = 1$, $\Theta \sim \Delta(0.2, 0.8)$). 194
8.5 Gaussian-approximation (SNR= 0dB, $\sigma_H^2 = 1$, $\Theta \sim \Delta(0.2, 0.8)$). . . 194
8.6 Error exponent vs. transmission rate. 197
8.7 Gaussian approximation. 197
8.8 Error probability vs. transmission rate, non-zero mean fading. . . 199
8.9 Error probability vs. transmission rate, zero-mean fading. 200

LIST OF SYMBOLS

$E_{\mathcal{G}}$	Set of edges of graph $\mathcal{G}$	12
(i, j)	Undirected edge between i and j	12
$\text{nbd}(i; \mathcal{G})$	Neighborhood of node i in graph $\mathcal{G}$	12
$\text{Deg}(i)$	Degree of node i	12
$\lvert A \rvert$	Cardinality of set A	12
$\lvert \mathbf{A} \rvert$	Determinant of matrix $\mathbf{A}$	12
$\mathbf{Y}_{\mathbf{v}}$	Random measurement vector at positions $\mathbf{v}$	13
$\perp$	Statistical independence	13
$L(\mathbf{Y}_{\mathbf{v}})$	Log-likelihood ratio for binary hypothesis	20
$\mathcal{C}$	Clique set for LLR computation	21
ϕ_c	Potential function for clique c	21
$\mathcal{E}_{i,j}$	Cost of routing from i to j	28
N_g	Network Graph for Communication	28
$\bar{N}_g$	Metric closure of Network Graph	28
$\mathcal{E}(N_g)$	Total energy for using links in N_g	29
$\langle i, j \rangle$	A directed edge from i to j	29
$\text{Proc}(c)$	Processor node for clique c	32
FG	Forwarding subgraph for LLR computation	32
AG	Aggregation subgraph for LLR computation	32
$\mathcal{F}_{\Upsilon}$	Fusion policy digraph	32
Υ^*	Minimum cost fusion policy	34
$\mathbf{L}(\mathbf{Y}_{\mathbf{v}})$	Log-likelihood ratio vector for M-ary hypothesis testing	50
$\mathcal{E}^{\text{SP}}(i, j; N_g)$	Cost of shortest-path routing from i to j using links in N_g	56
opt_tradeoff	Optimal Cost-Performance Tradeoff	61
π	Penalty for non-selection	61
γ	Cost-Performance Tradeoff Parameter	62
obj	Objective Function for Tradeoff	62
$\mathcal{D}$	Detection Error Exponent	63
D	Kullback Leibler Distance	64
π^{iid}	Penalty Function Under IID Measurements	65
ν	Path-loss Exponent	94
$\bar{\mathcal{E}}(\Upsilon_n(\mathbf{V}_n))$	Avg. Energy Under Policy Υ	95
$\Upsilon^*(\mathbf{V}_n)$	Min. Energy Fusion Policy	97
$\mathfrak{A}$	Set of Lossless Fusion Policies	97
λ	Sensor Node Density	99
Q_1	Unit Area Square Around Origin	99
τ	PDF for IID Node Placement	99
$\mathcal{P}_a$	Homogeneous Poisson Distribution with Intensity a	100
$\zeta(\nu; \text{MST})$	Scaling constant for ν-weighted MST Edges	102
$C_g(\mathbf{V}_n)$	Complete Graph Over Nodes $\mathbf{V}_n$	106

CHAPTER 1

INTRODUCTION

We are living in an increasingly networked world with networks of varying scales: the nodes in the network can comprise of billions of tiny devices, our personal mobile gadgets, or even our friends. The nature of links is also varied; they can be wireless, wire-line, or social links. There is rich interaction and information flow between these networks - for instance, between the computer and the social networks. So far, these different networks have been mostly studied as independent entities.

Another feature of these networks is the massive scale of the data they generate. Analysis of such large data sets requires scalable algorithms whose computational complexity does not grow with data. Moreover, since data is generated at a large number of nodes, the communication requirements of an algorithm is a key parameter. Depending on the application, algorithms need to undertake distributed computations at various nodes for communication requirements to be scalable in the data size and in the number of nodes in the network.

Many network applications involve collaborative processing of network data. For instance, in distributed statistical inference, the goal is to reach a decision about some common underlying phenomenon. Examples include intrusion detection, anomaly detection, temperature field estimation, and so on. We consider distributed inference where nodes communicate their data to a more powerful decision node called the fusion center, which then makes the final decision. We explicitly model the costs and constraints (e.g., energy, bandwidth) posed by the communication network to move data to the fusion center for inference.

If the nodes were to communicate all their raw data to the fusion centers, then such a scheme has a high communication cost, and is not scalable in the network size. However, if the end goal is inference, there is no need to communicate all the raw data; instead, we should compute and communicate a *sufficient statistic*, a function of the raw data, which ensures that there is no loss in inference accuracy at the fusion center. At the same time, the sufficient statistic has dimensionality reduction resulting in savings of communication costs.

We look at two scenarios for distributed computation of the sufficient statistic. In the first, we consider multi-hop routing with energy constraints, and develop in-network processing schemes for inference. In the second, we consider random access over a multiple access channel with energy and bandwidth constraints, and develop channel-aided computation schemes. These schemes are instances of cross-layer optimization, where we exploit the inference application through the sufficient statistic to optimize routing and medium-access control (MAC). Below, I present my thesis research on scalable algorithms for distributed inference, based on the works in [1–9].

1.1 Multihop In-network Processing

Dependency graph is an effective model for describing relationships between nodes in a network based on some attribute, and needs to be inferred from the data generated by the nodes. For inference of the correct dependency graph model, the sufficient statistic has a compact form based on local dependency graph properties. In [1–3], we propose schemes for distributed computation of the sufficient statistic by exploiting the dependency graph structure.

Our scheme is scalable - it has strictly bounded average communication costs, even as the network grows, for a wide range of dependency graph models. Intuitively, when the dependency graph has only short-range edges between nearby nodes, the computation of the sufficient statistic can be undertaken locally with low communication costs. We provide a precise definition of such local dependency graphs based the concept of graph *stabilization* using the recent results on random geometric graphs. Such local dependency graphs occur in many scenarios - for example, the dependency between the location-based search queries and internet users; users near a particular location are more likely to query about that location than the ones further away. Another example is a sensor network measuring temperature of a field where nearby sensors tend to record similar temperatures.

We also provide a closed-form expression for average communication cost for inference under our scheme, and it has a nice representation in terms of the dependency graph, signal attenuation model and node placement. We use the expression to design efficient node placement strategies with low communication costs in [4]. We also address the related issue of selecting informative nodes for inference (sub-sampling) in [6] to further reduce the communication costs.

1.2 Medium-Access Control

We consider medium-access control (MAC) schemes for communication between the nodes in a network and the fusion center in [7–9]; the end goal is inference about a common underlying phenomenon measured by the nodes.

Traditionally, MAC schemes allocate transmission from different nodes to or-

thogonal channels (such as in time or frequency) to avoid interference. Instead, we propose a MAC scheme where nodes may interfere with one another, yet achieve good inference accuracy in the end. We allocate orthogonal channels to data levels: all nodes reaching the same local decision use the same orthogonal channel to transmit, if they decide to do so. This is an instance of channel-aided computation where we use the multiple access channel to compute a noisy histogram or the *type* of the local decisions, which serves as the sufficient statistic for inference. The bandwidth requirement of this scheme is independent of the number of transmitting nodes, and is hence, scalable for large networks.

The extent to which interference aids inference depends on the nature of the multiple-access channel. Coherent channels add energy of the interfering signals more efficiently than canceling channels, and we quantify this behavior of the fading channels through a compact parameter, called the channel *coherence index*.

If the channel is canceling, then in our scheme, transmissions on independent orthogonal channels are more likely in order to avoid any interference. On the other hand, if the channel is coherent, simultaneous transmissions are more likely in our scheme. More specifically, we establish that for low coherence-index channels, our scheme has a finite optimal rate which maximizes inference performance. A sharp contrast is the extreme case when the channel is fully coherent (no random fading). In this case, we prove that the optimal rate is unbounded, which means that there should be simultaneous transmissions, in order to exploit the channel coherency.

Hence, our scheme adapts medium-access control based on the channel conditions to maximize inference performance, and it outperforms the classical or-

thogonal transmission scheme in terms of inference accuracy and bandwidth efficiency under the same energy budget.

1.3 Related Work on In-network Processing for Inference

The seminal work of Gupta and Kumar [10] on the capacity of wireless networks has stimulated extensive studies covering a broad range of networking problems with different performance metrics. See also [11]. Here, we limit ourselves to the related works on energy consumption and data fusion for statistical inference.

Results on scaling laws for energy consumption are limited. In [12], energy scaling laws for multihop wireless networks (without any data fusion) are derived under different routing strategies. The issue of node placement for desirable energy scaling has been considered in [13, 14], where it is argued that uniform node placement, routinely considered in the literature, has poor energy performance when there is no data fusion. It is interesting to note that, for fusion networks, uniform sensor distribution is in fact optimal among a general class of distributions. See Chapter 5.

Energy-efficient data fusion has received a great deal of attention over the past decade. See a few recent surveys in [15,16]. It has been recognized that sensor observations tend to be correlated, and that correlations should be exploited through data fusion. One line of approach is the use of distributed compression with the aim of reconstructing all the measurements at the fusion center. Examples of such approaches can be found in [17–19].

While sending data from all sensors to the fusion center is certainly sufficient to ensure optimal inference, it is not necessary. More relevant to our work on in-network processing for inference is the idea of data aggregation, e.g., [20–22]. Finding aggregation policies for correlated data, however, is nontrivial; it depends on the specific applications for which the sensor network is designed. Perhaps a more precise notion of aggregation is in-network function computation where certain functions are computed by passing intermediate values among nodes [23–26]. However, these works are mostly concerned with computing symmetric functions such as the sum function, which in general, do not satisfy the constraint of optimal statistical inference at the fusion center.

In the context of statistical inference using wireless sensor networks, the idea of aggregation and in-network processing has been explored by several authors. See [27–33]. Most relevant to our work are [27–31, 34] where the Markovian correlation structures of sensor measurements are exploited explicitly. These results mostly deal with one-dimensional node placements, and do not deal with randomly placed nodes or energy scaling laws.

We also consider sub-sampling of the sensor field to achieve tradeoff between fusion costs and quality of inference in Chapter 4. Sensor selection algorithms have been considered in a variety of contexts, such as for control [35], for target tracking [36], multimedia streams [37], fixed number selection [38], region selection [39], for information maximization [40], in dynamical systems [41, 42], and so on. However, to the best of our knowledge, the problem of optimal node selection (e.g., see survey [43]) has not been considered in conjunction with in-network fusion before. Indeed in single-hop networks, there is no need for data fusion. But most large networks are multi-hop, and routing costs are substan-

tially reduced through fusion at intermediate nodes, as seen in simulations in Section 4.4. Many works on node selection assume perfect sensing of a region (e.g., [39]). In contrast, our result in Chapter 4 explicitly models correlated imprecise measurements via a Markov random field, and is the basis for selecting "informative" sensors for inference.

In Chapter 6, we tackle the related issue of performance of sensor networks, in terms of inference accuracy in large networks. When inference accuracy decays exponentially with the sample size, the rate of decay is given by the error exponent. In Chapter 6, we derive error exponents for hypothesis testing of Markov random fields. In this context, we list some related work. The large-deviation analysis for the test of simple hypotheses with general distributions exists [44, 45], but closed-form expressions are possible only for certain cases. such an analysis for homogeneous Gauss-Markov random fields on lattices have been considered by Sung et al in [31, 46]. However, their techniques are not easily generalized to arbitrarily placed nodes with spatially-dependent fields, considered here. In [47], an expression for the Kullback-Leibler (KL) divergence rate is derived when the two distributions are Markov chains of arbitrary order, which is a special case of the formulation here.

The scaling laws for energy consumption and inference accuracy derived in this thesis rely heavily on several results on the law of large numbers for geometric random graphs. We have extensively borrowed the formulations and techniques of Penrose and Yukich [48,49]. See Appendix 5.A for a brief description and [50–52] for detailed expositions of these ideas.

1.4 Related Work on MAC Design for Statistical Inference

In Chapter 8, we investigate the design of medium access schemes for statistical inference over fading multiple access channel. In this context, the earlier results on classical distributed detection [53–55]. assume perfect channels between sensors and the fusion center. In the context of power and bandwidth-constrained wireless sensor networks, Chamberland and Veeravalli used large-deviation techniques for the optimal design of local quantization rules [56, 57]. See also Aldosari and Moura [58]. We too use large-deviation techniques, but for the design of multi-access communications. Distributed detection in the presence of channel fading is considered in [59,60], where each user has a dedicated channel to the fusion center.

The problem of distributed detection on multi-access channels are more recent [61–64]. The transmission scheme used is the so-called type-based multiple access (TBMA) proposed independently by Mergen and Tong [61,64] and by Liu and Sayeed [62]. The positive result of TBMA is that when there is no fading, the asymptotic performance of TBMA (as the number of sensors approaches infinity) is same as that when the fusion center has direct access to sensor observations. The negative result, however, is that when the channel has zero-mean fading, TBMA fails to be consistent for a single data collection. Furthermore, these results apply only for a fixed number of sensor. In [9], we proposed type-based random access (TBRA) as a multi-access scheme for non-zero mean fading channels, incorporating random number of sensors. We used large-deviation approaches and compared the detection error exponents of TBRA and TBMA for non-zero mean fading channels.

The TBRA used here in Chapter 8 differs from the other existing approaches in several significant aspects. TBRA allows the fading channels to have zero mean and detectors to be non-coherent. This scenario is relevant since it may be difficult to estimate a large number of fading coefficients at the receiver. Also, it may be difficult to synchronize transmissions among geographically distributed nodes to achieve phase coherency at the receiver. By having the expected number of transmissions ρ go to infinity, the exponential decay of error probabilities is achieved. Under the formulation of this chapter, the large-deviation approaches considered in [9, 61, 62, 64] are not applicable.

1.5 Organization of Thesis

The thesis is organized as follows: in Chapter 2, we introduce spatial data correlation model of Markov random field. In Chapter 3, we exploit this model to analyze fusion schemes for optimal inference with minimum routing cost. In Chapter 4, we allow fusion schemes to incorporate sub-sampling to achieve optimal cost-performance tradeoff for inference. In Chapter 5, we build on the results of Chapter 3 and derive energy scaling laws for optimal inference in random networks. In Chapter 6, we provide inference accuracy scaling laws in random networks (error-exponent analysis). In Chapter 7, we unify the results of Chapters 5 and 6 to obtain optimal node density for energy-constrained inference in random networks. In Chapter 8, we consider a different problem on medium access design for statistical inference. Chapter 9 concludes the thesis and proposes some extensions to be pursued in future.

CHAPTER 2

SPATIAL DATA CORRELATION MODEL

In many realistic scenarios the sensor measurements are correlated, and our framework takes this into account. Examples of correlated signals include temperature and humidity sensors, and magnetometric sensors tracking a moving vehicle. Acoustic data are rich in spatial correlations due to the presence of echoes caused by multipath reflections. We use a *Markov random field* (MRF) model which incorporates correlation in terms of a graph, known as the *dependency graph*. The model for spatial data correlation crucially affects in-network processing and fusion policies. Various assumptions on correlation have been made in the literature. Before we describe the actual MRF model, we describe some spatial correlation models considered in the literature.

2.1 Literature on Spatial Correlation Models

Joint-Gaussian distributions and distance-based correlation function are widely assumed in the literature due to their simplicity [65–68]. Alternatively, diffusion-based [69] and joint-entropy based models [70] have also been employed. The use of remote-sensing data, proposed in [71], may not meet the resolution requirements. The model proposed in [72] is a special case of a Markov random field (MRF).

Markov random fields, as a class of parametric models for spatial data, were introduced by Besag [73, 74], and were known as conditional auto-regressions in his works. Prior to these works, Hammersley and Clifford formulated their now famous theorem on the equivalence of MRF to a Gibbs field [75]. However,

the manuscript was never published, and a sketch of the original proof can be found in [76], along with further discussion on the historical aspects of research on MRF.

The use of the MRF model for spatial data in sensor networks is relatively new (e.g., [77]), although it is widely used in image processing [78] and geostatistics [79]. This could be due to the complexity of the model for arbitrarily-placed nodes. We will see that the use of a Markov random field model leads to the formation of "clusters" that are based on the statistical dependence, rather than other considerations such as residual energy [80,81]. The notion of clustering has been used extensively in sensor networks, where nodes send their data to one member of the cluster, which then processes and forwards to the destination. However, here, the issues are complicated by the fact that measurements processed in these statistical "clusters" have to be further aggregated rather than simply being forwarded to the destination.

In general, spatial signals are acausal in contrast to causal temporal signals. In the literature, the two are usually distinguished by referring to acausal signals as random fields and to causal signals as random processes. An example of exploiting correlation in a causal propagation setting can be found in [27,82].

We assume that all the sensors know the Markov random field model. In practice, the dependency structure and the model parameters of the Markov random field model can be estimated by incorporating a training phase. The seminal work of Chow and Liu in [83] considers the problem of approximating an unknown distribution from its samples using a procedure for learning the tree model that maximizes the likelihood of the training samples among the set of all possible tree models . Recently, learning graphical models from data sam-

ples specifically for binary hypothesis testing has been considered in [84]. Their procedure learns each hypothesis model from both sets of training samples.

In this chapter, we employ the Markov random field model for spatial correlation, taking into account only its graphical dependency structure; but no parametric correlation function is assumed. Moreover, any general random field without special properties can be represented as a MRF with a complete dependency graph (called *saturated models* [85]).

2.2 Definition and Properties of MRF

An undirected graph $\mathcal{G}$ is a tuple $\mathcal{G} = (\mathbf{v}, E_{\mathcal{G}})$, where $\mathbf{v}$ is the vertex set and $E_{\mathcal{G}}$= { (i, j)} is the edge set. We allow graphs to have multiple or parallel edges, but no loops. The neighborhood function nbd$(i; \mathcal{G})$ of a node i is the set of all other nodes having an edge with it in $\mathcal{G}$. Let Deg(i) denote the degree of node i. A subgraph induced by $\mathbf{v}' \subset \mathbf{v}$ on $\mathcal{G}$ is denoted by $\mathcal{G}(\mathbf{v}')$, and a complete subgraph or a clique has edges between any two nodes in $\mathbf{v}'$. A maximal clique is one that is not contained in any other clique. Throughout this chapter, a clique refers to a maximal clique, unless otherwise mentioned. For a directed graph (digraph), we denote the edges (arcs) by $< i, j >$, where the direction is from i to j, and node j belongs to the set of immediate successors of i, and i is in the set of immediate predecessor of j. The above graph functions are extended to sets, for example, (i, A) denotes the set of edges between i and members of A. For sets A and B, let $A \setminus B = \{i : i \in A, i \notin B\}$ and let $|A|$ denote cardinality of a set A. For a matrix $\mathbf{A}$, $A(i, j)$ is the element in the i^{th} row and j^{th} column and $|\mathbf{A}|$ its determinant.

The MRF falls under the framework of acausal graphical models and satisfies

conditional-independence properties, based on an undirected graph known as the *dependency graph* and is defined below.

Definition 1 (Markov Random Field) *Let* $\mathbf{Y_v} = [Y_i, i \in \mathbf{v}]^T$ *denote the random vector of measurements at positions given by set* $\mathbf{v}$. $\mathbf{Y_v}$ *is a Markov random field with an (undirected) dependency graph* $\mathcal{G} = (\mathbf{v}, E_{\mathcal{G}})$*, if* $\forall\, i \in \mathbf{v}$*,*

$$Y_i \perp \mathbf{Y}_{\mathbf{v}\setminus\{i,nbd(i)\}} | \mathbf{Y}_{nbd(i)}, \tag{2.1}$$

where $\perp$ *denotes conditional independence.*

In words, the above definition states that the value at any node, given the values at its neighbors, is conditionally independent of the rest of the network.

Example: One Dimensional MRF

A simple example is the first order auto-regressive (AR-1) process, given by

$$Y_t = A_{t-1} Y_{t-1} + \epsilon_{t-1}, \qquad Y_{t-1} \perp \epsilon_{t-1}, \quad \forall t \in \mathbf{v} = \{1, \ldots, n\}. \tag{2.2}$$

Since Y_t is conditionally independent of the past, given the measurement Y_{t-1}, we write

$$Y_t \perp Y_{1,\ldots,t-2} | Y_{t-1}, \quad 2 < t \leq n.$$

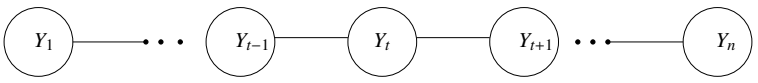


Figure 2.1: Linear dependency graph for first-order AR process.

Similarly, we can write

$$Y_{t+2,\ldots,n} \perp Y_t | Y_{t+1}, \quad 1 \le t < n.$$

This implies that

$$Y_t \perp \mathbf{Y}_{\mathbf{v}\setminus\{t-1,t,t+1\}} | \{Y_{t-1}, Y_{t+1}\}, \quad \forall t = 2, \ldots, n-1, \qquad Y_1 \perp \mathbf{Y}_{\mathbf{v}\setminus\{1,2\}} | Y_2, \quad Y_n \perp \mathbf{Y}_{\mathbf{v}\setminus\{n,n-1\}} | Y_{n-1}.$$

Hence, we have the dependency graph with neighborhood function

$$\text{nbd}(t) = \{t-1, t+1\}, \quad \text{for } t \neq 1, n, \quad \text{nbd}(1) = 2, \quad \text{nbd}(n) = n-1.$$

In other words, the dependency graph is a linear chain, as shown in Fig.2.1. Hence, the conditional independence relations of the AR-1 process have a simple graphical representation which is not apparent in (2.2). However, the dependency graph does not capture all the information of the AR-1 process, in particular, that the process is causal. On the other hand, the dependency graph can be used to model more general acausal dependencies, typically found in spatial random fields.

Properties of a general MRF

For a Markov random field, in fact, three types of Markov properties can be defined:

1. Local Markov Property: $Y_i \perp\!\!\!\perp \mathbf{Y}_{\mathbf{v}\setminus(i\cup\text{nbd}(i))} | \mathbf{Y}_{\text{nbd}(i)}$, $\forall i \in \mathbf{v}$.
2. Global Markov Property: $\mathbf{Y}_A \perp\!\!\!\perp \mathbf{Y}_B | \mathbf{Y}_C$, where A, B, C are disjoint sets. A, B are non-empty and C separates A, B. See Fig.2.2.
3. Pairwise Markov Property: $Y_i \perp\!\!\!\perp Y_j | \mathbf{Y}_{\mathbf{v}\setminus\{i,j\}} \iff (i,j) \notin E$

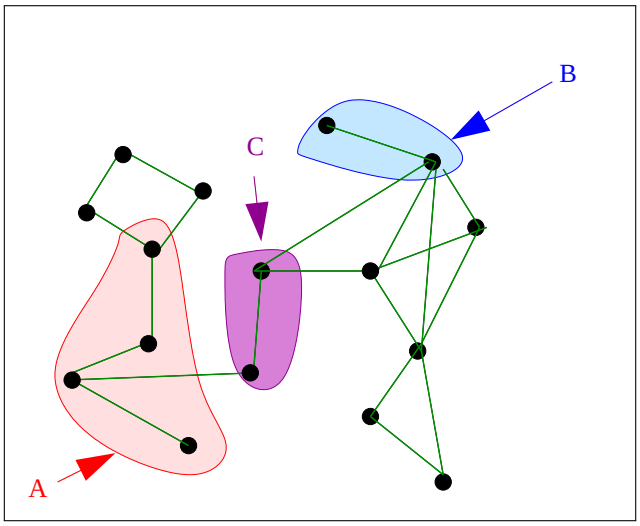


Figure 2.2: Global Markov Property: $\mathbf{Y}_A \perp\!\!\!\perp \mathbf{Y}_B | \mathbf{Y}_C$.

In definition 1, we have used the local Markov property. We can immediately see that the global Markov property implies the local Markov property, since we can set

$$A = \{i\}, B = \mathbf{v} \setminus \{i, \text{nbd}(i)\}, C = \text{nbd}(i).$$

Similarly, the global Markov property implies the pairwise Markov property, since we can set

$$A = \{i\}, B = \{j\}, C = \mathbf{v} \setminus \{i, j\}, \quad \forall (i, j) \notin E_{\mathcal{G}}.$$

The three properties can be shown to be equivalent under the positivity condition [85]. The positivity condition is as follows: for all $A \subset \mathbf{v}$ with samples $\mathbf{y}_A$, $\mathbf{y}_{\mathbf{v}\setminus A}$ such that $f(\mathbf{y}_A), f(\mathbf{y}_{\mathbf{v}\setminus A}) > 0$, the conditional is also positive

$$f(\mathbf{y}_A | \mathbf{y}_{\mathbf{v}\setminus A}) > 0,$$

where f is the density function. An equivalent condition for positivity is

$$(f(\mathbf{y}_{\mathbf{v}}) = 0) \Rightarrow (f(y_i) = 0), \quad \forall i \in \mathbf{v}.$$

An example that does *not* satisfy positivity is the fully correlated case: $Y_1 = Y_2 \ldots = Y_n$. In this case, the joint likelihood is zero whenever all the samples are not equal, but the marginal likelihood is not necessarily zero.

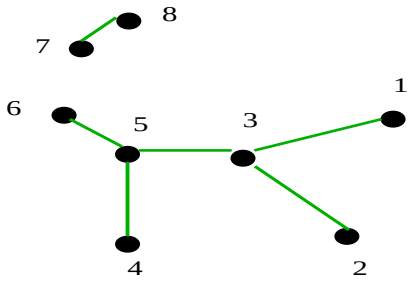
The Hammersley-Clifford theorem [76] states that for a MRF $\mathbf{Y}_\mathbf{v}$ with dependency graph $\mathcal{G} = (\mathbf{v}, E_\mathcal{G})$, the joint pdf f, under the positivity condition, can be expressed as

$$-\log f(\mathbf{Y}_\mathbf{v}|\mathcal{G}(\mathbf{v})) = \sum_{c \in C} \psi_c(\mathbf{Y}_c), \tag{2.3}$$

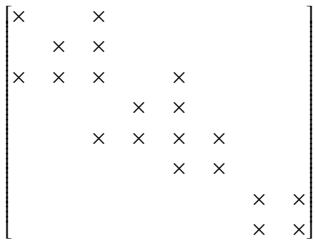
where C is a collection of (maximal) cliques in $\mathcal{G}$, the functions ψ_c, known as *clique potentials*, are real valued, non-negative and not zero everywhere on the support of $\mathbf{Y}_c$. Thus, the tuple $\Xi = \{\mathcal{G}, C, \psi\}$ specifies the MRF in (2.3). We assume that the normalization constant is already incorporated in the potential functions, in order to ensure that we have a valid pdf. For general potentials, finding the normalizing constant (called the *partition function*) is NP-hard, but approximate algorithms have been proposed in [86].

From (2.3), we see that the complexity of the likelihood function is vastly reduced for sparse dependency graphs; here, the conditional-independence relations in (2.1) results in the factorization of the joint likelihood into a product of components, each of which depends on a small set of variables. This form is already exploited by distributed algorithms such as belief propagation [87] for local inference of hidden measurements. In this chapter, we exploit the MRF model for a global inference problem, explained in Section 2.3.

In this chapter, we assume that the number of cliques $|C|$ of the MRF is polynomial in the number of nodes. This is satisfied by many graph families such as



(a) A labeled simple undirected graph.



(b) × : Non-zero potential matrix elements.

Figure 2.3: Fill pattern of potential matrix is same as the dependency graph.

bounded-degree graphs [88]. Note that in (2.3), the set of cliques C contains only those cliques with non-zero potentials. For example, for independent measurements, C is the vertex set, and we have the likelihood function as a weighted sum function,

$$-\log f(\mathbf{Y}_{\mathbf{v}}) = -\sum_{i\in\mathbf{v}} \log f_i(Y_i), \quad \mathbf{Y}_{\mathbf{v}} \sim \prod_{i\in\mathbf{v}} f_i,$$

where f_i is the marginal pdf of Y_i. Besag's auto-model [74] is a special MRF with only pairwise dependencies, and hence, the clique set C is the set of edges $E_{\mathcal{G}}$. This leads to a simplified expression for the likelihood function,

$$-\log f(\mathbf{Y}_{\mathbf{v}}; \{\mathcal{G}, E_{\mathcal{G}}, \psi\}) = \sum_{(i,j)\in E_{\mathcal{G}}} \psi_{i,j}(Y_i, Y_j). \tag{2.4}$$

Multi-parameter exponential family of conditional probabilities can be used to define such pairwise Markov random fields [89]. An example of Besag's model is the Ising Model, which was first introduced to study phase transition in ferromagnetic materials.

Gauss-Markov Random Field

The Gauss-Markov Random Field (GMRF) has some special properties. In this case, (2.3) is equivalent to (2.4), since the likelihood function of $\mathbf{Y}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is given by

$$\log f(\mathbf{Y}_\mathbf{v}; \mathbf{A}) = \frac{1}{2}(-n \log 2\pi + \log |\mathbf{A}| + \sum_{i \in \mathbf{v}} A(i,i) Y_i^2 + \sum_{i,j \in \mathbf{v}} A(i,j) Y_i Y_j), \tag{2.5}$$

where $\mathbf{A}{:=}\mathbf{\Sigma}^{-1}$ is the inverse of the covariance matrix. For a given dependency graph $\mathcal{G} = (\mathbf{v}, E_\mathcal{G})$, the GMRF should also satisfy (2.4). Hence, comparing the two equations (2.4) and (2.5), we have

$$A(i,j) = 0 \iff (i,j) \notin E_\mathcal{G}.$$

Hence, there is a one-to-one correspondence between the non-zero elements of $\mathbf{A}$ and the dependency graph edges $E_\mathcal{G}$, and is illustrated in Fig.2.3. Since $\mathbf{A}$ is associated with the potentials, it is called the *potential matrix*. Hence, for the Gaussian distribution, we only need the edges of the dependency graph and not the higher-order cliques. Moreover, for the Gaussian case, the edge potential $\psi_{i,j}(Y_i, Y_j)$ in (2.4) reduces to the sum of squares and cross-products of the measurements, weighted by the coefficients of the potential matrix $\mathbf{A}$. When the dependency graph is acyclic, we can additionally obtain a closed form for the elements of the potential matrix $\mathbf{A}$, in terms of the elements of the covariance matrix $\mathbf{\Sigma}$.

Fact 1 (GMRF with Acyclic Dependency Graph) *The coefficients of the potential matrix* $\mathbf{A}{:=}\mathbf{\Sigma}^{-1}$*, with zero mean and covariance matrix* $\mathbf{\Sigma}$ *and acyclic dependency graph* $\mathcal{G} = (\mathbf{v}, E_\mathcal{G})$*, are*

$$A(i,i) = \frac{1}{\Sigma(i,i)}\Big(1 + \sum_{j \in nbd(i)} \frac{\Sigma(i,j)^2}{\Sigma(i,i)\Sigma(j,j) - \Sigma(i,j)^2}\Big), \tag{2.6}$$

$$A(i,j) = \begin{cases} \dfrac{-\Sigma(i,j)}{\Sigma(i,i)\Sigma(j,j) - \Sigma(i,j)^2} & \text{if } (i,j) \in E_{\mathcal{G}}, \\ 0 & o.w. \end{cases} \tag{2.7}$$

The determinant of the potential matrix of $\mathbf{A}$ *is given by*

$$|\mathbf{A}| = \frac{1}{|\mathbf{\Sigma}|} = \frac{\prod_{i \in \mathbf{v}} \Sigma(i,i)^{Deg(i)-1}}{\prod\limits_{\substack{(i,j) \in E_{\mathcal{G}} \\ i<j}} [\Sigma(i,i)\Sigma(j,j) - \Sigma(i,j)^2]}. \tag{2.8}$$

Proof: See Appendix 2.A. □

In fact, for any MRF with acyclic dependency graph $\mathcal{G}$, the joint pdf $f_{\mathbf{Y_v}}$ can be expressed in terms of marginals at nodes f_{Y_i} and pairwise joint pdfs f_{Y_i,Y_j} as

$$f_{\mathbf{Y_v}}(\mathbf{y_v}) = \prod_{i \in \mathbf{v}} f_{Y_i}(y_i) \prod_{(i,j) \in E_{\mathcal{G}}} \frac{f_{Y_i,Y_j}(y_i,y_j)}{f_{Y_i}(y_i) f_{Y_j}(y_j)}. \tag{2.9}$$

See [90] for details.

2.3 Statistical Inference of Markov Random Fields

The problem of distributed detection considers a set of sensors, one of them designated as the fusion center or the decision node, and all the sensor observations are ultimately routed (in some form) to it. This setup is relevant when we need

to make a global decision on the phenomenon (contrasting to local inference algorithms such as belief propagation). We consider the binary hypothesis-testing problem with two given hypotheses, the null hypothesis $\mathcal{H}_0$ and the alternative $\mathcal{H}_1$. We limit ourselves to only simple hypothesis testing, i.e., the probability measures under both the hypotheses are known to all the sensors.

In statistical theory, a *sufficient statistic* is a well-behaved function of the data, which is as informative as the raw data for inference. Formally, a function $T(Y)$ is said to be a sufficient statistic for model P_θ, if conditioned on $T(Y)$, $Y \sim P_\theta$ does not depend on θ. It is said to be *minimal* if it is a function of every other sufficient statistic for P_θ [91]. A minimal sufficient statistic for inference represents the maximum possible reduction in dimensionality of the raw data, without destroying information about the underlying phenomenon [91]. The log-likelihood ratio (LLR) is the minimal sufficient statistic for hypothesis testing [92]. Let $f(\mathbf{Y}_\mathbf{v}; \mathcal{H}_j)$ be the pdf of the measurements $\mathbf{Y}_V$ under hypothesis j. The optimal decision rule at the fusion center is a threshold test based on the log-likelihood ratio (LLR), denoted by $L(\mathbf{Y}_\mathbf{v})$,

$$L(\mathbf{Y}_\mathbf{v}) := \log \frac{f(\mathbf{Y}_\mathbf{v}; \mathcal{H}_0)}{f(\mathbf{Y}_\mathbf{v}; \mathcal{H}_1)}. \tag{2.10}$$

The result is also true for the M-ary hypothesis testing problem, where the LLR vector

$$\left[\log \frac{f(\mathbf{Y}_\mathbf{v}; \mathcal{H}_0)}{f(\mathbf{Y}_\mathbf{v}; \mathcal{H}_1)}, \ldots, \log \frac{f(\mathbf{Y}_\mathbf{v}; \mathcal{H}_0)}{f(\mathbf{Y}_\mathbf{v}; \mathcal{H}_{M-1})}\right]^T$$

is minimally sufficient.

Form of Log-Likelihood Ratio for MRF

In this chapter, we assume that the measurement samples are drawn from distributions specified by distinct Markov random fields, defined on the same node set. In particular, we consider

$$\mathcal{H}_0 : \Xi_0 = \{\mathcal{G}_0(\mathbf{v}), C_0, \psi_0\} \quad \text{vs.} \quad \mathcal{H}_1 : \Xi_1 = \{\mathcal{G}_1(\mathbf{v}), C_1, \psi_1\}. \tag{2.11}$$

From (2.3) and (2.10), the LLR is given by the difference of the respective clique potentials,

$$L(\mathbf{Y}_\mathbf{v}) = \sum_{a \in C_1} \psi_{1,a}(\mathbf{Y}_a) - \sum_{b \in C_0} \psi_{0,b}(\mathbf{Y}_b). \tag{2.12}$$

It is easily seen that the LLR can be expressed as the sum of potentials of an "effective" Markov random field $\Xi = \{\mathcal{G}, C, \phi\}$ specified as follows: the effective dependency graph $\mathcal{G} = (\mathbf{v}, E_\mathcal{G})$, has the edge set $E_\mathcal{G} = E_{\mathcal{G}_0} \cup E_{\mathcal{G}_1}$; the effective clique set is $C = C_0 \cup C_1$, with only the resulting maximal cliques retained; the effective potential functions ϕ_c are given by

$$\phi_c(\mathbf{Y}_c) := \sum_{a \in C_1, a \subset c} \psi_1(\mathbf{Y}_a) - \sum_{b \in C_0, b \subset c} \psi_0(\mathbf{Y}_b), \quad \forall\, c \in C. \tag{2.13}$$

Therefore, the LLR has a succinct form, which will be used in the rest of this chapter,

$$L(\mathbf{Y}_{V_n}; \Xi) = \sum_{c \in C} \phi_c(\mathbf{Y}_c). \tag{2.14}$$

In order to quantify inference performance, we consider the *Neyman-Pearson* criterion [91], where for a fixed false-alarm probability (type-I error), the detector at the fusion center is optimal in terms of the type-II error probability P_M.

2.A Proofs

Proof of Theorem 1

Using the expression $\mathbf{A\Sigma} = \mathbf{I}$, we have the following identities:

$$A(i,i) \quad + \quad \sum_{j \in \text{nbd}(i)} A(i,j) \frac{\Sigma(i,j)}{\Sigma(i,i)} = \frac{1}{\Sigma(i,i)}, \tag{2.15}$$

$$A(i,i) \quad + \quad A(i,j) \frac{\Sigma(j,j)}{\Sigma(i,j)} + \sum_{\substack{k \in \text{nbd}(i) \\ k \neq j}} A(i,k) \frac{\Sigma(j,k)}{\Sigma(i,j)} = 0, \ \forall j \in \text{nbd}(i), \tag{2.16}$$

where (2.15) is obtained by the sum-product of i^{th} row and i^{th} column of $\mathbf{A}$ and $\mathbf{\Sigma}$. Similarly, (2.16) is obtained by sum-product of i^{th} row of $\mathbf{A}$ and j^{th} column of $\mathbf{\Sigma}$ and dividing by $\Sigma(i,j)$. In (2.16), by acyclicity for $k \in \text{nbd}(i)$ and $k \neq j$, we have $j \nsim k$. From MRF assumption, we have

$$\frac{\Sigma(j,k)}{\Sigma(i,j)} = \frac{\Sigma(i,k)}{\Sigma(i,i)}, \quad \forall\, j,k \in \text{nbd}(i), k \neq j.$$

Subtracting (38) from (37), only the terms with $A(i,j)$ survive and hence, we obtain $A(i,j)$. Substituting all the $A(i,j)$'s in (2.15), we obtain $A(i,i)$. Hence, all the coefficients of potential matrix $\mathbf{A}$ are given by (2.7).

Let $|\mathbf{A}^{(n)}|$ be the determinant of the potential matrix of n nodes. Assume $n > 1$, since we have $|\mathbf{A}^{(1)}| = \Sigma(1,1)^{-1}$. The determinant of the potential matrix is the product of determinants of the connected components. We therefore consider only one component $\mathcal{G}'(\mathbf{v}', E') \subseteq \mathcal{G}$. Assume $\mathcal{G}'$ has at least one edge, otherwise we have for diagonal matrix $|\mathbf{A}^{(n)}| = \prod_{i \in \mathbf{v}'} \Sigma(i,i)^{-1}$. Since $\mathcal{G}'$ is acyclic, it has a leaf, i.e., there is some vertex a with degree 1. Let b be its only neighbor. We assume the vertices have been ordered $\mathbf{v}' = \{V_1, \ldots, V_n\}$ so that $V_{n-1} = b, V_n = a$. Then $\mathbf{A}^{(n)}$ has the following form

$$\mathbf{A}^{(n)} = \begin{bmatrix} \cdot & \cdots & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdots & \cdot & \cdot & 0 \\ \cdot & \cdots & \cdot & A(n-1,n-1) & A(n-1,n) \\ 0 & \cdots & 0 & A(n,n-1) & A(n,n) \end{bmatrix},$$

where we have from (2.7),

$$\begin{aligned} A(n,n) &= \frac{\Sigma(n-1,n-1)}{[\Sigma(n,n)\Sigma(n-1,n-1) - \Sigma(n,n-1)^2]}, \\ A(n-1,n) &= \frac{-\Sigma(n,n-1)}{[\Sigma(n,n)\Sigma(n-1,n-1) - \Sigma(n,n-1)^2]}, \\ A(n-1,n-1) &= \frac{1}{\Sigma(n-1,n-1)} - A(n-1,n)\frac{\Sigma(n,n-1)}{\Sigma(n-1,n-1)} + C, \end{aligned}$$

where C represents contributions from nodes in $\mathbf{v}' \backslash V_n$ i.e., with node V_n removed, and having an edge with V_{n-1}. Multiplying the n^{th} column by

$$\frac{A(n,n-1)}{A(n,n)} = \frac{-\Sigma(n,n-1)}{\Sigma(n-1,n-1)}$$

and subtracting it from $(n-1)^{\text{th}}$ column and using the determinant rule, we have

$$|\mathbf{A}^{(n)}| = \begin{vmatrix} \cdot & \cdots & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdots & \cdot & \cdot & 0 \\ \cdot & \cdots & \cdot & A'(n-1,n-1) & A(n-1,n) \\ 0 & \cdots & 0 & 0 & A(n,n) \end{vmatrix}, \tag{2.17}$$

where

$$\begin{aligned} A'(n-1,n-1) :=& A(n-1,n-1) \\ &+ \frac{\Sigma(n,n-1)}{\Sigma(n-1,n-1)} A(n,n-1). \end{aligned} \tag{2.18}$$

Hence, we have

$$|\mathbf{A}^{(n)}| = A(n,n)|\mathbf{M}_n|, \quad \text{for} \quad n > 1,$$

where $\mathbf{M}_n$ is the minor of $A(n,n)$ in (2.17). Substituting in (2.18), we have $A'(n-1,n-1) = C$, where as noted before, C is the contributions from nodes in $\mathbf{v}'\backslash V_n$ and having an edge with V_{n-1}. This implies that $A'(n-1,n-1)$ is the coefficient in the potential matrix for the subgraph induced by $\mathbf{v}'\backslash V_n$. Since only V_{n-1} has an edge with V_n, coefficients of nodes other than V_n and V_{n-1} are unaffected by the removal of V_n. Hence, $\mathbf{M}_n$ is the potential matrix for the subgraph induced by $\mathbf{v}'\backslash V_n$,

$$\mathbf{M}_n = \mathbf{A}^{(n-1)}.$$

Since $\mathbf{v}'\backslash V_n$ is acyclic, a leaf is always present, rearrange the rows such that $\mathbf{A}^{(n-1)}$ has a leaf in the last two rows, i.e., it has the same structure as in (2.17). Remove

a leaf in each step of the recursion, until all the edges are removed, then find the determinant with the diagonal matrix consisting of the remaining nodes and we obtain (2.8). □

CHAPTER 3

MINIMUM COST IN-NETWORK PROCESSING FOR OPTIMAL INFERENCE

Routing in communication networks, both wireline and wireless, has been a subject of extensive and in-depth study over the last few decades. It is a subject that is fairly well understood. Its "state-of-the-art" status can be summarized as follows; If a well-defined performance measure can be translated to a link metric, then there are low-complexity, efficient, robust, fast-converging, and often distributed algorithms for finding the optimal routes. Note the important distinction regarding the possibility of mapping the performance measure to a link metric. For example, in the internet, if end-to-end latency is the performance measure, then the link metric is delay over the link. Bellman-Ford-type algorithms then perform very well and quickly discover the best routes [93]. By contrast, on the traditional circuit-switched voice telephone network, where the performance measure is blocking probability, there is no known link metric that captures the performance measure and, hence, up to this day we only have heuristic routing algorithms for assigning routes to accepted calls.

At this point it is also important to note that the routing problem, being basically a discrete optimization task, has always a default solution that consists of the exhaustive search over the finite number of possible routes. The only reason this solution is unattractive is the prohibitive complexity of this search as the network size increases.

Another example of successful mapping of a performance measure to a link metric that allows the use of efficient algorithms is energy consumption in a wireless network. The energy consumption on a single link is then the right met-

ric. That link energy consumption, depending on the assumptions on the network operation, consists of the transmission energy (proportional to the transmission power needed to reach the destination at a given rate and bit-error-rate target for chosen modulation and coding schemes, as well as to the channel attenuation), the energy expended for reception at the receiving end of the link, and, finally, the residual energy at the battery of the node at that end.

What all routing problems to date share is the traditional IP paradigm of store-and forward, which treats the source packets as "sacrosanct" monoliths that must be carried through the network intact until they are received at the destination node. Already, the idea of network coding has shown how it is possible to improve performance if this paradigm is reconsidered [94]. In this chapter we will examine a different issue that arises in specialized routing that shows equally well the inadequacy of traditional packet forwarding.

Our focus will be the case of wireless sensor networks. The unique characteristic of such networks is that the performance measure is typically associated with the "mission" of the network. For example, if the sensor network is deployed for the purpose of detecting the presence of a target, then the objective is to maximize the probability of correct detection, subject to the usual constraint o the false alarm rate. In other words, the mission of the network is statistical inference. Thus, the collected measurement data at the source sensors need not be forwarded to the fusion center (i.e. the ultimate destination node) in their entirety. Of course, such complete forwarding remains an option (just as the exhaustive search over all possible routes was an option in ordinary routing). But it is an inefficient option that is highly undesirable in networks that must also prolong their lifetime as much as possible.

In this chapter we aim at a comprehensive presentation of this new aspect of wireless sensor networking and at a unified study of routing, inference, and energy consumption. Inherent in this presentation is the notion of combinatorial optimization (which remains the underpinning element of the routing task) and of spatial information modeling (which defines the information dependencies in the data the sensor nodes gather).

3.1 Network and Cost Model

We assume the presence of a medium-access control (MAC) that eliminates collisions or interferences among the nodes. The network is connected, i.e., communication is feasible via a multi-hop route between any two nodes in the network. We assume that communication is bidirectional. We consider the unicast mode of routing, where a packet from a node is routed to a single destination and the intermediate nodes do not perform any processing or store the packet for future use.

In our formulation, the processing costs are assumed constant, thus ignored in the optimization. Usually the routing costs reflect transmission energy, but it could also represent, for example, delay, bandwidth, or a combination of these considerations. We represent the routing of a real number by a packet. A routing cost function is assumed to be known, and is denoted by $\mathcal{E}_{i,j} > 0$ between i and j. The metric closure on communication/network graph N_g, is defined as the complete graph where the cost of each edge (i, j) in the metric closure is the cost of the shortest path between i and j in N_g [95, p. 58]. Henceforth, we only consider the metric closure of the communication graph, denoted by $\bar{N}_g$, and denote the metric costs by $\mathcal{E}^{\text{SP}}_{i,j}$. There is no loss of generality, since the edges of

the metric closure can be replaced with the corresponding shortest paths. For any graph $N_g \subset \bar{N}_g$, let $\mathcal{E}(N_g)$ denote the total cost of its links,

$$\mathcal{E}(N_g) := \sum_{e \in E_{N_g}} \mathcal{E}_e, \tag{3.1}$$

where $\mathcal{E}_e$ is the cost of the link e and E_{N_g} is the set of links in N_g; if a link is used m times, then E_{N_g} contains m parallel links to incorporate the costs in our formulation.

We assume that it is connected but not necessarily fully connected, and that it contains the Euclidean minimum spanning tree over the node set $\mathbf{v}_n$ and directed towards the fusion center v_1, denoted by DMST($\mathbf{v}_n; v_1$). Usually in the literature, in order to incorporate the maximum power constraints at the nodes, the network graph is assumed to be a disc graph with radius above the connectivity threshold [11], but we do not limit to this model. Transmissions are scheduled so as to not interfere with one other. Nodes are capable of adjusting their transmission power depending on the location of the receiver.

A fusion policy $\Upsilon(\mathbf{v}_n)$ consists of a transmission schedule with the transmitter-receiver pairs and the aggregation algorithm that allows a node to combine its own and received values to produce a new communicating value. We model a fusion policy Υ by a *fusion-policy digraph*, $\mathcal{F}_\Upsilon := (\mathbf{v}_n, \overrightarrow{E}_\Upsilon)$, and $\overrightarrow{E}_\Upsilon$ contains *directed links*. A directed[1] link $\langle i, j \rangle$ denotes a direct transmission from i to j and is required to be a member in the network graph $N_g(\mathbf{v}_n)$ for transmissions to be feasible. If one node communicates with another node k times, k direct links are present between these two nodes in the edge set $\overrightarrow{E}_\Upsilon$ of the fusion

[1] We denote a directed link by $\langle i, j \rangle$ and an undirected link by (i, j).

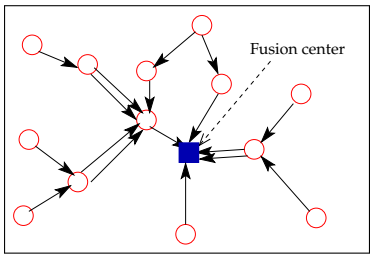


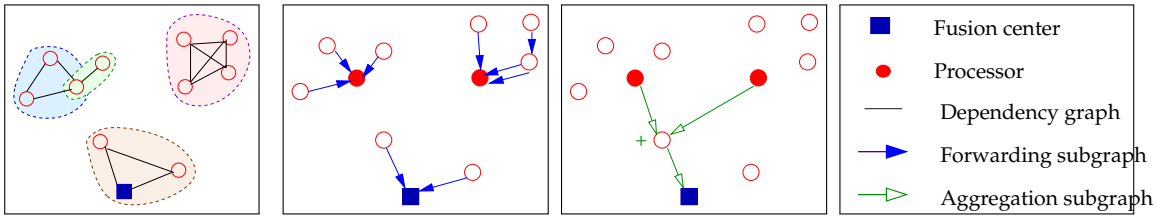
Figure 3.1: Fusion policy digraph $\mathcal{F}_\Upsilon$: each edge routes one real number.

policy Υ. Since we are only interested in characterizing the overall energy expenditure, the order of transmissions is not important; we only need to consider the associated cost with each link in $\overrightarrow{E}_\Upsilon$ and calculate the sum cost for Υ.

Nodes communicate in the form of packets. Each packet contains bits for at most one (quantized) real variable and other overhead bits independent of the network size. We assume that all real variables[2] are quantized to K bits, and K is independent of network size and is sufficiently large that quantization errors can be ignored.

In our formulation all real numbers are quantized with sufficiently high precision to ignore the quantization error and all nodes function as both sensors and routers. Quantization is indeed an important issue for detection and communication. However, even in the classical distributed setup, optimal quantization is not tractable for the correlated case. The recent works on this topic consider conditionally i.i.d. measurements with a fixed network topology of bounded-height tree [96] or a tandem network [97].

[2]In principle, the raw and aggregated data may require different amount of energy for communication, and can be incorporated into our framework.



(a) Maximal cliques of dependency graph

(b) Forwarding subgraph computes clique potentials

(c) Aggregation subgraph adds computed potentials

(d) Legend

Figure 3.2: Fusion policy DFMRF for inference at fusion center.

3.2 Formulation of Minimum Cost Fusion

By optimal routing for inference, we mean the fusion scheme that minimizes the total costs of routing under the constraint that the likelihood function in (2.14) is delivered to the fusion center.

$$\mathcal{E}(\Upsilon^*(\mathbf{v}_n)) = \inf_{\Upsilon \in \mathfrak{F}_{\mathcal{G}}} \sum_{i \in \mathbf{v}_n} \mathcal{E}_i(\Upsilon(\mathbf{v}_n)), \tag{3.2}$$

where $\mathfrak{F}_{\mathcal{G}}$ is the set of valid data-fusion policies

$$\mathfrak{F}_{\mathcal{G}} := \{\Upsilon : L_{\mathcal{G}}(\mathbf{y}_{\mathbf{v}_n}) \text{ computable at the fusion center}\}.$$

Recall the succinct form of LLR in (2.14),

$$L(\mathbf{Y}_{\mathbf{v}}; \Xi) = \sum_{c \in C} \phi_c(\mathbf{Y}_c). \tag{3.3}$$

Hence, the LLR consists of the sum of the clique potential functions ϕ and is amenable to localized processing within the cliques of the Markov random field. Hence, we propose a hierarchical order of processing the LLR. See Fig.3.2. In the first stage, raw data are forwarded to compute all the potential functions

at various nodes in the network. In the second stage, the computed values are summed up and delivered to the fusion center.

For the first stage of LLR computation, each clique potential function ϕ_c is assigned a unique computation site, known as the *processor* for clique c, denoted by $\text{Proc}(c)$. Once the processor for clique c is assigned, measurement Y_i of each clique member $i \subset c$ (other than the processor) is routed to $\text{Proc}(c)$ along a path of feasible communication links. Since we are considering unicast mode of communication, the minimum cost is along the shortest path represented by the link $< i, \text{Proc}(c) > \in \bar{N}_g$ with cost $\mathcal{E}^{\text{SP}}(i, \text{Proc}(c))$, where $\bar{N}_g$ is the metric closure of the communication graph. The set of all links used by a fusion scheme in the first stage of computation to forward raw data to the processors is called the *forwarding* subgraph, denoted by FG,

$$\text{FG}:=\{< i, \text{Proc}(c) >: i \subset c, i \neq \text{Proc}(c), c \in \mathcal{C}\}.$$

In the second stage of LLR computation, all the computed potential functions are summed up to obtain the LLR which is then delivered to the fusion center. The set of links used by a fusion scheme in the second stage of LLR computation to sum up the computed potential values is known as the *aggregation* subgraph, denoted by AG. The tuple with the forwarding and aggregation subgraphs of a fusion scheme Υ is referred to as the *fusion digraph*, $\mathcal{F}_\Upsilon$:=$\{\text{FG}_\Upsilon, \text{AG}_\Upsilon\}$. A schematic of a fusion scheme is shown in Fig.3.2. The total routing costs of a fusion scheme is given by

$$\mathcal{E}(\mathcal{F}_\Upsilon) = \mathcal{E}(\text{FG}_\Upsilon) + \mathcal{E}(\text{AG}_\Upsilon).$$

Hence, any fusion scheme Υ in our setup is specified by a processor-assignment mapping Proc_Υ and a fusion digraph $\mathcal{F}_\Upsilon = \{\text{FG}_\Upsilon, \text{AG}_\Upsilon\}$, and we represent the

scheme by the tuple Υ:={Proc, FG, AG}. Note that we do not explicitly specify the sequence in which data is transported and processed by a fusion scheme; we impose constraints to ensure that such a feasible sequence exists.

We first need the constraint that the scheme delivers the LLR to the fusion center

$$AggVal(v_1; \Upsilon) = L(\mathbf{Y_v}; \Xi), \tag{3.4}$$

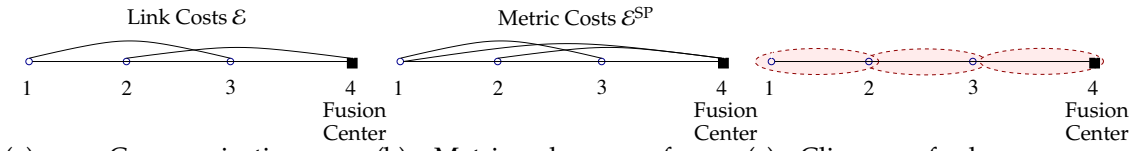
where $AggVal(i; \Upsilon)$ is the value at node i at the end of fusion.

3.2.1 Local Processor Assignment

We now make the following additional assumption which simplifies the fusion scheme: each clique potential function ϕ_c is assigned a "local" processor, which is one of the clique members,

$$\text{Proc}(c) \subset c, \quad \forall\, c \in C. \tag{3.5}$$

The local processor assignment also implies that local knowledge of potential function parameters is sufficient, i.e., each sensor i only needs to know the potential functions ϕ_c of the cliques c to which it belongs, and hence, the storage requirement at the sensors is considerably reduced. In practice, the potential function parameters are sent to the nodes by the fusion center after empirically estimating the joint-pdf of the measurements. Through this, the nodes



(a) Communication graph. (b) Metric closure of comm. graph. (c) Cliques of dependency graph.

Figure 3.3: Inputs to the problem of minimum cost fusion for inference.

also implicitly receive information about their clique memberships. Hence, local processor assignment can also reduce the communication overhead during the learning stage. Localized processing can be especially efficient when the dependency graph of the Markov random field is a proximity graph, where edges are based on local point configuration [98]. We now formally define the minimum-cost fusion scheme Υ^* which minimizes the total routing costs

$$\Upsilon^* := \arg\min_{\Upsilon} \mathcal{E}(\mathcal{F}_\Upsilon), \tag{3.6}$$

subject to the constraints in (3.4) and (3.5). Hence, the problem of minimum cost fusion takes the metric closure of communication graph and the maximal cliques of the dependency graph as inputs and provides a processor assignment and fusion digraph as outputs. An example of the problem of minimum cost fusion is illustrated in Fig.3.3, with the communication graph in Fig.3.3a and the path graph as the dependency graph in Fig.3.3c, which are independent of one another. The resulting metric closure of communication graph in Fig.3.3b and cliques of dependency graph are taken as the inputs for the problem of minimum cost fusion.

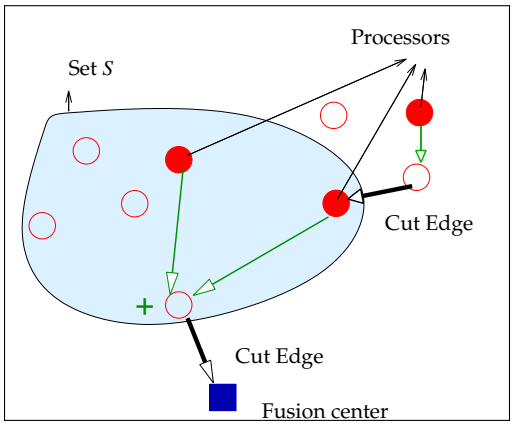


Figure 3.4: Cut set S separating set of processors and fusion center.

3.2.2 0-1 Integer Programming Formulation

We now write a 0-1 integer program whose optimal solution provides the minimum cost fusion scheme in (3.6) for computing the LLR and delivering it to the fusion center v_1. We can map any valid fusion digraph $\mathcal{F} = \{\text{FG}, \text{AG}\}$ and the processor assignment mapping Proc to variables $\mathbf{y}$ and $\mathbf{z}$, defined as

$$z(j,c) := I[\text{Proc}(c) == j], \quad y(i,j) := I[< i,j > \in \text{AG}],$$

where I is the indicator function. Once the processor assignment is fixed, the set of shortest paths from clique members to the processors minimizes the routing costs in the forwarding subgraph. Hence, we can set the forwarding subgraph as

$$\text{FG} \leftarrow \{< i,j >: I(\sum_{c:i \subset c} z(j,c) \geq 1)\},$$

where we ensure that every node i forwards its measurement to node j, whenever j is the processor of cliques c that contain node i along the link in the metric closure (which has the same cost as the shortest path). Hence, the total routing costs of the fusion digraph can be expressed as,

$$\mathcal{E}(\mathcal{F}) = \mathcal{E}(\text{FG}) + \mathcal{E}(\text{AG}) = \frac{1}{2} \sum_{i,j \in \mathbf{v}} [I(\sum_{c:i \subset c} z(j,c) \geq 1) + y(i,j)] \mathcal{E}^{\text{SP}}(i,j),$$

where the factor of $\frac{1}{2}$ ensures that each edge is counted only once. We now write a constraint equivalent to the local processor constraint in (3.8) and ensuring that at least one processor is selected,

$$\sum_{j \subset c} z(j,c) \geq 1, \quad \forall c \in C.$$

We now need a constraint on the aggregation subgraph to ensure that the sum of the potential functions is delivered to the fusion center, and hence, the constraint in (3.4) is satisfied. To this end, we define that *A separates B* if $\mathbf{A} \cap \mathbf{B} \neq \emptyset$ and $\mathbf{A} \cap \mathbf{B} \neq \mathbf{B}$. We consider all sets $\mathbf{s} \subset \mathbf{v}$ separating the union of the set of processors and the fusion center. A cut edge of set $\mathbf{s}$ is one that has exactly one endpoint in $\mathbf{s}$. As illustrated in Fig.3.4, since all the values at the processors contained within $\mathbf{s}$ can be summed up to a single packet, for the information to flow out of $\mathbf{s}$ (or into $\mathbf{s}$), at least one cut edge of $\mathbf{s}$ is needed. Hence, we write the constraint that

$$\sum_{i \in \mathbf{s}, j \notin \mathbf{s}} y(i,j) \geq 1, \forall \mathbf{s} \subset \mathbf{v} \text{ separating } \{\bigcup_{c \in C} \text{Proc}(c) \bigcup v_1\}.$$

We now have the integer program,

$$\frac{1}{2} \min_{\mathbf{y},\mathbf{z}} \quad \sum_{i,j \in \mathbf{v}} [I(\sum_{c: i \subset c} z(j,c) \geq 1) + y(i,j)] \mathcal{E}^{\text{SP}}(i,j) \qquad \text{(IP-1)}, \tag{3.7}$$

$$s.t. \quad \sum_{j \subset c} z(j,c) \geq 1, \quad \forall c \in C, \quad \text{let Proc}(c) := \{j : z(j,c) = 1\}, \tag{3.8}$$

$$\sum_{i \in \mathbf{s}, j \notin \mathbf{s}} y(i,j) \geq 1, \forall \mathbf{s} \subset \mathbf{v} \text{ separating } \{\bigcup_{c \in C} \text{Proc}(c) \bigcup v_1\}, \tag{3.9}$$

$$y(i,j), z(j,c) \in \{0,1\}. \tag{3.10}$$

Hence, the optimal solutions to (3.6) and (3.7) are the same.

3.3 Special Case: IID Measurements

In the special case when the measurements are i.i.d. conditioned on either hypothesis, the log-likelihood ratio (LLR) in (2.10) is the sum of the log-likelihoods of individual sensor measurementsi.e.,

$$L(\mathbf{Y}_{\mathbf{v}}) = \sum_{v_i \in \mathbf{v}} L(Y_{v_i}), \quad \mathbf{Y}_{v_i} \overset{i.i.d.}{\sim} \mathcal{H}_0 \text{ or } \mathcal{H}_1. \tag{3.11}$$

Theorem 1 (Lower bound on minimum energy expenditure) *The following results hold:*

1. *the energy cost for the optimal fusion policy Υ^* in (3.2) satisfies*

$$\mathcal{E}(\Upsilon^*(\mathbf{v}_n)) \geq \mathcal{E}(MST(\mathbf{v}_n)) := \sum_{e \in \mathbb{E}_{\theta_i} N_{smst}(\lambda)(\mathbf{v}_n)} |e|^{\nu}, \tag{3.12}$$

2. *the lower bound (3.12) is achieved (i.e., equality holds) when the observations are independent under both hypotheses. In this case, the optimal fusion policy Υ^* aggregates data along $DMST(\mathbf{v}_n; v_1)$, the directed minimum spanning tree, with all the edges directed toward the fusion center v_1. Hence, the optimal fusion digraph $\mathcal{F}_{\Upsilon^*}$ is the $DMST(\mathbf{v}_n; v_1)$.*

Proof: We first prove part 2), for which we consider the case when observations are independent, and the log-likelihood ratio is given by

$$L_{\mathcal{G}}(\mathbf{y}_{\mathbf{v}_n}) = \sum_{i \in \mathbf{v}_n} L_i(y_i), \quad L_i(y_i) := \log \frac{f_{1,i}(y_i)}{f_{0,i}(y_i)},$$

where $f_{k,i}$ is the marginal pdf at node i under $\mathcal{H}_k$. Consider MST($\mathbf{v}_n$), whose links minimize $\sum_{e \in \text{Tree}(\mathbf{v}_n)} |e|^\gamma$. It is easy to check that at the fusion center, the log-likelihood ratio can be computed using the following aggregation policy along the DMST($\mathbf{v}_n; v_1$) as illustrated in Fig.3.5: each node i computes the aggregated variable $q_i(\mathbf{y}_{\mathbf{v}_n})$ from its predecessor and sends it to its immediate successor. The variable q_i is given by the summation

$$q_i(\mathbf{y}_{\mathbf{v}_n}) := \sum_{j \in \mathcal{N}_p(i)} q_j(\mathbf{y}_{\mathbf{v}_n}) + L_i(y_i), \tag{3.13}$$

where $\mathcal{N}_p(i)$ is the set of immediate predecessors of i in DMST($\mathbf{v}_n; v_1$).

To show part 1), we note that any data-fusion policy must have each node transmit at least once and that the transmission must ultimately reach the fusion center. This implies that the fusion digraph must be connected with the fusion center and the DMST with edge-weight $|e|^\gamma$ minimizes the total energy under the above constraints. Hence, we have (3.12). □

Note that the above lower bound in (3.12) is achievable when the measurements are independent under both hypotheses. It is interesting to note that data correlations, in general, increase the energy consumption under the constraint of optimal inference performance since the log-likelihood ratio in (3.3) cannot be decomposed fully in terms of the individual node measurements.

3.3.1 Data Fusion for Markov Random Fields (DFMRF) Scheme

We first propose a simple heuristic (DFMRF), based on the minimum spanning tree. Here, we separate the design of processor selection and aggregation tree. We arbitrarily assign a clique member as the clique processor and then exploit

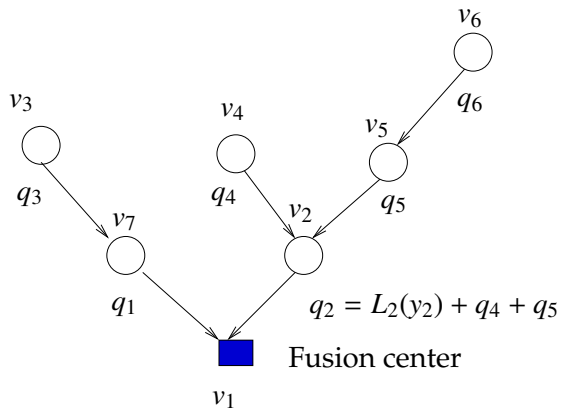


Figure 3.5: The optimal fusion graph DMST for independent observations.

the fact that it is feasible to compute the sum of the potentials along the MST. Of course, here only the processors have useful information in the form of potential functions and the other nodes just forward the aggregated information. This heuristic is simple to implement since there are efficient distributed algorithms for finding the MST [99,100].

We specify the DFMRF scheme in Fig.3.6. For a clique c, the processor is assigned arbitrarily to the clique member with the lowest index (line 3). Other suitable factors such as residual energy can instead be used for the assignment. The shortest-path routes from other members of c to the processor are added to the forwarding subgraph FG (line 5), and the raw data is routed along these links to enable the computation of the clique potentials. Note that the construction of the FG can be implemented in a localized manner whenever the dependency graph is local (e.g., k nearest-neighbor graph, disk graph). The aggregation subgraph AG is DMST($\mathbf{v}$), the minimum spanning tree, directed towards the fusion center (line 9) and potentials are added hierarchically along AG.

We now quantify the performance of the DFMRF scheme for a special scenario that allows us to utilize the lower-bound in Theorem 1.

Require: $\mathbf{v} = \{v_1, \ldots, v_n\}$, v_1: Fusion center, $C = \{c_0, \ldots, c_{|C|-1}\}$: maximal cliques of MRF, DMST($\mathbf{v}$): Minimum spanning tree, direct toward v_1

SP(i, j)= (Directed) shortest path from i to j
for $j \leftarrow 0, |C| - 1$ **do**
    $\text{Proc}(c_j) \leftarrow \min_{v_i \in c_j} v_i$ ▷ Arbitrary processor assignment
    if $|c_j| > 1$ **then**
        Add $\mathcal{E}^{\text{SP}}(c_j \setminus \text{Proc}(c_j), \text{Proc}(c_j))$ to FG
    end if
end for
AG $\leftarrow$ DMST($\mathbf{v}$), $\Upsilon \leftarrow \{\text{Proc}, \text{FG}, \text{AG}\}$
return Υ

Figure 3.6: Data fusion for Markov random fields (DFMRF) policy.

Theorem 2 (Approximation) *For the case when the routing costs are Euclidean and the dependency graph is a subgraph of the Euclidean MST (e.g., 1-nearest neighbor graph), the DFMRF scheme has an approximation ratio of* 2.

Proof: The MST in the lower bound (Theorem 1) is Euclidean, since the transmission costs are Euclidean. Since the dependency graph is a subgraph of the Euclidean MST, all the links in DFMRF are contained in the Euclidean MST. Hence, we have the approximation ratio of 2. To show that the bound is tight, we note that the case of extended equilateral triangles on the Euclidean plane achieves this bound. □

In Chapter 5, we generalize the above result to any stabilizing Euclidean dependency graph and provide approximation guarantees for random node sets with the number of nodes $n \to \infty$.

Require: $\mathbf{v} = \{v_1, \ldots, v_n\}$, v_1: Fusion center, $C = \{c_0, \ldots, c_{|C|-1}\}$: maximal cliques of MRF,
1: N_g= Metric closure of comm. graph, $\mathcal{E}^{\text{SP}}$ = Link costs in $\bar{N}_g$,
2: ST$(G, \mathcal{L})$ = δ-approx. Steiner tree on G, terminal set $\mathcal{L}$
3: $G', \mathbf{v}_c \leftarrow$ Map-All$(\bar{N}_g; \mathcal{E}^{\text{SP}}, C)$
4: DST = ST$(G', \mathbf{v}_c \cup v_1)$ and directed towards v_1
5: $\Upsilon \leftarrow$ RevMap-All(DST; $\mathbf{v}_c, \mathbf{v}, C$)
6: **return** Υ

Figure 3.7: δ-approximate fusion policy Υ (AggApprox).

3.4 Steiner Tree Reduction

In this section, we show that optimal fusion has a Steiner-tree reduction under local processing constraints. We specify the graph transformations required for such a reduction and finally obtain a valid fusion scheme with processor assignment and fusion digraph. We also show that the Steiner-tree reduction is approximation factor preserving. This implies that any approximation algorithm for Steiner tree provides the same ratio for minimum cost fusion.

3.4.1 Simplified Integer Program

We first note that if the processor assignment is already predetermined and not part of the routing cost optimization, then we can easily characterize the optimal solution. In practice, a predetermined processor assignment might be enforced by considering other factors such as processing capabilities or residual energies of different nodes. In this case, the forwarding subgraph is also predetermined by the shortest paths to the processors. The optimal aggregation subgraph is the Steiner tree with the set of processors and the fusion center, as the terminals. This is because the sum of the potential function values at the processors is

computed optimally along the Steiner tree.

We next consider a modified cost optimization problem, where we ignore the routing costs of the forwarding subgraph, incurred in transporting the raw measurements to a processor. In [2, Lemma 3], we show that the minimum cost aggregation subgraph is the group Steiner tree [101], with nodes in each clique of the Markov random field forming a group.

The presence of processor assignment in cost optimization in (3.7) makes the problem harder than the above versions. It influences the costs of both the forwarding and aggregation subgraphs in a fusion scheme. It is not immediately clear that there is a Steiner tree reduction for (3.7). In fact, if we directly relax the integers to $\mathbf{y}, \mathbf{z} \geq 0$ in (3.7), the program is non-linear. We now use the local processor assignment constraint in (3.8) to write an equivalent integer program with a linear relaxation. Let $\mathbf{z}^*$ be the optimal solution to (3.7). We have

$$\begin{aligned}
\sum_{i,j\in\mathbf{v}} I(\sum_{c:i\subset c} z^*(j,c) \geq 1)\mathcal{E}^{\text{SP}}(i,j) &= \sum_{i,j\in\mathbf{v}} I(\sum_{c:i,j\subset c} z^*(j,c) \geq 1)\mathcal{E}^{\text{SP}}(i,j), \\
&= \sum_{i,j\in\mathbf{v}} \sum_{c:i,j\subset c} z^*(j,c)\mathcal{E}^{\text{SP}}(i,j), \\
&= \sum_{\substack{c\in C \\ |c|>1}} \sum_{i,j\subset c} z^*(j,c)\mathcal{E}^{\text{SP}}(i,j), \qquad (3.14)
\end{aligned}$$

where the first equality is from local processor assignment constraint, the second equality is due to the fact that we need to assign only one processor and that there is a unique maximal clique c, if it exists, containing both i and j. Note that if the local assignment constraint is removed, then j might be assigned as the processor to many cliques c and hence, the equality does not hold. Interchanging the sums in the last equality is possible since the terms are non-zero

function MAP-ALL($\bar{N}_g(\mathbf{v}); \mathcal{E}^{\text{SP}}, C$)
  nbd($v; G$) = Neighborhood of v in G
  Initialize $G' \leftarrow \bar{N}_g$, $\mathbf{v}_c \leftarrow \emptyset$,
  for $j \leftarrow 0, |C| - 1$ **do** ▷ Let $\mathbf{v}_n$ and C be ordered
    if $|c_j| > 1$ **then**
      $\mathbf{v}_c \leftarrow v_{n+j}$, Add new node v_{n+j} to G',
      for all $v_i \subset c_j$ **do**
        Add node v_i to nbd($v_{n+j}; G'$)
        $\mathcal{E}^{\text{SP}}(v_n, v_i; G') \leftarrow \sum_{v_k \subset c_j, k \neq i} \mathcal{E}^{\text{SP}}(v_i, v_k; \bar{N}_g)$
      end for
    else
      $\mathbf{v}_c \leftarrow v_i$, for $v_i \subset c_j$ ▷ For trivial cliques
    end if
  end for
return G', $\mathbf{v}_c$
end function

Figure 3.8: Map-All($\bar{N}_g; \mathcal{E}^{\text{SP}}, C$) adds virtual nodes for each non-trivial clique.

when there is a clique c containing both i and j, and this implies that $|c| > 1$. Hence, we can now write an equivalent IP for minimum cost fusion under local processor assignment

$$\min_{\mathbf{y},\mathbf{z}} [\sum_{\substack{c \in C \\ |c|>1}} \sum_{i,j \subset c} z(j,c) \mathcal{E}^{\text{SP}}(i,j) + \sum_{i,j \in \mathbf{v}} y(i,j) \mathcal{E}^{\text{SP}}(i,j)] \qquad \text{(IP-2)}, \tag{3.15}$$

subject to the same constraints (3.8)-(3.10). Upon relaxation of the integer constraints, IP-2 is a linear program.

We now show that a Steiner tree on the transformed communication graph is the optimal solution to IP-2 in (3.15). To this end, we define an operation Map-All($\bar{N}_g$) in Fig.3.8 which involves adding new virtual clique-representative nodes $\mathbf{v}_c$ for each non-trivial clique ($|c| > 1$) and adding edges between $\mathbf{v}_c$ and

function REVMAP-ALL($G'; \mathbf{v}_c, \mathbf{v}, C$)
$\mathcal{N}_s(i; G)$, $\mathcal{N}_p(i; G)$ = Imm. successor, predecessor of i
Initialize $G \leftarrow G'$
for all $v_j \in \mathbf{v}_c$ **do**
if $j > n$ **then**
$k \leftarrow j - n$,
Proc$(c_k) \leftarrow \mathcal{N}_s(v_j; G')$, for $c_k \in C$,
$v_j \leftarrow c_k \setminus$ Proc(c_k), Replace $< v_j, \text{Proc}(c_k) >$ in G with edges $< v_j, \text{Proc}(c_k) >$, mark them
if $\mathcal{N}_p(v_j; G) \neq \emptyset$ **then** Replace $< \mathcal{N}_p(v_j), v_j >$ in G with edges $< \mathcal{N}_p(v_j), \text{Proc}(c_k) >$
end if
else
Proc$(c_l) \leftarrow v_j$, for $v_j \subset c_l$ ▷ For trivial cliques
end if
end for
FG $\leftarrow$ Marked edges of G, AG $\leftarrow G \setminus$ FG
$\Upsilon \leftarrow$ {Proc, FG, AG}
return Υ
end function

Figure 3.9: RevMap-All($G; \mathbf{v}_c, \mathbf{v}, C$) maps tree G to fusion scheme Υ.

all the members of clique c with costs,

$$\mathcal{E}^{\text{SP}}(\mathbf{v}_c, j) := \sum_{i \subset c} \mathcal{E}^{\text{SP}}(i, j), \quad \forall j \subset c.$$

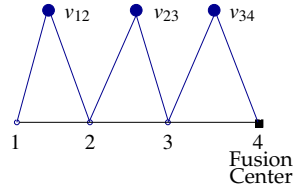
The above cost represents the cost incurred in the forwarding subgraph upon assigning a node j as the processor for clique c. Let the set of all added clique representative vertices be $\mathbf{v}'$. Hence, IP-2 in (3.15) is now equivalent to

$$\frac{1}{2} \min_{\mathbf{y},\mathbf{z}} \sum_{\mathbf{v}_c \in \mathbf{v}', j \in \mathbf{v}} z(j, c) \mathcal{E}^{\text{SP}}(\mathbf{v}_c, j) + \sum_{i,j \in \mathbf{v}} y(i, j) \mathcal{E}^{\text{SP}}(i, j),$$

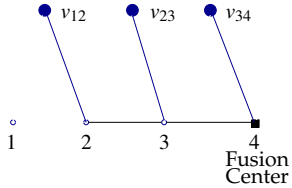
subject to the same constraints (3.8)-(3.10). For the final step, we define the set of nodes $\mathbf{v}''$ to account for trivial cliques

$$\mathbf{v}'' := \{i : i \in \mathbf{v}, i \subset c, \text{ for some } c \in C, |c| = 1\}.$$

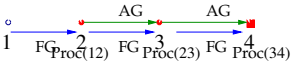
The set of clique representative nodes is $\mathbf{v}_c := \mathbf{v}' \cup \mathbf{v}''$, the set of newly added virtual nodes and the trivial cliques. We now write the equivalent IP which is



(a) Expanded communication graph Map-All($\bar{N}_g$) with costs under different scenarios of processor selection.



(b) Steiner tree on Map-All($\bar{N}_g$) with nodes v_{12}, v_{23}, v_{34} and fusion center 4 as terminals.



(c) Fusion scheme with processor assignment, forwarding and aggregation subgraphs after RevMap-All.

Figure 3.10: Example 1: Minimum cost fusion for chain dependency graph.

the Steiner tree with the set of clique representatives $\mathbf{v}_c$ and the fusion center v_1 as the terminals,

$$\frac{1}{2}\min_{\mathbf{x}} \quad \sum_{i,j\in\mathbf{v}} x(i,j)\mathcal{E}^{\text{SP}}(i,j), \tag{3.16}$$

$$s.t. \quad \sum_{i\in\mathbf{s},\, j\notin\mathbf{s}} x(i,j) \geq 1, \forall \mathbf{s} \subset \mathbf{v}\cup\mathbf{v}' \text{ separating } \{\mathbf{v}_c \cup v_1\}, \quad x(i,j)\in\{0,1\} \tag{3.17}$$

The equivalence holds since in the above Steiner tree, each clique representative node $\mathbf{v}_c \in \mathbf{v}'$ has to be connected to at least one clique member and hence, the local processor assignment constraint in (3.8) is satisfied, and the constraint (3.17) which ensures that all the terminals $\mathbf{v}' \cup v_1$ are connected implies that all the processors and the fusion center are connected and hence, the constraint in (3.9) is satisfied. Hence, the optimal solution to minimum cost routing for inference is a Steiner tree on the transformed graph Map-All($\bar{N}_g$).

In order to obtain the fusion scheme, we need another transformation after finding the Steiner tree in (3.16) on the transformed graph Map-All($\bar{N}_g$). We first direct the Steiner tree towards the fusion center, denoted by DST. The reverse mapping RevMap-All(DST) in Fig.3.9 assigns the unique immediate successor

of every clique-representative node v_c in DST as the processor of the clique c. The edges from the representative nodes in DST are replaced by links in the metric closure from other clique members to the processor and added to the forwarding subgraph of the fusion scheme. All other edges, not belonging to representative nodes in DST, are assigned as the aggregation subgraph.

In the above discussion, we have shown that the optimal solution is a Steiner tree involving transformations Map-All and RevMap-All, summarized in Fig.3.7. We now prove in addition that the above Steiner-tree reduction is approximation-factor preserving. To this end, we state the conditions under which the reduction preserves the approximation ratio [102, A.3.1].

Definition 2 (Approximation-factor preserving reduction) *Let Π_1 and Π_2 be two minimization problems, with $opt\_tradeoff_{\Pi_i}$ denoting the values of their optimal solutions. An approximation factor preserving reduction from Π_1 to Π_2 consists of two polynomial time algorithms, f and g, such that,*

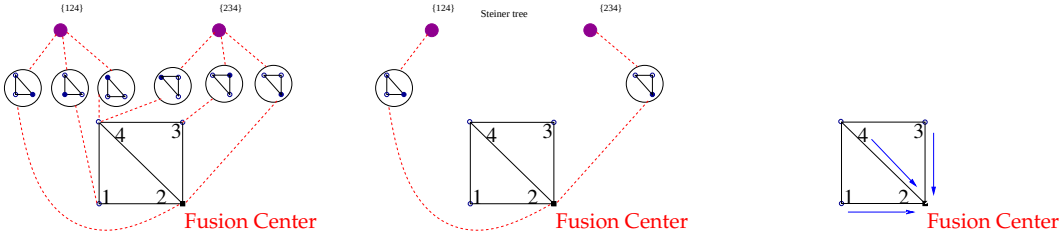
- *for any instance I_1 of Π_1, $I_2 = f(I_1)$ is an instance of Π_2 such that*

$$opt\_tradeoff_{\Pi_2}(I_2) \leq opt\_tradeoff_{\Pi_1}(I_1). \tag{3.18}$$

- *for any solution t of I_2, $s = g(I_1, t)$ is a solution of I_1 such that*

$$obj_{\Pi_1}(I_1, s) \leq obj_{\Pi_2}(I_2, t). \tag{3.19}$$

We now note that AggApprox results in a feasible fusion and runs in polynomial time since there are polynomial number of cliques. For any feasible solution to



(a) Expanded communication graph Map-All($\bar{N}_g$) with costs under different scenarios of processor selection.

(b) Steiner tree on Map-All($\bar{N}_g$) with nodes v_{12}, v_{23}, v_{34} and fusion center 4 as terminals.

(c) Fusion scheme with processor assignment, forwarding and aggregation subgraphs after RevMap-All.

Figure 3.11: Example 2: Minimum cost fusion for given dependency graph.

Steiner tree, replacement of links in line 9 of RevMap-All in Fig.3.9 reduces the sum cost, and hence, (3.19) holds.

The approximation-ratio preserving Steiner tree reduction implies that any approximation algorithm for Steiner tree provides the same approximation ratio for minimum cost fusion, when applied with the above transformations. Since currently the best known ratio for Steiner tree is 1.55, it is also the best possible approximation for minimum cost fusion for inference.

3.4.2 Examples

We now illustrate the optimal fusion scheme through Steiner-tree reduction for simple examples of a chain dependency graph in Fig.3.10, where the link communication costs and the metric closure are implicit and not shown. For this simple example, we can intuitively see that the optimal scheme first forwards raw data in the direction of fusion center. Upon computing the potential functions at the processors, the values are added along the chain, starting with the

farthest processor. In Fig.3.10c, this optimal fusion scheme with forwarding and aggregation subgraphs is shown along with the values transported along the links. We now illustrate that the Steiner tree with transformations provides the same optimal solution. In Fig.3.10a, the expanded communication graph Map-All($\bar{N}_g$) is shown with added clique-representative nodes and edges. The added edges represent the costs in the forwarding subgraph on choosing a node as a processor. In Fig.3.10b, the optimal Steiner tree on the expanded graph is shown with the clique representative nodes and the fusion center as terminals. Using RevMap-All, the Steiner tree is mapped to a fusion scheme by first directing the tree towards the fusion center, and then, assigning the immediate successor of clique representative nodes as processors. Hence, the member closer to the fusion center is chosen as the processor in this example. The edges from clique representative nodes are replaced with forwarding subgraph edges, and we can see that the costs are conserved. The remaining edges in the Steiner tree form the aggregation subgraph. Hence, the RevMap-All operation provides the optimal fusion scheme shown in Fig.3.10c. A similar Steiner tree solution is obtained for another example in Fig.3.11 and in this case, the optimal scheme reduces to shortest-path routing policy.

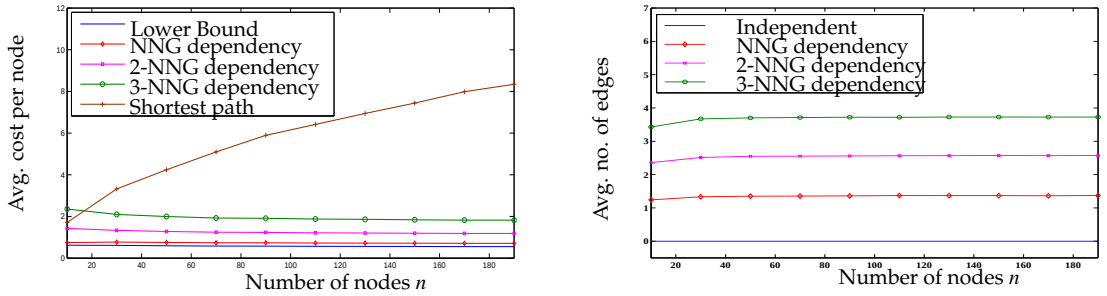
3.5 Simulation Results

We now plot some simulation results in Fig.3.12 under uniform random placement of nodes and conduct 500 independent simulation runs. We fix a constant node density and consider routing cost on link as $(i, j) \propto \text{dist}(i, j)^2$. We see that savings due to aggregation are considerable compared to shortest-path routing for k-nearest neighbor graphs (k-NNG), at low values of k. These graphs are

probably the best candidates, after the independent data case, for in-network processing of the likelihood function. We also observe that there is direct correspondence between the number of cliques and the routing cost for fusion. Hence, it appears that the number of cliques is a good measure for judging the effectiveness of in-network processing. The gap between the heuristics and the lower bound, represents the overhead arising due to correlation. A dense dependency graph has high routing costs due to the complexity of its likelihood function. This is unlike the case of compression with the aim of routing all the raw data to a destination, where a dense dependency graph (more correlation) implies redundancy and hence, reduction in routing costs.

The use of localized processing constraint and unicast mode of communication are crucial to obtaining the above Steiner-tree reduction. They lead to the separation of costs of routing raw measurements (in the forwarding subgraph) to compute different potential functions. On the other hand, in the absence of these constraints, the edge costs in the forwarding graph are no longer independent, and finding the optimal scheme requires the use of hyper-edges. However, once a scheme is designed under the unicast setup, the broadcast nature of the wireless medium could be exploited to further reduce costs by broadcasting raw data from each node to all its processors.

We have so far considered minimum cost routing for optimal inference. A relaxation of this problem is where we only select a subset of measurements for routing and fusion, and we aim to achieve optimal tradeoff between routing costs and end detection performance. This problem requires first the characterization of the detection performance, and one possibility is to use the detection error exponent, which is the asymptotic rate of exponential decay of error prob-



(a) Avg. routing cost per node.

(b) Avg. no. of Edge potentials calculated.

Figure 3.12: Simulation results for k nearest-neighbor dependency graphs.

ability. It will be interesting to explore if this problem has reduction to well known optimization problems, as it turned out in the case of optimal inference with local processing.

3.6 Extension to M-ary Hypothesis Testing

The problem of binary simple hypothesis testing is extended in a different direction by considering $M > 2$ number of hypotheses. This is known as the M-ary hypothesis-testing problem. In this case, the LLR vector, denoted by $\mathbf{L}(\mathbf{Y_v})$ with respect to $\mathcal{H}_0$ is defined as

$$\mathbf{L}(\mathbf{Y_v}) := \log[\frac{f(\mathbf{Y_v};\mathcal{H}_0)}{f(\mathbf{Y_v};\mathcal{H}_1)}, \ldots, \frac{f(\mathbf{Y_v};\mathcal{H}_0)}{f(\mathbf{Y_v};\mathcal{H}_{M-1})}]^T. \tag{3.20}$$

The above LLR vector is the minimal sufficient statistic for M-ary hypothesis testing [92].

Given that the measurements $\mathbf{Y_v}$ are drawn from distinct Markov random

fields under each hypothesis $\mathcal{H}_i$, let $\Xi^i = \{\mathcal{G}_i(\mathbf{v}), \mathcal{C}_i, \phi_i\}$ be the effective MRF between $\mathcal{H}_0$ and $\mathcal{H}_i$, defined for the binary hypothesis in (3.3). The LLR vector has the form

$$\mathbf{L}(\mathbf{Y}_\mathbf{v}) = [\sum_{c\in\mathcal{C}_1} \phi_{1,c}(\mathbf{Y}_c), \ldots, \sum_{c\in\mathcal{C}_{M-1}} \phi_{M-1,c}(\mathbf{Y}_c)]^T. \tag{3.21}$$

We can now apply the Steiner-tree reduction independently for each dimension of the LLR vector to obtain $M-1$ fusion schemes, each computing the function independent of the other. However, this can be wasteful since the forwarded raw measurements can be used to compute many functions simultaneously, i.e., by assigning common processors. To this end, we now define the combined clique set $\mathcal{C}:= \cup_{i=1}^{M-1} \mathcal{C}_i$, with only the resulting maximal cliques retained. We assume that $|\mathcal{C}|$ is polynomial in the number of vertices. We slightly relax the localized processing constraint, in that it is only with respect to cliques in $\mathcal{C}$, i.e., we assume that the processor for any clique c in $\mathcal{C}$ is a member of c. On the other hand, for sub-cliques $c' \subset c$, for $c' \in \mathcal{C}_i$ for some $i = 1, \ldots, M-1$, the processor can lie outside c' but still should belong to c. In the lemma below, we prove a simple result that unique processor assignment reduces routing costs of raw-data forwarding.

Lemma 1 (Unique Processor Assignment) *Given a clique $c \in \mathcal{C}$ and cliques $c' \subset c$, for $c' \in \mathcal{C}_i$ for some $i = 1, \ldots, M-1$, assigning a unique processor for all the cliques c' minimizes the routing cost of forwarding raw measurements.*

Proof: See Appendix 3.B. □

We first consider a special case of M-ary hypothesis testing

$$\begin{aligned} \mathcal{H}_i &: \mathbf{Y}_{\mathbf{v}} \text{ are independent for } i = 0, M-2, \text{ and} \\ \mathcal{H}_{M-1} &: \text{MRF } \Xi = \{\mathcal{G}, C, \phi\}. \end{aligned} \tag{3.22}$$

Hence, the LLR vector in (3.20) is now of the form

$$\mathbf{L}(\mathbf{Y}_{\mathbf{v}}) = [\sum_{i \in \mathbf{v}} w_{1,i} Y_i, \ldots, \sum_{i \in \mathbf{v}} w_{M-2,i} Y_i, \sum_{c \in C} \phi_c(\mathbf{Y}_c)]^T. \tag{3.23}$$

In the lemma below, we establish the structure of optimal fusion scheme that delivers the above LLR vector to the fusion center.

Lemma 2 (M-ary Hypothesis : special case) *For the special case of M-ary simple hypothesis testing in 3.22, the optimal fusion scheme is only dependent on* Υ^**, the optimal scheme has the same aggregation subgraphs for computing different components of the LLR vector in (3.23).*

Proof: Let Proc be the processor assignment for computing the MRF Ξ and let AG_Υ be its aggregation subgraph. Now, since other dimensions of the LLR vector have identical structure of a weighted sum function, there is no need to assign processors for them and their aggregation graphs will be identical, denoted by AG'. Now, AG' is required to at least span Proc, the set of processors; otherwise, it is not feasible to compute the sum function. Hence, the optimal AG' is given by the Steiner tree with Proc as the terminals, which is given by AG_Υ. □

Lemma 2 may not hold for a general M-ary hypothesis. However, in order to obtain a similar simplified scheme for any M-ary hypothesis testing prob-

lem, we impose the constraint of a unique processor assignment. This reduces routing costs in general, since raw measurements are forwarded only once. This also implies that the aggregation subgraph for computing each dimension of the LLR vector is identical. This is because once the processors are assigned, the aggregation subgraph is a Steiner tree over the processors. Hence, this constraint simplifies the design of fusion scheme significantly and implies that even for M-ary hypothesis, we are designing only one fusion scheme, instead of $M-1$ schemes, one for each dimension. Hence, generalization from binary hypothesis to M-ary hypothesis retains the approximability and adds no additional complexity under these additional constraints. Also note that when $M = 2$, the schemes are identical to those in section 3.2.

Theorem 3 (Steiner Reduction for M-ary hypothesis) *The optimal fusion scheme with localized processing constraints for testing M-ary hypothesis has an approximation-factor preserving Steiner-tree reduction (AggApprox_M) shown in Fig.3.13.*

Proof: Due to the local processing constraint, the optimal fusion scheme has unique processor for a clique $c \in C$ and identical aggregation subgraphs for computing each dimension of the LLR vector. Such a scheme is given by (AggApprox_M) if optimal Steiner tree is used. The cost of optimal Steiner tree in this case is $\frac{1}{M-1}$ times the cost of the resulting optimal fusion scheme. For other approximations, the cost of Steiner tree on expanded graph is at least $\frac{1}{M-1}$ cost of the resulting fusion scheme on applying RevMap-All. Hence, the approximation factor is preserved. It runs in polynomial time since $|C|$ is polynomial in number of vertices. □

Require: $\mathbf{v} = \{v_1, \ldots, v_n\}$, v_1= Fusion center,
$C = \{c_0, \ldots, c_{|C|-1}\}$= combined maximal clique set for M-ary hypothesis, $M > 1$
$\bar{N}_g$= Metric closure of comm. graph, $\mathcal{E}^{\text{SP}}$ = Link costs in $\bar{N}_g$,
$\text{ST}(G, \mathcal{L})$ = δ-approx. Steiner tree on $\mathcal{G}$, terminal set $\mathcal{L}$
$G', \mathbf{v}_c \leftarrow \text{Map-All}(\bar{N}_g; \frac{\mathcal{E}^{\text{SP}}}{M-1}, C)$
$\text{DST} = \text{ST}(G', \mathbf{v}_c \cup v_1)$ and directed towards v_1
$\{\text{Proc}, \text{FG}, \text{AG}\} \leftarrow \text{RevMap-All}(\text{DST}; \mathbf{v}_c, \mathbf{v}, C)$
$\text{AG} \leftarrow$ AG with each edge replicated $M - 2$ times
return $\Upsilon \leftarrow \{\text{Proc}, \text{FG}, \text{AG}\}$

Figure 3.13: δ-approximate fusion Υ for M-ary hypothesis (AggApprox_M).

3.7 Conclusion

In this chapter, we have presented an instance of cross-layer design where information from the application layer is used to reduce the routing costs for a statistical inference application. We employ the machinery of approximation algorithms to prove a Steiner tree reduction, enabling us to use any Steiner tree approximation algorithm for minimum cost fusion. Our simulations show a significant saving in cost due to in-network processing compared to routing all the data to the fusion center for proximity-based sparse dependency graph models. In the next chapter, we further reduce routing costs by allowing for sub-sampling of the sensor field to achieve optimal tradeoff between routing costs and the resulting quality of inference.

3.A Overview of Steiner Trees and Approximation Algorithms

In this section, we briefly define the Steiner tree and study its properties. These will be employed to describe our results in the subsequent sections. The material in this section is mainly from [102]. We first define the *Steiner minimal*

tree [95, p. 148] below.

Definition 3 (Steiner tree) *Let $G(\mathbf{v})$ be an undirected graph with non-negative edge weights. Given a set $\mathbf{w} \subset \mathbf{v}$ of terminals, a Steiner tree (ST) is the tree $T \subset G$ of minimum total edge weight such that T includes all vertices in L.*

Finding the Steiner tree is NP-hard and there has been extensive work on finding approximation algorithms. A $0-1$ integer program to find the Steiner tree can be written as

$$\min_{\mathbf{y}} \quad \frac{1}{2}\sum_{i,j\in\mathbf{v}} y(i,j)\mathcal{E}^{\text{SP}}(i,j), \tag{3.24}$$

$$s.t. \quad \sum_{i\in\mathbf{s}, j\notin\mathbf{s}} y(i,j) \geq 1, \forall S \subset V \text{ separating } \mathbf{w}, y(i,j) \in \{0,1\}, \tag{3.25}$$

where we say that $\mathbf{A}$ *separates* $\mathbf{B}$ if $\mathbf{A} \cap \mathbf{B} \neq \emptyset$ and $\mathbf{A} \cap \mathbf{B} \neq \mathbf{B}$. This condition ensures that all the terminals are connected, as illustrated in Fig.3.4.

Definition 4 (Approximation algorithm) *Let Π be a minimization problem and let $\delta : \mathbf{Z}^+ \to \mathbf{Q}^+$, with $\delta \geq 1$. An algorithm A is said to be a factor δ approximation algorithm for Π, if on each instance I, A produces a feasible solution s for I, such that for cost function $\mathcal{E}$,*

$$\mathcal{E}(a) \leq \delta(|I|)opt\_tradeoff(I), \tag{3.26}$$

and the running time of A is bounded by a fixed polynomial in $|I|$.

Many approximation algorithms for finding the Steiner tree require the cost function to be a metric. If the function is not a metric, its *metric closure* [95, p. 58] is considered instead.

Definition 5 (Metric closure) *Let $\mathcal{E}^{SP}(i, j; N_g)$ denote the cost of the shortest path from i to j on N_g. The metric closure on graph N_g, denoted by $\{C_g, \mathcal{E}^{SP}\}$, is defined as the complete graph where the cost of each edge (i, j) is equal to $\mathcal{E}^{SP}(i, j)$.*

A simple *MST heuristic* approximates the Steiner tree over G and terminal set L with the minimum spanning tree spanning the set L, over the metric closure of G. The MST heuristic has an approximation bound of 2 [103]. The best known approximation bound for Steiner tree on graphs is 1.55, derived in [104]. The Steiner tree can be generalized to group Steiner tree, introduced by Reich and Widmayer [101].

Definition 6 (Group Steiner tree) *Let G be an undirected graph with non-negative edge weights. Given groups of vertices $g_i \subset \mathbf{v}$ of terminals, a group Steiner tree is the tree $T \subset G$ of minimum total edge weight such that T includes at least one vertex from each group g_i.*

Since the group Steiner tree is a generalization of the Steiner tree, it is also NP-hard. For a group Steiner tree, polylogarithmic (in the number of groups) approximation algorithms have been proposed [105]. A series of polynomial-time heuristics are described in [106] with worst-case ratio of $O(|g|^{\epsilon})$ for $\epsilon > 0$. The *prize-collecting Steiner minimal tree* (PCST) [107] is a generalization of the Steiner tree, is defined as the tree rooted at a specified vertex v_1 that minimizes the sum of the costs of the edges in the tree plus the penalties of the vertices not spanned by the tree. It is formally defined below.

Definition 7 (PCST) *Given an undirected graph $G = (\mathbf{v}, E)$, a root vertex $v_1 \in \mathbf{v}$, nonnegative edge costs $\mathcal{E}_e^{SP} \geq 0, e \in E$, and nonnegative vertex penalties $\pi_i > 0, i \in \mathbf{v}$, the PCST is the tree $T^* = (\mathbf{v}^*, E^*)$ rooted at v_1 such that*

$$T^* = \arg \min_{T=(\mathbf{v}', E')} \{ \sum_{e \in E'} \mathcal{E}_e^{SP} + \sum_{i \notin \mathbf{v}'} \pi_i \}. \tag{3.27}$$

Note that in the PCST when we set the penalties for the set of terminals as infinity and zero for other nodes, it reduces to a Steiner tree. The penalty associated with a node represents forgone profits by not selecting it. To cast the fusion schemes in terms of a Steiner tree or prize collecting Steiner tree problem, we utilize some *approximation-factor preserving reductions* [102].

3.B Proof of Lemma 1

Assume the lemma is not true. Let c be a non-trivial clique in the combined set C. This implies that c occurs in one of the effective MRF, say C_i for some $i = 1, \ldots, M-1$, since C is the union of all C_i with only maximal cliques retained. Let Proc(c') represent the processor assigned to compute the potential of some non-trivial clique $c' \subset c$ and $c' \in \Xi_j$, the effective MRF Ξ_j, for $j = 1, \ldots, M-1$. By the local processing constraint, we have Proc(c') $\subset c$. This implies that $\cap_{c'}$Proc(c') is either Proc(c) or an empty set. For the first case, since measurements from all the nodes in c except Proc(c) have to be routed to Proc(c), routing costs in the forwarding subgraph (FG) are reduced by replacing Proc(c') with Proc(c) as the sole processor for $j = 1, \ldots, M-1$. This is because the new FG is contained in the original FG. Even for the second case, the routing costs are reduced. This is

because

$$
\begin{aligned}
& \sum_{c' \in C_j, c' \subset c, |c'|>1} \sum_{k \in c', k \neq \mathrm{Proc}(c')} \mathcal{E}^{\mathrm{SP}}(k, \mathrm{Proc}(c')) \\
= & \sum_{c' \in C_j, c' \subset c, |c'|>1, c' \neq c} \sum_{k \in c', k \neq \mathrm{Proc}(c')} \mathcal{E}^{\mathrm{SP}}(k, \mathrm{Proc}(c')) \\
& + \sum_{k \in c, k \neq \mathrm{Proc}(c)} \mathcal{E}^{\mathrm{SP}}(k, \mathrm{Proc}(c)) \\
\geq & \sum_{k \in c, k \neq \mathrm{Proc}(c)} \mathcal{E}^{\mathrm{SP}}(k, \mathrm{Proc}(c)) \qquad (3.28)
\end{aligned}
$$

□

CHAPTER 4

COST-PERFORMANCE TRADEOFF FOR INFERENCE

This chapter considers selection of sensors to achieve optimal cost-performance tradeoff for inference. The costs are incurred in routing and aggregating the selected subset of sensor measurements, and the performance is in terms of the probability of error in inferring the correct hypothesis at the fusion center, given the aggregated data. The contributions are three fold. First, we propose a formulation for optimal sensor selection and in-network fusion known as the *prize-collecting data fusion* (PCDF). Second, we prove its reduction to a known optimization problem for certain correlation structures. Third, for general correlation, we propose two heuristics, and study their performance through simulations.

When the sensor measurements are i.i.d. and the number of sensors goes to infinity, PCDF reduces to an optimization problem known as the *prize-collecting Steiner tree* (PCST) [107]. It is defined as the sub-tree rooted at a specified vertex (fusion center in our case) that minimizes the sum of edge costs in the tree plus the penalties of the nodes not spanned by it. For PCDF with i.i.d. data, the node penalties are uniform, and given by the single-letter Kullback-Leibler divergence (KLD).

For a special class of dependency graphs of Markov random field models, a constrained form of PCDF asymptotically reduces to PCST on an augmented graph, where the augmentation involves adding new nodes and edges to account for increase in aggregation costs due to the presence of correlation. In general, finding the constrained PCDF is NP-hard and we resort to approximations via the PCST reduction. The approximation ratio ρ of any polynomial-time

algorithm guarantees that its output is no worse than ρ times the optimal value. We give an approximation algorithm where the approximation ratio depends only on the number of "profitable" cliques in the dependency graph.

We then develop group selection heuristics for general correlation structures based on the above approximation, viz., component selection and clique selection, and study their performance through simulations. It is observed that the heuristics perform substantially better than the optimal selection scheme which routes the selected measurements to the fusion center without any aggregation at the intermediate nodes. Hence, our approach of incorporating aggregation into the sensor selection formulation substantially reduces routing costs leading to efficient selection policies. We then study the influence of node topology and observe that at sparse spatial dependencies, a clustered node placement achieves better cost-performance tradeoff compared to a uniform placement. These results have direct implications on designing good node placement strategies for cost-performance tradeoff.

4.0.1 Problem Formulation

The goal of this chapter is to select an optimal sensor subset[1] $\mathbf{v}_s \subset \mathbf{v}$, given the entire set $\mathbf{v}$, and to incorporate in-network aggregation of the measurements $\mathbf{Y}_{\mathbf{v}_s}$ before delivery to the fusion center $v_1 \in \mathbf{v}$. It is not possible to quantify inference performance under arbitrary aggregation. Hence, we limit ourselves to aggregation schemes which guarantee the same inference performance as the centralized scheme, i.e., *as if* the fusion center had direct access to the selected measurements $\mathbf{Y}_{\mathbf{v}_s}$. In this case, there is no performance loss due to aggregation at the

[1]The unselected nodes can still function as routers and forward data.

intermediate nodes. In statistical theory, a *sufficient statistic* is a well-behaved function of the data, which is as informative as the raw data for inference [108]. Hence, a scheme which computes and delivers a sufficient statistic results in no loss of inference performance due to aggregation.

We assume that the optimal Neyman-Pearson (NP) detector is used at the fusion center, and that the inference performance is measured by the NP type-II error probability P_M. We are thus interested in subset selection $\mathbf{v}_s \subset \mathbf{v}$ and design of aggregation scheme $\Upsilon(\mathbf{v}_s)$ delivering a sufficient statistic of its measurements $\mathbf{Y}_{\mathbf{v}_s}$ such that optimal linear tradeoff is achieved between the total routing costs $\mathcal{E}(\Upsilon(\mathbf{v}_s))$ and a *penalty* function π, based on the NP type-II error $P_M(\mathbf{v}_s)$,

$$\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \gamma\pi) := \min_{\mathbf{v}_s \subset \mathbf{v}, \Upsilon(\mathbf{v}_s)} \Big[\mathcal{E}(\Upsilon(\mathbf{v}_s)) + \gamma\pi(\mathbf{v} \setminus \mathbf{v}_s)\Big], \ \gamma > 0 \tag{4.1}$$

where $\mathbf{v} \setminus \mathbf{v}_s := \{i : i \in \mathbf{v}, i \notin \mathbf{v}_s\}$ and π is given by

$$\pi(\mathbf{v} \setminus \mathbf{v}_s) := \log \frac{P_M(\mathbf{v}_s)}{P_M(\mathbf{v})} > 0, \quad \forall \mathbf{v}_s \subset \mathbf{v}. \tag{4.2}$$

When we select all the sensors ($\mathbf{v}_s = \mathbf{v}$), (4.2) evaluates to zero, and there is no loss in performance since no measurement is dropped. On the other hand, for a proper subset ($\mathbf{v}_s \subsetneq \mathbf{v}$), we incur a loss in performance and hence, pay a positive penalty in terms of the fraction of increase in error probability due to non-selection of nodes in $\mathbf{v} \setminus \mathbf{v}_s$. Since we collect prizes or penalties for nodes not selected, and incorporate fusion over the selected data, we will henceforth refer to the optimal solution in (4.1) as the prize-collecting data fusion (PCDF) scheme.

The parameter γ is known as the *tradeoff factor*, and is used to adjust the relative importance of cost and performance. Note that the optimization in (4.1) is the Lagrangian dual for the problem of finding the optimal fusion scheme under a constraint on the inference performance or vice versa. Hence, once we have an algorithm to find the (approximate) solution to (4.1), we can use it in the constrained optimization problems. This aspect is however not studied in this chapter, and we will limit to finding solutions to (4.1). Denote the objective in (4.1) as

$$\text{obj}(\mathbf{v}_s, \Upsilon(\mathbf{v}_s); \mathbf{v}, \mathcal{E}, \gamma\pi) := \Big[\mathcal{E}(\Upsilon(\mathbf{v}_s)) + \gamma\pi(\mathbf{v} \setminus \mathbf{v}_s)\Big], \tag{4.3}$$

and the optimal node subset and fusion scheme by

$$[\mathbf{v}_*, \Upsilon_*(\mathbf{v}_*)] := \arg \min_{\mathbf{v}_s \subset \mathbf{v}, \Upsilon(\mathbf{v}_s)} \text{obj}(\mathbf{v}, \mathcal{E}, \gamma\pi). \tag{4.4}$$

When the tradeoff factor is sufficiently large ($\gamma \to \infty$), the optimal tradeoff problem in (4.1) reduces to minimum cost fusion, considered in [2], where optimal inference is required and hence, all the nodes are selected, and the goal is to find the fusion scheme which minimizes the total routing costs while ensuring delivery of a sufficient statistic to the fusion center. When the tradeoff factor is sufficiently small ($\gamma \to 0$), none of the nodes are selected.

$$\lim_{\gamma \to 0} \mathbf{v}_*(\mathbf{v}, \mathcal{E}, \gamma\pi) \to \emptyset, \quad \lim_{\gamma \to \infty} \mathbf{v}_*(\mathbf{v}, \mathcal{E}, \gamma\pi) \to \mathbf{v}.$$

4.0.2 Preliminary Observations & Results

For binary hypothesis testing, the log-likelihood ratio (LLR) is *minimally* sufficient and represents maximum reduction in dimensionality of raw data. It is given by

$$L(\mathbf{Y}_{\mathbf{v}_s}) := \log \frac{f_{\mathbf{v}_s}(\mathbf{Y}_{\mathbf{v}_s}; \mathcal{H}_0)}{f_{\mathbf{v}_s}(\mathbf{Y}_{\mathbf{v}_s}; \mathcal{H}_1)}, \tag{4.5}$$

where $f_{\mathbf{v}_s}(\mathbf{Y}_{\mathbf{v}_s}; \mathcal{H}_j)$ is the pdf of the measurements $\mathbf{Y}_{\mathbf{v}_s}$ under hypothesis $\mathcal{H}_j$. Hence, the optimal aggregation scheme in (4.1), for a given node subset $\mathbf{v}_s$, is a scheme $\Upsilon(\mathbf{v}_s)$ computing and delivering $L(\mathbf{v}_s)$ to the fusion center with minimum total cost $\mathcal{E}(\Upsilon(\mathbf{v}_s))$.

For the penalty function in (4.2), in general, the error probability P_M does not have a closed form, and hence, an analytical solution to (4.1) is not tractable. We focus on the large-network scenario, where the error probability P_M can be approximated by the *error exponent* [108]. When the type-II error $P_M(\mathbf{v})$ decays exponentially with the sample size $|\mathbf{v}|$, for a fixed type-I error, the NP error exponent is given by

$$\mathcal{D} := -\lim_{|\mathbf{v}| \to \infty} \frac{1}{|\mathbf{v}|} \log P_M(\mathbf{v}). \tag{4.6}$$

We will see that we can replace the error probability P_M in (4.2) by an expression based on the error exponent in (4.6), and yet achieve optimality with respect to (4.1), as the number of nodes goes to infinity.

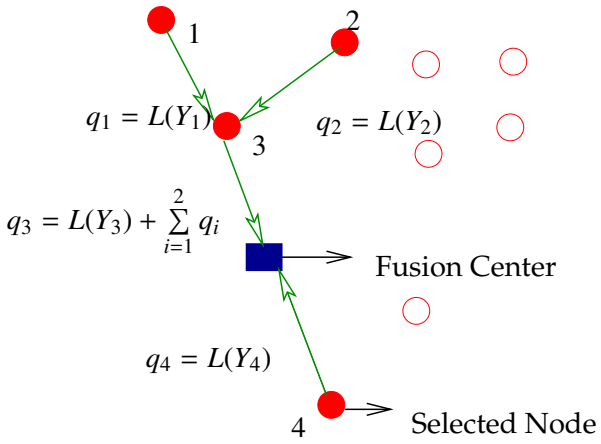


Figure 4.1: Aggregation of i.i.d. measurements along the PCST.

4.1 IID Measurements

We now consider the case when all the sensor measurements are i.i.d. under each hypothesis, $Y_i \overset{i.i.d.}{\sim} f(Y;\mathcal{H}_j)$, for $j = 0, 1$. We first solve a different optimization problem based on (4.6) and then prove its asymptotic convergence to (4.1).

For i.i.d. data, from Stein's Lemma [108, Thm. 12.8.1], the exponent $\mathcal{D}$ in (4.6) is the Kullback-Leibler divergence (KLD)

$$\mathcal{D} = D(f(Y_1;\mathcal{H}_0)||f(Y_1;\mathcal{H}_1)) := \int_y \log \frac{f(y;\mathcal{H}_0)}{f(y;\mathcal{H}_1)} f(y;\mathcal{H}_0)dy$$

We now consider a new penalty function which assigns uniform penalty to each unselected node equal to the KLD D. Hence, if $\mathbf{v}_s$ is the selected subset, the penalty is given by

$$\pi^{iid}(\mathbf{v} \setminus \mathbf{v}_s) := [|\mathbf{v}| - |\mathbf{v}_s|]D, \tag{4.7}$$

First, we establish that the optimal solution under the penalty function π in (4.2) is the same as the optimal solution with penalty π^{iid}, as the number of nodes goes to infinity.

Theorem 4 (Asymptotic optimality of PCST for i.i.d. data) *Under bounded link costs, we have*

$$\lim_{|\mathbf{v}|\to\infty} \frac{opt\_tradeoff(\mathbf{v}, \mathcal{E}, \gamma\pi)}{opt\_tradeoff(\mathbf{v}, \mathcal{E}, \gamma\pi^{iid})} \to 1, \quad \forall \gamma > 0. \tag{4.8}$$

Proof: See Appendix 4.A. □

Hence, it suffices to solve the optimization with π^{iid} instead of π for asymptotic networks, given by

$$\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \gamma\pi^{iid}) := \min_{\mathbf{v}_s \subset \mathbf{v}, \Upsilon(\mathbf{v}_s)} \Big[\mathcal{E}(\Upsilon(\mathbf{v}_s)) + \gamma[|\mathbf{v}| - |\mathbf{v}_s|]D\Big]. \tag{4.9}$$

In order to incorporate in-network aggregation in (4.9), we need an explicit form for $L(\mathbf{Y}_{\mathbf{v}_s})$ since it needs to be computed by the fusion scheme. For i.i.d. data, it is

$$L(\mathbf{Y}_{\mathbf{v}_s}) = \sum_{i \in \mathbf{v}_s} \log \frac{f(Y_i; \mathcal{H}_0)}{f(Y_i; \mathcal{H}_1)}, \quad \forall \mathbf{v}_s \subset \mathbf{v}, \tag{4.10}$$

which is a simple sum function in the selected nodes. In the theorem below, we prove that the optimal solution to (4.9) is the prize-collecting Steiner tree (PCST).

Theorem 5 (Selection & aggregation of i.i.d. data) *The optimal solution to (4.9) is aggregation along the prize-collecting Steiner tree rooted at the fusion center v_1, and edges directed towards v_1: each node i in the PCST computes and transmits q_i to its immediate successor, given by*

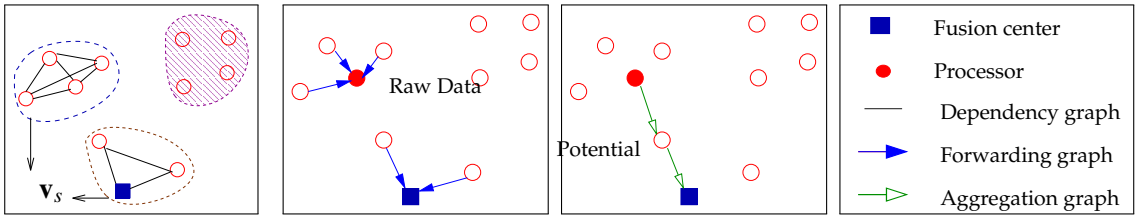
$$q_i = L(Y_i) + \sum_{j \in \mathcal{N}_p(i)} q_j, \tag{4.11}$$

where $\mathcal{N}_p(i)$ is the set of immediate predecessors of i in the directed PCST.

Proof: The LLR sum function in (4.10) over a selected subset $\mathbf{v}_s$ can be computed along the edges of a tree spanning $\mathbf{v}_s$, rooted at and directed towards the fusion center, and $\mathbf{v}_s$ should be selected so as to achieve optimality in (4.9). By definition, it is given by the PCST. □

Hence, the optimal aggregation for i.i.d. data is along the directed PCST. A schematic of the scheme is shown in Fig.4.1. In general, finding the PCST is NP-hard. In [107], an approximation algorithm for the PCST with approximation ratio $2-(|\mathbf{v}|-1)^{-1}$ for any node set $\mathbf{v}$ is proposed, and is referred to as the *Goemans-Williamson* (GW) algorithm.

Theorem 5 establishes the optimality of PCST for the penalty function π^{iid} in (4.7). From Theorem 4, the PCST is also optimal for the penalty function π in (4.2), when the network size goes to infinity. Hence, the PCDF in (4.1) reduces to aggregation along the PCST for i.i.d data, as the network size goes to infinity, and the GW-algorithm approximates the PCST with a proven guarantee of $2 - (|\mathbf{v}| - 1)^{-1}$.



(a) Cliques of dependency graph (b) Forwarding subgraph (c) Aggregation subgraph (d) Legend

Figure 4.2: Computation of the log-likelihood ratio $L(\mathbf{Y}_{\mathbf{v}_s})$ for $\mathbf{v}_s \subset \mathbf{v}$.

4.2 Correlated Measurements: MRF Model

We now generalize the results to the case when the measurements are correlated according to a Markov random field model. Several new challenges arise here. First, the LLR is no longer a simple sum function as in the i.i.d. case in (4.10). Hence, the structure of fusion schemes computing the LLR is not clear. Second, the error exponent $\mathcal{D}$ is no longer the single-letter KLD as for i.i.d data, and hence, the exponent-based penalty may not be separable in the nodes. Third, nodes cannot be assigned uniform penalties as in the i.i.d. case, since they affect inference performance differently in the presence of correlation.

With the above challenges, it is not tractable to solve the PCDF problem, defined in (4.1). Instead, we solve (4.1) under an additional constraint that the subsets $\mathbf{v}_s$ considered are only those that span a sub-collection of cliques of the dependency graph $C_s \subset C$, and is referred to as the *constrained* PCDF,

$$\text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \gamma\pi) := \min_{\substack{\mathbf{v}_s \subset C_s \subset C \\ \Upsilon(\mathbf{v}_s)}} \Big[\mathcal{E}(\Upsilon(\mathbf{v}_s)) + \gamma\pi(\mathbf{v} \setminus \mathbf{v}_s)\Big]. \tag{4.12}$$

In other words, the selection policy is coarser since it selects or rejects cliques of nodes instead of individual ones. Since we are ruling out certain subsets for

selection, we cannot guarantee optimality with respect to (4.1).

4.2.1 In-network Aggregation of LLR

In order to design a fusion scheme for computing the LLR, we use its explicit characterization in (2.14) as,

$$L(\mathbf{Y}_{\mathbf{v}};\Xi) = \sum_{c\in C} \phi_c(\mathbf{Y}_c). \tag{4.13}$$

based on the clique set C of the joint dependency graph, $\mathcal{G}(\mathbf{v}):=\mathcal{G}_0(\mathbf{v}) \cup \mathcal{G}_1(\mathbf{v})$. Comparing the above form with that for i.i.d data in (4.10), we see that correlation increases the complexity of the L.

For any subset $\mathbf{v}_s \subset \mathbf{v}$, its marginal LLR can also be expressed based on the clique set C' of its dependency graph $\mathcal{G}'(\mathbf{v}_s)$

$$L(\mathbf{Y}_{\mathbf{v}_s}) = \sum_{c\in C'} \phi'_c(\mathbf{Y}_c), \tag{4.14}$$

where $\mathcal{G}'(\mathbf{v}_s):=\mathcal{G}'_0(\mathbf{v}_s)\cup\mathcal{G}'_1(\mathbf{v}_s)$, and $\mathcal{G}'_j(\mathbf{v}_s)$ is the dependency graph of the marginal pdf $f_{\mathbf{v}_s}(\mathbf{Y}_{\mathbf{v}_s};\mathcal{H}_j)$, for $j = 0, 1$. In general, $\mathcal{G}'(\mathbf{v}_s)$ is not a subgraph of $\mathcal{G}(\mathbf{v})$ and C' is not contained in C. Hence, the structure of the marginal LLR and its fusion scheme change with the selected set $\mathbf{v}_s$.

We now describe the structure of fusion schemes computing the LLR of a given subset $\mathbf{v}_s$. See Fig.4.2. The issue of optimal selection of $\mathbf{v}_s$ will be considered later. Given the dependency graph $\mathcal{G}'(\mathbf{v}_s)$, the computation is in two stages.

First, the data $\mathbf{Y}_c$ are forwarded from all the members of clique $c \in C'$ to compute its potential $\phi'_c(\mathbf{Y}_c)$ at an assigned *processor*, denoted by Proc(c). The set of links used for such data forwarding in all the cliques form the *forwarding* graph (FG).

In the second stage of LLR computation, all the clique potentials are summed up and delivered to the fusion center, using a set of links referred to as the *aggregation* subgraph (AG). The tuple with the forwarding and aggregation subgraphs of a fusion scheme is the *fusion digraph*, $\mathcal{F}$:={FG, AG}, since it is the complete set of links used by the fusion scheme. The total routing costs of the fusion scheme is

$$\mathcal{E}(\mathcal{F}) = \mathcal{E}(\text{FG}) + \mathcal{E}(\text{AG}). \tag{4.15}$$

For finding the constrained PCDF in (4.12), we thus need to find a fusion scheme which minimizes the sum of routing costs in the two stages of LLR computation.

4.2.2 Error Exponent & Penalty Function

Along the lines of our approach for i.i.d. data, in the constrained PCDF problem in (4.12), we replace the error-probability based penalty π with the error exponent $\mathcal{D}$ for MRF hypothesis testing.

We now provide results for the error exponent $\mathcal{D}$, which is then used to define a penalty function π^{clq} in (4.17) approximating the function π in (4.2), based on the inference error probability.

Theorem 6 (Error Exponent for MRF) *When the sequence of normalized log-likelihood ratio variables is uniformly integrable and converges in probability under the null hypothesis $\mathcal{H}_0$, the error exponent in (4.6) is*

$$\mathcal{D} = p \lim_{n\to\infty} \frac{1}{n} \sum_{c\in C} \mathbb{E}(\phi_c(\mathbf{Y}_c)|\mathbf{v}; \mathcal{H}_0), \tag{4.16}$$

where ϕ_c is the potential function for clique c, C is the MRF clique collection in (4.13) and $\mathbb{E}$ is the expectation under $\mathcal{H}_0$.

Proof: We use the form of LLR in (4.13). See Appendix 4.A. □

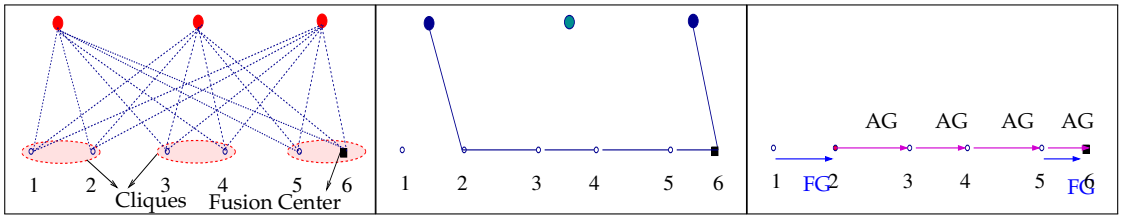
Hence, the exponent is given by the limit of the normalized sum of functions over the dependency cliques. We define a new penalty function π^{clq} based on the error exponent to be used in the optimization in (4.12), where the unselected cliques are assigned penalty

$$\pi^{clq}(C \setminus C_s) := \sum_{c\in C\setminus C_s} (\mathbb{E}(\phi_c(\mathbf{Y}_c)|\mathbf{v}; \mathcal{H}_0))^+, \tag{4.17}$$

and use it instead of the original penalty function π in (4.2) based on the error probability.

4.2.3 Special Case of MRF: Disjoint Cliques

We now provide approximation guarantees and convergence results for (4.12) under a special class of dependency graphs. This in turn inspires the development of a general class of heuristics for any dependency graph in Section 4.3.



(a) Augmented graph using Map-All (b) PCST on augmented graph (c) Scheme to compute LLR via RevMap-Sub

Figure 4.3: Clique selection and fusion via PCST reduction for binary cliques.

We consider the special case when all the cliques in the joint dependency graph $\mathcal{G}(\mathbf{v})$ are disjoint. This can occur for instance, when nodes are placed according to a cluster process and the dependency graph is given by a disk graph. See Section 4.4. Here, the form of the LLR in (4.14) and the exponent in (4.16) are simplified further.

For disjoint cliques, the dependency graph $\mathcal{G}'(\mathbf{v}_s)$ is a subgraph of $\mathcal{G}(\mathbf{v})$, for any node subset $\mathbf{v}_s$ spanning a sub-collection of cliques $C_s \subset C$, and hence,

$$L(\mathbf{Y}_{\mathbf{v}_s}) = \sum_{c \in C_s} \phi_c(\mathbf{Y}_c). \tag{4.18}$$

Hence, it is simpler to design fusion schemes in this case since the dependency structure does not change for different nodes subsets, as long as the nodes span a sub-collection of cliques.

For disjoint cliques, the penalty function for each clique in (4.17) simplifies to the KLD of measurements in clique $c \in C$

$$\pi^{clq}(c) = D(f_c(\mathbf{Y}_c; \mathcal{H}_0) \| f_c(\mathbf{Y}_c; \mathcal{H}_1)) := D_c. \tag{4.19}$$

Hence, if nodes in a clique c is not selected, then a penalty equal to its KLD D_c is paid.

We now prove the asymptotic optimality of using the exponent-based penalty function π^{clq} in (4.19), instead of the original penalty function π in (4.2) in (4.12).

Theorem 7 (Asymptotic Optimality) *When the number of cliques grows with network size* ($|C| \to \infty, \quad as\ |\mathbf{v}| \to \infty$), *and the link costs are bounded, we have*

$$\lim_{|\mathbf{v}|\to\infty} \frac{opt\_tradeoff\_clique(\mathbf{v}, \mathcal{E}, \gamma\pi)}{opt\_tradeoff\_clique(\mathbf{v}, \mathcal{E}, \gamma\pi^{clq})} = 1, \quad \forall \gamma > 0. \tag{4.20}$$

Proof: Along the lines of Theorem 4. See 4.A. □

Hence, using the penalty function π^{clq} in (4.19) instead of π is suitable for networks with large number of cliques. An example where this does *not* occur is when the dependency graph is complete, and has a single clique. We therefor need a sparse dependency graph to guarantee the asymptotic convergence of the constrained PCDF in (4.12) to the optimal solution under penalty π^{clq}. Along the lines of our approach for the i.i.d. case, we now prove that under π^{clq}, the optimal solution reduces to a PCST.

Theorem 8 (PCST Reduction) *$opt\_tradeoff\_clique(\mathbf{v}, \mathcal{E}, \gamma\pi^{clq})$ has an approximation-ratio preserving PCST reduction.*

Proof: By simplifying an integer program. See Appendix 4.A. □

The above result implies that any approximation algorithm for the PCST can be transformed to an approximation for opt_tradeoff_clique($\mathbf{v}, \mathcal{E}, \gamma\pi^{clq}$), with its approximation ratio preserved. One such instance, called the approximate prize-collecting data fusion (Approx_PCDF), is given in Fig.4.4. It builds an approximate PCST on an augmented graph using the GW-algorithm [107].

The augmented graph is given by the function Map-Sub in Fig.4.5, where for each non-trivial clique c (size greater than one) of the dependency graph, it adds a virtual node v_c and connects it to the nodes $v \in \mathbf{v}$. The costs of new edges reflect the cost of forwarding raw data to candidate processors to compute the clique potentials in the first stage of LLR computation, which is not needed for i.i.d. data. Hence, the routing costs are increased in the presence of correlation due to additional complexity of the LLR.

The penalty of each virtual node v_c is $\pi^{clq}(c)$ in (4.19) and the penalties of all nodes $v \in \mathbf{v}$ are set to zero. After building the approximate PCST on the augmented graph, the function RevMap-Sub in Fig.4.6 maps it to a valid output, viz., the set of selected cliques and the fusion scheme to compute its LLR. An example of the PCST reduction is shown in Fig.4.3.

As in the i.i.d. case, an approximate PCST is built on the augmented graph using the GW-algorithm [107]. Since the augmented graph has $|\mathbf{v}| + |C^{nt}|$ number of nodes, where C^{nt} is the set of non-trivial cliques, the approximation ratio of Approx_PCDF(Map-Sub) with respect to opt_tradeoff_clique($\mathbf{v}, \mathcal{E}, \gamma\pi^{clq}$) is $2 - (|\mathbf{v}| + |C^{nt}| - 1)^{-1}$.

We now improve its approximation ratio based on some simple observations regarding the GW-algorithm. Define the collection of *profitable cliques* $C_p \subset C$ as

those generating a net "profit" after reducing their scaled KLD by the costs of raw-data routing to any processor

$$C_p := \{c : c \in C, |c| = 1 \text{ or } |c| > 1 \text{ and } \gamma D_c \geq \min_{i \in \mathbf{v}} \sum_{v_k \subset c_j, k \neq i} \mathcal{E}^{\text{SP}}(v_i, v_k)\}, \tag{4.21}$$

and let Map-Sub′ be the modified version of Map-Sub which only adds virtual nodes for non-trivial profitable cliques, i.e., $c \in C_p, |c| > 1$, instead of adding for all non-trivial cliques, $c \in C^{nt}$, as done by Map-Sub. Below, we give the improved approximation ratio.

Theorem 9 (Improved Approx. Ratio) *On using the Map-Sub′ function, the approximation ratio for Approx_PCDF with respect to opt_tradeoff_clique*$(\mathbf{v}, \mathcal{E}, \gamma\pi^{clq})$ *is*

$$\rho(\textit{Approx\_PCDF(Map-Sub}')) = 2 - \frac{1}{\max(|C_p| - I(v_1 \in C_p), 1)}.$$

Proof: Only profitable cliques can be selected in the optimal solution. See Appendix 4.A. □

Hence, the approximation ratio for Approx_PCDF(Map-Sub′) depends only on the number of profitable cliques $|C_p|$, which may be substantially smaller than the size of the augmented graph $|\mathbf{v}| + |C^{nt}|$ leading to improved approximation guarantees. In fact, when there are no profitable cliques ($C_p = \emptyset$), the algorithm outputs the optimal solution ($\rho = 1$) of not selecting any of the nodes.

Require: $\mathbf{v} = \{v_1, \ldots, v_n\}$ nodes, v_1= Fusion center,
$\mathcal{M} = \{c_0, \ldots, c_{|\mathcal{M}|-1}\}$= Candidate node groups
For Algo=Clique Selection, $\mathcal{M} = C$ is the clique set of $\mathcal{G}(\mathbf{v})$ and $\Pi = \pi^{clq}$ in (4.17).
For Algo=Component Selection, $\mathcal{M}$ is the set of components of $\mathcal{G}(\mathbf{v})$ and $\Pi = \pi\_\text{cmp}$ in (4.22).
N_g= Metric closure of network, $\mathcal{E}^{\text{SP}}$ = Link costs
Π_m = Penalty of group m, γ = tradeoff factor
$\{G', \mathbf{v}_m, \pi\} \leftarrow$ Map-Sub(N_g; $\mathcal{M}, \mathcal{E}^{\text{SP}}, \Pi, \gamma$)
PCST($G; \mathcal{E}^{\text{SP}}, \pi$) = (Approx.) Prize-collecting Steiner tree on G using GW algorithm with cost $\mathcal{E}^{\text{SP}}$, node penalty fn. π
DPCST = PCST(G') directed towards v_1
$\{\mathcal{M}_s, \Upsilon\} \leftarrow$ RevMap-Sub(DPCST; $\mathbf{v}_m$, $\mathbf{v}$, $\mathcal{M}$, Algo)
return $\{\mathcal{M}_s, \Upsilon\}$

Figure 4.4: Approx_PCDF(Map-Sub,Algo) selects groups $\mathcal{M}_s$ and policy Υ.

4.3 Node Selection Heuristics

The results in the previous section inspire the development of two heuristics for a general dependency graph, viz., clique selection and component selection. The Approx_PCDF algorithm in the previous section, based on the PCST reduction, can be generalized as follows: form groups of nodes according to some criterion as candidates for selection, and define a penalty function for not selecting each group. Apply the PCST reduction as before by augmenting the graph with virtual nodes for each group. Using the RevMap-Sub, the output is a selected sub-collection of groups and a fusion scheme which computes a sum function over the selected groups.

The desired output for cost-performance tradeoff is however *not* a fusion scheme for computing a sum function, but for computing the marginal LLR of the selected nodes. As we discussed in Section 4.2.2, the LLR structure (dependency graph) changes with the selected node set in general. We now overcome this hurdle by grouping nodes in such a manner that the LLR of any selected

function MAP-SUB($N_g(\mathbf{v}); \mathcal{M}, \mathcal{E}^{\text{SP}}, \Pi, \gamma$)
  nbd($v; G$) = Neighborhood of v in undirected $\mathcal{G}$
  $G' \leftarrow N_g$, $\mathbf{v}_m \leftarrow \emptyset$, $\pi(v_i) \leftarrow 0$, $\forall v_i \in \mathbf{v}$
  for $j \leftarrow 0$ to $|\mathcal{M}| - 1$ **do** ▷ Let $\mathbf{v}$ and $\mathcal{M}$ be ordered
    if $|m_j| > 1$ **then**
      $\mathbf{v}_m \leftarrow v_{n+j}$
      Add new node v_{n+j} to G'
      Assign penalty $\gamma\pi(v_{n+j}) \leftarrow \gamma\Pi_{m_j}$,
      for all $v_i \in \mathbf{v}$ **do**
        Add node v_i to nbd($v_{n-1+j}; G'$)
        $\mathcal{E}^{\text{SP}}(v_{n-1+j}, v_i; G') \leftarrow \sum_{v_k \subset c_j, k \neq i} \mathcal{E}^{\text{SP}}(v_i, v_k; N_g)$
      end for
    else
      $V_m \leftarrow v_i$, $\pi(v_i) \leftarrow \gamma\Pi_{m_j}$, $v_i \subset m_j$ ▷ 1-groups
    end if
  end for
return $\{G', \mathbf{v}_m, \pi\}$
end function

Figure 4.5: Map-Sub($N_g; \mathcal{M}, \mathcal{E}^{\text{SP}}, \Pi, \gamma$) adds nodes for each non-trivial group.

sub-collection of groups is indeed a sum function over those groups.

For general dependency graphs, such groups are given by the components of the dependency graph, i.e., if all or none of the nodes belonging to each component of the graph are selected, then the LLR of the selected subset is a simple sum function over the selected components

$$L(\mathbf{Y}_{\mathbf{v}_s}) = \sum_{v \subset m, m \in \mathcal{M}} L(\mathbf{Y}_m),$$

where $m \in \mathcal{M}$ is a component in the dependency graph. Moreover, we can define penalty for each component by collecting the terms of the error exponent in (4.16) consisting of all the cliques contained in it, given by

$$\pi\_\text{cmp}(m) := \sum_{c \subset m, c \in \mathcal{C}} \mathbb{E}[\phi_c(\mathbf{Y}_c); \mathcal{H}_0] = D_m, \tag{4.22}$$

function REVMAP-SUB(G'; $\mathbf{v}_c$, $\mathbf{v}$, $\mathcal{M}$,Algo)
  $\mathcal{N}_s(v;\mathcal{G})$, $\mathcal{N}_p(v;\mathcal{G})$ = Imm. successor, predecessor
  $< i, j >$= Directed edge from i to j
  Initialize $G \leftarrow G'$, $\mathcal{M}_s \leftarrow \emptyset$
  for all $v_j \in \mathbf{v}_c$ with $\mathcal{N}_s(v_j; G') \neq \emptyset$ **do**
    if $j > n-1$ **then**
      $k \leftarrow j - n + 1$, $\mathcal{M}_s \leftarrow \mathcal{M}_s \cup m_k$
      Proc(m_k) $\leftarrow \mathcal{N}_s(v_j; G')$, for $m_k \in \mathcal{M}$,
      $v_j \leftarrow c_k \setminus$ Proc(m_k), Delete $< v_j, \text{Proc}(m_k) >$ in G, add $< v_j, \text{Proc}(m_k) >$, mark them
      if $\mathcal{N}_p(v_j; G) \neq \emptyset$ **then** Replace $< \mathcal{N}_p(v_j), v_j >$ in $\mathcal{G}$ with edges $< \mathcal{N}_p(v_j), \text{Proc}(m_k) >$
      end if
    else
      Proc(m_l) $\leftarrow v_j$, for $v_j \subset m_l$, $\mathcal{M}_s \leftarrow \mathcal{M}_s \cup m_l$
    end if
  end for
  FG $\leftarrow$ Marked edges of G, AG $\leftarrow G \setminus$ FG
  Retain only one edge in FG if there are parallel links
  Let $\mathbf{v}$(Proc) be set of all processors
  Let $\mathcal{A} \leftarrow$ nodes in $\mathbf{v}$ spanning the groups $\mathcal{M}_s$
  if Algo=Clique Selection **then**
    Let C' be clique set of $\mathcal{G}'(\mathcal{A})$
    for all $c \in C' \setminus \mathcal{M}_s$ **do**
      $\text{Proc}(c) \leftarrow \arg \min_{i \in \mathbf{v}(\text{Proc})} \sum_{\substack{j: j \subset c \\ < j,i > \notin \text{FG}}} \mathcal{E}^{\text{SP}}(i, j)$
      Add $< j, \text{Proc}(c) >$, $j \subset c \setminus \text{Proc}(c)$ to FG if not already present
    end for
    $\mathcal{M}_s \leftarrow C'$
  end if
  $\Upsilon \leftarrow \{\text{Proc}, \text{FG}, \text{AG}\}$
return $\{\mathcal{M}_s, \Upsilon\}$
end function

Figure 4.6: RevMap-Sub(G'; $\mathbf{v}_c$, $\mathbf{v}$, $\mathcal{M}$) selects groups $\mathcal{M}_s$ and policy Υ.

where D_m is the KLD of the component m, and the penalties of different components are additive. We term such a policy considering different components of the dependency graph as candidates for selection as the *component selection* heuristic.

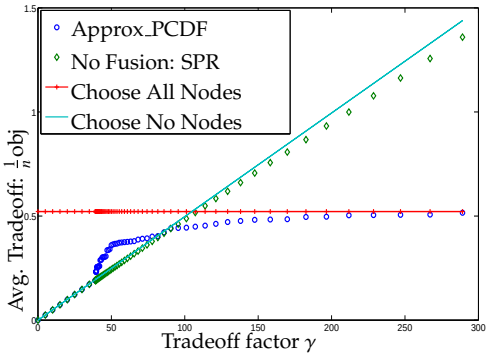


Figure 4.7: Objective Value obj vs. Tradeoff Factor under Uniform Placement

Optimal cost-performance tradeoff is however not guaranteed for the component selection heuristic since we may be severely limiting our choices of node subsets for selection. For instance, if the graph has a single component, then the heuristic reduces to a binary decision of selecting all or none of the nodes. We now propose another heuristic which may perform better in such instances.

As in the previous section, we consider the cliques of the dependency graph as the groups, i.e., candidates for selection, and the penalty function for each clique in (4.17). This is referred to as the clique selection heuristic. However, as noted, the output fusion scheme is not guaranteed to compute the marginal LLR of the selected node set which is a requirement for inference. In Fig.4.6, we add additional lines from (17) to (26) to ensure that the marginal LLR is indeed computed. For each new clique in the marginal dependency graph, not present in the dependency graph over all the nodes $\mathcal{G}(\mathbf{v})$, we ensure that its clique potential is computed by adding edges from its members to a processor to the forwarding subgraph (FG) of the fusion scheme. However, since new edges are added, routing costs increase, and we can no longer provide optimality results for the clique selection heuristic for a general MRF, as we did in the previous section.

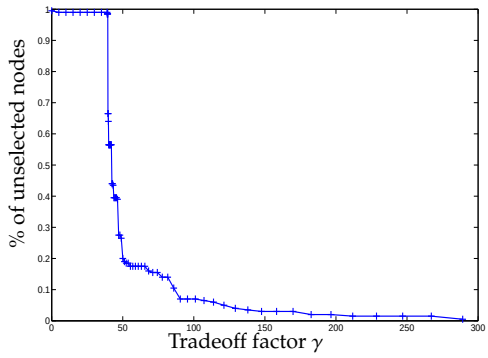


Figure 4.8: % of Nodes Not Selected vs. Tradeoff Factor, Uniform Placement

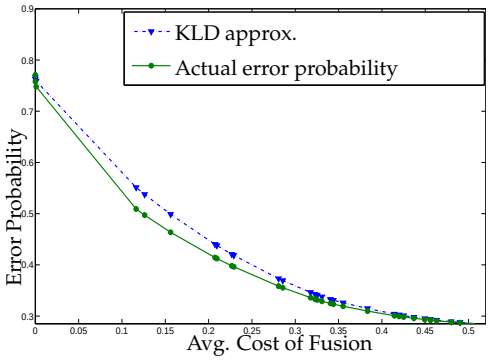


Figure 4.9: Cost vs. Performance for Selected Set $\mathbf{v}_s$ under Uniform Placement

The component and clique selection policies represent group selection of nodes with aggregation for efficient cost-performance tradeoff. The component selection heuristic can be viewed as coarse selection or rejection of nodes as a full component, while the clique selection heuristic is more fine-grained, depending on the graph. For graphs having very few components, and yet, a large number of cliques, we expect the clique selection policy to have better cost-performance tradeoff than component selection, since there are more candidates for selection. On the other hand, for sparse graphs with large number of components, we expect the component selection policy to do better, and this is validated by our simulations.

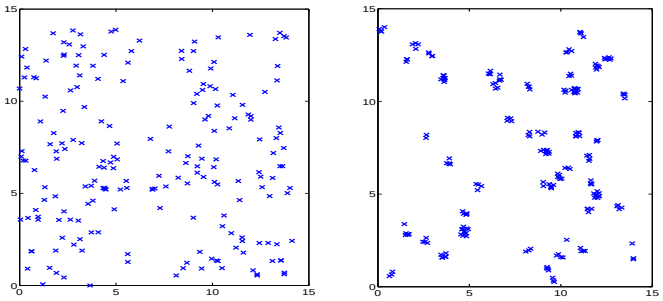


Figure 4.10: Samples of i.i.d uniform placement and Matern cluster process.

4.4 Numerical Analysis

4.4.1 Simulation Environment

We assume that the sensor measurements are Gaussian under either hypothesis with the same covariance matrix

$$\mathbf{Y}_{\mathbf{v}} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\mathbf{v}}), \quad \text{under } \mathcal{H}_i, \ i = 0, 1. \tag{4.23}$$

This scenario arises when the sensors measure a deterministic signal with additive (correlated) Gaussian noise under each hypothesis. The KLD D and the type-II error probability P_M have closed forms for Gaussian variables [91, 108]. We fix $\boldsymbol{\mu}_0 = \mathbf{0}, \boldsymbol{\mu}_1 = 0.1\mathbf{I}$ and the type-I error $\alpha = 0.2$.

In our setup, n (expected) number of nodes are distributed in a square. We consider two node placement distributions: uniform and Matern cluster process[2] [109]. See Fig.4.10. The routing cost between any two nodes i and j for

[2]Here, a parent Poisson process first generates points. A child Poisson process then generates nodes in a disc around each point of the parent process.

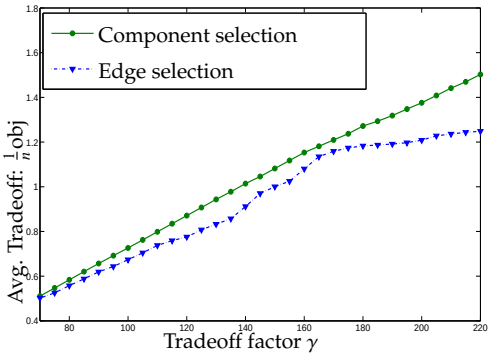


Figure 4.11: Group Selection for Gaussian data, 60 runs, $\delta = 1.2, n = 50$

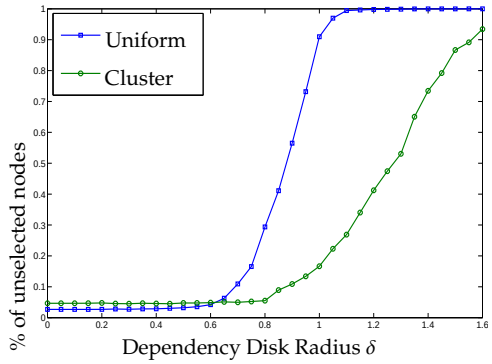


Figure 4.12: Component Selection for Gaussian data, 60 runs, $\gamma = 140, n = 200$

direct transmission is given by the power-weighted distance $|i, j|^{\nu}$. We present the result when the set of feasible direct connections is the complete graph and the path-loss $\nu = 2$: similar trends were observed for any connected graph and $\nu \in [2, 4]$.

4.4.2 Results: IID Measurements

We first consider the case when all the measurements are i.i.d. conditioned on each hypothesis with unit variance ($\mathbf{\Sigma_v} = \mathbf{I}$). We compare the performance of our fusion scheme Approx_PCDF in Fig. 4.4 with the following simple schemes:

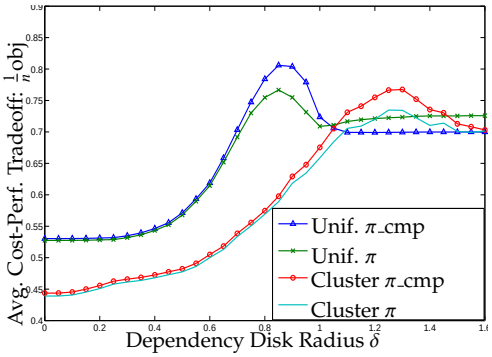


Figure 4.13: Component Selection for Gaussian data, 60 runs, $\gamma = 140, n = 200$

choosing all the nodes and conducting fusion along the MST, choosing none of the nodes (paying penalty for all the nodes), and additionally, optimal selection with no aggregation, i.e., routing all the selected data to the fusion center via the shortest path routes (SPR). It is given by the set of "profitable" nodes

$$\mathbf{v}_*^{\mathrm{SPR}}(V, \mathcal{E}, \gamma\pi^{iid}) = \{i : i \in \mathbf{v}, \gamma D > \mathcal{E}^{\mathrm{SP}}(i, v_1)\}, \tag{4.24}$$

where $\mathcal{E}^{\mathrm{SP}}$ is the cost of shortest path. In Fig.4.7, we find that the tradeoff function obj in (4.3) for Approx_PCDF is significantly better than those for the other schemes. Hence, incorporating fusion into cost-performance tradeoff significantly reduces the costs and achieves better tradeoff.

Fig.4.8 shows that more nodes are selected by Approx_PCDF as the tradeoff factor γ increases, since the penalty is given by $\gamma\pi$. In Fig.4.9, we plot the average (per-node) routing cost for aggregation of selected measurements versus the resulting error probability for Approx_PCDF under different γ. We see that the exponent-based approximation e^{-nD} is close to the actual error probability P_M.

4.4.3 Results: Correlated Measurements

We employ the GMRF model in [110], where the dependency graph $\mathcal{G}(\mathbf{v})$ is a disk graph[3] with radius δ and the coefficients of the potential matrix $\mathbf{A_v}$:=$\mathbf{\Sigma_v^{-1}}$ are given by

$$A_{\mathbf{v}}(i,j) = \begin{cases} 1 - \sum\limits_{k:(i,k)\in\mathcal{G}(\mathbf{v})} A(i,k), & i = j, \\ -2(1 - \frac{|i,j|}{\delta}), & j \neq i, \text{dist}(i,j) \leq \delta, \\ 0, & \text{o.w.} \end{cases} \tag{4.25}$$

We find that the positive definiteness is ensured in the above model since $\mathbf{A}$ is diagonally dominant. For Gaussian measurements, the maximum clique size is two and higher order clique potentials are zero (see Chapter 2). Hence, the clique selection heuristic in Fig.4.4 reduces to selection of the dependency edges, and is called the *edge selection policy*.

We find that for the above model, the penalty for the entire node set given by the KLD $D_{\mathbf{v}}$ does not change with the disc radius δ or the node placement. However, the configuration of cliques and their KLD indeed depend on these factors and influence the nature of selected set.

In Fig.4.11, we compare the component and edge selection heuristics under uniform placement. We fix the disk radius $\delta = 1.2$ and here, the disk graph is connected (single component). We expect the edge selection heuristic to perform better since it has more choices here when compared to component selection, which has to make a binary choice whether to select all or none of the

[3] A disk graph has edges between nodes within δ inter-node distance.

nodes. We find that for γ shown in the figure, this indeed is the case; the edge selection heuristic performs better and selects some nodes, while the component selection heuristic selects none of the nodes thereby incurring high penalty in terms of error probability.

In Fig.4.12 and Fig.4.13, we study the influence of node placement on our heuristics, and consider uniform and Matern cluster process with component selection heuristic. We observe that at low values of δ, the clustered process is more efficient; here, more nodes are chosen, and the tradeoff function obj is lower. However, as δ increases, the two processes have nearly the same performance. As in the i.i.d. case, the exponent-based penalty $\pi_$cmp is close to π, based on the error probability in all the instances.

We can provide an intuitive explanation for the above behavior. At low dependency (small values of the disk radius δ), clustering the nodes is more efficient than uniform placement since it leads to significantly smaller number of components, thereby providing more choices to the component selection heuristic. Moreover, the routing costs within the components are also significantly reduced upon clustering since nodes are nearer, and hence, more nodes are selected leading to improved tradeoff. However, as δ increases, there are fewer and larger components, leading to increased routing costs and fewer choices for selection. Hence, the cluster process is a good node-placement strategy for achieving efficient cost-performance tradeoff at sparse spatial dependencies, and our heuristic has good performance in this regime.

4.A Proofs

Proof of Theorem 4

It is easy to see that $|\mathbf{v}_*(\mathbf{v}, \mathcal{E}, \gamma\pi'')|$ is monotonic in the tradeoff factor $\gamma > 0$, for both penalty functions $\pi'' = \pi, \pi^{iid}$ in (4) and (10). Hence, $\exists\gamma_1$ such that $\forall\gamma \geq \gamma_1$, we have

$$\lim_{|\mathbf{v}|\to\infty} \frac{|\mathbf{v}_*(\mathbf{v}, \mathcal{E}, \gamma\pi'')|}{|\mathbf{v}|} = 1,$$

for both functions $\pi'' = \pi, \pi^{iid}$. The actual value of γ_1 indeed depends on the system parameters. For $\gamma \geq \gamma_1$, the average penalty goes to zero for both functions $\pi'' = \pi, \pi^{iid}$ since almost all nodes are selected and all edge costs to be bounded. Hence,

$$\lim_{|\mathbf{v}|\to\infty} \frac{1}{|\mathbf{v}|}\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \gamma\pi'') = \lim_{|\mathbf{v}|\to\infty} \frac{1}{|\mathbf{v}|}\mathcal{E}(\Upsilon_*(\mathbf{v}_*(\mathbf{v}, \mathcal{E}, \pi''))) = \lim_{|\mathbf{v}|\to\infty} \frac{1}{|\mathbf{v}|}\mathcal{E}(\Upsilon_*(\mathbf{v})),$$

since each edge cost is assumed bounded. Hence, we have for

$$\lim_{|\mathbf{v}|\to\infty} \frac{\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \gamma\pi)}{\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \gamma\pi^{iid})} = 1, \quad \forall\gamma > \gamma_1 > 0.$$

Now for a fixed $m < 1$, consider $\gamma \leq \gamma_2(m)$ such that

$$\limsup_{|\mathbf{v}|\to\infty} \frac{|\mathbf{v}_*(\mathbf{v}, \mathcal{E}, \gamma\pi''|)}{|\mathbf{v}|} = m < 1, \quad \pi'' = \pi, \pi^{iid}.$$

Hence, we limit our search over a collection of sets $\mathcal{A}_m$:=$\{\mathbf{v}_s : \frac{|\mathbf{v}_s|}{|\mathbf{v}|} \leq m\}$ for the optimal solution opt_tradeoff$(V, \mathcal{E}, \gamma\pi'')$ for both $\pi'' = \pi, \pi^{iid}$ in this case. For i.i.d. measurements, from the existence of exponent we have

$$[|\mathbf{v}| - |\mathbf{v}_s|]\mathcal{D} - \epsilon \leq \log \frac{P_M(\mathbf{v}_s)}{P_M(\mathbf{v})} \leq [|\mathbf{v}| - |\mathbf{v}_s|]\mathcal{D} + \epsilon, \tag{4.26}$$

Define new penalty functions

$$\pi^{\pm}(\mathbf{v}\backslash\mathbf{v}_s) := [|\mathbf{v}| - |\mathbf{v}_s|](\mathcal{D} \pm \frac{\delta}{|\mathbf{v}|}), \quad \forall \mathbf{v}_s \in \mathcal{A}$$

where $\delta(m) := \limsup_{|\mathbf{v}|\to\infty, \mathbf{v}_s \in \mathcal{A}_m} \frac{\epsilon|\mathbf{v}|}{|\mathbf{v}|-|\mathbf{v}_s|} = \frac{\epsilon}{1-m} < \infty$.

For the same edge costs, a uniformly smaller penalty function for each node subset results in a lower value of the optimal solution. Hence, we have

$$\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi^-) \leq \text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi'') \leq \text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi^+),$$

for $\pi'' = \pi, \pi^{iid}$. We now claim that if all the edge costs are unique and satisfy $\mathcal{E}_e \neq \gamma\mathcal{D}$, then for some n_0

$$\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi^-) = \text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi^+), \quad \forall |\mathbf{v}| > n_0. \tag{4.27}$$

Note that if we substitute the penalty function π^+ with π^-, we uniformly reduce the node penalties by $\frac{2\delta}{n}$, where $n = |\mathbf{v}|$. This implies that some nodes from the optimal node set with penalty function π^+ (abbreviated as $\mathbf{v}_*^+$) may be potentially removed. We claim that none of the nodes are removed for all $n > n_0$, for some n_0 when the edge costs are all unique and not equal to node penalty. In this case, we can always find a small perturbation of the node penalty without changing the optimal solution. For example, consider a leaf node in $\mathbf{v}_*^+$, from cardinality one test [111], if its edge $\mathcal{E}_e > \gamma(\mathcal{D} - \frac{\delta}{n})$, then it cannot be in $\mathbf{v}_*^-$. But since it is in $\mathbf{v}_*^+$, we have

$$\mathcal{E}_e \leq \gamma(\mathcal{D} + \frac{\delta}{n}).$$

Since we have assumed that $\mathcal{E}_e \neq \gamma\mathcal{D}$, we can find some n_0 such that for all $n > n_0$

$$\mathcal{E}_e \leq \gamma(\mathcal{D} \pm \frac{\delta}{n}).$$

Hence, the leaf nodes are the same in $\mathbf{v}_*^-$ and $\mathbf{v}_*^+$ for $n > n_0$. Similarly, we can apply general cardinality tests in [111] such that for large n, the vertices in $\mathbf{v}_*^+$ are not eliminated. Even in the case when some of the edge costs and node penalties are non-unique, the change in the objective value goes to zero asymptotically. Therefore,

$$\lim_{|\mathbf{v}|\to\infty} \frac{\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi^-)}{\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi^+)} \to 1, \quad \forall \gamma \leq \gamma_2(m), m < 1.$$

By sandwich theorem, we have

$$\lim_{|\mathbf{v}|\to\infty} \frac{\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi)}{\text{opt\_tradeoff}(\mathbf{v}, \mathcal{E}, \pi^{iid})} \to 1, \quad \forall \gamma \leq \gamma_2(m), m < 1.$$

Note that when $m \to 1$, $\gamma_2(m) \to \gamma_1$, and hence, we can make the gap between γ_1 and $\gamma_2(m)$ arbitrarily small. □

Proof of Theorem 6

When the sequence of normalized LLR converges in probability under null hypothesis[4] , the NP type-II error exponent under a fixed type-I error bound is [44, Theorem 1]

$$\mathcal{D} = \text{p} \lim_{|\mathbf{v}|\to\infty} \frac{1}{|\mathbf{v}|} L(\mathbf{Y}_V), \quad \mathbf{Y}_\mathbf{v} \sim \mathcal{H}_0, \tag{4.28}$$

$$= \text{p} \lim_{|\mathbf{v}|\to\infty} \frac{1}{|\mathbf{v}|} \mathbb{E}[L(\mathbf{Y}_\mathbf{v}); \mathcal{H}_0], \tag{4.29}$$

where p lim denotes convergence in probability. The reduction from (4.28) to (4.29) holds when the sequence of the normalized LLR variables is uniformly integrable [112, (16.21)]. Using the form of LLR for a MRF in (17),

[4]Random variables X_n converge in probability to X, if $\lim_n \mathbb{P}[|X_n - X| \geq \epsilon] = 0$, for each positive ϵ. [112, p. 268].

$$\mathbb{E}[L(\mathbf{Y}_\mathbf{v}); \mathcal{H}_0] = \sum_{c \in C} \mathbb{E}[\phi_c(\mathbf{Y}_c); \mathcal{H}_0]. \tag{4.30}$$

□

Proof of Theorem 7

As in proof of Theorem 2, for a sequence of node sets $\mathbf{v}$ with clique collection C and another sequence of node subsets $\mathbf{v}_s \subsetneq \mathbf{v}$ with sub-collection $C_s \subsetneq C$, when $\limsup_{|\mathbf{v}|\to\infty} \frac{|C_s|}{|C|} = 1$, the result holds as in the i.i.d. case.

Assume that $\limsup_{|\mathbf{v}|\to\infty} \frac{|C_s|}{|C|} = m < 1$. From Theorem 3,

$$\sum_{c \in C \setminus C_s} D_c - \epsilon \leq \log \frac{P_M(\mathbf{v}_s)}{P_M(\mathbf{v})} \leq \sum_{c \in C \setminus C_s} D_c - \epsilon,$$

for some $\epsilon > 0$. Define new penalty functions

$$\pi^{\pm}(C \setminus C_s) := \sum_{c \in C \setminus C_s} D_c \pm \frac{\delta}{|C|},$$

where $\delta := \frac{\epsilon}{1-m}$ is finite since $m < 1$.

For the same edge costs, a uniformly smaller penalty function for each node subset results in a lower value of the optimal solution. Hence, we have

$$\begin{aligned} \text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \pi^-) &\leq \text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \pi'') \\ &\leq \text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \pi^+), \end{aligned}$$

for $\pi'' = \pi, \pi\_\text{cmp}$. Since the number of cliques grows as the number of nodes, $\frac{\delta}{|C|} \to 0$ as $|\mathbf{v}| \to \infty$ and π^- and π^+ can be made close to one another. On lines of the proof of Theorem 2, we can show that

$$\lim_{|\mathbf{v}|\to\infty} \frac{\text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \pi^-)}{\text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \pi^+)} \to 1.$$

By sandwich theorem, we have

$$\lim_{|\mathbf{v}|\to\infty} \frac{\text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \pi)}{\text{opt\_tradeoff\_clique}(\mathbf{v}, \mathcal{E}, \pi\_\text{cmp})} \to 1.$$

□

Proof of Theorem 8

We now write a 0-1 integer program whose optimal solution provides the optimal clique selection and fusion scheme in (11) for computing its marginal LLR and delivering it to the fusion center v_1.

As explained in Section 3.2.2, we can map any valid fusion digraph $\mathcal{F} = \{\text{FG}, \text{AG}\}$ and the processor assignment mapping Proc to variables $\mathbf{y}$ and $\mathbf{z}$, defined as

$$z(j, c) := I[\text{Proc}(c) == j], \quad y(i, j) := I[< i, j > \in \text{AG}],$$

where I is the indicator function and, the total routing costs of the fusion digraph in (19) can be expressed as,

$$\mathcal{E}(\mathcal{F}) = \frac{1}{2} \sum_{i,j\in\mathbf{v}} [I(\sum_{c:i\subset c} z(j, c) \geq 1) + y(i, j)]\mathcal{E}^{\text{SP}}(i, j).$$

We now need to incorporate the inference performance into the integer program. From (11), it is equivalent to imposing penalties for not selecting a set of cliques $X \subset C$ for processing and data fusion. This can happen in two ways, viz., the clique may not be assigned a processor or the computed clique potential may not be aggregated and delivered to the fusion center. Hence, (11) is equivalent to the following integer program:

$$\min_{\mathbf{y},\mathbf{z},\mathbf{u}} \quad \frac{1}{2}\sum_{i,j\in\mathbf{v}}[I(\sum_{c:i\subset c} z(j,c) \geq 1) + y(i,j)]\mathcal{E}^{\text{SP}}(i,j) + \sum_{X\subset C} u(X)\pi(X) \qquad \text{(IP-1)}, \tag{4.31}$$

$$s.t. \quad \text{let Proc:}=\{j : z(j,c) = 1, \text{ for } j \in \mathbf{v}, c \in C\}, \tag{4.32}$$

$$\sum_{c:c\in\mathbf{s}, j\in\mathbf{v}} z(j,c) + \sum_{X:X\supset\mathbf{s}} u(X) \geq 1, \forall \mathbf{s} \subset C, \tag{4.33}$$

$$\sum_{i\notin\mathbf{s}, j\in\mathbf{s}} y(i,j) + \sum_{\substack{X:X\supset A \\ A=\{c:\text{Proc}(c)\in\mathbf{s}\}}} u(X) \geq 1, \forall \mathbf{s} \subset \mathbf{v}, \mathbf{s} \cap \text{Proc} \neq \emptyset, \tag{4.34}$$

$$\mathbf{y}, \mathbf{z}, \mathbf{u} \in \{0,1\}, \tag{4.35}$$

where $\pi(X):=\gamma \sum_{c\in X} D_c$.

For the case of clique selection, we have

$$\begin{aligned}\sum_{i,j\in\mathbf{v}} I(\sum_{c:i\subset c} z^*(j,c) \geq 1)\mathcal{E}^{\text{SP}}(i,j) &= \sum_{i,j\in\mathbf{v}}\sum_{c:i\subset c} z^*(j,c)\mathcal{E}^{\text{SP}}(i,j), \\ &= \sum_{\substack{c\in C \\ |c|>1}}\sum_{i\subset c, j\in\mathbf{v}} z^*(j,c)\mathcal{E}^{\text{SP}}(i,j),\end{aligned}$$

where the two equalities hold since there is a unique clique c containing node i, since c is a clique. Adding the constraint that $|c| > 1$ does not affect the optimal solution. Hence, we have the equivalent IP,

$$\min_{\mathbf{y},\mathbf{z},\mathbf{u}} \quad \frac{1}{2}[\sum_{c\in C,|c|>1}\sum_{i\subset c, j\in\mathbf{v}} z(j,c)\mathcal{E}^{\text{SP}}(i,j) + \sum_{i,j\in\mathbf{v}} y(i,j)\mathcal{E}^{\text{SP}}(i,j)] + \sum_{X\subset C} u(X)\pi(X) \qquad \text{(IP-2)} \tag{4.36}$$

We can now add new nodes v_c and define new edge costs as

$$\mathcal{E}^{\text{SP}}(v_c, j) := \sum_{i \subset c} \mathcal{E}^{\text{SP}}(i, j), \quad \forall j \subset c,$$

and the new penalties are π'

$$\pi'(X) = \sum_{c: v_c \in X \text{or } |c|=1, i \in X, i \subset c} \gamma D_c, \quad \forall X \subset \mathbf{v} \cup \mathbf{v}'$$

Hence, we have

$$\begin{aligned}
\min_{\mathbf{y},\mathbf{z},\mathbf{u}} \quad & \frac{1}{2}\Big[\sum_{v_c \in \mathbf{v}', j \in \mathbf{v}} z(j,c)\mathcal{E}^{\text{SP}}(v_c, j) + \sum_{i,j \in \mathbf{v}} y(i,j)\mathcal{E}^{\text{SP}}(i,j)\Big] \\
& + \sum_{X \subset \mathbf{v} \cup \mathbf{v}'} \pi'(X)u(X) \qquad \text{(IP-3)}, \qquad (4.37) \\
s.t. \quad & \text{let Proc}:=\{j : z(j,c) = 1, \text{ for } j \in \mathbf{v}, v_c \in \mathbf{v}'\}, \\
& \sum_{c: v_c \in \mathbf{s}, j \in \mathbf{v}} z(j,c) + \sum_{X: X \supset \mathbf{s}} u(X) \geq 1, \forall \mathbf{s} \subset \mathbf{v} \cup \mathbf{v}', \\
& \sum_{i \notin \mathbf{s}, j \in \mathbf{s}} y(i,j) + \sum_{X: X \supset \mathbf{s}} u(X) \geq 1, \forall \mathbf{s} \subset \mathbf{v} \cup \mathbf{v}', \mathbf{s} \cap \text{Proc} \neq \emptyset, \\
& \mathbf{y}, \mathbf{z}, \mathbf{u} \in \{0,1\},
\end{aligned}$$

where the constraints are redefined since the penalty π' is defined over the entire set $\mathbf{v} \cup \mathbf{v}'$. In the final step, we $\mathbf{z}$ and $\mathbf{y}$ as variables $\mathbf{x}$ and this turns out to be the IP for the PCST.

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & \sum_{i,j \in \mathbf{v} \cup \mathbf{v}'} \frac{1}{2} x(i,j)\mathcal{E}^{\text{SP}}(i,j) + \sum_{X \subset \mathbf{v} \cup \mathbf{v}'} \pi'(X)u(X), \quad \text{(IP-4)} \\
s.t. \quad & \sum_{i \notin \mathbf{s}, j \in \mathbf{s}} x(i,j) + \sum_{X: X \supset \mathbf{s}} u(X) \geq 1, \forall \mathbf{s} \subset \mathbf{v} \cup \mathbf{v}', \\
& \mathbf{x}, \mathbf{u} \in \{0,1\}.
\end{aligned}$$

□

Proof of Theorem 9

We first show that the approximation factor of the GW-algorithm is only dependent on the number of vertices with strictly positive penalty.

Lemma 3 (Approx. Factor of GW-Algorithm) *Given node set* $\mathbf{v}$*, root* v_1 *and subset* $\mathbf{v}' \subset \mathbf{v}$ *with all nodes with non-zero penalty, the GW-algorithm for PCST in [107] has an approximation factor*

$$2 - \frac{1}{\max[|\mathbf{v}'| - I(v_1 \in \mathbf{v}'), 1]}, \tag{4.38}$$

where I *is the indicator function.*

Proof: The approximation factor is based on the upper bound on the number of active nodes in any iteration of the algorithm in [107, Thm. 4.1]. Since only nodes in $\mathbf{v}'$ have non-zero penalties, the number of active cliques is at most $|\mathbf{v}'|$ in any iteration. Moreover, the root v_1 is set inactive by the algorithm and if $v_1 \in \mathbf{v}'$, the number of active nodes is at most $|\mathbf{v}'| - 1$. □

Hence, for Tradeoff_Approx, only the nodes corresponding to the cliques have non-zero penalties. This implies that the approximation ratio is improved to

$$\rho(\text{Tradeoff\_Approx(Map-Sub)}) = 2 - (|C| - I(v_1 \in C))^{-1}, \tag{4.39}$$

where the indicator function is over the event that the fusion center is a 1-clique.

We can further improve the approximation ratio by modifying the function Map-Sub by using the result below about the optimal solution.

Lemma 4 (Profitable Components) *In the optimal solution opt_tradeoff_clique*$(\mathbf{v}, \mathcal{E}, \pi^{clq})$ *only the cliques in the sub-collection* $C_p \subset C$ *are potentially selected, with* C_p *defined as*

$$C_p := \{c : c \in C, |c| = 1 \text{ or } |c| > 1 \text{ and } \gamma D_c \geq \min_{i \in \mathbf{v}} \sum_{v_k \subset c_j, k \neq i} \mathcal{E}^{SP}(v_i, v_k)\}. \tag{4.40}$$

Proof: First note that all the selected clique representative nodes are leaves in the PCST. This is because if a zero-penalty node is a leaf in the PCST, then the cost is lowered by removing it. For a clique $c \notin C_p$, let vertex v_c be its representative in the augmented network graph Map-Sub($N_g(\mathbf{v})$) and say it is spanned in PCST and connected to some node i. By construction of Map-Sub($\bar{N}_g$), $i \subset c$. But the value of the objective function of the PCST can be lowered by removing the edge (v_c, i), since the penalty is less than any edge cost

$$\gamma D_c < \mathcal{E}^{\text{SP}}(v_c, i), \quad \forall\, i \in \mathbf{v}, c \notin C_p.$$

Hence, $v_c \notin$ PCST for $c \notin C_p$. □

The above lemma implies that only cliques generating a net "profit" after reducing their scaled KL-distance by the costs of raw-data routing to the processor are candidates for optimal selection. This implies that there is no need to add virtual nodes for non-profitable cliques in the augmented graph and hence, approximation factor on using Map-Sub' holds from Lemma 3 and 4. □

CHAPTER 5

ENERGY SCALING LAWS FOR OPTIMAL INFERENCE IN RANDOM NETWORKS

We consider the problem of distributed statistical inference in a network of randomly located sensors taking measurements and transporting the locally processed data to a designated fusion center. The fusion center then makes an inference about the underlying phenomenon based on the data collected from all the sensors.

For statistical inference using wireless sensor networks, energy consumption is an important design parameter. The transmission power required to reach a receiver distance d away with a certain signal-to-noise ratio (SNR) scales in the order of d^{ν}, where $2 \leq \nu \leq 6$ is the path loss [113]. Therefore, the cost of moving data from sensor locations to the fusion center, either through direct transmissions or through multihop forwarding, significantly affects the lifetime of the network.

5.1 Scalable data fusion

We investigate the cost of data fusion for inference, and its scaling behavior with the size of the network and the area of deployment. In particular, for a network of n random sensors located at points $\mathbf{V}_n = \{V_1, \cdots, V_n\}$ in $\mathbb{R}^2$, a *fusion policy* Υ_n maps $\mathbf{V}_n$ to a set of scheduled transmissions and computations. The average cost (e.g., energy) of a policy is given by

$$\bar{\mathcal{E}}(\Upsilon_n(\mathbf{V}_n)) := \frac{1}{n} \sum_{i \in \mathbf{V}_n} \mathcal{E}_i(\Upsilon_n(\mathbf{V}_n)), \tag{5.1}$$

where $\mathcal{E}_i(\Upsilon_n(\mathbf{V}_n))$ is the cost at node i under policy Υ_n. The above average cost is random, and we are interested in its scalability in random networks as $n \to \infty$.

Definition 8 (Scalable Policy) *A sequence of policies* $\Upsilon := (\Upsilon_n)_{n \geq 1}$ *is scalable on average if*

$$\lim_{n \to \infty} \mathbb{E}(\bar{\mathcal{E}}(\Upsilon_n(\mathbf{V}_n))) = \bar{\mathcal{E}}_\infty(\Upsilon) < \infty$$

where the expectation $\mathbb{E}$ *is with respect to the random locations* $\mathbf{V}_n$*, and* $\bar{\mathcal{E}}_\infty(\Upsilon)$ *is referred to as the* scaling constant*. A sequence of policies* Υ_n *is* weakly scalable *if*

$$p \lim_{n \to \infty} \bar{\mathcal{E}}(\Upsilon(\mathbf{V}_n))) = \bar{\mathcal{E}}_\infty(\Upsilon) < \infty,$$

where $p\lim$ *denotes convergence in probability. It is* strongly scalable *if the above average energy converges almost surely and is* L^2 (mean-squared) scalable *if the convergence is in mean square.*

Hence, a scalable fusion policy implies a finite average energy expenditure even as the network size increases. We focus mostly on the L^2 scalability of the fusion policies, which implies weak and average scalability [112]. Further, we are interested in *lossless* data-fusion policies which enable the fusion center to perform optimal statistical inference with the best inference accuracy *as if* all the raw sensor data were available.

To motivate this study, first consider two simple fusion policies: the direct transmission policy (DT) in which all sensors transmit directly to the fusion center (single hop), and the shortest-path (SP) policy, where each node forwards

its raw data to the fusion center using the shortest-path route without any data combination at the intermediate nodes.

We assume, for now, that n sensor nodes are uniformly distributed in a square of area n. It is perhaps not surprising that neither of the above two policies is scalable as $n \to \infty$. For the DT policy[1], intuitively, the average transmission range from the sensors to the fusion center scales as $\sqrt{n}$, thus $\bar{\mathcal{E}}(\mathrm{DT}(\mathbf{V}_n))$ scales as $n^{\frac{\gamma}{2}}$. On the other hand, we expect the SP policy to have better scaling since it chooses the best multi-hop path to forward data from each node to the fusion center. However, even in this case, there is no finite scaling. Here, the average number of hops in the shortest path from a node to the fusion center scales in the order of $\sqrt{n}$, and thus, $\bar{\mathcal{E}}(\mathrm{SP}(\mathbf{V}_n))$ scales in the order of $\sqrt{n}$. Rigorously establishing the scaling laws for these two non-scalable policies is not crucial at this point since the same scaling laws can be easily established for regular networks when sensor nodes are on two-dimensional lattice points. See [114].

Are there scalable policies for data fusion? Among all the fusion policies not performing data combination at the intermediate nodes, the shortest-path (SP) policy minimizes the total energy. Thus, no scalable policy exists unless nodes cooperatively combine their information, a process known as *data aggregation*. Data aggregation, however, must be considered in conjunction with the performance requirements of specific applications. In this chapter, we assume that optimal statistical inference is performed at the fusion center *as if* all the raw sensor data were available, and this places a constraint on data aggregation. For instance, it rules out sub-sampling of the sensor field, considered in [6].

[1]The direct transmission policy may not even be feasible, depending on the maximum transmission power constraints at the sensors.

5.2 Summary of Scaling Results

In this chapter, we investigate the energy scaling laws of lossless fusion policies which are allowed to perform data aggregation at the intermediate nodes, but ensure that the fusion center achieves the same inference accuracy *as if* all the raw observations were collected without any data combination. We assume that the underlying binary hypotheses for the sensor measurements can be modeled as Markov random fields (MRF).

For sensor locations $\mathbf{V}_n$ and possibly correlated sensor measurements, finding the minimum energy fusion policy under the constraint of optimal inference is given by

$$\mathcal{E}(\Upsilon^*(\mathbf{V}_n)) = \inf_{\Upsilon \in \mathfrak{A}} \sum_{i \in \mathbf{V}_n} \mathcal{E}_i(\Upsilon(\mathbf{V}_n)), \tag{5.2}$$

where $\mathfrak{A}$ is the set of valid lossless data-fusion policies

$$\mathfrak{A} := \{\Upsilon : \text{optimal inference is achieved at the fusion center}\}.$$

In general, the above optimization is NP-hard [2], and hence, studying its energy scaling behavior directly is intractable. We establish upper and lower bounds on the energy of this optimal policy Υ^* and analyze the scaling behavior of these bounds. The lower bound is obtained via a policy conducting fusion along the Euclidean minimum spanning tree (MST), which is shown to be optimal when the sensor measurements are statistically independent under both hypotheses. The upper bound on the optimal fusion policy is established through a specific suboptimal fusion policy, referred to as Data Fusion over Markov Random Fields (DFMRF). DFMRF becomes optimal when observations are independent under either hypothesis, where it reduces to fusion along the MST. For certain spatial dependencies among sensor measurements of practi-

cal significance, such as the Euclidean 1-nearest neighbor graph, DFMRF has an approximation ratio 2, i.e., its energy is no more than twice that of the optimal fusion policy, independent of the size and configuration of the network.

We then proceed to establish a number of asymptotic properties of the DFMRF policy in Section 5.4, including its energy scalability, its performance bounds, and the approximation ratio with respect to the optimal fusion policy when the sensor measurements have dependencies described by a k-nearest neighbor graph or a disc graph (continuum percolation). Applying techniques developed in [49–51, 115], we provide a precise characterization of the scaling bounds as a function of sensor density and sensor placement distribution. These asymptotic bounds for DFMRF, in turn, imply that the optimal fusion policy is also scalable. Hence, we use the DFMRF policy as a vehicle to establish scaling laws for optimal fusion. Additionally, we use the energy scaling constants to optimize the distribution of the sensor placements. For independent measurements conditioned on each hypothesis, we show that the uniform distribution of the sensor nodes minimizes the asymptotic average energy consumption over all i.i.d spatial placements when the path-loss exponent of transmission is greater than two ($\nu > 2$). For $\nu \in [0, 2)$, we show that the uniform distribution is, in fact, the most expensive[2] node configuration in terms of routing costs. We further show that the optimality of the uniform node distribution applies for both the lower and upper bounds on the average energy consumption of the optimal fusion policy under Markov random field measurements with k-nearest neighbor dependency graph or the disc dependency graph under certain conditions.

To the best of our knowledge, our results are the first to establish the energy scalability of data fusion for certain correlation structures of the sensor measure-

[2]The path-loss exponent for wireless transmissions satisfies $\nu > 2$.

ments. The use of energy scaling laws for the design of efficient sensor placement is new and has direct engineering implications. The fusion policy DFMRF first appeared in [116], and is made precise here with detailed asymptotic analysis using the weak law of large numbers (WLLN) for *stabilizing* Euclidean graph functionals. One should not expect that scalable data fusion is always possible, and at the end of Section 5.4, we discuss examples of correlation structures where scalable lossless data-fusion policy does not exist.

5.3 Random Fusion Network Model

5.3.1 Stochastic model of sensor locations

We assume that n sensor nodes (including the fusion center) are placed randomly with sensor i located at $V_i \in \mathbb{R}^2$. By convention, the fusion center is denoted by $i = 1$, and is located at $V_1 \in \mathbb{R}^2$. We denote the set of locations of the n sensors by $\mathbf{V}_n$:=$\{V_1, \ldots, V_n\}$. For our scaling law analysis, we consider a sequence of sensor populations placed in expanding square regions $Q_{\frac{n}{\lambda}}$ of area $\frac{n}{\lambda}$ and centered at the origin $\mathbf{0} \in \mathbb{R}^2$, where we fix λ as the overall sensor density and let the number of sensors $n \to \infty$.

To generate sensor locations V_i, first let $Q_1 := [-\frac{1}{2}, \frac{1}{2}]^2$ be the unit-area square[3], and $X_i \overset{i.i.d.}{\sim} \tau, 1 \leq i \leq n$, be a set of n independent and identically distributed (i.i.d.) random variables distributed on support Q_1 according to τ. Here, τ is a probability density function (pdf) on Q_1 which is bounded away from zero and infinity. We then generate V_i by scaling X_i accordingly: $V_i = \sqrt{\frac{n}{\lambda}} X_i \in Q_{\frac{n}{\lambda}}$. A useful

[3]The results in this chapter hold for τ defined on any convex unit area.

special case is the uniform distribution ($\tau \equiv 1$). Let $\mathcal{P}_a$ be the homogeneous Poisson distribution on $\mathbb{R}^2$ with intensity $a > 0$.

5.3.2 Random dependency graphs

We consider the statistical inference problem of simple binary hypothesis testing, $\mathcal{H}_0$ vs. $\mathcal{H}_1$, on a pair of Markov random fields. Under regularity conditions [117], a MRF is defined by its (undirected) dependency graph $\mathcal{G}$ and an associated pdf $f(\cdot \mid \mathcal{G})$.

Under hypothesis $\mathcal{H}_k$ and sensor location set $\mathbf{V}_n = \{V_1, \cdots, V_n\}$ generated according to the stochastic model in Section 5.3.1, we assume that the dependency graph $\mathcal{G}_k := (\mathbf{V}_n, E_k)$ models the correlation among the sensor observations. Note that the node location set $\mathbf{V}_n$ under the two hypotheses are identical. Set E_k is the set of edges of the dependency graph $\mathcal{G}_k$, and it defines the correlations of the sensor observations, as described in the next section.

We restrict our attention to *proximity*-based Euclidean dependency graphs. In particular, we consider two classes of dependency graphs[4]: the (undirected) k-nearest neighbor graph (k-NNG) and the disc graph, also known as the continuum percolation graph. We expect that our results extend to other locally-defined dependency structures such as the Delaunay, Voronoi, the minimum spanning tree, the sphere of influence and the Gabriel graphs. An important property of the aforementioned graphs is a certain *stabilization* property (discussed in Appendix 5.A.) facilitating asymptotic scaling analysis.

[4]The k-nearest neighbor graph (k-NNG) has edges (i, j) if i is one of the top k nearest neighbors of j or viceversa, and ties are arbitrarily broken. The disc graph has edges between any two points within a certain specified Euclidean distance (radius).

5.3.3 Cost Model

For node i to transmit data to node j which is distance $|i, j|$ away, we assume that node i spends energy[5] $\gamma|i, j|^\nu$. Without loss of generality, we assume $\gamma = 1$. Hence, given a fusion policy $\mathcal{F}_\Upsilon = (\mathbf{v}_n, \overrightarrow{E}_\Upsilon)$ of network size n, the average energy consumption is given by

$$\bar{\mathcal{E}}(\Upsilon(\mathbf{v}_n)) = \frac{1}{n}\mathcal{E}(\Upsilon(\mathbf{v}_n)) = \frac{1}{n}\sum_{\langle i,j\rangle\in\overrightarrow{E}_\Upsilon} |i, j|^\nu, \quad 2 \le \nu \le 6. \tag{5.3}$$

5.4 Energy Scaling Laws

We now establish the scaling laws for optimal and suboptimal fusion policies. From the expression of average energy cost in (5.3), we see that the scaling laws rely on the law of large numbers (LLN) for stabilizing graph functionals. An overview of the LLN is provided in Appendix 5.A.

We recall some notations and definitions used in this section. $X_i \overset{i.i.d.}{\sim} \tau$, where τ is supported on $\mathcal{B}_1$, the unit square centered at the origin $\mathbf{0}$. The node location-set is $\mathbf{V}_n := \sqrt{\frac{n}{\lambda}}(X_i)_{i=1}^n$ and the limit is obtained by letting $n \to \infty$ with fixed $\lambda > 0$.

5.4.1 Energy scaling for optimal fusion: independent case

We first provide the scaling result for the case when the measurements are independent under either hypothesis. From Theorem 1, the optimal fusion policy minimizing the total energy consumption in (3.2) is given by aggregation along

[5]Since nodes only communicate a finite number of bits, we use energy instead of power as the cost measure.

the directed minimum spanning tree. Hence, the energy scaling is obtained by the asymptotic analysis of the MST.

For the random node-location set $\mathbf{V}_n$, the average energy consumption of the optimal fusion policy for independent measurements is

$$\bar{\mathcal{E}}(\Upsilon^*(\mathbf{V}_n)) = \bar{\mathcal{E}}(\mathrm{MST}(\mathbf{V}_n)) = \frac{1}{n} \sum_{e \in \mathbb{E}_{\theta_i} N_{\mathrm{smst}}(\lambda)(\mathbf{V}_n)} |e|^\nu. \tag{5.4}$$

Let $\zeta(\nu; \mathrm{MST})$ be the constant arising in the asymptotic analysis of the MST edge lengths, given by

$$\zeta(\nu; \mathrm{MST}) := \mathbb{E}\Big[\sum_{e \in E(\mathbf{0}; \mathbb{E}_{\theta_i} N_{\mathrm{smst}}(\lambda)(\mathcal{P}_1 \cup \{\mathbf{0}\}))} \frac{1}{2} |e|^\nu \Big], \tag{5.5}$$

where $\mathcal{P}_a$ is the homogeneous Poisson process of intensity $a > 0$, and $E(\mathbf{0}; \mathrm{MST}(\mathcal{P}_1 \cup \{\mathbf{0}\}))$ denotes the set of edges incident to the origin in $\mathrm{MST}(\mathcal{P}_1 \cup \{\mathbf{0}\})$. Hence, the above constant is half the expectation of the power-weighted edges incident to the origin in the minimum spanning tree over a homogeneous unit intensity Poisson process, and is discussed in Appendix 5.A in (5.30). Although $\zeta(\nu; \mathrm{MST})$ is not available in closed form, we evaluate it through simulations in Section 5.5.

We now provide the scaling result for the optimal fusion policy when the measurements are independent based on the LLN for the MST obtained in [49, Thm 2.3(ii)].

Theorem 10 (Scaling for independent data [49]) *When the sensor measurements*

are independent under each hypothesis, the limit of the average energy consumption of the optimal fusion policy in (5.4) is given by

$$\lim_{n\to\infty} \bar{\mathcal{E}}(\Upsilon^*(\mathbf{V}_n)) \overset{L^2}{=} \lambda^{-\frac{\nu}{2}} \zeta(\nu; MST) \int_{\mathcal{B}_1} \tau(x)^{1-\frac{\nu}{2}} dx. \tag{5.6}$$

Hence, asymptotically the average energy consumption of optimal fusion is a constant (independent of n) in the mean-square sense for independent measurements. In contrast, forwarding all the raw data to the fusion center according to the shortest-path (SP) policy has an unbounded average energy growing in the order of $\sqrt{n}$. Hence, significant energy savings are achieved through data fusion.

The scaling constant for average energy in (5.6) brings out the influence of several factors on energy consumption. It is inversely proportional to the node density λ. This is intuitive since placing the nodes with a higher density (i.e., in a smaller area) decreases the average inter-node distances and hence, also the energy consumption.

The node-placement pdf τ influences the asymptotic energy consumption through the term

$$\int_{\mathcal{B}_1} \tau(x)^{1-\frac{\nu}{2}} dx.$$

When the placement is uniform ($\tau \equiv 1$), the above term evaluates to unity. Hence, the scaling constant in (5.6) for uniform placement simplifies to

$$\lambda^{-\frac{\nu}{2}} \zeta(\nu; \mathrm{MST}).$$

The next theorem shows that the energy under uniform node placement ($\tau \equiv$

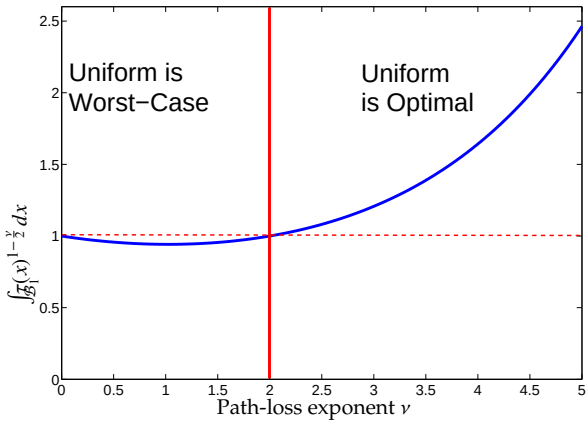


Figure 5.1: Ratio of energy under node pdf τ and uniform pdf.

1) optimizes the scaling limit in (5.6) when the path-loss exponent $\nu > 2$. Also, see Fig.5.1.

Theorem 11 (Minimum energy placement: independent case) *For any pdf τ supported on the unit square $\mathcal{B}_1$, we have*

$$\int_{\mathcal{B}_1}\tau(x)^{1-\frac{\nu}{2}}dx \geq 1, \quad \forall\, \nu > 2, \tag{5.7}$$

$$\int_{\mathcal{B}_1}\tau(x)^{1-\frac{\nu}{2}}dx \leq 1, \quad \forall\, \nu \in [0,2). \tag{5.8}$$

Proof: We have the Hölder inequality

$$\|f_1 f_2\|_1 \leq \|f_1\|_p \|f_2\|_q, \quad \forall p > 1, q = \frac{p}{p-1}, \tag{5.9}$$

where for any positive function f,

$$\|f\|_p := \Big(\int_{\mathcal{B}_1} f(x)^p dx\Big)^{\frac{1}{p}}.$$

When $\nu > 2$, in (5.9), substitute $f_1(x)$ with $\tau(x)^{\frac{1}{p}}$, $f_2(x)$ with $\tau(x)^{-\frac{1}{p}}$, and p with $\frac{\nu}{\nu-2} \geq 1$ which ensures that $p > 1$, to obtain (5.7).

For $\nu \in [0, 2)$, in (5.9), substitute $f_1(x)$ with $\tau(x)^{\frac{1}{p}}$, $f_2(x)$ with 1, $p = \frac{2}{2-\nu} > 1$ to obtain (5.8). □

The above result implies that, in the context of i.i.d. node placements, it is asymptotically energy-optimal to place the nodes uniformly when the path-loss exponent $\nu > 2$, which is the case for wireless transmissions. The intuitive reason is as follows: without loss of generality, consider a clustered distribution in the unit square, where nodes are more likely to be placed near the origin. The MST over such a point set has many short edges, but a few very long edges, since a few nodes are placed near the boundary with finite probability. On the other hand, for uniform point sets, the edges of the MST are more likely to be all of similar lengths. Since for energy consumption, we have power-weighted edge-lengths with path-loss exponent $\nu > 2$, long edges are penalized harshly, leading to higher energy consumption for clustered placement when compared with uniform node placement.

5.4.2 Energy scaling for optimal fusion: MRF case

We now evaluate the scaling laws for energy consumption of the DFMRF policy for a general Markov random field dependency among the sensor measurements. The DFMRF aggregation policy involves the cliques of the dependency graph which arise from correlation between the sensor measurements. The total energy consumption of DFMRF for random sensor locations $\mathbf{V}_n$ is given by

$$\begin{aligned}\mathcal{E}(\mathrm{DFMRF}(\mathbf{V}_n)) &= \sum_{c\in\mathcal{C}(\mathbf{V}_n)}\sum_{i\subset c}\mathcal{E}^{\mathrm{SP}}(i,\mathrm{Proc}(c);N_g)\\ &+ \mathcal{E}(\mathrm{MST}(\mathbf{V}_n)),\end{aligned} \tag{5.10}$$

where $\mathcal{E}^{\mathrm{SP}}(i,j;N_g)$ denotes the energy consumption for the shortest path between i and j using the links in the network graph $N_g(\mathbf{V}_n)$ (set of feasible links for direct transmission).

We now additionally assume that the network graph $N_g(\mathbf{V}_n)$ is a *local* u-energy spanner. In the literature [118], a graph $N_g(\mathbf{V}_n)$ is called a u-energy spanner, for some constant $u > 0$ called its *energy stretch factor*, when it satisfies

$$\max_{i,j\in\mathbf{V}_n}\frac{\mathcal{E}^{\mathrm{SP}}(i,j;N_g)}{\mathcal{E}^{\mathrm{SP}}(i,j;C_g)} \le u, \tag{5.11}$$

where $C_g(\mathbf{V}_n)$ denotes the complete graph on $\mathbf{V}_n$. In other words, the energy consumption between any two nodes is no worse than u-times the best possible value, i.e., over the shortest path using links in the complete graph. Intuitively, the u-spanning property ensures that the network graph possesses sufficient set of communication links to ensure that the energy consumed in the forwarding stage is bounded. Examples of energy u-spanners include the Gabriel graph[6] (with stretch factor $u = 1$ when the path-loss exponent $\nu \ge 2$), the Yao graph, and its variations [118]. In this chapter, we only require a weaker version of the above property that asymptotically there is at most u-energy stretch between the neighbors in the dependency graph

[6]The longest edge in Gabriel graph is $O(\sqrt{\log n})$, the same order as that of the MST [119]. Hence, the maximum power required at a node to ensure u-energy spanning property is of the same order as that needed for critical connectivity.

$$\limsup_{n\to\infty} \max_{(i,j)\in\mathcal{G}(\mathbf{V}_n)} \frac{\mathcal{E}^{\text{SP}}(i,\, j; N_g(\mathbf{V}_n))}{\mathcal{E}^{\text{SP}}(i,\, j; C_g(\mathbf{V}_n))} \leq u. \tag{5.12}$$

From (5.12), we have

$$\begin{aligned} \mathcal{E}(\text{FG}(\mathbf{V}_n)) &\leq u \sum_{c\in\mathcal{C}(\mathbf{V}_n)} \sum_{i\subset c} \mathcal{E}^{\text{SP}}(i, \text{Proc}(c); C_g), \\ &\leq u \sum_{c\in\mathcal{C}(\mathbf{V}_n)} \sum_{i\subset c} |i, \text{Proc}(c)|^{\nu}, \end{aligned} \tag{5.13}$$

where we use the property that the multihop shortest-path route from each node i to Proc(c) consumes no more energy than the direct one-hop transmission.

In the DFMRF policy, recall that the processors are members of the respective cliques, i.e., Proc(c) $\subset c$, for each clique c in the dependency graph. Hence, in (5.13), only the edges of the processors of all the cliques are included in the summation. This is upper bounded by the sum of all the power-weighted edges of the dependency graph $\mathcal{G}(\mathbf{V}_n)$. Hence, we have

$$\mathcal{E}(\text{FG}(\mathbf{V}_n)) \leq u \sum_{e\in\mathcal{G}(\mathbf{V}_n)} |e|^{\nu}. \tag{5.14}$$

From (5.10), for the total energy consumption of the DFMRF policy, we have the upper bound,

$$\mathcal{E}(\text{DFMRF}(\mathbf{V}_n)) \leq u \sum_{e\in\mathcal{G}(\mathbf{V}_n)} |e|^{\nu} + \mathcal{E}(\text{MST}(\mathbf{V}_n)). \tag{5.15}$$

The above bound allows us to draw upon the general methods of asymptotic analysis for graph functionals presented in [49, 120].

From (5.15), the DFMRF policy scales whenever the right-hand side of (5.14) scales. By Theorem 10, the energy consumption for aggregation along the MST scales. Hence, we only need to establish the scaling behavior of the first term in (5.14).

We now prove scaling laws governing the energy consumption of DFMRF and we also establish its asymptotic approximation ratio with respect to the optimal fusion policy. This in turn also establishes the scaling behavior of the optimal policy.

Theorem 12 (Scaling of DFMRF Policy) *When the dependency graph $\mathcal{G}$ of the sensor measurements is either the k-nearest neighbor or the disc graph, the average energy of DFMRF policy satisfies*

$$\begin{aligned}
&\limsup_{n\to\infty} \bar{\mathcal{E}}(DFMRF(\mathbf{V}_n)) \\
&\overset{a.s.}{\leq} \limsup_{n\to\infty} \Big(\frac{1}{n} \sum_{e\in\mathcal{G}(\mathbf{V}_n)} u\,|e|^{\nu} + \bar{\mathcal{E}}(MST(\mathbf{V}_n)) \Big) \\
&\overset{L^2}{=} \frac{u}{2} \int_{\mathcal{B}_1} \mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}(\mathcal{P}_{\lambda\tau(x)}\cup\{\mathbf{0}\})} |\mathbf{0}, j|^{\nu} \Big] \tau(x) dx \\
&\quad + \lambda^{-\frac{\nu}{2}} \zeta(\nu; MST) \int_{\mathcal{B}_1} \tau(x)^{1-\frac{\nu}{2}} dx. \qquad (5.16)
\end{aligned}$$

Proof: See Appendix 5.B. □

Hence, the above result establishes the scalability of the DFMRF policy. In the theorem below, we use this result to prove the scalability of the optimal

fusion policy and establish asymptotic upper and lower bounds on its average energy.

Theorem 13 (Scaling of Optimal Policy) *When the dependency graph $\mathcal{G}$ is either the k-nearest neighbor or the disc graph, the limit of the average energy consumption of the optimal policy Υ^* in* (3.2) *satisfies the upper bound*

$$\limsup_{n\to\infty} \bar{\mathcal{E}}(\Upsilon^*(\mathbf{V}_n)) \overset{a.s.}{\leq} \limsup_{n\to\infty} \bar{\mathcal{E}}(DFMRF(\mathbf{V}_n)), \tag{5.17}$$

where the right-hand side satisfies the upper bound in (5.16). Also, Υ^* satisfies the lower bound given by the MST

$$\begin{aligned} \liminf_{n\to\infty} \bar{\mathcal{E}}(DFMRF(\mathbf{V}_n)) &\overset{a.s.}{\geq} \liminf_{n\to\infty} \bar{\mathcal{E}}(\Upsilon^*(\mathbf{V}_n)) \\ \overset{a.s.}{\geq} \lim_{n\to\infty} \bar{\mathcal{E}}(MST(\mathbf{V}_n)) &\overset{L^2}{=} \lambda^{-\frac{\nu}{2}} \zeta(\nu; MST) \int_{\mathcal{B}_1} \tau(x)^{1-\frac{\nu}{2}} dx. \end{aligned} \tag{5.18}$$

Proof: From (3.12), the DFMRF and the optimal policy satisfy the lower bound given by the MST. □

Hence, the limiting average energy consumption for both the DFMRF policy and the optimal policy is strictly finite, and is bounded by (5.16) and (5.18). These bounds also establish that the approximation ratio of the DFMRF policy is asymptotically bounded by a constant, as stated below. Define the constant $\rho := \rho(u, \lambda, \tau, \nu)$, given by

$$\rho := 1 + \frac{u \int_{\mathcal{B}_1} \frac{1}{2} \mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}(\mathcal{P}_{\lambda\tau(x)}\cup\{\mathbf{0}\})} |\mathbf{0}, j|^{\gamma} \Big] \tau(x) dx}{\lambda^{-\frac{\gamma}{2}} \zeta(\gamma; \mathrm{MST}) \int_{\mathcal{B}_1} \tau(x)^{1-\frac{\gamma}{2}} dx}. \tag{5.19}$$

Lemma 5 (Approximation Ratio for DFMRF) *The approximation ratio of DFMRF is given by*

$$\begin{aligned} &\limsup_{n\to\infty} \frac{\mathcal{E}(DFMRF(\mathbf{V}_n))}{\mathcal{E}(\Upsilon^*(\mathbf{V}_n))} \\ &\overset{a.s.}{\leq} \limsup_{n\to\infty} \frac{\mathcal{E}(DFMRF(\mathbf{V}_n))}{\mathcal{E}(MST(\mathbf{V}_n))} \overset{L^2}{=} \rho, \end{aligned} \tag{5.20}$$

where ρ is given by (5.19).

Proof: Combine Theorem 12 and Theorem 13. □

We further simplify the above results for the k-nearest neighbor dependency graph in the corollary below by exploiting its scale invariance. The results are expected to hold for other *scale-invariant* Euclidean stabilizing graphs as well. The edges of a scale-invariant graph are invariant under a change of scale, or put differently, $\mathcal{G}$ is scale invariant if scalar multiplication by any positive constant α from $\mathcal{G}(\mathbf{V}_n)$ to $\mathcal{G}(\alpha\mathbf{V}_n)$ induces a graph isomorphism for all node sets $\mathbf{V}_n$.

Along the lines of (5.5), let $\zeta(\gamma; k\text{-NNG})$ be the constant arising in the asymptotic analysis of the k-NNG edge lengths, that is

$$\zeta(\gamma; k\text{-NNG}) := \mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in k\text{-NNG}(\mathcal{P}_1\cup\{\mathbf{0}\})} \frac{1}{2} |\mathbf{0}, j|^{\gamma} \Big]. \tag{5.21}$$

Corollary 1 (*k*-NNG Dependency Graph) *We obtain a simplification of Theorem 12 and 13 for average energy consumption, namely*

$$\begin{aligned}
&\limsup_{n\to\infty} \bar{\mathcal{E}}(\Upsilon^*(\mathbf{V}_n)) \overset{a.s.}{\leq} \limsup_{n\to\infty} \bar{\mathcal{E}}(DFMRF(\mathbf{V}_n)) \\
&\overset{a.s.}{\leq} \limsup_{n\to\infty} \Big(\frac{1}{n} \sum_{e\in\mathcal{G}(\mathbf{V}_n)} u\,|e|^{\nu} + \bar{\mathcal{E}}(MST(\mathbf{V}_n)) \Big) \\
&\overset{L^2}{=} \lambda^{-\frac{\nu}{2}} [u\,\zeta(\nu; k\text{-}NNG) + \zeta(\nu; MST)] \int_{\mathcal{B}_1} \tau(x)^{1-\frac{\nu}{2}} dx. \qquad (5.22)
\end{aligned}$$

The approximation ratio of DFMRF satisfies

$$\begin{aligned}
\limsup_{n\to\infty} \frac{\mathcal{E}(DFMRF(\mathbf{V}_n))}{\mathcal{E}(\Upsilon^*(\mathbf{V}_n))} &\overset{a.s.}{\leq} \limsup_{n\to\infty} \frac{\mathcal{E}(DFMRF(\mathbf{V}_n))}{\mathcal{E}(MST(\mathbf{V}_n))} \\
&\overset{L^2}{=} \Big(1 + u\frac{\zeta(\nu; k\text{-}NNG)}{\zeta(\nu; MST)}\Big). \qquad (5.23)
\end{aligned}$$

Proof: This follows from [49, Thm 2.2]. □

Hence, the expressions for the energy scaling bounds and the approximation ratio are further simplified when the dependency graph is the k-nearest neighbor graph. A special case of this scaling result for the 1-nearest-neighbor dependency under uniform node placement was proven in [4, Thm 2]. The constants in (5.5) and (5.21) can be computed numerically.

It is interesting to note that the approximation factor for the k-NNG dependency graph in (5.23) is independent of the node placement pdf τ and node density λ. Hence, DFMRF has the same efficiency relative to the optimal policy under different node placements. The results of Theorem 11 on the optimality

of the uniform node placement are also applicable here, but for the lower and upper bounds on energy consumption. We formally state it below.

Theorem 14 (Minimum energy bounds for k-NNG) *Uniform node placement* ($\tau \equiv 1$) *minimizes the asymptotic lower and upper bounds on average energy consumption in (5.18) and (5.22) for the optimal policy under the k-NNG dependency graph over all i.i.d. node placement pdfs* τ.

Proof: From Theorem 11 and (5.22). □

We also prove the optimality of uniform node-placement distribution under the disc-dependency graph, but over a limited set of node placement pdfs τ.

Theorem 15 (Minimum energy bounds for disc graph) *Uniform node placement* ($\tau \equiv 1$) *minimizes the asymptotic lower and upper bounds on the average energy consumption in (5.18) and (5.22) for the optimal fusion policy under the disc dependency graph over all i.i.d. node-placement pdfs* τ *satisfying the lower bound*

$$\tau(x) > \frac{1}{\lambda}, \quad \forall x \in \mathcal{B}_1, \tag{5.24}$$

where $\lambda > 1$ *is the (fixed) node placement density.*

Proof: We use the fact that for the disc graph $\mathcal{G}$ with a fixed radius, more edges are added as we scale down the area. Hence, for Poisson processes with intensities $\lambda_1 > \lambda_2 > 0$,

$$\mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}(\mathcal{P}_{\lambda_1}\cup\{\mathbf{0}\})} |\mathbf{0},j|^\gamma\Big] \geq \mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}(\mathcal{P}_{\lambda_2}\cup\{\mathbf{0}\})} |\mathbf{0},j|^\gamma\Big]\left[\frac{\lambda_2}{\lambda_1}\right]^{\frac{\gamma}{2}},$$

where the right-hand side is obtained by merely rescaling the edges present under the Poisson process at intensity λ_2. Since, new edges are added under the Poisson process at λ_1, the above expression is an inequality, unlike the case of k-NNG where the edge set is invariant under scaling. Substituting λ_1 with $\lambda\tau(x)$, and λ_2 by 1 under the condition that $\lambda\tau(x) > 1$, $\forall x \in \mathcal{B}_1$, we have

$$\begin{aligned}
&\int_{\mathcal{B}_1} \mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}(\mathcal{P}_{\lambda\tau(x)}\cup\{\mathbf{0}\})} |\mathbf{0},j|^{\nu}\Big]\tau(x)dx \\
\geq\ & \lambda^{-\frac{\nu}{2}}\mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}(\mathcal{P}_1\cup\{\mathbf{0}\})} |\mathbf{0},j|^{\nu}\Big]\int_{\mathcal{B}_1}\tau(x)^{1-\frac{\nu}{2}}dx, \\
\geq\ & \lambda^{-\frac{\nu}{2}}\mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}(\mathcal{P}_1\cup\{\mathbf{0}\})} |\mathbf{0},j|^{\nu}\Big], \quad \nu > 2.
\end{aligned}$$

□

Hence, uniform node placement is optimal in terms of the energy scaling bounds under the disc dependency graph if we restrict to pdfs τ satisfying (5.24).

We have so far established the finite scaling of the average energy when the dependency graph describing the correlations among the sensor observations is either the k-NNG or the disc graph with finite radius. However, we cannot expect finite energy scaling under any general dependency graph. For instance, when the dependency graph is the complete graph, the log-likelihood ratio in (3.3) is a function of only one clique containing all the nodes. In this case, the optimal policy in (3.2) consists of a unique processor chosen optimally, to which all the other nodes forward their raw data along shortest paths, and the processor then forwards the value of the computed log-likelihood ratio to the fusion center. Hence, for the complete dependency graph, the optimal fusion policy

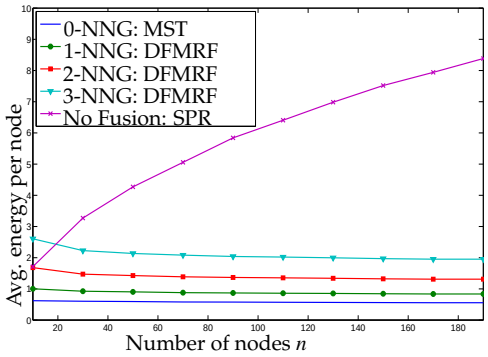


Figure 5.2: Avg. energy, k-NNG dependency, $\tau \equiv 1$, 500 runs, $\nu = 2$.

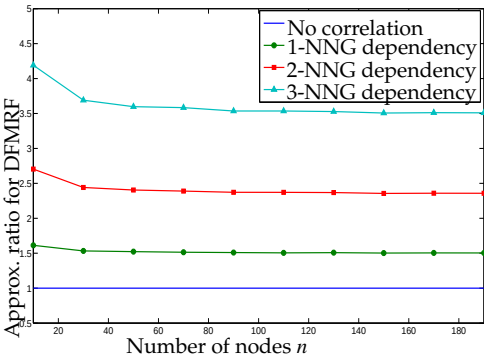


Figure 5.3: Approx. ratio, k-NNG dependency, $\tau \equiv 1$, 500 runs, $\nu = 2$.

reduces to a version of the shortest-path (SP) routing, where the average energy consumption grows as $\sqrt{n}$ and does not scale with n.

5.5 Numerical Illustrations

As described in Section 5.3.1, n nodes are placed in area $\frac{n}{\lambda}$ and one of them is randomly chosen as the fusion center. We conduct 500 independent simulation runs and average the results. We fix node density $\lambda = 1$. We plot results for two cases of dependency graph, viz., the k-nearest neighbor graph and the disc graph with a fixed radius δ.

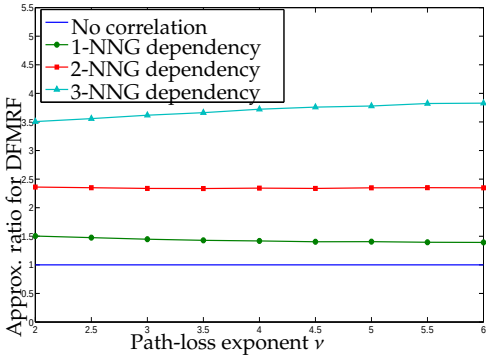


Figure 5.4: Approx. ratio, k-NNG dependency, $\tau \equiv 1$, 500 runs, $n = 190$.

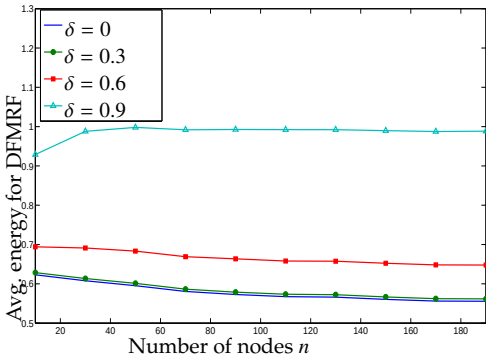


Figure 5.5: Disk Dependency graph, $\nu = 2$, uniform ($\tau \equiv 1$).

In Figs.5.2,5.3 and 5.4, we plot the simulation results for the k-nearest neighbor dependency graph and uniform node placement. Recall in Corollary 1, we established that the average energy consumption of the DFMRF policy in (5.22) is finite and bounded for asymptotic networks under k-NNG dependency. On the other hand, we predicted in Section 5.1 that the average energy under no aggregation (SP policy) increases without bound with the network size. The results in Fig.5.2 agree with our theory and we note that the convergence to asymptotic values is quick, and occurs in networks with as little as 30 nodes. We also see that the energy for DFMRF policy increases with the number of neighbors k in the dependency graph since the graph has more edges leading to computation of a more complex likelihood ratio by the DFMRF policy.

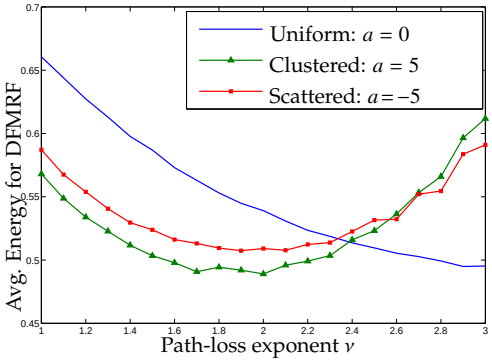


Figure 5.6: Disk Dependency graph, radius $\delta = 0$, $n = 190$.

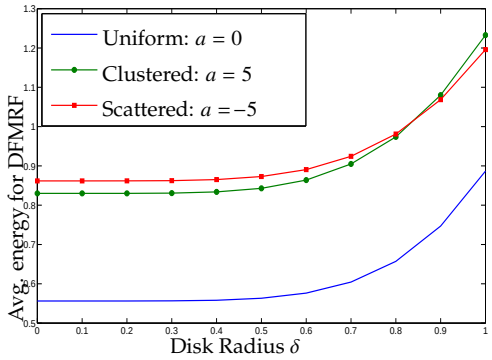
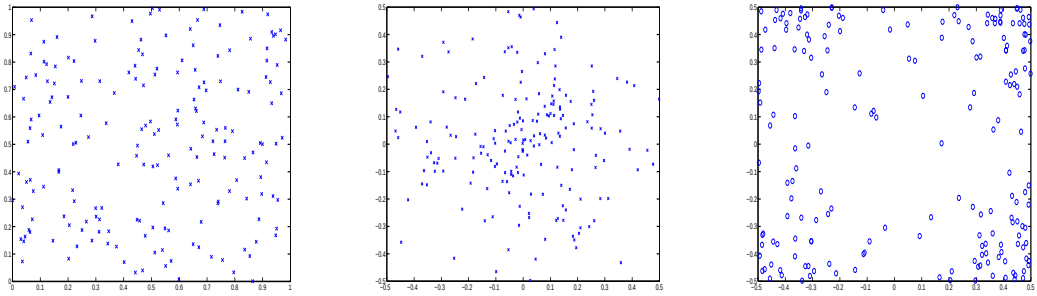


Figure 5.7: Disk Dependency graph, radius δ, $\nu = 4$.

We plot the approximation ratio of the DFMRF policy for k-NNG in (5.23) against the number of nodes in Fig.5.3 and against the path-loss exponent ν in Fig.5.4. As established by Corollary 1, the approximation ratio is a constant for large networks, and we find a quick convergence to this value in Fig.5.3 as we increase the network size. In Fig.5.4, we also find that the approximation ratio is fairly insensitive with respect to the path-loss exponent ν.

In Fig.5.5, we plot the average energy consumption of DFMRF in (5.16) under uniform node placement and the disc dependency graph with radius δ. The average energy is bounded, as established by Theorem 12. As in the k-NNG case, on increasing the network size, there is a quick convergence to the asymp-

totic values. Moreover, as expected, energy consumption increases with the radius δ of the disc graph since there are more edges. Note that the energy consumption at $\delta = 0$ and $\delta = 0.3$ are nearly the same, since at $\delta = 0.3$, the disc graph is still very sparse, and hence, the energy consumed in the forwarding stage of the likelihood-ratio computation is small.



(a) Uniform $a \to 0$. (b) Clustered $a = 5$. (c) Spread-out $a = -5$.

Figure 5.8: Sample realization of $n = 190$ points on unit square under pdf τ.

We now study the effect of i.i.d. node-placement pdf τ on the energy consumption of both DFMRF policy and shortest-path policy with no data aggregation. In Fig.5.6, Fig.5.7 and Fig.5.9, we consider a family of truncated-exponential pdfs τ_a given by

$$\tau_a(x) = \xi_a(x(1))\xi_a(x(2)), \quad x \in \mathbb{R}^2, \tag{5.25}$$

where, for some $a \neq 0$, ξ_a is given by the truncated exponential

$$\xi_a(z) := \begin{cases} \dfrac{ae^{-a|z|}}{2(1-e^{-\frac{a}{2}})}, & \text{if } z \in [-\frac{1}{2}, \frac{1}{2}], \\ 0, & \text{o.w.} \end{cases} \tag{5.26}$$

Note that as $a \to 0$, we obtain the uniform distribution in the limit ($\tau_0 \equiv 1$). A positive a corresponds to clustering of the points with respect to the origin and

viceversa. In Fig.5.8, a sample realization is shown for the cases $a = \pm 5$ and $a \to 0$.

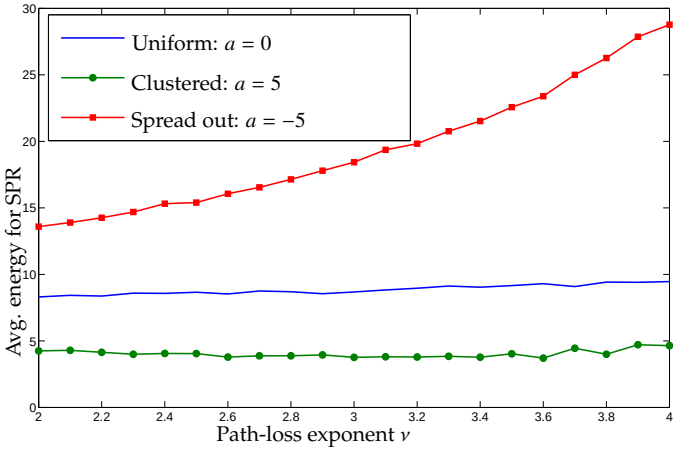


Figure 5.9: Avg. energy for shortest-path routing. 500 runs and $n = 190$.

Intuitively, for shortest-path (SP) policy where there is no data aggregation, the influence of node placement on the energy consumption is fairly straightforward. If we cluster the nodes close to one another, the average energy consumption decreases. On the other hand, spreading the nodes out towards the boundary increases the average energy. Indeed, we observe this behavior in Fig.5.9, for the placement pdf τ_a defined above in (5.25) and (5.26). However, as established in the previous sections, optimal node placement for the DFMRF policy does not follow this simple intuition.

In Theorem 11, we established that the uniform node placement ($\tau_0 \equiv 1$) minimizes the asymptotic average energy consumption of the optimal policy (which turns out to be the DFMRF policy), when the path-loss exponent $\nu \geq 2$. For $\nu \in [0, 2]$, the uniform distribution has the worst-case value. This is verified in Fig.5.6, where for $\nu \in [1, 3]$, the uniform distribution initially has high energy consumption but decreases as we increase the path-loss exponent ν. We see that at threshold of around $\nu = 2.4$, the uniform distribution starts

having lower energy than the non-uniform placements (clustered and spread-out), while according to Theorem 11, the threshold should be $\nu = 2$. Moreover, Theorem 11 also establishes that the clustered and spread-out distributions ($a \pm 5$) have the same energy consumption since the expressions $\int_{\mathcal{B}_1} \tau_a(x)^{1-\frac{\nu}{2}} dx$ for $a = 5$ and $a = -5$ are equal for τ_a given by (5.25) and (5.26), and this approximately holds in Fig.5.6.

We now study the energy consumption of the DFMRF policy in Fig.5.7 under the disc dependency graph and the node placements given in Fig.5.8. In Fig.5.7, for path-loss exponent $\nu = 4$, we find that the uniform node placement ($\tau_0 \equiv 1$) performs significantly better than the non-uniform placements for the entire range of the disc radius δ. Intuitively, this is because at large path-loss exponent ν, communication over long edges consumes a lot of energy and long edges occur with higher probability in non-uniform placements (both clustered and spread-out) compared to the uniform placement. Hence, uniform node placement is significantly energy-efficient under high path-loss exponent of communication.

5.A Functionals on random points sets

In [48, 49, 121], Penrose and Yukich introduce the concept of stabilizing functionals to establish weak laws of large numbers for functionals on graphs with random vertex sets. As in this thesis, the vertex sets may be marked (sensor measurements constituting one example of marks), but for simplicity of exposition we work with unmarked vertices. We briefly describe the general weak law of large numbers after introducing the necessary definitions.

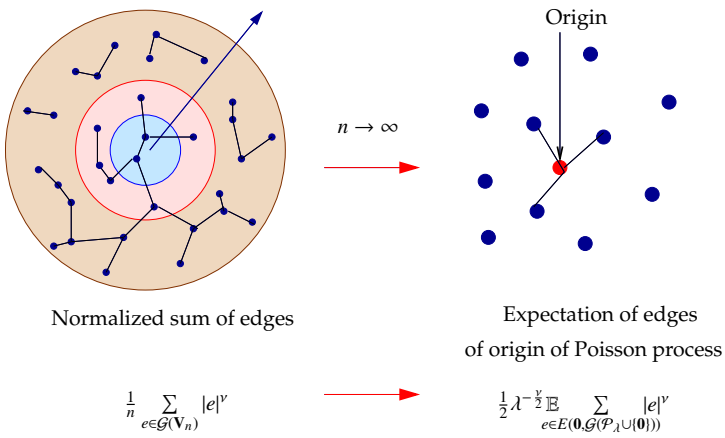


Figure 5.10: LLN for sum graph edges on uniform point sets ($\tau \equiv 1$).

Graph functionals on a vertex set $\mathbf{V}$ are often represented as sums of spatially dependent terms

$$\sum_{x \in \mathbf{V}} \xi(x, \mathbf{V}),$$

where $\mathbf{V} \subset \mathbb{R}^2$ is locally finite (contains only finitely many points in any bounded region), and the measurable function ξ, defined on all pairs $(x, \mathbf{V})$, with $x \in \mathbf{V}$, represents the interaction of x with other points in $\mathbf{V}$. We see that the functionals corresponding to energy consumption can be cast in this framework.

When $\mathbf{V}$ is random, the range of spatial dependence of ξ at node $x \in \mathbf{V}$ is random, and the purpose of *stabilization* is to quantify this range in a way useful for asymptotic analysis. There are several similar notions of stabilization, but the essence is captured by the notion of stabilization of ξ with respect to homogeneous Poisson points on $\mathbb{R}^2$, defined as follows. Recall that $\mathcal{P}_a$ is a homogeneous Poisson point process with intensity $a > 0$.

We say that ξ is translation invariant if $\xi(x, \mathbf{V}) = \xi(x + z, \mathbf{V} + z)$ for all $z \in \mathbb{R}^2$. Let $\mathbf{0}$ denote the origin of $\mathbb{R}^2$ and let $B_r(x)$ denote the Euclidean ball centered at x with radius r. A translation-invariant ξ is *homogeneously stabilizing* if for all

intensities $a > 0$ there exists almost surely a finite random variable $R := R(a)$ such that

$$\xi(\mathbf{0}, (\mathcal{P}_a \cap B_R(\mathbf{0})) \cup \mathcal{A}) = \xi(\mathbf{0}, \mathcal{P}_a \cap B_R(\mathbf{0}))$$

for all locally finite $\mathcal{A} \subset \mathbb{R}^2 \setminus B_R(\mathbf{0})$. Thus ξ stabilizes if the value of ξ at $\mathbf{0}$ is unaffected by changes in point configurations outside $B_R(\mathbf{0})$.

ξ satisfies the moment condition of order $p > 0$ if

$$\sup_{n \in \mathbb{N}} E[\xi(n^{\frac{1}{2}} X_1, n^{\frac{1}{2}} \{X_i\}_{i=1}^n)^p] < \infty. \tag{5.27}$$

We use the following weak laws of large numbers throughout. Recall that X_i are i.i.d. with density τ.

Theorem 16 (WLLN [49, 120]) *Put $q = 1$ or $q = 2$. Let ξ be a homogeneously stabilizing translation-invariant functional satisfying the moment condition (5.27) for some $p > q$. Then*

$$\begin{aligned} &\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \xi\Big(\sqrt{\frac{n}{\lambda}} X_i,\ \sqrt{\frac{n}{\lambda}} \{X_j\}_{j=1}^n \Big) \\ = &\int_{\mathcal{B}_1} E[\xi(\mathbf{0}, \mathcal{P}_{\lambda\tau(x)})] \tau(x) dx \ \ in \ \ L^q. \end{aligned} \tag{5.28}$$

We interpret the right-hand side of the above equation as a weighted average of the values of ξ on homogeneous Poisson point processes $\mathcal{P}_{\lambda\tau(x)}$. When ξ satisfies scaling such as $E[\xi(\mathbf{0}, \mathcal{P}_a)] = a^{-\alpha} E[\xi(\mathbf{0}, \mathcal{P}_1)]$, then the limit on the right-hand side of (5.28) simplifies to

$$\lambda^{-\alpha} E[\xi(\mathbf{0}, \mathcal{P}_1)] \int_{\mathcal{B}_1} (\tau(x))^{1-\alpha} dx \text{ in } L^q, \tag{5.29}$$

a limit appearing regularly in problems in Euclidean combinatorial optimization. For uniform node placement ($\tau(x) \equiv 1$), the expression in (5.28) reduces to $E[\xi(\mathbf{0}, \mathcal{P}_\lambda)]$, and the LLN result for this instance is pictorially depicted in Fig.5.10.

For example, if $\xi(x, \mathbf{V})$ is one half the sum of the ν-power weighted edges incident to x in the MST (or any scale-invariant stabilizing graph) on $\mathbf{V}$, i.e.,

$$\xi(x, \mathbf{V}) := \frac{1}{2} \sum_{e \in E(x, \mathbb{B}_{\theta_i} N_{\text{smst}}(\lambda)(\mathbf{V}))} |e|^\nu,$$

then substituting α with $\frac{\nu}{2}$ in (5.29),

$$\begin{aligned} & \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \xi\Big(\sqrt{\frac{n}{\lambda}} X_i, \sqrt{\frac{n}{\lambda}} \{X_i\}_{i=1}^n \Big) \\ = \; & \lambda^{-\frac{\nu}{2}} E[\xi(\mathbf{0}, \P_1)] \int_{Q_1} (\tau(x))^{1-\frac{\nu}{2}} dx \\ = \; & \lambda^{-\frac{\nu}{2}} \zeta(\nu; \text{MST}) \int_{Q_1} (\tau(x))^{1-\frac{\nu}{2}} dx, \end{aligned} \tag{5.30}$$

where $\zeta(\nu; \text{MST})$ is defined in (5.5).

5.B Proofs

Proof of Theorem 12

The energy consumption of DFMRF satisfies the inequality in (5.16). For the MST we have the result in Theorem 10. We now use stabilizing functionals to

show that

$$\frac{1}{n} \sum_{e \in \mathcal{G}(\mathbf{V}_n)} |e|^\nu$$

converges in L^2 to a constant. For all locally finite vertex sets $\mathcal{X} \subset \mathbb{R}^2$ supporting some dependency graph $\mathcal{G}(\mathcal{X})$ and for all $x \in \mathcal{X}$, define the functional $\eta(x, \mathcal{X})$ by

$$\eta(x, \mathcal{X}) := \sum_{y:(x,y) \in \mathcal{G}(\mathcal{X})} |x, y|^\nu. \tag{5.31}$$

Notice that $\sum_{x \in \mathcal{X}} \eta(x, \mathcal{X}) = 2 \sum_{e \in \mathcal{G}(\mathcal{X})} |e|^\nu$.

From [49, Thm 2.4], the sum of power-weighted edges of the k-nearest neighbors graph is a stabilizing functional and satisfies the bounded-moments condition (5.27). Hence, the limit in (5.28) holds when the dependency graph is the k-NNG.

Finally, the sum of power-weighted edges of the continuum percolation graph is a stabilizing functional which satisfies the bounded-moments condition (5.27), thus implying that the limit in (5.28) holds.

Indeed, η stabilizes with respect to $¶_a$, $a \in (0, \infty)$, since points distant from x by more than the deterministic disc radius do not modify the value of $\eta(x, ¶_a)$. Moreover, η satisfies the bounded moments condition (5.27) since each $|x, y|$ is bounded by the deterministic disc radius and the number of nodes in $n^{\frac{1}{2}}\{X_i\}_{i=1}^n$ which are joined to $n^{\frac{1}{2}} X_1$ is a random variable with moments of all orders. □

CHAPTER 6

INFERENCE ACCURACY SCALING LAWS FOR RANDOM NETWORKS

We have so far analyzed energy consumption for achieving optimal inference at the fusion center. In this chapter, we derive laws for scaling of inference accuracy with the network size under random node placement. We consider the Neyman-Pearson (NP) formulation, where the detector is optimal at a fixed false-alarm probability. We focus on the large-network scenario, where the number of observations goes to infinity. Under Neyman-Pearson formulation, for any positive level of the false alarm or the type-I error probability, when the mis-detection or the type-II error probability $P_M(n)$ of the NP detector decays exponentially with the sample size n, we have the error exponent

$$\mathcal{D} := -\lim_{n\to\infty} \frac{1}{n} \log P_M(n). \tag{6.1}$$

In this chapter, we are interested in evaluating the error exponent in (6.1) for random networks under MRF hypotheses.

Given the node locations $\mathbf{V}_n = \mathbf{v}_n$, let $D_{\mathbf{v}_n}$ denote the Kullback-Leibler divergence between the conditional pdfs $f(\mathbf{y}\mathbf{v}_n \mid \mathcal{G}_0(\mathbf{v}_n), \mathcal{H}_0)$ and $f(\mathbf{y}_{\mathbf{v}_n} \mid \mathcal{G}_1(\mathbf{v}_n), \mathcal{H}_1)$,

$$D_{\mathbf{v}_n} := \int_{\mathbf{y}_{\mathbf{v}_n}} \log \frac{f(\mathbf{y}_{\mathbf{v}_n} | \mathcal{G}_0(\mathbf{v}_n), \mathcal{H}_0)}{f(\mathbf{y}_{\mathbf{v}_n} | \mathcal{G}_1(\mathbf{v}_n), \mathcal{H}_1)} f(\mathbf{y}_{\mathbf{v}_n} | \mathcal{G}_0(\mathbf{v}_n), \mathcal{H}_0) \mathbf{d}\mathbf{y}_{\mathbf{v}_n}. \tag{6.2}$$

In Section 6.2, we relate the error exponent $\mathcal{D}$ in (6.1) to the KL-divergence in

(6.2).

6.0.1 Additional Assumptions on the Inference Model

A key modeling feature in this chapter is to incorporate the spatial dependence of sensor measurements. This is achieved by explicitly specifying the influence of (random) node locations on the MRF dependency graph and the conditional distributions of the measurements given the node locations.

We restrict our attention to proximity-based local dependency graphs such as the (undirected) (k-NNG) or the disk graph (also known as continuum percolation). An important localization property of these graphs is *stabilization* facilitating asymptotic scaling analysis.

We assume that a set of clique potentials $\psi_{m,c} > 0$ under either hypothesis can be parameterized locally by the sensor locations of the clique members and their l-hop neighbors, for some finite l, in a translation-invariant manner, i.e.,

$$\psi_{m,c}(\mathbf{y}_c; \mathbf{v}_n) = \psi_{m,c}(\mathbf{y}_c; \mathbf{v}_n + v), \quad \forall c \in C_m, v \in \mathbb{R}, \tag{6.3}$$

$$\psi_{m,c}(\mathbf{y}_c; \mathbf{v}_n) = \psi_{m,c}(\mathbf{y}_c; \{v_i : \mathrm{nbd}^l(i) \in c\}), \ \forall c \in C_m, \tag{6.4}$$

where nbd^l is the set of all 0 to l-hop neighbors. Further conditions are imposed for acyclic graphs in Section 6.1.1.

6.1 Error Exponent as a Graph Functional

The *spectrum* of the log-likelihood ratio [44, 45] is defined as the distribution of the normalized LLR evaluated under the null hypothesis

$$\frac{L(\mathbf{Y}_{\mathbf{V}_n})}{n}, \quad [\mathbf{Y}_{\mathbf{V}_n}, \mathbf{V}_n] \text{ under } \mathcal{H}_0,$$

where $L(\mathbf{Y}_{\mathbf{V}_n})$ is given by (3.3). In [44, 45] it is proven[1] that for Neyman-Pearson detection under a fixed type-I error bound, the LLR spectrum can fully characterize the type-II error exponent of the hypothesis-testing system, and is independent of the type-I bound.

When LLR spectrum converges in probability to a constant $\mathcal{D}$, the error exponent $\mathcal{D}$ of NP detection in (6.1) is [44]

$$\mathcal{D} = \text{p} \lim_{n\to\infty} \frac{1}{n} L(\mathbf{Y}_{\mathbf{V}_n}), \quad [\mathbf{Y}_{\mathbf{V}_n}, \mathbf{V}_n] \text{ under } \mathcal{H}_0, \tag{6.5}$$

where p lim denotes the limit in probability, assuming it exists.

When $\mathbf{Y}_{\mathbf{V}_n}$ are i.i.d. conditioned under both $\mathcal{H}_0$ and $\mathcal{H}_1$, the result in (6.5) reduces to Stein's lemma [108, Theorem 12.8.1] and the limit in (6.5) is the node Kullback-Leibler (KL) divergence, i.e., when $Y_{V_i} \overset{i.i.d.}{\sim} g_k$ under $\mathcal{H}_k$,

$$\mathcal{D} = D_{V_1} := \int_y \log \frac{g_0(y)}{g_1(y)} g_0(y) dy. \tag{6.6}$$

In Section 6.2, we evaluate the error exponent for MRF hypotheses through

[1] The generalization to an exponential type-I error bound [44, 45] is not tractable since a closed-form cumulative distribution of the LLR is needed.

the limit in (6.5). Due to random node placement and spatial dependence of the MRF hypotheses, the error exponent in (6.5) is the limit of a random-graph functional, and we can appeal to the LLN for graph functionals [49].

6.1.1 Acyclic Dependency Graphs

We consider the case when the dependency graphs under either MRF hypothesis $\mathcal{G}_0$ and $\mathcal{G}_1$ are *acyclic* and also stabilizing, such as the Euclidean 1-nearest neighbor graph.

Given a fixed set of points $\mathbf{v}_n$, the joint pdf of MRF for an acyclic dependency graph $\mathcal{G}(\mathbf{v}_n)$ admits a factorization [117]

$$f(\mathbf{y}_{\mathbf{v}_n}) = \prod_{i \in \mathbf{v}_n} f_i(y_i) \prod_{\substack{(i,j) \in \mathcal{G}(\mathbf{v}_n) \\ i<j}} \frac{f_{i,j}(y_i, y_j)}{f_i(y_i) f_i(y_j)}, \tag{6.7}$$

where f_i are the node marginal pdfs and $f_{i,j}$ are the pairwise pdfs on the edges. Recall that instead of fixed node locations, we have random locations $\mathbf{V}_n$ here, and hence, we consider the conditional pdf $f(\mathbf{y}_{\mathbf{V}_n} | \mathcal{H}_m, \mathcal{G}_m(\mathbf{V}_n))$ under each hypothesis $\mathcal{H}_m$. From (6.7), for an acyclic dependency graph $\mathcal{G}_m(\mathbf{V}_n)$, we can specify the conditional pdf $f(\mathbf{y}_{\mathbf{V}_n} | \mathcal{H}_m, \mathcal{G}_m(\mathbf{V}_n))$ through the conditional node pdfs $f_i(y_i | \mathcal{H}_m, \mathcal{G}_m)$ and the conditional pairwise edge pdfs $f_{i,j}(y_i, y_j | \mathcal{H}_m, \mathcal{G}_m)$.

We consider here a special form of spatial dependence in (6.4) by having identical node marginal pdfs for all node locations and edge marginal pdfs which are dependent only on the respective edge lengths. Under hypothesis

$\mathcal{H}_m$, for $m = 0, 1$,

$$f_i(y_i \mid \mathcal{G}_m, \mathcal{H}_m) = g_m(y_i),\ i \in \mathbf{V}_n, \tag{6.8}$$

$$f_{i,j}(y_i, y_j \mid \mathcal{G}_m, \mathcal{H}_m) = h_m(y_i, y_j \mid R_{ij}),\ (i,j) \in \mathcal{G}_m, \tag{6.9}$$

where g_m is the node pdf and h_m is the pairwise pdf at the edges conditioned on R_{ij}, the Euclidean length of edge (i, j).

By using (6.7), (6.8) and (6.9), we simplify (6.5) as

$$\mathcal{D} = \operatorname*{p\,lim}_{n\to\infty} \frac{1}{n}\Big[\sum_{i\in\mathbf{V}_n} \log \frac{g_0(Y_i)}{g_1(Y_i)} + \sum_{\substack{(i,j)\in\mathcal{G}_0 \\ i<j}} \log \frac{h_0(Y_i, Y_j \mid R_{ij})}{g_0(Y_i)g_0(Y_j)} - \sum_{\substack{(i,j)\in\mathcal{G}_1 \\ i<j}} \log \frac{h_1(Y_i, Y_j \mid R_{ij})}{g_1(Y_i)g_1(Y_j)}\Big],\ [\mathbf{Y}_{\mathbf{v}_n}, \mathbf{V}_n] \text{ under } \mathcal{H}_0, \tag{6.10}$$

Note that the above expression is a graph functional, based on the edge lengths of random graphs $\mathcal{G}_0$ and $\mathcal{G}_1$ with additional randomness from the conditional distribution of the sensor measurements given the edge lengths.

6.2 Detection Error Exponent

In this section, we derive the error exponent for general MRF hypotheses.

6.2.1 Testing Against Independence

We first provide the closed-form error exponent for the special case when the null hypothesis has i.i.d. measurements with no spatial dependence, $f(\mathbf{y}_{\mathbf{v}_n} | \mathcal{G}_0(\mathbf{v}_n), \mathcal{H}_0) = \prod_{i\in\mathbf{v}_n} g_0(y_i)$. Here, the dependency graph is trivial, $\mathcal{G}_0 = \emptyset$,

and the error exponent in (6.10) simplifies as

$$\mathcal{D} = \mathrm{p}\lim_{n\to\infty} \frac{1}{n}\Big[-\sum_{\substack{(i,j)\in\mathcal{G}_1\\ i<j}} \log \frac{h_1(Y_i, Y_j \mid R_{ij})}{g_1(Y_i)g_1(Y_j)} + \sum_{i\in\mathbf{V}_n} \log \frac{g_0(Y_i)}{g_1(Y_i)}\Big], \quad Y_{V_i} \overset{i.i.d.}{\sim} g_0,\ \sqrt{\tfrac{\lambda}{n}}V_i \overset{i.i.d.}{\sim} \tau. \tag{6.11}$$

The above expression is a graph functional defined over a marked point process, where the marks are the sensor measurements Y_{V_i} drawn i.i.d from the pdf g_0.

We can now appeal directly to the LLN for marked point processes [49, Thm. 2.1] to simplify (6.11). Define a functional on the edge lengths

$$\begin{aligned}\xi_1(r_{ij}) :=& \mathbb{E}\Big[-\log \frac{h_1(Y_i, Y_j)}{g_1(Y_i)g_1(Y_j)}\Big| R_{ij} = r_{ij}, \mathcal{H}_0\Big], \\ =& -\int_{y_i}\int_{y_j} \log \frac{h_1(y_i, y_j \mid R_{ij} = r_{ij})}{g_1(y_i)g_1(y_j)} g_0(y_i)g_0(y_j)dy_j dy_i,\end{aligned} \tag{6.12}$$

where the expectation is over the measurements conditioned on the node locations.

Recall from the moments condition in (5.27), ξ_1 is said to satisfy moments condition of order $p > 0$ if

$$\sup_{n\in\mathbb{N}} E[\sum_{j\in \mathrm{nbd}(\mathbf{0}), j\in\mathbf{V}_n} \xi_1(R_{\mathbf{0}j})^p] < \infty, \tag{6.13}$$

where nbd($\mathbf{0}$) denotes the neighbors of the origin in $\mathcal{G}_1$ and the expectation is over the node locations. We require that $p = 1$ or 2. In Section 6.3, we prove that ξ_1 satisfies the moment condition for the Gaussian distribution under some simple constraints on the covariance matrix.

Recall that $\mathcal{P}_\lambda$ is the homogeneous Poisson distribution on $\mathbb{R}^2$ with density λ. We now provide the result below.

Lemma 6 (Testing Acyclicity Against Independence) *When ξ_1 satisfies the moments condition in* (6.13)*, the error exponent for testing against independence has the form*

$$\mathcal{D} = D_{V_1} + \frac{1}{2}\int_{\mathcal{B}_1} \mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}_1(\mathcal{P}_{\lambda\tau(x)}\cup\{\mathbf{0}\})} \xi_1(R_{\mathbf{0}j})\Big]\tau(x)dx, \tag{6.14}$$

where D_{V_1} is the node KL-divergence given by (6.6)*.*

Proof: The first term follows from LLN for i.i.d variables. For the second term, ξ_1 is a stabilizing functional since it is a functional of edges of a stabilizing graph $\mathcal{G}_1$ and bounded-moments condition in (6.13) holds. Hence, the LLN in [49] guarantees L^2 convergence to the above constant, which in turn implies convergence in probability. □

Remark 1 *When the node locations are uniform ($\tau(x) \equiv 1$), the error exponent in (6.14) simplifies as*

$$\mathcal{D} = D_{V_1} + \frac{1}{2}\mathbb{E}\Big[\sum_{j:(\mathbf{0},j)\in\mathcal{G}_1(\mathcal{P}_\lambda\cup\{\mathbf{0}\})} \xi_1(R_{\mathbf{0}j})\Big]. \tag{6.15}$$

6.2.2 General Hypothesis Testing

In this section, we extend the results to any general distribution under the null hypothesis. For such cases, we cannot directly use the LLN for marked point process to evaluate (6.10), since the marks are required to be i.i.d. for the LLN to hold.

We now additionally assume *uniform integrability* [112, (16.21)] to convert the functional on a marked point process in (6.5) to a functional on an unmarked process. In Section 6.3, we show that the Gaussian distribution satisfies uniform integrability.

Proposition 1 (Uniform Integrability) *When the normalized spectrum, given by the sequence* $\{\frac{1}{n}L(\mathbf{Y}_{\mathbf{V}_n})\}_{n\geq 1}$ *is uniformly integrable and converges in probability under* $\mathcal{H}_0$*, the error exponent in (6.5) is the KL-divergence rate,*

$$\mathcal{D} = \lim_{n\to\infty} \frac{D_{\mathbf{V}_n}}{n}, \tag{6.16}$$

$$= p \lim_{n\to\infty} \frac{1}{n}\Big[\sum_{a\in C_1} \mathbb{E}(\psi_{1,a}(\mathbf{Y}_a) \mid \mathbf{V}_n, \mathcal{H}_0) - \sum_{b\in C_0} \mathbb{E}(\psi_{0,b}(\mathbf{Y}_b) \mid \mathbf{V}_n, \mathcal{H}_0) + \log \frac{Z_1(\mathbf{V}_n)}{Z_0(\mathbf{V}_n)}\Big], \tag{6.17}$$

where $D_{\mathbf{V}_n}$ *is the KL-divergence in* (6.2)*,* $\psi_{i,c}$ *is potential of clique* $c \in C_i$ *of the MRF under hypothesis* $\mathcal{H}_i$ *in (3.3).*

$$\textit{Proof: } \mathcal{D} = \lim_{n\to\infty} \frac{D_{\mathbf{V}_n}}{n} = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}[L(\mathbf{Y}_{\mathbf{V}_n}) \mid \mathcal{H}_0], \tag{6.18}$$

$$= \mathrm{p} \lim_{n\to\infty} \frac{1}{n}\mathbb{E}[L(\mathbf{Y}_{\mathbf{V}_n}) \mid \mathbf{V}_n, \mathcal{H}_0]. \tag{6.19}$$

Now evaluating the conditional expectation using the form of LLR for a MRF in (3.3), we have the result. □

Hence, we have (6.17), which is a functional on an unmarked process. Since the clique potential functions in (6.17) are parameterized by the node locations, (6.17) is a functional over a random graph. Note that we do not need the dependency graphs to be acyclic for the above result. We now specialize the above result for acyclic dependency graphs.

Lemma 7 (Acyclic Graphs) *For acyclic graphs $\mathcal{G}_0$ and $\mathcal{G}_1$,*

$$\mathcal{D} = D_{V_1} + p \lim_{n\to\infty} \frac{1}{n}\Big[\sum_{\substack{(i,j)\in\mathcal{G}_1\setminus\mathcal{G}_0 \\ i<j}} \xi_1(R_{ij}) + \sum_{\substack{(i,j)\in\mathcal{G}_0\cap\mathcal{G}_1 \\ i<j}} \xi_2(R_{ij}) \sum_{\substack{(i,j)\in\mathcal{G}_0\setminus\mathcal{G}_1 \\ i<j}} \xi_3(R_{ij})\Big], \tag{6.20}$$

where D_{V_1} and ξ_1 are given by (6.6) and (6.12), and the edge functionals ξ_2 and ξ_3 are defined as

$$\xi_2(r_{ij}) := \mathbb{E}\Big[\log \frac{h_0(Y_i, Y_j | R_{ij}=r_{ij})}{h_1(Y_i, Y_j | R_{ij}=r_{ij})} \Big| R_{ij}=r_{ij}, \mathcal{H}_0\Big] - 2D_{V_1} \tag{6.21}$$

$$\xi_3(r_{ij}) := I(Y_i; Y_j \mid R_{ij}= r_{ij}, \mathcal{H}_0), \tag{6.22}$$

where $I(X;Y)$ is mutual information between X and Y and $I(X;Y \mid Z = z)$ is mutual information conditioned on $Z = z$.

Proof: From (6.10) and Proposition 1. □

We now provide the error exponent for MRF hypotheses.

Theorem 17 (Exponent For Stabilizing Acyclic Graphs) *When ξ_i for $i = 1, 2, 3$ satisfy the bounded-moments condition in* (6.13)*, the error exponent for stabilizing acyclic dependency graphs is given by*

$$\mathcal{D} = D_{V_1} + \frac{1}{2} \sum_{i=1}^{3} \int_{\mathcal{B}_1} \mathbb{E}\left[\sum_{j:(\mathbf{0},j)\in E_{i,\tau(x)}} \xi_i(R_{\mathbf{0}j}) \right] \tau(x) dx, \tag{6.23}$$

where $E_{1,\tau(x)} := \mathcal{G}_1 \setminus \mathcal{G}_0(\mathcal{P}_{\lambda\tau(x)} \cup \{\mathbf{0}\})$, $E_{2,\tau(x)} := \mathcal{G}_0 \cap \mathcal{G}_1(\mathcal{P}_{\lambda\tau(x)} \cup \{\mathbf{0}\})$, and $E_{3,\tau(x)} := \mathcal{G}_0 \setminus \mathcal{G}_1(\mathcal{P}_{\lambda\tau(x)} \cup \{\mathbf{0}\})$.

Proof: Since $\mathcal{G}_0$ and $\mathcal{G}_1$ are stabilizing, its subgraphs with edges $E_{i,\tau(x)}$ for $i = 1, 2, 3$ can be shown to be stabilizing. The moments condition in (6.13) holds. Hence, the LLN follows. □

Remark 2 *When the node locations are uniform* ($\tau(x) \equiv 1$)*, the error exponent in (6.23) simplifies as*

$$\mathcal{D} = D_{V_1} + \frac{1}{2} \sum_{i=1}^{3} \mathbb{E} \sum_{j:(\mathbf{0},j)\in E_{i,1}} \xi_i(R_{\mathbf{0}j}). \tag{6.24}$$

6.3 Gaussian Distribution on Acyclic Graphs

In this section, we simplify the results of the previous section on acyclic graphs when the distribution under each hypothesis $\mathcal{H}_m$ is Gaussian $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_{m,\mathbf{v}_n})$, given the node locations $\mathbf{V}_n = \mathbf{v}_n$. In this case, the MRF factorization in (2.3) leads to a special relationship between the coefficients of the covariance matrix and its inverse, called the *potential matrix* . Specifically, there is a one-to-one correspondence between the non-zero elements of the potential matrix $\boldsymbol{\Sigma}^{-1}_{m,\mathbf{v}_n}$ and the dependency graph edges $\mathcal{G}_m(\mathbf{v}_n)$. Moreover, for acyclic graphs $\mathcal{G}_m(\mathbf{v}_n)$, further simplifications are possible [5, Thm. 1].

The additional constraints of spatial dependence for acyclic graphs in (6.8) and (6.9) imply that under each hypothesis, the mean and the variance at all the nodes are equal and that the correlation coefficient between any two neighboring nodes is only dependent on the inter-node distance, i.e., under hypothesis $\mathcal{H}_m$, for $m = 0, 1$, we have $\boldsymbol{\mu}_m = \mu_m \mathbf{I}$, $\Sigma_{m,\mathbf{v}_n}(i, i,) = \sigma^2_m$, and for $(i, j) \in \mathcal{G}_m(\mathbf{v}_n)$, we

have $\Sigma_{m,\mathbf{v}_n}(i,j) = \rho_m(R_{ij})\sigma_m^2$. Here, the correlation function $\rho_m(\cdot) < 1$ is positive and monotonically decreasing in the edge length, for each $m = 0, 1$.

Moreover, $\rho_m(0) < 1$, or the so-called nugget effect, according to geo-statistics literature [122, 123]. It has been observed in mining applications, where the micro-scale variation is assumed to be caused by the existence of small nuggets of the enriched ore. Many other ecological phenomena such as soil bacteria population [124], aquatic population [125] etc. also exhibit this behavior. Note that the presence of nugget effect has the same effect on correlation as imposing an exclusion region on how near two nodes can be placed. However, for such an exclusion constraint to hold, we need more complicated node placement distributions than the uniform or Poisson assumption. Although such distributions can be handled in principle, they are not analytically tractable. Some examples of the correlation function are

$$\rho_m(R) = \rho_m(0)\mathrm{e}^{-aR},\ \rho(R) = \frac{\rho_m(0)}{1+R^a},\ a \geq 0, 0 \leq \rho_m(0) < 1.$$

With the above assumptions, the covariance matrix under hypothesis $\mathcal{H}_m$ is given by

$$\Sigma_{m,\mathbf{v}_n}(i,j) = \begin{cases} \sigma_m^2 > 0, & i = j, \quad (6.25a) \\ \sigma_m^2 \prod_{(a,b)\in \mathrm{Path}(i,j;\mathcal{G}_m(\mathbf{v}_n))} \rho_m(R_{a,b}), & \text{o.w.} \quad (6.25b) \end{cases}$$

where $\mathrm{Path}(i, j; \mathcal{G}_m(\mathbf{v}_n))$ is the set of edges of the acyclic graph $\mathcal{G}_m(\mathbf{v}_n)$ belonging to the *unique* path[2] connecting the nodes i and j. It can be shown that $\Sigma_{m,\mathbf{v}_n}$ in (6.25) is positive definite for any node configuration $\mathbf{v}_n$ when $\rho_m(\cdot) < 1$.

[2] $\Sigma_{m,\mathbf{v}_n}(i,j) = 0$ if no path exists between i and j in $\mathcal{G}_m$.

Under the above assumptions, we now provide closed-form expression for the Gaussian error exponent. Recall that Path($\mathbf{0}, j; \mathcal{G}_0$) denotes the set of edges in $\mathcal{G}_0$ connecting the origin $\mathbf{0}$ with some node j. Let $\Delta\mu := \mu_1 - \mu_0, \quad K := \frac{\sigma_1^2}{\sigma_0^2}$.

Theorem 18 (Gaussian Error Exponent) *For Gaussian distribution under each hypothesis, the error exponent is given by* (6.23)*, with the terms simplifying as*

$$D_{V_1} = \frac{1}{2}\Big(\log(K) + \frac{1}{K} - 1 + \frac{\Delta\mu^2}{\sigma_1^2}\Big), \tag{6.26}$$

$$\xi_1(R_{\mathbf{0}j}) = \frac{\rho_1(R_{\mathbf{0}j})[\rho_1(R_{\mathbf{0}j}) - \prod_{(k,l)\in Path(\mathbf{0},j;\mathcal{G}_0(\mathcal{P}_{\lambda\tau(x)}\cup\mathbf{0}))} \rho_0(R_{kl})]}{[1-\rho_1^2(R_{\mathbf{0}j})]K} + \frac{\log[1-\rho_1^2(R_{\mathbf{0}j})]}{2}, \tag{6.27}$$

$$\xi_2(R_{\mathbf{0}j}) = \frac{\rho_1(R_{\mathbf{0}j})[\rho_1(R_{\mathbf{0}j}) - \rho_0(R_{\mathbf{0}j})]}{[1-\rho_1^2(R_{\mathbf{0}j})]K} + \frac{1}{2}\log\frac{1-\rho_1^2(R_{\mathbf{0}j})}{1-\rho_0^2(R_{\mathbf{0}j})} - \frac{\Delta\mu^2\rho_1(R_{\mathbf{0}j})}{\sigma_1^2(1+\rho_1(R_{\mathbf{0}j}))}, \tag{6.28}$$

$$\xi_3(R_{\mathbf{0}j}) = -\frac{\log[1-\rho_0^2(R_{\mathbf{0}j})]}{2}. \tag{6.29}$$

Proof: From [5, Thm. 1], we have the expressions for determinant and potential matrix coefficients for acyclic graphs, and we use them to simplify terms in the error exponent.

The moments condition in (6.13) holds for ξ_m for $m = 1, 2, 3$ since the terms are bounded for correlation functions $\rho_k(R_{ij})$ which are decreasing in edge lengths and $\rho_k(0) < 1$.

A function $a_n \to a$ is said to be uniformly integrable over measure ν if

$$\lim_{\alpha\to\infty} \sup_n \int_{|a_n|\geq\alpha} |a_n| d\nu = 0. \tag{6.30}$$

For uniform integrability [112, (16.21)] of normalized spectrum, it is sufficient to show that for any $n > 0$

$$\lim_{\alpha\to\infty} \int_{|\mathbf{y}^T(\boldsymbol{\Sigma}_0^{-1}-\boldsymbol{\Sigma}_1^{-1})\mathbf{y}|\geq n\alpha} \frac{1}{n}|\mathbf{y}^T(\boldsymbol{\Sigma}_0^{-1}-\boldsymbol{\Sigma}_1^{-1})\mathbf{y}| \exp\left[-\frac{\mathbf{y}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{y}}{2}\right]\mathbf{dy} = 0$$

From positive definiteness, this reduces to showing

$$\lim_{\alpha\to\infty} \int_{|\mathbf{y}^T(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}_1^{-1})\mathbf{y}|\geq \alpha} |\mathbf{y}^T(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}_1^{-1})\mathbf{y}| \exp\left[-\frac{\mathbf{y}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{y}}{2}\right]\mathbf{dy} = 0,$$

which is true. □

6.3.1 1-Nearest Neighbor Dependency

We now provide further simplifications when the dependency graphs are 1-nearest neighbor graphs. The simplification arises from the fact that limits in Theorem 18 can be further simplified for 1-nearest neighbor graph. For simplicity assume that $\mathcal{G}_1$ under $\mathcal{H}_1$ is the 1-nearest neighbor graph while the measurements are i.i.d are $\mathcal{H}_0$. We limit to uniform node placement ($\tau \equiv 1$) in this section. For this special case, the error exponent given by Theorem 18 simplifies as

$$\mathcal{D} = \frac{1}{2}\Big[\frac{1}{2}\mathbb{E} \sum_{\substack{\mathbf{X}:\mathbf{X}\in\mathcal{P}_\lambda,\\ (\mathbf{0},\mathbf{X})\in 1-\mathrm{NNG}(\mathbf{X}\cup\mathbf{0})}} f(\rho_1(R_{\mathbf{0},\mathbf{X}})) + \log K + \frac{1}{K} - 1\Big], \tag{6.31}$$

where

$$f(x) := \log[1-x^2] + \frac{2x^2}{K[1-x^2]}. \tag{6.32}$$

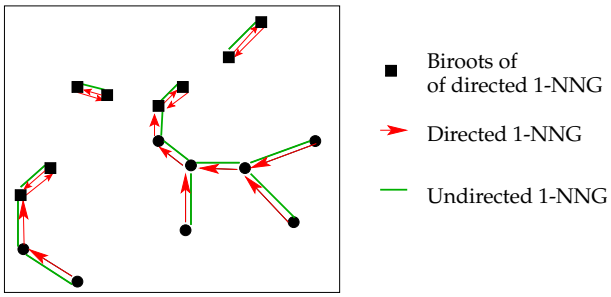


Figure 6.1: Directed & undirected versions of nearest-neighbor graph.

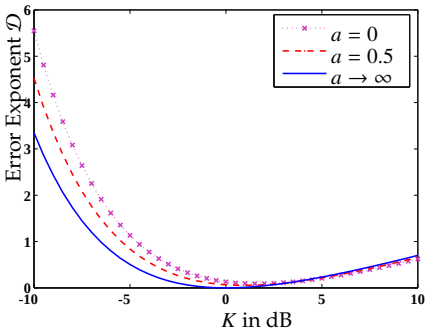
The 1-NNG has a number of important properties. It is acyclic with a maximum[3] node degree of 6 [127]. It turns out that we need to analyze the directed 1-NNG, in order to obtain the final form of the error exponent. We now mention some of its special properties. The directed 1-NNG $\mathcal{G}'(\mathbf{v}, \mathcal{E}')$ is defined by

$$E' = \{< i, \mathrm{nn}(i) >, i \in \mathbf{v}\}, \tag{6.33}$$

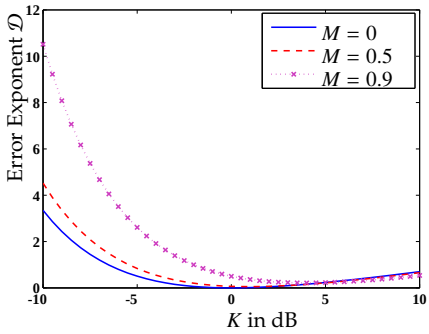
where nn denotes 1-nearest neighbor function. For a directed 1-NNG with at least two nodes, each connected component contains exactly one 2-cycle. This is known as the biroot of the component [127]. See Fig.6.1. Also note, the directed 1-NNG counts the edges from these biroots twice, while the undirected version counts only once.

We therefore split the sum of edge functionals in (8.29), using the fact that the directed 1-NNG counts the weights from biroots or mutual neighbors twice, while the undirected version counts only once. See Fig.6.1. We therefore split the sum of the edge functionals of the undirected 1-NNG as

[3]The node degree is finite for 1-NNG in any dimension and is called the kissing number [126].



(a) Different values of correlation coefficient a, nugget $M = 0.5$.



(b) Different values of nugget M, correlation coefficient $a = 0.5$.

Figure 6.2: Error exponent $\mathcal{D}$ vs. ratio of variances K, node density $\lambda = 1$.

$$\sum_{e \in 1-\mathrm{NNG}(\mathbf{v})} f(\rho_1(R_e)) = \sum_{e \in 1-\mathrm{DNNG}(\mathbf{v})} f(\rho_1(R_e)) - \frac{1}{2} \sum_{e \in 1-\mathrm{MNNG}(\mathbf{v})} f(\rho_1(R_e)), \tag{6.34}$$

where $1-\mathrm{NNG}(\mathbf{v})$, $1-\mathrm{DNNG}$ and $1-\mathrm{MNNG} \subset 1-\mathrm{DNNG}$ are the undirected 1-NNG, the directed 1-NNG, and edges between the biroots or the mutual neighbors of the directed 1-NNG, respectively. Now, we evaluate the expectation for the two terms separately, since expectation is linear. A similar approach is employed in [128].

We now provide an expression for the limit of the edge functional based on the distribution of distances of the directed 1-NNG, which are related to hitting or vacancy probabilities of the spatial point process, which are typically exponential or gamma distributed, similar to their one-dimensional counterparts [129].

Lemma 8 (Expectation of Edge Functional) *The expectation term of the edge functional in (6.31) is given by*

$$\frac{1}{2}\mathbb{E}\sum_{\substack{\mathbf{X}:\mathbf{X}\in\mathcal{P}_\lambda,\\(\mathbf{0},\mathbf{X})\in 1-NNG(\mathbf{X}\cup\mathbf{0})}} f(\rho_1(R_{\mathbf{0},\mathbf{X}})) = \mathbb{E}f(\rho_1(Z_1)) - \frac{\pi}{2\omega}\mathbb{E}f(\rho_1(Z_2)), \tag{6.35}$$

where Z_1 and Z_2 are Rayleigh distributed with variances $(2\pi\lambda)^{-1}$ and $(2\omega\lambda)^{-1}$, and ω is given by

$$\omega = \frac{4\pi}{3} + \frac{\sqrt{3}}{2} \approx 5.06, \tag{6.36}$$

and is the area of the union of two unit- radii circles with centers unit distant apart.

Proof : See Appendix 6.A. □

In the theorem below, we obtain the final form of the error exponent.

Theorem 19 (Expression for $\mathcal{D}$) *For a GMRF with 1-NNG dependency and correlation function ρ_1 and nodes drawn from the binomial or the Poisson process with node density λ and region area $\frac{n}{\lambda}$, the error exponent $\mathcal{D}$ for Neyman-Pearson detection is*

$$\begin{aligned}\mathcal{D}_{\rho_1}(K, M, \lambda) \quad = \quad & \frac{1}{2}[\mathbb{E}f(\rho_1(Z_1), K) - \frac{\pi}{2\omega}\mathbb{E}f(\rho_1(Z_2), K) \\ & + \log K + \frac{1}{K} - 1],\end{aligned} \tag{6.37}$$

where

$$f(x, K) := \log[1 - x^2] + \frac{2x^2}{K[1 - x^2]}. \tag{6.38}$$

Z_1 and Z_2 are Rayleigh distributed with second moments $(2\pi\lambda)^{-1}$ and $(2\omega\lambda)^{-1}$.

The above theorem holds for any general correlation function. In (6.37), except for the first two f-terms which capture the correlation structure of the GMRF, the remaining terms represent the detection error exponent for two IID Gaussian processes. In the corollary below, we specialize (6.37) to the case of constant correlation. In this case, the two f-terms reduce to a single term.

Corollary 2 (Constant Correlation) *For constant values of the correlation, the error exponent $\mathcal{D}$ is independent of the node density λ and*

1. *for constant positive correlation or $\rho_1(R_e) \equiv M < 1$, $\forall e \in E$, we have*

$$\begin{aligned} \mathcal{D}(K, M) \quad = \quad & \frac{1}{2}\Big[\log K + \frac{1}{K} - 1 \\ & + (1 - \frac{\pi}{2\omega}) f(M, K)\Big], \end{aligned} \tag{6.39}$$

   where f and ω are given by (6.32) and (6.36).

2. *for the independent case or $\rho_1(R_e) \equiv 0$, $\forall e \in E$, we have*

$$\mathcal{D}(K, 0) = \frac{1}{2}\Big[\log K + \frac{1}{K} - 1\Big]. \tag{6.40}$$

In the above corollary, we verify that (6.39) reduces to (6.40), on substituting $M = 0$. In (6.39), the effect of correlation can be easily analyzed through the sign of the function $f(M, K)$. Also,

$$f(M,K)\begin{cases} < 0, & \text{for } K > \frac{2}{1-M^2}, \quad (6.41a) \\ > 0, & \text{for } K < 2. \quad (6.41b) \end{cases}$$

Therefore, at large variance-ratios, the presence of correlation hurts the asymptotic performance, when compared with the independent case. But the situation is reversed at low values of the variance ratio and the presence of correlation helps in detection performance. In the next section, we will draw similar conclusions when the correlation function is the exponential function through numerical evaluations.

6.3.2 Numerical Results

In this section, we focus on a specific correlation function namely the exponential-correlation function,

$$\rho_1(R) = Me^{-aR}, \ a > 0, 0 < M < 1. \quad (6.42)$$

Using Theorem 19, we numerically evaluate $\mathcal{D}$ through Monte-Carlo runs. In (6.37), the error exponent is an implicit function of the correlation coefficient a, through the correlation function ρ_1. For fixed values of K and M, we have

$$\mathcal{D}(K,M,\lambda,a) = \mathcal{D}(K,M,1,\frac{a}{\sqrt{\lambda}}), \quad (6.43)$$

which we obtain by changing the integration variable in the expectation term in (6.37). Therefore, in terms of the error exponent, increasing the node density

λ is equivalent to a lower correlation coefficient at unit density. Here, we plot only the effects of correlation coefficient a and nugget M on $\mathcal{D}$.

In Fig.6.2(a), we plot the error exponent at $\lambda = 1$ and $M = 0.5$, for different values of correlation coefficient a. Note, the cases $a = 0$ and $a \to \infty$ correspond to (6.39) and (6.40). We notice that a more correlated GMRF or the one with smaller a, has a higher exponent at low value of K, whereas the situation is reversed at high K. Equivalently, increasing the node density λ improves the exponent at low value of K, but not at high K. Also, when the variance ratio K is large enough, $\mathcal{D}$ appears to increase linearly with K (in dB), and the correlation coefficient a and nugget M appear to have little effect, as expected from Theorem 19. In Fig.6.2(b), we plot the exponent at constant correlation coefficient $a = 0.5$ for different values of the nugget M. Also note, $M = 0$ reduces to the independent case. We notice a similar behavior as the correlation coefficient. A higher value of M results in a higher exponent at low K, but not at high K.

6.A Proofs

Proof of Lemma 8

We use an approach similar to [128]. Let $B_z(\mathbf{X})$ denote a circle of radius z, centered at $\mathbf{X}$. We take expectation on both sides of (6.34) for graphs over all the Poisson points $\mathbf{X} \cup \mathbf{0}$. Let $1-\mathrm{NNG}(\mathbf{v})$, $1-\mathrm{DNNG}$ and $1-\mathrm{MNNG} \subset 1-\mathrm{DNNG}$ be the undirected 1-nearest neighbor graph, the directed nearest-neighbor graph, and edges between the biroots or the mutual neighbors of the directed 1-nearest neighbor graph. See Fig.6.1.

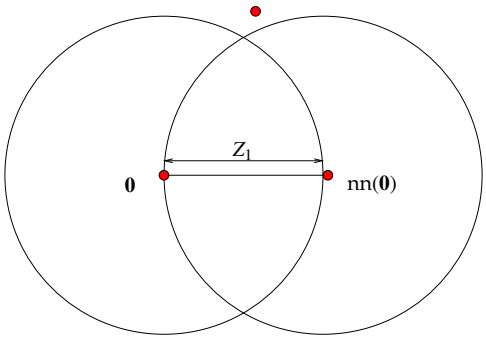


Figure 6.3: Event that the origin is a biroot in the directed 1-NNG.

$$\mathbb{E}[\sum_{\substack{\mathbf{X}:\mathbf{X}\in\mathcal{P}_\lambda \\ (\mathbf{0},\mathbf{X})\in 1-\mathrm{NNG}(\mathbf{X})}} f(\rho_1(R_{\mathbf{0},\mathbf{X}}))]=\mathbb{E}[\sum_{\substack{\mathbf{X}:\mathbf{X}\in\mathcal{P}_\lambda, \\ (\mathbf{0},\mathbf{X})\in 1-\mathrm{DNNG}(\mathbf{X})}} f(\rho_1(R_{\mathbf{0},\mathbf{X}}))] - \frac{1}{2}\mathbb{E}[\sum_{\substack{\mathbf{X}:\mathbf{X}\in\mathcal{P}_\lambda \\ (\mathbf{0},\mathbf{X})\in 1-\mathrm{MNNG}(\mathbf{X})}} f(\rho_1(R_{\mathbf{0},\mathbf{X}}))]. \tag{6.44}$$

The first term on the right-hand side in (6.44) simplifies as

$$\mathbb{E}[\sum_{\substack{\mathbf{X}:\mathbf{X}\in\mathcal{P}_\lambda, \\ (\mathbf{0},\mathbf{X})\in 1-\mathrm{DNNG}(\mathbf{X})}} f(\rho_1(R_{\mathbf{0},\mathbf{X}}))] = \mathbb{E}[f(\rho_1(Z_1))], \tag{6.45}$$

where Z_1 is the unique directed nearest-neighbor distance of the origin with points distributed according to $\mathcal{P}_\lambda$, the Poisson point process of intensity λ on $\Re^2$. The random variable Z_1 is like a waiting time, and can be visualized as the time taken for an inflating circle to first touch a point from the Poisson process. We therefore have $Z_1 > z$ iff. $B_z(\mathbf{0})$ does not contain any points from the Poisson process, i.e.,

$$\mathbb{P}[Z_1 > z] = \mathbb{P}[\nexists \mathbf{X} \neq \mathbf{0} \in B_z(\mathbf{0}) \cap \mathcal{P}_\lambda] = \mathrm{e}^{-\lambda\pi z^2}. \tag{6.46}$$

Therefore, Z_1 is Rayleigh with second moment $(2\pi\lambda)^{-1}$.

Similarly, for the second term, we need to find the PDF of the nearest-neighbor distance of the origin when the origin is a biroot or a mutual nearest neighbor. This event occurs when the union of the circles centered at origin and its nearest neighbor contains no other Poisson point. See Fig.6.3. Let $\mathcal{A}$ be the intersection of the events that the directed nearest-neighbor distance of origin lies in the interval $[z, z + dz]$ and the event that origin is a biroot

$$\begin{aligned} \mathcal{A} \;:=\; & (\mathcal{P}_\lambda \cap (B_z(\mathbf{0}) \cup B_z(\mathrm{nn}(\mathbf{0})))\backslash\{\mathbf{0}, \mathrm{nn}(\mathbf{0})\} = \emptyset) \\ & \cap (Z_1 \in [z, z + dz]). \end{aligned} \tag{6.47}$$

Its probability is given by,

$$\begin{aligned} \mathbb{P}[\mathcal{A}] &= \mathbb{P}(\text{origin is biroot}|Z_1)\mathbb{P}(Z_1 \in [z, z + dz]) \\ &= \mathrm{e}^{-(\omega-\pi)\lambda z^2} 2\lambda\pi z \mathrm{e}^{-\lambda\pi z^2} dz \qquad (6.48) \\ &= 2\lambda\pi z \mathrm{e}^{-\omega\lambda z^2} dz = \frac{\lambda}{\omega}[2\omega\pi z \mathrm{e}^{-\omega\lambda z^2} dz] \qquad (6.49) \\ &= \frac{\lambda}{\omega}\mathbb{P}(Z_2 \in [z, z + dz]), \qquad (6.50) \end{aligned}$$

where $\mathrm{nn}(\mathbf{0})$ is the nearest-neighbor of the origin and $\omega := |B_1(\mathbf{0}) \cup B_1(\mathbf{1})| = \frac{4\pi}{3} + \frac{\sqrt{3}}{2}$, the area of the union of circles unit distant apart and Z_2 is a Rayleigh variable

with variance $(2\pi\omega)^{-1}$. Hence, the second term on the right-hand side in (6.44) simplifies as

$$\frac{1}{2}\mathbb{E}[\sum_{\substack{\mathbf{X}:\mathbf{X}\in\mathcal{P}_\lambda \\ (\mathbf{0},\mathbf{X})\in 1-\mathrm{MNNG}(\mathbf{X})}} f(\rho_1(R_{\mathbf{0},\mathbf{X}}))] = \frac{\pi}{2\omega}\mathbb{E}[f(\rho_1(Z_2))]. \tag{6.51}$$

From (6.34, 6.46, 6.50), we obtain (6.35). □

CHAPTER 7

ENERGY-CONSTRAINED INFERENCE IN RANDOM NETWORKS

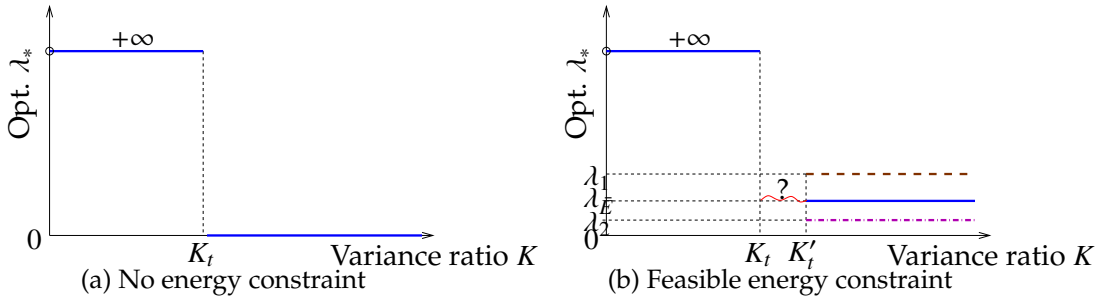
In this chapter, we consider the optimization of sensor density of a multi-hop sensor network in the context of energy-constrained distributed detection. For large sensor networks, it is unrealistic to optimize individual sensor locations. It is then natural to consider random deployment of sensors where the sensor density becomes the key design parameter. Optimizing sensor density is not only important for sensor deployment but also gives a simple decentralized sensor transmission strategy by deciding to transmit under independent coin flips.

To characterize the detection performance, we consider the Neyman-Pearson (NP) error exponent $\mathcal{D}$, discussed in the previous chapter. Our objective is to find an optimal node density λ_* that maximizes the detection error exponent $\mathcal{D}_\lambda$, under a constraint $\bar{E}$ on the average (per node) energy consumption $\bar{\mathcal{E}}(\lambda)$, when the number of nodes goes to infinity.

$$\lambda_* := \arg\max_{\lambda>0} \mathcal{D}_\lambda \quad \text{subject to } \bar{\mathcal{E}}(\lambda) \le \bar{E}. \tag{7.1}$$

We address the following questions: does an optimal node density exist? And if so, what is its value? Is it one of the extremes, viz., zero or infinity? This is an important question, since if the optimal node density is either zero or infinity, then we can simply place the nodes in as small or large an observation area as possible.

We use the Gaussian inference model with nearest-neighbor dependency



(a) No energy constraint (b) Feasible energy constraint

Figure 7.1: Optimal node density vs. variance ratio K. See Theorem 21.

graph, discussed in the previous chapter in Section 6.3.1 with an additional assumption that the correlation function under $\mathcal{H}_1$, given by $\rho(r)$ is convex, and employ the error exponent results, derived in Theorem 19. We limit to uniform distribution for node placement in this chapter.

7.1 Overview of Results and Approach

Given a per-node energy budget for data fusion, we aim to find the sensor density that optimizes the detection performance. Assuming that all the nodes have the same measurement variance σ_j^2 under each hypothesis $\mathcal{H}_j$, recall the variance ratio as

$$:= Variance Ratio of Two Gaussian Hypotheses K := \frac{\sigma_1^2}{\sigma_0^2}. \tag{7.2}$$

The main results demonstrate the presence of a threshold K_t effect on the variance ratio K. As shown in Figure 7.1, when K is below a threshold K_t, the optimal density is unbounded, and thus it is optimal to concentrate sensors near the fusion center. Moreover, this result is independent of the energy constraint implying that imposing an energy constraint does not degrade detection per-

formance. On the other hand, when K is above the threshold K_t and no energy constraint is imposed, the optimal density tends to zero (Fig. 7.1a), which in practice, implies that it is optimal to disperse sensors in the largest possible area. In this regime of K, imposing an energy constraint leads to a strictly finite optimal density and we provide bounds on this optimal value (Fig. 7.1b).

We give a closed-form expression for the threshold K_t on the variance ratio K which decides the behavior of the optimal density. The threshold K_t is independent of the energy constraint. Moreover, somewhat surprisingly, it depends on the correlation structure only through the limiting correlation of two sensors as their separation distance vanishes.

We also investigate the use of an energy density constraint where the total energy consumption in a given area is constrained instead of the average energy at each node. We show that an optimal node density under this formulation exists, and is strictly finite for all values of the variance ratio K. This is in sharp contrast to the threshold behavior under a per-node energy constraint. We prove analytical bounds for the optimal density when the variance ratio $K < 2$.

This chapter is organized as follows. Results on optimal density are in section 7.2 and numerical analysis in section 7.2.5. The energy-density formulation is dealt in section 7.3. Section 7.4 concludes the chapter.

7.2 Optimal Node Density

Recall that our objective is to find an optimal node density λ_* maximizing the detection error exponent under a constraint on the average energy consumption

of data fusion.

$$\lambda_* := \arg\max_{\lambda>0} \mathcal{D}_\lambda \quad \text{subject to } \bar{\mathcal{E}}(\lambda) \leq \bar{E}.$$

We have so far derived a closed-form expression for the error exponent $\mathcal{D}_\lambda$ (short form for $\mathcal{D}(\lambda, K; \rho)$ in (7.8)) and bounds for optimal average energy consumption $\bar{\mathcal{E}}_*(\lambda)$. In this section, we exploit these expressions to derive the optimal node density.

Before proceeding to the actual derivations, it is useful to consider a special case, viz., when both the hypotheses have the same measurement variance ($K = (\frac{\sigma_1}{\sigma_0})^2 = 1$) and there is no energy constraint ($\bar{E} \to \infty$). Since there is no correlation under $\mathcal{H}_0$, the two hypotheses can be distinguished only by the presence of correlation under $\mathcal{H}_1$. Correlation is maximized when all the nodes are clustered close to one another, since correlation decays with distance. Hence, the optimal density should be infinite. We prove that this is indeed true and also characterize the optimal density for general K and energy constraint $\bar{E}$. To this end, we first characterize the set of feasible node densities which can support data fusion under the given energy constraint $\bar{E}$.

7.2.1 Detection Error Exponent

In this section, we rewrite the error exponent derived in Theorem 19 in a more convenient form, in terms of the variables and functions defined below.

$$
\begin{aligned}
f_1(x) &:= \log[1-x^2], f_2(x) := \frac{2x^2}{[1-x^2]}, && (7.3)\\
f(x,K) &:= f_1(x) + \frac{1}{K} f_2(x) && (7.4)\\
h_i(x;\rho) &:= f_i(\rho(x)) - \frac{\pi}{2\omega} f_i(\rho(\sqrt{\frac{\pi}{\omega}}x)), && (7.5)\\
h(x,K;\rho) &:= h_1(x;\rho) + \frac{1}{K} h_2(x;\rho), && (7.6)
\end{aligned}
$$

where $\rho(\cdot)$ is the correlation function. Let Z denote the Rayleigh random variable with variance $(2\pi)^{-1}$ as in the previous chapter and recall that in (6.36) ω is the area of the union of two unit- radii circles with centers unit distance apart, given by

$$
\omega = \frac{4\pi}{3} + \frac{\sqrt{3}}{2} \approx 5.06. \tag{7.7}
$$

Theorem 20 (Expression for $\mathcal{D}$) *For a GMRF on NNG with correlation function ρ, with the nodes drawn from the binomial or the Poisson process with node density λ and region area $\frac{n}{\lambda}$, the error exponent $\mathcal{D}$ for Neyman-Pearson detection is*

$$
\mathcal{D}(\lambda, K; \rho) = \frac{1}{2}\Big[\mathbb{E}_\lambda\, h(Z\lambda^{-0.5}, K; \rho) + \log K + \frac{1}{K} - 1\Big], \tag{7.8}
$$

where $\mathbb{E}_\lambda$ is the expectation over the random variable Z.

Proof: Note that

$$\mathcal{D}(\lambda, K; \rho) = \frac{1}{2} \Big[\lim_{n\to\infty} \frac{1}{n} \sum_{e\in \mathrm{NNG}} f(\rho(R_e), K; \rho) + \log K + \frac{1}{K} - 1 \Big] \tag{7.9}$$

□

Note that in (7.8), the expectation term captures the correlation structure of the GMRF and the remaining terms represent the detection error exponent for two i.i.d. Gaussian processes with variance ratio K, i.e.,

$$\mathcal{D}(\lambda, K; \rho) = \frac{1}{2} \mathbb{E}_\lambda \, h(Z\lambda^{-0.5}, K; \rho) + \mathcal{D}_{IID}(K), \tag{7.10}$$

where $\mathcal{D}_{IID}(K)$ the error exponent when the measurements are conditionally i.i.d., and is given by

$$\mathcal{D}_{IID}(K) = \frac{1}{2} \Big[\log K + \frac{1}{K} - 1 \Big]. \tag{7.11}$$

Hence, the effect of correlation on the error exponent is quantified in a compact form. It can be easily verified that the expectation term is zero, when $\rho(0) = M = 0$ (no correlation).

It is easy to see that $\mathcal{D}_{IID}(K)$ is independent of the node density λ. Hence, the issue of optimal node density for error exponent arises only in the presence of correlation. Moreover, intuitively, as we vary node density λ, the edge lengths in the resulting NNG scale by the factor $\lambda^{-0.5}$, on an average. Hence, we see that this factor $\lambda^{-0.5}$ appears in the expectation term in (7.10).

7.2.2 Feasible Node Density Set

In order to incorporate the energy constraint $\bar{E}$, we need to find a feasible set Λ of node densities for minimum energy routing that satisfies the average energy constraint $\bar{E}$ and delivers the LLR to the fusion center,

$$\Lambda(\bar{E}):=\{\lambda : \bar{\mathcal{E}}_*(\lambda) \leq \bar{E}\}. \tag{7.12}$$

When the density goes to infinity, the optimal average energy consumption for routing goes to zero

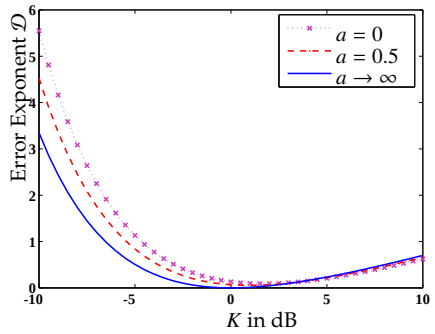
$$\lim_{\lambda\to\infty} \bar{\mathcal{E}}_*(\lambda) = 0.$$

This is derived from the bounds on $\bar{\mathcal{E}}_*$. Hence, the energy constraint $\bar{E}$ is satisfied at infinite density or in other words, $\lambda \to \infty \in \Lambda$ for Λ defined in (7.12). Hence, whenever Λ is non-empty, it is of the form

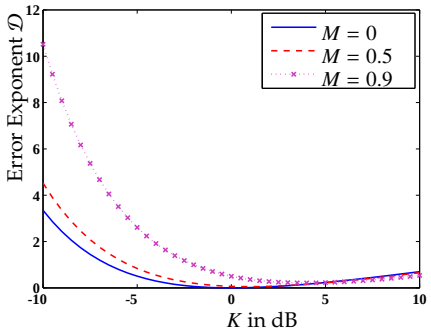
$$\Lambda(\bar{E}) = [\lambda_{\bar{E}}, \infty], \tag{7.13}$$

where $\lambda_{\bar{E}}$ is defined as the minimum node density under the energy constraint $\bar{E}$ at which it is feasible to perform data fusion. However, as discussed in Chapter 3, finding the minimum energy scheme is NP-hard. Hence, finding an expression for $\lambda_{\bar{E}}$ is analytically intractable. We instead provide bounds on $\lambda_{\bar{E}}$.

We first consider feasible node density for DFMRF scheme. Recall from Corollary 1 that the average energy consumption DFMRF under 1-NNG dependency and uniform placement satisfies

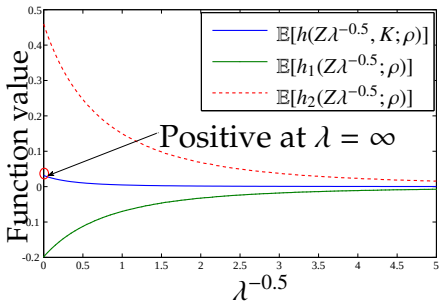


(a) Different values of correlation coefficient a, limiting correlation $M = 0.5$.

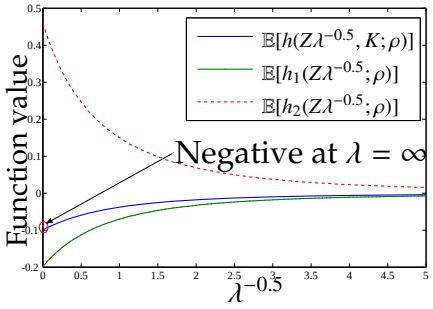


(b) Different values of limiting correlation M, correlation coefficient $a = 0.5$.

Figure 7.2: Error exponent $\mathcal{D}$ vs. variance ratio K.



(a) Correlation provides positive contribution. $K = 2 < K_t$.



(b) Correlation provides negative contribution. $K = 2K_t$.

Figure 7.3: Trend under correlation $\rho(R) = M \exp[-aR]$, $M = 0.5, a = 1$.

$$\limsup_{n\to\infty} \bar{\mathcal{E}}(\text{DFMRF}(\mathbf{V}_n)) \leq \lambda^{-\frac{\nu}{2}} [u\,\zeta(\nu; \text{1-NNG}) + \zeta(\nu; \text{MST})]. \tag{7.14}$$

Hence, imposing the energy constraint $\bar{E}$ translates to a constraint on λ under the DFMRF,

$$\bar{\mathcal{E}}(\text{DFMRF}(\lambda)) \leq \bar{E} \iff \lambda \geq \lambda_1(\bar{E}) := \Big(\frac{\bar{E}}{u\,\zeta(\nu; \text{1-NNG}) + \zeta(\nu; \text{MST})}\Big)^{\frac{2}{\nu}}. \tag{7.15}$$

Similarly, we use scaling results for lower bound on optimal energy consumption, given by the MST, derived in Theorem 10

$$\lim_{n\to\infty} \bar{\mathcal{E}}(\mathrm{MST}(\mathbf{V}_n)) = \lambda^{-\frac{\nu}{2}}\zeta(\nu;\mathrm{MST}). \tag{7.16}$$

Hence,

$$\bar{\mathcal{E}}(\mathrm{MST}(\lambda)) \le \bar{E} \iff \lambda \ge \lambda_2(\bar{E}) := \Big(\frac{\bar{E}}{\zeta(\nu;\mathrm{MST})}\Big)^{\frac{2}{\nu}}. \tag{7.17}$$

From the bounds on optimal energy, we have bounds on the feasible set of node densities Λ for the minimum energy scheme,

$$\{\lambda : \lambda \ge \lambda_2(\bar{E})\} \subset \Lambda(\bar{E}) \subset \{\lambda : \lambda \ge \lambda_1(\bar{E})\}. \tag{7.18}$$

From the definition of $\lambda_{\bar{E}}$ in (7.13), we have

$$\lambda_2(\bar{E}) \le \lambda_{\bar{E}} \le \lambda_1(\bar{E}). \tag{7.19}$$

Hence, although we are unable to evaluate $\lambda_{\bar{E}}$, we have bounds that are easily evaluated.

7.2.3 Infinite Node Density

We first analyze the error exponent $\mathcal{D}$, when the node density λ goes to infinity. This will provide insights for finding the optimal density. As $\lambda \to \infty$, we have

$$\mathcal{D}(\infty, K; \rho) = \mathcal{D}_{IID}(K) + \frac{1}{2}h(0, K, M),$$

where $h(0, K; \rho)$ depends on ρ only through $\rho(0) = M$, and

$$h(0, K, M) = (1 - \frac{\pi}{2\omega})\Big(\log[1 - M^2] + \frac{2M^2}{K[1 - M^2]}\Big). \tag{7.20}$$

In the theorem below, we prove that the presence of correlation can either improve or degrade the error exponent, depending on the variance ratio K. We establish a threshold on K that determines the transition.

Lemma 9 (Behavior at Infinite Density $(\lambda \to \infty)$**)** *At $\lambda \to \infty$, the correlation term $h(0, K, M)$ in (7.20) is positive, if the variance ratio K is below a threshold value $K_t(M)$,*

$$h(0, K, M) \begin{cases} \geq 0, & \text{for } K < K_t(M), \quad (7.21a) \\ < 0, & \text{for } K > K_t(M). \quad (7.21b) \end{cases}$$

For a fixed $\rho(0) = M < 1$, the threshold $K_t(M)$ is

$$K_t(M) = -\frac{1}{\log(1 - M^2)} \frac{2M^2}{1 - M^2}, \tag{7.22}$$

and $2 < K_t(M) < \frac{2}{1-M^2}$.

Proof: From (7.20) and $(1 - \frac{\pi}{2\omega}) > 0$. □

Hence, we obtain a somewhat surprising result that at infinite node density, the effect of correlation on error exponent is different based on the variance ratio K and is determined by a threshold K_t on K. For values of K below the threshold K_t, the presence of correlation improves the error exponent in (7.21a). On other hand, above the threshold K_t, the presence of correlation degrades the error exponent in (7.21b). Moreover, at infinite density since the inter-node distances go to zero, the correlation function is given by $\rho(0) = M$, and hence, the threshold K_t in (7.21) is only a function of M.

Although the results in Lemma 9 are valid only at infinite density, we can utilize them to compare with the other extreme scenario when the density $\lambda \to 0$. In this case, the error exponent $\mathcal{D}_\lambda \to \mathcal{D}_{IID}$, i.e., the conditionally i.i.d. case. From Lemma 9, we can conclude that below the threshold K_t, it is better to cluster the nodes close to one another ($\lambda \to \infty$) rather than place them as far as possible ($\lambda \to 0$). On the other hand, above the threshold, the opposite is true. Hence, the results for infinite node density in Lemma 9 provide guidelines on the effect of correlation on the error exponent. In the next section, we will generalize these results to prove that the optimal node density displays a threshold behavior.

7.2.4 Threshold Behavior of Optimal Density

In this section, we provide the results for optimal density. From (7.13), we can rewrite density optimization in (7.1) as

$$\lambda_* = \arg\max_{\lambda \geq \lambda_{\bar{E}}} \mathcal{D}_\lambda = \arg\max_{\lambda \geq \lambda_{\bar{E}}} \mathbb{E}_\lambda \, h(Z\lambda^{-0.5}, K; \rho), \tag{7.23}$$

where for the last equality, we use the fact that $\mathcal{D}_{IID}(K)$ is independent of λ in (7.10).

To analyze the behavior of the expectation term in (7.23), we first focus on the function f in (7.4) given by

$$f(x, K) = \log(1 - x^2) + \frac{2}{K} \frac{x^2}{1 - x^2}, \quad \forall x \in [0, M]. \tag{7.24}$$

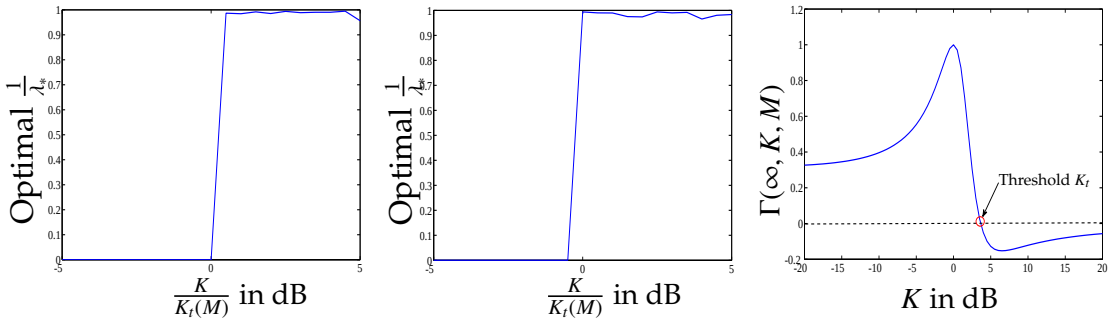
Since $\rho(0) = M$ and $\rho(\infty) = 0$, we have $x \in [0, M]$. In the lemma below, we provide results on the behavior of f.

Lemma 10 (Behavior of f) *The function $f(x, K)$ in (7.24) with $M < 1$ satisfies*

$$\arg\max_{x \in [0,M]} f(x, K) = \begin{cases} M, & K < K_t, \quad (7.25a) \\ 0, & o.w. \quad (7.25b) \end{cases}$$

Proof: See Appendix 7.A. □

Hence, the function $f(x, K)$ attains its maximum only at one of the boundary points for $x \in [0, M]$. The particular boundary point is determined by a threshold K_t on K, as seen in (7.25) and also K_t depends only on M, the limiting correlation.



(a) Optimal density λ_* decided by threshold $K_t(M)$ on variance ratio K. $M = 0.1$

(b) Optimal density λ_* decided by threshold $K_t(M)$ on variance ratio K. $M = 0.8$

(c) Sensitivity analysis: Correlation contribution to $\mathcal{D}$ vs. variance ratio K at $\lambda = \infty$ and $M = 0.5$. See (7.32).

Figure 7.4: Optimal density λ_* decided by threshold $K_t(M)$.

When the point sets are drawn from binomial or Poisson processes, and R_e are the edge-lengths of NNG , consider the edge functional

$$\frac{1}{n} \sum_{e \in \text{NNG}} f(\rho(R_e), K; \rho).$$

From (7.8) and (7.9), we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{e \in \mathcal{E}_d} f(\rho(R_e), K; \rho) = \mathbb{E}[h(Z\lambda^{-0.5}, K; \rho)]. \tag{7.26}$$

Hence, we can use the result on the maximum of function f in Lemma 10 to find the corresponding optimal density maximizing the expectation term in (7.23). In the theorem below, we provide such a result on the optimal density and show that its behavior is determined only by the thresholds K_t and K_t' on K.

Theorem 21 (Result on $\lambda_*(K, \bar{E})$) *The optimal density in (7.23) that maximizes the error exponent, under feasible average energy constraint $\bar{E}$, is given by*

$$\lambda_*(K, \bar{E}) = \infty, \quad \forall\, K < K_t(M), \tag{7.27}$$

where the threshold K_t is given by (7.22), and

$$\lambda_*(K, \bar{E}) = \lambda_{\bar{E}} < \infty, \quad \forall\, K > K_t'(M), \tag{7.28}$$

where $\lambda_{\bar{E}}$ is defined in (7.13), and satisfies bounds in (7.19), and

$$K_t'(M) = \frac{2}{1 - M^2} > K_t(M), \tag{7.29}$$

where $M = \rho(0) < 1$ is the correlation function as the inter-node distance goes to zero. Also, when the energy constraint is infinite $\bar{E} \to \infty$, we have $\lambda_{\bar{E}} = 0$, and the result in (7.28) is improved to

$$\lambda_*(K, \infty) = 0, \quad \forall\, K > K_t(M). \tag{7.30}$$

Proof: See Appendix 7.A. □

The above theorem states that when the variance ratio K is below the threshold K_t, for any feasible energy constraint $\bar{E}$, optimality is attained at infinite density. On the other hand, above another threshold $K_t' > K_t$, the minimum feasible node density $\lambda_{\bar{E}}$ which supports data fusion under constraint $\bar{E}$ attains the optimal value.

In the special case, when there is infinite energy ($\bar{E} \to \infty$), we have $\lambda_1(\bar{E}) = \lambda_2(\bar{E}) = \lambda_{\bar{E}} = 0$. In this case, we prove that optimal value is zero even in the gap region between the two thresholds K_t and K_t'. Numerical investigation suggests that under any finite energy constraint, the optimal density is $\lambda_{\bar{E}}$ even in the gap region between the two thresholds K_t and K_t'.

Also interestingly, the thresholds K_t and K_t' depend on correlation $\rho(R)$ only through the limiting value $\rho(0) = M$. This is because of the behavior of the function f, elucidated in Lemma 10. We also note that as $M \to 1$, the thresholds $K_t, K_t' \to \infty$. Hence, when we approach full correlation as the inter-node distance goes to zero, the optimal node density tends towards infinity for all values of K.

7.2.5 Numerical Analysis

In this section, we plot the error exponent and optimal node density for a specific correlation function namely the exponential-correlation function,

$$\rho(R) = M\mathrm{e}^{-aR}, \ a > 0, 0 < M < 1. \tag{7.31}$$

Using Theorem 20, we numerically evaluate $\mathcal{D}$ through Monte-Carlo runs. In (7.8), the error exponent is an implicit function of the correlation coefficient a, through the correlation function ρ. We plot the effects of correlation coefficient a and limiting correlation M on $\mathcal{D}$ in Fig.7.2.

In Fig.7.2(a), we plot the error exponent at $\lambda = 1$ and $M = 0.5$, for different values of correlation coefficient a. We notice that a more correlated GMRF or

the one with smaller a, has a higher exponent at low value of K, whereas the situation is reversed at high K. Also, when the variance ratio K is large enough, $\mathcal{D}$ appears to increase linearly with K (in dB), and the correlation coefficient a and the limiting correlation M appear to have little effect, as expected from Theorem 20. In Fig.7.2(b), we plot the exponent at constant correlation coefficient $a = 0.5$ for different values of the limiting correlation M. Also note, $M = 0$ reduces to the independent case. We notice a similar behavior as the correlation coefficient a in Fig.7.2(a). A higher value of M results in a higher exponent at low K, but not at high K.

In Fig.7.3, we fix the correlation coefficient $a = 1$ in (7.31), and plot the expectations of functions h, h_1 and h_2 against $\lambda^{-0.5}$. In Fig.7.3a and Fig.7.3b, the value of K is below and above the threshold K_t. We observe that the behavior at $\lambda = \infty$ is different in the two plots. Note that the functions h_1 and h_2 are independent of K, but K affects their scaling in h.

In Fig.7.4a and Fig.7.4b, we numerically evaluate the optimal $\lambda_*(K, \lambda_{\bar{E}})$ for different values of M and variance ratio K. It is convenient to plot the results in terms of $\lambda^{-0.5}$, since the optimal λ_* is infinite when K is below the threshold K_t. We observe the threshold behavior at K_t, as predicted in Theorem 21: when $K < K_t$, we have $\lambda_* = \infty$ and for $K > K_t$, in Fig.7.4a and Fig.7.4b, optimality is mostly attained at the other extreme point $\lambda_{\bar{E}}$. This is consistent with Theorem 21.

7.2.6 Sensitivity Analysis

In Theorem 21, we proved the result on the optimal density λ_*. In this section, we analyze the extent to which the error exponent is dependent on λ. This

enables us to gauge the usefulness of finding λ_*. To this end, we define

$$\Gamma(\lambda, K; \rho) := \frac{\mathbb{E}[Z\lambda^{-0.5}, K; \rho]}{2\mathcal{D}(\lambda, K; \rho)}, \tag{7.32}$$

which is the fraction of contribution coming from correlation to the error exponent in (7.10), and hence, it is the part influenced by λ. Note, $0 \leq \Gamma(\lambda, K; \rho) \leq 1$ and $\Gamma(0, K; \rho) = 0$.

Lemma 11 (Sensitivity of $\mathcal{D}_\lambda$ to λ) *At $K = 1$, the fraction of contribution from correlation to the error exponent $\Gamma(\lambda, K; \rho)$ is maximum,*

$$\Gamma(\lambda, 1; \rho) = 1, \quad \forall\, \lambda > 0. \tag{7.33}$$

Also, in the large-K regime,

$$\Gamma(\lambda, K; \rho) \to 0, \quad \text{as } K \to \infty, \forall\, \lambda \in \mathfrak{R}^+. \tag{7.34}$$

Hence, node density greatly influences detection performance at $K = 1$. Intuitively, this is because at $K = 1$, the two hypotheses can only be distinguished through the presence of correlation under $\mathcal{H}_1$. We also see that it decays to zero as $K \to \infty$. Hence, the error exponent is insensitive to changes in density at high K. In Fig.7.4c, we plot $\Gamma(\lambda, K; \rho)$ as a function of K.

7.3 Energy-density Constraint

We now analyze the optimal node density under a different formulation. Instead of having fixed number of nodes placed in varying areas under a given constant node density, in this section, we fix areas of deployment A according to a constant node density λ and then, let $A \to \infty$. This means that we now have the flexibility of placing few powerful nodes or many cheaper nodes in a given area. On the other hand, earlier we had the option of choosing the area of deployment for a fixed number of nodes.

Under this formulation, the processing energy C_p at each node needs to be incorporated. We impose an energy density constraint κ

$$\kappa \geq \lim_{\mathcal{E},A\to\infty} \frac{\mathcal{E}}{A}, \tag{7.35}$$

where $\mathcal{E}$ is the total (routing + processing) energy consumption $\mathcal{E}$ in area A. Hence, we impose a constraint on the energy consumption per unit area, instead of a constraint on the average energy consumption per node, as in (7.1).

The error exponent $\mathcal{D}_\lambda^p$ with respect to total energy $\mathcal{E}$ is

$$\mathcal{D}_\lambda^p := \lim_{\mathcal{E}\to\infty} -\frac{1}{\mathcal{E}} \log P_M(\mathcal{E}). \tag{7.36}$$

The optimal density λ_*^p under the energy-density constraint is

$$\lambda_*^p := \max_{\lambda\in\Re^+} \mathcal{D}_\lambda^p, \tag{7.37}$$

subject to the energy-density constraint κ in (7.35).

Note that the error exponent $\mathcal{D}_\lambda^p$ with respect to the total energy $\mathcal{E}$ can be expressed as

$$\mathcal{D}_\lambda^p = \frac{\mathcal{D}_\lambda}{\bar{\mathcal{E}}(\lambda)}, \tag{7.38}$$

where $\mathcal{D}_\lambda$ is the error exponent with respect to number of nodes and $\bar{\mathcal{E}}(\lambda)$ is the average energy per node, dealt in the previous sections. Similarly, the energy density constraint simplifies to

$$\kappa \geq \lim_{\mathcal{E},A\to\infty} \frac{\mathcal{E}}{A} \approx \lambda \frac{\mathcal{E}}{\lfloor \lambda A \rfloor} = \lambda \bar{\mathcal{E}}(\lambda), \quad \lambda \in \mathfrak{R}^+, \tag{7.39}$$

since $\lfloor \lambda A \rfloor$ is the number of nodes and the approximation consists of ignoring the integer requirement for the number of nodes. We again define Λ for this setup to be the set of the feasible node densities under minimum energy routing

$$\Lambda := \{\lambda : \lambda \bar{\mathcal{E}}_*(\lambda) \leq \kappa\}. \tag{7.40}$$

In the theorem below, we show the finiteness of λ_p^*. We note that this result is on lines of the results in [28] for a one dimensional GMRP.

Theorem 22 (Finite λ_*^p) *The optimal density λ_*^p in (7.37) exists whenever the set Λ in (7.40) is non-empty and λ_*^p is additionally finite if the per-node processing energy $C_p > 0$.*

Proof: Note that a compact set possesses a finite extremum point. To show that Λ is a compact set, we first note that Λ is a closed set. To show the bounded property, note that

$$\lambda\bar{\mathcal{E}}(\lambda) \sim \lambda[C_p + c\lambda^{-\frac{\nu}{2}}] \to \infty, \text{ as } \lambda \to \infty.$$

Hence, the constraint κ in (7.39) is violated as $\lambda \to \infty$ and hence, ∞ cannot be in Λ. Hence, Λ is closed and bounded, i.e., compact. This implies that $\lambda_*^p \in \Lambda$ is finite. □

In the above theorem, note that for the feasible set Λ to be non-empty, the constraint κ has to be sufficiently large. In other words, we need a large enough energy density κ to support energy consumption involved in processing and routing of measurements.

We observe that there is no threshold effect when there is a energy-density constraint and the optimal node density λ_*^p is always finite. This is in sharp contrast with the results in the previous section, where under the per-node energy constraint, λ_* can be unbounded depending on the regime of K. This is because here, the energy-density constraint κ limits the energy consumption in a given area thereby making infinite density infeasible. This implies that λ_*^p cannot be unbounded. In the theorem below, we provide bounds for λ_*^p when the variance ratio $K < 2$, based on the monotonicity of the error exponent in this regime.

Theorem 23 (Optimal λ_*^p) *Let $\lambda_i^p(\kappa)$ be the largest (positive real) root of the equation,*

$$\lambda C_p + \lambda^{1-\frac{\nu}{2}} c_i(\nu) - \kappa = 0, \quad \textit{for } i = 1, 2. \tag{7.41}$$

where $c_1(\nu)$:=$u\,\zeta(\nu;$ 1-$NNG) + \zeta(\nu; MST)$ and $c_2(\nu)$:=$\zeta(\nu; MST)$ are the energy-scaling

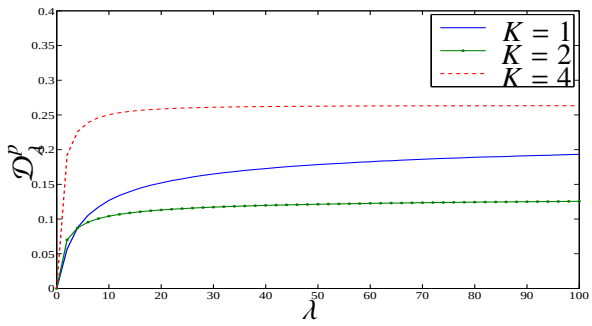


Figure 7.5: Error exponent vs. total energy. $\nu = 2, C_p = c_1(2), M = 0.6$.

constants. If the per-node processing energy $C_p > 0$, the variance ratio $K < 2$ and the constraint κ is such that both $\lambda_1^p(\kappa)$ and $\lambda_2^p(\kappa)$ exist then the optimal λ_*^p satisfies

$$\lambda_2^p(\kappa) \leq \lambda_*^p \leq \lambda_1^p(\kappa). \tag{7.42}$$

Proof: On the lines of the arguments in the previous section, it can be shown that $\mathcal{D}_\lambda$ is increasing in λ for $K < 2$. $\bar{\mathcal{E}}_*(\lambda)$ is decreasing in λ. Hence, the exponent $\mathcal{D}_\lambda^p$ in (7.36) is increasing in λ. From the energy-density constraint in (7.39) and Theorem 22, the feasible set Λ is bounded. From the bounds on $\bar{\mathcal{E}}_*(\lambda)$, if the largest real roots of (7.41), for $i = 1, 2$, exist, then the maximum value in Λ is bounded by these roots. □

Hence, in the above theorem, we obtain bounds on the optimal density λ_*^p under energy-density formulation. We prove this by first showing that the error exponent $\mathcal{D}_\lambda^p$ is increasing with λ when the variance ratio $K < 2$, as seen in Fig. 7.5. This implies that λ_*^p is the largest feasible density under constraint κ that supports the energy consumption for data fusion. In the end, we provide the bounds in (7.42) on this largest feasible density through the bounds for average energy consumption $\bar{\mathcal{E}}_\lambda$. Although we analytically prove the bounds in (7.42)

only when the variance ratio $K < 2$, the behavior of the error exponent $\mathcal{D}_\lambda^p$ in Fig.7.5 suggests that the bound in (7.42) may be valid for all values of K.

7.4 Conclusions

The tradeoff between the energy consumption in data fusion and the resulting detection performance at the fusion center is an important problem in the context of sensor networks. In this chapter, we incorporated correlation between the measurements through the Gauss-Markov random field model. We characterized the density of node deployment that maximizes the detection error exponent subject to a constraint on the average energy consumption. The measurement variance is crucial in determining whether the optimal node density is limited by the fusion energy constraint and displays a threshold behavior. We derived the threshold analytically and verified it with simulations.

7.A Proofs

Proof of Lemma 10

Since $\rho(0) = M$ and $\rho(\infty) = 0$, we have $x \in [0, M]$.

$$\frac{\partial f}{\partial x} = \frac{2x}{1-x^2}(-1 + \frac{2}{K(1-x^2)}).$$

Therefore, f has only one critical point in $(0, M]$. For $K < 2$, $\frac{\partial f}{\partial x} > 0$ and for $K > K_t'$, $\frac{\partial f}{\partial x} < 0$, $\forall x \in [0, M]$. There are no critical points. For $2 < K < K_t'$, the critical point is a minimum. Hence, maximum is attained at one of the boundary

points $\{0, M\}$. For $K < K_t$, it is at $x = M = \rho(0)$ and hence,

$$f(\rho(R_e), K) \leq f(\rho(0), K; \rho), \quad \forall\, R_e \geq 0.$$

Similarly, for $K > K_t$, we have

$$f(R_e, K; \rho) \leq f(\infty, K; \rho) = 0, \quad \forall\, R_e \geq 0.$$

□

Proof of Theorem 21

From Lemma 10, when $K < K_t$, $\rho(0) = M$ attains the maximum of $f(\rho(R_e), K)$. Hence, we have for $R_e \geq 0$,

$$\frac{1}{n} \sum_{e \in \mathcal{E}_d} f(\rho(R_e), K; \rho) \leq \frac{1}{n} \sum_{e \in \mathcal{E}_d} f(\rho(0), K; \rho), \quad \forall K < K_t.$$

Letting $n \to \infty$ on both sides, from (7.26)

$$\mathbb{E}[h(Z\lambda^{-0.5}, K; \rho)] \leq h(0, K, M), \quad \forall\, K < K_t.$$

Hence, the optimal density in this regime is given by

$$\lambda_*(K, \bar{E}) = \arg\max_{\lambda \geq \lambda_{\bar{E}}} \mathbb{E}[h(Z\lambda^{-0.5}, K; \rho)] = \infty \quad \forall\, K < K_t.$$

From Lemma 10, when $K > K_t$, $\rho(\infty) = 0$ attains the maximum of $f(\rho(R_e), K)$. For the case of infinite energy $\bar{E} \to \infty$, $\lambda_{\bar{E}} = \lambda_2(\bar{E}) = \lambda_1(\bar{E}) = 0$. Hence,

$$\lambda_*(K, \infty) = \arg\max_{\lambda \geq \lambda_{\bar{E}} = 0} \mathbb{E}[h(Z\lambda^{-0.5}, K; \rho)] = 0 \quad \forall\, K > K_t.$$

For finite constraint $\bar{E}$ and $K > K_t'$, $f(\rho(R_e), K)$ is increasing in R_e. We have $R_e = \lambda^{-0.5} R_e'$, where R_e' is the edge-length in unit area and is independent of

λ. Hence, $f(\rho(R_e), K)$ is non-increasing in λ and the limit $\mathbb{E}[h(Z\lambda^{-0.5}, K; \rho)]$ is also non-increasing in λ. Hence,

$$\lambda_*(K, \bar{E}) = \arg\max_{\lambda \geq \lambda_{\bar{E}}} \mathbb{E}[h(Z\lambda^{-0.5}, K; \rho)] = \lambda_{\bar{E}} \quad \forall\, K > K_t'.$$

□

CHAPTER 8

MEDIUM ACCESS DESIGN FOR STATISTICAL INFERENCE

The design of large wireless-sensor networks (WSN) must deal with challenges beyond the optimization of the local and the global decision rules, as is the case for classical distributed inference. Bandwidth has to be allocated to accommodate a large number of sensor nodes; transmissions must be made energy efficient to prolong network lifetime. Wireless transmissions make the medium-access control a crucial component. To this end, well-known deterministic scheduling schemes such as the time-division multiple access (TDMA) may not be appropriate; nodes may be sleeping, faulty, or placed in locations with poor transmission conditions. It is thus desirable to consider MAC schemes in the context of detection and estimation, that facilitate effective delivery of information from a random number of nodes to the fusion center.

We consider the number of reporting sensors to be random. This may arise in large-scale wireless sensor networks, where random access may be the preferred medium access, as it does not require any centralized scheduling. Examples of random access include the ALOHA scheme, where sensors decide to transmit based on a simple coin-flip. Alternatively sensors may undertake a more sophisticated scheme and decide to transmit only significant data. Another scenario is when the fusion center is a mobile-access point and travels to different geographic locations, with nodes dispersed according to a point process. In this chapter, we focus on the design of energy-optimal random-access schemes for distributed detection and estimation.

We consider in this chapter the problem of distributed detection over a wireless-fading channel via random access. We will not deal with the design

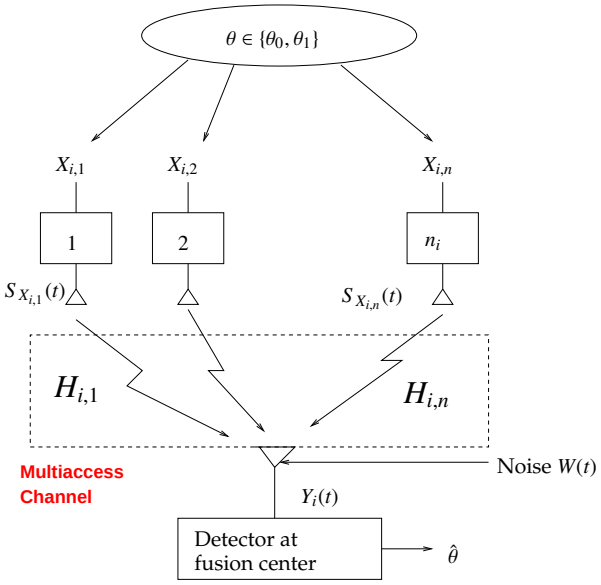


Figure 8.1: Distributed detection over multi-access fading channel.

of local quantization rules, which is a challenging problem even for the classical distributed detection. Our focus is on the communication (or the random access) aspect of the distributed detection, which to our knowledge has not been treated in the past.

We model the number of sensors involved in each transmission as random with a certain average transmission rate λ. There are several reasons to consider random access. The sensors may use a simple probabilistic wake-up strategy in which a sensor decides to participate in transmission based on a simple coin-flip. The sensor may also decide if a transmission is warranted according to its measurement, transmitting only when the data is "significant" [130]. Yet another possibility is that the fusion center is a mobile-access point, and it travels to different regions of the field to collect data, in which case the number of sensors involved in each collection is random.

A schematic of our problem is illustrated in Fig.8.1 with detailed model definition and assumptions given in section 8.1. The fusion center collects data in multiple slots, each involving a random number of transmitting sensors. We couple the so-called type-based multiple access (TBMA) [62, 64] with a simple random access protocol analogous to the ALOHA. Referred to as the type-based random access (TBRA), sensors transmit probabilistically using a set of orthogonal waveforms keyed to their measurements. Specifically, sensors with the same data value will transmit (if they decide to do so) using the same waveform on a multi-access fading channel. The bandwidth requirement of TBMA in the absence of fading, is proportional to the number of local quantization levels, not to the number of sensors. The use of orthogonal waveforms eliminates interference among users with different data values and makes it possible to have coherent combining of transmissions in the absence of fading. We will see, however, that simultaneous transmissions in fading is much more complicated, and it may not always be desirable.

8.0.1 Summary of main results

Given the fixed local quantization rule and the available set of orthogonal waveforms for transmission, the design of TBRA reduces to the optimal choice of the mean transmission rate λ. Intuitively, if λ is too small, not enough sensors transmit, and performance suffers. On the other hand, if too many sensors transmit, since they transmit on a multi-access channel, it is not obvious that the transmissions will not interfere with each other, resulting in poor detection performance.

In searching for the optimal transmission rate λ_*, we use the detection error exponent M_λ, a function of λ, to characterize performance. We first establish

that given the expected number of transmissions ρ in l collections, $\rho{:=}\lambda l$, the detection error probability P_e decays exponentially in the form

$$P_e = e^{-\rho M_\lambda + o(\rho)} \qquad (\rho \to \infty), \tag{8.1}$$

where $\frac{o(\rho)}{\rho} \to 0$ as $\rho \to \infty$.

The form of M_λ varies depending on the type of detectors (Bayesian or Neyman-Pearson) and the fading characteristics of the multi-access channel.

Next, we characterize the behavior of the error exponent M_λ for different cases. It turns out that M_λ crucially depends on the *coherence index* γ defined by

$$\gamma = \frac{|\mathbb{E}(H)|^2}{\mathrm{Cov}(H)}, \tag{8.2}$$

where H is the effective fading coefficient between a sensor and the fusion center.[1] Intuitively, higher γ leads to better SNR gain from simultaneous transmissions at the fusion center.

Illustrated in Fig.8.2 are sketches of error exponents as functions of λ and γ. The shapes of these curves will be justified by analytical and numerical results in section 8.2 and section 8.4. We see that for low coherence indices, there exists an optimal λ_* for which the error exponent is maximized. This implies that there is an optimal sensor-activation probability so that the average number of transmitting sensors is optimal. The intuition is that for fading channels with zero-mean ($\gamma = 0$), sensors transmitting simultaneously using the same waveform tend to cancel each other (in the mean), which is the reason that TBMA schemes involving a single data collection fail [61,62,64]. A sharp contrast is the extreme

[1] The dependencies of inference performance on the coherence index has been observed in the past [64,131].

case when the channel is deterministic without fading ($\gamma = \infty$). We show that there does not exist an optimal λ_*, which means that the optimal strategy is to have simultaneous transmissions, in order to take advantage of the channel coherency. This chapter aims to provide insights into the optimal tradeoff, for the case when the expected number of transmissions ρ goes to infinity.

We show the existence of an optimal average transmission rate λ_* when the channel-coherence index γ is small. We also provide the characterization of the error exponent when λ is large. It is in fact the behavior of M_λ as $\lambda \to \infty$ that helps us to describe the shape of error exponent curve in Fig.8.2. By letting $\lambda \to \infty$, we employ a version of the central limit theorem (CLT) involving a random number of summands. The limiting distribution allows us to characterize M_∞ analytically. For large transmission rates λ, Gaussian approximation can be used to obtain estimates of the error exponent. Perhaps more importantly in practice, the Gaussian approximation provides $\tilde{\lambda}_*$, an approximation to the optimal rate λ_*.

Our numerical evaluation and simulations are also informative. We numerically evaluate M_λ under different conditions to confirm our theory. We present a performance comparison between TBRA and TDMA, under a fixed energy constraint. The simulation confirms the analysis and our intuition: the two schemes have different operation regimes for the zero-mean ($\gamma = 0$) multi-access channels (if complexity is not part of the consideration). At low SNR, TBRA performs considerably better than TDMA because of its optimal allocation of transmissions over time and across sensors, to obtain a significant SNR gain. At high SNR, on the other hand, SNR gain is not needed and the deterministic scheduling of TDMA shows an advantage as it avoids the possibility of inter-

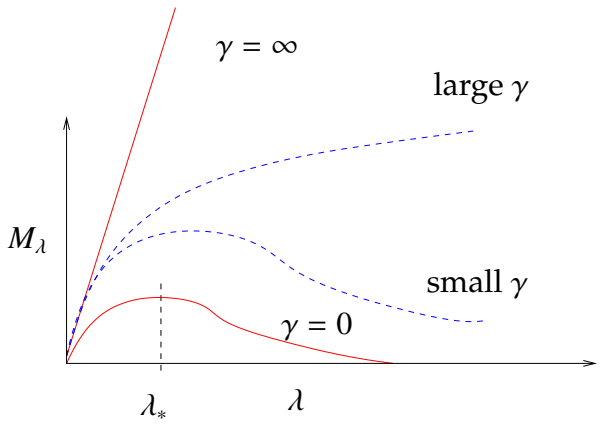


Figure 8.2: Performance M_λ under transmission rate λ and coherence index γ.

fering transmissions due to random access. On the other hand, under large-γ regime, TBRA performs better than TDMA for a wide range of SNR values, by exploiting channel coherency.

In section 8.1, we explain the system model in detail and give the problem statement. In section 8.2, we explain the receiver structure and present the expressions of the error exponent with respect to the expected number of transmissions. Both Neyman-Pearson and Bayesian detectors are considered. We present a characterization of asymptotic behavior of error exponents when the average number of transmissions goes to infinity. We also discuss the use of Gaussian approximation to investigate behaviors of error exponents. In section 8.4, we provide numerical evaluations of error exponents and simulation results on the detection error probability where we compare TBRA with deterministic TDMA scheduling. Conclusions and comments are made in section 8.5.

8.1 Model and problem formulation

8.1.1 TBRA: Transceiver and sufficient statistics

We assume the global statistical model as simple binary hypotheses

$$\mathcal{H}_0 : \theta = \theta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \theta = \theta_1.$$

As illustrated in Fig.8.1, the fusion center collects data in multiple time slots indexed by i. In each collection, there are N_i sensors involved in the transmission, where N_i is a random variable with mean λ and probability-mass function (PMF) $g(n, \lambda) := \Pr(N_i = n)$. We assume that the sequence N_i is IID.

In the i^{th} data collection, a sensor involved in the transmission[2], say sensor k, has measurement $X_{i,k} \in \{1, \cdots, K\}$ i.e., quantized to K levels. We assume that the sensor data $\{X_{i,k}\}$ are conditionally IID across time and sensors, given θ, with PMF $p_\theta(\cdot)$. In vector notation we have,

$$X_{i,k} \overset{\text{i.i.d}}{\sim} \mathbf{p}_\theta = (p_\theta(1), \cdots, p_\theta(K)), \quad \theta \in \{\theta_0, \theta_1\}.$$

In the i^{th} collection, sensor k encodes its measurement $X_{i,k}$ to a certain waveform and transmits it over a multi-access fading channel. As in TBMA, a set of K orthonormal waveforms $\{\phi_m(t), m = 1, \cdots, K\}$ are used, each corresponding to a specific data value. Specifically, the baseband signal transmitted by sensor k in collection i is given by

$$S_{i,k}(t) = \sqrt{\mathcal{E}}\phi_{X_{i,k}}(t),$$

where $\mathcal{E}$ is the energy of the transmission.

[2]Without loss of generality, we will only consider those sensors involved in the transmission.

The channel coefficients ($\tilde{H}_{i,k} \in \mathbb{C}$) are time-varying, IID across sensors and time. We assume coarse synchronization in the sense that at the fusion center, there is no inter-collection interference. Such synchronization can be derived by letting fusion center transmit a synchronization beacon. It can also be accomplished by adding sufficient guard time between consecutive data collections. For low rate applications, this assumption is reasonable.

The received complex-baseband signal after collecting l samples is given by

$$Y_i(t) = \sum_{k=1}^{N_i} \tilde{H}_{i,k} S_{i,k}(t - \tau_{i,k}) + W_i(t), \quad i = 1, \cdots, l, \tag{8.3}$$

where we assume that the channel-state information $\{\tilde{H}_{i,k}\}$ is not known at the receiver. $\tau_{i,k}$ are the random delays for different sensor transmissions and the noise $W_i(t)$ is assumed to be complex white zero-mean Gaussian, with power density σ^2. We define the sensor signal-to-noise ratio by SNR:$=\frac{\mathcal{E}}{\sigma^2}$.

Under the narrow-band signal assumption, the flat-fading approximation which neglects the time dispersion in the signal is valid. Therefore, the delay is only through the carrier phase i.e., $S_{i,k}(t - \tau_{i,k}) \approx S_{i,k}(t) \exp(-j2\pi f_c \tau_{i,k})$, where f_c is the carrier frequency. Denoting the effective fading statistic by $H_{i,k}$:$=\tilde{H}_{i,k} \exp(-j2\pi f_c \tau_{i,k})$ with mean μ_H:$=\mathbb{E}(H_{i,k})$ and covariance σ_H^2:$=\text{Cov}(H_{i,k})$, the received signal is thus given by

$$Y_i(t) = \sum_{k=1}^{N_i} H_{i,k} S_{i,k}(t) + W_i(t), \quad i = 1, \cdots, l, \tag{8.4}$$

where we assume that $\{H_{i,k}\}$ are proper complex Gaussian, and are unknown at the fusion center.

Sufficient statistics $\{\mathbf{Y}_i\}$ are generated from the bank of filters matched to the orthogonal basis $\{\phi_m(t)\}$. For the i^{th} collection,

$$\begin{aligned}\mathbf{Y}_i &:= \frac{1}{\sqrt{\mathcal{E}}}\Big[\langle Y_i(\cdot), \phi_1(\cdot)\rangle, \cdots, \langle Y_i(\cdot), \phi_K(\cdot)\rangle\Big] \\ &= \sum_{k=1}^{N_i} H_{i,k}\mathbf{e}_{X_{i,k}} + \mathbf{W}_i, \end{aligned} \tag{8.5}$$

where $\langle Y_i(\cdot), \phi_m(\cdot)\rangle$ is the output of the matched filter corresponding to $\phi_m(t)$, $\mathbf{e}_m$ the unit vector with non-zero entry at the m^{th} position, and $\mathbf{W}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}_c(0, \frac{1}{\text{SNR}}\mathbf{I})$.

To see the intuition behind the coherence index γ defined in (8.2), we explicitly write the m^{th} entry of $\mathbf{Y}_i = [Y_{i,1}, \cdots, Y_{i,K}]^{\text{T}}$

$$Y_{i,m} = \sum_{k=1}^{N_i} H_{i,k} 1_{\{X_{i,k}=m\}} + W_{i,m}, \tag{8.6}$$

where $1_{\mathcal{A}}$ is the event-indicator function. The extreme case is when the channel is deterministic with $H_{i,k} \equiv 1$ ($\gamma \to \infty$). Transmissions from those sensors observing data value m add up coherently, and $Y_{i,m}$ is the number of sensors that observe data level m (plus noise), which gives rise to notion of type-based transmission[3]. On the other hand, when $\gamma = 0$, ($\mu_H = 0$), the transmissions add up non-coherently, and the mean of $\mathbf{Y}_i$ contains no information of the model.

[3]Given $X_{i,k} = x_{i,k}$, $N_i = n_i$ and the observation $\mathbf{Y}_i = \mathbf{y}_i$, in the absence of noise, the type of $x_{i,k}$ is $\frac{1}{n_i}\mathbf{y}_i$. [108, 132].

8.1.2 Spatio-temporal tradeoff and problem formulation

The design of TBRA reduces to finding the optimal-activation strategy that minimizes detection-error probability. For sensors that are activated probabilistically either by themselves or by the beacon from the fusion center, a TBRA scheme reduces to finding the mean number of transmissions $\lambda = \mathbb{E}(N_i)$. To this end, we need to connect λ with the detection error probability.

If the fusion center collects data using TBRA for l time slots, the expected number of transmissions is $\rho{:=}\lambda l$, which is also proportional to the total energy consumption. Fixing ρ, there is a spatio-temporal tradeoff between the average number of transmissions per slot and the total number of time slots. Should energy be allocated mostly to simultaneous transmissions by making λ large? Or should we rely on taking more data collections by choosing a large l. The optimal design of TBRA is to achieve optimal tradeoff between λ and ł.

We will consider two types of detector: the Bayesian detector and the Neyman-Pearson detector. The explicit characterizations of error probabilities for these two cases are not tractable. We thus examine the case when the expected number of transmissions ρ is large. Let $P_e(\rho, \lambda)$ be the detection-error probability (either the miss-detection probability of the Neyman-Pearson detector or the average of the miss detection and the false-alarm probabilities in the Bayesian setup). We will optimize TBRA through the error exponent

$$M_\lambda{:=} -\lim_{\rho\to\infty} \frac{1}{\rho} \log P_e(\rho, \lambda), \tag{8.7}$$

which is equivalent to say that $P_e(\rho, \lambda)$ decays exponentially with respect to ρ

with rate M_λ, a function of λ, as in (8.1). The justification of exponential decay of P_e will be given in section 8.2 using standard arguments involving the Cramér's theorem and the Stein's lemma. Next, we optimize TBRA by seeking

$$\lambda_* = \arg\sup_{\lambda>0} M_\lambda. \tag{8.8}$$

Although M_λ can be evaluated numerically for a given statistical model of hypotheses and fading, it is of theoretical and practical significance to establish that λ_* is finite and bounded. To this end, we need to characterize M_λ as $\lambda \to 0$ and $\lambda \to \infty$.

8.2 Optimal type-based random access

The key step towards optimal TBRA is the characterization of detection error exponent defined in (8.7). The form of error exponent is well known in the theory of large-deviation analysis [132]: the Chernoff information for the Bayesian detector, and the Kullback-Leibler (KL) distance (relative entropy) for the Neyman-Pearson detector. We first present the optimal detector and then give the general characterizations of the error exponents with respect to the expected number of transmissions ρ. Next, we state a result on the existence of optimal λ_* that maximizes the error exponent. We then consider the limiting case when $\lambda \to \infty$. The asymptotic analysis not only gives the key argument for the existence of a finite optimal λ_*, but also provides a qualitative assessment of the error exponents and a computationally tractable way of estimating λ_*.

8.2.1 Optimal non-coherent detector

The optimal detector given the matched filter output $\{\mathbf{Y}_i\}$ is the likelihood-ratio detector under both the Bayesian and Neyman-Pearson settings. With the IID assumption, the detector is given by

$$T_l = \frac{1}{l}\sum_{i=1}^{l} \log \frac{f_{1,\lambda}(\mathbf{Y}_i)}{f_{0,\lambda}(\mathbf{Y}_i)} \gtrless \tau. \tag{8.9}$$

where $f_{k,\lambda}(\mathbf{y})$ is the PDF[4] of $\mathbf{Y}_i$ under hypothesis $\mathcal{H}_k$, l is the number of data collections and the threshold τ is chosen according to the prior for the Bayesian detector or the false-alarm rate for the Neyman-Pearson detector.

The receiver only needs to compute the likelihood ratio in (8.9). In practice, the likelihood ratio may not have a closed-form expression[5], and numerical evaluation is necessary. However, since the receiver is non-coherent, it does not have a RAKE structure. Therefore, the complexity of the receiver is not limited by the number of simultaneous transmissions.

8.2.2 Detection-error exponents

In the following theorem, we give the expressions for the detection-error exponents. These are direct applications of the Cramér's theorem and the Stein's lemma with only a trivial modification that changes time index in the standard setting to the expected number of transmissions ρ.

[4]We will assume that the likelihood function is well defined.

[5]For the special case of Poisson number of sensors with Gaussian fading, an infinite-sum expression is available.

Theorem 24 (Error exponents) *Given expected number of transmissions ρ and mean transmission rate λ, let $P_e^B(\rho,\lambda)$ be the average error probability of the Bayesian detector under any prior, and $P_e^{NP}(\rho,\lambda)$ be the miss detection-error probability of the Neyman-Pearson detector under any fixed size α. The error exponents for the two detectors are given by*

$$M^{NP}(\lambda) := -\lim_{\rho\to\infty}\frac{1}{\rho}\log P_e^{NP}(\rho,\lambda) = \frac{1}{\lambda}D_\lambda(f_0\|f_1), \tag{8.10}$$

$$M^{B}(\lambda) := -\lim_{\rho\to\infty}\frac{1}{\rho}\log P_e^{B}(\rho,\lambda) = \frac{1}{\lambda}C_\lambda(f_0,f_1), \tag{8.11}$$

where $D_\lambda(f_0\|f_1)$ is the Kullback-Leibler distance and $C_\lambda(f_0,f_1)$ the Chernoff information.

Proof: See [108, 132, p. 92-94] . □

While the above theorem provides the basis for investigating error exponents, it says little about the behavior of error exponents as functions of λ, especially about whether there exists an optimal λ_*. The following theorem gives the results for the two extreme cases: $\gamma = 0$ and $\gamma = \infty$.

Theorem 25 (Existence of optimal λ) *Let λ be the mean transmission rate and let $f_{i,\lambda}(\mathbf{y})$ be the PDF of the matched-filter output $\mathbf{Y}$, under the hypothesis $\mathcal{H}_i$. Assume the following:*

1. *for the Neyman-Pearson detection, PDF $f_{0,\lambda}$ and $f_{1,\lambda}$ are differentiable functions of λ almost everywhere,*
2. *for the Bayesian detection, the above assumption and in addition, the optimizing parameter ν_* is differentiable in λ almost everywhere, given by,*

$$\nu_*(\lambda) = \arg\min_{\nu\in[0,1]}\log\int_{\mathbf{y}} f_{0,\lambda}^{\nu}(\mathbf{y})f_{1,\lambda}^{1-\nu}(\mathbf{y})\mathbf{dy}. \tag{8.12}$$

In addition, assume that the PMF of N, $g(n,\lambda)$ is differentiable in λ and satisfies the following properties,

$$\lim_{\lambda\to 0} g(n,\lambda) = 1_{\{n=0\}}, \tag{8.13}$$

$$\lim_{\lambda\to 0} \frac{d}{d\lambda} g(n,\lambda) = -a1_{\{n=0\}} + a1_{\{n=1\}},\ a>0, \tag{8.14}$$

where $1_{\mathcal{A}}$ is the event-indicator function. The following results hold:

1. *if the channel has zero-mean fading, i.e., $\gamma = 0$, then*

$$\lim_{\lambda\to 0} M^{NP}(\lambda) = \lim_{\lambda\to\infty} M^{NP}(\lambda) = 0, \tag{8.15}$$

$$\lim_{\lambda\to 0} M^{B}(\lambda) = \lim_{\lambda\to\infty} M^{B}(\lambda) = 0, \tag{8.16}$$

which imply that there exist $\lambda_*^{NP}, \lambda_*^{B}$ such that $0 < \lambda_*^{NP}, \lambda_*^{B} < \infty$ and

$$\sup_{\lambda>0} M^{NP}(\lambda) = \frac{1}{\lambda_*^{NP}} D(f_{0,\lambda_*^{NP}} \| f_{1,\lambda_*^{NP}}), \tag{8.17}$$

$$\sup_{\lambda>0} M^{B}(\lambda) = \frac{1}{\lambda_*^{B}} C(f_{0,\lambda_*^{B}} \| f_{1,\lambda_*^{B}}), \tag{8.18}$$

2. *if the channel is deterministic i.e., $Cov(H) = 0$ or $\gamma = \infty$, then there does not exist a bounded optimal λ_* that maximizes the error exponents. In particular,*

$$M^{NP}(\lambda) = \Theta(\lambda), \quad M^{B}(\lambda) = \Theta(\lambda), \quad as \quad \lambda\to\infty, \tag{8.19}$$

where the notation Θ means that λ is an exponentially-tight bound[6]

Proof: The proof for the above theorem relies on the analysis of the extreme cases under the regularity assumptions of the KL distance and the Chernoff information. See Appendix 8.A. □

[6] $\Theta(a(\lambda)) = \{b(\lambda) : 0 \le c_1 a(\lambda) \le b(\lambda) \le c_2 a(\lambda), \forall \lambda > \lambda_o\}$ for some $c_1, c_2, \lambda_o > 0$.

In the above theorem, the assumptions on the PMF of N (8.13) and (8.14) imply that at low λ, there is utmost one transmission. Examples include Poisson distribution and the binomial distribution $B(n, p)$ for a fixed n with $p \to 0$.

The continuity assumption for the Neyman-Pearson detection is easily satisfied for many well-behaved distributions. On the other hand, the assumption for Bayesian detection in (8.12) is harder to satisfy in practice.

Theorem 25 establishes the general shape of M_λ as shown in Fig.8.2, for the extreme values of coherence index, $\gamma = 0$ and $\gamma = \infty$. The role of γ in M_λ is embedded in the KL distance or the Chernoff information through the PDFs $(f_{i,\lambda}(\mathbf{y}), \quad i = 0, 1)$, which are typically continuous functions of γ. Therefore, one can infer the behavior of M_λ for small and large γ.

8.2.3 Asymptotic distribution and Gaussian approximation

A key step in proving theorem 25 is the investigation of M_λ as $\lambda \to \infty$. The idea is to use the continuity argument coupled with a version of the central limit theorem (CLT) to calculate the KL distance and the Chernoff information. We elaborate this approach here for two reasons. First, the calculation of M_∞ is needed in proving theorem 25; this is accomplished by the use of CLT. Second, from a practical stand-point, the Gaussian approximation via CLT gives a computationally tractable way to approximate the error exponent. This is especially useful when we try to find the optimal sensor activation rate λ_*. The accuracy of such an approximation of course, depends on the specific distributions of the sensor measurements and the channel, and we will demonstrate its performance in section 8.4.

We shall focus in this section on the single collection model, and evaluate the error exponents using the limiting distribution as $\lambda \to \infty$. For ease of notation, we drop the time index i in (8.5), and consider the model

$$\mathbf{Y} = \sum_{k=1}^{N} H_k \mathbf{e}_{X_k} + \mathbf{W}, \tag{8.20}$$

where we have a random summand N, with PMF $g(n, \lambda)$ and mean $\mathbb{E}(N) = \lambda$.

Theorem 26 (Limiting distribution of Y) *Assume that the effective channel gains $\{H_k\}$ are IID distributed with mean μ_H and covariance σ_H^2, and that the number of sensors N has PMF $g(n, \lambda)$, with mean λ. Also, assume that N/λ converges to a constant $\eta > 0$ in distribution, i.e.,*

$$\frac{N}{\lambda} \xrightarrow{d} \eta, \quad \text{as } \lambda \to \infty. \tag{8.21}$$

When the PMF of sensor measurements $p_\theta(m) > 0$, $\forall\ m = \{1, \ldots, K\}$ and $\theta = \{\theta_0, \theta_1\}$, the shifted and scaled matched-filter output $\mathbf{Y}$ satisfies the central limit theorem and has the limiting complex-normal distribution according to

$$\frac{\mathbf{Y} - \eta\lambda\mu_H \mathbf{p}_\theta}{\sqrt{\eta\lambda}} \xrightarrow{d} \mathcal{N}_c(0, Cov(H_1 \mathbf{e}_{X_1})) \quad as \quad \lambda \to \infty. \tag{8.22}$$

In addition, if N is Poisson, then each entry of vector $\mathbf{Y}$ is independently asymptotically Gaussian, given by

$$\frac{Y(m) - \lambda\mu_H p_\theta(m)}{\sigma_H \sqrt{\lambda p_\theta(m)}} \xrightarrow{d} \mathcal{N}_c(0,1), \quad as \quad \lambda \to \infty, \tag{8.23}$$

$\forall\ m = 1, \ldots, K.$

Proof : The proof of the above is an application of CLT with a random summand. We use the fact that $\eta = 1$ for Poisson. See Appendix 8.A for details. □

Evaluating the covariance matrix in (8.22), we have

$$\begin{aligned} \mathrm{Cov}(H_1 \mathbf{e}_{X_1}) &= \sigma_H^2 \mathrm{Diag}(\mathbf{p}_\theta) \\ &\quad + |\mu_H|^2 (\mathrm{Diag}(\mathbf{p}_\theta) - \mathbf{p}_\theta \mathbf{p}_\theta^T). \end{aligned} \tag{8.24}$$

However, the result for the Poisson distribution in (8.23) is stronger than for a general PMF $g(n, \lambda)$ in (8.22), since the asymptotic distribution is independent across the quantization levels. This is due to the property of marking, which implies that the number of sensors transmitting a particular data-level is independently Poisson. Moreover, by (8.24), under zero-mean fading ($\mu_H = 0$), the asymptotic distribution is independent across the data-levels for any general PMF $g(n, \lambda)$. In this section, we assume that N is Poisson, for the ease of evaluation of the exponents under the limiting distribution.

Since $\mathbf{Y}$ is asymptotically Gaussian, in the large-λ regime, the hypothesis-testing problem can be cast as the testing of binary hypotheses with

$$\tilde{\mathcal{H}}_i \quad : \quad \mathbf{Y} \sim \mathcal{N}_c(\mu_i, \Sigma_i), i = 0, 1 \tag{8.25}$$

$$\mu_i = \lambda\mu_H \mathbf{p}_{\theta_i}, \ \ \Sigma_i = \lambda\sigma_H^2 \mathrm{Diag}(\mathbf{p}_{\theta_i}) + \frac{\sigma^2}{\mathcal{E}}\mathbf{I}. \tag{8.26}$$

The effect of the channel coherency is more evident in the asymptotic distribution. For any positive γ ($\mu_H > 0$), the mean of $\mathbf{Y}$ in (8.26) contains information about the underlying hypothesis $\mathcal{H}_i$. As λ increases, the two hypotheses are more separated and the error probability decays exponentially. When $\mu_H = 0$ ($\gamma = 0$), on the other hand, the information is in the covariance, and the error probability for a single-data collection converges to a constant value as $\lambda \to \infty$ [64]. However, as mentioned in section 8.2.2, in this chapter, we let the expected number of transmissions ρ go to infinity. In this case, there is exponential decay of error probability, and we therefore need to characterize the exponents.

We now characterize the error exponents, defined by (8.10) and (8.11), for the Gaussian distribution. It turns out that the Gaussian Chernoff-information $\tilde{C}$ and the Kullback-Leibler distance $\tilde{D}$ have closed-form expressions, enabling us to evaluate the asymptotic limits of the true exponents.

Lemma 12 (Gaussian error exponents) *Let σ_H^2 be the channel variance, γ the channel-coherence index, SNR be the sensor SNR and $\Delta{:=}p_{\theta_0}(k) - p_{\theta_1}(k)$. Denote,*

$$\alpha_k \;:=\; \frac{\lambda\sigma_H^2\, SNR\, p_{\theta_0}(k) + 1}{\lambda\sigma_H^2\, SNR\, p_{\theta_1}(k) + 1}, \;\; \omega_k{:=}\lambda\gamma\Delta \tag{8.27}$$

and let β_k be the positive root of the quadratic equation[7]*,*

$$\omega_k\alpha_k\beta_k^2 + (\alpha_k - 1)\beta_k - \log\alpha_k - \omega_k = 0. \tag{8.28}$$

[7]Exactly one positive root exists for the case when $p_{\theta_0}(k) \neq p_{\theta_1}(k)$.

The error exponents $\tilde{E}^{NPD}(\lambda)$ and $\tilde{E}^{BD}(\lambda)$ under the Neyman-Pearson and the Bayesian settings, for the Gaussian distribution in (8.25), are given by

$$\begin{aligned}\tilde{E}^{NPD}(\lambda) = \frac{\tilde{D}_{\lambda,\gamma}}{\lambda} &= \frac{1}{\lambda}\sum_{k=1}^{K}\Big(-\log\alpha_k \\ &\quad + (1+\omega_k)(\alpha_k - 1)\Big), \qquad (8.29)\\ \tilde{E}^{BD}(\lambda) = \frac{\tilde{C}_{\lambda,\gamma}}{\lambda} &= \frac{1}{\lambda}\sum_{k=1}^{K}\Big(-\log\beta_k + \beta_k \\ &\quad + \frac{(\beta_k-1)^2\alpha_k\omega_k}{(\alpha_k-1)} - 1\Big). \qquad (8.30)\end{aligned}$$

Proof : The proof is derived using Kullback-Leibler distances for Gaussian distributions. See Appendix 8.A. □

Given the closed-form expressions for the Gaussian error exponents $\tilde{E}^{\text{NPD}}(\lambda)$ and $\tilde{E}^{\text{BD}}(\lambda)$, we can evaluate the various limits.

Theorem 27 (Limiting properties of error exponents) *The Chernoff information $\tilde{C}_{\lambda,\gamma}$ and the KL distance $\tilde{D}_{\lambda,\gamma}$ are monotonically increasing functions of the coherence index γ, transmission rate λ and sensor SNR. For finite γ, the error exponents $\tilde{E}^{NPD}(\lambda)$ and $\tilde{E}^{BD}(\lambda)$ converge to a finite limit, proportional to the coherence index γ, as $\lambda \to \infty$ given by*

$$\begin{aligned}\lim_{\lambda\to\infty}\tilde{E}^{NPD}(\lambda) &= \gamma\sum_{k=1}^{K}\frac{\Delta}{p_{\theta_1}(k)}, \qquad (8.31)\\ \lim_{\lambda\to\infty}\tilde{E}^{BD}(\lambda) &= \gamma\sum_{k=1}^{K}\big(\sqrt{p_{\theta_0}(k)} - \sqrt{p_{\theta_1}(k)}\big)^2. \qquad (8.32)\end{aligned}$$

Proof : We establish the monotonicity and evaluate the limits using expressions in (8.29) and (8.30). Also note, (8.32) is a scaled version of Hellinger distance [133]. See Appendix 8.A. □

We also investigate the case when the channel is perfectly coherent: $\mu_H = 1$ and $\sigma_H \to 0$, or $\gamma \to \infty$.

Theorem 28 (Error exponent for perfectly coherent channels) *In the absence of fading, the error exponents of the NP and Bayesian detectors for the limiting distribution are given by*

$$\lim_{\gamma\to\infty} \tilde{E}^{NPD}(\lambda) = \lambda SNR \sum_{k=1}^{K} \Delta^2 \tag{8.33}$$

$$\lim_{\gamma\to\infty} \tilde{E}^{BD}(\lambda) = \frac{\lambda SNR}{4} \sum_{k=1}^{K} \Delta^2 \tag{8.34}$$

Proof : Substituting $\sigma_H = 0$, we derive the expressions by finding the KL distance and Chernoff information between the distributions, $\mathcal{N}_c(\lambda \mathbf{p}_{\theta_0}, \sigma^2)$ and $\mathcal{N}_c(\lambda \mathbf{p}_{\theta_1}, \sigma^2)$. □

To contrast the perfectly coherent case, we examine the case when the channel is non-coherent, i.e., $\mu_H = 0$ ($\gamma = 0$). Interestingly, the dependencies of the Chernoff information and the KL distance on the transmission rate λ, the sensor SNR, and the channel variance σ_H^2 can be summarized using a single parameter—the average SNR at the receiver,

$$\chi := \lambda \sigma_H^2 \text{SNR}. \tag{8.35}$$

Theorem 29 (Error exponents for non-coherent channels) *For the non-coherent channels ($\mu_H = 0$), the Chernoff information $\tilde{C}_\chi$ and the KL distance $\tilde{D}_\chi$ are functions of χ:=$\lambda\sigma_H^2 SNR$, and have the following properties:*

1. *$\tilde{C}_\chi$ and $\tilde{D}_\chi$ are monotonically-increasing concave functions of χ.*
2. *As $\chi \to \infty$, $\tilde{C}_\chi$ and $\tilde{D}_\chi$ converge to finite limits when $p_{\theta_i}(k) > 0 \ \forall i, k$.*
3. *Normalized functions $\frac{\tilde{C}_\chi}{\chi}$ and $\frac{\tilde{D}_\chi}{\chi}$ have unique maxima, which are only functions of $\mathbf{p}_{\theta_0}$ and $\mathbf{p}_{\theta_1}$.*

Proof : See Appendix 8.A. □

The compact parameter χ provides additional insights and design options. The optimal error exponents can be achieved with a combination of choices of sensor activation and sensor SNR. At small SNR, for example, more sensors are needed to obtain SNR gain. On the other hand, at high SNR, fewer sensors transmit to avoid the non-coherent cancelation of the signals (on an average). The optimal λ_* is chosen so as to balance these opposing effects of the multi-access fading channel.

The error exponents $\tilde{E}^{\text{NPD}}(\lambda)$ and $\tilde{E}^{\text{BD}}(\lambda)$ in (8.29) and (8.30) of the limiting Gaussian distributions are good approximations for the true exponents at large λ, due to the continuity property. We shall demonstrate this with a numerical example below, which is a qualitative representation for the general case.

8.3 Optimal TBRA for Parameter Estimation

8.3.1 Bayesian Cramér-Rao Bound

We define the performance metric for estimation as the normalized Bayesian Cramér-Rao lower bound (BCRB) [134]. Given expected number of transmissions ρ and mean transmission rate λ per data collection, let $\hat{\Theta}$ be a Bayesian estimator. Under some regularity conditions [134, p. 72], we have

$$\mathbb{E}(\hat{\Theta} - \Theta)^2 \geq \frac{1}{\frac{\rho}{\lambda}\mathbb{E}[I_\lambda(\Theta)] + A_\pi}, \tag{8.36}$$

with equality iff conditional PDF of $\mathbf{U}$, $f_U(\theta|\mathbf{u}^l)$, is Gaussian; and $I_\lambda(\theta)$ is the Fisher information of a single data collection of the sufficient statistic $\mathbf{U}$, for a given θ and A_π only depends on the PDF of Θ i.e., $\pi(\cdot)$. To obtain design guidelines, we define the normalized expected Fisher information, given by

$$M_\lambda := \frac{\mathbb{E}[I_\lambda(\Theta)]}{\lambda}, \tag{8.37}$$

where the expectation is taken over Θ. Maximizing the normalized Bayesian information with respect to λ, gives the least BCRB. In general, the BCRB is not achieved by the MMSE estimator. Note if we instead formulate θ as a deterministic parameter, then the optimal TBRA scheme would depend on θ. In addition to the regularity conditions for the existence of BCRB, we assume that the PDF $f_U(\mathbf{u}|\theta;\lambda)$ is differentiable up to second order ($\mathbb{C}^2$) in $\mathbf{y}, \theta$ and λ.

Having defined the performance metric for estimation, the design of optimal

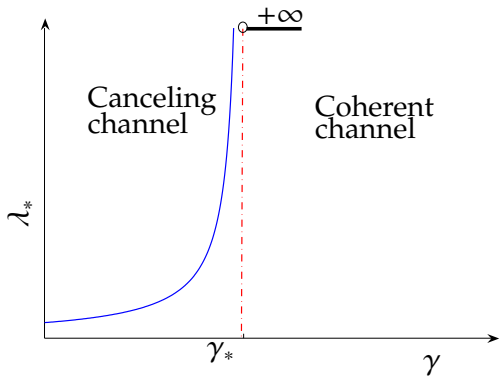


Figure 8.3: Optimal transmission rate λ_* under coherence index γ.

TBRA now reduces to finding an optimal transmission rate as before, with mean number of transmissions ρ fixed,

$$\lambda_* := \arg \sup_{\lambda \in \Re^+} M_\lambda. \tag{8.38}$$

The results of Theorems 25-29 also hold for estimation. See Appendix 8.A for details. We now prove an additional result on the presence of a critical coherence index which holds for parameter estimation.

8.3.2 Critical Coherence Index γ_*

In theorem 25, we have characterized the behavior of the metric $M_{\lambda,\gamma}$ and thereby the optimal transmission rate $\lambda_*(\gamma)$, for extreme values of the coherence index i.e., $(\gamma = 0)$ and $(\gamma = \infty)$. For finite positive γ, we expect smooth transition between these extreme behaviors, especially for well-behaved distributions. To study the nature of λ_*, it is crucial to characterize the slope of M_λ, since a negative slope at large-λ implies that λ_* is bounded. However, we can only numerically evaluate M_λ for finite λ.

If we impose an additional regularity condition that conditional PDF $f_\lambda(\mathbf{y}|\theta)$ is continuously differentiable to second order, then the partial derivatives up to the second derivative are continuous [135]. Therefore,

$$\frac{\partial}{\partial\lambda}M_\lambda \to \frac{\partial}{\partial\lambda}\tilde{M}_\lambda, \quad \text{as} \quad \lambda \to \infty. \tag{8.39}$$

This condition is satisfied by well-behaved distributions. For the Poisson-Gaussian distribution, we can express the conditional PDF $f_\lambda(\mathbf{y}|\theta)$ as an infinite sum. On evaluating the limits, we find that it satisfies (8.39).

Therefore, at large-λ, we can reasonably approximate the slope of the actual metric by the slope of the Gaussian metric i.e.,

$$\frac{\partial}{\partial\lambda}M_\lambda \approx \frac{\partial}{\partial\lambda}\tilde{M}_\lambda. \tag{8.40}$$

Rewriting the Gaussian performance metric,

$$\tilde{M}_\lambda = 2\lambda\,\mathrm{SNR}\,\sigma_H^2\mathbb{E}\Big[\sum_{k=1}^{K}\frac{\gamma\,p'_\Theta(k)^2}{\lambda\sigma_H^2\,\mathrm{SNR}\,p_\Theta(k)+1} + \sum_{k=1}^{K}\frac{\sigma_H^2\mathrm{SNR}\,p'_\Theta(k)^2}{(\lambda\sigma_H^2\,\mathrm{SNR}\,p_\Theta(k)+1)^2}\Big], \tag{8.41}$$

we note that the two terms signify the opposing effects of coherence and cancelation respectively. This is because at large values of λ, the first terms approaches a constant, proportional to γ; whereas the second term decays to zero. Moreover, for all values of λ, the first term is increasing in λ and the second term

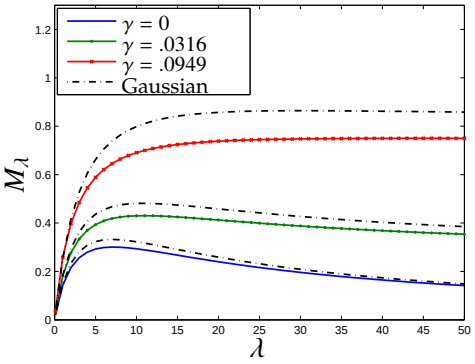


Figure 8.4: Performance metric (SNR= -5dB, $\sigma_H^2 = 1$, $\Theta \sim \Delta(0.2, 0.8)$).

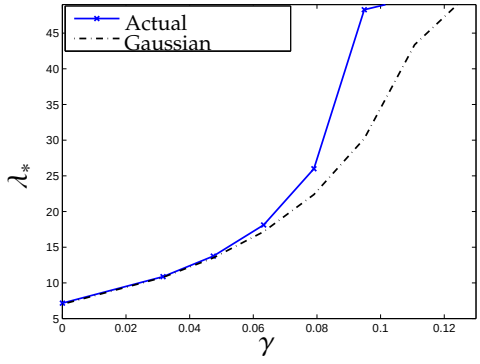


Figure 8.5: Gaussian-approximation (SNR= 0dB, $\sigma_H^2 = 1$, $\Theta \sim \Delta(0.2, 0.8)$).

is decreasing. Hence, if the first term dominates to such an extent that $\tilde{M}_\lambda$ is always increasing in λ, then the optimal λ_* is infinite. If the first term dominates for some value γ_*, then it dominates for all $\gamma > \gamma_*$. In the following theorem, we establish such a critical coherence index γ_*; signifying transition between these opposing effects.

Theorem 30 *For the Gaussian metric $\tilde{M}_{\lambda,\gamma}$ given by (8.41), suppose the optimal transmission rate $\tilde{\lambda}_*(\gamma)$ is given by*

$$\tilde{\lambda}_*(\gamma) := \arg \sup_{\lambda \in \Re^+} \tilde{M}_{\lambda,\gamma}. \tag{8.42}$$

Then there exists a critical coherence index γ_* such that

$$\tilde{\lambda}_*(\gamma) = \infty \,, \forall \gamma > \gamma_*. \tag{8.43}$$

Additionally for $\gamma < \gamma_*$, the metric $\tilde{M}_\lambda$ is unimodal.

The critical coherence index γ_* given by,

$$\gamma_* = \sigma_H^2 SNR. \tag{8.44}$$

Proof : We evaluate the sign of derivative of $\tilde{M}_\lambda$ with respect to λ. See Appendix 8.A for details. □

In the above theorem, we characterized the nature of optimal λ_* for finite positive γ. For well behaved distributions, the optimal $\lambda_*(\gamma)$ is a continuous function of γ (Fig.8.3). The critical coherence index γ_* divides the channels into two categories, viz.,

- coherent channels ($\gamma > \gamma_*$) : the optimal λ_* is unbounded, which implies that increasing the number of simultaneous transmissions always improves the performance metric.
- canceling channels ($\gamma < \gamma_*$) : λ_* is bounded and unique, which implies that increasing the number of simultaneous transmissions beyond a point degrades the performance metric.

Hence, for the canceling channels, we need to design sleeping strategies to limit interference. On the other hand for coherent channels, the sensors simply need to transmit simultaneously, in order to maximize performance.

8.4 Numerical results and simulations

In this section, we resort to numerical and simulation techniques to validate the theories developed in this chapter. We consider binary quantized measurements

with PMF

$$\mathbf{p}_{\theta_0} = [0.7, 0.3], \quad \mathbf{p}_{\theta_1} = [0.3, 0.7].$$

For the Bayesian setting, we assume equally likely priors.

We assume that the channel fading is Gaussian[8] $H_{i,k} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_H, \sigma_H^2)$ with the mean and the variance varying according to different simulation conditions. The number of sensors involved in each transmission N_i is IID Poisson.

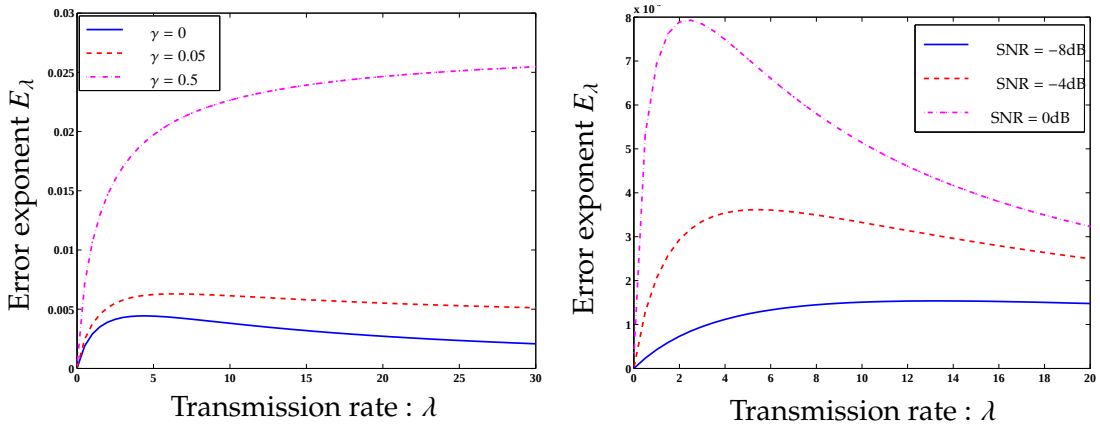
The error exponents are evaluated numerically (without using the Gaussian approximation), and the detection-error probabilities are estimated using Monte-Carlo simulations.

8.4.1 Evaluation of error exponents

Since the central limit theorem is applicable only in the large-λ regime, in order to draw conclusions for finite λ we numerically evaluate the Chernoff information and the Kullback-Leibler distance. We found that the Chernoff information and Kullback-Leibler distance have similar shapes. Therefore, only the behavior of Chernoff information is presented here.

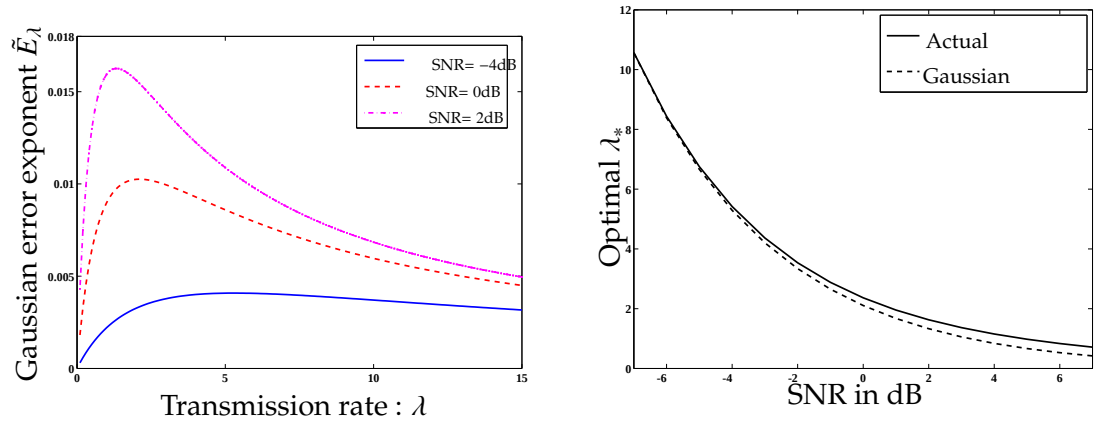
Fig.8.6a and Fig.8.6b show the behavior of the actual error exponent E_λ vs. λ, one (Fig.8.6a) with varying channel-coherence index γ for a fixed SNR, and the other with varying SNR for $\gamma = 0$. The existence of optimal λ_* is evident for small γ. To see the similarity and difference between the actual E_λ and the Gaussian approximated $\tilde{E}_\lambda$, we plot $\tilde{E}_\lambda$ in Fig.8.7a. The curves in Fig.8.6b and Fig.8.7a have similar shapes and share the same trend with respect to both λ

[8]For proper Gaussian variables, the real and imaginary parts are independent. Therefore, it suffices to limit to real variables



(a) SNR = 7dB, $\sigma_H^2 = 0.1$ (b) $\gamma = 0$, $\sigma_H^2 = 1$

Figure 8.6: Error exponent vs. transmission rate.



(a) Gaussian error exponent vs. transmission rate. ($\gamma = 0$, $\sigma_H^2 = 1$.) (b) Optimal transmission rate vs. SNR in dB. ($\gamma = 0$, $\sigma_H^2 = 1$.)

Figure 8.7: Gaussian approximation.

and SNR. The actual values of the error exponents are indeed different, with the Gaussian approximation giving a more conservative estimate of the true error exponent.

Fig.8.7b shows the actual optimal transmission rate λ_* and the sub-optimal rate $\tilde{\lambda}_*$ (obtained via optimizing the Gaussian error exponents in (8.29) and (8.30)). The optimal $\tilde{\lambda}_*$ obtained from the Gaussian approximation appears to

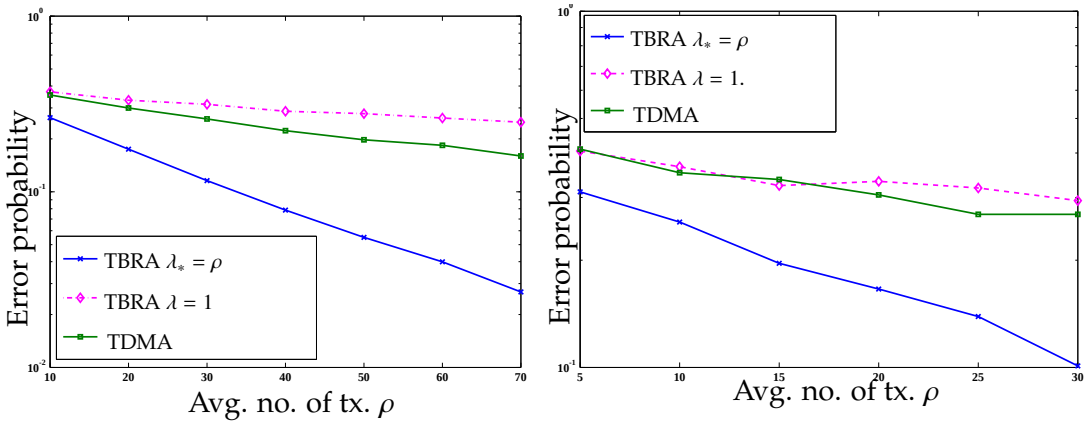
provide a reasonable approximation for the true λ_*. We observe that as SNR decreases, the sub-optimal $\tilde{\lambda}_*$ approaches the true λ_*. This behavior is the result of CLT, since the value of λ_* increases as the SNR decreases. Moreover, even at high SNR, the absolute error of approximation is fortunately quite small, since the value of λ_* decreases as the SNR increases.

8.4.2 Performance and discussion

We compare the TBRA scheme with the conventional TDMA scheme, in which one sensor is scheduled to transmit in a data collection, with energy $\mathcal{E}$. We fix the expected number of transmissions $\rho = l\lambda$ in the comparison. Our comparison is fair since both TBRA and TDMA have the same total energy budget.

We run simulations with values specified in the beginning of the section 8.4. We consider two regimes of γ and SNR, with four possible scenarios. For the TBRA scheme, from section 8.4.1, the optimal performance is at λ_*. We also consider the performance of TBRA under $\tilde{\lambda}_*$, obtained by Gaussian approximation. We also include the TBRA scheme with $\lambda = 1$, which enables us to study the random-access aspect of TBRA under different conditions.

Fig.8.8a and Fig.8.8b are simulations for non-zero mean fading channels. For the case shown in Fig.8.8a, the optimal strategy is single-shot transmission (all sensors transmitting simultaneously), since $\lambda_* > \rho$, for the values of ρ used in the simulation. We see that the optimal TBRA scheme performs better than TDMA. However, for the TBRA scheme with $\lambda = 1$ the performance is similar to TDMA. This suggests that the gain for TBRA comes mainly from coherence. In Fig.8.8b, at high SNR, the optimal strategy is still single-shot transmission, due



(a) $\gamma = 0.5$, SNR $= -4$dB, $\sigma_H^2 = 1$ (b) $\gamma = 0.5$, SNR $= 7$dB, $\sigma_H^2 = 0.1$

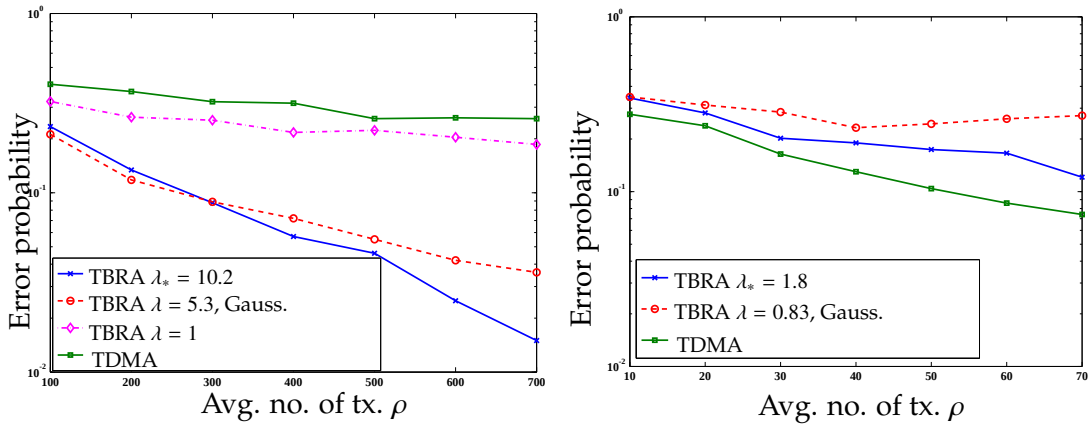
Figure 8.8: Error probability vs. transmission rate, non-zero mean fading.

to the high value of γ. This suggests that at large γ, the optimal strategy is to have as many simultaneous transmissions as allowed by the network, in order to exploit the channel coherency.

Fig.8.9a and Fig.8.9b are simulations under zero-mean fading. Fig.8.9a shows that TBRA performs better than TDMA at low SNR. In Fig.8.9b, we see that TDMA does slightly better at high SNR, under zero-mean fading.

We observe that there is not much performance gap between the optimal rate λ_* and sub-optimal $\tilde{\lambda}_*$, from Gaussian approximation. Also, we have ignored the communication overheads involved in each data collection. This is significantly higher for TDMA, since it has more data collections than TBRA, for a fixed ρ.

Some intuitions on the comparison of TBRA and TDMA are in order. At large γ, the optimal rate λ_* is also large. The gain from coherence suggests that at large γ, TBRA will do better than TDMA for a wide range of SNR values. Moreover, the total number of dimensions used by TBRA is just the number of data levels K (for single-shot transmission), which is far less than the number



(a) $\gamma = 0$, $\sigma_H^2 = 1$, SNR = −4dB.

(b) $\gamma = 0$, $\sigma_H^2 = 1$, SNR = 4dB.

Figure 8.9: Error probability vs. transmission rate, zero-mean fading.

of dimensions used by TDMA, given by $K\rho$. Thus, TBRA is also bandwidth efficient at large γ.

Under the zero-mean fading, λ_* is inversely proportional to the sensor SNR. Therefore, at high SNR, λ_* is small, in order to avoid canceling effects of zero-mean fading (Fig.8.7b). However, there are still some collisions between the transmissions, due to random access. In contrast, TDMA schedules exactly one sensor to transmit. We therefore expect TDMA to do better at high SNR, under zero-mean fading.

However, at low SNR, simultaneous transmissions counter noise more effectively. Therefore, we expect TBRA to perform better than TDMA. Moreover, at low SNR, TBRA has significantly lower number of data collections than TDMA leading to a quicker detection at fusion center. Also, the total number of dimensions used by TBRA ($\frac{K\rho}{\lambda_*}$) is far less than the number of dimensions used by TDMA ($K\rho$). Thus, at low SNR, TBRA is also bandwidth efficient under zero-mean fading.

8.4.3 Numerical Analysis for Parameter Estimation

The channel fading is proper complex Gaussian $H_{i,k} \overset{\text{i.i.d}}{\sim} \mathcal{N}_c(\mu_H, \sigma_H^2)$ and number of sensors involved in each transmission N_i is IID Poisson. Θ is drawn from triangular distribution $\Delta(0.2, 0.8)$ with 0.2 and 0.8 as the end-points. We consider the estimation of Bernoulli-distributed data at the sensors with Θ as the mean i.e.,

$$\mathbf{p}_\Theta = [\Theta; 1 - \Theta].$$

Since CLT is applicable only in large-λ regime, to draw conclusions for finite λ, we numerically evaluated the expected Fisher information. Fig.8.4 shows the plot of both true M_λ (without Gaussian approximation) and $\tilde{M}_\lambda$ (Gaussian approximation) for different values of coherence indices. We find that the true M_λ and $\tilde{M}_\lambda$ from the Gaussian approximation have similar shapes and share the same trend with respect to λ, γ and SNR. For larger values of γ, the Gaussian approximation does not appear to be good and needs large values of λ to converge. Fig.8.5 shows the accuracy of the Gaussian approximation in determining the optimal $\lambda_*(\gamma)$ for different values of γ. We find the Gaussian estimate to be quite close, especially at low values of γ, which are the practical cases of interest.

8.5 Conclusions

In this chapter, we focus on the communication aspect—random access in particular—of distributed detection and estimation for large sensor networks. We employ TBRA which inherits most of the attractive features of TBMA, e.g.,

the efficient bandwidth scaling and the asymptotic optimality under ideal channel conditions. The main advance of this work is the removal of the requirement of channel coherency and the ability to handle random number of sensors, transmitting simultaneously in a slot. By examining a number of extreme cases, we are able to obtain a general characterization of the error exponents as illustrated in Fig.8.2. From a practical stand-point, the approaches using the Gaussian approximation, presented in section 8.2.3, seem to give the correct insight into an optimal design. Such a characterization is a valuable guide, as a network designer pursues practical solutions.

We have left several important problems open. We have considered a spatio-temporal allocation scheme under the total energy constraint. For applications under other constraints (e.g., time), our formulation does not hold. We have not dealt with the design of local quantization rule. Given that the large-deviation analysis is used in this chapter, as well as in several related work [57, 58, 62] dealing with different aspects of the problem a "cross-layer" optimization of local quantization, communications and global inference should be of interest.

8.A Proofs

Proof of theorem 25

Let $\mathfrak{o}(\lambda)$ represent a function such that $\frac{\mathfrak{o}(\lambda)}{\lambda} \to 0 \quad \text{as} \quad \lambda \to 0$.

For the PMF of N, $g(n, \lambda)$, applying Taylor's expansion for λ near zero, we have $\mathbb{P}(N_i = 1) = a\lambda + \mathfrak{o}(\lambda) \quad \text{and} \quad \mathbb{P}(N_i > 1) = \mathfrak{o}(\lambda)$.

Define the conditional PDF of matched filter output $\mathbf{Y}_i$ under hypothesis $\mathcal{H}_k$ as $f_k(\cdot|N=0):=w(\cdot)$, $f_k(\cdot|N=1),:=h_k(\cdot)$ and $f_k(\cdot|N>1):=c_k(\cdot)$, where $w(\cdot)$ is the PDF of White Gaussian noise. Marginalizing over N, we have the PDF of $\mathbf{Y}_i$ under hypothesis $\mathcal{H}_k$,

$$\begin{aligned} f_k(\cdot) &= (1-\lambda a - \mathfrak{o}(\lambda))w(\cdot) \\ &\quad + (\lambda a + \mathfrak{o}(\lambda))h_k(\cdot) + \mathfrak{o}(\lambda)c_k(\cdot). \end{aligned} \tag{8.45}$$

Error Exponents

Now we have the KL distance [108]

$$E^{\text{NPD}}(\lambda) = \frac{D(f_0||f_1)}{\lambda} = \frac{1}{\lambda}\int_{\mathbf{y}} f_0(\mathbf{y}) \log \frac{f_0(\mathbf{y})}{f_1(\mathbf{y})}\mathbf{dy}. \tag{8.46}$$

Using (8.45) we have,

$$\lim_{\lambda\to 0} E^{\text{NPD}}(\lambda) = a\int_{\mathbf{y}} \Big(h_0(\mathbf{y}) - h_1(\mathbf{y})\Big)\mathbf{dy} = 0.$$

For Chernoff information we have [108],

$$\begin{aligned} C &= D(f_{v_*}||f_0) = D(f_{v_*}||f_1), \\ f_v(\cdot) &= \frac{f_0^v(\cdot)f_1^{1-v}(\cdot)}{\int_{\mathbf{y}} f_0^v(\mathbf{y})f_1^{1-v}(\mathbf{y})\mathbf{dy}}. \end{aligned} \tag{8.47}$$

Thus, we have, for $k=0,1$,

$$\lim_{\lambda\to 0}\frac{D(f_v\|f_k)}{\lambda}=\int_{\mathbf{y}}va\Big(h_0(\mathbf{y})+h_1(\mathbf{y})-2w(\mathbf{y})\Big)\mathbf{dy}=0.$$

Therefore from (8.47), we have $\lim_{\lambda\to 0}E^{\text{BD}}(\lambda)=0$. For the case when $\lambda\to\infty$, we first show that the limit coincides with exponents of limiting distribution. From assumption (1) stated in theorem 25, the integrand in (8.46) is differentiable with respect to λ implying that the integral is differentiable [135]. This implies that the limit in (8.46) exists as $\lambda\to\infty$ and coincide with the Gaussian exponent. Similar argument holds for (8.47). For expressions of Gaussian exponents for $\gamma=0$ and $\gamma=\infty$, refer Theorems 29 and 28. Although these theorems have results for Poisson N, the exponents for any general PMF $g(n,\lambda)$ coincide with Poisson, for $\gamma=0$. For $\gamma=\infty$, the claim can be easily shown for a general PMF $g(n,\lambda)$.

Parameter Estimation

Define the conditional PDF of the sufficient statistic $\mathbf{U}$ given $N=0,1$ and θ as

$$\begin{aligned}f_U(\mathbf{u}|N=0,\Theta=\theta;\lambda) &:= w(\mathbf{u}),\\ f_U(\mathbf{u}|N=1,\Theta=\theta;\lambda) &:= h_\theta(\mathbf{u}),\\ f_U(\mathbf{u}|N>1,\Theta=\theta) &:= c_\theta(\mathbf{u}),\end{aligned}$$

where $w(\cdot)$ is the PDF of white-Gaussian noise, independent of θ. Marginalizing over N, for small λ we have the PDF of $\mathbf{U}$ given θ as

$$f_U(\mathbf{u}|\theta;\lambda) = (1-\lambda-\mathfrak{o}(\lambda))w(\mathbf{y}) + (\lambda+\mathfrak{o}(\lambda))h_\theta(\mathbf{y}) + \mathfrak{o}(\lambda)c_\theta(\mathbf{y}).$$

Differentiating with respect to θ,

$$\frac{\partial}{\partial\theta}f_U(\mathbf{u}|\theta;\lambda) = (\lambda+\mathfrak{o}(\lambda))\frac{\partial}{\partial\theta}h_\theta(\mathbf{u}) + \mathfrak{o}(\lambda)\frac{\partial}{\partial\theta}c_\theta(\mathbf{u}).$$

From the definition of Fisher information

$$\frac{I_\lambda(\theta)}{\lambda} = \frac{1}{\lambda}\int_{\mathbf{u}}(\frac{\partial}{\partial\theta}\log f_U(\mathbf{u}|\theta;\lambda))^2 f_U(\mathbf{u}|\theta;\lambda)\mathbf{dy}.$$

Since $f_U(\mathbf{u}|\theta;\lambda)$ is a differentiable function of λ and $\mathbf{u}$, M_λ is continuous in λ [135]. Substituting for $f_U(\mathbf{u}|\theta;\lambda)$ and taking the limit,

$$\lim_{\lambda\to 0} M_\lambda = \lim_{\lambda\to 0}\mathbb{E}\frac{I_\lambda(\Theta)}{\lambda} = 0.$$

For the case when $\lambda\to\infty$, a limiting conditional distribution exists, by theorem 26. Let $\mathbf{Z}:=\frac{\mathbf{Y}}{\sqrt{\lambda}}$. Therefore, the sufficient statistic is $\mathbf{U} = \frac{\mathbf{Y}}{\sqrt{\lambda^3}} = \frac{\mathbf{Z}}{\lambda}$. Let $f_Z(\mathbf{z}|\theta;\lambda)$ be the conditional PDF of $\mathbf{Z}$ and Θ respectively.

$$\begin{aligned} M_\lambda &= \frac{\mathbb{E}[I_\lambda^U(\Theta)]}{\lambda}, \\ &= \frac{\mathbb{E}[I_\lambda^Z(\Theta)]}{\lambda^2}, \end{aligned}$$

where I_λ^Z is the Fisher information of $\mathbf{Z}$ at a given λ. Let $\mathbf{G} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_\theta)$, where $\mathbf{\Sigma}_\theta := \sigma_H^2 \text{Diag}(\mathbf{p}_\theta)$. The Gaussian Fisher information is given by

$$\begin{aligned} I^G(\theta) &= tr[\mathbf{\Sigma}_\theta^{-1} \frac{\partial \mathbf{\Sigma}_\theta}{\partial \theta} \mathbf{\Sigma}_\theta^{-1} \frac{\partial \mathbf{\Sigma}_\theta}{\partial \theta}], \\ &= \sum_{i=1}^{K} \frac{p_\theta'^2(i)}{p_\theta^2(i)}. \end{aligned}$$

We define $\mathbf{V}$ as

$$\mathbf{V} := \mathbf{Z} - \boldsymbol{\mu}_\theta, \quad \boldsymbol{\mu}_\theta := \lambda \mu_H \mathbf{p}_\theta.$$

and let $f_V(\mathbf{v}|\theta; \lambda)$ be the PDF of $\mathbf{V}$. From the local limit theorem for the densities [136], with the assumption that $\mathbb{E}[\mathbf{V}^k] < \infty$, for some $k \geq 3$, we have

$$\lim_{\lambda \to \infty} f_V(\mathbf{z}|\theta; \lambda) = f_G(\mathbf{z}|\theta).$$

Under the assumption of double differentiability of f_Z with respect of λ, θ and $\mathbf{z}$, the partial derivatives are also continuous.

$$\begin{aligned} \lim_{\lambda \to \infty} \frac{\partial}{\partial \theta} f_V(\mathbf{z}|\theta; \lambda) &= \lim_{\lambda \to \infty} \lim_{h \to 0} \frac{f_V(\mathbf{v}|\theta + h; \lambda) - f_V(\mathbf{v}|\theta; \lambda)}{h} \\ &= \frac{\partial}{\partial \theta} f_G(\mathbf{z}|\theta), \end{aligned}$$

where the limits can be interchanged, since f is assumed to be continuous in both λ and θ. Since the functions are continuous with respect to $\lambda \in \mathfrak{R}$, the limits and the expectations can also be interchanged. Therefore,

$$\lim_{\lambda\to\infty} I^V_\lambda(\theta) \to I^G(\theta).$$

Similarly, the expectation with respect to θ is also continuous. Now, in order to relate the Fisher information of $\mathbf{V}$ and $\mathbf{Z}$, we have

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log f_Z(\mathbf{z}|\theta;\lambda) &= \frac{\partial}{\partial\theta}\log f_V(\mathbf{z}-\boldsymbol{\mu}_\theta|\theta;\lambda) \\
&= \frac{\partial}{\partial\theta}\log f_V(\mathbf{v}|\theta;\lambda)\Big|_{\mathbf{v}=\mathbf{z}-\mu_\theta} \\
&\quad -\frac{\partial}{\partial\mathbf{v}}\log f_V(\mathbf{v}|\theta;\lambda)\Big|_{\mathbf{v}=\mathbf{z}-\mu_\theta}\frac{\partial\boldsymbol{\mu}_\theta^T}{\partial\theta}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
I^Z_\lambda(\theta) &= \mathbb{E}[\frac{\partial}{\partial\theta}\log f_Z(\mathbf{Z}|\theta;\lambda)]^2 \\
&= I^V_\lambda(\theta) + \mathbb{E}[\frac{\partial}{\partial\mathbf{v}}\log f_V(\mathbf{V}|\theta;\lambda)\Big(\frac{\partial\boldsymbol{\mu}_\theta}{\partial\theta}\Big)^T]^2 \\
&\quad -2\frac{\partial\boldsymbol{\mu}_\theta}{\partial\theta}\mathbb{E}[\frac{\partial}{\partial\mathbf{v}}\log f_V(\mathbf{V}|\theta;\lambda)\frac{\partial}{\partial\theta}\log f_V(\mathbf{V}|\theta;\lambda)].
\end{aligned}$$

The last term, under regularity conditions, is

$$\begin{aligned}
&\mathbb{E}[\frac{\partial}{\partial\mathbf{v}}\log f_V(\mathbf{V}|\theta;\lambda)\frac{\partial}{\partial\theta}\log f_V(\mathbf{V}|\theta;\lambda)] \\
= &\int_{\mathbf{v}}\frac{\partial}{\partial\mathbf{v}}\log f_V(\mathbf{V}|\theta;\lambda)\frac{\partial}{\partial\theta}f_V(\mathbf{V}|\theta;\lambda)\mathbf{dv} \\
= &\frac{\partial}{\partial\theta}\int_{\mathbf{v}}[\frac{\partial}{\partial\mathbf{v}}\log f_V(\mathbf{V}|\theta;\lambda)]f_V(\mathbf{V}|\theta;\lambda)\mathbf{dv} \\
= &\frac{\partial}{\partial\theta}\int_{\mathbf{v}}\frac{\partial}{\partial\mathbf{v}}f_V(\mathbf{V}|\theta;\lambda)\mathbf{dv} \\
= &f_V(\infty) - f_V(-\infty) = 0,
\end{aligned}$$

where we assume that the density is zero at infinity. For the second term, we have

$$\begin{aligned}
&\lim_{\lambda\to\infty} \mathbb{E}[\frac{\partial}{\partial \mathbf{v}} \log f_V(\mathbf{V}|\theta;\lambda)\Big(\frac{\partial \boldsymbol{\mu}_\theta}{\partial \theta}\Big)^T]^2 \\
= \;& \mathbb{E}[\frac{\partial}{\partial \mathbf{v}} \log f_G(\mathbf{V}|\theta)\Big(\frac{\partial \boldsymbol{\mu}_\theta}{\partial \theta}\Big)^T]^2, \\
= \;& 2\sum_{i=1}^{k} \mathbb{E}[\frac{V_i^2}{\Sigma_\theta(i)^2}(\frac{\partial \mu_\theta(i)}{\partial \theta})^2], \\
= \;& 2\sum_{i=1}^{k} \frac{1}{\Sigma_\theta(i)}(\frac{\partial \mu_\theta(i)}{\partial \theta})^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{\lambda\to\infty} M_\lambda &= \lim_{\lambda\to\infty} \frac{\mathbb{E}[I_\lambda^Z(\Theta)]}{\lambda^2}, \\
&= \lim_{\lambda\to\infty} \frac{\mathbb{E}[I^G(\Theta)]}{\lambda^2} + 2\gamma \sum_{i=1}^{K} \mathbb{E}[\frac{p'_\Theta(i)^2}{p_\Theta(i)}], \\
&= 2\gamma \sum_{i=1}^{K} \mathbb{E}[\frac{p'_\Theta(i)^2}{p_\Theta(i)}].
\end{aligned}$$

□

Proof of theorem 26

Recall the CLT with random number of summands [112, p. 369]. Let $X_1, X_2, \ldots,$ be IID random variables with mean 0 and variance σ^2, and $S_n = \sum_{i=1}^{n} X_i$. For each positive t, let ν_t be a random variable assuming positive integers as values; not necessarily independent of X_n. Suppose, there exist positive constants a_t and η such that $a_t \to \infty$, $\frac{\nu_t}{a_t} \xrightarrow{d} \eta$ as $t \to \infty$. Then

$$\frac{S_{\nu_t}}{\sigma\sqrt{\nu_t}} \xrightarrow{d} \mathcal{N}(0,1), \quad \frac{S_{\nu_t}}{\sigma\sqrt{\eta a_t}} \xrightarrow{d} \mathcal{N}(0,1). \tag{8.48}$$

In our case, parameter a_t corresponds to λ, ν_t to N. We have $\frac{N}{\lambda} \xrightarrow{d} \eta > 0$ and $\frac{1}{\sqrt{\eta\lambda}}\mathbf{W} \xrightarrow{p} \mathbf{0}$ as $\lambda \to \infty$. By Slutsky's theorem [112], $\frac{\mathbf{W}}{\sqrt{\eta\lambda}} \xrightarrow{d} \mathbf{0}$ as $\lambda \to \infty$. Extending the above to complex domain and to random vectors using *multivariate central limit theorem* [112, p. 385] we obtain,

$$\frac{\sum_{k=1}^{N}(H_k\mathbf{e}_{X_k} - \mu_H\mathbf{p}_\theta)}{\sqrt{\eta\lambda}} \xrightarrow{d} \mathcal{N}_c(0, \mathrm{Cov}(H_1 e_{X_1}))$$

When N is Poisson, let $N^{(m)}$ be the number of sensors transmitting data level m. Since $N^{(m)}$ is a thinning Poisson process [137, p. 317], $N^{(m)}$ is independent for different data levels and

$$N^{(m)} \sim \mathrm{Poiss}(\lambda p_\theta(m)).$$

Therefore, the vector $\mathbf{Y}$ has independent entries. Applying the above mentioned central limit theorem for random summands, to each entry of the vector we obtained the needed result. □

Proof of lemma 12

Let $f_i \sim \mathcal{N}_c(\mu_i, \Sigma_i)$. The KL distance $D(f_0\|f_1)$ is given by,

$$D(f_0||f_1) = \sum_k \Big(\log \frac{\Sigma_1(k,k)}{\Sigma_0(k,k)} + \big(\frac{\Sigma_0(k,k) + |\mu_0(k) - \mu_1(k)|^2}{\Sigma_1(k,k)} - 1\big)\Big).$$

Define density function f_γ by,

$$f_\gamma(\mathbf{y}) = \frac{f_0(\mathbf{y})^\gamma f_1(\mathbf{y})^{1-\gamma}}{\int f_0(\mathbf{y})^\gamma f_1(\mathbf{y})^{1-\gamma} d\mathbf{y}}.$$

The Chernoff information is given by [108],

$$C(f_0, f_1) = D(f_{\gamma^*}||f_0) = D(f_{\gamma^*}||f_1).$$

Solving the above equation yields the expression for Chernoff information. For our setup we find expressions for μ_i and Σ_i. □

Proof of lemma 27

(a) To prove the monotonicity of $\tilde{D}_{\lambda,\gamma}$ and $\tilde{C}_{\lambda,\gamma}$ with respect to λ and γ, we see that

$$\frac{\partial \tilde{D}}{\partial \lambda} = \sum_{k=1}^{M} \Big(\frac{\partial \tilde{D}}{\partial \alpha_k}\frac{\partial \alpha_k}{\partial \lambda} + \frac{\partial \tilde{D}}{\partial \omega_k}\frac{\partial \omega_k}{\partial \lambda}\Big).$$

From the expressions of α_k and ω_k we find that

$$\frac{\partial \tilde{D}}{\partial \alpha_k} = (-\frac{1}{\alpha_k} + 1 + \omega_k), \ \frac{\partial \tilde{D}}{\partial \omega_k} = \alpha_k - 1.$$

We see that $\Delta < 0$ implies $\alpha_k < 1$, $\omega_k < 0$, $\frac{\partial \alpha_k}{\partial \lambda} < 0$ and $\frac{\partial \omega_k}{\partial \lambda} < 0$. Similar results can be obtained for $\alpha_k > 1$. Combining the above, $\frac{\partial \tilde{D}}{\partial \lambda} > 0$. Similarly we obtain

$$\begin{aligned} \frac{\partial \tilde{D}}{\partial \gamma} &= \sum_k (\alpha_k - 1)\lambda\Delta > 0 \\ \frac{\partial \tilde{D}}{\partial \text{SNR}} &= \sum_k \Big(\frac{\partial \tilde{D}}{\partial \alpha_k} \frac{\partial \alpha_k}{\partial \text{SNR}} \Big) > 0. \end{aligned}$$

Similar results can be obtained for $\tilde{C}_{\lambda,\gamma}$.

(b) As $\lambda \to \infty$, $\alpha_k \to \frac{p_{\theta_0}(k)}{p_{\theta_1}(k)}$ and $\lim_{\lambda\to\infty} \frac{\omega_k(\alpha_k - 1)}{\lambda} = \frac{\gamma\Delta^2}{p_{\theta_1}(k)}$. This gives the result for $\tilde{E}^{\text{NPD}}(\lambda)$. As $\lambda \to \infty, \beta_k \to \frac{1}{\sqrt{\alpha_k}}$ and hence the result for $\tilde{E}^{\text{BD}}(\lambda)$. □

Proof of theorem 29

Define

$$\begin{aligned} f(x) &\triangleq (-\log x + x - 1) \\ g(x) &\triangleq \frac{\log x}{x - 1}, \ h(x, a, b) \triangleq \frac{xa + 1}{xb + 1} \end{aligned}$$

For the zero-mean fading, we have

$$\begin{aligned} D &= \sum_{k=1}^{M} f(\alpha_k) \quad \alpha_k = h(\chi, p_{\theta_0}(k), p_{\theta_1}(k)), \\ C &= \sum_{k=1}^{M} f(\beta_k) \quad \beta_k = g(\alpha_k). \end{aligned}$$

(a) *Monotonicity :* $g'(x) < 0$ and,

$$f'(x)\begin{cases} < 0 & x < 1, \\ \geq 0 & o.w. \end{cases} \qquad \frac{\partial h}{\partial x}(x,a,b)\begin{cases} > 0 & a > b \\ \leq 0 & o.w \end{cases}$$

Combining the above results, we obtain $C_\chi^{'} > 0$ Similarly, we obtain $C_\chi^{''} < 0$.

(b) *Limits :* We find C_∞ by substituting, $\lim_{\chi\to\infty} \alpha_k = \frac{p_{\theta_0}(k)}{p_{\theta_1}(k)}$ and see that $C_\infty < \infty$ for $p_{\theta_i}(k) > 0 \ \forall i,k$.

(c) *Extremal Points:* Let $M_\chi = \frac{C_\chi}{\chi}$. Now, $M_\chi' = 0$ implies

$$\chi\, C_\chi^{'} = C_\chi. \tag{8.49}$$

We have $M_0 = 0$ and $M_\infty = 0$ and M is differentiable. This implies that solution exists for (8.49). Let χ_* be the solution. Then the double derivative at χ_* is given by

$$M_{\chi_*}^{''} = \frac{C_{\chi_*}^{''}}{\chi_*} < 0.$$

Therefore, χ_* is the unique maximum. Similar results follow for D_χ. □

Proof of theorem 30

The sign of the derivative is crucial in determining the bounded nature of optimal $\tilde{\lambda}$. Differentiating (8.41) we obtain

$$\frac{\partial \tilde{M}}{\partial \lambda} = 2\mathbb{E}\sum_{k=1}^{K}\Big(\frac{\sigma_H^2 \mathrm{SNR} p'_\Theta(k)}{(\lambda\sigma_H^2 \mathrm{SNR} p_\Theta(k)+1)}\Big)^2 \Big(\frac{\gamma}{\sigma_H^2 \mathrm{SNR}} - \frac{\lambda\sigma_H^2 \mathrm{SNR} p_\Theta(k)-1}{\lambda\sigma_H^2 \mathrm{SNR} p_\Theta(k)+1}\Big).$$

Therefore, the sign of the function inside the expectation is determined by

$$\frac{\gamma}{\sigma_H^2 \mathrm{SNR}} - \frac{\lambda\sigma_H^2 \mathrm{SNR} p_\Theta(k)-1}{\lambda\sigma_H^2 \mathrm{SNR} p_\Theta(k)+1}, \quad k = 1, \ldots, K.$$

The term $\frac{\lambda\sigma_H^2 \mathrm{SNR}_{p_\Theta(k)-1}}{\lambda\sigma_H^2 \mathrm{SNR}_{p_\Theta(k)+1}}$ is an increasing function of λ for $\lambda > 0$ and attains maximum of 1 as $\lambda \to \infty$. The value of γ at which the sign reverses is therefore given by (8.44). This also implies the unimodality for $\gamma < \gamma_*$. □

CHAPTER 9

OUTLOOK FOR THE FUTURE

In this thesis, we considered distributed schemes with scalable resource utilization in terms of energy consumption and bandwidth requirement for transmissions towards achieving statistical inference. In Chapters 2-7, we considered routing schemes with in-network computation to fuse data as they are routed towards the fusion center. In Chapter 8, we considered random-access scheme with channel-aided computations to reduce energy and bandwidth requirements of transmissions.

The central theme of this thesis is that it is neither sustainable nor scalable to send all the data in the network to the fusion center; the key is to then reduce communications through computational thinking [138]. Scalable algorithms are particularly relevant in an era where data are readily available, networks are ubiquitous, and computation and communication are reaching new speeds. Scalable algorithms have complexity and communication requirements not growing rapidly with the ever-expanding data domains.

The results of this thesis are also a step towards a unified network theory. While point-to-point communication has been completely characterized by the Shannon capacity [108], its extension to networks has not been very successful. Moreover, unlike the optimistic results of point-to-point links, the network counterpart results have been so far discouraging. For instance, the seminal work of Gupta and Kumar [10] states that the capacity of wireless networks scales as $O(\sqrt{\frac{n}{\log n}})$ for n nodes which has been further improved to $O(\sqrt{n})$ in [139]. In other words, the capacity of the wireless networks do not scale well with the number of nodes. The key to overcoming this capacity hurdle is then

to incorporate computations instead of sending all the data, as demonstrated in this thesis.

In terms of future work, a unified mathematical and a computational framework for scalable information processing is needed. This involves (i) design of scalable algorithms for statistical inference and learning of high dimensional data (ii) data selection and in-network computation to achieve tradeoffs between communications costs, and the resulting accuracy of inference and learning.

9.1 Scalable Learning & Inference of High-Dimensional Data

In Chapter 5, we showed that scalable inference can be achieved through in-network computation for a special scenario of binary hypothesis testing of Markov dependency graphs. However, scalable inference is not always possible: we establish in Chapter 5 that *stabilizing* graphical models are scalable while *saturated* models (with fully connected dependency graphs) are not. Hence, it is crucial to establish necessary and sufficient conditions for scalable inference based on the structure of graphical models and the communication cost functions in the network.

Understanding scalability for other instances of inference and learning are also of interest. Some specific problems are inference and tracking of dynamically changing dependency graphs and network connections, designing scalable and consistent learning algorithms through dimensionality reduction [140] and through estimation of the intrinsic data dimension [141, 142], thereby circumventing the "curse of dimensionality" encountered in large data sets. An-

other extension is replacing the Markov graphical model considered in [1] with the more general multi-resolution model [143], and investigating scalable algorithms in such scenarios.

9.2 Cost-Performance Tradeoff for Inference & Learning

In Chapter 4, we considered data selection and fusion for binary hypothesis testing by assigning the Kullback-Leibler divergence between the hypotheses as the penalty function. In future, I plan to expand this framework to incorporate more general forms of inference and learning, and this requires characterizing penalty functions based on general inference and learning accuracies. Relevant here, is the recent work on analysis of accuracy (error exponents) in learning dependency graph structures [144]. Extensions towards designing efficient approximation algorithms for cost-performance tradeoff for general inference and learning are of interest. An interesting specialization is when the data nodes are in Euclidean space. Previously, efficient Steiner tree algorithms have been designed for Euclidean costs [145], and we can similarly exploit the Euclidean geometry to obtain efficient prize-collecting data fusion algorithms in future.

9.3 Relevance

A unified mathematical foundation to analyze scalable inference and learning in large networks needs to be established, and this spans multiple areas such as networking, communications, signal processing, machine learning, statistics, approximation theory, and information theory. A theory towards scalable infer-

ence and learning has a potential transformative effect on applications involving large complex networks such as smart energy grids, large social networks, biological networks, and financial infrastructures [146, 147]. This will in turn pave way towards fully exploiting the benefits of information revolution with emphasis on knowledge than mere data.

BIBLIOGRAPHY

[1] A. Anandkumar, J. Yukich, L. Tong, and A. Swami, "Energy Scaling Laws for Distributed Inference in Random Networks," *Accepted to IEEE J. Selec. Area Comm., available on Arxiv*, Dec. 2008.

[2] A. Anandkumar, L. Tong, A. Swami, and A. Ephremides, "Minimum Cost Data Aggregation with Localized Processing for Statistical Inference," in *Proc. of INFOCOM*, (Phoenix, USA), pp. 780–788, April 2008.

[3] A. Anandkumar, A. Ephremides, A. Swami, and L. Tong, "Routing for Statistical Inference in Sensor Networks," in *Handbook on Array Processing and Sensor Networks* (S. Haykin and R. Liu, eds.), ch. 23, John Wiley & Sons, 2009.

[4] A. Anandkumar, L. Tong, and A. Swami, "Optimal Node Density for Detection in Energy Constrained Random Networks," *IEEE Tran. Signal Proc.*, vol. 56, pp. 5232–5245, Oct. 2008.

[5] A. Anandkumar, L. Tong, and A. Swami, "Detection of Gauss-Markov Random Fields with Nearest-neighbor Dependency," *IEEE Tran. Information Theory*, vol. 55, pp. 816–827, Feb. 2009.

[6] A. Anandkumar, M. Wang, L. Tong, and A. Swami, "Prize-Collecting Data Fusion for Cost-Performance Tradeoff in Distributed Inference," in *Proc. of IEEE INFOCOM*, (Rio De Janeiro, Brazil), April 2009.

[7] A. Anandkumar and L. Tong, "Type-Based Random Access for Distributed Detection over Multiaccess Fading Channels," *IEEE Tran. Signal Proc.*, vol. 55, pp. 5032–5043, Oct. 2007.

[8] A. Anandkumar, L. Tong, and A. Swami, "Distributed estimation via random access," *Information Theory, IEEE Transactions on*, vol. 54, pp. 3175–3181, July 2008.

[9] A. Anandkumar and L. Tong, "A Large Deviation Analysis of Detection over Multi-Access Channels with Random Number of Sensors," in *Proc. of ICASSP'06*, vol. IV, (Toulouse, France), pp. 1097–1101, May 2006.

[10] P. Gupta and P. R. Kumar, "The Capacity of Wireless Networks," *IEEE Tran. Information Theory*, vol. 46, pp. 388–404, March 2000.

[11] M. Franceschetti and R. Meester, *Random Networks for Communication: From Statistical Physics to Information Systems*. Cambridge University Press, 2008.

[12] Q. Zhao and L. Tong, "Energy Efficiency of Large-Scale Wireless Networks: Proactive vs. Reactive Networking," *IEEE JSAC Special Issue on Advances in Military Wireless Communications*, May 2005.

[13] X. Liu and M. Haenggi, "Toward Quasiregular Sensor Networks: Topology Control Algorithms for Improved Energy Efficiency," *IEEE Tran. on Parallel and Distributed Systems*, pp. 975–986, 2006.

[14] X. Wu, G. Chen, and S. Das, "Avoiding Energy Holes in Wireless Sensor Networks with Nonuniform Node Distribution," *IEEE Tran. on Parallel and Distributed Systems*, vol. 19, pp. 710–720, May 2008.

[15] Q. Zhao, A. Swami, and L. Tong, "The Interplay Between Signal Processing and Networking in Sensor Networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 84–93, 2006.

[16] A. Giridhar and P. Kumar, "Toward a Theory of In-network Computation in Wireless Sensor Networks," *IEEE Comm. Mag.*, vol. 44, no. 4, pp. 98–107, 2006.

[17] R. Cristescu, B. Beferull-Lozano, M. Vetterli, and R. Wattenhofer, "Network Correlated Data Gathering with Explicit Communication: NP-Completeness and Algorithms," *IEEE/ACM Transactions on Networking (TON)*, vol. 14, no. 1, pp. 41–54, 2006.

[18] P. von Rickenbach and R. Wattenhofer, "Gathering Correlated Data in Sensor Networks," in *Joint workshop on Foundations of Mobile Computing*, pp. 60–66, 2004.

[19] H. Gupta, V. Navda, S. Das, and V. Chowdhary, "Efficient gathering of correlated data in sensor networks," in *Proc. of ACM Intl. symposium on Mobile ad hoc networking and computing*, pp. 402–413, 2005.

[20] S. Madden, M. Franklin, J. Hellerstein, and W. Hong, "TinyDB: an acquisitional query processing system for sensor networks," *ACM Transactions on Database Systems*, vol. 30, no. 1, pp. 122–173, 2005.

[21] C. Intanagonwiwat, R. Govindan, and D. Esterin, "Directed Diffusion

: A Scalable and Robust Paradigm for Sensor Networks," in *Proc. 6th ACM/Mobicom Conference*, (Boston,MA), pp. pp 56–67, 2000.

[22] B. Krishnamachari, D. Estrin, and S. Wicker, "Modeling Data-centric Routing in Wireless Sensor Networks," in *IEEE INFOCOM*, (New York, USA), 2002.

[23] A. Giridhar and P. Kumar, "Maximizing the functional lifetime of sensor networks," in *Proc. of IPSN*, 2005.

[24] A. Giridhar and P. Kumar, "Computing and Communicating Functions over Sensor Networks," *IEEE JSAC*, vol. 23, no. 4, pp. 755–764, 2005.

[25] S. Subramanian, P. Gupta, and S. Shakkottai, "Scaling Bounds for Function Computation over Large Networks," in *IEEE ISIT*, June 2007.

[26] O. Ayaso, D. Shah, and M. Dahleh, "Counting Bits for Distributed Function Computation," in *Proc. ISIT*, (Toronto, Canada), pp. 652–656, July 2008.

[27] Y. Sung, S. Misra, L. Tong, and A. Ephremides, "Cooperative Routing for Signal Detection in Large Sensor Networks," *IEEE JSAC*, vol. 25, no. 2, pp. 471–483, 2007.

[28] J. Chamberland and V. Veeravalli, "How Dense Should a Sensor Network Be for Detection With Correlated Observations?," *IEEE Tran. on Information Theory*, vol. 52, no. 11, pp. 5099–5106, 2006.

[29] S. Misra and L. Tong, "Error Exponents for Bayesian Detection with Randomly Spaced Sensors," *IEEE Tran. on Signal Processing*, vol. 56, no. 8, 2008.

[30] Y. Sung, X. Zhang, L. Tong, and H. Poor, "Sensor Configuration and Activation for Field Detection in Large Sensor Arrays," *IEEE Tran. on Signal Processing*, vol. 56, no. 2, pp. 447–463, 2008.

[31] Y. Sung, H. Yu, and H. V. Poor, "Information, Energy and Density for Adhoc Sensor Networks over Correlated Random Fields: Large-deviation Analysis," in *IEEE ISIT*, pp. 1592–1596, July 2008.

[32] N. Katenka, E. Levina, and G. Michailidis, "Local Vote Decision Fusion

for Target Detection in Wireless Sensor Networks," in *Joint Research Conf. on Statistics in Quality Industry and Tech.*, (Knoxville, USA), June 2006.

[33] L. Yu, L. Yuan, G. Qu, and A. Ephremides, "Energy-driven Detection Scheme with Guaranteed Accuracy," in *Proc. IPSN*, pp. 284–291, 2006.

[34] O. Kreidl and A. Willsky, "Inference with Minimal Communication: a Decision-Theoretic Variational Approach," in *Advances in Neural Information Processing Systems*, 2006.

[35] S. Jiang, R. Kumar, and H. Garcia, "Optimal sensor selection for discrete-event systems with partial observation," *Automatic Control, IEEE Tran. on*, vol. 48, no. 3, pp. 369–381, 2003.

[36] V. Isler and R. Bajcsy, "The sensor selection problem for bounded uncertainty sensing models," in *Proc. IPSN*, pp. 151–158, April 2005.

[37] P. Atrey, M. Kankanhalli, and J. Oommen, "Goal-oriented optimal subset selection of correlated multimedia streams," *ACM Tran. on Multimedia Comp. Comm. & App.*, vol. 3, no. 1, 2007.

[38] S. Joshi and S. Boyd, "Sensor Selection via Convex Optimization," *Accepted to IEEE Tran. on Signal Proc.*, 2008.

[39] Y. Nakamura, K. Tei, Y. Fukazawa, and S. Honiden, "Region-Based Sensor Selection for Wireless Sensor Networks," pp. 326–331, 2008.

[40] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg, "Near-optimal sensor placements: maximizing information while minimizing communication cost," in *Proc. of IPSN*, pp. 2–10, 2006.

[41] V. Gupta, T. Chung, B. Hassibi, and R. Murray, "On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage," *Automatica*, vol. 42, no. 2, pp. 251–260, 2006.

[42] R. Debouk, S. Lafortune, and D. Teneketzis, "On an Optimization Problem in Sensor Selection," *Discrete Event Dynamic Systems*, vol. 12, no. 4, pp. 417–445, 2002.

[43] H. Rowaihy, S. Eswaran, M. Johnson, D. Verma, A. Bar-Noy, T. Brown, and T. La Porta, "A survey of sensor selection schemes in wireless sensor networks," in *Proc. of SPIE*, 2007.

[44] P.-N. Chen, "General formulas for the Neyman-Pearson type-II error exponent subject to fixed and exponential type-I error bounds," *IEEE Tran. on Information Theory*, vol. 42, pp. 316–323, Jan. 1996.

[45] T. Han, "Hypothesis Testing with the General Source," *IEEE Tran. Inform. Theory*, vol. 46, pp. 2415–2427, Nov. 2000.

[46] H. Künsch, "Thermodynamics and Statistical Analysis of Gaussian Random Fields," *Probability Theory and Related Fields*, vol. 58, no. 3, pp. 407–421, 1981.

[47] Z. Rached, F. Alajaji, and L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *Information Theory, IEEE Transactions on*, vol. 50, no. 5, pp. 917–921, 2004.

[48] M. Penrose and J. Yukich, "Limit Theory For Random Sequential Packing And Deposition," *Annals of Applied probability*, vol. 12, no. 1, pp. 272–301, 2002.

[49] M. Penrose and J. Yukich, "Weak Laws Of Large Numbers In Geometric Probability," *Annals of Applied Probability*, vol. 13, no. 1, pp. 277–303, 2003.

[50] J. Steele, "Growth Rates of Euclidean Minimal Spanning Trees with Power Weighted Edges," *The Annals of Probability*, vol. 16, no. 4, pp. 1767–1787, 1988.

[51] J. Yukich, "Asymptotics for Weighted Minimal Spanning Trees on Random Points," *Stochastic Processes and their Applications*, vol. 85, no. 1, pp. 123–138, 2000.

[52] M. Penrose, *Random Geometric Graphs*. Oxford University Press, 2003.

[53] P. K. Varshney, *Distributed Detection and Data Fusion*. New York, NY: Springer, 1997.

[54] R. Viswanathan and P.K.Varshney, "Distributed Detection with Multiple Sensors: Part I-Fundamentals," *Proceedings of the IEEE*, vol. 85, pp. 54–63, Jan. 1997.

[55] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed Detection with Multiple Sensors: Part II-Advanced Topics," *Proceedings of the IEEE*, vol. 85, pp. 64–79, Jan. 1997.

[56] J.-F. Chamberland and V. V. Veeravalli, "Asymptotic results for decentralized detection in power constrained wireless sensor networks," *IEEE JSAC*, 2004.

[57] J.-F. Chamberland and V. V. Veeravalli, "Decentralized Detection in Sensor Networks," *IEEE Tran. Signal Processing*, vol. 51, pp. 407–416, February 2003.

[58] S. Aldosari and J. Moura, "Detection in decentralized sensor networks," in *Proc. of ICASSP 04 Conf.*, (Montreal, Canada), May 2004.

[59] B. Chen, R. Jiang, T. Kasetkasem, and P. Varshney, "Channel aware decision fusion in wireless sensor networks," *IEEE Tran. on Signal Processing*, vol. 52, pp. 3454–3458, Dec. 2004.

[60] B. Chen and P.Willett, "On the Optimality of the Likelihood-Ratio Test for Local Sensor Decision Rules in the Presence of Nonideal Channels," *IEEE Tran. on Information Theory*, vol. 51, pp. 693–700, Feb 2005.

[61] G. Mergen and L. Tong, "Estimation Over deterministic multiaccess channels," in *Proceedings of the 42nd Allerton Conf. on Communications, Control, and Computing*, (Monticello, IL), Sep.28-Oct.01 2003.

[62] K. Liu and A. M. Sayed, "Optimal distributed detection strategies for wireless sensor networks,," in *42nd Annual Allerton Conf. on Commun., Control and Comp.*, Oct. 2004.

[63] G. Mergen, V. Naware, and L. Tong, "Asymptotic detection performance of type-based multiple access in sensor networks," in *Proc. of SPAWC'05*, (New York, NY), June 2005.

[64] G. Mergen, V. Naware, and L. Tong, "Asymptotic detection performance of Type-Based Multiple Access over multiaccess fading channels," *to appear in IEEE Tran. on Signal Processing*, 2007.

[65] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *VLDB*, 2004.

[66] R. Cristescu and M. Vetterli, "On the optimal density for real-time data gathering of spatio-temporal processes in sensor networks," in *IPSN*, pp. 159–164, 2005.

[67] D. Marco, E. Duarte-Melo, M. Liu, and D. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," in *Proc. IPSN*, pp. 1–16, 2003.

[68] S. Yoon and C. Shahabi, "The Clustered Aggregation technique leveraging spatial and temporal correlations in wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 3, no. 1, 2007.

[69] J. Faruque and A. Helmy, "RUGGED: RoUting on finGerprint Gradients in sEnsor Networks," in *IEEE/ACS Intl. Conf. on Pervasive Services (ICPS)*, pp. 179–188, 2004.

[70] S. Pattem, B. Krishnamachari, and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," in *Proceedings of the third international symposium on Information processing in sensor networks*, pp. 28–35, 2004.

[71] D. Ganesan, B. Greenstein, D. Perelyubskiy, D. Estrin, and J. Heidemann, "An evaluation of multi-resolution search and storage in resource-constrained sensor networks," in *Proceedings of the First ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2003.

[72] A. Jindal and K. Psounis, "Modeling spatially correlated data in sensor networks," *ACM Tran. on Sensor Net.*, vol. 2, no. 4, pp. 466–499, 2006.

[73] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of Royal Stat. Soc.*, vol. 36, no. B, pp. 192–225, 1974.

[74] J. Besag, "Statistical analysis of non-lattice data," *The Statistician*, vol. 24, no. 3, pp. 179–195, 1975.

[75] J. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," *Unpublished manuscript*, 1971.

[76] P. Clifford, "Markov Random Fields in Statistics," *Disorder in Physical Systems*, pp. 19–32, 1990.

[77] M. Cetin, L. Chen, J. Fisher, A. Ihler, O. Kreidl, R. Moses, M. Wainwright, J. Williams, and A. Willsky, "Graphical Models and Fusion in Sensor Networks," in *Wireless sensor networks: signal processing & comm. perspectives*, ch. 9, pp. 215–250, J. Wiley & Sons, 2007.

[78] S. Li, *Markov random field modeling in computer vision*. Springer-Verlag London, 1995.

[79] N. Cressie, *Statistics for spatial data*. J. Wiley, 1993.

[80] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Tran. on W. Comm.*, vol. 1, pp. 660–670, Oct. 2002.

[81] T. Kwon and M. Gerla, "Clustering with power control," in *Military Communications Conference Proceedings, IEEE*, vol. 2, 1999.

[82] Y. Sung, L. Tong, and H. Poor, "Neyman-Pearson Detection of Gauss-Markov Signals in Noise: Closed-Form Error Exponent and Properties," *IEEE Tran. Information Theory*, vol. 52, pp. 1354–1365, April 2006.

[83] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Tran. on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[84] S. Sanghavi, V. Tan, and A. Willsky, "Learning Graphical Models for Hypothesis Testing," in *IEEE Workshop on Stat. Signal Proc.*, 2007.

[85] S. Lauritzen, *Graphical models: Clarendon Press*. Clarendon Press, 1996.

[86] K. Jung and D. Shah, "Approximate message-passing inference algorithm," in *IEEE ITW*, pp. 224–229, 2007.

[87] J. Pearl, *Probabilistic Reasoning in Intelligent Systems—Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[88] D. Eppstein, "All maximal independent sets and dynamic dominance for sparse graphs," in *Proc. of ACM-SIAM symp. on discrete algorithms*, pp. 451–459, 2005.

[89] N. Cressie and S. Lele, "New Models for Markov Random Fields," *Journal of Applied Probability*, vol. 29, no. 4, pp. 877–884, 1992.

[90] R. Cowell, *Probabilistic Networks and Expert Systems*. Springer, 1999.

[91] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1994.

[92] E. Dynkin, "Necessary and Sufficient Statistics for a Family of Probability Distributions," *Tran. Math, Stat. and Prob.*, vol. 1, pp. 23–41, 1961.

[93] D. P. Bertsekas and R. Gallager, *Data Networks*. Prentice Hall, 1992.

[94] R. Koetter and M. Medard, "An algebraic approach to network coding," *IEEE/ACM Tran. on Networking*, vol. 11, no. 5, pp. 782–795, 2003.

[95] B. Wu and K. Chao, *Spanning Trees and Optimization Problems*. Chapman & Hall, 2004.

[96] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "Data fusion trees for detection: Does architecture matter?," *IEEE Tran. on Information Theory*, 2007. submitted for publication.

[97] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "On the sub-exponential decay of detection error probabilities in long tandems," *IEEE Tran. on Information Theory*, 2007. submitted for publication.

[98] L. Devroye, "The expected size of some graphs in computational geometry.," *Comp. & math. with app.*, vol. 15, no. 1, pp. 53–64, 1988.

[99] R. Gallager, P. Humblet, and P. Spira, "A Distributed Algorithm for Minimum-Weight Spanning Trees," *ACM Tran on Prog. Languages and Sys. (TOPLAS)*, vol. 5, no. 1, pp. 66–77, 1983.

[100] P. Humblet, "A Distributed Algorithm for Minimum Weight Directed Spanning Trees," *IEEE Tran. on Communications*, vol. 31, no. 6, pp. 756–762, 1983.

[101] G. Reich and P. Widmayer, "Beyond steiner's problem: a vlsi oriented generalization," in *Proc. of Intl. workshop on Graph-theoretic concepts in computer science*, pp. 196–210, 1990.

[102] V. Vazirani, *Approximation Algorithms*. Springer, 2001.

[103] L. Kou, G. Markowsky, and L. Berman, "A fast approximation algorithm for Steiner trees," *Acta Informatica*, vol. 15, pp. 141–145, 1981.

[104] G. Robins and A. Zelikovsky, "Improved Steiner tree approximation in graphs," in *Proc. of ACM-SIAM symposium on Discrete algorithms*, pp. 770–779, 2000.

[105] N. Garg, G. Konjevod, and R. Ravi, "A polylogarithmic approximation algorithm for the group steiner tree problem," *J. Algorithms*, vol. 37, no. 1, pp. 66–84, 2000.

[106] C. S. Helvig, G. Robins, and A. Zelikovsky, "Improved approximation bounds for the group steiner problem," in *DATE '98: Proceedings of the conference on Design, automation and test in Europe*, pp. 406–413, 1998.

[107] M. Goemans and D. Williamson, "A General Approximation Technique for Constrained Forest Problems," *SIAM J. on Computing*, vol. 24, p. 296, 1995.

[108] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[109] A. Lawson and D. Denison, *Spatial Cluster Modelling*. CRC Press, 2002.

[110] A. Pettitt, I. Weir, and A. Hart, "A Conditional Autoregressive Gaussian Process for Irregularly Spaced Multivariate Data with Application to Modelling Large Sets of Binary Data," *Statistics and Computing*, vol. 12, no. 4, pp. 353–367, 2002.

[111] A. Lucena and M. Resende, "Strong lower bounds for the prize collecting Steiner problem in graphs," *Discrete Applied Mathematics*, vol. 141, no. 1-3, pp. 277–294, 2004.

[112] P.Billingsley, *Probability and Measure*. New York, NY: Wiley Inter-Science, 1995.

[113] A. Ephremides, "Energy Concerns in Wireless Networks," *IEEE Wireless Communications*, pp. 48–59, August 2002.

[114] W. Li and H. Dai, "Energy-Efficient Distributed Detection Via Multihop Transmission in Sensor Networks," *IEEE Signal Processing Letters*, vol. 15, pp. 265–268, 2008.

[115] D. Aldous and J. Steele, "The objective method: probabilistic combinatorial optimization and local weak convergence," *Probability on Discrete Structures*, vol. 110, pp. 1–72, 2004.

[116] A. Anandkumar, L. Tong, and A. Swami, "Energy Efficient Routing for

Statistical Inference of Markov Random Fields," in *Proc. of CISS '07*, (Baltimore, USA), pp. 643–648, March 2007.

[117] P. Brémaud, *Markov Chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer, 1999.

[118] X. Li, "Algorithmic, Geometric and Graphs Issues in Wireless Networks," *Wireless Comm. and Mobile Computing*, vol. 3, March 2003.

[119] P. Wan and C. Yi, "On the Longest Edge of Gabriel Graphs in Wireless Ad Hoc Networks," *IEEE Tran. on Parallel and Distributed Systems*, pp. 111–125, 2007.

[120] M. Penrose, "Laws Of Large Numbers In Stochastic Geometry With Statistical Applications," *Bernoulli*, vol. 13, no. 4, pp. 1124–1150, 2007.

[121] M. Penrose and J. Yukich, "Central Limit Theorems For Some Graphs In Computational Geometry," *Annals of Applied Probability*, vol. 11, no. 4, pp. 1005–1041, 2001.

[122] J. Møller, *Spatial statistics and computational methods*. Springer New York, 2003.

[123] M. Jaksa, P. Brooker, and W. Kaggwa, "Inaccuracies Associated with Estimating Random Measurement Errors," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 123, no. 5, pp. 393–401, 2006.

[124] G. Grundmann and D. Debouzie, "Geostatistical analysis of the distribution of NH 4 and NO 2-oxidizing bacteria and serotypes at the millimeter scale along a soil transect," *FEMS Microbiol. Ecol*, vol. 34, pp. 57–62, 2000.

[125] M. Barange and I. Hampton, "Spatial structure of co-occurring anchovy and sardine populations from acoustic data: implications for survey design," *Fisheries Oceanography*, vol. 6, no. 2, pp. 94–108, 1997.

[126] C. Zong, "The kissing numbers of convex bodies- a brief survey," *Bulletin of the London Mathematical Society*, vol. 30, no. 01, pp. 1–10, 2000.

[127] D. Eppstein, M. S. Paterson, and F. F. Yao, "On Nearest-Neighbor Graphs," *Discrete and Computational Geometry*, vol. 17, pp. 263–282, 1997.

[128] A. Wade, "Explicit laws of large numbers for random nearest-neighbour type graphs," *Adv. Appl. Probab*, vol. 39, 2007.

[129] A. Baddeley, *A crash course in stochastic geometry*. London: Chapman and Hall, 1999. Stochastic Geometry: Likelihood and Computation, eds O.E. Barndorff-Nielsen, W.S. Kendall, M.N.M. van Lieshout.

[130] P. Willett and L. Tong, "One aspect to cross-layer design in sensor networks," in *Proceedings of MILCOM*, (Monterey, Ca), October 2004.

[131] G. Mergen and L. Tong, "Type based estimation over multiaccess channels," *IEEE Tran. Signal Processing*, vol. 54, pp. 613–626, February 2006.

[132] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications, 2nd ed.* NY: Springer, 1998.

[133] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Tran. on Info. Theory*, vol. 46, Jul 2000.

[134] H. V. Trees, *Detection, Estimation and Modulation Theory*, vol. 1. New York: Wiley, 1968.

[135] R.S.Strichartz, *The Way of Analysis*, vol. 2. Sudbury, MA: Jones and Bartlett, 2000.

[136] V. Petrov, "On Local Limit Theorems for Sums of Independent Random Variables," *Theory of Probability and its Applications*, vol. 9, p. 312, 1964.

[137] S. Resnick, *Adventures in Stochastic Processes*. New York: Springer Verlag, 2002.

[138] J. Wing, "Computational Thinking," *Comm. of the ACM*, vol. 49, no. 3, p. 33, 2006.

[139] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Tran. on Information Theory*, vol. 53, no. 10, pp. 3549–3572, 2007.

[140] K. Varshney and A. Willsky., "Learning Dimensionality-Reduced Classifiers for Information Fusion'," in *Proc. of Intl. Conf. on Information Fusion*, (Seattle, Washington), July 2009.

[141] E. Levina and P. Bickel, "Maximum Likelihood Estimation of Intrinsic Dimension," in *Proc. of NIPS*, vol. 17, 2004.

[142] J. Costa and A. Hero III, "Determining Intrinsic Dimension and Entropy of High-Dimensional Shape Spaces," *Statistics And Analysis of Shapes*, p. 231, 2006.

[143] M. Choi, "Multiscale Gaussian Graphical Models and Algorithms for Large-Scale Inference," Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, June 2007.

[144] V. Tan, A. Anandkumar, L. Tong, and A. Willsky, "A Large-Deviation Analysis for the Maximum Likelihood Learning of Tree Structures," *submitted to IEEE Tran. on Information Theory, available on Arxiv*, April 2009.

[145] S. Arora, "Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems," *Journal of the ACM (JACM)*, vol. 45, no. 5, pp. 753–782, 1998.

[146] A. Anandkumar, C. Bisdikian, and D. Agrawal, "Tracking in a Spaghetti Bowl: Monitoring Transactions Using Footprints," in *Proc. ACM Intl. Conf. on Measurement & Modeling of Computer Systems (Sigmetrics)*, (Annapolis, Maryland, USA), June 2008.

[147] A. Anandkumar, C. Bisdikian, T. He, and D. Agrawal, "Designing A Fine Tooth Comb Frugally: Selectively Retrofitting Monitoring in Distributed Systems," (Seattle, USA), June 2009. Workshop on Mathematical Performance Modeling and Analysis (MAMA).