DECENTRALIZED ESTIMATION UNDER COMMUNICATION CONSTRAINTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MURAT ÜNEY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

AUGUST 2009

Approval of the thesis:

DECENTRALIZED ESTIMATION UNDER COMMUNICATION CONSTRAINTS

submitted by **MURAT ÜNEY** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İsmet Erkmen
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. Kemal Leblebicioğlu
Supervisor, **Electrical and Electronics Engineering Dept., METU** _____

Assist. Prof. Dr. Müjdat Çetin
Co-supervisor, **FENS, Sabancı University** _____

Examining Committee Members:

Prof. Dr. Mübeccel Demirekler
Electrical and Electronics Engineering Dept., METU _____

Prof. Dr. Kemal Leblebicioğlu
Electrical and Electronics Engineering Dept., METU _____

Prof. Dr. Yasemin Yardımcı Çetin
Informatics Institute, METU _____

Prof. Dr. Uğur Halıcı
Electrical and Electronics Engineering Dept., METU _____

Assist. Prof. Dr. Sinan Gezici
Electrical and Electronics Engineering Dept., Bilkent University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: MURAT ÜNEY

Signature :

ABSTRACT

DECENTRALIZED ESTIMATION UNDER COMMUNICATION CONSTRAINTS

Üney, Murat

M.S., Department of Electrical and Electronics Engineering

Supervisor : Prof. Dr. Kemal Leblebicioğlu

Co-Supervisor : Assist. Prof. Dr. Müjdat Çetin

August 2009, 136 pages

In this thesis, we consider the problem of decentralized estimation under communication constraints in the context of Collaborative Signal and Information Processing. Motivated by sensor network applications, a high volume of data collected at distinct locations and possibly in diverse modalities together with the spatially distributed nature and the resource limitations of the underlying system are of concern. Designing processing schemes which match the constraints imposed by the system while providing a reasonable accuracy has been a major challenge in which we are particularly interested in the tradeoff between the estimation performance and the utilization of communications subject to energy and bandwidth constraints.

One remarkable approach for decentralized inference in sensor networks is to exploit graphical models together with message passing algorithms. In this framework, after the so-called information graph of the problem is constructed, it is mapped onto the underlying network structure which is responsible for delivering the messages in accordance with the schedule of the inference algorithm. However it is challenging to provide a design perspective that addresses the tradeoff between the estimation accuracy and the cost of communications. Another approach has been performing the estimation at a fusion center based on the quantized

information provided by the peripherals in which the fusion and quantization rules are sought while taking a restricted set of the communication constraints into account.

We consider two classes of in-network processing strategies which cover a broad range of constraints and yield tractable Bayesian risks that capture the cost of communications as well as the penalty for estimation errors. A rigorous design setting is obtained in the form of a constrained optimization problem utilizing the Bayesian risks. These processing schemes have been previously studied together with the structures that the solutions exhibit in the context of decentralized detection in which a decision out of finitely many choices is made.

We adopt this framework for the estimation problem. However, for the case, computationally infeasible solutions arise that involve integral operators that are impossible to evaluate exactly in general. In order not to compromise the fidelity of the model we develop an approximation framework using Monte Carlo methods and obtain particle representations and approximate computational schemes for both the in-network processing strategies and the solution schemes to the design problem. Doing that, we can produce approximating strategies for decentralized estimation networks under communication constraints captured by the framework including the cost. The proposed Monte Carlo optimization procedures operate in a scalable and efficient manner and can produce results for any family of distributions of concern provided that samples can be produced from the marginals. In addition, this approach enables a quantification of the tradeoff between the estimation accuracy and the cost of communications through a parameterized Bayesian risk.

Keywords: Collaborative Signal and Information Processing, Decentralized Estimation, Communication Constrained Inference, Random Fields, Message Passing Algorithms, Graphical Models, Sensing Architectures, Monte Carlo Methods.

ÖZ

İLETİŞİM KISITLARI ALTINDA DAĞITIK KESTİRİM

Üney, Murat
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Kemal Leblebicioğlu
Ortak Tez Yöneticisi : Yrd. Doç. Dr. Müjdat Çetin

Ağustos 2009, 136 sayfa

Bu tez çalışmasında, İmeceli Sinyal ve Bilgi İşleme kapsamında iletişim kısıtları altında dağıtık kestirim problemini ele almaktayız. Algılayıcı ağı uygulamalarıyla gündeme gelen yüksek hacimli, ayrı ayrı konumlardan toplanan ve muhtemelen farklı yapılardaki veri ile beraber altta yatan sistemin uzamsal dağıtık doğası ve kaynak sınırlarının göz önünde bulundurulması önemlidir. Başlıca bir zorluğu bir yandan kabul edilebilir doğrulukta sonuçlar sağlarken öte yandan sistemin dayattığı kısıtlara uyan işleme yöntemlerinin tasarlanması olan bu kapsamda özel olarak kestirim başarımı ile enerji ve bantgenişliği kısıtlarına tabi olan haberleşme kullanımı arasındaki ödünleşim ile ilgilenmekteyiz.

Algılayıcı ağlarında dağıtık çıkarsama için dikkat çeken yaklaşımlardan biri çizge modelleri ve ileti geçme algoritmalarından faydalanmaktır. Bu çerçevede problemin bilgi çizgesi olarak da anılan gösterimi kurgulandıktan sonra söz konusu çizge çıkarsama algoritmasının çizelgesi uyarınca iletileri dağıtmakla sorumlu olan temeldeki ağ yapısına eşlenir. Ancak açık bir biçimde kestirim doğruluğu ve iletişimin maliyeti arasındaki ödünleşimi dikkate alan tasarımlar yapmak zordur. Bir başka yaklaşım ise kestirimi yan birimlerden sağlanan nicemlenmiş bilgiye dayalı olarak bir kaynaşım merkezinde gerçekleştirmektir ki dar bir dizi

iletişim kısıtı gözetilerek kaynaşım ve nicemleme kuralları bulunmaya çalışılır.

Biz ise geniş bir kısıtlar yelpazesini kapsayan ve de hem kestirim hataları için ceza hem de iletişim maliyetini içeren çözümlenebilir bir Bayesçi risk tanımlanmasına olanak veren iki sınıf ağda işleme stratejisi ele almaktayız. Söz konusu Bayesçi riskler kullanılarak, kısıtlı eniyileme problemi biçiminde, matematiksel kesinlikli bir tasarım çerçevesi elde edilmektedir. Bu işleme yöntemleri, çözümlerin sergilediği yapılar ile birlikte, sonlu sayıda seçenekten bir kararın söz konusu olduğu dağıtık sezim kapsamında henüz çalışılmıştır.

Bu çerçeveyi kestirim problemine uyarlamaktayız. Ancak, bu durumda, değerlemesi genel olarak imkansız tümlev işleçleri içeren, hesapsal olurluğu olmayan çözümler ortaya çıkmaktadır. Biz ise modelin sadakatinden ödün vermemek için Monte Karlo yöntemlerini kullanarak bir yaklaşıklama çerçevesi geliştirmekte ve hem ağda işleme stratejileri hem de tasarım probleminin çözümü için parçacık temsilleri ile yaklaşık hesapsal yöntemler elde etmekteyiz. Böylece bu çerçevenin kapsadığı ve maliyeti de içeren iletişim kısıtları altında dağıtık kestirimci ağları için yaklaşıklayan stratejiler üretebilmekteyiz. Önerilen Monte Karlo eniyileme yordamları ölçeklenebilir ve verimli bir şekilde işlemekte ve bileşen dağılımlarından örneklem elde edilebilen herhangi dağılım aileleri için sonuç vermektedir. Ek olarak, bu yaklaşımla parametrik bir Bayesçi risk kullanılarak kestirim doğruluğu ve iletişim maliyeti arasındaki ödünleşim nicemsel olarak gözlemlenebilmektedir.

Anahtar Kelimeler: İmeceli Sinyal ve Bilgi İşleme, Dağıtık Kestirim, İletişim Kısıtları Altında Çıkarsama, Rasgele Alanlar, İleti Geçme Algoritmaları, Çizge Modelleri, Algılayıcı Mimarileri, Monte Karlo Yöntemleri.

ACKNOWLEDGMENTS

There are many people whom I shall acknowledge for their help and support during my Ph.d. studies. I thank to my advisor Prof. Kemal Leblebicioğlu for the enthusiasm and encouragement he has provided throughout this period, his knowledge and experience he has shared, his "never no" policy to me and his always left-open door even at the weekends. Without any less I want to thank Dr. Müjdat Çetin of Sabancı University who has led to transform this research to a whole new level with his vision. I am gratefull for his guidance, the motivation and the opportunity he has provided in Signal and Information Processing Laboratory of Sabancı University. He is the best teacher I have ever met with his keen ability of importance ordering and delicately keeping the focus which reflect his solid knowledge and researcher identity. Together with his personal traits it is an ivaluable experience to work with him.

I want to acknowledge Dr. O. Patrick Kreidl of MIT for his sincere help throughout many discussions on the subject. Besides, I had the opportunity to contact with Prof. Pramod Varshney of Syracuse University. I want to thank to him for helping me receive one of his old works and sharing his vision while he was in Turkey during spring of 2008.

I want to thank Prof. Uğur Halıcı and Prof. Yasemin Yardımcı Çetin of METU for their solicitude during the whole period and generously letting me over-allocate their time for the Progress Committee meetings. Fellow grad. students Hüseyin Yiğitler and Sedat Doğru has always been patient while listening my never ending intricate beurocratic issues and related complaints. I also want to thank fellow RA Fatih Bayramoğlu of METU EEE, Dr. Ozgur Ertuğ of Gazi University, and Tansu Filik and Serkan Karakütük of METU Sensor Array and Multi-channel Signal Processing Group for their friendship and support throughout.

All the staff of Vision and Pattern Analysis Laboratory of Sabancı University (SU) has been supportive and the affiliated faculty has been generous in mentoring. I indebt special thanks to Dr. Hakan Erdoğan and acknowledge letting me use the computational resources of the Speech Processing Group. I also want to acknowledge Barış Tümer of IT Dept. of SU for providing constant support in cases of emergencies. I want to thank Octavian&Diana Soldea,

Ozben, Sureyya, Emrecan and the FENS2020 staff. I owe you an appologize for the “long” seminars I gave during the meetings. I also want to acknowledge Sarper Göktürk of SU for reviewing parts of the thesis.

I apologize to the people whom I undeliberately fail acknowledging due to the already missed time schedule. I also kindly require understanding for any typos and inaccurate sentences left. I want to acknowledge the partial support provided by the Scientific and Technological Research Council of Turkey under grant 105E090, and by the European Commission under grant MIRG-CT-2006-041919.

Finally, I want to thank to my family and my love Şükran D. Çavdar for their support and love that keep me alive and provide me the motivation to make any accomplishments.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Statistical inference captures a variety of problems in signal and information processing in which a system is expected to draw results on a physical phenomenon of interest based on the observations that contain relevant information together with uncertainities. Detection and Estimation Theory covers such problems where the phenomenon is described by a quantity that can take values from a well defined set and we are able to provide a probabilistic model describing a characterization of the uncertainities of or between this quantity and the observations induced.

As the problem size grows, the performance of the physical systems which realize the solutions of such problems heavily depends on the constraints driven by the available resources of which limited energy, communication bandwidth and computational power are the most eminent. Consider, for example, a typical scenario that arise in wireless sensor network applications in which there are many such observations due to multiple quantities of interest, collected by devices of various modes and located at various places [1]. In addition to the spatially distributed nature of the underlying physical system, it is often the case that there is no infrastructure provided by the environment and the components rely on limited energy stored in batteries [2].

It is possible to handle various modes of measurement in a unified manner under a probabilistic framework. Since the observations are characterized by corresponding likelihood functions, fundamental rules of the Probability Theory yield a natural fusion scheme through these functions and apparent dependecies of the variables. However, a straightforward treatment of the corresponding inference problem yields processing schemes that become extremely expensive with an increase in the number of variables. Considering the inherent computational

power limitations of the system together with the high volume of data, scalability is one of the key characteristics of feasible solutions for which to be achieved necessary computations are systematically reduced and preferably carried out in a distributed manner while providing reasonably accurate results. In addition, the apparent need for communications has to be supported by the bandwidth limited links provided by the underlying communication network structure. Another fundamental issue is the limited energy budget and it is often the case that transmitting bits is far more costly than computing them in terms of energy dissipation [3]. Collaborative signal and information processing is involved extracting useful information from observations in such a scenario while taking the aforementioned tradeoffs into account [4].

In this thesis we employ a probabilistic framework and consider decentralized estimation. We are interested in processing schemes that match a certain description of communication constraints including the set of available links and their capacity. We consider the trade of between the estimation accuracy and the cost of communications in particular.

Graphical models together with message passing algorithms has proved useful for decentralized statistical inference in sensor networks (see e.g. [5] and the references therein). In this framework, a graph representation that reveal the information structure beared by the set of variables associated with the probabilistic model of the problem is provided together with message passing algorithms that yield efficient statistical inference. On the other hand, the underlying system maintains a communication network which renders a distinct graph in terms of available links. After mapping the former onto the later, a decentralized inference scheme is obtained that can be realized by the physical system provided that the capacity of the communication network supports that required by the messagings. However it is hard to solve the problem given the available links and capacities together with cost of transmission over them.

We consider two classes of decentralized estimation strategies that cover a broad range of communication constraints and yield graph representations. The vertices of the graphs correspond to platforms and the edges are subsets of available links that render a Directed Acyclic Graph for the first, and an Undirected Graph for the second. Associated with each edge is a set of admissible sysmbols that comply with the link capacity. The processing strategies represented by the graphs are such that a tracktable Bayesian objective is achieved. Having the

Bayesian cost penalize the estimation errors as well as the transmissions, a communication constrained design setting is obtained through constraining the set of feasible strategies by the graph. We explore the solution strategies for the design problem exhibiting scalability and efficiency.

1.1 Motivation

1.1.1 Networking Sensor Platforms

Sensor Networks have provided a technology base for large scale distributed sensory systems constituted of platforms that have limited capabilities. Mounted on each device are circuitries for communication, processing and sensing as well as batteries for energy storage. A conceptual block diagram of the hardware is given in Figure 1.1. Simple devices such as thermometers and more complicated devices similar to one which performs multi-array processing and find the direction-of-arrival of an incoming signal are referred to as sensors as well as peripherals such as GPS receivers and gyroscopes.

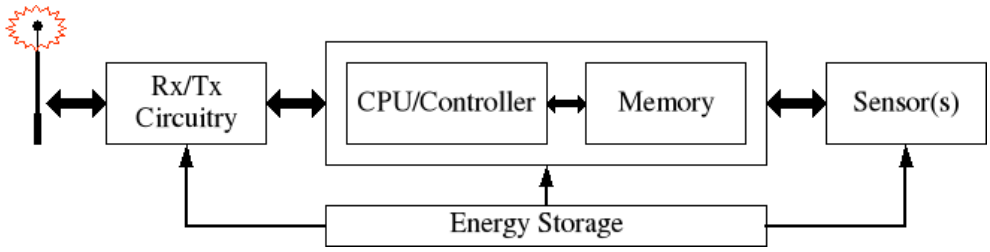
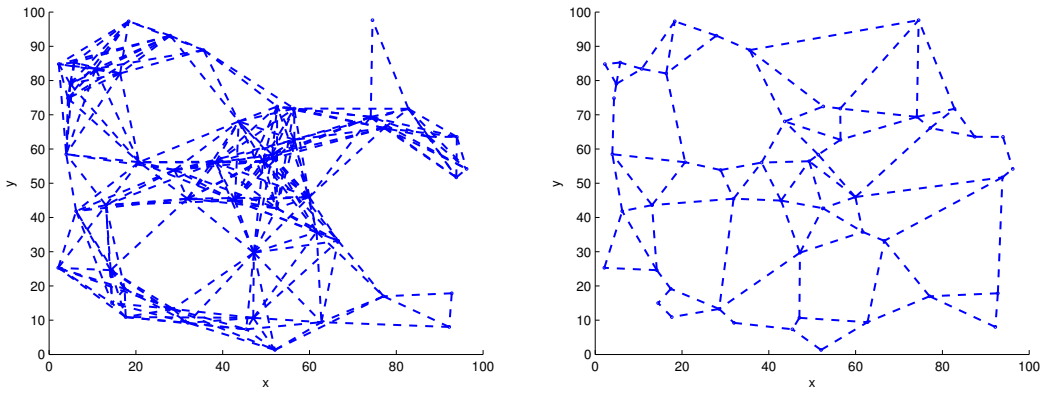


Figure 1.1: Conceptual block diagram of a sensor platform hardware.

The communication circuitry is in accordance with the physical layer specifications for wireless ad-hoc networking (e.g. IEEE 802.15.4) which can maintain limited capacity links with the peers that fall in a certain range. Connections established among nearby platforms yield a graph in which the edges are communication links maintained by the circuitries. Also associated with each edge is the capacity which depends on many factors including the condition of the channels[1] and the transmitter powers. The connectivity of the graph can be controlled by adequate adjustment of the power levels [6]. Over a connected topology, it is possible to transmit messages between any two nodes through appropriate multi-hop routing protocols in

[1] The condition of the channels refers to the issues affecting the capacity such as the fading the regime (e.g. Rayleigh or Rician) and interference considering the multiple-access medium.

sensor networks. Moreover, the network layer can provide services for aggregating or broadcasting data as well as maintaining higher level topologies over the connectivity graph referred to as multi-tier architectures (see, for example, [7][8][9]). An example deployement scenario is illustrated in Figure 1.2 through the connectivity graph in Figure 1.2a. In Figure 1.2b, the corresponding Gabriel Graph is presented. Gabriel Graphs are geometric graphs which are considered by routing protocols. Hence, it is possible for the network layer to provide a view of the network to the upper layers in the OSI model as the Gabriel Graph itself or clusters constructed through in which an edge does not necessarily correspond to a 1-hop link.



(a) Connectivity Graph (b) Gabriel Graph

Figure 1.2: We consider an example deployement in which 50 platforms are randomly spread over a square region of edge length 100 meters. Each platform is assumed to be able to communicate with peers located within a distance of 30 meters. We illustrate the corresponding Connectivity Graph (a) in which the edges are 1-hop links, and the Gabriel Graph (b) which is considered by some routing algorithms. Note that the Gabriel Graph is not necessarily a subgraph of the Connectivity Graph and the edges do not necessarily correspond to 1-hop links.

Let us consider the OSI model with a simple twist of treating all the entities above the network as a single application layer which bears the processing tasks (Figure 1.3). Hence, it is possible to consider the network as an entity which offers some data services that respond upon requests of the processing block and provide input from the other platforms in the network. From the processing point of view, the underlying ad-hoc network may appear in various topologies which are not necessarily subgraphs of the connectivity graph. Moreover, the edges do not necessarily correspond to 1-hop neighbors.

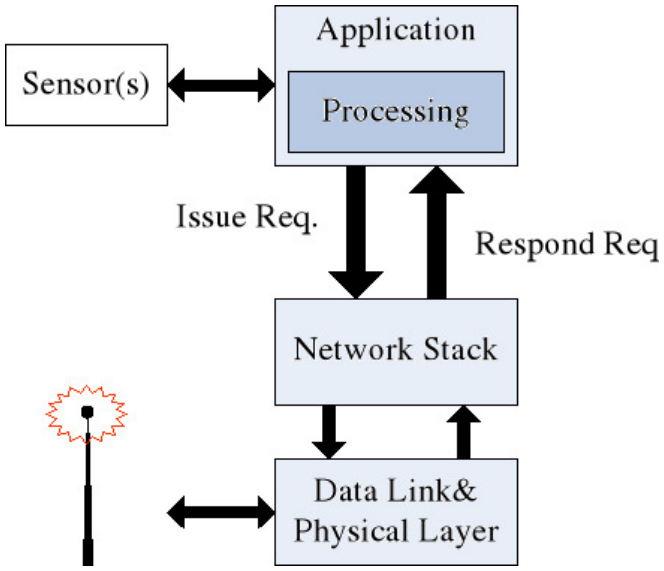


Figure 1.3: An example architecture for the sensor platforms. The network layer reponds to the requests issued by the application which bears the processing entities. Hence the input to the processing block is provided both from the sensors and from other platforms through the network layer.

1.1.2 The Cost of Communications

The sensor platforms rely on stored energy and prudent use of it is crucial in order to extend the operational lifetime of a sensor platform. Therefore, energy-aware operation at all levels is crucial for which the energy cost of communications is of concern among others.

The energy consumption of the communication circuitry of a sensor platform is considerably high in comparison with the sensors and the processing unit when not turned off[2]. The communication circuitry is constituted of a processor for tasks such as coding-decoding, power amplifiers and the antenna. The energy dissipation depends on factors such as the coding scheme, modulation, channel fading regime and the output power (see, for example Chp. 2 of [10]). The radiated power decreases with $1/d^2$ and $1/d^4$ in free space and near ground respectively, due to the ground reflecting waves [3]. The relaibility of communication depends on the received signal power. Hence, it is possible to vary the maximum hop distance by adjusting the output power and together with code selection and voltage/frequency selection for the communication processor, topology control and power-aware communications are

[2] For example, a Rockwell WINS node dissipates 383.3 mW of Power when the communication circuitry is removed and both the sensors and the processing unit is active. Setting the transceiver in idle mode results with 727.5 mW which is nearly twice that of the former case. Switching the transceiver to the receiver mode and highest power transmitter mode results with 751.6 mW and 1080.5 mW of power dissipation respectively [2].

achieved [11].

A first order radio model in order to estimate the energy consumption for both transmitting and receiving k bits at d meters distance is given in [12]. Consider Figure 1.4. For the transmission of k bits over a distance d requires $E_{elec} * k + e_{amp} k d^2$ joules at the transmitter and $E_{elec} * k$ joules at the reciever where $E_{elec} = 50nJ/bit$ and $e_{amp} = 100pJ/bit/m^2$.

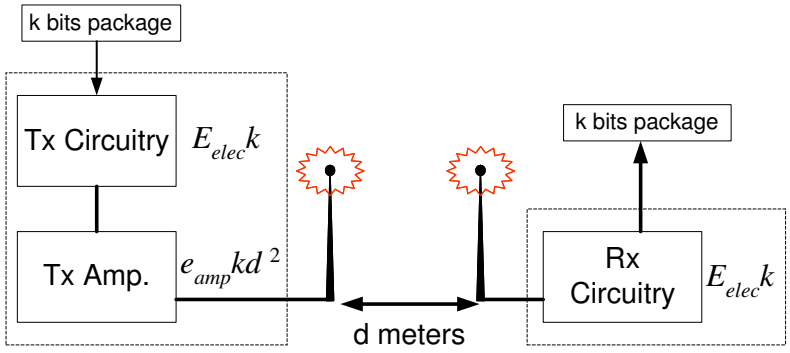


Figure 1.4: First order radio transmission model as appeared in [12].

Considering this model, it is possible to estimate the energy cost of a k bit packet over the path between two nodes maintained by the routing protocol from the network entity. These paths correspond to available links and together with their capacities, the energy cost constitute a considerably broad set of the communication constraints.

1.1.3 Random Field Estimation

A motivating example for decentralized estimation under communication constraints is random field estimation. Consider the scenario in which a number of platforms are spread over a region to monitor turbulent flow. Mounted on each platform is a sensor that measures the velocity of the flow along the direction of its normal at that spatial location. The magnitudes at proximal locations are similar and it is possible to capture this fact in terms of a Markov Random Field (MRF) representation with a Gaussian joint density for turbulent flow (Chp. 12 of [13]). Statistical models based on covariance descriptions of spatial variables are also used for geostatistics data [14] such as temperature measurements over a field (Chp. 1 in [15]).

The platforms are wirelessly networked in an ad hoc manner such that communication links are maintained which render a connected planar graph such that any messages sent between

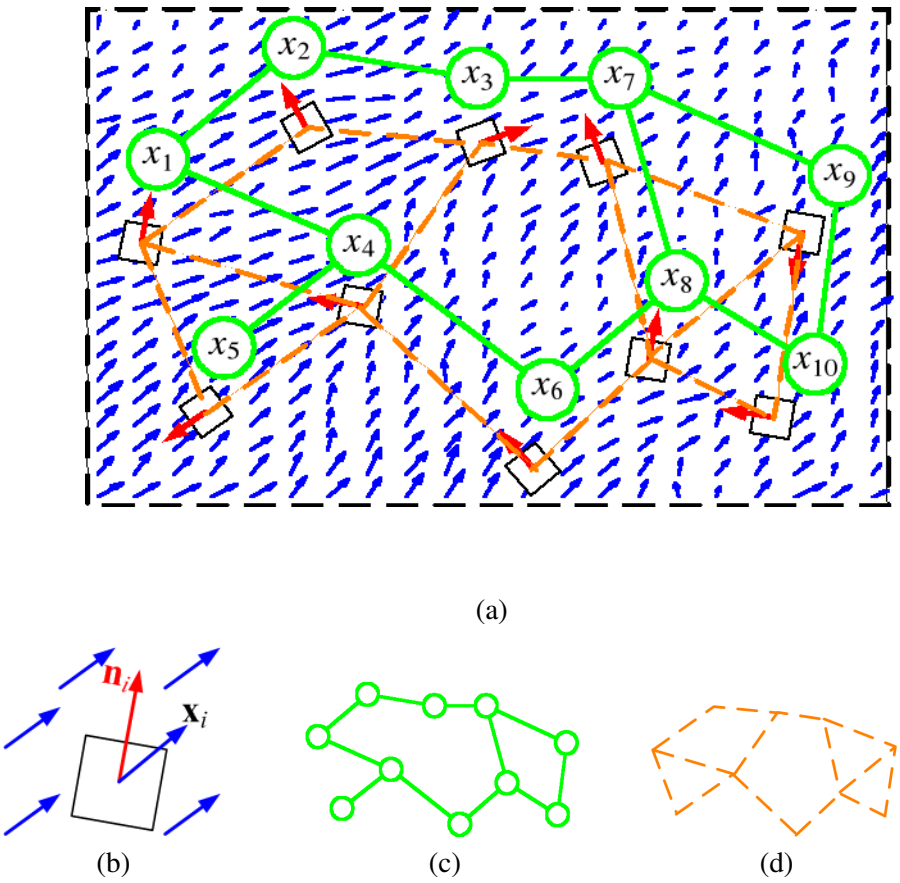


Figure 1.5: Illustration of a random field estimation scenario: (a) Sensor platforms (black squares) spread over a region to monitor turbulent flow presented by the blue vector field. (b) The observation model: Each sensor measures the magnitude of the flow at its spatial location along the direction of its normal, i.e. $y_i = x_i + n_i$ where $n_i \sim \mathcal{N}(0, \sigma_i^2)$. (c) An example Markov Random Field prior for $x = (x_1, x_2, ..., x_{10})$. (d) The underlying network structure rendered by the available links.

two nodes will travel along paths in this graph. The components of this scenario is illustrated in Figure 1.5.

We are particularly interested in the estimation task itself and do not consider conditions on how the result is stored by the network. From the global point of view, it is possible to treat the example network constituted of N platforms evaluating a function $\gamma : \mathrm{R}^N \to \mathrm{R}^N$ such that an estimate for the random vector of concern $x \in \mathrm{R}^N$, i.e. $\hat{x} = \gamma(y)$, is output. In one possible case, after the evaluation of γ is completed, one of the platforms hold all the components of the result, i.e. $\{x_i\}$ for $i = 1, 2, ..., N$. It can also be the case that the fields of x are stored at distinct platforms. For both cases, it is possible to treat the sensor network as a distributed database system from which queries including that for the results can be responded through

appropriate schemes (see e.g. Chp.6 of [16]). This abstraction has the advantage of isolating the inference scheme in the system architecture and enabling detail hiding among entities. Consider, for example, an application in which the rate of requests regarding the value of $\hat{x}$ is well below the rate at which it is produced. Another example is a case in which the user is interested in the occurance of a certain event defined by a set of conditions on x and in this respect, the estimation result is an intermediate level output. Therefore we restrict the discussion with schemes which result with the network to store $\hat{x}$ without any conditions on the association between the platforms and the components of $\hat{x}$.

A straightforward estimation strategy is to have each platform i estimate its local variable x_i based on y_i. This so-called myopic approach does not demand any communications and it scales well with the number of platforms and hence variables. Considering the dependency of flow magnitudes at nearby locations, the estimation accuracy will be worse than that the conventional approach achieves in which the inference is based on all the observations, i.e. $y = (y_1, y_2, ..., y_{10})$. One of the major drawbacks is that it is required to constantly forward all the observations to a center without taking the corresponding energy dissipation into account and creating a tremendous traffic in the network. Second, the center platform undertakes a huge amount of computations (consider, e.g. the number of platforms on the order of thousands) while the rest of the platforms do not contribute to the evaluation of γ.

We are interested in schemes that fall between the myopic and the conventional approach in the sense of achieving a reasonable estimation accuracy while arising an acceptable communications and computations demand in an energy-aware manner. Since communications is the most energy consuming action, we consider introducing its cost into the design problem involving specification of γ.

1.1.4 Target Localization in Tracking

Target tracking in a surveillance scenario has been another envisioned application of sensor networks [17]. Objects moving in the coverage of platforms spread over a region are of concern. After an object is detected, a record of its location at successive time steps is required for the time window it stays in the coverage. Conventionally, the uncertainities in this scenario has been handled in a Bayesian framework and the localization at a certain time instant is performed through the posterior distribution of the location based on all the relevant ob-

servations since detection. Assuming a Markov Chain model for the target maneuver, it can easily be shown that the posterior is obtained by recursive Bayesian filtering of observations collected at consecutive time steps which specializes to the celebrated Kalman filtering in the case of Gaussian uncertainities and linear dynamics and particle filtering in the case of sampling based and Monte Carlo approximations to the distributions and the integrations involved respectively (Such approximations are needed in order to handle non-linear dynamics and/or non-Gaussian uncertainities) [18].

The recursive filtering involves an update stage in which the current observations are fused with the track information and the posterior density of location is obtained, and a prediction phase in which a predictive distribution is obtained based on the posterior and the dynamic model of the target. Hence, the statistical inference problems involved are in multiple levels such as detecting a target and initiating a track, based on the previous record and the current observations updating the track with the current estimation of the target location and a predictive distribution. In the multiple target case, it is also necessary to discriminate the observations from different targets and associate them with the tracks accordingly, which is referred to as the data association problem [19]. All these tasks are required to be carried out in accordance with the in-network processing paradigm [20] in sensor networks. This is possible by exploiting the local structures that the problem exhibit and distributing the processing accordingly.

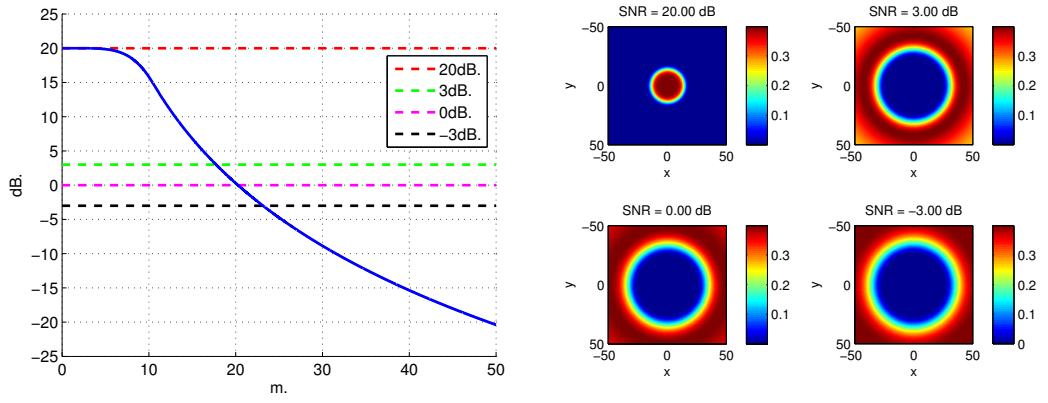
In our motivating example, at a certain time instant, not all the sensors collect observations of a certain quality. The most informative measurement is likely to be made by the platform that is nearest to the target. The set of measurements which could provide additional information is collected by a cluster of platforms located in the vicinity. It is possible to select the platform that is expected to make the most informative measurement in the next step based on the target dynamics, observations models and the predictive density captured by the track [21]. A collaborative tracking scheme is obtained through having the current leader platform, at each time step, hand off the track to the next leader which will be responsible for the next update [22]. It is also possible to arrange clusters for the leadership of each platform and invoke platforms to contribute to the track update (and hence to the localization of the target)[23].

Consider the case in which acoustic intensity sensor are mounted on the platforms. At time

instant n, the measurement of a sensor located at l_s due to a target at $l_t[n]$ is given by

$$y_s[n] = Af(d[n]^{\alpha}) + \omega \tag{1.1}$$

where A is the nominal intensity of the acoustic signal emitted by the target, $d \triangleq \|l_s - l_t\|_2$ is the Euclidean distance between the sensor and the target, f is a monotonically decreasing function with $f(0) = 1$, α is the attenutation coefficient and ω is a zero mean Gaussian random variable with variance σ^2 [24]. The attenuation coefficient, depending on the physical environment, varies between 2 and 4. It is apparent that such a measurement bears information on the displacement but it is possible to fuse information from distinct platforms and infer the position of the target. The monotone behaviour of f reveals that the SNR decreases as the distance between the sensor and the target gets larger. In other words, the contribution of observations collected by platforms located farther than a certain distance is negligible. We present the displacement versus SNR for unity noise variance and $A = 10$ in Figure 1.6a. The maximum SNR is achieved when the displacement is zero and platforms located more than 20 meters to the target can not provide useful information (Figure 1.6b).



(a) distance vs. SNR

(b) Observation Likelihoods

Figure 1.6: We consider an acoustic intesity sensor located at the origin and select the nominal amplitude $A = 10$, noise variance of unity and attenuation coefficient $\alpha = 2.6$. (a) the Graph of the sensor-target displacement versus SNR. Note that the sensor has an SNR of $20, 3, 0$ and -3dBs for displacements of $0, 17.75, 20.772$ and 23.152 meters. (b) Likelihoods of the target location x corresponding to the observation $y_s = Af(d^{\alpha})$ for $d \in \{0, 17.75, 20.772, 23.152\}$. Note that, the likelihood mask for an observation made under condition of below 0dB SNR reveals only the information that the target is somewhere farther than 20 meters.

In Figure 1.7 we present a localization scenario with a sensor network constituted of platforms equipped with acoustic intensity sensors. The boundaries of the sets of points that are closest to sensors render a voronoi tesellation of the region. It is possible to predict a target' s location

and hence the patch it would reside, based on the track information and have the associated platform assigned as the new leader and handed off the track information. The influence region of the target is a circle that is centered at the target location with a radius of 20 meters considering the discussion above. The platforms that fall within this region collect observations that would increase the accuracy of the location estimation. A straightforward approach would be forwarding the observations over the available links. Such a communication structure can be represented by a directed ayclic graph implying the direction of information flow from the parentless nodes to the leader node. The concerns on the cost of communications yield another approach in which only the measurement of the leader node is used to update the track [20]. The communications required regarding only the localization is still not zero since it is necessary to execute a protocol to make the nodes in the target influence region aware that the target they are sensing is led by another node [23].

The concerns on the trade-off between the utilization of communications and the accuracy of localization becomes more complicated considering that the exact target position is not known and information is required from all nodes that fall in the convex hull rendered by the influence regions of all possible target locations within in the voronoi patch. It is apparent from Figure 1.7 that some of these nodes can not provide any useful information that worths the cost paid for communications. Similar to the discussion in Section 1.1.3, we treat the estimation task as the evaluation of a function γ that maps the space of observations to the space of locations. For the functions γ yielding a certain form that enables distributed evaluation, it is possible to find an information collection mechanism which yields a relatively reasonable network traffic through the construction of a graph representation for the communication structure together with local functions for the nodes that constitute γ restricted on a discretization of both its domain and range sets [25].

On the other hand, given a certain communication structure, the question of how to design γ that achieves a reasonable estimation accuracy for a reasonable communication cost while taking the communication constraints into account including the limited capacity of the links and the energy cost required to transmit over them remains. For example, over the example communication structure presented in Figure 1.7, the nodes can perform a variety of computational schemes that would result with the position of the target with a certain accuracy. We are interested in design procedures that would allow for graceful degradation of this accuracy when it is preferred to reduce the utilization of communications in order to lower power

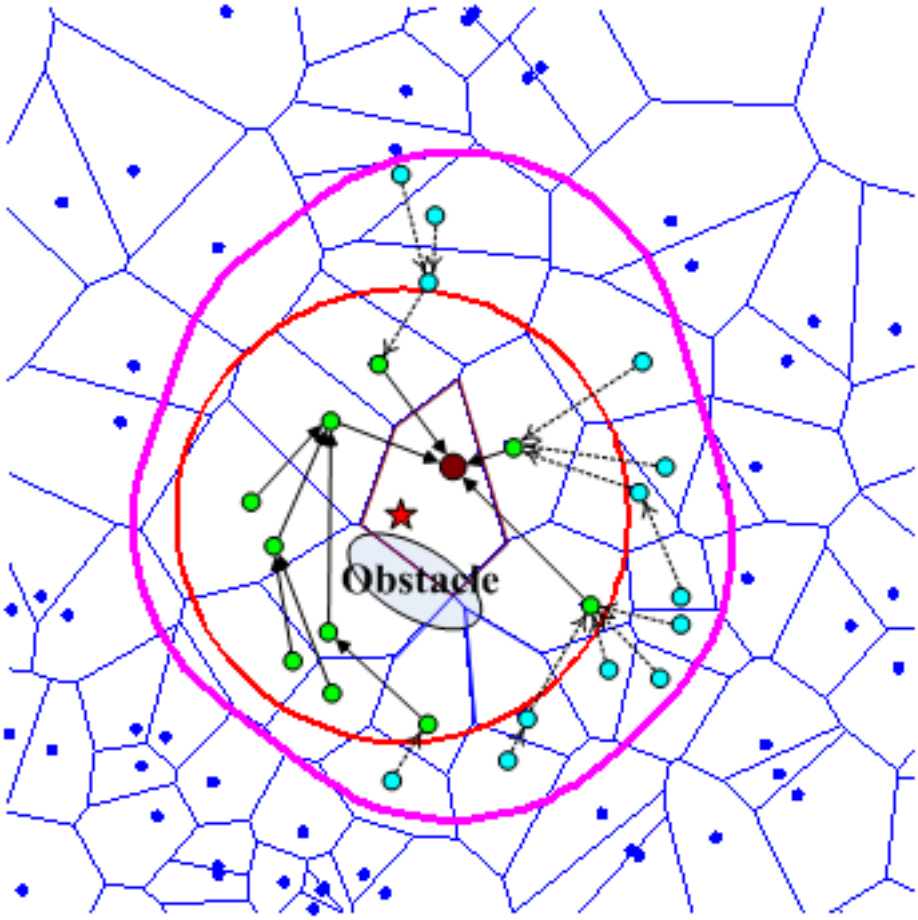


Figure 1.7: Illustration of a target localization scenario in sensor networks: A target (red star) is located in the brown voronoi patch and the associated node (brown dot) is assigned as the leader. The influence region of the target is bounded by the red circle centered at the target location with a radius of 20 m.s. The union of influence regions of possible target locations in the brown patch renders a convex hull (region within the magenta boundary). The nodes that fall in the influence region of the target (green dots) would contribute to the localization task. A possible communication structure over the available links is shown by solid arrows rendering a directed acyclic graph in which starting from the parentless nodes and ending at the leader node, upon receiving messages from its parents, each node send messages to the child nodes. Notice that the underlying network topology might not allow for nearest-neighbor communication in a single hop due to possible obstacles in the environment such as rocks or walls. Moreover, since the target location is not known exactly, it can be the case that information from all nodes that fall in the convex hull, including the non-informative ones (cyan dots), is required. A possible communication structure is the extension of the former with the dashed arrows.

dissipation and hence increase operational lifetime of the network.

1.2 Contributions

In Sections 1.1.1 and 1.1.2, we have introduced an overview of networked sensing with an emphasis on the communication constraints including the energy cost as well as the layered architecture together with the distinction between the underlying channels and the links maintained for the processing entities by the network layer. The examples in Sections 1.1.3 and 1.1.4 present scenarios in which an estimation task is to be carried out in accordance with the in-network processing paradigm and underline the distinction between the information structure exhibited by the probabilistic model of the problem and the communication structure that the system relies upon.

We are concerned with the design of decentralized estimation networks under the communications constraints including the topology and the capacity of the links as well as the cost, possibly due to energy consumption, of transmitting over them. We consider two classes of decentralized strategies that are composed of local communication and computation rules and yield graph representations in which the nodes represent the platforms and the edges correspond to available links. Under reasonable assumptions, these strategies yield tractable Bayesian objectives which capture penalty for inference errors as well as cost of communications. A rigorous design problem is obtained in the form of a constrained optimization problem after employing the Bayesian risk and constraining the feasible decentralized strategies to the space defined through the graphs. The two classes, together with the structures that the solution exhibits have been recently studied in [26] and [27] in the context of decentralized detection in which a decision from a finite set is to be made.

We adopt this framework for the decentralized estimation in which a decision from a denumerable set is of concern. Similar to the detection case, a Team Theoretic investigation for each class yields an iterative solution which starting from an initial strategy converges to a set of person-by-person optimal local rules while decreasing the Bayesian risk.

However, due to the fact that the estimation problem is involved with denumerable sets, the iterations contain expressions with integral operators that have no closed form solutions in general. We introduce an approximation framework utilizing Monte Carlo methods under

which the particle representations of the local rules and approximate computational schemes for the operators replaces the original expressions. As a result, the iterative solutions turn to Monte Carlo optimization methods which also maintain the appealing features of the former.

We provide the following benefits:

- The framework we utilize enables us to consider a broad range of communication and computation structures for the design of decentralized estimation networks. Since the approach is Bayesian, it is possible to introduce prior information and the cost of communications captured by the Bayesian risk provides a graceful degredation for the estimation accuracy when it is required to communicate less and hence energy-aware operation. Under reasonable assumptions the optimization procedures scale with the number of platforms as well as number of variables involves and also admit message passing interpretations which render them fit for network self-organization.

- The Monte Carlo optimization schemes we propose maintain all the features above and scale with the number of samples required. In addition, they are valid for any set of distributions provided that samples can be generated from the marginals. Hence they are not restricted with standard distributions such as Gaussians and do not require coordinated sampling schemes such as Gibbs sampling.

- Employing a Bayesian risk that captures penalties for both the estimation errors and the utilization of communications, we are able to address the tradeoff between the communication cost and the estimation accuracy through considering the performance points while varying a parameter that admits the interpretation of equivalent estimation penalty per unit cost of communications.

The main contributions of the thesis are the Monte Carlo optimization schemes summarized in Algorithm 4 presented in Section 3.4 and Algorithm 5 presented in Section 4.3. We provide a design methodology for decentralized estimation networks under a broad range of communication constraints with the benefits discussed.

1.3 Thesis Organization

The organization of the thesis is as follows; in Chapter 2 we introduce notation and background that provides a common language for the rest of the thesis. We start with an overview of mathematical preliminaries and introduce the terms and notation used throughout. Then we introduce the two classes of decentralized strategies in the context of detection in Section 2.7. We discuss the online communication and computation schemes represented by directed and undirected graphs together with the construction of the constrained optimization problem regarding communication constrained inference. Then the Team Theoretic investigation and the iterative solutions are given.

In Chapter 3, we introduce decentralized estimation strategies with a directed acyclic communication structure. We discuss the conditions of optimality and point out that unlike the detection case, we are not able to achieve useful representations of rules local to nodes. It is also not possible to evaluate the operators that appear in the iterations. Then we introduce the Monte Carlo optimization framework for this class of strategies in Section 3.4 with the corresponding strategy optimization scheme. Through examples, we present the features of the proposed approach and address the tradeoff between the estimation accuracy and utilization of communication.

In Chapter 4 we consider the class of strategies over undirected graphs. We similarly discuss the conditions of optimality and the implications exhibited for this case. We present the approximation framework for the undirected case in Section 4.3 together with the Monte Carlo optimization scheme. We discuss that the proposed approximation framework is also valid for two-tiered networks in which a higher level undirected strategy is performed among nodes of a lower level directed acyclic graph.

Finally we present future extensions of the work including possible research directions in Chapter 5.

CHAPTER 2

BACKGROUND

The material in this chapter is arranged with the intention of providing a background on decentralized Bayesian inference motivated by sensor network applications. Starting with a review of the conventional centralized setting a progressive treatment of the problem is developed by pointing out inherent intricacies regarding Bayesian inference and introducing constraints that arise in a decentralized setting.

We begin with presenting mathematical preliminaries in order to introduce the nomenclature used throughout. In Section 2.1.1, an overview of basic definitions in Graph Theory is given which are often referred in the proceeding discussion. In the following section, elements of Probability Theory are introduced in a similar fashion. The focus is on the characterization of a number of variables through their joint probability density function and the properties that this characterization exhibits as a result of certain relations among the variables such as independence and conditional independence. In addition, the law of total expectation and its utilization for statistical Bayesian inference is given. In Section 2.2 it is pointed out that a number of conventional Bayesian detection and estimation problem can be handled through this law which results a variational form for the detector or estimator that acieves minimum risk.

A straightforward utilization of Bayesian inference exhibits exponential increase in complexity with the problem size. Probabilistic graphical models, reviewed in Section 2.3, provide a means to both represent the problem structure and exploit possible sparsities for efficient inference through message passing algorithms. We explain how this perspective is used in order to handle decentralized inference problems in Section 2.4.

Another subtlety of Bayesian inference is the evaluation of integrals that arise in the case of

continous random variables. In Section 2.5 we present Monte Carlo methods which are numerical approximations that are successfully employed for expressions involving expectations of functions over denumerable sets. As an example, particle filters are algorithms obtained by utilizing Monte Carlo approximations in recursive Bayesian filtering.

We start discussing the communication constrained setting in Section 2.6 and elaborate approaches addressing the tradeoff between the utilization of communication and estimation accuracy. A remarkable design approach that exhibits graceful degredation of the global inference performance has been recently introduced in the context of decentralized detection and we summarize this perspective Section 2.7. Specifically, we consider [28] together with [27],[26] and [29], but unlike the problem setting considered in these work we assume error-free communication links and present the overview accordingly. Two classes of decentralized inference strategies are of concern. The first class, discussed in Section 2.7.1, can be represented with a directed acyclic graph and yields a rigorous design problem in the communication constrained setting. The second class is the two stage decentralized strategies over undirected graphs and provide similar benefits as presented in Section 2.7.2. In the following chapters, we utilize these strategies in the context of estimation.

2.1 Mathematical Preliminaries

In this section basic elements from graph theory together with probability theory and statistics are presented in order to establish a common notation and nomenclature throughout. Some notation for sets, index sets and functions over multiple arguments is introduced first.

Given a set $\mathcal{V}$, we denote with $\setminus$ the difference operation. Given $i \in \mathcal{V}$ the complementary of i with respect to $\mathcal{V}$, i.e. the set $\mathcal{V} \setminus \{i\}$, is denoted by $\setminus i$ where it is obvious from the context that the universal set of concern is $\mathcal{V}$. Let $|.|$ denote the set cardinality. $\mathcal{P}(\mathcal{V})$ is the power set of $\mathcal{V}$, i.e. the set of all subsets of $\mathcal{V}$, and it holds that $|\mathcal{P}(\mathcal{V})| = 2^{|V|}$.

We use $\{.\}$ to denote that the elements are unordered whereas $(.)$ implies an ordering. When a recursive operation based on an index set is of concern, we assume that the choice of elements follows the trivial ordering. In this respect, let $\otimes$ denote consecutive Cartesian Products, then, e.g. $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ and $\mathcal{X} = \bigotimes_{i \in \{1,2,3\}} \mathcal{X}_i$ are synonymous noting that $\otimes$ is based on $\{1, 2, 3\}$.

Given a set $\mathcal{A} \subset \mathcal{V}$ where $\mathcal{V} = \{1, 2, ..., N\}$, and an indexed set $X = \{X_1, X_2, ..., X_N\}$, $X_{\mathcal{A}}$ denotes a subset of X given by $X_{\mathcal{A}} = \{X_i | i \in \mathcal{A}\}$. Similarly $X_{\setminus i}$ denotes the set $\{X_j | j \in \mathcal{V} \setminus \{i\}\}$ and $X_{\setminus \mathcal{A}}$ is $\{X_j | j \in \mathcal{V} \setminus \mathcal{A}\}$.

Given an n-tuple $(x_1, x_2, ..., x_N)$ we denote with $[.]_i$ the i^{th} component of its n-tuple argument, e.g. $[(x_1, x_2, ..., x_N)]_i = x_i$. Similarly, given a group of variables $\{x_1, x_2, ..., x_N\}$ such that $x_i \in \mathcal{X}_i$, we denote with $[.]_{X_i}$ the element of its argument that takes values from the set X_i, e.g. $[\{x_1, x_2, ..., x_N\}]_i = x_i$.

Finally, we sometimes let any permutation of arguments of a function to refer to the same value, e.g. given $f(x_1, x_2)$ and $f(x_2, x_1)$ refer to the same value when it is obvious from the context that $f(x_1, x_2)$ and $f(x_2, x_1)$ denotes $f_{X_1,X_2}(x_1, x_2)$ and $f_{X_2,X_1}(x_2, x_1)$ respectively such that $f_{X_1,X_2}(x_1, x_2) = f_{X_2,X_1}(x_2, x_1)$ holds. In this sense if $\mathcal{A} = \{1, 2\}$ then $f(x_{\mathcal{A}})$ denotes any of $f_{X_1,X_2}(x_1, x_2)$ or $f_{X_2,X_1}(x_2, x_1)$.

2.1.1 An Overview of Graph Theory

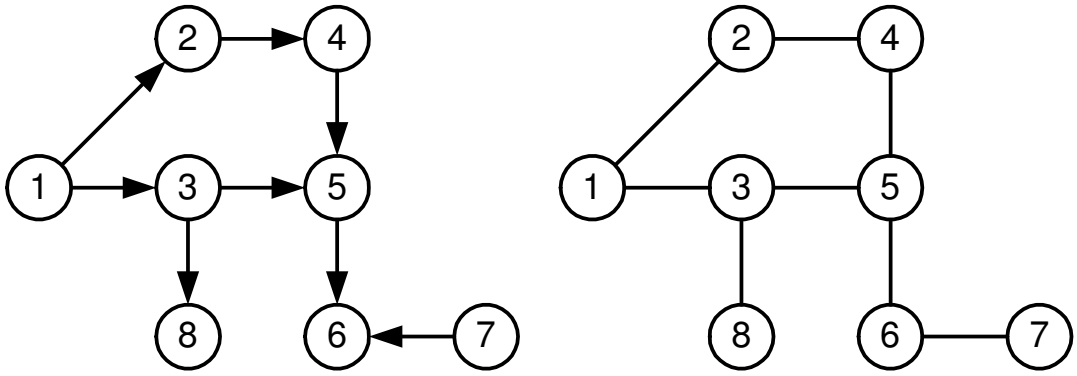
A *graph* is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of *vertices* (or *nodes*) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of *edges*. $\mathcal{G}$ is said to be a *graph on* $\mathcal{V}$ whereas an edge $(i, j) \in \mathcal{E}$ connects vertices $i, j \in \mathcal{V}$, initiating from i and terminating at j.

Two vertices $i, j \in \mathcal{V}$ are *neighbours* if they are connected, i.e. (i, j) or $(j, i) \in \mathcal{E}$. Hence the set of neighbours of a vertex j is defined by $ne(j) \triangleq \{i \in \mathcal{V} | (i, j) \in \mathcal{E} \vee (j, i) \in \mathcal{E}\}$, in other words all vertices $v \in \mathcal{V}$ *adjacent* to j. The graph $\mathcal{G}$ is *simple* if it does not contain any self-adjacent vertices, i.e. $(v, v) \notin \mathcal{E}$, $\forall v \in \mathcal{V}$, which we will be dealing with throughout the thesis.

The set of *parents* of a vertex j, denoted by $\pi(j)$, is the collection of vertices that are connected with j through edges terminating at j, i.e. $\pi(j) \triangleq \{i \in \mathcal{V} | (i, j) \in \mathcal{E}\}$. Similarly the set of *children* of a vertex j, denoted by $\chi(j)$, is constituted of vertices that are connected with j through edges initiating from j, i.e. $\chi(j) \triangleq \{k \in \mathcal{V} | (j, k) \in \mathcal{E}\}$.

The *indegree* of a vertex j is the number of its parents, i.e. $|\pi(j)|$ and similarly the *outdegree* of j is the number of its children, i.e. $|\chi(j)|$. The *degree* of a vertex is defined as the number of its neighbours, i.e. $|ne(j)|$.

A graph $\mathcal{G}$ is said to be directed if there is no pair of edges connecting the same vertices such that one terminates at the vertex the other initiates from, i.e. if $(i, j) \in \mathcal{E} \Rightarrow (j, i) \notin \mathcal{E}, \forall (i, j) \in \mathcal{E}$. In Figure (2.1a), we illustrate an example directed graph on $\mathcal{V} = \{1, 2, ..., 8\}$ with an edge set $\mathcal{E} = \{(1, 2), (1, 3), (2, 4), (3, 5), (3, 8), (4, 5), (5, 6), (7, 6)\}$. Similarly, a graph



(a) (b)

Figure 2.1: Illustration of (a) the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, ..., 8\}$ and $\mathcal{E} = \{(1, 2), (1, 3), (2, 4), (3, 5), (3, 8), (4, 5), (5, 6), (7, 6)\}$, (b) its undirected counterpart as an example to an undirected graph.

is said to be undirected if it holds that two vertices i and j connected by an edge (i, j) implies that they are also connected by (j, i), i.e. if $(i, j) \in \mathcal{E} \Rightarrow (j, i) \in \mathcal{E}, \forall (i, j) \in \mathcal{E}$. If a graph is declared to be undirected, for convenience, the pair of edges connecting the same pair of vertices will be denoted by only one element in the edge set $\mathcal{E}$, either by (i, j) or (j, i) relaxing the orderedness of these pairs. This interpretation also yields a natural definition for the *undirected counterpart* of a directed graph. In Figure (2.1b) we give the illustration of an undirected graph which is the undirected counterpart of the example above. We note that in an undirected graph, the indegree, outdegree and degree of a vertex are equal.

A *path* $\mathcal{P} = (\mathcal{V}, \mathcal{E})$ is a graph where the set of vertices and edges are of the form $\mathcal{V} = \{j_1, j_2, ..., j_n\}$ and $\mathcal{E} = \{(j_1, j_2), (j_2, j_3), ..., (j_{n-1}, j_n)\}$ respectively. The number of edges is the *length* of $\mathcal{P}$. A *cycle* (or *loop*) is a path with an additional edge (j_n, j_1). Hence a cycle of length n is a path from j_1 back to itself.

A *subgraph* $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ of $\mathcal{G}$ is a graph such that $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$. If $\mathcal{E}'$ contains all edges $(i, j) \in \mathcal{E}$ such that $i, j \in \mathcal{V}'$, then $\mathcal{G}'$ is referred to as the $\mathcal{V}'$ *induced* subgraph of $\mathcal{G}$. A path (or cycle) is said to be contained in $\mathcal{G}$ if it is a subgraph of $\mathcal{G}$. A *chord* of a cycle in $\mathcal{G}$ is an edge $(i, j) \in \mathcal{E}$ which is not in that cycle but connects its two vertices in $\mathcal{G}$.

Given three disjoint sets $\mathcal{A}, \mathcal{B}$ and $\mathcal{C} \subset \mathcal{V}$, $\mathcal{B}$ is said to *separate* $\mathcal{A}$ and $\mathcal{C}$ in $\mathcal{G}$ if $\mathcal{G}\backslash\mathcal{B}$ induced subgraph of $\mathcal{G}$ does not contain any path between any two vertices selected from $\mathcal{A}$ and $\mathcal{B}$ respectively.

Consider an undirected graph $\mathcal{G}$. It is *triangular* if every cycle of length greater than four has a chord. It is *connected* if there exists a path linking any pair of its vertices and *acyclic* if it contains no cycles. An undirected graph which is acyclic and connected is called a *tree*. Tree graphs have the property that there is only one path between any two vertices. On a tree, vertices with degree 1 are called *leaf* nodes. An undirected path which is also a tree is called a *chain*. A *clique* is a set of nodes $\mathcal{C} \subseteq \mathcal{V}$ such that any pair $i, j \in \mathcal{C}$ is connected in $\mathcal{G}$. A clique is *maximal* if there are no nodes $i \in \mathcal{V}\backslash\mathcal{C}$ such that $\mathcal{C} \cup \{i\}$ is a clique. $\mathcal{G}$ is a *complete* graph if it is a maximal clique itself.

Given a collection of subsets of $\mathcal{V}$, i.e. $\mathcal{V}_\mathcal{T} \subseteq \mathcal{P}(\mathcal{V})$, a *junction tree* $\mathcal{G}_\mathcal{T} = (\mathcal{V}_\mathcal{T}, \mathcal{E}_\mathcal{T})$ is a tree on $\mathcal{V}_\mathcal{T}$ such that every vertex on the (unique) path between $i_\mathcal{T}$ and $j_\mathcal{T} \in \mathcal{V}_\mathcal{T}$ contains $i_\mathcal{T} \cap j_\mathcal{T}$. This condition is also known as the *running intersection property*. The *seperator set* between two adjacent vertices $i_\mathcal{T}, j_\mathcal{T} \in \mathcal{V}_\mathcal{T}$ is given by $\mathcal{S}_{i_\mathcal{T}, j_\mathcal{T}} = i_\mathcal{T} \cap j_\mathcal{T}$ and the set of all seperators by $\mathcal{S}_\mathcal{T} = \{\mathcal{S}_{i_\mathcal{T}, j_\mathcal{T}} | (i_\mathcal{T}, j_\mathcal{T}) \in \mathcal{E}_\mathcal{T}\}$.

A *bipartite graph* has the property that its set of vertices $\mathcal{V}$ can be separated into two non-empty class sets such that any edge $(i, j) \in \mathcal{E}$ connects vertices of different classes.

Consider a directed graph $\mathcal{G}$. The set of *ancestors* of a vertex j, denoted by $an(j)$, is the maximal subset of $\mathcal{V} \setminus j$ such that initiating from each of its members there exists a directed path to j in $\mathcal{G}$. Similarly, the set of *descendants* of j is denoted by $de(j)$ and defined as the maximal subset of $\mathcal{V} \setminus j$ such that for any given member, there exists a directed path initiating from j and terminating at that member in $\mathcal{G}$.

A directed acyclic graph admits a partial ordering; a forward partial ordering proceeds after assigning an initial count to the parentless nodes, e.g. $l = 0$, and after removing them from $\mathcal{V}$ by increasing the count by one, and recursively continuing the same procedure. A similar approach starting with childless vertices yield a backward partial ordering. Moreover $an(j)$ and $de(j)$ are disjoint, i.e. $an(j) \cap de(j) = \{\}$. Note that, an acyclic directed graph does not necessarily have an acyclic undirected counterpart. If the undirected counterpart of a directed graph is also acyclic, it is called a *polytree*. In this case, for any vertex j no two parent vertices

have a common ancestor and no two child vertices have a common descendant.

Altough the definitions might slightly vary in different resources on graphs (e.g. a clique is maximal by definition in [30] whereas this property is not required in [31]) we refer to the definitions above throughout.

2.1.2 An Overview of Probability Theory

Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ where Ω is a sample space, $\mathcal{F}$ is an event space and $\mathcal{P}$ is a probabiliy measure, and a measurable space $(\mathcal{X}, \mathcal{B})$, a *random variable* X is a measurable function $X : \Omega \to \mathcal{X}$ with the probability measure $\mathcal{P}$.

When $\mathcal{X}$ is a finite set, we can talk about the probability that X takes a specific value x. This probability is captured in the *probability mass function* $p_X : \mathcal{X} \to [0, 1]$ given by $p_X(x) = \mathcal{P}(\{\omega \in \Omega | X(\omega) = x\})$. Consider the case where $\mathcal{X} = \mathbb{R}$ and $\mathcal{B}$ the Borel field on $\mathbb{R}$, in which we can talk about the probability that X takes a value in an interval (x_1, x_2) with $x_1 < x_2$. This probability is conveyed by the *cumulative distribution function* $P_X : \mathcal{X} \to [0, 1]$ given by $P_X(x) = \mathcal{P}(\{\omega \in \Omega | X(\omega) \leq x\})$ whereas the *probability density function* $p_X : \mathcal{X} \to [0, \infty)$ is given by $p_X(x) = dP(x)$. Given a measurable function $f : \mathcal{X} \to \mathcal{X}$, $Z = f(X)$ is also a random variable with a cumulative distribution $P_Z(z) = \mathcal{P}(\{\omega \in \Omega | f(X(\omega)) \leq z\})$. With sufficient care, we can treat a random variable x taking values from a discrete (or continous) set $\mathcal{X}$ as characterized by its probability mass (or density) function, without further considerations on the underlying probability space. We refer to X as a *continous random variable* if the distribution function P_X is continous and as a *discrete random variable* if P_X is piecewise constant.

A *random vector* X is constituted of N random variables, i.e. $X = (X_1, X_2, ..., X_N)^T$ and takes values from $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_N$ where $\mathcal{X}_i$ is the range of X_i for $i = 1, 2, ..., N$. Consider a continous random vector with a *joint probability density* $p_X(x_1, x_2, ..., x_N)$. The individual density functions of X_i are called *marginal densities* of X and given by $p(x_i) = \int_{\mathcal{X}_{\setminus i}} \mathrm{d}x_{\setminus i} p_X(x_1, x_2, ..., x_N)$. Similarly, the joint distribution of a subset of components of X, denoted by $X_{\mathcal{A}}$ is given by $p_{X_{\mathcal{A}}}(x_{\mathcal{A}}) = \int_{\mathcal{X}_{\setminus \mathcal{A}}} \mathrm{d}x_{\setminus \mathcal{A}} p_X(x_1, x_2, ..., x_N)$. These definitions are also valid for a discrete random vector where we replace integrations with summations.

Given X_i fixed to x_i such that $p(x_i) \neq 0$, the remaining components $X_{\setminus i}$ in the context of X are characterized by the *conditional density function* given X_i which is equal to $p_{X_{\setminus i}|X_i}(x_{\setminus i}|x_i) =$

$p_X(x_1, x_2, ..., x_N)/p_{X_i}(x_i)$ through the celebrated Bayes' rule. Note that $p_{X_{\setminus i}|X_i}$ can be interpreted as a family of distributions from which one is selected by x_i. Similarly, given some components fixed, i.e. $X_{\mathcal{A}} = x_{\mathcal{A}}$ such that $p(x_{\mathcal{A}}) \neq 0$, the set of remaining variables $X_{\mathcal{B}} = X_{\setminus \mathcal{A}}$ are characterized by the *conditional density function* over $X_{\mathcal{B}}$ given $X_{\mathcal{A}}$ which is equal to $p_{X_{\mathcal{B}}|X_{\mathcal{A}}}(x_{\mathcal{B}}|x_{\mathcal{A}}) = p_{X_{\mathcal{B}},X_{\mathcal{A}}}(x_{\mathcal{B}}, x_{\mathcal{A}})/p_{X_{\mathcal{A}}}(x_{\mathcal{A}})$.

Starting from a joint density p_X of a random vector X, probabilistic characterizations of individual and grouped components of X for marginal and conditional cases are introduced. Additional relations which provide further conceptuallization are characterized through certain conditions on the functions defined above.

Any two components where $p_{X_i,X_j}(x_i, x_j) = p_{X_i}(x_i)p_{X_j}(x_j)$ holds are said to be *independent*. A stronger concept is *mutual independence* for which to exist given $X = \{X_1, X_2, ..., X_N\}$, it holds that the joint density $p_X(x_1, x_2, ..., x_N) = \prod_{i\in\{1,...,N\}} p(x_i)$. Similarly, given 3 non-empty sets $X_{\mathcal{A}}, X_{\mathcal{B}}$ and X_C, the components $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ are said to be *conditionally independent* given X_C if and only if $p(x_{\mathcal{A}}, x_{\mathcal{B}}|x_C) = p(x_{\mathcal{A}}|x_C)p(x_{\mathcal{B}}|x_C)$ holds for all values of $x_C \in \mathcal{X}_C$ with nonzero probability[1]. This relation is denoted by $\mathcal{X}_{\mathcal{A}} \perp\!\!\!\perp \mathcal{X}_{\mathcal{B}} \,|\, \mathcal{X}_C$.

Given a continous random vector, its *expected value*, also referred to as the *mean vector*, is given by $E_X\{X\} = \int_{\mathcal{X}} \mathrm{d}x\, x p_X(x)$ and denoted by μ_X. It can easily be shown that the components of μ_X are expected values of the components of X, i.e. $\mu_X = (\mu_{X_1}, \mu_{X_2}, ..., \mu_{X_N})$ where $\mu_{X_i} = E_{X_i}\{X_i\}$. Given a function c over $\mathcal{X}$, the expected value of the function $E_X\{c(X)\}$ is given by replacing x in the integrand with $c(x)$. The *variance* of a single random variable is the exptected value of the function $c(x_i) = (x_i - m_{X_i})^2$. The *covariance* $\sigma_{X_iX_j}$ of two random variables X_i and X_j is given as the expected value of the function, $c(x_i, x_j) = (x_i - \mu_{X_i})(x_j - \mu_{X_j})$ whereas the *covariance matrix* C_X of the random vector X is the expected value of $c(X) = (X - \mu_X)(X - \mu_X)^T$ and has the covariance of X_i and X_j at its $(i, j)^{th}$ field. Two random variables are called *uncorrelated* if their covariance is zero. Hence, the components of X are *mutually uncorrelated* if C_X is diagonal. Through an inner product interpretation of $E_{X_iX_j}\{X_iX_j\}$, two random variables are said to be *orthogonal* if $E_{X_iX_j}\{X_iX_j\} = 0$, denoted by $X_i \perp X_j$.

The *conditional expectation* is defined in a similar fashion by substituting the relevant conditional distribution in the integrand and setting the integration domain appropriately. Suppose that the conditions we have previously discussed for the existence of $p_{x_{\mathcal{B}}|x_{\mathcal{A}}}$ hold. Then the

[1] Note that the subscripts of the distributions are dropped since it is possible to identify them from the context.

conditional expectation of $X_{\mathcal{B}}$ given $X_{\mathcal{A}} = x_{\mathcal{A}}$ is given by

$$E_{X_{\mathcal{B}}|x_{\mathcal{A}}}\{X_{\mathcal{B}}|X_{\mathcal{A}} = x_{\mathcal{A}}\} = \int_{\mathcal{X}_{\mathcal{B}}} \mathrm{d}x_{\mathcal{B}} x_{\mathcal{B}} p_{X_{\mathcal{B}}|X_{\mathcal{A}}}(x_{\mathcal{B}}|x_{\mathcal{A}})$$

and denoted by $\mu_{X_{\mathcal{B}}|x_{\mathcal{A}}}$. Similarly, $E_X\{c(X)\} = E_{\mathcal{X}_{\mathcal{A}}}\{E_{\mathcal{X}_{\mathcal{B}}|\mathcal{X}_{\mathcal{A}}}\{c(X)|X_{\mathcal{A}}\}\}$ where the conditional expectation in the right hand side is interpreted as a function over $X_{\mathcal{A}}$ given by

$$E_{\mathcal{X}_{\mathcal{B}}|\mathcal{X}_{\mathcal{A}}}\{c(X)|X_{\mathcal{A}}\} = \int_{\mathcal{X}_{\mathcal{B}}} \mathrm{d}x_{\mathcal{B}} c(X_{\mathcal{A}}, x_{\mathcal{B}}) p_{X_{\mathcal{B}}|X_{\mathcal{A}}}(x_{\mathcal{B}}|X_{\mathcal{A}})$$

The celebrated *law of total expectation* is obtained considering $c : \mathcal{X}_{\mathcal{B}} \to \mathbb{R}$ and in a similar fashion with the discussion above as $E_{X_{\mathcal{B}}}\{c(X_{\mathcal{B}})\} = E_{\mathcal{X}_{\mathcal{A}}}\{E_{\mathcal{X}_{\mathcal{B}}|\mathcal{X}_{\mathcal{A}}}\{c(X_{\mathcal{B}})|X_{\mathcal{A}}\}\}$.

The definitions above are valid for discrete random vectors and variables with the difference that integrations are replaced with summations where necessary. Equivalently, we can stay within the contionus framework given by the expressions above and consider a probability density function constructed by superpositioning impulses scaled to integrate to the corresponding value of the probability mass function and shifted to the corresponding value of $x \in \mathcal{X}$, i.e. given the probability mass function p_X, it is possible to proceed with the corresponding continous distribution $p_X^c(x)$ given by

$$p_X^c(x) = \sum_{\tilde{x} \in \mathcal{X}} p_X(\tilde{x}) \delta(x - \tilde{x})$$

Several examples with an extended discussion related with the subject can be found in several textbooks including [32]. We refer to the definitions and notation introduced above throughout the thesis, with an emphasis on conditional independence and expectations of functions.

2.2 Bayesian Detection and Estimation

Statistical inference deals with drawing conclusions about some quantities based on the *observation vector* $y = (y_1, y_2, ..., y_n)^T$ (or simply the "data"). Detection and estimation is a well established area of statistical inference, also known as the statistical decision theory, in which the value of an unknown quantity $x \in \mathcal{X}$ is of concern given observations induced under uncertainities. The mathematical model asserts that we are given observations $y \in \mathcal{Y}$ due to a probabilistic mapping from $\mathcal{X}$ to $\mathcal{Y}$ where $\mathcal{X}$ is the set of possible values of x. Evaluating a function $\gamma : \mathcal{Y} \to \mathcal{X}$ at the observation y, a guess $\hat{x}$ is produced (Figure 2.2). The problem is to find γ^* among the possible functions that results with the desired quality of inference. This

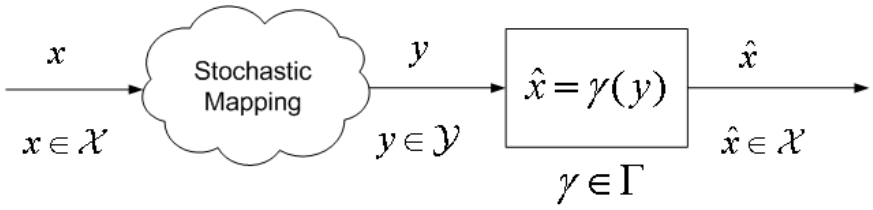


Figure 2.2: Illustration of the conventional scenario in detection and estimation; a random vector x which takes values from the set $\mathcal{X}$ induces a measurement $y \in \mathcal{Y}$ where the observer is to decide on the value of x based on y. Hence the mathematical description of the system is a function $\gamma : \mathcal{Y} \to \mathcal{X}$ which maps an observation $y \in \mathcal{Y}$ to a value in $\mathcal{X}$ called an estimate of x, i.e. $\hat{x} = \gamma(y)$. The set of all possible such γ is given by $\Gamma = \{\gamma | \gamma : \mathcal{Y} \to \mathcal{X}\}$.

scenario is referred to as a detection problem if $\mathcal{X}$ is of finite cardinality, i.e. $|\mathcal{X}| < \infty$, and an estimation problem if $\mathcal{X}$ is denumerable. The uncertainities of the observation process such as sensor noise is captured by the probabilistic mapping and characterized by the *observation likelihood* function $p_{Y|X}$. In addition, it is possible to treat x as a realization of a random vector X which takes values from $\mathcal{X}$ in accordance with the *a priori density* function p_X. Therefore, the vector constructed by aggregating the *latent variables* X and the observation variables Y, i.e. $Z = (X^T, Y^T)^T$, admits a characterization through the joint density (or mass) function $p_Z(x, y) = p_{Y|X}(y|x)p_X(x)$. Such a probabilistic model enables a characterization of the uncertainities related to the latent variables X based on the observed value y through the Bayes rule and a "rational guess" on the realized value of X can be drawn utilizing the law of total expectation.

In the Bayesian estimation setting, we assign costs to all possible $(x, \hat{x})$ pairs through a function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and select γ such that the *expected cost*, also referred to as the *Bayes risk* is mimimum. In order to proceed, we note that $\hat{X}$, being a function of Y, is a random variable and selection of γ where $\hat{x} = \gamma(y)$ induces the conditional distribution $p(\hat{x}|y)$ [2]. In addition, since γ operates on only Y, the conditional independence relation $X \perp\!\!\!\perp \hat{X} \,|\, Y$ holds and the joint distribution $p(x, \hat{x}, y; \gamma)$ is given by

$$p(x, \hat{x}, y; \gamma) = p(\hat{x}|y; \gamma)p(x|y)p(y) \tag{2.1}$$

Therefore, for any selection of $\gamma : \mathcal{Y} \to \mathcal{X}$, there exists an expected cost given by

[2] In other words, the distribution $p(\hat{x}|y)$ is specified by selecting the function γ. We denote such dependencies by "; γ" as in $p(\hat{x}|y; \gamma)$ throughout the thesis. Let us assume that γ is a deterministic rule, then $p(\hat{x}|y) = \delta(\hat{x} - \gamma(y))$ where δ is the Dirac' s delta distribution. A discussion regarding possible non-deterministic rules is introduced later in this section.

$J(\gamma) = E\{c(x, \hat{x}); \gamma\}$, where the underlying distribution is $p(x, \hat{x}; \gamma) = \int_{\mathcal{Y}} \mathrm{d}y p(x, \hat{x}, y; \gamma)$. Hence, the Bayesian estimation problem can be expressed as

$$\begin{aligned} \text{(P0):} \qquad & \min E\{c(x, \hat{x}); \gamma\} \\ & \text{subject to } \gamma \in \Gamma \end{aligned} \tag{2.2}$$

where Γ is the set of all functions mapping $\mathcal{Y}$ to $\mathcal{X}$, i.e. $\Gamma = \{\gamma | \gamma : \mathcal{Y} \to \mathcal{X}\}$.

Before proceeding with the solution to (P0), we note that it is possible to treat γ as a deterministic function as well as a randomized rule. Let $\bar{\Gamma}$ denote the set of deterministic estimation rules. A randomized mechanism proceeds as follows: Suppose that a probability density function $p(k)$ is given for the random variable K which takes values on the interval $(0, 1)$. It is also possible to define a mapping from $(0, 1)$ to $\bar{\Gamma}$. Therefore a denumerable set of deterministic rules $\{\gamma_k\}_{k\in(0,1)}$ can be obtained. As soon as y is received, a randomized $\tilde{\gamma}$ refers to drawing k from $p(k)$ and evaluating the corresponding deterministic rule, i.e. $\tilde{\gamma}(y) = \gamma_k(y)$. Let $\tilde{\Gamma}$ denote the set of all such $\tilde{\gamma}$. Note that for $\tilde{\gamma} \in \tilde{\Gamma}$ it holds that $J(\tilde{\gamma}) = \int_0^1 \mathrm{d}k p(k) J(\gamma_k)$. Moreover, $\gamma_k \in \bar{\Gamma}$ and $\bar{\Gamma} \subset \tilde{\Gamma}$ yielding

$$\begin{aligned} \inf_{\gamma\in\tilde{\Gamma}} J(\gamma) &= \inf_{\gamma\in\tilde{\Gamma}} \int_0^1 \mathrm{d}k p(k) J(\gamma_k) \\ &= \int_0^1 \mathrm{d}k p(k) \inf_{\gamma\in\bar{\Gamma}} J(\gamma_k) \\ &= \inf_{\gamma\in\bar{\Gamma}} J(\gamma_k) \end{aligned}$$

Therefore, for the problem (P0), we can confine ourselves to the set of deterministic rules and hence choose $\Gamma = \bar{\Gamma}$.

Let us denote the solution to (P0) by γ^* and consider the law of total expectation for $J(\gamma)$, i.e.

$$E\{c(x, \hat{x}); \gamma\} = E\{E\{E\left\{c(x, \hat{x}) | \hat{X}, Y; \gamma\right\} | Y\}\}$$

An explicit form for the expression above is obtained after substituting Eq.(2.1) in $J(\gamma)$ together with some arrangements, i.e. having

$$J(\gamma) = \int_{\mathcal{Y}} \mathrm{d}y p(y) \int_{\mathcal{X}} \mathrm{d}\hat{x} p(\hat{x}|y; \gamma) \int_{\mathcal{X}} \mathrm{d}x c(x, \hat{x}) p(x|y)$$

and noticing that $J(\gamma^*)$ is minimum if and only if $\gamma^*(y) = \arg\min_{\hat{x}\in\mathcal{X}} \int_{\mathcal{X}} \mathrm{d}x c(x, \hat{x}) p(x|y)$ for all $y \in \mathcal{Y}$ with nonzero probability. Equivalently, since $p(y|x)p(x) \propto p(x|y)$ for all such $y \in \mathcal{Y}$,

$$\gamma^*(Y) = \arg\min_{\hat{x}\in\mathcal{X}} \int_{\mathcal{X}} \mathrm{d}x c(x, \hat{x}) p(Y|x) p(x) \tag{2.3}$$

with probability one.

Hence, the solution to (P0) is obtained in a variational form through applying the law of total expectation to the Bayesian objective. The discussion above holds also for the detection problem through replacing integrations with summations where appropriate (and constructing a randomized rule through a positive finite integer K and considering a probability mass function $p(i)$ for $i = 1, 2, ..., K$ together with a finite set of deterministic decision rules [33]).

Example 2.2.1 *Consider problem (P0) with $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, and the squared error cost, i.e. $c(x, \hat{x}) = (x - \hat{x})^2$. The solution is obtained in closed form through differentiating the objective function in Eq.(2.3) with respect to $\hat{x}$ as*

$$\hat{X} = \gamma^*(Y) = \int_{\mathcal{X}} \mathrm{d}x \, x \, p(x|Y) \tag{2.4}$$

which is nothing but the expected value of the posterior density. This well known result in the literature on estimation is referred to as the Minimum Mean Squared Error (MMSE) Estimator [34]. It is possible to make an extension to the case where X and Y are random vectors, i.e. $\mathcal{X} = \mathbb{R}^N$ and $\mathcal{Y} = \mathbb{R}^M$ with the cost given by $c(x, \hat{x}) = \|x - \hat{x}\|^2 = (x - \hat{x})^T (x - \hat{x}) = \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$. The above steps yield a solution with the components equal to the expected values of posterior marginals, i.e. $\hat{X} = \gamma^*(Y) = (E\{X_1|Y\}, E\{X_2|Y\}, ..., E\{X_N|Y\})^T$.

Returning to the single variable case, another common choice for the cost function is the complementary ϵ-neighborhood indicator given by

$$c(x, \hat{x}) = \begin{cases} 0 \, , & |x - \hat{x}| \leq \epsilon \\ 1 \, , & |x - \hat{x}| > \epsilon \end{cases}$$

for which the solution is similarly obtained by

$$\hat{X} = \gamma^*(Y) = \arg\max_{x \in \mathcal{X}} p(x|Y) \tag{2.5}$$

as $\lim \epsilon \to 0$ and called the Maximum A Posteriori (MAP) Estimator [35]. This solution generalized to the random vectors case does not exhibit any simplifications for its components, but for $c(x, \hat{x}) = \sum_{i=1}^{N} c_i(x_i, \hat{x}_i)$ where c_i' s are complementary ϵ-neighborhood indicators, i^{th} the component of $\hat{X} = \gamma^*(Y)$ is equal to $\tilde{x}_i$ that maximizes the i^{th} marginal posterior, i.e. $\hat{x}_i = \arg\max_{\tilde{x}_i \in \mathcal{X}_i} p(\tilde{x}_i|Y)$ and γ^* is called the Maximum Posterior Marginal (MPM) Estimator.

Another well known problem is m-ary detection in which case $\mathcal{X} = \{1, 2, ..., M\}$ and Eq.(2.3) turns to $\gamma^*(Y) = \arg\min_{i \in \{1,2,...,M\}} [\mathbf{\Theta} p(Y|x)]_i$ where $\mathbf{\Theta}$ is a coefficients matrix such that the

$(i,j)^{th}$ entry is given by $[\mathbf{\Theta}]_{i,j} = c(\hat{x}=i, x=j)p(x=j)$ and $p(Y|x)$ is interpreted as a vector such that j^{th} field is given by $\left[p(Y|x)\right]_j = p(Y|x=j)$. This form further simplifies for the binary detection case, i.e. $|\mathcal{X}| = 2$, and the minimum Bayesian risk decision is found as a likelihood ratio test given by

$$\Lambda(y) \triangleq \frac{p(y|x=1)}{p(y|x=2)} \underset{x=2}{\overset{x=1}{\gtrless}} \frac{[\mathbf{\Theta}]_{1,2} - [\mathbf{\Theta}]_{2,2}}{[\mathbf{\Theta}]_{2,1} - [\mathbf{\Theta}]_{2,2}} \tag{2.6}$$

2.3 Probabilistic Graphical Models

Many statistical inference tasks including estimation (and detection) are involved with marginals of the joint density function introduced in Section 2.1.2. For example, MMSE and MPM estimators presented in Example 2.2.1 rely on marginal densities of the joint a-posteriori density function. Although the joint distribution provides a complete probabilistic model for the variables of concern, the complexity of performing Bayesian inference inhibits straightforward computation of marginals. Consider a random vector with N components each taking a value among M choices. With a straightforward approach, storage of the joint mass function and computing the marginal distribution of one of the variables through simple marginalization yield a storage and computation complexity of $O(M^N)$ and $O(M^{N-1})$ respectively.

On the other hand, if the *global* function $f_X(x)$ subject to marginalizations is in the form of product of functions over subsets of its arguments, i.e. $f_X(x) = \prod_{C\in\mathfrak{C}} f_{X_C}(x_C)$ where $\mathfrak{C} \subset \mathcal{P}(\mathcal{V})$, this complexity might be reduced through the application of the well known distributive law. Roughly speaking, the complexity of the solution to the "Marginalize Product of Functions" problem is related to the particular factorization of the global function in terms of *local* functions, i.e. $\{f_{X_C}(x_C)\}_{C\in\mathfrak{C}}$, and smaller domain set dimensions of these functions and fewer number of variables shared among them results with reduced complexity. In the extreme case, the global function is such that it is the product of functions over single components, i.e. $f_X(x) = f_{X_1}(x_1)f_{X_2}(x_2)...f_{X_N}(x_N)$. A generalization along these lines is presented in [36], together with an algorithm which corresponds to computation of marginals through invoking the distributive law in accordance with the particular factorization and results with reduced complexity. Still, in the worst case, it is shown that Bayesian inference is NP-hard [37].

In the context of statistical inference, given a set of random variables, independence relations defined in Section 2.1.2 lead factorizations of the joint density for which case it is the global function of interest. Hence, these relations describing the interactions among variables are fundamental to analysis and synthesis of probabilistic models as well as drawing inferences over them. Such relations are also referred to as Markov properties of the set of random variables of concern. Similar to the MPF discussion, Probabilistic Graphical Models provide means of representing these properties in terms of relations on graphs together with inference mechanisms acting over them which, in a sense, tailor the complexity to that of the interactions.

A Graphical Model consists of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ which represents a set of random variables $X = \{X_1, X_2, ..., X_N\}$ constituting the process to be modeled and Markov properties among them together with functions defined on a subset of random variables, i.e. $\left\{\psi_j(X_j)\right\}$ where $X_j \subset X$. They come in two varieties of directed and undirected models which differ in the way they express these properties and the corresponding factorization of the joint distribution.

Directed Graphical Models: Directed Graphical Models, also referred to as *Bayesian Networks* or *Belief Networks*, are directed acyclic graphs which are commonly interpreted to be presenting causality relationships among variables through conditional distributions. A directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ expresses that the joint distribution of concern satisfies

$$p_X(x) = \prod_{v \in \mathcal{V}} p(x_v | x_{\pi(v)}) \tag{2.7}$$

The factors of this product form are determined by the parent-child relationships in $\mathcal{G}$ and conditional distibutions over them. Considering Eq. (2.7), the example directed graph given in Figure (2.1a) corresponds to

$$p(x_1, x_2, ..., x_8) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3, x_4)p(x_6|x_7)p(x_8|x_3)$$

For efficient inference on Bayesian networks, it is possible to transform $\mathcal{G}$ to an undirected model $\mathcal{G}_{\mathcal{M}}$ through *moralization* and then proceed in the undirected framework [38]. This approach will be presented in the discussion of undirected graphical models.

Undirected Graphical Models: We will focus specifically on *Markov Random Fields* which are often mentioned throughout the thesis[3]. Markov Random Fields are undirected

[3] Another common undirected graphical model is *Factor Graphs* which, together with the sum-product and the max-product algorithms, render well known algorithms such as Kalman filtering on Markov chains and Viterbi decoding as special cases [39]. Also, for connections of factor graphs, Bayesian networks and Markov random fields, the interested reader is referred to this reference.

graphical models where each node $v \in \mathcal{V}$ is associated with a set of random variables $X_v \subset X$ and Markov properties of X are represented through separation relations on $\mathcal{G}$. The *pairwise Markov property* is that the absence of an edge between two nodes indicates conditional independence of the corresponding random variables given the remaining, i.e. if $(v_1, v_2) \notin \mathcal{E}$, then $X_{v_1} \perp\!\!\!\perp X_{v_2} \,|\, X_{\mathcal{V}\setminus\{v_1,v_2\}}$. Consequently, for any $v \in \mathcal{V}$, the *local Markov property*, $p(x_v|x_{\mathcal{V}\setminus\{v\}}) = p(x_v|x_{ne(v)})$, holds. In general, X is Markov with respect to graph $\mathcal{G}$ if given disjoint sets $\mathcal{A}, \mathcal{B}, C \subset \mathcal{V}$ satisfying the property that $\mathcal{B}$ separates $\mathcal{A}$ and C in $\mathcal{G}$, the random variables $X_{\mathcal{A}}$ and X_C are conditionally independent given $X_{\mathcal{B}}$.

Given X Markov with respect to $\mathcal{G}$, a theorem due to Hammersley and Clifford guarantees that the joint distribution factorizes to local functions over cliques of $\mathcal{G}$, also commonly called compatibility functions.

Theorem 2.3.1 *(Hammersley-Clifford) X is Markov with respect to $\mathcal{G}$ if and only if any positive and continuous joint probability density p_X factorizes according to $\mathcal{G}$, i.e.*

$$p_X(x) \propto \prod_{C \in \mathfrak{C}} \psi_C(x_C) \tag{2.8}$$

where $\mathfrak{C}$ is the set of all cliques of $\mathcal{G}$.

Proof. See, for example [31], where the proof sets the equivalence of pairwise and local Markov properties, X being Markov with respect to $\mathcal{G}$ and the factorization above. ■

This theorem provides us a basis for constructing a factorization for the joint density given a Markov Random Field, or conversely, a Markov Random Field given a certain factorization of the joint density. For the former case, we expect the compatibility functions to be in the form of product of local conditional and joint densities, however exact identification requires additional care.

If $\mathcal{G}$ is a tree, then the cliques are single vertices and pairs of vertices that are connected, i.e. $\mathfrak{C} = \mathcal{V} \cup \{\{v_1, v_2\} | (v_1, v_2) \in \mathcal{E}\}$ and the joint distribution satisfies

$$p(x) = \prod_{v \in \mathcal{V}} p_v(x_v) \prod_{(v_1,v_2) \in \mathcal{E}} \frac{p(x_{v_1}, x_{v_2})}{p(x_{v_1})p(x_{v_2})} \tag{2.9}$$

The compatibility functions are easily identified comparing this equation with Eq.(2.8). Hence, we achieve a means to describe the joint probability distribution of a large number of random

variables in terms of functions of lower dimensions in their domain set. Similar to the MPF case, inference algorithms on graphical models exploit this property for solving problems dealing with the computation of marginal distributions of a subset of variables or their value at the peak of the joint distribution.

On a tree graph $\mathcal{G}$, the marginal probabilities at any node $v \in \mathcal{V}$ is given by the product of the node compatibility function and functions regarding each neighbor node as

$$p_v(x_v) \propto \psi_v(x_v) \prod_{u \in ne(v)} m_{u \to v}(x_v) \tag{2.10}$$

where

$$m_{u \to v}(x_v) \propto \int_{x_u \in X_u} \mathrm{d}x_u \, \psi_{v,u}(x_v, x_u) \psi_u(x_u) \prod_{z \in ne(u) \setminus v} m_{z \to u}(x_u) \tag{2.11}$$

and represent a set of fixed point equations due to the acyclic structure of $\mathcal{G}_{\mathcal{T}}$ [40]. An exact solution to this set of equations yields a treatment where each node $\mathcal{G}$ is viewed as an entity which can compute its state given by Eq.(2.10) and *messages* given by Eq.(2.11) to be transmitted to neighboring nodes together with a messaging schedule, leading to a *Message Passing Algorithm*.

One possible solution for computing the marginal of a certain variable in a message passing fashion is to select the associated node as the root and start a message flow from the leaves directed to the root. For the leaf nodes, the term in Eq.(2.11) due to messages from neighbors drops and the schedule begins to proceed. As node u receives messages from all its neigbours except $v \in ne(u)$ it is triggered to compute Eq.(2.11). The marginal distribution of concern is achieved through Eq.(2.10) as soon as the root node receives messages from all of its neigbors.

In order to compute all marginals, *Belief Propagation* utilizes the same structure without any particular root. Node u sends only one message to any of its neighbour $v \in ne(u)$ by suppressing any trigger if it has already. Hence, after a finite number of steps, the message passing operations stop and the node states are the corresponding marginals. Similarly, the problem of finding the value of the variable associated with node v, i.e. x_v, at the maximum of the joint density is expressed by replacing the integration (or summation if X is discrete) in Eq.s (2.10) and (2.11) by max operations and the solution has the same schedule as that in the previous problem.

It is possible to treat Belief Propagation as a generalization of recursive Bayesian filtering

on chains which turn to Kalman smoothing with Gaussian distributions. Similarly, viterbi decoding can be described as an instance of the max-product algorithm.

Note that the nodes of an MRF is not necessarily associated with a single random variable. Suppose $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is not a tree and each node is associated with a single random variable. It is possible to obtain another representation $\mathcal{G}_\mathcal{P} = (\mathcal{V}_\mathcal{P}, \mathcal{E}_\mathcal{P})$ through an appropriate aggregation of nodes $v \in \mathcal{V}$ to super-nodes $v_\mathcal{P} \subset \mathcal{V}$ such that the corresponding factorization takes the form where each compatibility function is over a super-node $v_\mathcal{P} \in \mathcal{V}_\mathcal{P}$ or an edge $(u_\mathcal{P}, v_\mathcal{P}) \in \mathcal{E}_\mathcal{P}$ [41]. In other words

$$p(x) \propto \prod_{v_\mathcal{P} \in \mathcal{V}_\mathcal{P}} \psi_{v_\mathcal{P}}(x_{v_\mathcal{P}}) \prod_{(u_\mathcal{P}, v_\mathcal{P}) \in \mathcal{E}_\mathcal{P}} \psi_{u_\mathcal{P}, v_\mathcal{P}}(x_{u_\mathcal{P}}, x_{v_\mathcal{P}}) \tag{2.12}$$

holds which is similar to that for a tree given in Eq.(2.9) except that for the case it is not straightforward to identify the compatibility functions. Moreover, with the representation $\mathcal{G}_\mathcal{P}$, our resolution is decreased in a sense, since further Markov properties among single variables are not visible anymore.

A related approach which can be used to exploit the advantages that the tree structures exhibit is to construct the corresponding junction tree representation. There exist a junction tree over maximal cliques of a triangular graph and it is possible to triangulate $\mathcal{G}$ by adding chords such that no Markov property that has not been presented in $\mathcal{G}$ is added. Suppose $\mathcal{G}_\Delta = (\mathcal{V}, \mathcal{E}_\Delta)$ is such a triangulation of $\mathcal{G}$. Then we can construct a tree on the set of maximal cliques of $\mathcal{G}_\Delta$ denoted by $\mathfrak{C}_\mathcal{M}$, i.e. $\mathcal{J} = (\mathfrak{C}_\mathcal{M}, \mathcal{E}_\mathcal{M})$, such that the running intersection property is satisfied and the joint probability distribution is given by

$$p_X(x) = \frac{\prod_{C \in \mathfrak{C}_\mathcal{M}} p_C(x_C)}{\prod_{\mathcal{S} \in \mathcal{S}_\mathcal{T}} p_\mathcal{S}(x_\mathcal{S})} \tag{2.13}$$

where $\mathcal{S}_\mathcal{T}$ is the set of separators of $\mathcal{J}$ [42]. Then we can still apply the message passing algorithms on trees with simple twists.

Considering inference on a Bayesian Network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, together with Eq.(2.7) and Eq.(2.8) we notice that any node $v \in \mathcal{V}$ must lie in the same clique with its parents in the MRF representation. Therefore, after constructing the undirected counterpart of $\mathcal{G}$, we add edges between unconnected parents of each node with respect to the original graph. This procedure is known as *moralization* and the resulting undirected graph $\mathcal{G}_\mathcal{M}$ as the *moral graph* of $\mathcal{G}$. Then the discussion for constructing a junction tree is valid.

Example 2.3.2 *In Figure(2.3a) we illustrate the moral graph of $\mathcal{G}$ given in Figure (2.1a). Note that edges are added to connect parent nodes* 3, 4 *of* 5 *and* 5, 7 *of* 6 *respectively. There are two possible triangulations; obtained through adding a chord connecting nodes* 2 *and* 3 *or nodes* 1 *and* 4 *as given in Figure 2.3b and Figure 2.3c respectively. Also seen is the maximal cliques with dashed lines, on which the junction trees are constructed. Given the set of maximal cliques, any tree satisfying the running intersection property is valid. Some possible junction trees corresponding to the two possible triangulated graphs are given in Figure2.4a-c and Figure 2.4d-e respectively (Other junction trees can be constructed, for example, by swithcing the nodes* $1, 2, 4$ *and* $1, 3, 4$ *in Figure 2.4d-e). Note that the junction trees corresponding to a particular triangulation have the same set of separators and hence represent the same product form given by Eq. (2.13). They differ, for example, in the message schedule they imply.*

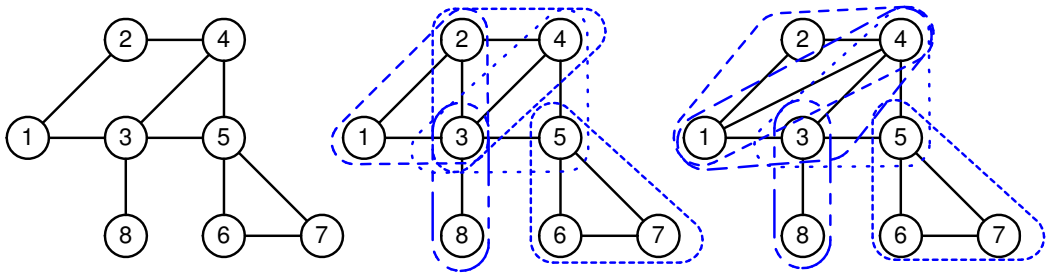


Figure 2.3: Illustration of (a) the moral graph of $\mathcal{G}$ given in Figure (2.1a) (b) a possible triangulation of the moral graph constructed by adding a chord connecting nodes 2 and 3, (c) another possible triangulation by adding a chord connecting nodes 1 and 4.

Although Equations (2.10) and (2.11) describe the situation for a loop free graph $\mathcal{G}_{\mathcal{T}}$, they are also well defined for any MRF $\mathcal{G}$ through the corresponding pairwise factorized form given in Eq.(2.12). Hence, given a pairwise MRF $\mathcal{G}_{\mathcal{P}}$, rather than constructing the junction tree representation and applying inference algorithms on junction trees, one may prefer to ignore the loops in $\mathcal{G}_{\mathcal{P}}$ and run (2.11) on a loopy graph. A number of reasons motivate this approach including the facts that aggregating the nodes to form a junction tree results in increased domain set dimensions for the corresponding compatibility functions which consequently inhibits suitability for distributed computations and complexity reduction and that construction of the corresponding compatibility functions is not straightforward, in general.

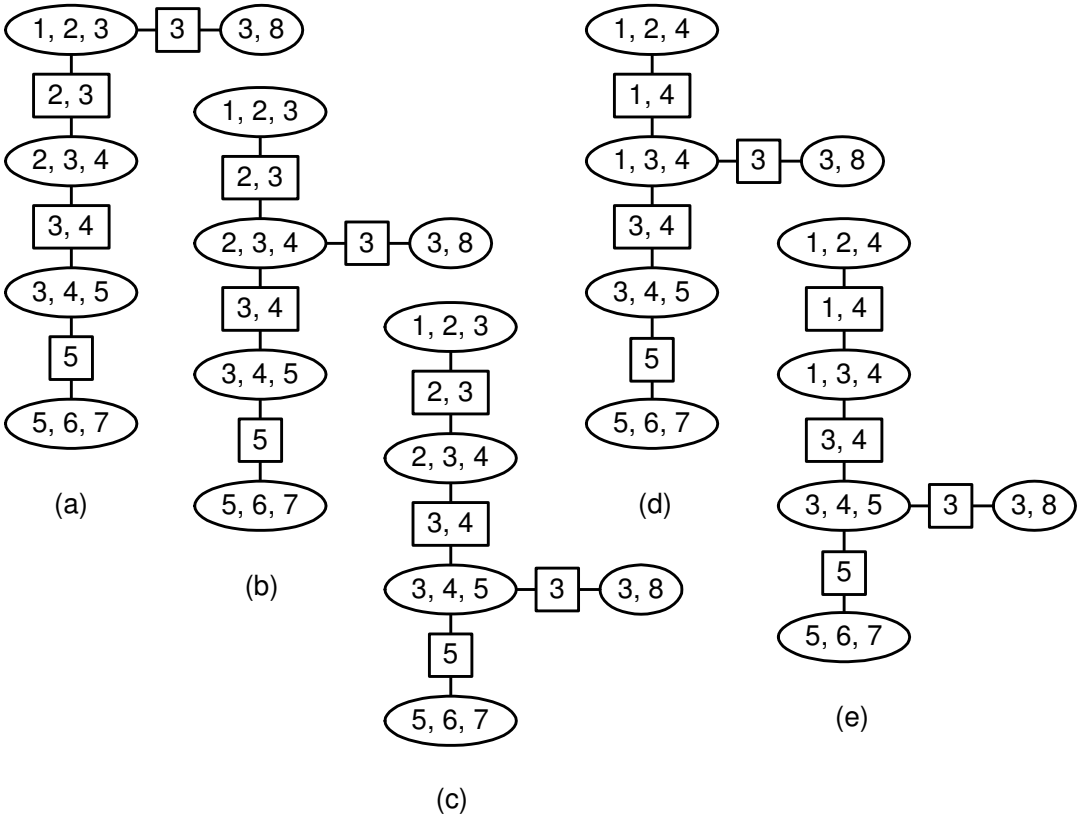


Figure 2.4: Illustration of some possible junction trees on the maximal cliques of the triangulated moral graphs constructed by connecting them such that the running intersection property is satisfied. The nodes which are maximal cliques are represented with the ellipse shapes and the separator sets although not a part of the graph, are shown by rectangles for convenience; (a),(b),(c) corresponds to the triangulation given in Figure (2.1b), (d),(e) corresponds to the triangulation given in Figure (2.1c).

Loopy variants of the sum-product rule have been successful in many applications including decoding of low-density parity check codes [43] which is characterized by graphs with long cycles. The error bounds and performance of loopy versions of inference algorithms have been studied further. Examples of the various papers on the subject are [44] and [45] just to mention a few. Inference algorithms on loopy graphs together with the assessment of their performance are of interest in the sensor network context where typically graphs with short cycles arise [5].

2.4 Graphical Models for Statistical Inference in Sensor Networks

The treatment of a sensing problem in the statistical inference framework is provided through a probabilistic characterization of uncertainities. However, the conventional approaches given in Section (2.2) necessitates the collection of observations at a processing center and fail to meet the requirements in such a scenario where the probabilistic model is composed of a large number of variables and the observations are collected by devices that are located apart, have limited processing and communication capabilities and energy budget. In sensor networks, it is preferable to utilize a framework which distributes the computational load over the nodes through a balanced and acceptable amount of communications and hence yields a collaborative scheme for the processing of the observations.

Considering sensor network applications, it is often the case that the quantities subject to inference are associated with physical locations and exhibit interactions that are local, i.e. each variable relies on the ones associated within a neighborhood and renders the remaining as nuisance variables. In addition, the noise processes associated with sensor platforms are independent and the observations collected by platforms are induced only by the quantities associated with that region. Provided that these interactions render a sparse enough information structure, it is possible to decompose the corresponding global processing scheme in terms of local procedures that are distributed in the network. Probabilistic graphical models introduced in Section 2.3 provide a rigorous base for representing the information structure of the problem and efficient processing schemes through message passing algorithms. In this respect, the information graph refers to any type of graphical model, which might vary regarding the specific application. For example, the random field estimation scenario presented in Section 1.1.3 is well suited for Markov random fields. The discussion of distributed statistical inference in sensor networks for different representations is similar, so let us consider the case in which the information structure is represented by a Markov random field and the joint density has the form given by Eq. (2.8). Observed variables introduced with label Y results with

$$p(x, y) \propto \prod_{C \in \mathfrak{C}} \psi_C(x_C, y_C) \tag{2.14}$$

The a-posteriori joint distribution of the variables constituting x given all observations y is proportional to the right hand side of Eq. (2.14) and without loss of generality, equality

is satisfied through appropriate scaling of the compatibility functions. In Section 2.3, we pointed out that it is possible to aggregate nodes of an MRF and obtain the joint density in a pair-wise factorized form given by with respect to the modified graph Eq. (2.12). For the case, this approach yields

$$p(x|y) \propto \prod_{v \in \mathcal{V}} \psi_v(x_v, y_v) \prod_{(s,t) \in \mathcal{E}} \psi_{s,t}(x_s, x_t, y_{st}) \tag{2.15}$$

where y_v and y_{st} are the set of noisy observations induced only by the set of random variables x_v and $\{x_s, x_t\}$ respectively.

If the measurements satisfy the observation locality property, i.e. for each observation y_v, y_v is independent of $x_{u \in \mathcal{V} \setminus \{v\}} \cup y_{u \in \mathcal{V} \setminus \{v\}}$ given x_v, then the information structure given by Eq. (2.15) reduces to

$$p(x|y) \propto \prod_{v \in \mathcal{V}} \psi_v(x_v, y_v) \prod_{(s,t) \in \mathcal{E}} \psi_{s,t}(x_s, x_t) \tag{2.16}$$

which is often the case in sensor network applications. Consider the example scenario presented in Section 1.1.3. The MRF representation of the prior for the latent variables together with the observed variables satisfying the locality property yield the MRF representation given in Figure 2.5.

Through mapping partitions of the graph $\mathcal{G}$ which represents the information structure onto real sensor platforms, a collaborative processing scheme for sensor networks is possible. Considering a certain message passing algorithm for inference, messages between nodes of the information graph that are hosted by two distinct sensor platforms correspond to real communication over the network for which the network layer should maintain links[4]. In addition to the link availability, the computational resources of the platforms should be enough to perform the message and state updates.

This mapping is trivial in the case that the locality of observations hold and consequently each factor $\psi_v(x_v, y_v)$ in Eq. (2.16) describes the contribution of a distinct sensor platform. The platform that collects the observation(s) y_v is then associated with y_v, and the variable(s) inducing y_v, i.e. x_v, together with the corresponding factor, e.g. $\psi_v(x_v, y_v)$ (Considering the random field estimation example, the partitions of $\mathcal{G}$ shown by dashed-ellipses in Figure 2.5

[4] It is not reasonable to assume that these links are available since, as also indicated in Section 1.1.1, the graph in which edges are available links between platforms, also referred to as the network graph, is a complete graph provided that the underlying channels render a connected graph.

are associated with the platforms that collect the observation variables contained in the partitions). In addition, two neighboring platforms should be provided enough information to evaluate edge compatibility functions, i.e. $\psi_{s,t}(x_s, x_t)$.

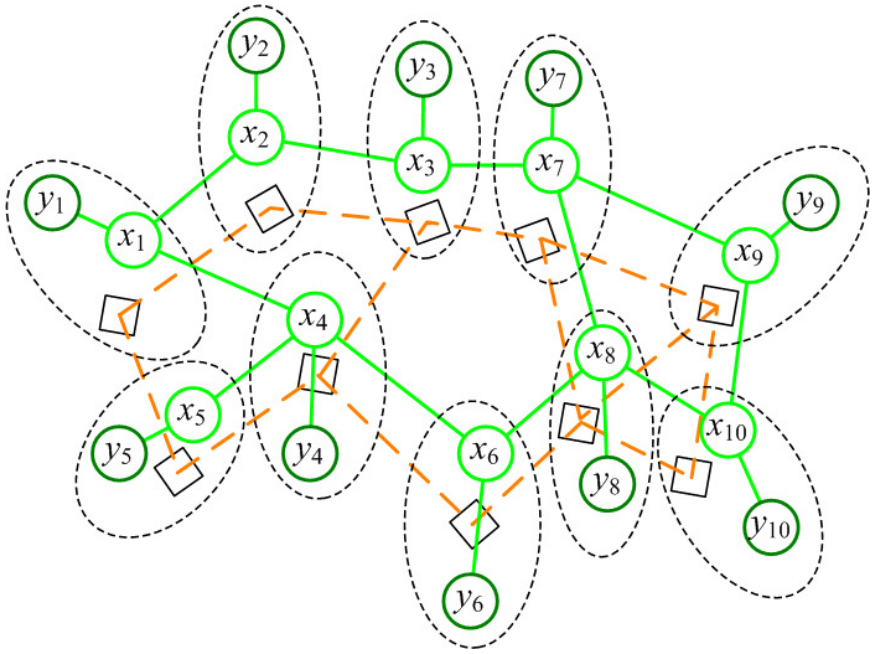


Figure 2.5: Illusration of the trivial mapping for the pairwise MRF representation for the case that the property of locality of observations is satisfied. The information structure of the problem is revealed by the MRF (seen in green, with the observed variables darker). Each physical platform (black square) observes y_v induced only by the random variable associated with its physical location, i.e. x_v. When the pairwise representation in Eq.(2.16) is valid, the partition of the information structure containing x_v and y_v is mapped to the corresponding platform (the partition of the information within the dashed-ellipses are associated with the platform falling in the same ellipse). The message passings between neigboring nodes x_u and x_v correspond to real communication between hosting platforms over the network (orange graph with dashed lines).

The accuracy of the result depends on a number of factors. First of all, due to the limited bandwidth of the links, it is not possible to transmit messages without errors (possibly due to discretization and quantization) resulting with a degraded performance [46]. Second, the problem structure might not yield exact inference through message passing schemes at all. Note that, in order to perform exact inference on the general structure given in Eq. (2.14), in principle, the corresponding junction tree $\mathcal{J}$ should be constructed. This approach exhibits the subtleties argued in Section 2.3 and most likely obscures the mapping onto the platforms and increases the computational resource demands on single platforms. The option to proceed with a loopy graph and perform a loopy message passing is more appealing considering

the low overhead for mapping and identification of the compatibility functions together with comparably lower computational demands on single platforms. On the other hand, this approach, yielding an infinite message schedule, often results with approximate inference and an additional factor degrading performance is the fact that the message schedule can be followed only for a finite number of iterations in a sensor network scenario. Nevertheless, it is not uncommon that the resulting accuracy is reasonable despite the aforementioned factors.

This perspective has proved useful for various problems in sensor network applications including multiple target tracking. Data association is central to multiple target tracking in which observations are associated with targets that has induced them. It is NP-hard in general and conventional approaches such as MHT [47] and JPDA [48] together with their extensions (see e.g. [49] [50]) lack scalability or can not handle non-Gaussian non-linear dynamics [51] in a sensor network scenario which requires in-network processing for all the relevant tasks for tracking [20]. In [52], association hypoteses are represented through discrete random variables which render a Markov random field representation with a pairwise factorization for the joint density in accordance with the sensor-target coverage information. The most likely hypothesis is found using an approximate MAP estimation scheme on graphs with loops. This scheme is amenable for distributed processing through the mapping of the information graph such that each platform contains the association variables related to itself. It is also preferable in a centralized setting for being suitable for parellel processing and producing approximate results of reasonable quality to an NP-hard problem.

Similar treatment of signal and information processing problems arising in applications such as self-localization in sensor networks, ambient temperature estimation via sensor platforms distributed over a region, target localization problems as well as conventional settings such as detection is possible(see for example [53] [54] [55] [24] [56] [5] [57]).

2.5 Monte Carlo Integration

The tradeoff between the soundness of the model and the corresponding computational demands is a fundamental issue in statistical inference. For example, throughout the previous discussion on probabilistic graphical models, we have addressed the scalability of the scheme in the number of variables. In this framework, one aspect of the model complexity is the de-

gree of connectedness which corresponds to the intensity of the dependency relations among the variables. The message passing computations exhibit a complexity in the number of variables in accordance with the sparsity of the graph. Another aspect involves the computation of the expressions involved. For example, in a Bayesian setting, estimation of a random vector requires computation of integrals with no closed form solutions in general. We can either compromise the model accuracy and restrict ourselves to standard distributions with well known expressions under certain operators or compromise point accuracy and proceed in a numerical approximation framework.

Conventional numerical integration based on quadratures yield exponential increase in the the number of grid points as well as computations with the dimensionality of the domain of integration. Hence they lack scalability with the number of random variables in our context. Numerical approaches that are based on grid points produced by random number generators, also called Monte Carlo Methods, attempt to aviod the situation together with possible intricacies related to the particular integrand [58] [59].

In the conventional Monte Carlo method, the value of an integral is of concern, i.e.

$$i = \int_{\mathcal{X}} dx\, p(x) f(x) \tag{2.17}$$

where i is bounded and $p(x)$ has a bounded integral over $\mathcal{X}$. It is often the case that $\mathcal{X}$ is of multiple dimensions and the right hand side of Eq. 2.17 is a multi-dimensional integral. Without loss of generality, it is possible to assume that $p(x)$ integrates to 1 over the set of all possible values of x, i.e. $\int_{\mathcal{X}} dx\, p(x) = 1$ and hence admits a joint probability density interpretation.

Such integrals often arise in Bayesian inference with continuous variables. For example, in recursive Bayesian filtering, the predictive density of the state of a dynamical system at time t, i.e. X^t, given posterior distribution of the previous state based on all the observations collected up to time $t-1$, i.e. $p(x^{t-1}|\{y^i\}_{i=1}^{t-1})$, and the state transition density $p(x^t|x^{t-1})$ is given by

$$p(x^t|\{y^i\}_{i=1}^{t-1}) = \int_{\mathcal{X}^{t-1}} \mathrm{d}x^{t-1}\, p(x^t|x^{t-1}) p(x^{t-1}|\{y^i\}_{i=1}^{t-1})$$

under the assumption that $X^t \perp\!\!\!\perp \{Y^i\}_{i=1}^{t-1} \,|\, X^{t-1}$. This expression is simply Eq. (2.17) for all $x^t \in \mathcal{X}^t$ with $f(x) = p(x^t|X^{t-1} = x)$ and $p(x) = p(X^{t-1} = x|\{y^i\}_{i=1}^{t-1})$. It is not possible to compute this integration exactly, in general, unless the distributions involved has adequate forms such as Gaussians.

2.5.1 The Conventional Monte Carlo Method

Provided that we are able to generate M independent samples $\{x^{(1)}, x^{(2)}, ..., x^{(M)}\}$ from the density $p(x)$, an estimate for $i = E_{p_X}\{f(X)\}$ is given by

$$\hat{i}_M = \frac{1}{M}\sum_{k=1}^{M} f(x^{(k)}) \tag{2.18}$$

and by the Strong Law of Large Numbers (SLLN) converges to i almost surely, i.e.

$$\lim_{M\to\infty} \hat{i}_M \to i \text{ with probability } 1$$

On the other hand, in order to estimate the density function $p(x)$ given M independent samples drawn from it, an appropriate approach is the Parzen Window estimate given by

$$p_M(x) = \frac{1}{M}\sum_{k=1}^{M} \frac{1}{h} K\left(\frac{x - x^{(k)}}{h}\right)$$

where $K(x)$ is a weighting function for which a possible choice is

$$K(x) = \begin{cases} 1/2 & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and $h = h(M)$ such that $\lim_{M\to\infty} h(M) \to 0$. Under some mild conditions

$$\lim_{M\to\infty} E\{p_M(x)\} = p(X)$$

with finite variance at all points x of continuity of $p(x)$ rendering $p_M(x)$ an unbiased and consistent estimator of $p(x)$[60]. The twist of the dependence of h on M and choosing $h \to 0$ regardless of M yields an emprical estimate

$$\tilde{p}_M(x) = \frac{1}{M}\sum_{k=1}^{M} \delta(x - x^{(k)})$$

which substituted in place of $p(x)$ in Eq.(2.17) yields exactly Eq.(2.18). Then the convergence of i_M follows the convergence of $p_M(x)$. The estimation error corresponding to Eq.(2.18) is asymptotically normal denoted by

$$\lim_{M\to\infty} \sqrt{M}(\hat{i}_M - i) \sim \mathcal{N}(0, \sigma_g^2)$$

where $\sigma_g^2 = E\{(g(x) - i)^2\}$. The rate of convergence is independent of the dimension of the integrand [61] rendering this approach particularly preferable to grid based methods in the case of multi-dimensional integrations.

2.5.2 Importance Sampling

The *importance sampling method* is used if it is not possile to sample from $p(x)$ but from the *importance sampling density* $g(x)$. An alternative representation of Eq.(2.17) as $i = \int_{\mathcal{X}} dx\, g(x) \frac{p(x)}{g(x)} f(x)$ together with M independent samples $\{x^{(1)}, x^{(2)}, ..., x^{(M)}\}$ generated from the density $g(x)$ lead to the estimator

$$\begin{aligned} \hat{i}_M &= \frac{1}{M} \sum_{k=1}^{M} \frac{p(x^{(k)})}{g(x^{(k)})} f(x^{(k)}) \\ &= \frac{1}{M} \sum_{k=1}^{M} \omega_{(k)} f(x^{(k)}) \end{aligned} \tag{2.19}$$

where $\omega_{(k)}$ is identifed as $\omega_{(k)} = p(x^{(k)})/g(x^{(k)})$. This estimate converges to $E_{p_X}\{f(X)\}$ almost surely for reasons similar to the convergence of the conventional Monte Carlo method (SLLN), provided that $g(x)$ is non zero for all values of x where $f(x)$ is non zero, i.e. $supp(g) \supset supp(f)$. The rate of convergence is related to the error variance and importance sampling densities achieving a finite variance are found out to be, loosely speaking, mimicing $p(x)$ [62].

An alternative estimator which relaxes this restriction on the choice of the importance sampling density is given by

$$\hat{i}_M = \frac{1}{\sum_{k=1}^{M} \omega_{(k)}} \sum_{k=1}^{M} \omega_{(k)} f(x^{(k)}) \tag{2.20}$$

which again converges almost surely to i.

In addition, if the conditions $E\{\omega(X)\} < \infty$ and $E\{f^2(X)\omega(X)\} < \infty$ are satisfied, then as $M \to \infty$, $\sqrt{M}(\hat{i}_M - i) \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = E\{(f(X) - i)^2 \omega(X)\}$ [63]. In other words, provided that the conditions hold, the estimates i_M for different sample sets are "consistent" such that as M increases, the estimation errors converge in distribution to a zero mean Gaussian with decreasing variance. For cases where most of the weights are relatively small and dominated by the rest, this esimator is preferable to the one given by Eq.(2.19) altough it is slightly biased for small M [62]. Similar to the conventional Monte Carlo Method, the rate of convergence is independent of the dimensionality of the integrand.

Monte Carlo methods (together with sample generation through Markov Chains) have been successfully applied for many Bayesian inference problems involving non-Gaussian distributions and/or nonlinear dynamics (see e.g. [64] [65] [66]) including particle filtering for

Bayesian recursive filtering, e.g. in target tracking [67], and non-parameteric belief propagation [68] for inference in non-Gaussian Markov random fields with continuous variables, e.g. in vision based tracking [69]. In this work, we employ Monte Carlo methods for optimization rendering a Monte Carlo optimization framework.

2.6 Decentralized Statistical Inference Under Communication Constraints

In the preceding sections, it is pointed out that Bayesian inference is involved with construction of distributions based on the observations (e.g. the joint a posteriori density functions or the posterior marginals) and integral operators over them (e.g. the expected values of certain posterior marginals). It is also often the case that exactly performing the required computations has high complexity and adequate approximations are inevitable. Additional constraints are introduced in the context of sensor networks due to the limited computation resources, communication bandwidth and energy budget.

The primary consideration of in-network processing is to have the platforms perform the necessary computations in a collaborative manner (preferably while taking the arising communication load into account). From a global point of view, it is possible to treat the network as evaluating a predetermined function γ at the set of observations $y = (y_1, y_2, ..., y_N)$ (periodically, considering the case in which the online processing proceeds after the measurements are collected at each time step) and decompose γ in terms of local computations yielding a directed acyclic graph representation in which the vertices are computation tasks and edges are the transmissions of the outputs from the vertices they are emanating from to the vertices they are terminated at (see e.g. Chp. 3 in [70] or Chp. 4 in [25]). For example, one trivial decomposition is $\gamma(y_1, y_2, ..., y_N) = \gamma_0(\gamma_1(y_1), \gamma_2(y_2), ..., \gamma_N(y_N))$ where the local function γ_i for $i \in \{1, 2, ..., N\}$ is the identity function and $\gamma_0 = \gamma$, and this decomposition has a "star" shaped tree representation in which each leaf node bears γ_i for $i \in \{1, 2, ..., N\}$ directed to the root node associated with γ_0. After mapping the vertices of the computation graph to the sensor platforms, in-network processing is achieved with a certain communication load since the observations are collected by distinct platforms. The centralized approach corresponds to mapping identity functions with all the platforms except associating γ_0 (together with the corresponding γ_i) with the designated fusion center in the previous example.

On the other hand, there is a number of factors that affect the communication load. First of all, in the case that some y_is take values from denumerable sets and γ has a denumerable range set, a discretization of γ through its domain and range sets are required in order the resulting scheme to be realizable by digital machinery networked over limited capacity links. Let us denote the resulting approximate scheme which is the realizable global function with $\tilde{\gamma}$. Another factor related to the communication load is then, the decomposition properties of $\tilde{\gamma}$. For each possible decomposition and each valid mapping, a certain communication and computation structure over the network arise with a certain load together with a certain cost of communications. The bounds related to the communication load arising for in-network computation has been investigated for different structures that $\tilde{\gamma}$ exhibits and for different network types(e.g. collocated networks in which every node is in 1-hop neighborhood with the remaining and broadcasts through an assigned time slot in the multiple access channel or random planar networks which is the case described in Section 1.1.1) [71]. It is also possible to consider, among some possible mappings of the computation graph onto the network graph, the one achieving less cost for the transmissions (see e.g. Chp. 4 in [70] and Chp. 4 and 5 in [25]). To the best knowledge of the author' s, a general problem setting is highly complicated providing small chances of finding a feasible solution that is optimal in some sense. It is rather often the case that the solution involves heuristics and it is not straigtforward to obtain a scheme which also complies with the in-network processing paradigm and is amenable for network self-organization.

The previous discussion aspires to provide a perspective in which given γ, a communication and computation strategy over the network structure is sought that results with a judicious utilization of communication resources. In this setting, the degredation in the accuracy is due to substituting $\tilde{\gamma}$ for the global function computed because of the fact that it is suitable to realization by the sensor network. However, it is not easy -if not impossible- to address the tradeoff between the utilization of communications and the resulting accuracy such that a gracefull degradation of the inference performance is achieved while decreasing the communication load.

For example, consider γ in the context of Bayesian inference where it is reasonable to assume it as one of the estimators (or detectors) presented in Example 2.2.1. In addition, it is possible to handle the complexity of computing γ and obtain $\tilde{\gamma}$ through probabilistic graphical models together with message passing algorithms considering the dependencies among variables and

Monte Carlo methods considering the representation and integration of non-parametric forms (e.g. Nonparametric Belief Propagation [72]). For the case, given a graphical model and a schedule regarding a certain message passing algorithm, a computation graph is obtained (see e.g. [73]) for which a trivial mapping onto the platforms exists, provided that the locality of observations hold (see the discussion in Section 2.4).

On the other hand, only for a limited number of cases it has been possible to consider the cost of communications while deciding on this mapping. For example, in [74], a junction tree representation is considered together with a communication network which provides links between any two sensor platforms with appropriately assigned costs. An energy-aware in-network processing scheme is obtained through handing-off variables associated with nodes of the tree while keeping the information structure valid (consider, e.g. Figure 2.4 where we present possible junction trees for the same information structure) such that the resulting communication cost is reduced (Recall that message passing over junctions trees result with exact clique potentials). In this setting, the trade-off is not addressed in the sense that it is not possible to have the estimation accuracy degrade gracefully as the utilization of communications is decreased.

In the case of loopy graphs for which a junction tree representation is not useful, it has been the usual practice to consider the communication constraints and energy-awareness after the (trivial) mapping. Recall that loopy versions of message passing algorithms yield an infinite schedule and converge asymptotically if they do, most often to approximate results. In the context of sensor network additional modifications are utilized. First of all, a finite schedule is selected which can be represented similarly by a computation graph and result with a successive approximation. One possible method to modify the computation graph during online processing is to censor the messages which has not been changed more than an amount with respect to a measure. This approach is utilized in [75] where decision of sending or censoring a message is given individually by sensor platforms regardless of any other information such as states of the neighboring platforms or penalty for communication, resulting a myopic behavior. Second, the limited bandwidth requires that finite representations should be employed for the messages which are non-parametric functions over denumerable domains in the case of continous variables, in general. A discretization is possible through Monte Carlo methods which yield an approximation framework for Belief Propagation (known as Nonparametric Belief Propagation) [72]. Moreover, encoding of the messages considering

the cost of transmissions and induced errors due to successive approximations has also been investigated (Chp. 5 in [46]).

Hence, it is possible to utilize successive approximations for message passing algorithms in order to have them suitable to be realized by a sensor network and analyze the effects of message errors introduced by the approximation framework on the resultant distributions [76]. However, it is not straightforward to address the estimation accuracy as well as its tradeoff between the utilization of communications. Note that, all the examples above presume a certain global function γ in the beginning and seek an approximation $\tilde{\gamma}$ that can be realized by the sensor network for a reasonable amount of communications.

On the other hand, statement of a design problem in a communication constrained environment is concern to us, i.e. given available communication links and their capacities together with costs of transmissions over them, we seek a communication and computation structure γ which achieves a certain accuracy for a certain cost of communications. We are also interested in schemes that exhibit graceful degredation of estimation accuracy by decreasing communication load which has not been provided by the perspectives described above.

One possible mathematical statement of the design setting is in the form of an optimization problem in which a global function γ is sought together with a communication and computation structure (or online processing scheme or strategy) which realizes γ over the network. Suppose that the available links and their capacities are taken into account by this structure. Provided that tractable measures for the estimation accuracy and the cost of transmissions arising in the network are obtained, the in-network strategy with a desired performance can be selected, in principle. However, we are not guaranteed to achieve a well-posed setting for strategy optimization. On the other hand, it is possible to improve the behaviour of the solution if a useful characterization of γ can be utilized through in a multiple objective framework.

In [29], such an approach is utilized for decentralized detection and three classes of decentralized strategies are considered. The first class is constituted by structures that yield a directed acyclic computation graph rendered over the set of available links. Each vertex corresponds to a platform and the uni-directional transmissions modelled with the directed edges comply with the corresponding link capacities. Strategies that allow bi-directional transmissions and render an undirected graph over the underlying links compose the second class such that the messaging and inference are performed in two successive stages. Strategies from both classes

yield tractable Bayesian risks that bears both penalty for decision errors and transmissions. Moreover, a Team Theoretic investigation of the problem yields that the strategies that satisfy an equilibrium condition admit a useful characterization through finite dimensional vectors and starting with an initial strategy, it is possible to find a better one iteratively regarding a dual objective Bayesian risk. The third class, namely "multi-stage architectures" extend the two-stage approach on undirected graphs and yield subtle strategy optimization schemes involving heuristics.

2.7 Optimization of Decentralized Detection Networks

There is a plethora of work on decentralized detection networks[5] including earlier investigations involving a "star" network (also referred to as a parallel network) in which peripheral nodes collect observations induced by one of a finite number of hypothesis and based on them, transmit messages to a fusion center which performs the decision and mappings local to the nodes are of concern [77, 78, 79, 80]. An overview of the general Bayesian setting given in [33] is as follows: Denoting with 0 the fusion center and letting the peripherals take labels from the set $\{1, 2, ..., N\}$, each peripheral sensor $i \in \{1, 2, ..., N\}$ makes an observation $y_i \in \mathcal{Y}_i$ due to a discrete random variable X which takes values from a finite set $\mathcal{X}$ and transmit a symbol u_i from a finite set $\mathcal{U}_i$ to the fusion center based on y_i. The fusion center produces $\hat{X}$, based on the incoming messages $\{u_i\}_{i\in\{1,2,...,N\}}$. The conventional centralized treatment given in Section 2.2, requires each peripheral sensor to transmit y_i to the fusion center. As soon as $|\mathcal{U}_i| << |\mathcal{Y}_i|$ the decentralized scheme requires much less communication bandwidth. Also similar to the discussion for the conventional setting, it is possile to treat each node of the network evaluating a local function given by $\gamma_i : \mathcal{Y}_i \to \mathcal{U}_i$ and $\gamma_0 : \mathcal{U}_1 \times \mathcal{U}_2 \ldots \times \mathcal{U}_N \to \mathcal{X}$ in the case of a peripheral and the fusion center respectively. Note that the processing is distributed to the sensor platforms through functions $\{\gamma_i\}_{i\in\mathcal{V}}$ whereas a single function $\gamma : \mathcal{Y} \to \mathcal{X}$ is of concern in the centralized approach. Optimization of the network in a Bayesian setting refers to the search for $\{\gamma_i^*\}_{i\in\mathcal{V}}$ where $\mathcal{V} = \{0, 1, 2, ..., N\}$ such that the expected value of a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ penalizing $(x, \hat{x})$ pairs, i.e. $E\left\{c(x, \hat{x}); \{\gamma_i^*\}_{i\in\mathcal{V}}\right\}$ is minimum.

This treatment leads to a problem definition similar to ($P0$) given by Expression (2.3) for

[5] We refer to both binary decision problems and M-ary hypothesis testing as detection throughout, emphasizing that in both a decision among finitely many choices are made.

which the observation space would be $\mathcal{Y} = \underset{i\in\mathcal{V}}{\otimes} \mathcal{Y}_i$. However, the argument of the optimization can take values from a proper subset of $\Gamma = \{\gamma|\gamma : \mathcal{Y} \to \mathcal{X}\}$, i.e. the set of functions mapping the set of observations to the set from which X takes values from as defined in Section 2.2. Let us define the strategy $\gamma_\star = (\gamma_0, \gamma_1, ..., \gamma_N)$ with $\{\gamma_i\}_{i\in\mathcal{V}}$ as defined above. It is clear that $\gamma_\star$ is a function which maps $\mathcal{Y}$ on to $\mathcal{X}$. But, defininining the set of possible local rules for platform i by $\Gamma_i = \{\gamma_i|\gamma_i : \mathcal{Y}_i \to \mathcal{U}_i\}$ and the set of possible strategies by $\Gamma^\star = \underset{i\in\mathcal{V}}{\otimes} \Gamma_i$ from which the strategy $\gamma_\star$ takes values , it is apparent that $\Gamma^\star \subset \Gamma$. Then, the problem of concern is

$$\begin{aligned} (\mathrm{P}\star): \qquad & \min E\left\{c(x, \hat{x}); \gamma\right\} \\ & \text{subject to } \gamma \in \Gamma^\star \end{aligned} \tag{2.21}$$

It is possible to treat (P$\star$) as a team decision problem in which it is often the case that the global optimal solution(s) can be found only for "highly structured" cases. It is rather feasible to find a solution that satisfies an equilibrium condition in an iterative manner yielding local rules in a variational form [33].

Besides the star topology, network settings such as tandem (or series) [81] and tree architectures [82] have been studied including the case in which multiple decision variables, also referred to as multiple (local) event structures, are of concern [83] (see e.g. [84] for a review). In [28], three strategy classes, two of which are of concern throughout, are presented that extend the literature on optimization of decetentralized detection networks in a number of aspects including the set of communication constraints introduced in the model. Similar to (P$\star$), the optimization of these strategies involve finding local rules for platforms that minimize a Bayesian risk which, in addition, captures the cost of transmissions as well as penalty for the decision errors. The structures that the problem exhibit together with that of corresponding Team Theoretic solutions under various assumptions are investigated.

The first class is constituted of strategies that admit a directed acyclic graph representation in which the edges correspond to unreliable uni-directional communication links (see Chp.3 in [28], [85, 27]). Extending the "star" topology, it is also possible to distribute the decision process in the network through associating random variables with the vertices which correspond to the platforms. Under certain assumptions, a team theoretic investigation yields an efficient optimization procedure which admits a message passing algorithm and exhbits a graceful degredation of the decision performance as the utilization of communication is re-

duced. We present an overview of this class in Section 2.7.1 which also captures the literature aforementioned.

The second class we consider covers strategies which utilize bi-directional communication and can be represented with an undirected graph (see Chp.4 in [28] and also [26]). In order to have a causal online processing scheme the local rules are restriced a two stage operation in which the communication rules are executed based on the observations first. As soon as a platform receives all incoming messages, it performs the decision rule based on both these messages and its observation. Under certain assumptions, the class of two stage strategies on undirected graphs, for which we provide a brief review in Section 2.7.2, exhibit properties similar to that of the strategies on directed acyclic graphs.

2.7.1 Directed Graph Constrained Networks

In this section we summarize optimization of decentralized strategies defined over directed acyclic graphs (the interested reader is referred to Chp. 3 in [28] for a detailed discussion). We exclude unrelaible channel models and assume that all links provide error-free communications.

2.7.1.1 Online Processing Model and Problem Definition

A directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents a communication and computation structure for a decentralized system where each platform is associated with a node $v \in \mathcal{V}$. An edge $(i, j) \in \mathcal{E}$ corresponds to the finite capacity communication link from platform i to j on which i can transmit a symbol $u_{i\to j}$ from the set of admissible symbols $\mathcal{U}_{i\to j}$ which is finite and the number of elements $\left|\mathcal{U}_{i\to j}\right|$ is in accordance with the link capacity capturing the bandwidth constraints. For example, it is possible to represent a link with capacity $\log_2 d_{ij}$ bits with $\mathcal{U}_{i\to j}$ such that $\left|\mathcal{U}_{i\to j}\right| = d_{ij} + 1$ where $0 \in \mathcal{U}_{i\to j}$ holds indicating no transmission and enabling a message cencoring scheme.

Let $u_{\pi(j)}$ denote the incoming messages to node j from the parent nodes $\pi(j)$, given by $u_{\pi(j)} \triangleq \{u_{i\to j} | i \in \pi(j)\}$. Let $\mathcal{U}_{\pi(j)}$ denote the set from which $u_{\pi(j)}$ takes values from. This set is constructed through consecutive Cartesian products given by $\mathcal{U}_{\pi(j)} \triangleq \bigotimes_{i\in\pi(j)} \mathcal{U}_{i\to j}$. The set of outgoing messages from node j to child nodes $\chi(j)$, given by $u_j \triangleq \{u_{j\to k} | k \in \chi(j)\}$

takes values from the set $\mathcal{U}_j$ which can be defined in a similar way with that for $\mathcal{U}_{\pi(j)}$ as $\mathcal{U}_j \triangleq \underset{k\in\chi(j)}{\otimes} \mathcal{U}_{j\to k}$.

In addition, each node j is associated with a random variable(s) X_j that takes values from the set $\mathcal{X}_j$ and hence the inference is distributed in the network. Note that this mapping is arbitrary, in principle, and provides a broad range of possibilities for decentralized inference. As node j measures $y_j \in \mathcal{Y}_j$ and receives $u_{\pi(j)} \in \mathcal{U}_j$, it evaluates a function, called its local rule, defined by $\gamma_j : \mathcal{Y}_j \times \mathcal{U}_{\pi(j)} \to \mathcal{U}_j \times \mathcal{X}_j$ which produces an estimate $\hat{x}_j \in \mathcal{X}_j$ as well as outgoing messages $u_j \in \mathcal{U}_j$. The space of rules local to node j is given by $\Gamma_j^{\mathcal{G}} \triangleq \{\gamma_j | \gamma_j : \mathcal{Y}_j \times \mathcal{U}_{\pi(j)} \to \mathcal{U}_j \times \mathcal{X}_j\}$ where the superscript $\mathcal{G}$ denotes that the definition of the set relies on $\mathcal{G}$ together with the set $\{\mathcal{U}_{i\to j} | (i, j) \in \mathcal{E}\}$.

As pointed out in Section 2.1.1 a DAG implies a partial ordering, e.g. in accordance with the reachability relation. In Figure 2.6 we present an example DAG together with forward and backward partial orderings in which the parentless and the childless nodes has the smallest order respectively. The directed acyclic nature of $\mathcal{G}$ leads a causal online processing of the observations when the nodes execute their local rules in accordance with the forward partial order, i.e. starting from the parentless nodes, at each step, nodes with the corresponding order evaluate their local rules and stop after the childless nodes perform theirs which do not involve deciding on symbols to transmit. The example DAG illustrated in Figure 2.6 yields a processing completed in 4 steps. Note that, all edges corresponding to communication links are invoked once in this scheme and the process from node j' s point of view is illustrated in Figure 2.7.

The aggregation of local rules denoted by γ is called a strategy, i.e. $\gamma = (\gamma_1, \gamma_2, ..., \gamma_N)$ and takes values from the set of feasible strategies given by $\Gamma^{\mathcal{G}} = \underset{v\in\mathcal{V}}{\otimes} \Gamma_v^{\mathcal{G}}$. Considering (P0) given by Expression (2.3) we note that $\Gamma^{\mathcal{G}} \subset \Gamma$. The communication load of the system is the set of all transmitted symbols during the processing given by $u \triangleq \{u_{i\to j} | (i, j) \in \mathcal{E}\}$ and takes values from the set $\mathcal{U} \triangleq \underset{(i,j)\in\mathcal{E}}{\otimes} \mathcal{U}_{i\to j}$. Also note that the random vector of concern, i.e. X, takes values from the set $\mathcal{X} = \underset{j\in\mathcal{V}}{\otimes} \mathcal{X}_j$ and similarly, the observation Y takes values from the set $\mathcal{Y} = \underset{j\in\mathcal{V}}{\otimes} \mathcal{Y}_j$. The global view of the scenario is illustrated in Figure 2.7.

The Bayesian risk function c for the network described by the graph $\mathcal{G}$ is selected such that a decision error penalty for the pair $(x, \hat{x})$ and a cost due to the corresponding communication load u are assigned. Hence $c : \mathcal{U} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and the objective function to be minimized

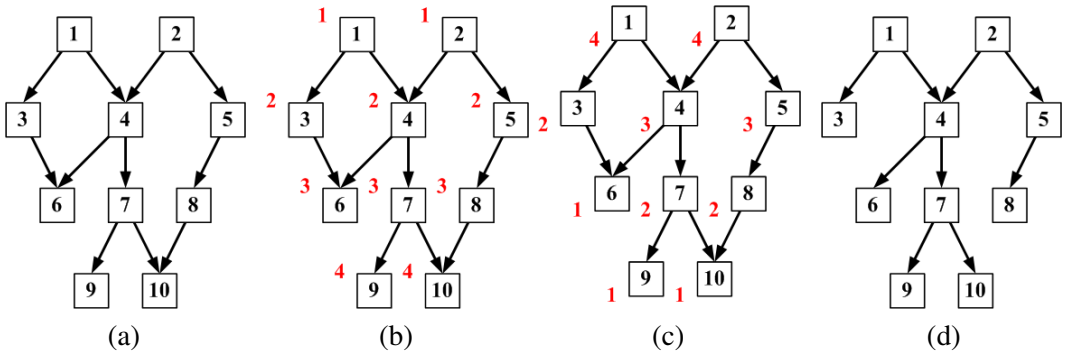


Figure 2.6: (a) An example directed acyclic graph $\mathcal{G}$. (b) $\mathcal{G}$ with the forward orders on the upper left side of the nodes. Ordering starts with assigning 1 to the parentless nodes and proceeds with removing them from $\mathcal{G}$, increasing the count by one and repeating the same procedure untill all nodes are counted. (c) $\mathcal{G}$ with the backward orders on the lower left side of the nodes. Ordering starts with assigning 1 to the childless nodes and proceeds similar to the forward ordering. (d) A polytree in $\mathcal{G}$. Note that no nodes has two parents that have common ancesors or two children that has common descendants.

is given by $J(\gamma) = E\{c(u, x, \hat{x}); \gamma\}$ where the expectation is over $p(u, \hat{x}, x; \gamma)$. Therefore, the problem of finding the best strategy for inference under communication constraints described by $\mathcal{G}$ turns to a constrained optimization problem given by

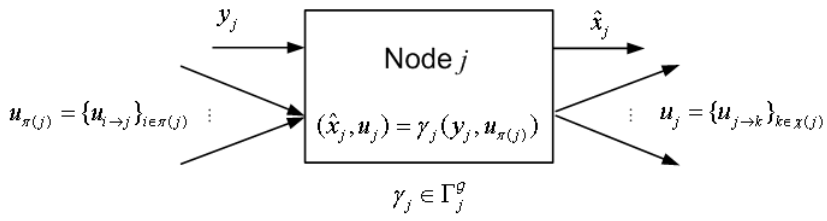
$$\begin{aligned} (\text{P1}): \quad & \min E\{c(u, x, \hat{x}); \gamma\} \\ & \text{subject to } \gamma \in \Gamma^{\mathcal{G}} \end{aligned} \tag{2.22}$$

In order to show that for a given deterministic strategy $\gamma \in \Gamma^{\mathcal{G}}$, there exists an expected cost value $J(\gamma)$, consider the underlying distribution. The equality given by

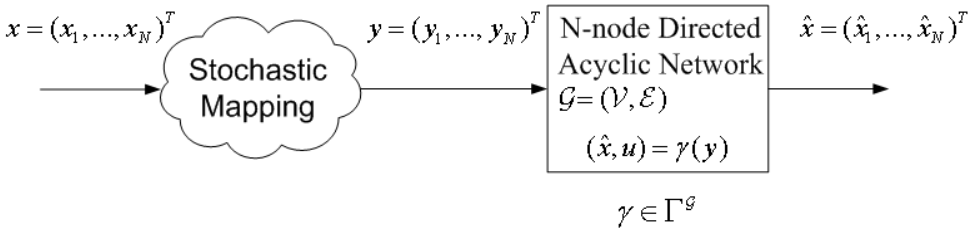
$$p(u, x, \hat{x}; \gamma) = \int_{y \in \mathcal{Y}} \mathrm{d}y \; p(u, \hat{x}|x, y; \gamma) p(x, y) \tag{2.23}$$

is satisfied which reveals the connection between the strategy γ and the objective function through the distribution $p(u, \hat{x}|x, y; \gamma)$. We note that the strategy γ can be viewed as a function given by $\gamma : \mathcal{Y} \to \mathcal{X} \times \mathcal{U}$. Therefore it specifies the corresponding distribution $p(u, \hat{x}|x, y; \gamma)$. The causal online processing provided by the partial ordering on the directed acyclic graph starting from the parentless nodes implies that the local rules are coupled into this expression by

$$p(u, \hat{x}|x, y; \gamma) = \prod_{j=1}^{N} p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j) \tag{2.24}$$



(a)Node j point of view.



(b) Global view.

Figure 2.7: Online processing scheme modelled with a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$: (a) The viewpoint of node j in $\mathcal{G}$ which evaluates its local rule γ_j on its measurement y_j as well as the received messages $u_{\pi(j)}$ and produces an inference on the value of the random variable it is associated with, i.e. $\hat{x}_j$, together with outgoing messages u_j to its children, (b) The global view of the decentralized decision strategy on $\mathcal{G}$ where a random process X takes the value x as the outcome of an experiment and induces observations y on the decentralized system represented with $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

where the node labels recalled by j, without loss of generality, are in accordance with the forward partial ordering on $\mathcal{V}$ (The interested reader is referred to Section A.1 for further details). We note similarly that $p(u_j, \hat{x}_j | y_j, u_{\pi(j)}; \gamma_j)$ for each $j \in \mathcal{V}$ is specified by γ_j. Since both $\mathcal{X}_j$ and $\mathcal{U}_j$ are of finite cardinality, this specification is given by

$$p(u_j, \hat{x}_j | y_j, u_{\pi(j)}; \gamma_j) = \delta_{\hat{x}_j, [\gamma_j(y_j, u_{\pi(j)})]_{\mathcal{X}_j}} \delta_{u_j, [\gamma_j(y_j, u_{\pi(j)})]_{\mathcal{U}_j}} \tag{2.25}$$

where δ is the Kronecker delta.

In conclusion, given a set of local rules $\{\gamma_j\}$ for $j \in \mathcal{V}$ constituting a strategy γ in accordance with a DAG $\mathcal{G}$, an online processing is implied in a decentralized fashion. In addition, it is possible to compute the corresponding expected cost due to both decision errors and cost of communications by substituting Eq.s (2.23– 2.25) in $J(\gamma)$.

2.7.1.2 Team Theoretic Solution

In a team decision problem, N members constitute a team such that each can take actions $\gamma_j \in \Gamma_j$ with a single cost function $J(\gamma_1, \gamma_2, ..., \gamma_N)$. The problem is to find $\gamma_j^* \in \Gamma_j$ for $j = 1, 2, ..., N$ such that $J(\gamma_1^*, \gamma_2^*, ..., \gamma_N^*)$ is minimum. It is often the case that finding a solution to this problem is hard. A relaxation is to find a Nash equilibirium solution where no change in a single member' s action yields a better cost, i.e. Find $\gamma^* = (\gamma_1^*, ..., \gamma_n^*)$ such that

$$\gamma_j^* = arg \min_{\gamma_j \in \Gamma_j} J(\gamma_j, \gamma_{\setminus j}^*) \tag{2.26}$$

for all $j \in \{1, 2, ..., N\}$. Such a solution is also said to be person-by-person optimal.

It can easily be shown that, given $\gamma^0 = (\gamma_1^0, ..., \gamma_N^0)$ where $\gamma^0 \in \Gamma$ and $\Gamma = \underset{j \in \{1,2,...,N\}}{\otimes} \Gamma_j$ with the corresponding cost $J(\gamma^0)$, $J(\gamma^1) \leq J(\gamma^0)$ where $\gamma_{\setminus j}^1 = \gamma_{\setminus j}^0$ and $\gamma_j^1 = \arg\min_{\gamma_j \in \Gamma_j} J(\gamma_j, \gamma_{\setminus j}^0)$ for any j. Hence, starting from an initial strategy, an iteration that converges to a person by person optimal strategy in a coordinate descent manner is given by Algorithm 1.

Algorithm 1 Iterations converging to a person-by-person optimal strategy.

0) `(Initiate)` $l = 0$;

Choose $\gamma^0 = (\gamma_1^0, \gamma_2^0, ..., \gamma_N^0)$ such that $\gamma_j \in \Gamma_j$ for $j = 1, 2, ..., N$

1) `(Update)` $l = l + 1$;

`for` $j = N, N-1, ..., 1$

$\gamma_j^l = \arg \min_{\gamma_j \in \Gamma_j} J(\gamma_1^{l-1}, ..., \gamma_{j-1}^{l-1}, \gamma_j, \gamma_{j+1}^l, ..., \gamma_N^l)$

2) `(Check)` `If` $J(\gamma^{l-1}) - J(\gamma^l) < \varepsilon$ `STOP, else GO TO 1;`

Team Decision framework has been introduced in various fields such as optimal control and decentralized detection (see for example [86] and [33]). Note that (P1) given by Expression (2.22) is already suitable for such a treatment. In the next section, the corresponding person-by-person optimality conditions in the context of the decentralized detection networks constrained by graphs and expressions corresponding to steps of Algorithm 1 are presented.

2.7.1.3 Optimizing Directed Graph Constrained Networks

Consider $J(\gamma)$, and the decentralized detection problem under communication constraints, i.e. (P1). The directed acyclic network constraints guarantee that for the underlying distribu-

tion the equality given by Eq.(2.24) holds. Moreover, it can be shown that further structure is introduced in problem (P1) through certain assumptions which lead to a scalable coordinate descent procedure that converges to a person-by-person optimal solution.

Assumption 1*(Conditional Independence)***:** This assumption states that, given the state of the process X, the observations are conditionally independent, i.e. $p(x, y) = p(x) \prod_{i=1}^{N} p(y_i|x)$, which is a consequence of the fact the noise processes of sensors are mutually independent.

Proposition 2.7.1 *Consider (P1). Under Assumption 1, given a person-by-person optimal strategy*

$\gamma^* = (\gamma_1^*, \gamma_2^*, ..., \gamma_N^*)$ *and fixing the local rules other than the* j^{th} *at the optimal, i.e.* $\gamma_{\setminus j} = \gamma_{\setminus j}^*$, *the* j^{th} *optimal rule given by Eq.(2.26) is equivalent to*

$$\gamma_j^*(Y_j, U_{\pi(j)}) = arg \min_{(u_j, \hat{x}_j) \in \mathcal{U}_j \times \mathcal{X}_j} \sum_{x \in \mathcal{X}} p(Y_j|x) \theta_j^*(u_j, \hat{x}_j, x; U_{\pi(j)}) \tag{2.27}$$

with probability 1*, where*

$$\theta_j^*(u_j, \hat{x}_j, x, u_{\pi(j)}) = p(x) \sum_{u_{\setminus \{j\} \cup \pi(j)}} \sum_{\hat{x}_{\setminus j} \in \mathcal{X}_{\setminus j}} c(u, \hat{x}, x) \prod_{i \neq j} \int_{\mathcal{Y}_i} \mathrm{d}y_i \, p(y_i|x) p(u_i, \hat{x}_i|y_i, u_{\pi(i)}; \gamma_i^*) \tag{2.28}$$

Proof. See e.g. [28]. We provide a similar proof which differs in that it relies on the correspondance between the j^{th} local rule, i.e. γ_j, and the conditional density $p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j)$. Consider the expected risk $J(\gamma) = J(\gamma_j, \gamma_{\setminus j})$ after substituting the mathematical statement of Asumption 1, Eq.(2.24) and $\gamma_{\setminus j} = \gamma_{\setminus j}^*$ in Eq.(2.23), i.e.

$$\begin{aligned} J(\gamma_j, \gamma_{\setminus j}^*) &= \sum_{x \in \mathcal{X}} \sum_{\hat{x} \mathcal{X}} \sum_{u \in \mathcal{U}} c(u, x, \hat{x}) p(x) p(u_j, \hat{x}_j|x, u_{\pi(j)}; \gamma_j) \prod_{i \neq j}^{N} p(u_i, \hat{x}_i|x, u_{\pi(i)}; \gamma_i^*) \\ &= \int_{\mathcal{Y}_y} \mathrm{d}y_j \sum_{\hat{x}_j \in \mathcal{X}_j} \sum_{u_j \in \mathcal{U}_j} \sum_{u_{\pi(j)} \in \mathcal{U}_{\pi(j)}} p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j) \sum_{x \in \mathcal{X}} p(y_j|x) p(x) \\ &\qquad \sum_{u_{\setminus \{j\} \cup \pi(j)}} \sum_{\hat{x}_{\setminus j} \in \mathcal{X}_{\setminus j}} c(u, x, \hat{x}) \prod_{i \neq j}^{N} p(u_i, \hat{x}_i|x, u_{\pi(i)}; \gamma_i^*) \end{aligned} \tag{2.29}$$

Considering the fact that "if there exists an optimal rule, then there exist an optimal deterministic rule" which is discussed in Section 2.2; given $Y_j = y_j$ and $U_{\pi(j)} = u_{\pi(j)}$, and assigning $(u_j, \hat{x}_j)$ to $(y_j, u_{\pi(j)})$ in γ_j corresponds to selecting $p(U_j = u_j, \hat{X} = \hat{x}_j|Y_j = y_j, U_{\pi(j)} = u_{\pi(j)}; \gamma_j) = 1$ and zero for $U_j \neq u_j, \hat{X} \neq \hat{x}_j$. $(u_j, \hat{x}_j) \in \mathcal{U}_j \times \mathcal{X}_j$ that yields a minimum value for the integrand in Eq.(2.29) is such that its weight in Eq.(2.29), i.e.

$$\sum_{x \in \mathcal{X}} p(y_j|x) p(x) \sum_{u_{\setminus j} \in \mathcal{U}_{\setminus j}} \sum_{\hat{x}_{\setminus j} \in \mathcal{X}_{\setminus j}} c(u, \hat{x}, x) \prod_{i \neq j} \int_{\mathcal{Y}_i} \mathrm{d}y_i p(y_i|x) p(u_i, \hat{x}_i|y_i, u_{\pi(i)}; \gamma_i^*) \tag{2.30}$$

is minimum[6]. Hence, the conditional distribution constructed by utilizing the procedure above for all $(y_j, u_{\pi(j)})$ with non-zero probability minimizes $J(\gamma_j, \gamma^*_{\setminus j})$ while specifying a deterministic local rule γ_j as

$$\gamma^*_j(y_j, u_{\pi(j)}) = \arg \min_{(u_j, \hat{x}_j) \in \mathcal{U}_j \times \mathcal{X}_j} \sum_{x \in \mathcal{X}} p(y_j|x) \theta^*_j(u_j, \hat{x}_j, x, u_{\pi(j)})$$

with probability 1, where $\theta^*_j(u_j, \hat{x}_j, x, u_{\pi(j)})$ is identified as

$$p(x) \sum_{u_{\setminus \{j\} \cup \pi(j)}} \sum_{\hat{x}_{\setminus j} \in \mathcal{X}_{\setminus j}} c(u, x, \hat{x}) \prod_{i \neq j}^{N} \int_{\mathcal{Y}_i} \mathrm{d}y_i \, p(y_i|x) p(u_i, \hat{x}_i | u_{\pi(i)}, y_i; \gamma^*_i) \tag{2.31}$$

considering the weights of the assignments composing γ_j given by Eq. (2.30). ■

It is possible to regard $\theta_j(u_j, \hat{x}_j, x; u_{\pi(j)})$ as coefficients that provide a finite parameterization for γ_j that minimizes the Bayesian risk given any set of local rules $\gamma_{\setminus j}$ for nodes other than the j^{th} not necessarily fixed at the equilibrium. In other words, it is possible to treat the right hand side of Eq.(2.28) as an operator ψ acting on any set of local rules $\gamma_{\setminus j}$ and produces a finite dimensional coefficient vector θ_j of $|\mathcal{U}_j| \times |\mathcal{X}_j| \times |\mathcal{X}| \times |\mathcal{U}_{\pi(j)}|$ dimensions, i.e $\theta_j = \psi(\gamma_{\setminus j})$. Together with Eq.(2.27) which couples θ_j to the local rule γ_j, it is possible to utilize Proposition 2.7.1 for the `Update` step of Algorithm 1 and obtain an iterative scheme that starts with an initial strategy and converges to a person-by-person optimal one.

However, the person-by-person optimal local rule requires marginalization over X and hence does not scale with the number of random variables of concern. Moreover, evaluation of ψ does not scale with the number of nodes. A further investigation in order to obtain strategies for efficient online processing of observations as well as scalable iterative optimization schemes yields that additional assumptions should hold which corresponds to a structured information architecture.

Assumption 2*(Measurement Locality)***:** Every node j observes y_j due to only x_j, i.e. $p(y_j|x) = p(y_j|x_j)$.

Assumption 2 substitutes the marginalization over $\mathcal{X}$ in Eq.(2.28) with $\mathcal{X}_j$. Therefore the person-by-person optimal j^{th} local rule given by Proposition 2.7.1 can be restated as

$$\gamma^*_j(Y_j, U_{\pi(j)}) = arg \min_{(u_j, \hat{x}_j) \in \mathcal{U}_j \times \mathcal{X}_j} \sum_{x_j \in \mathcal{X}_j} p(Y_j|x_j) \phi^*_j(u_j, \hat{x}_j, x_j; U_{\pi(j)}) \tag{2.32}$$

[6] If the cardinality of the set of minimizers is greater than 1, then one of its elements is selected arbitrarily since we are considering a deterministic local rule and a corresponding distribution.

with probability 1, where

$$\phi_j^*(u_j, \hat{x}_j, x_j; u_{\pi(j)}) = \sum_{x_{\setminus j} \in \mathcal{X}_{\setminus j}} \theta_j^*(u_j, \hat{x}_j, x; u_{\pi(j)}) \tag{2.33}$$

holds. Hence, scalability for online processing in the number of components is achieved. In order the optimization procedure to be efficient, consider the following:

Assumption 3*(Cost Locality)***:** The Bayesian cost function is additive over the nodes $j \in \mathcal{V}$, i.e.

$$c(u, \hat{x}, x) = \sum_{j \in \mathcal{V}} c_j(u_j, \hat{x}_j, x_j) \tag{2.34}$$

Assumption 4*(Polytree Topology)***:** Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a polytree, i.e. $\mathcal{G}$ is a directed acyclic graph with an acyclic undirected counterpart (See the example in Figure 2.6 for comparison).

Proposition 2.7.2 *Consider (P1) and Proposition 2.7.1. If Assumptions 3 and 4 hold in addition to Assumption 1 and 2, then Eq.(2.32) applies with a proportionality expression for* $\phi_j^*(u_j, \hat{x}_j, x_j; u_{\pi(j)})$ *given by*

$$\phi_j^*(u_j, \hat{x}_j, x_j; u_{\pi(j)}) \propto p(x_j) P_j^*(u_{\pi(j)}|x_j) \left[c_j(u_j, \hat{x}_j, x_j) + C_j^*(u_j, x_j) \right] \tag{2.35}$$

where

$$P_j^*(u_{\pi(j)}|x_j) = \begin{cases} 1 & , \text{ if } \pi(j) \text{ empty} \\ \sum\limits_{x_{\pi(j)} \in \mathcal{X}_{\pi(j)}} p(x_{\pi(j)}|x_j) \prod\limits_{i \in \pi(j)} P_{i \to j}^*(u_{i \to j}|x_i) & , \text{ otherwise} \end{cases} \tag{2.36}$$

with terms regarding to influence of $i \in \pi(j)$ *on* j*, i.e.* $P_{i \to j}^*(u_{i \to j}|x_j)$

$$P_{i \to j}^*(u_{i \to j}|x_i) = \sum_{u_{\chi(i) \setminus j} \in \mathcal{U}_{\chi(i) \setminus j}} \sum_{u_{\pi(i)} \in \mathcal{U}_{\pi(i)}} P_i^*(u_{\pi(i)}|x_i) \sum_{\hat{x}_i \in \mathcal{X}_i} \int_{\mathcal{Y}_i} \mathrm{d}y_i \, p(u_i, \hat{x}_i|y_i, u_{\pi(i)}; \gamma_i^*) p(y_i|x_i) \tag{2.37}$$

and the $C_j^*(u_j, x_j)$ *term which is added to the local cost and given by*

$$C_j^*(u_j, x_j) = \begin{cases} 0 & , \text{ if } \chi(j) \text{ empty} \\ \sum_{k \in \chi(j)} C_{k \to j}^*(u_{j \to k}, x_j) & , \text{ otherwise} \end{cases} \tag{2.38}$$

with the terms regarding the influence of j *over* $k \in \chi(j)$*, i.e.* $C_{k \to j}^*(u_{j \to k}, x_j)$

$$C_{k \to j}^*(u_{j \to k}, x_j) = \sum_{x_k \in \mathcal{X}_k} \sum_{\hat{x}_k \in \mathcal{X}_k} \sum_{u_k \in \mathcal{U}_k} \left[c_k(u_k, \hat{x}_k, x_k) + C_k^*(u_k, x_k) \right] Q_{k \to j}^*(u_k, \hat{x}_k, x_k | u_{j \to k}, x_j) \tag{2.39}$$

with

$$Q^*_{k\to j}(u_k, \hat{x}_k, x_k|u_{j\to k}, x_j) = \sum_{x_{\pi(k)\setminus j}\in\mathcal{X}_{\pi(k)\setminus j}} p(x_{\pi(k)\setminus j}, x_k|x_j) R^*_{k\to j}(u_k, \hat{x}_k|x_k, x_{\pi(k)\setminus j})$$

$$R^*_{k\to j}(u_k, \hat{x}_k|x_k, x_{\pi(k)\setminus j}) = \sum_{u_{\pi(k)\setminus j}\in\mathcal{U}_{\pi(k)\setminus j}} \int_{\mathcal{Y}_k} \mathrm{d}y_k\, p(y_k|x_k)\, p(u_k, \hat{x}_k|y_k, u_{\pi(k)}; \gamma^*_k) \prod_{m\in\pi(k)\setminus j} P^*_{m\to k}(u_{m\to k}|x_m)$$

Proof. We will present an overview of the proof whereas the interested reader is referred to [28] for details. The proof starts with substituting Eq.(2.31) in Eq.(2.33) as well as mathematical statements of Assumption 3. Next, we recognize that under directed acyclic topology and Assumption 2, the set of incoming messages to node j, i.e. $U_{\pi(j)}$, does not depend on local rules of the nodes other than the acestors of node j. Moreover, under Assumption 3, the decision $(u_j, \hat{x}_j)$ local to node j does not affect the costs of nodes other than the descendants of j. Under Assumption 4, given γ_j, the information flow among ancestors of j is independent of the information flow among the descendats and leads to the form in Eq.(2.35). Moreover, since Assumption 4 implies that no parents of j share a common ancestor and no child of j share a common descendant, the likelihood terms, starting from the parentless ascendants takes the recursive form in Eq.(2.36) and (2.37) down to node j and cost terms influenced by the local decision of j, starting from the childless descendants takes the form in Eq.(2.38) and (2.39) up to node j. ■

Considering Eq.(2.36) and (2.37) we note that $P^*_{i\to j}(u_{i\to j}|x_i)$ is the likelihood of x_i based on the particular message $u_{i\to j}$ on the link from node i to j and under Assumption 4 $P^*_j(u_{\pi(j)}|x_j)$ is the likelihood of x_j for the particular incoming message vector $u_{\pi(j)}$, i.e. $p(u_{\pi(j)}|x_j; \gamma_{an(j)})$. A similar treatment of Eq.(2.38) and (2.39) reveals that $C^*_{k\to j}(u_{j\to k}, x_j)$ terms are the expected cost if the actual value of the random variable associated with node j takes the value x_j and node j sends the message $u_{j\to k}$ on the link to its child k. Hence, under a polytree topology $C^*_j(u_j, x_j)$ is the total expected cost induced on the descendats of j for transmitting u_j. This cost is added to the local cost $c_j(u_j, \hat{x}_j, x_j)$ in Eq.(2.35) which also penalizes the transmission cost. Also considering Eq.(2.32) and (2.35), and noting that under these assumptions $p(x_j)p(y_j|x_j)P(u_{\pi(j)}|x_j) \propto p(x_j|y_j, u_{\pi(j)})$, we conclude that given the measurement y_j and the incoming messages $u_{\pi(j)}$, node j chooses the output with the minimum expected cost where the cost terms are sum of that due to the rules local to j and its descendants and the underlying distribution is determined by the ascendants.

In the discussion above, person-by-person optimal local rules $\{\gamma^*_j\}$ are characterized by finite

dimensional vectors $\{\phi_j^*\}_{j\in\mathcal{V}}$. Also considering Algorithm 1, it is possible to start with an initial strategy $\gamma^0 = (\gamma_1^0, ..., \gamma_N^0)$ corresponding to $\{\phi_j^0\}_{j\in\{1,...,N\}}$ and then proceed in a coordinate descent fashion, i.e. $\phi_j^l = \psi_j(\phi_1^{l-1}, ..., \phi_{j-1}^{l-1}, \phi_{j+1}^l, ..., \phi_N^l)$ for $j = N, N-1, ..., 1$ and $l = 1, 2, ...$, and consequently obtain a non-increasing, convergent sequence $\left\{J(\gamma^l)\right\}$. Such an algorithm is achieved through interpreting the right hand side of Eq.s(2.35)-(2.39) as operators that can be evaluated at any $\gamma = (\gamma_1, ..., \gamma_N) \in \Gamma^{\mathcal{G}}$ not necessarily optimal. Consider d_j, f_j, g_j and h_j, which summarize this treatment and given by

$$\begin{aligned}
\phi_j &= d_j(P_j, C_{\chi(j)\to j}) \\
P_j &= f_j(P_{\pi(j)\to j}) \\
P_{j\to\chi(j)} &= g_j(\phi_j, P_j) \\
C_{j\to\pi(j)} &= h_j(\phi_j, P_{\pi(j)\to j}, C_{\chi(j)\to j})
\end{aligned}$$

where $P_{\pi(j)\to j} = \{P_{i\to j}\}_{i\in\pi(j)}$, $C_{\chi(j)\to j} = \{C_{k\to j}\}_{k\in\chi(j)}$ and $C_{j\to\pi(j)} = \{C_{j\to i}\}_{i\in\pi(j)}$. Given an arbitrary $\{\phi_j\}_{j\in\mathcal{V}}$, it is possible to calculate the incoming message likelihoods for every node through f_j and g_j starting with the parentless nodes (since they are never subject to receiving $P_{i\to j}$ terms for $i \in \pi(j)$) and proceeding in accordance with the forward partial ordering. The childless nodes can then update their local rules using d_j (since they are never subject to receive any $C_{k\to j}$ terms for $k \in \chi(j)$) which corresponds to holding all the local rules other than the childless nodes' fixed and update the corresponding coordinates. Consecutively, it becomes possible to compute the conditional expected cost terms for the parents by utilizing h_j and continue with other nodes in accordance with the backward partial ordering. This forward-backward scheme exactly substitutes the `Update` step in Algorithm 1 such that at the l^{th} step the following computations take place:

$$\begin{aligned}
P_j^l &:= f_j(P_{\pi(j)\to j}^l) \\
P_{j\to\chi(j)}^l &:= g_j(\phi_j^{l-1}, P_j^l)
\end{aligned}$$

for $j = 1, 2, ..., N$ and

$$\begin{aligned}
\phi_j^l &:= d_j(P_j^l, C_{\chi(j)\to j}^l) \\
C_{j\to\pi(j)}^l &:= h_j(\phi_j^l, P_{\pi(j)\to j}^l, C_{\chi(j)\to j}^l)
\end{aligned}$$

for $j = N, N-1, ..., 1$.

Performing the computations above for $l = 0, 1, ...$ a convergent sequence $\left\{J(\gamma^l)\right\}$ is obtained

where the expected cost is given by

$$J(\gamma^l) = \sum_{j \in \mathcal{V}} G_j(\gamma_j^l) \tag{2.40}$$

such that

$$G_j(\gamma_j^l) = \sum_{x_j \in \mathcal{X}_j} p(x_j) \sum_{u_j \in \mathcal{U}_j} \sum_{\hat{x}_j \in \mathcal{X}_j} c_j(u_j, \hat{x}_j, x_j) \sum_{u_{\pi j} \in \mathcal{U}_{\pi j}} P_j^{l+1}(u_{\pi(j)}|x_j) p(u_j, \hat{x}_j|x_j, u_{\pi(j)}; \phi_j^l) \tag{2.41}$$

The iterative scheme which is a specialized form of Algorithm 1 for optimization of decentralized detection networks under communication constraints in a Bayesian setting is given by Algorithm 2. It is possible to perform this algorithm in a message passing fashion treating

Algorithm 2 Iterations converging to a person-by-person optimal decentralized detection strategy over a directed acyclic graph $\mathcal{G}$.

0) `(Initiate)` $l = 0$;

Choose $\gamma^0 = (\gamma_1^0, \gamma_2^0, ..., \gamma_N^0)$ such that $\gamma_j^0 \in \Gamma_j^{\mathcal{G}}$ for $j = 1, 2, ..., N$

1) `(Update)` $l = l + 1$;

`i) (Forward pass)`

`for` $j = 1, 2, ..., N$

$P_j^l = f_j(\left\{P_{i \to j}^l(u_{i \to j}|x_i)\right\}_{i \in \pi(j)})$

$\left\{P_{j \to k}^l(u_{j \to k}|x_j)\right\}_{k \in \chi(j)} = g_j(\phi_j^{l-1}, P_j^l)$

`ii) (Backward pass)`

`for` $j = N, N-1, ..., 1$

$\phi_j^l = d_j(P_j^l, \left\{C_{k \to j}^l(u_{j \to k}, x_j)\right\}_{k \in \chi(j)})$

$\left\{C_{j \to i}^l(u_{i \to j}, x_i)\right\}_{i \in \pi(j)} = h_j(\phi_j^l, \left\{P_{i \to j}^l(u_{i \to j}|x_i)\right\}_{i \in \pi(j)}, \left\{C_{k \to j}^l(u_{j \to k}, x_j)\right\}_{k \in \chi(j)})$

2) `(Check) If` $J(\gamma^{l-1}) - J(\gamma^l) < \varepsilon$ `STOP, else GO TO 1;`

each node $j \in \mathcal{V}$ as an entity which can perform communications and computations. Each node $j \in \mathcal{V}$ starts only with the knowledge of $p(x_j, x_{\pi(j)})$ and $c(u_j, \hat{x}_j, x_j)$ and an initial local rule $\gamma_j^0 \in \Gamma_j^{\mathcal{G}}$ which determines $p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j^0)$. In the forward pass, each node recieves $P_{i \to j}$ from its parents $i \in \pi(j)$, compute $P_{j \to k}$ for its children $k \in \chi(j)$ and transmit. The message schedule proceeds in accordance with the forward partial ordering. In the backward pass, each node receives $C_{k \to j}$ from its children $k \in \chi(j)$ and computes $C_{j \to i}$ for its parents $i \in \pi(j)$ which involves updating the local rule. The messaging proceeds, this time, regarding the backward partial ordering.

In conclusion, owing to the information structure introduced under Assumptions 1-4, an efficient online processing strategy is achieved. In other words, the decentralized strategy scales well with the number of variables. In addition, the optimization of the local rules in a person-by-person sense admits a message passing algorithm which scales both with the number of variables and the number of platforms. The resulting iterative scheme given as Algorithm 2 is amenable for network self-organization and the communication cost of this procedure is reasonable for a network that would execute the resulting strategy for a certain amount of time after initialization [26].

2.7.2 Undirected Graph Constrained Networks

A possible extension to the class of decentralized strategies over directed acyclic graphs which utilize a uni-directional communication regime is obtained through allowing bi-directional transmissions. Such a structure can be considered as more preferable in matching the ad-hoc and multi-hop nature of the underlying network and has been introduced in the context of decentralized detection in [26]. It is possible to achieve an online processing scheme which yields a finite message schedule without any deadlock and a tracktable information structure through separating the communication and decision actions in two stages. In this section, we summarize two stage strategies over undirected networks and conditions under which an efficient optimization procedure that also admits a message passing interpretation is possible.

2.7.2.1 Online Processing Model and Problem Definition

Owing to the partial ordering property of directed acyclic graphs (DAG), the implied online processing scheme exhibits tracktability in a rather straightforward manner (Consider Eq.s(2.24) and (2.23) which complement the problem definition (P1) given by Expression (2.22)). On the other hand, the introduction of -loosely speaking- feedback into the processing scheme through bi-directional communications arises a number of concerns. First, after a finite step, termination should be guaranteed. Second, existence of possible deadlock[7] The approach utilized in [26] fullfills these requirements by considering a strategy in which each platform performs in two stages: In the first stage, each platform evaluates a communica-

[7] This term is used to refer to a situation in which some nodes wait to receive the messages from some neighbours before evaluating their local rules that in turn prevents the neigbors from evaluating their local rules and decide on the symbols admitting a chicken-and-egg problem.

tion rule which is, unlike the DAG case, based on only the observation and not the incoming messages. The second stage is performed successively, as soon as all expected messages are received, and involves evaluating the decision functions based on the observations, incoming messages and outcoming messages, in general.

Let us represent a set of platforms with the index set $\mathcal{V} = \{1, ..., N\}$. Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each edge $(i, j) \in \mathcal{E}$ corresponds to a communication link of finite capacity. Provided that $\mathcal{G}$ is an undirected graph $(i, j) \in \mathcal{E} \Leftrightarrow (j, i) \in \mathcal{E}$ holds enabling to model a bi-directional setting. Similar to the DAG case, associated with each edge is the set of admissible symbols $\mathcal{U}_{i\to j}$ for which, e.g. $\left|\mathcal{U}_{i\to j}\right| = d_{ij} + 1$ holds considering a capacity of $\log_2 d_{ij}$ bits such that $0 \in \mathcal{U}_{i\to j}$ indicates no transmission.

Each $j \in \mathcal{V}$ is associated with a random variable X_j that takes values from a finite set $\mathcal{X}_j$, i.e. $\left|\mathcal{X}_j\right| < \infty$. In the first stage, having observed $y_j \in \mathcal{Y}_j$, node j transmits a message $u_{j\to i}$ taking values from $\mathcal{U}_{j\to i}$ to each of its neighbors $i \in ne(j)$ constituting $u_j = \{u_{j\to i} | i \in ne(j)\}$. The set of all possible outgoing messages is given by $\mathcal{U}_j \underset{i\in ne(j)}{\otimes} \mathcal{U}_{j\to i}$. In the second stage, an inference on the value of X_j is drawn based on the observation y_j, the outgoing messages produced in the first stage, i.e. u_j, and the incoming messages from neighboring nodes given by $u_{ne(j)} = \{u_{i\to j} | i \in ne(j)\}$. The set of all possible incoming messages is given by $\mathcal{U}_{ne(j)} = \underset{i\in ne(j)}{\otimes} \mathcal{U}_{i\to j}$.

A causal online processing of measurements $y = (y_1, y_2, ..., y_N)$ where $\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_N$ takes place when each $j \in \mathcal{V}$, first performs its local communication rule $\mu_j : \mathcal{Y}_j \to \mathcal{U}_j$ acting on only y_j, and after $u_{ne(j)}$ is received, proceeds with the local decision rule $\nu_j : \mathcal{Y}_j \times \mathcal{U}_j \times \mathcal{U}_{ne(j)} \to \mathcal{X}_j$. Hence, the local rule of node j is a pair given by $\gamma_j = (\mu_j, \nu_j)$.

Similar to the discussion for the DAG case, the space of all stage one communication rules is given by $\mathcal{M}_j^{\mathcal{G}} = \{\mu_j | \mu_j : \mathcal{Y}_j \to \mathcal{U}_j\}$ and the stage two decision rule space is defined as $\mathcal{N}_j^{\mathcal{G}} = \{\nu_j | \nu_j : \mathcal{Y}_j \times \mathcal{U}_j \times \mathcal{U}_{ne(j)} \to \mathcal{X}\}$. The local rule spaces $\Gamma_j^{\mathcal{G}} = \mathcal{M}_j^{\mathcal{G}} \times \mathcal{N}_j^{\mathcal{G}}$ for $j \in \mathcal{V}$ construct the strategy space $\Gamma^{\mathcal{G}} = \underset{v\in V}{\otimes} \Gamma_v^{\mathcal{G}}$ and the problem definition (P1) given by Expression 2.22 holds.

The causal processing provided through the two stage scheme described above implies that the equation that couples the local rules for the directed case, i.e. Eq.(2.24), has the corre-

spondance of

$$p(u, \hat{x}|x, y; \gamma) = \prod_{j=1}^{N} p(u_j, \hat{x}_j|y_j, u_{ne(j)}; \gamma_j) \tag{2.42}$$

with

$$p(u_j, \hat{x}_j|y_j, u_{ne(j)}; \gamma_j) = p(u_j|y_j; \mu_j) p(\hat{x}_j|y_j, u_j, u_{ne(j)}; \nu_j) \tag{2.43}$$

in addition. Here, it holds that

$$p(u_j|y_j; \mu_j) = \delta_{u_j, \mu_j(y_j)} \tag{2.44}$$

$$p(\hat{x}_j|y_j, u_j, u_{ne(j)}; \delta_j) = \delta_{\hat{x}_j, \nu_j(y_j, u_j, u_{ne(j)})} \tag{2.45}$$

where δ is the Kronecker' s delta. It is possible to substitute the expressions above in the underlying distribution of the Bayesian risk $J(\gamma) = E\{c(u, x, \hat{x}); \gamma\}$ which is given by Eq.(2.23). There corresponds, then an objective value for every given two stage strategy $\gamma \in \Gamma^{\mathcal{G}}$ where $\mathcal{G}$ is an undirected graph. Therefore, under the two stage local rules over an undirected graph interpretation, the communication constrained problem (P1) given by Expression (2.22) is valid.

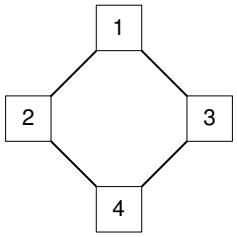
2.7.2.2 Optimizing Undirected Graph Constrained Networks

The communication constrained problem (P1) exhibits a different nature in the case of two-stage strategies over undirected graphs compared to that for strategies over a directed acyclic graph (DAG). As discussed in Section 2.7.1.3, in the case of a DAG, the conditional independence assumption alone enables the person-by-person optimal strategy to be in the form a finite set of likelihood-ratio tests (Proposition 2.7.1). For two-stage strategies over undirected graphs, an additional condition should hold:

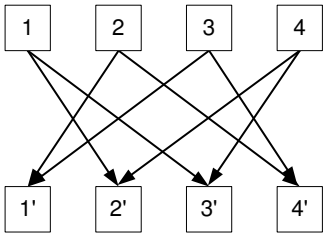
Assumption 5*(Separable Costs)***:** The global cost function is the sum of costs due to the stage-one communication rules and stage-two decision rules, which are in turn additive over the nodes, i.e.

$$c(u, \hat{x}, x) = \sum_{i=1}^{N} \left[c_i^d(\hat{x}_i, x) + \lambda c_i^c(u_i, x) \right] \tag{2.46}$$

It is shown that, if only Assumptions 1 and 5 hold, then a person-by-person optimal two-stage strategy over undirected graph in the form of a finite collection of likelihood-ratio tests is obtained which is similar to Proposition 2.7.1 for the DAG case(See Chp. 4 in [28] or



(a) Constraining undirected graph.



(b) Unwrapped DAG counterpart.

Figure 2.8: (a) A loopy undirected graph of 4 nodes given by $\mathcal{G} = (\mathcal{V}, \mathcal{X})$ where $\mathcal{V} = \{1, 2, 3, 4\}$ and $\mathcal{E} = \{(1, 2), (2, 1), (1, 3), (3, 1), (2, 4), (4, 2), (4, 3), (3, 4)\}$, (b) the DAG counterpart regarding the two-stage online processing. Nodes $1 - 4$ correspond to platforms $1 - 4$ but only performing the stage one communication rules, whereas nodes $1' - 4'$ correspond to platforms $1 - 4$ but only performing the stage two decision rules.

[26]). In addition, it is possible to inspect multi-stage schemes on undirected graphs through unwrapping them on directed graphs [8]. Consider, for example, the undirected graph and its unwrapped directed counterpart in Figure 2.8. Under Assumptions 1 and 5, nodes $1 - 4$ perform only the stage one rules, i.e. μ_js, and nodes $1' - 4'$ are associated only with the stage two rules, i.e. ν_js. Node j and j' correspond to the same physical platform but different processing tasks, in this respect. Hence, Proposition 2.7.1 for the DAG case can be applied to the corresponding unwrapped graphs. It is then possible to show that given a person-by-person optimal two-stage strategy γ^* over an undirected graph and fixing all local rules other than the j^{th} at the optimal, the j^{th} person-by-person optimal stage-one rules lie in the set $\{\mu_j | \mu_j : \mathcal{Y}_j \to \mathcal{U}_j\}$ and stage-two rules reside in $\{\nu_j | \nu_j : \mathcal{Y}_j \times \mathcal{U}_{ne(j)} \to \mathcal{X}_j\}$.

In order to reduce the complexity further and provide scalability for both the local rules in the number of random variables (i.e. dimensionality of $\mathcal{X}$) and for the iterative optimization scheme in the number of platforms, assumptions of the measurement and cost locality, i.e. Assumptions 2 and 3, should also hold. Under these conditions, given a person-by-person optimal strategy on the undirected graph, the j^{th} rule is found -similar to that given in Proposition 2.7.2- in a variational form and depending on the remaining:

Proposition 2.7.3 *Consider (P1) and let Assumptions 1-3 and 5 hold. Given an undirected graph $\mathcal{G}$ and considering a person-by-person optimal two-stage strategy γ^*, the j^{th} optimal*

[8] See for example [73] for the utilization of an approach along these lines for examining the Loopy Belief Propagation in which a computation tree is constructed such that the regular Belief Propagation over this graph corresponds to m-step of the loopy counterpart.

rule when the rest is fixed at the optimal, i.e. $\gamma_{\setminus j} = \gamma^*_{\setminus j}$*, reduces to*

$$\mu_j^*(Y_j) = \arg\min_{u_j \in \mathcal{U}_j} \sum_{x_j \in \mathcal{X}_j} \alpha_j^*(u_j, x_j) p(Y_j|x_j) \tag{2.47}$$

for the stage-one communication rule, where the rule coefficients $\alpha_j^* \in \mathbb{R}^{|\mathcal{U}_j \times \mathcal{X}_j|}$ *are given by*

$$\alpha_j^*(u_j, x_j) \propto p(x_j) \left[\lambda c_j^c(u_j, x_j) + \sum_{i \in ne(j)} C_{i \to j}^*(u_{j \to i}, x_j) \right] \tag{2.48}$$

and

$$\nu_j^*(Y_j, U_{ne(j)}) = \arg\min_{\hat{x}_j \in \mathcal{X}_j} \beta_j^*(\hat{x}_j, x_j; U_{ne(j)}) p(Y_j|x_j) \tag{2.49}$$

for the stage two decision rule where the rule coefficients $\beta_j \in \mathbb{R}^{|\mathcal{X}_j \times \mathcal{X}_j \times \mathcal{U}_{ne(j)}|}$

$$\beta_j^*(\hat{x}_j, x_j; u_{ne(j)}) \propto p(x_j) c_j^d(\hat{x}_j, x_j) P_j^*(u_{ne(j)}|x_j) \tag{2.50}$$

with the incoming message likelihood and the likelihood message from neighbor node $i \in ne(j)$*given by*

$$P_j^*(u_{ne(j)}|x_j) = \sum_{x_{ne(j)} \in \mathcal{X}_{ne(j)}} p(x_{ne(j)}|x_j) \prod_{i \in ne(j)} P_{i \to j}^*(u_{i \to j}|x_j) \tag{2.51}$$

and

$$\begin{aligned} P_{i \to j}^*(u_{i \to j}|x_j) &= \sum_{u_i \setminus u_{i \to j}} p(u_i|x_i; \mu_i^*) \\ &= \sum_{u_i \setminus u_{i \to j}} \int_{\mathcal{Y}_i} \mathrm{d}y_i p(y_i|x_i) p(u_i|y_i; \mu_i^*) \end{aligned} \tag{2.52}$$

respectively. The conditional expected cost term from each neighbor node $i \in ne(j)$ *is given by*

$$\begin{aligned} C_{i \to j}^*(u_{j \to i}, x_j) = \sum_{x_i \in \mathcal{X}_i} \sum_{\hat{x}_i \in \mathcal{X}_i} c_i^d(\hat{x}_i, x_i) \sum_{x_{ne(i)\setminus j} \in \mathcal{X}_{ne(i)\setminus j}} p(x_i, x_{ne(i)\setminus j}|x_j) \times \\ \sum_{u_{ne(i)\setminus j}} p(\hat{x}_i|x_i, u_{ne(i)}; \nu_i^*) \prod_{j' \in ne(i)\setminus j} P_{j' \to i}^*(u_{j' \to i}|x_{j'}) \end{aligned} \tag{2.53}$$

Proof. Note that under Assumptions 1-3 and 5, Eq.(2.43) specializes to

$$p(u_j, \hat{x}_j|y_j, u_{ne(j)}; \gamma_j) = p(u_j|y_j; \mu_j) p(\hat{x}_j|y_j, u_{ne(j)}; \nu_j)$$

which in turn yields the information structure of the strategy $\gamma_j \in \Gamma^{\mathcal{G}}$ as

$$p(u, \hat{x}|x, y; \gamma) = \prod_{j=1}^{N} p(u_j|y_j; \mu_j) \prod_{j=1}^{N} p(\hat{x}_j|y_j, u_{ne(j)}; \nu_j)$$

This structure admits the interpretation of being the information structure of a decentralized scheme over a DAG (corresponding to the unwrapped counterpart of the originial undirected graph) such that the rules local to parentless nodes are μ_js and the rules local to childless nodes are ν_js, i.e.

$$p(u, \hat{x}|x, y; \gamma) = \prod_{j\in\{1,...,N\}} p(u_j|y_j; \gamma_j) \prod_{j'\in\{1',...,N'\}} p(\hat{x}_{j'}|y_{j'}, u_{\pi(j')}; \gamma_{j'})$$

where $\gamma_j = \mu_j$ for $j \in \{1, ..., N\}$ and $\gamma_j = \nu_j$ for $j \in \{1', ..., N'\}$. Hence Proposition 2.7.2 applies to the unwrapped directed counterpart of the undirected graph with the set of nodes $\mathcal{V} = \{1, ..., N, 1', ..., N'\}$ and the corresponding set of local rules $\{\mu_1, ..., \mu_N, \nu_1, ..., \nu_N\}$ yielding the above expressions. For a detailed proof, see [28] under the assumption of error-free links.

■

Algorithm 3 Iterations converging to a person-by-person optimal decentralized two-stage strategy over an undirected graph $\mathcal{G}$.

---

0) (Initiate) $l = 0$;

Choose $\gamma^0 = (\gamma_1^0, \gamma_2^0, ..., \gamma_N^0)$ such that $\gamma_j^0 \in \Gamma_j^{\mathcal{G}}$ for $j = 1, 2, ..., N$

1) (Update) $l = l + 1$;

i) for $j = 1, 2, ..., N$

$P^l_{j\to ne(j)} = g_j(\alpha_j^{l-1})$

ii) for $j = 1, 2, ..., N$

$P^l_j = f_j(P^l_{ne(j)\to j})$

$\beta^l_j = q_j(P^l_j)$ % Update the stage-two rule

$C^l_{j\to ne(j)} = h_j(\beta_j, P^l_{ne(j)\to j})$

iii) for $j = 1, 2, ..., N$

$\alpha^l_j = r^l_j(C^l_{ne(j)\to j})$ % Update the stage-one rule

2) (Check) If $J(\gamma^{l-1}) - J(\gamma^l) < \varepsilon$ STOP, else GO TO 1;

Considering Proposition 2.7.3, it is possible to obtain an iterative scheme which starting with an initial two-stage strategy, converges to a person-by-person optimal one in a similar fashion discussed in Section 2.7.1.3 for the directed acyclic case. The treatment of the right hand sides of Eq.s (2.48), (2.50)-(2.53) as operators that can act on any set of their arguments, not necessarily corresponding to optimal local rules, is summarized by r_j and q_j together with

f_j, g_j and h_j given by

$$\begin{aligned} \alpha_j &= r_j(C_{ne(j)\to j}) \\ \beta_j &= q_j(P_j) \\ P_j &= f_j(P_{ne(j)\to j}) \\ P_{j\to ne(j)} &= g_j(\alpha_j) \\ C_{j\to ne(j)} &= h_j(\beta_j, P_{ne(j)\to j}) \end{aligned}$$

where $P_{ne(j)\to j} = \{P_{i\to j}\}_{i\in ne(j)}$, $C_{ne(j)\to j} = \{C_{i\to j}\}_{i\in ne(j)}$ and $C_{j\to ne(j)} = \{C_{j\to i}\}_{i\in ne(j)}$.

Considering the expressions above and following the steps similar to that presented in Section 2.7.1.3, the iterative scheme given in Algorithm 3 is obtained for offline optimization of two-stage strategies under communication constraints. In addition, the cost for any given strategy γ^l is easily found in terms of the expressions above as

$$J(\gamma^l) = \sum_{i\in\mathcal{V}} \sum_{x_i\in\mathcal{X}_i} p(x_i)\left[\lambda G_i^c(\mu_i^l|x_i) + G_i^d(\nu_i^l|x_i)\right]$$

where

$$G_i^c(\mu_i^l|x_i) = \sum_{u_i\in\mathcal{U}_i} c_i^c(u_i, x_i) p(u_i|x_i;\mu_i^l)$$

$$G_i^d(\nu_i^l|x_i) = \sum_{\hat{x}_i\in\mathcal{X}_i} c_i^d(\hat{x}_i, x_i) \sum_{u_{ne(i)}\in\mathcal{U}_{ne(i)}} p(\hat{x}_i|x_i, u_{ne(i)};\nu_i^l) \sum_{x_{ne(i)}\in\mathcal{X}_{ne(i)}} p(x_{ne(i)}|x_i) \prod_{j\in ne(i)} P_{j\to i}^{l+1}(u_j|x_j)$$

Note that, similar to that in the directed acyclic case, the `Update` step of Algorithm 3 also admits a message passing interpretation. In the first pass, all nodes compute and send forward likelihood terms to their neighbors. In the second pass, upon receiption of the likelihood messages, all nodes update their stage-two estimation rules and compute and send expected cost messages to their neighbors. After receiving cost messages from neighbors, each node update its stage-one communication rule. This structure of the optimization scheme renders it suitable for network self-organization.

CHAPTER 3

DECENTRALIZED ESTIMATION NETWORKS UNDER DIRECTED GRAPH CONSTRAINTS

In this chapter, we start presenting our investigation on decentralized estimation networks under communication constraints. Consider the motivations for the problem introduced in Section 1.1 including the possible applications in Section 1.1.3 and 1.1.4. A general treatment of decentralized statistical inference has been presented in Section 2.6, considering a system structure subject to resource constraints among which the emphasis is on the availability of communication links, their finite capacity and transmission costs (possibly due to energy dissipation). A number of approaches in the literature is addressed including the distributed function evaluation framework and message passing algorithms under communication constraints.

We are particularly interested in design perspectives and the tradeoff between the estimation accuracy and the communication cost. Similar challenges have been considered in the context of decentralized detection for which we present an overview in Section 2.7. In this chapter, we consider the class of decentralized strategies over directed acyclic graphs which is summarized in Section 2.7.1 for detection. In Section 3.1 we provide an overview of literature work on optimization of estimation networks. We adopt the communication constrained setting in Section 3.2 and present the Team Theoretic investigation in Section 3.3 which yields an iterative optimization scheme. In the estimation setting, unlike the detection problem, the random variables of concern take values from denumerable sets, which prevents us to utilize the iterative scheme in practice. We introduce a Monte Carlo framework in Section 3.4 and provide particle representations together with approximate computational schemes that yield an approximation to a person-by-optimal strategy. In Section 3.5 we utilize the frame-

work in examples and present the benefits including the smooth degredation of the estimation performance with the decrease in utilization of the communications.

3.1 Introduction

In this chapter, we consider the estimation of a random vector that takes values from an N-dimensional Euclidean space through a system with a communication and computation structure that exhibits collaborative processing and better matches the underlying communication topology. This scenario captures, e.g. the estimation of a common parameter (as in e.g.[87]) as well as samples of a field (as in e.g.[88]) with an ad-hoc sensor network.

Similar to the case in the detection setting, the literature on decentralized estimation often considers a structure in which a fusion center responsible for performing the estimation task exploits messages from the peripheral platforms in order to improve the accuracy (see for example [89][90] with [91] and the references therein). In this perspective, the limited bandwidth is addressed through having the peripherals perform a quantization of the observation, and the number of quanta is determined by the link capacity. The design problem involves selecting the quantizers for the peripherals in terms of informativeness together with a fusion rule that achieves a reasonable accuracy. Motivated by sensor networks, similar problem settings under different domain knowledge such as the noise distribution and quantization level constraints have been investigated (see e.g. [87] [92] [93] [94]) as well as the case in which samples of a field are subject to estimation (see e.g.[95] [96] [88] and the references therein).

Altough these treatments consider keeping the communication demand as low as possible, they are limited in capturing certain aspects of the problem. First of all, the communication structures for which results can be produced for are restricted to star topologies (See Section 2.7 for the discussion of such communication and computation structures in the context of detection). The cost of transmissions from peripherals to the fusion center which possibly vary considering the multi-hop nature is not explicitly accounted for. Often, a common random variable is of concern and the estimation is performed only at the fusion center. This restricts the collaboration among platforms for online processing of observations and opens up a possibility for a computational bottleneck in the case of multiple random variables.

Unlike the canonical inference approaches mentioned above, we employ a design perspective

in which the cost of communications and estimation errors are considered explicitly in a Bayesian setting as well as the constraints including the availability and capacity of links. Similar challenges are of concern in decentralized detection for which a general treatment has been presented in [28] and summarized in Section 2.7.1. In this setting, the available links between sensor platforms render a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes and edges correspond to platforms and uni-directional links between two platforms respectively. The inference task is distributed through associating random variables with sensor platforms. Each node evaluates its local rule, given the incoming messages and its own measurement, producing an inference on the associated random variable(s) and outgoing messages. The design problem involves finding the set of local rules, which is referred to as the strategy, that minimizes an expected cost which captures contributions of both cost of communications and detection errors in a Bayesian setting with the set of feasible strategies constrained by $\mathcal{G}$. Decentralized detection is NP-hard in general, nevertheless necessary (but not sufficient) optimality conditions yield nonlinear Gauss-Seidel iterations which converge to a person-by-person optimal strategy [97]. In [27], this treatment is utilized for a directed acyclic topology and an iterative solution together with conditions under which the iterations admit a message passing interpretation that is scalable with the number of nodes are established.

We generalize this framework to decentralized estimation, and address some of the limitations of the canonical distributed estimation algorithms mentioned above ([89, 87, 91]). However this approach leads to an iterative scheme that involves integral equations that have no closed form solutions in general. Such a problem arise in message passing algorithms over continous Markov random fields with general distributions and has been the motivation for algorithms relying on particle representations together with approximate computational schemes including Non-parametric Belief Propagation [68, 72] which has been successfully applied for articulated visual object tracking [69, 98].

In order not to compromise model accuracy, we develop an approximation framework using Monte Carlo integration methods. In the resulting network, the platforms perform computations which correspond to approximations to an approximately person-by-person optimal rule. We maintain the scalability of the solution both in the number of nodes and sample sizes and we can produce results for any set of distributions as long as samples can be generated from them. In this respect, we present an efficient Monte Carlo optimization algorithm for decentralized estimation networks subject to communication constraints in a Bayesian setting.

The algorithm can be carried out in a message passing fashion making it also suitable for network self-organization. Also, the approach is valid for any distribution families as soon as samples can be generated.

3.2 Online Processing Modelled With Directed Acyclic Graphs

We consider the class of decentralized strategies presented in Section 2.7.1.1 which comply with directed acyclic graphs (DAGs). The nodes represent the sensor platforms and each edge corresponds to a low capacity link on which error-free transmission of an element from the set of admissible symbols is possible. The operation of the platforms are constrained by the graph such that each node, based on its measurement and incoming messages from parents, produces a local estimate and outgoing messages to children.

A causal online processing of observations is achieved through executing the local rules of the platforms in accordance with the forward partial ordering regarding the reachability relation over the DAG. The information structure imposed by the strategy in the context of estimation is similar to that in the detection setting except that the set $\mathcal{X}$ from which the random variables of concern take values from is denumerable.

Considering the graph constrained problem setting (P1) given by Expression(2.22). The Bayesian risk capturing the penalty both for the estimation error and communication load u is defined as $c : \mathcal{U} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The objective function is similarly given by $J(\gamma) = E\{c(u, \hat{x}, x); \gamma\}$ where the underlying distribution is given by

$$p(u, x, \hat{x}; \gamma) = \int_{y \in \mathcal{Y}} \mathrm{d}y \; p(u, \hat{x}|x, y; \gamma) p(x, y) \tag{3.1}$$

with the conditional imposed by the directed acyclic structure

$$p(u, \hat{x}|x, y; \gamma) = \prod_{j=1}^{n} p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j) \tag{3.2}$$

Further details including the proof is presented in Section A.1. Unlike the detection case, given γ_j, it is convenient to treat $p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j)$ as a finite set of distributions parameterized on u_j, i.e.

$$p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j) = p_{u_j}(\hat{x}_j|y_j, u_{\pi(j)}; \gamma_j) \tag{3.3}$$

where

$$p_{[\gamma_j(y_j,u_{\pi(j)})]_{\mathcal{U}_j}}(\hat{x}_j|y_j,u_{\pi(j)};\gamma_j)=\delta(\hat{x}_j-\left[\gamma_j(y_j,u_{\pi(j)})\right]_{\mathcal{X}_j}) \tag{3.4}$$

and δ is the Dirac' s delta distribution. Hence, the local rule γ_j and the distribution family $p_{u_j}(\hat{x}_j|y_j,u_{\pi(j)};\gamma_j)$ specify each other accordingly.

It follows that given a decentralized estimation strategy constrained by the directed acyclic graph $\mathcal{G}$, i.e. $\gamma\in\Gamma^{\mathcal{G}}$, there corresponds an expected Bayesian risk $J(\gamma)=E\{c(u,x,\hat{x});\gamma\}$ that can be computed by substituting Eq.s(3.1)-(3.4) in the expectation for which unlike the detection case integrations over $\mathcal{X}$ appear. The communication constrained inference problem (P1) given by Expression (2.22) is still valid and admits the interpratation of the problem of decentralized estimation constrained by directed acyclic graphs.

This framework provides the benefits of extending the star-topology to directed acyclic graphs while taking into account the availability and capacity of links together with a penalty due to transmissions and estimation errors. However, it is not straightforward to choose a sensible cost function for every realization of the process $(U,X,\hat{X})$ unlike the detection problem in which both the sets $\mathcal{U}$ and $\mathcal{X}$ are finite. In the estimation setting, it is unlikely that an arbitrary selection of the function $c:\mathcal{U}\times\mathcal{X}\times\mathcal{X}\rightarrow\mathbb{R}$ would achieve consistency in penalyzing estimation errors and the communication cost. A convenient approach is to consider the additive form

$$c(u,x,\hat{x})=c^d(x,\hat{x})+\lambda c^c(u,x) \tag{3.5}$$

admitting the interpretation that c^d is the cost of estimation errors and c^c is the cost due to communication together with a unit conversion coefficient λ which admits the interpretation of the equivalent estimation error per unit communication cost. Note that for the case

$$J(\gamma)=J_d(\gamma)+\lambda J_c(\gamma) \tag{3.6}$$

also holds where $J_d(\gamma)$ and $J_c(\gamma)$ are the expected penalties due to estimation errors and communication respectively. When $\lambda=0$, communications has no cost and the only consideration is the constraint on the feasible strategy space which is based on the available links and the corresponding capacities together with $\mathcal{G}$. This multi-objective optimization setting yields that $\gamma^*=\arg\min_{\gamma\in\gamma^{\mathcal{G}}}J_d(\gamma)+\lambda J_c(\gamma)$ vary with λ and $\{(J_c(\gamma^*),J_d(\gamma^*))|\lambda\geq 0\}$ describes the achievable performance points which, in our case, is similar to a pareto-optimal front and enables smoothly trading off estimation accuracy for less transmissions.

One possible choice for the penalty of estimation error is $c^d(x, \hat{x}) = (x - \hat{x})^T(x - \hat{x})$ where T denotes the transpose. For $\lambda = 0$, the minimum mean squared error strategy $\gamma \in \Gamma^{\mathcal{G}}$ is of concern.

In addition, it is possible to represent a selective communication scheme in this framework through selecting the set of admissible symbols on a link $(i, j) \in \mathcal{E}$ with capacity d_{ij} as $\mathcal{U}_{i\to j} = \{0, 1, 2, ..., 2^{d_i j} - 1\}$ without loss of generality and select the communication cost c^c such as

$$c^c(u) = \sum_{j\in\mathcal{V}} \sum_{i\in\chi(j)} c^c(u_{j\to i}, x_j) \tag{3.7}$$

where $c^c(u_{j\to i}, x_j) = 0$ if $u_{j\to i} = 0$. In this case, the symbol 0 represents no transmission but still an informative signalling to the corresponding child node. If $c^c(u_{j\to i}, x_j) = 1$ for $u_{j\to i} \neq 0$ $\forall(i, j) \in \mathcal{E}$, then $J_c(\gamma)$ is the expected network load under strategy $\gamma \in \Gamma^{\mathcal{G}}$.

3.3 Team Theoretic Solution Under DAG Constraints

In principle, given any communication and computation strategy for decentralized inference, the Bayesian framework discussed in Section 3.2 is useful provided that the underlying distribution $p(u, x, \hat{x}; \gamma)$ together with the contribution of the strategy $p(u, \hat{x}|x, y; \gamma)$ are tractable. On the other hand, further considerations on scalability of γ in the number of variables and the optimization scheme in the number of both platforms and variables arise.

The online processing scheme we utilize yields an efficient detection strategy and offline optimization under certain conditions as presented in Section 2.7.1.3 (recall that if the conditional independence of observations does not hold, then the decentralized detection problem is NP-hard [97]). Considering (P1) in the context of detection, if the sensor noise processes are independent and equivalenty observations are conditionally independent (Assumption 1), a team theoretic investigation yields a tractable solution to (P1) based on Proposition 2.7.1. This scheme is iterative and corresponds to a set of non-linear fixed point equations that converge to a person-by-person optimal strategy. Provided that the measurement locality(Assumption 2) assumptions also hold, the decentralized detection strategy scales with the number of variables. Together with the cost locality (Assumption 3) and polytree topology (Assumption 4), it is possible to find a person-by-person optimal decentralized strategy such that the set of fixed point equations admit a message passing interpretation which also scales with the number of nodes.

The intricacies in the detection setting described above suggest to endorse to the team theoretic investigation for the estimation problem as well. Consider the graph constrained problem (P1) given in Expression 2.22 with X and $\hat{X}$ taking values from a denumerable set $\mathcal{X}$. Proposition 2.7.1 restated for the estimation problem is as follows:

Proposition 3.3.1 *Consider (P1) such that X and $\hat{X}$ take values from a denumerable set $\mathcal{X}$. Under Assumption 1, given a person-by-person optimal strategy $\gamma^* = (\gamma_1^*, \gamma_2^*, ..., \gamma_N^*)$, and fixing all the local rules other than the j^{th} at the optimum, i.e. $\gamma_{\setminus j} = \gamma_{\setminus j}^*$, the j^{th} person-by-person optimal rule is given by*

$$\gamma_j^*(Y_j, U_{\pi(j)}) = arg \min_{(u_j, \hat{x}_j) \in \mathcal{U}_j \times \mathcal{X}_j} \int_{\mathcal{X}} \mathrm{d}x p(Y_j|x) \theta_j^*(u_j, \hat{x}_j, x; U_{\pi(j)}) \tag{3.8}$$

where

$$\theta_j^*(u_j, \hat{x}_j, x, u_{\pi(j)}) = p(x) \sum_{u_{\setminus \{j\} \cup \pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u, \hat{x}, x) \prod_{i \neq j} \int_{\mathcal{Y}_i} \mathrm{d}y_i\, p(y_i|x) p(u_i, \hat{x}_i | y_i, u_{\pi(i)}; \gamma_i^*) \tag{3.9}$$

Proof. The proof is similar to that for Proposition 2.7.1. After substituting $\gamma_{\setminus j} = \gamma_{\setminus j}^*$, Eq.s(3.2), (3.1) and the mathematical statement of Assumption 1 in $J(\gamma) = J(\gamma_j, \gamma_{\setminus j})$ we obtain

$$\begin{aligned} J(\gamma_j, \gamma_{\setminus j}^*) &= \int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{X}} \mathrm{d}\hat{x} \sum_{u \in \mathcal{U}} c(u, x, \hat{x}) p(x) p(u_j, \hat{x}_j | x, u_{\pi(j)}; \gamma_j) \prod_{i \neq j}^{N} p(u_i, \hat{x}_i | x, u_{\pi(i)}; \gamma_i^*) \\ &= \int_{\mathcal{Y}_y} \mathrm{d}y_j \int_{\mathcal{X}_j} \mathrm{d}\hat{x}_j \sum_{u_j \in \mathcal{U}_j} \sum_{u_{\pi(j)} \in \mathcal{U}_{\pi(j)}} p(u_j, \hat{x}_j | y_j, u_{\pi(j)}; \gamma_j) \int_{\mathcal{X}} \mathrm{d}x p(y_j|x) p(x) \\ &\qquad \sum_{u_{\setminus \{j\} \cup \pi(j)}} \int_{\mathcal{X}_{\setminus j}} c(u, x, \hat{x}) \prod_{i \neq j}^{N} p(u_i, \hat{x}_i | x, u_{\pi(i)}; \gamma_i^*) \end{aligned} \tag{3.10}$$

We similarly consider a deterministic local rule γ_j. Given $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$ and $y_j \in \mathcal{Y}_j$ the solution γ_j^* minimizing Eq.(3.10) with probability 1 corresponds to selecting $p_{u_j}(\tilde{x}_j | y_j, u_{\pi(j)}; \gamma_j) = \delta(\hat{x}_j - \tilde{x}_j)$ for $(u_j, \hat{x}_j)$ such that its weight in Eq.(3.10), i.e.

$$\int_{\mathcal{X}} \mathrm{d}x p(y_j|x) p(x) \sum_{u_{\setminus \{j\} \cup \pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j} c(u, x, \hat{x}) \prod_{i \neq j} \int_{\mathcal{Y}_i} \mathrm{d}y_i p(u_i, \hat{x}_i | u_{\pi(i)}, y_i; \gamma_i^*) p(y_i|x) \tag{3.11}$$

is minimum. Hence

$$\gamma_j^*(Y_j, U_{\pi(j)}) = arg \min_{(u_j, \hat{x}_j) \in \mathcal{U}_j \times \mathcal{X}_j} \int_{\mathcal{X}} \mathrm{d}x p(Y_j|x) \theta_j^*(u_j, \hat{x}_j, x; U_{\pi(j)})$$

where θ_j^* is identified as

$$\theta_j^*(u_j, \hat{x}_j, x, u_{\pi(j)}) = p(x) \sum_{u_{\setminus j} \in \mathcal{U}_{\setminus j}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u, \hat{x}, x) \prod_{i \neq j} \int_{\mathcal{Y}_i} \mathrm{d}y_i\, p(y_i|x) p(u_i, \hat{x}_i|y_i, u_{\pi(i)}; \gamma_i^*) \tag{3.12}$$

■

We note that altough Equations (3.8) and (3.9) are similar to Equations (2.27) and (2.28) in that the observation likelihood $p(y_j|x_j)$ appears as sufficient statistics and θ_j^* admits the interpretation of being optimal parameters specified by the local rules of nodes other than j. Also considering Eq.(A.6), it is apparent that

$$\int_{\mathcal{X}} \mathrm{d}x p(Y_j|x) \theta_j^*(u_j, \hat{x}_j, x; U_{\pi(j)}) \propto E\{c(u_{\setminus j}, u_j, x, \hat{x}_{\setminus j}, \hat{x}_j)|Y_j, U_{\pi(j)}; \gamma_{\setminus j}^*\}$$

Consider Algorithm 1 which is a general iterative scheme that converges to a person-by-person optimal strategy (Section 2.7.1.2). It is useful to treat the right hand side of Eq.(3.9) as an operator ψ such that given any set of local rules for nodes other than the j^{th}, i.e. $\gamma_{\setminus j}$, fixed not necessarily at an optimum, ψ produces θ_j, i.e. $\theta_j = \psi_j(\gamma_{\setminus j})$. Together with Eq.(3.8) the local rule for the j^{th} node is obtained which equivalently minimizes the conditional expected cost. However, unlike the detection case in which θ_j being a finite dimensional vector provides a useful parameterization for the j^{th} person-by-person optimal rule, for the estimation case, since $\mathcal{X}$ is denumerable, the corresponding fixed point equations utilizing ψ in the `Update` step of Algorithm 1 are not practically solvable in general.

Optimality in a person-by-person sense has also been the canonical approach in the decentralized estimation literature. Proposition 3.3.1 applied for the conventional setting of decentralized estimation, yields the same person-by-person optimality conditions given in [90]. The system structure considered in this setting is similar to that for the conventional decentralized detection (Section 2.7). A strategy on a star-topology $\gamma_\star \in \Gamma^\star$ is considered in which node 0 is the fusion center and performs only estimation based on the symbols from finite sets transmitted by the peripherals $\mathcal{V} \setminus 0$, i.e. $\gamma_0^\star : \mathcal{U}_{1\to 0} \times \mathcal{U}_{2\to 0} \times ... \times \mathcal{U}_{N\to 0} \to \mathcal{X}$ and the peripherals perform quantization through evaluating $\gamma_j : \mathcal{Y}_j \to \mathcal{U}_{j\to 1}$ given the number of quanta $d_{j\to 0}$ for $j \in \mathcal{V}\setminus 0$ and $\left|\mathcal{U}_{j\to 0}\right| = d_{j\to 0}$. The cost due to the communications is not considered but rather MSE and MAP strategies are of concern. For $c(u, \hat{x}, x) = (\hat{x} - x)^2$ the person-by-person optimal MSE fusion rule that is performed by node 1 reduces to

$$\gamma_0(u_{1\to 0}, u_{2\to 0}, ..., u_{N\to 0}) = E\{x|u_{1\to 0}, u_{2\to 0}, ..., u_{N\to 0}\}$$

and in this sense of optimality, the quantization rule of the peripheral node j is given by

$$\gamma_j(y_j) = \arg\min_{u_{j\to 0}} E\{E\{x|y_j, y_{\setminus j}\} - \gamma_0(u_{1\to 0}, ..., u_{j-1\to 0}, u_{j\to 0}, u_{j+1\to 0}, ..., u_{N\to 0})|y_j\}$$

which are equivalently represented by Eqs.(3.8) and (3.9). Moreover for the particular case in which there are two nodes; one peripheral and one fusion node with no observation noise involved, the problem definition together with the necessary conditions of optimality reduces to that for the conventional MSE quantizer design problem presented in [99][100]. The MAP setting is similarly covered by Proposition 3.3.1 utilizing the complementary ϵ-neighborhood indicator cost introduced in Example 2.2.1.

However, the structure of the local rules given above do not yield closed form representations in general, altough relatively straightforward numerical computations are involved when the joint density $p(x, y_1, ..., y_N)$ is Gaussian and x is of dimension 1. The fact that the fusion rule is not scalable in the number of peripherals opens up a possibility for computational bottlenecks. This consideration has led to a fusion rule which is linear in the received symbols [89]. Further work has been motivated by certain aspects of the vision of sensor networks such as that the prior distribution for X and noise may not be known except the fact that they take values from bounded sets, variable threshold quantizers are not cheap, and it is more reasonable to assume a cheap device of a single imprecise threshold and consider a framework in which multiple level quantizers can be captured in terms of two-level quantizers (see e.g. [87, 92, 93]).

On the other hand, certain aspects such as the cost of communications, structures beyond the star topology and the tradeoff between the estimation accuracy and the transmission costs are not addressed by these approaches. In addition, the case in which the dimensionality of X is greater than one is endorsed to the fusion center alone which does not yield a scalable nor a collaborative strategy.

The information structure we obtain by utilizing a strategy over a directed acyclic graph extends the set of possible structures as well as providing means of introducing the cost of communication. Moreover, under certain conditions, scalability with the dimensionality of X and collaborative processing among nodes are achieved.

Under Assumptions 1 and 2, the decentralized strategies over DAGs exhibit scalability in the number of variables (and platforms). For the case, the local rules evaluate marginalizations only over the set from which the associated variable that takes values, i.e. $\mathcal{X}_j$, rather than $\mathcal{X}$.

Hence, Proposition 3.3.1 reduces to[1]

$$\gamma_j^*(Y_j, U_{\pi(j)}) = \arg \min_{u_j \times \hat{x}_j \in (\mathcal{U}_j \times \mathcal{X}_j)} \int_{\mathcal{X}_j} \mathrm{d}x_j p(Y_j|x_j) \phi^*(u_j, \hat{x}_j, x_j; U_{\pi(j)}) \tag{3.13}$$

where

$$\phi_j^*(u_j, \hat{x}_j, x_j; u_{\pi(j)}) = \int_{x_{\setminus j} \in \mathcal{X}_{\setminus j}} \mathrm{d}x_j \theta_j^*(u_j, \hat{x}_j, x; u_{\pi(j)}) \tag{3.14}$$

introducing scalability in the number of variables during online iterations, since given ϕ_j^*, the j^{th} local rule involves marginalization over only $\mathcal{X}_j$.

The scalability in the number of nodes while constructing ϕ_j^* requires additional assumptions to hold. We will restate Proposition 2.7.2 with expressions corresponding to Equations (2.36)-(2.39) in the estimation setting with the difference that we provide a more convenient form for the Monte Carlo approximations that will be presented later.

Proposition 3.3.2 *Consider (P1) such that X and $\hat{X}$ take values from a denumerable set $\mathcal{X}$. Under Assumptions 1-4, Eq.(3.13) applies with a proportional expression for $\phi_j^*(u_j, \hat{x}_j, x_j; u_{\pi(j)})$ given by*

$$\phi_j^*(u_j, \hat{x}_j, x_j; u_{\pi(j)}) \propto p(x_j) P_j^*(u_{\pi(j)}|x_j) \left[c_j(u_j, \hat{x}_j, x_j) + C_j^*(u_j, x_j) \right] \tag{3.15}$$

where

$$P_j^*(u_{\pi(j)}|x_j) = \begin{cases} 1 & , \text{if } \pi(j) \text{ empty} \\ \int_{\mathcal{X}_{\pi(j)}} dx_{\pi(j)} p(x_{\pi(j)}|x_j) \prod_{i \in \pi(j)} P_{i \to j}^*(u_{i \to j}|x_i) & , \text{otherwise} \end{cases} \tag{3.16}$$

with terms regarding influence of $i \in \pi(j)$ on j, i.e. $P_{i \to j}^*(u_{i \to j}|x_i)$

$$P_{i \to j}^*(u_{i \to j}|x_i) = \sum_{u_{\chi(i)\setminus j} \in \mathcal{U}_{\chi(i)\setminus j}} \sum_{u_{\pi(i)} \in \mathcal{U}_{\pi(i)}} P_i^*(u_{\pi(i)}|x_i) \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i \int_{\mathcal{Y}_i} \mathrm{d}y_i \, p(u_i, \hat{x}_i|y_i, u_{\pi(i)}; \gamma_i^*) p(y_i|x_i) \tag{3.17}$$

and the $C_j^*(u_j, x_j)$ term which is added to the local cost and given by

$$C_j^*(u_j, x_j) = \begin{cases} 0 & , \text{if } \chi(j) = \emptyset \\ \sum_{k \in \chi(j)} C_{k \to j}^*(u_{j \to k}, x_j) & , \text{otherwise} \end{cases} \tag{3.18}$$

with terms regarding the influence of $k \in \chi(j)$ on j, i.e. $C_{k \to j}^*(u_{j \to k}, x_j)$

$$C_{k \to j}^*(u_{j \to k}, x_j) = \int_{\mathcal{X}_{\pi(k)\setminus j}} dx_{\pi(k)\setminus j} \int_{\mathcal{X}_k} dx_k p(x_{\pi(k)\setminus j}, x_k|x_j) \sum_{u_{\pi(k)\setminus j} \in \mathcal{U}_{\pi(k)\setminus j}} \prod_{m \in \pi(k)\setminus j} P_{m \to k}^*(u_{m \to k}|x_m) \times$$

$$I_k^*(u_{\pi(k)}, x_k; \gamma_k^*) \tag{3.19}$$

[1] In a similar fashion with the reduction of Proposition 3.3.1 to Eq.s (2.32) and (2.33).

with

$$I_k^*(u_{\pi(k)}, x_k; \gamma_k^*) = \int_{\mathcal{Y}_k} \mathrm{d}y_k \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k \sum_{u_k \in \mathcal{U}_k} \left[c_k(u_k, \hat{x}_k, x_k) + C_k^*(u_k, x_k) \right] p(u_k, \hat{x}_k | y_k, u_{\pi(k)}; \gamma_k^*) p(y_k | x_k) \tag{3.20}$$

Proof. For an overview of the proof, consider that for Proposition 2.7.2 in Section 2.7.1.3 together with that X and $\hat{X}$ take values from a denumerable set $\mathcal{X}$. A detailed proof is provided in Section A.2 in the Appendix. ■

The proposition above being an adoptation of Proposition 2.7.2 for the estimation problem is subject to similar interpretations with thereunto including the treatment of Eq.s (3.16)-(3.20) as operators for any given strategy $\gamma_{\setminus j} \in \Gamma_{\setminus j}^{\mathcal{G}}$ not necessarily optimal. Similarly, it is possible to summarize this treatment by d_j, f_j, g_j and h_j such that

$$\begin{aligned} \phi_j &= d_j(P_j, C_{\chi(j)\to j}) \\ P_j &= f_j(P_{\pi(j)\to j}) \\ P_{j\to\chi(j)} &= g_j(\phi_j, P_j) \\ C_{j\to\pi(j)} &= h_j(\phi_j, P_{\pi(j)\to j}, C_{\chi(j)\to j}) \end{aligned}$$

where $P_{\pi(j)\to j} = \{P_{i\to j}\}_{i\in\pi(j)}$, $C_{\chi(j)\to j} = \{C_{k\to j}\}_{k\in\chi(j)}$ and $C_{j\to\pi(j)} = \{C_{j\to i}\}_{i\in\pi(j)}$. Note that d_j, f_j, g_j and h_j are specified by the rigth hand sides (RHS) of Eq.s(3.15) and (3.18), Eq.(3.16), Eq.(3.17) and finally Eq.s(3.19) and (3.20) respectively.

Therefore, the forward backward scheme given by Algorithm 2 also applies for (P1) in the estimation setting, in principle.

In Section 3.2 it is pointed out that it is unlikely for an arbitrary selection of the cost function $c : \mathcal{U} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to achieve consistency in penalyzing the estimation errors and communication costs. In addition additive structures render it possible to select an appropriate c which yields the estimation accuracy smoothly degrade as the cost paid for communications is decreased. Considering Proposition 2.7.2, we present the additivity of the cost as an assumption and investigate the reductions arise in Eq.s (3.16)-(3.20).

Assumption 5*(Separable Cost)***:** The global cost function $c(u, \hat{x}, x)$ is separable to functions penalyzing estimation errors and communications given by $c(u, \hat{x}, x) = c^d(\hat{x}, x) + \lambda c^c(u, x)$ where λ is a constant. Hence $J(\gamma) = J_d(\gamma) + \lambda J_c(\gamma)$ here $J_d(\gamma) = E\{c^d(\hat{x}, x); \gamma\}$ and $J_c(\gamma) = E\{c^c(u, x); \gamma\}$.

Assumption 5, together with Assumption 3 implies that the local cost functions are seperable, i.e.

$$c_j(u_j, x_j, \hat{x}_j) = c_j^d(x_j, \hat{x}_j) + \lambda c_j^c(u_j, x_j) \tag{3.21}$$

Corollary 3.3.3 *Consider Proposition 3.3.2, if the local costs are separable, i.e. Assumption 5 holds in addition to Assumptions 1-4, then the person-by-person optimal local rule in the variational form given by Eq.(3.13) is separated into two rules for estimation and communication as* $\gamma_j^* = (\nu_j^*, \mu_j^*)$ *given by*

$$\hat{x}_j = \nu_j^*(Y_j, U_{\pi(j)}) = \arg\min_{\hat{x}_j \in \mathcal{X}_j} \int_{x_j \in \mathcal{X}_j} dx_j p(x_j) p(Y_j|x_j) P_j^*(U_{\pi(j)}|x_j) c_j^d(\hat{x}_j, x_j) \tag{3.22}$$

$$u_j = \mu_j^*(Y_j, U_{\pi(j)}) \arg\min_{u_j \in U_j} \int_{x_j \in \mathcal{X}_j} dx_j p(x_j) p(Y_j|x_j) P_j^*(U_{\pi(j)}|x_j) \left[\lambda c_j^c(x_j, u_j) + C_j^*(u_j, x_j)\right] \tag{3.23}$$

Moreover, the corresponding distribution $p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j^*)$ *takes the form*

$$p(u_j \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j^*) = p(\hat{x}_j|y_j, u_{\pi(j)}; \nu_j^*) p(u_j|y_j, u_{\pi(j)}; \mu_j^*) \tag{3.24}$$

Proof. After substituting the separable local cost in Eq.(3.15) and Eq.(3.13), the optimization is separated into two problems over arguments $\hat{x}_j \in \mathcal{X}$ and $u_j \in \mathcal{U}_j$. This separation also implies that $U_j \perp\!\!\!\perp \hat{X}_j \,|\, (Y_j, U_{\pi(j)})$ yielding Eq.(3.24) by definition. ■

Example 3.3.4 *Consider a separable local cost where the contribution of the decision errors is given by* $c_j^d(\hat{x}_j, x_j) = (\hat{x}_j - x_j)^2$ *in reminiscent of the mean square estimator. We obtain a closed form expression for the estimation rule regarding he variational form in Eq.(3.22) after differentiating with respect to* $\hat{x}$ *and finding the zero solution as*

$$\hat{x}_j = \nu_j^*(Y_j, U_{\pi(j)}) = \frac{\int_{\mathcal{X}_j} \mathrm{d}x_j \, x_j p(x_j) p(Y_j|x_j) P_j^*(U_{\pi(j)}|x_j)}{\int_{\mathcal{X}_j} \mathrm{d}x_j \, p(x_j) p(Y_j|x_j) P_j^*(U_{\pi(j)}|x_j)} \tag{3.25}$$

Note that the information structure implies that $P_j^*(U_{\pi(j)}|x_j) = p(u_{\pi(j)}|x_j; \gamma_{\setminus j}^*)$ *holds which in turn is equal to* $p(u_{\pi(j)}|x_j; \gamma_{an(j)}^*)$ *due to the polytree topology. In addition the conditional independence relation* $U_{\pi(j)} \perp\!\!\!\perp Y_j \,|\, X_j$ *holds such that equivalently* $p(x_j, y_j, u_{\pi(j)}) = p(x_j) p(y_j|x_j) p(u_{\pi(j)}|x_j)$*. Hence the denominator in Eq.(3.25) is nothing but* $p(y_j, u_{\pi(j)}) = p(y_j, u_{\pi(j)}; \gamma_{an(j)}^*)$ *and the estimator is given by*

$$\hat{x}_j = \delta_j^*(Y_j, U_{\pi(j)}) = \int_{\mathcal{X}_j} \mathrm{d}x_j \, x_j p(x_j|Y_j, U_{\pi(j)}; \gamma_{an(j)}^*)$$

which is the center of gravity of the posterior density conditioned on both the observation and the incoming messages (this density is determined by the rules local to ancestors of j under Assumption 1-4, which are fixed at the optimum). Therefore, under this particular choice of the decision cost, any selection of the communication rules for ancestors manifest themselves in the optimal estimation rule for node j through the likelihood $P_j^*(U_{\pi(j)}|x_j)$. Moreover, this is also equivalent to treating $U_{\pi(j)}$ as another conditionally independent observation and utilizing the MSE estimator using the posterior.

If the local cost functions are separable, then with the Corollaries above, certain simplifications in Proposition 3.3.2 takes place.

Corollary 3.3.5 *Consider Proposition 3.3.2, if the local costs are separable, then $I(u_{\pi(k)}, x_k; \gamma_k^*)$ given by Eq.(3.20) takes the form*

$$I(u_{\pi(k)}, x_k; \gamma_k^*) = J_{d|x_k,u_{\pi(k)}} + J_{c|x_k,u_{\pi(k)}} \tag{3.26}$$

where $J_{d|x_k,u_{\pi(k)}}$ is the local expected decision cost conditioned on x_k and $u_{\pi(k)}$ given by

$$J_{d|x_k,u_{\pi(k)}} = \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k \, c_k^d(\hat{x}_k, x_k) p(\hat{x}_k|x_k, u_{\pi(k)}; \delta_k^*) \tag{3.27}$$

and $J_{c|x_k,u_{\pi(k)}}$ is the total expected cost of transmitting the symbol u_k conditioned on x_k and $u_{\pi(k)}$, including costs induced on the descendats, i.e. $C_k^*(u_k, x_k)$, and the transmission cost captured in $c_k^c(u_k, x_k)$;

$$J_{c|x_k,u_{\pi(k)}} = \sum_{u_k \in \mathcal{U}_k} \left(\lambda c_k^c(u_k, x_k) + C_k^*(u_k, x_k)\right) p(u_k|x_k, u_{\pi(k)}; \mu_k^*) \tag{3.28}$$

The distribution $p(\hat{x}_k|x_k|, u_{\pi(k)}; \delta_k^*)$ is the conditional probability density function of the estimations specified by δ_k^* given by

$$p(\hat{x}_k|x_k, u_{\pi(k)}; \delta_k^*) = \int_{\mathcal{Y}_k} \mathrm{d}y_k \, p(\hat{x}_k|y_k, u_{\pi(k)}; \delta_k^*) p(y_k|x_k) \tag{3.29}$$

and $p(u_k|x_k, u_{\pi(k)}; \mu_k^*)$ is the conditional probability mass function of the outgoing messages specified by μ_k^* given by

$$p(u_k|x_k, u_{\pi(k)}; \mu_k^*) = \int_{\mathcal{Y}_k} \mathrm{d}y_k \, p(u_k|y_k, u_{\pi(k)}; \mu_k^*) p(y_k|x_k) \tag{3.30}$$

Proof. After substituting the separable local cost for node k given by Eq.(3.21) in Eq.(3.20) and rearranging terms

$$I_k^*(u_{\pi(k)}, x_k; \gamma_k^*) = \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k c_k^d(\hat{x}_k, x_k) \int_{\mathcal{Y}_k} \mathrm{d}y_k p(\hat{x}_k|y_k, u_{\pi(k)}; \delta_k^*) p(y_k|x_k)$$
$$+ \lambda \sum_{u_k \in \mathcal{U}_k} \left[\lambda c_k^c(u_k, x_k) + C_k^*(u_k, x_k)\right] \int_{\mathcal{Y}_k} \mathrm{d}y_k p(u_k, \hat{x}_k|y_k, u_{\pi(k)}; \gamma_k^*) p(y_k|x_k) \quad (3.31)$$

is obtained. ■

Providing the sufficient conditions of optimality in a person-by-person sense Equations (3.15)-(3.20) under Assumtions 1-4 or Equations (3.15)-(3.19) together with Equations (3.26)-(3.30) under Assumptions 1-5 imply an iterative optimization scheme for decentralized estimation networks subject to communication constraints represented with a directed acyclic graph together with a Bayesian cost that penalizes both estimation errors and transmission over the links. In principle, once the operators implied by these expressions are utilized in Algorithm 2, it is possible to find a person-by-person optimal decentralized estimation strategy starting with an initial one.

Finally, the `Check` step of Algorithm 2 requires the objective value achieved by the strategy at the l^{th} step, i.e. $J(\gamma^l)$. This terms is obtained after replacing the summations in Eq.(2.41) with integrals where necessary and rearranging terms as

$$J(\gamma^l) = \sum_{j \in \mathcal{V}} G_j(\gamma_j^l) \quad (3.32)$$

where

$$G_j(\gamma_j^l) = \int_{\mathcal{X}_j} \mathrm{d}x_j p(x_j) \sum_{u_{\pi j} \in \mathcal{U}_{\pi j}} P_j^{l+1}(u_{\pi(j)}|x_j) \int_{\mathcal{Y}_j} \mathrm{d}y_j \int_{\mathcal{X}_j} \mathrm{d}\hat{x}_j \sum_{u_j \in \mathcal{U}_j} c_j(u_j, \hat{x}_j, x_j)$$
$$p(u_j, \hat{x}_j|y_j, u_{\pi(j)}; \gamma_j^l) p(y_j|x_j) \quad (3.33)$$

3.4 Monte Carlo Approximations for Directed Communication Constraints

In Section 3.3 we have provided conditions of optimality in a person-by-person sense rendering Algorithm 2 suitable also for estimation networks under directed network constraints. However, it is not possible to evaluate the right hand side of Eq.s (3.16)-(3.20) and correspondingly d_j, f_j, g_j and h_j exactly, in general, for arbitrary prior marginals $p(x_j)$, observation

likelihoods $p(y_j|x_j)$ and rules local to nodes other than j, i.e. $\gamma_{\setminus j}$. In this section, we propose particle based representations together with approximate computational schemes so that this conceptual iterative algorithm can be realized.

We exploit the Monte Carlo methods presented in Section 2.5 such that independent samples generated from only the marginal distributions of X and Y are required, i.e.

$$S_{x_j} \triangleq \{x_j^{(1)}, x_j^{(2)}, ..., x_j^{(M_j)}\} \text{ such that } x_j^{(m)} \sim p(x_j) \text{ for } m = 1, 2, ..., M_j \tag{3.34}$$

and

$$S_{y_j} \triangleq \{y_j^{(1)}, y_j^{(2)}, ..., y_j^{(P_j)}\} \text{ such that } y_j^{(p)} \sim p(y_j) \text{ for } p = 1, 2, ..., P_j \tag{3.35}$$

for $j \in \mathcal{V}$. Altough the sizes of the sample sets defined above might vary for each $j \in \mathcal{V}$, we assume that $M_j = M$ and $P_j = P$ for $j \in \mathcal{V}$ for brevity of discussion throughout.

This provides scalability in the number of variables and the number of samples together with ease of application for a number of reasons. First, considering a single random variable, it is a relatively straightforward task to generate pseudo random numbers from an arbitrary probability density function provided that the inverse of the corresponding cumulative distribution can be evaluated (see, e.g. Chp. 2 in [62]). In addition, the neccessary knowledge of distributions in order to utilize Algorithm 2, i.e. $p(x_{\pi(i)}, x_i)$ and $p(y_i|x_i)$ for all $i \in \mathcal{V}$, implies that the marginals are already known and hence we are not required the knowledge of any additional distributions. Besides, we consider independent generations that require no coordinations. For the case in which we consider scalability with the number of random variables involved, sampling from the joint distribution is cumbersome where scalability can be maintained up to a degree with coordinated generation schemes, which require the evaluation of characterizing densities such as the conditionals. For example Gibbs sampling introduced in [101] requires the so called full conditionals $\{p(x_j|x_{\setminus j})\}_{j\in\mathcal{V}}$ whereas the Substitution Sampling method requires $N(N-1)$ conditionals for N components [102].

We proceed by considering the sufficient condition of person-by-person optimality for the j^{th} rule given by Proposition 3.3.2. The Monte Carlo optimization algorithm we propose follows successive approximations to the expressions involved. In other words, we consider the j^{th} node point of view and perform progressive approximations to the local rule until all the components that involve arguments from denumerable sets are replaced with particle representations including the node-to-note terms $P^*_{i\to j}$ and $C^*_{k\to j}$ for $i \in \pi(j)$ and $k \in \chi(j)$ respectively.

3.4.1 Approximating the Person-by-Person Optimal Local Rule

Given a person-by-person optimal strategy $\gamma^* \in \Gamma^{\mathcal{G}}$, consider the j^{th} optimal local rule given by Eq.(3.15) and Eq.(3.13) in the case that the remaining are fixed at the optimum $\gamma_{\setminus j} = \gamma^*_{\setminus j}$. After substitution we obtain

$$\gamma^*_j(Y_j, U_{\pi(j)}) = \arg \min_{u_j \times \hat{x}_j \in (\mathcal{U}_j \times \mathcal{X}_j)} R^*_j(u_j, \hat{x}_j; Y_j, U_{\pi(j)}) \tag{3.36}$$

where

$$R^*_j(u_j, \hat{x}_j; y_j, u_{\pi(j)}) = \int_{\mathcal{X}_j} \mathrm{d}x_j p(x_j) p(y_j|x_j) P^*_j(u_{\pi(j)}|x_j) \left[c_j(u_j, \hat{x}_j, x_j) + C^*_j(u_j, x_j) \right] \tag{3.37}$$

for all $u_j \in \mathcal{U}_j$, $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$, $y_j \in \mathcal{Y}_j$ and $\hat{x}_j \in \mathcal{X}_j$ where unlike the detection case $\mathcal{X}_j$ is a denumerable set and the right hand side of Eq.(3.37) involves an integral over $\mathcal{X}_j$. It is reasonable to assume that the observation likelihood $p(y_j|x_j)$ and the cost $c_j(u_j, \hat{x}_j, x_j)$ are known. However, the incoming message likelihood, i.e. $P^*_j(u_{\pi(j)}|x_j)$, together with the conditional cost induced on the descendants, i.e. $C^*_j(u_j, x_j)$, depend on the remaining local rules $\gamma^*_{\setminus j}$ (see Section A.2) and do not necessarily admit closed form expressions for arbitrary $\gamma_{\setminus j} \in \Gamma^{\mathcal{G}}_j$.

Suppose that for all $x_j \in \mathcal{X}_j$, $P^*_j(u_{\pi(j)}|x_j)$ and $C^*_j(u_j, x_j)$ are known, i.e. it is possible to evaluate them for all elements of their domains. The integral on the right hand side of Eq.(3.37) still prevents R^*_j to be evaluated exactly, in general. However, an approximation is possible through the classical Monte Carlo method given by Eq.(2.18). Given M independent samples generated from $p(x_j)$, i.e. S_{x_j} given by Eq.(3.34),

$$\tilde{R}^*_j(u_j, \hat{x}_j; y_j, y_{\pi(j)}) = \frac{1}{\left|S_{x_j}\right|} \sum_{x_j \in S_{x_j}} p(y_j|x_j) P^*_j(u_{\pi(j)}|x_j) \left[c_j(u_j, \hat{x}_j, x_j) + C^*_j(u_j, x_j) \right] \tag{3.38}$$

where tilde denotes that it is an approximation, i.e. $\tilde{R}^*_j(u_j, \hat{x}_j; y_j, y_{\pi(j)}) \approx R^*_j(u_j, \hat{x}_j; y_j, y_{\pi(j)})$ over all its domain.

Since we have assumed that P^*_j and C^*_j are known, it is implied that they can be evaluated at $x_j \in S_{x_j}$, for all $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$ and $u_j \in \mathcal{U}_j$ respectively. $\tilde{R}^*_j$ substituted in Eq.(3.36) in place of R^*_j corresponds to a local rule, which is an approximation to γ^*_j. Let us represent the approximation to the optimal local rule by $\tilde{\gamma}^{*1}_j$ where the superscript 1 denotes that this is a 1-step approximation, then $\tilde{\gamma}^{*1}_j(y_j, u_{\pi(j)}) \approx \gamma^*_j(y_j, u_{\pi(j)})$ for all $y_j \in \mathcal{Y}_j$ and for all $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$ with nonzero probability.

Consider Corollary 3.3.3. The objective of minimization in the variational form of the j^{th} local rule given by Eq.(3.36) is seperable, i.e. $R^*_j(u_j, \hat{x}_j; y_j, u_{\pi(j)}) = R^*_{j,d}(\hat{x}_j; y_j, u_{\pi(j)}) + R^*_{j,c}(u_j; y_j, u_{\pi(j)})$, under a separable cost function local to node j and yields two separate problems and corresponding rules for estimation and communication denoted by ν_j and μ_j respectively. Similarly the approximation $\tilde{R}^*_j$ given by Eq.(3.38) splits trivially underlying two separate approximations, i.e. $\tilde{\nu}_j{}^1$ and $\tilde{\mu}_j{}^1$.

Example 3.4.1 *Consider Example 3.3.4. The explicit solution of the variational form for the quadratic estimation error cost is given by Eq.(3.25). The conventional Monte Carlo approximation given by Eq.(3.38) substituted in Eq.(3.36) specializes to*

$$\hat{x}_j = \tilde{\nu}_j{}^1(y_j, u_{\pi(j)}) = \frac{\sum_{m=1}^{M} x_j^{(m)} p(y_j|x_j^{(m)}) P^*_j(u_{\pi(j)}|x_j^{(m)})}{\sum_{m=1}^{M} p(y_j|x_j^{(m)}) P^*_j(u_{\pi(j)}|x_j^{(m)})} \tag{3.39}$$

3.4.2 Approximating the Message Likelihood and the Conditional Expected Cost

In the previous section, we proposed an approximation to the j^{th} optimal rule under the assumption that the incoming message likelihood and the conditional expected cost $P^*_j(u_{\pi(j)}|x_j)$ and $C^*_j(u_j, x_j)$ are known at $x_j = x_j^{(m)}$ for $m = 1, 2, ..., M$, for all $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$ and for all $u_j \in \mathcal{U}_j$ respectively. Since it is not possible to express these functions in closed form for an arbitrary set of local rules $\gamma_j \in \Gamma^{\mathcal{G}}_j$, in this step, we approximate them considering Eq.(3.16) and Eq.(3.18) .

We start the discussion by considering Eq.(3.16) for the case in which $\pi(j) \neq \emptyset$. Suppose that the forward node-to-node terms, i.e. $P^*_{i\to j}(u_{i\to j}|x_i^{(m)})$ for $i \in \pi(j)$ are known such that we can evaluate them at $x_i = x_i^{(m)}$ where $x_i^{(m)} \sim p(x_i)$ for $m = 1, 2, ..., M$ and for all $u_{i\to j} \in \mathcal{U}_{i\to j}$. This assumption is justified by the fact that if the 1-step approximation described in Section 3.4.1 were to be applied to the rules local to nodes $i \in \pi(j)$, then S_{x_i} would be utilized.

We next, note that it is possible to treat the concatenation of the elements of the parent sample sets as a sample set that is drawn by the product of distributions associated with the former sets. In other words, consider $x^{(m)}_{\pi(j)} \triangleq (x_i^{(m)})_{i\in\pi(j)}$ for $m = 1, 2, ..., M$ where $x_i^{(m)} \in S_{x_i}$ for $i \in \pi(j)$. These elements constitute a sample set $S_{\pi(j)} \triangleq \{x^{(m)}_{\pi(j)} | x^{(m)}_{\pi(j)} = (x_i^{(m)})_{i\in\pi(j)}\}$ and it holds that $x^{(m)}_{\pi(j)}\, x^{(m)}_{\pi(j)} \sim \prod_{i\in\pi(j)} p(x_i)$.

This observation enables the Importance Sampling approximation given in Eq.(2.20) for P_j^* through the importance sampling distribution $\prod_{i\in\pi(j)} p(x_i)$. Then the importance weights are given by

$$\omega_j^{(m)(m')} = p(x_{\pi(j)}^{(m')}|x_j^{(m)}) / \prod_{i\in\pi(j)} p(x_i^{(m')})$$

with the corresponding approximation

$$\tilde{P}_j^{\,*}(u_{\pi(j)}|x_j^{(m)}) = \frac{1}{\sum_{m'=1}^{M} \omega_j^{(m)(m')}} \sum_{m'=1}^{M} \omega_j^{(m)(m')} \prod_{i\in\pi(j)} P_{i\to j}^*(u_{i\to j}|x_i^{(m')}) \tag{3.40}$$

for $m = 1, 2, ..., M$ and for all $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$.

Similarly, considering Eq.(3.18) for the case in which $\chi(j) \neq \emptyset$ we assume that the node-to-node backward cost terms, i.e. for all $k \in \chi(j)$, $C_{k\to j}^*(u_{j\to k}, x_j)$, are known at $x_j = x_j^{(m)}$ for $m = 1, 2, ..., M$ and for all $u_{j\to k} \in \mathcal{U}_{j\to k}$. Then the required values, i.e. $C_j^*(u_j, x_j^{(m)})$ for $m = 1, 2, ..., M$ and for all $u_j \in \mathcal{U}_j$ can be computed exactly using Eq.(3.18).

From node j' s point of view, given node-to-node terms $P_{i\to j}^*$ and $C_{k\to j}^*$ evaluated at points generated from the appropriate marginal distributions, a further approximation to the j^{th} person-by-person optimal rule is obtained by computing $\tilde{P}_j^*$ and C_j^* at values of their argument required in Eq.(3.38) and substituting $\tilde{P}_j^*$ in place of P_j^*. Let $\tilde{\gamma}_j^{*2}$ denote the corresponding rule, then $\tilde{\gamma}_j^{*2}(y_j, u_{\pi(j)}) \approx \tilde{\gamma}_j^{*1}(y_j, u_{\pi(j)}) \approx \gamma_j^*(y_j, u_{\pi(j)})$ for all $y_j \in \mathcal{Y}_j$ and for all $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$ with nonzero probability.

3.4.3 Approximating the Node-to-Node Terms

In the previous section, the 2-steps approximation to the j^{th} local rule is introduced provided that for all $i \in \pi(j)$, $P_{i\to j}^*(u_{i\to j}|x_i)$ is known for all $u_{i\to j} \in \mathcal{U}_{i\to j}$ and $x_i = x_i^{(m)}$ for $m = 1, 2, ..., M$. Another requirement is the evaluation of $C_{k\to j}^*(u_{j\to k}, x_j)$ for all $u_{j\to k} \in \mathcal{U}_{j\to k}$ and $x_j = x_j^{(m)}$ where $x_j^{(m)} \in S_{x_j}$. A further step involves approximating the node-to-node terms $P_{i\to j}^*$ and $C_{k\to j}^*$ evaluated at the discretization of their domains provided by the sample sets.

We consider the parent nodes $i \in \pi(j)$ and consider evaluation of Eq.(3.17) at the required values of its arguments. Suppose that γ_i^* is fixed at the optimum, implying also that $p(u_i, \hat{x}_i|y_i, u_{\pi(i)}; \gamma_i^*)$ is specified through Eq.s(3.3) and (3.4) for all $i \in \pi(i)$. The multiple

integral term in Eq.(3.17), rewritten as

$$p(u_i|x_i, u_{\pi(i)}; \gamma_i^*) = \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i \int_{\mathcal{Y}_i} \mathrm{d}y_i \, p(u_i, \hat{x}_i|y_i, u_{\pi(i)}; \gamma_i^*) p(y_i|x_i)$$

for convenience, should be evaluated at $x_i = x_i^{(m)}$ for $m = 1, 2, ..., M$, for all $u_i \in \mathcal{U}_i$ and for all $u_{\pi(i)} \in \mathcal{U}_{\pi(i)}$. Since there is no closed form solution for arbitrary γ_i^* and the observation likelihood $p(y_i|x_i)$, we perform an Importance Sampling approximation through the importance sampling distribution $p(y_i)$. Utilizing $y_i^{(p)} \in S_{y_i}$ and the importance weights given by

$$\omega_i^{(m)(p)} = p(y_i^{(p)}|x_i^m) / p(y_i^{(p)})$$

an importance sampling approximation to $p(u_i|x_i^{(m)}, u_{\pi(i)}; \gamma_i^*)$ for $m = 1, 2, ..., M$, for all $u_i \in \mathcal{U}_i$ and for all $u_{\pi(i)} \in \mathcal{U}_{\pi(i)}$ is given by

$$\tilde{p}(u_i|x_i^{(m)}, u_{\pi(i)}; \gamma_i^*) = \frac{1}{\sum_{p=1}^{P} \omega_i^{(m)(p)}} \sum_{p=1}^{P} \omega_i^{(m)(p)} \delta_{u_i, [\gamma_i^*(y_i^{(p)}, u_{\pi(i)})] u_i} \tag{3.41}$$

where δ denotes the Kronecker's delta. Note that, if Assumption 5 holds, the discussion applies with the equality $p(u_i|x_i, u_{\pi(i)}; \gamma_i^*) = p(u_i|x_i, u_{\pi(i)}; \mu_i^*)$

Regarding Eq.(3.17), having approximated the multiple integral term, we similarly assume that $P_i^*(u_{\pi(i)}|x_i^{(m)})$ is known for $x_i = x_i^{(m)}$ where $x_i^{(m)} \sim p(x_i)$ for $m = 1, 2, ..., M$ and for all $u_{\pi(i)} \in \mathcal{U}_{\pi(i)}$. Note that the approximation introduced in the previous section acounts for this function evaluated at such values of its arguments. Together with Eq.(3.41) we obtain

$$\tilde{P}_{i\to j}^*(u_{i\to j}|x_i^{(m)}) = \sum_{u_{\chi(i)\setminus j} \in U_{\chi(i)\setminus j}} \sum_{u_{\pi(i)} \in u_{\pi(i)}} P_i^*(u_{\pi(i)}|x_i^{(m)}) \tilde{p}(u_i|u_{\pi(i)}, x_i^{(m)}; \gamma_i^*) \tag{3.42}$$

for $m = 1, 2, ..., M$ and for all $u_{i\to j} \in \mathcal{U}_{i\to j}$. It is possible to replace the node-to-node terms assumed to be known in Eq.(3.40) with Eq.(3.42) corresponding to a further step in the progressive approximations to γ_j^*.

The remaining term for approximation is the conditional expected cost induced on descendats of j on the branch initiating with the child k, i.e. $C_{k\to j}^*(u_{j\to k}, x_j)$, for all $k \in \chi(j)$, evaluated at $x_j = x_j^{(m)}$ where $x_j^{(m)} \in S_{x_j}$ and for all $u_{j\to k} \in \mathcal{U}_{j\to k}$. A similar reasoning leads to approximating the required values through utilizing Monte Carlo methods on the right hand side of the expression obtained by subsituting Eq.(3.20) in Eq.(3.19).

Consider (3.20) and suppose that γ_k^* is known also implying that $p(u_k, \hat{x}_k|y_k, u_{\pi(k)}; \gamma_k^*)$ is deter-

mined for all $k \in \chi(j)$. Substituting Eq.(3.3) and (3.4) in Eq.(3.20) yields

$$I^*(u_{\pi(k)}, x_k; \gamma_k^*) = \int_{\mathcal{Y}_k} \mathrm{d}y_k \, [\ c_k([\gamma_k^*(y_k, u_{\pi(k)})]_{\mathcal{U}_k}, [\gamma_k^*(y_k, u_{\pi(k)})]_{\mathcal{X}_k}, x_k) + C_k^*([\gamma_k^*(y_k, u_{\pi(k)})]_{\mathcal{U}_k}, x_k)\ ] p(y_k|x_k) \quad (3.43)$$

evaluation of which can be approximated at $x_k = x_k^{(m)}$ for all $x_k^{(m)} \in S_{x_k}$ and for all $u_{\pi(k)} \in \mathcal{U}_{\pi(k)}$ by the Importance Sampling method, using the importance density $p(y_k)$. Assuming $C_k^*(u_k, x_k)$ is known at $x_k = x_k^{(m)}$ where $x_k^{(m)}$ is an element of the usual sample set local to k, i.e. S_{x_k}, and for all $u_k \in \mathcal{U}_k$ and utilizing $y_k^{(p)} \in S_{y_k}$ together with the importance weights

$$\omega_k^{(m)(p)} = p(y_k^{(p)}|x_k^{(m)}) / p(y_k^{(p)})$$

we obtain

$$\tilde{I}^*(u_{\pi(k)}, x_k^{(m)}; \gamma_k^*) = \frac{1}{\sum_{p=1}^{P} \omega_k^{(m)(p)}} \sum_{p=1}^{P} \omega_k^{(m)(p)} [\ c_k([\gamma_k^*(y_k^{(p)}, u_{\pi(k)})]_{\mathcal{U}_k}, [\gamma_k^*(y_k^{(p)}, u_{\pi(k)})]_{\mathcal{X}_k}, x_k) + C_k^*([\gamma_k^*(y_k^{(p)}, u_{\pi(k)})]_{\mathcal{U}_k}, x_k)\ ] \quad (3.44)$$

for $m = 1, 2, ..., M$ and for all $u_{\pi(k)} \in \mathcal{U}_{\pi(k)}$ such that $\tilde{I}^*(u_{\pi(k)}, x_k^{(m)}; \gamma_k^*) \approx I^*(u_{\pi(k)}, x_k^{(m)}; \gamma_k^*)$ holds. Note that $[\gamma_k^*(y_k^{(p)}, u_{\pi(k)})]_{\mathcal{U}_k}$ and $[\gamma_k^*(y_k^{(p)}, u_{\pi(k)})]_{\mathcal{X}_k}$ are simply the communication symbol and estimation output of γ_k^* evaluated at the tuple $(y_k^{(p)}, u_{\pi(k)})$.

In addition, if Assumption 5 holds, we consider Corollary 3.3.5 and Importance Sampling approximations to Eq.(3.27) and Eq.(3.28) evaluated at $x_k^{(m)}$ and $u_{\pi(k)}$ are similarly obtained as

$$\tilde{J}_{d|x_k^{(m)}, u_{\pi(k)}} = \frac{1}{\sum_{p=1}^{P} \omega_k^{(m)(p)}} \sum_{p=1}^{P} \omega_k^{(m)(p)} c_k^d(\delta_k^*(y_k^{(p)}, u_{\pi(k)}), x_k^{(m)}) \quad (3.45)$$

and

$$\tilde{J}_{c|x_k^{(m)}, u_{\pi(k)}} = \sum_{u_k} \left(\lambda c_k^c(u_k, x_k^{(m)}) + C_k^*(u_k, x_k^{(m)}) \right) \tilde{p}(u_k|x_k^{(m)}, u_{\pi(k)}; \mu_k^*) \quad (3.46)$$

yielding $\tilde{I}^*(u_{\pi(k)}, x_k^{(m)}; \gamma_k^*) = \tilde{J}_{d|x_k^{(m)}, u_{\pi(k)}} + \tilde{J}_{c|x_k^{(m)}, u_{\pi(k)}}$.

Eq.(3.19) requires message likelihood terms from all parents of node k except node j and it is reasonable to assume that for all $j' \in \pi(k) \setminus j$, $P_{j' \to k}^*(u_{j' \to k}|x_j)$ is known at $x_j = x_{j'}^{(m)}$ for $m = 1, 2, ..., M$ where $x_{j'}^{(m)}$ is an element from the usual sample set local to node j' and for all $u_{j' \to k} \in \mathcal{U}_{j' \to k}$. Similarly, we observe that the set which is constituted of elements that are concetanation of elements from the usual sample sets local to $j' \in \pi(k) \setminus j$ is distributed from the product of the corresponding marginals. In other words, let us define

$x^{(m)}_{\pi(k)\setminus j} \triangleq (x^{(m)}_{j'})_{j'\in\pi(k)\setminus j}$. Then it holds that $x^{(m)}_{\pi(k)\setminus j} \sim \prod_{j'\in\pi(k)\setminus j} p(x_{j'})$ and an importance sampling approximation to Eq.(3.19) is possible through the importance distribution $\prod_{j'\in\pi(k)\setminus j} p(x_{j'})$. Having computed $\tilde{I}^*(u_{\pi(k)}, x^{(m)}_k; \gamma^*_k)$ and utilizing the usual sample sets local to nodes $j' \in \pi(k)\setminus j$ together with the importance sampling weigths

$$\omega^{(m)(m')} = p(x^{(m')}_{\pi(k)\setminus j}, x^{(m')}_k | x^{(m)}_j) / p(x^{(m')}_k) \prod_{j'\in\pi(k)\setminus j} p(x^{(m')}_{j'})$$

we obtain

$$\tilde{C}^*_{k\to j}(u_{j\to k}, x^{(m)}_j) = \frac{1}{\sum_{m'=1}^{M} \omega^{(m)(m')}} \sum_{m'=1}^{M} \omega^{(m)(m')} \sum_{u_{\pi(k)\setminus j}} \prod_{j'\in\pi(k)\setminus j} P^*_{j'\to k}(u_{j'\to k} | x^{(m')}_{j'}) \tilde{I}^*(u_{\pi(k)}, x^{(m')}_k; \gamma^*_k) \quad (3.47)$$

for $m = 1, 2, ..., M$ and for all $u_{j\to k} \in \mathcal{U}_{j\to k}$ which[2], after substituting in place of $C^*_{k\to j}$ in the right hand side of Eq.(3.18) for $\chi(j) \neq \emptyset$ yields $\tilde{C}^*_j$, i.e.

$$\tilde{C}^*_j(u_j, x^{(m)}_j) = \sum_{k\in\chi(j)} \tilde{C}^*_{k\to j}(u_{j\to k}, x^{(m)}_j) \quad (3.48)$$

for $m = 1, 2, ..., M$ and for all $u_j \in \mathcal{U}_j$.

Similarly, after substituting $\tilde{P}^*_{i\to j}$ in place of $P^*_{i\to j}$ in the right hand side of Eq.(3.40), we obtain a further approximation to P^*_j,

$$\tilde{P}_j{}^*(u_{\pi(j)} | x^{(m)}_j) = \frac{1}{\sum_{m'=1}^{M} \omega^{(m)(m')}_j} \sum_{m'=1}^{M} \omega^{(m)(m')}_j \prod_{i\in\pi(j)} \tilde{P}^*_{i\to j}(u_{i\to j} | x^{(m')}_i) \quad (3.49)$$

for $m = 1, 2, ..., M$ and for all $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$, which substituted in $\tilde{R}^*_j$ together with $\tilde{C}^*_{k\to j}$ given by Eq.(3.48) yields $\tilde{\gamma}^{*3}_j(y_j, u_{\pi(j)}) \approx \tilde{\gamma}^{*2}_j(y_j, u_{\pi(j)}) \approx \tilde{\gamma}^{*1}_j(y_j, u_{\pi(j)}) \approx \gamma^*_j(y_j, u_{\pi(j)})$ for all $y_j \in \mathcal{Y}_j$ and for all $u_{\pi(j)} \in \mathcal{U}_{\pi(j)}$ with nonzero probability.

3.4.4 Monte Carlo Optimization of Decentralized Estimation Networks Constrained by Directed Graphs

In Section 3.4.1-3.4.3 we have introduced a Monte Carlo approximation framework regarding the sufficient conditions of person-by-person optimality given in Proposition 3.3.2. Considering a person-by-person optimal decentralized estimation strategy constrained by a polytree $\mathcal{G}$,

[2] Note that we have approximated the forward likelihood terms regarding node j and its parents, i.e. $P^*_{i\to j}$ for $i \in \pi(j)$. However, we still assume that node-to-node terms regarding other nodes including $P^*_{j'\to k}$ for $j' \in \pi(k)\setminus j$ where $k \in \chi(j)$ are known over all their domains.

i.e. $\gamma^* \in \Gamma^{\mathcal{G}}$ and having $\gamma_{\setminus j}$ fixed at the corresponding set of optimal local rules, i.e. $\gamma_{\setminus j} = \gamma^*_{\setminus j}$, we have constructed a rule local for j, $\tilde{\gamma}^{*3}_j(y_j, u_{\pi(j)})$ such that it is an approximation to the optimal rule γ^*_j given by Eq.(3.13) following the progression

$$\tilde{\gamma}^{*1}_j(y_j, u_{\pi(j)}) = \arg \min_{u_j \times \hat{x}_j \in (\mathcal{U}_j \times \mathcal{X}_j)} \sum_{x_j \in S_{x_j}} p(y_j|x_j) P^*_j(u_{\pi(j)}|x_j) \left[c_j(u_j, \hat{x}_j, x_j) + C^*_j(u_j, x_j) \right]$$

where S_{x_j} is given by Eq.(3.34),

$$\tilde{\gamma}^{*2}_j(y_j, u_{\pi(j)}) = \arg \min_{u_j \times \hat{x}_j \in (\mathcal{U}_j \times \mathcal{X}_j)} \sum_{x_j \in S_{x_j}} p(y_j|x_j) \tilde{P}^*_j(u_{\pi(j)}|x_j) \left[c_j(u_j, \hat{x}_j, x_j) + C^*_j(u_j, x_j) \right]$$

where $\tilde{P}^*_j$ is given by Eq.(3.40),

$$\tilde{\gamma}^{*3}_j(y_j, u_{\pi(j)}) = \arg \min_{u_j \times \hat{x}_j \in (\mathcal{U}_j \times \mathcal{X}_j)} \sum_{x_j \in S_{x_j}} p(y_j|x_j) \tilde{P}^*_j(u_{\pi(j)}|x_j) \left[c_j(u_j, \hat{x}_j, x_j) + \tilde{C}^*_j(u_j, x_j) \right] \quad (3.50)$$

where $\tilde{C}^*_j(u_j, x_j)$ and $\tilde{P}^*_j$ are given by Eq.s (3.48) and (3.49) respectively. Hence, in order to obtain $\tilde{\gamma}^{*3}_j$ we have utilized the proposed particle representations and approximate computational schemes for all terms that depend on $\gamma^*_{\setminus j}$ including the node-to-node terms, altough it is assumed that $\gamma^*_{\setminus j}$ are known exactly.

On the other hand, given S_{x_j} and S_{y_j}, the approximation framework is valid for the rules local to any node $j \in \mathcal{V}$. Moreover, owing to fusing the message likelihoods via importance sampling, the node-to-node terms would utilize the same discretization of their domains regardless of which local rule is subject to approximation. In addition, it is possible to treat the RHS of the expressions within the framework as operators valid for any strategy $\gamma \in \Gamma^{\mathcal{G}}$ including those in the "approximating" form given by Eq. (3.50). For the rest of this chapter, we refer to as an approximation, one of the appropriate full-step approximations appearing in Eq. (3.50) and denote with a single tilde, e.g. we denote $\tilde{\gamma}^{*3}_j$ with $\tilde{\gamma}^*_j$. Let us summarize the Monte Carlo framework with

$$\begin{aligned}
\tilde{\phi}_j(S_{x_j}, \hat{x}_j) &= \tilde{d}_j(\tilde{P}_j(S_{x_j}), \tilde{C}_{\chi(j)\to j}) \\
\tilde{P}_j(S_{x_j}) &= \tilde{f}_j(\tilde{P}_{\pi(j)\to j}) \\
\tilde{P}_{j\to\chi(j)} &= \tilde{g}_j(\tilde{\phi}_j(S_{x_j}, \hat{x}_j), \tilde{P}_j(S_{x_j})) \\
\tilde{C}_{j\to\pi(j)} &= \tilde{h}_j(\tilde{\phi}_j(S_{x_j}, \hat{x}_j), \tilde{P}_{\pi(j)\to j}, \tilde{C}_{\chi(j)\to j})
\end{aligned}$$

where

$$\tilde{P}_j(S_{x_j}) = \{(\tilde{P}_j(u_{\pi(j)}|x_j), u_{\pi(j)}, x_j) | u_{\pi(j)} \in \mathcal{U}_{\pi(j)} \wedge x_j \in S_{x_j}\}$$

$$\tilde{P}_{\pi(j)\to j} = \{\tilde{P}_{i\to j}(S_{x_i})\}_{i\in\pi(j)}$$

$$\tilde{P}_{i\to j}(S_{x_i}) = \{(\tilde{P}_{i\to j}(u_{i\to j}, x_i), u_i, x_i) | u_{i\to j} \in \mathcal{U}_{i\to j} \wedge x_i \in S_{x_i}\}$$

$$\tilde{P}_{j\to\chi(j)} = \{\tilde{P}_{j\to k}(S_{x_j})\}_{k\in\chi(j)}$$

$$\tilde{C}_{\chi(j)\to j} = \{\tilde{C}_{k\to j}(S_{x_j})\}_{k\in\chi(j)}$$

and $\tilde{\phi}_j(S_{x_j}, \hat{x})$ is given by

$$\left\{(p(y_j|x_j)\tilde{P}_j(u_{\pi(j)}|x_j)\left[c(u_j, \hat{x}_j, x_j) + \tilde{C}_j(u_j, x_j)\right], u_j, x_j) | u_j \in \mathcal{U}_j, u_{\pi(j)} \in \mathcal{U}_{\pi(j)}, x_j \in S_{x_j}\right\}$$

Note that $\tilde{\phi}_j(S_{x_j}, \hat{x}_j)$ is not exactly a discretization [3] of ϕ_j and $\tilde{C}_{k\to j}(S_{x_j})$ can be defined in a similar fashion to $\tilde{P}_{i\to j}(S_{x_i})$.

It is immediately possible to employ this framework in Algorithm 2 and achieve a Monte Carlo optimization algorithm which, starting with initial local rules, iteratively results a strategy that corresponds to performing computations to approximate a person-by-person optimal one. Given by Algorithm 4, this scheme maintains the message passing interpretation admitted in the `Update` step.

Starting with $\mathcal{G}$, $\{\mathcal{U}_{i\to j} | (i, j) \in \mathcal{E}\}$, each node initially maintains the knowledge of $p(x_{\pi(j)}, x_j)$ and $c(u_j, \hat{x}_j, x_j)$. As soon as samples from the marginal distributions, i.e. S_{x_j}, together with samples from the marginal distributions of the observation processes, i.e. S_{y_j}, are generated for all $j \in \mathcal{G}$ and an initial local rule $\gamma_j^0 \in \Gamma^{\mathcal{G}}$ is selected, the iterative scheme yields a set of local rules such that each node performs computations corrresponding to an approximation to a person-by-person optimum.

The approximate computation of the expected cost required in the `Check` step of Algorithm 4 for any given strategy, i.e. $\tilde{J}(\gamma)$ is obtained through a Monte Carlo approximation $\tilde{G}_j(\gamma_j^l)$ to Eq.(3.33) using the usual sample sets, i.e.S_{x_j} and S_{y_j}, as

$$\tilde{G}_j(\tilde{\gamma}_j^l) = \frac{1}{M}\sum_{m=1}^{M}\sum_{u_{\pi(j)}\in\mathcal{U}_{\pi(j)}} \tilde{P}_j^{l+1}(u_{\pi(j)}|x_j^{(m)})\times \frac{1}{\sum_{p=1}^{P}\omega_k^{(m)(p)}}\sum_{p=1}^{P}\omega_k^{(m)(p)} c_j(\,[\gamma_j(y_j^{(p)}, u_{\pi(j)})]_{\mathcal{U}_j},\, [\gamma_j(y_j^{(p)}, u_{\pi(j)})]_{\mathcal{X}_j}, x_j^{(m)}) \quad (3.51)$$

[3] Considering Eq.(3.15), for the evaluation of $\phi_j(u_j, \hat{x}_j, x_j; u_{\pi(j)})$, the argument $\hat{x}_j$ needs not to be discretized since only $c(u_j, \hat{x}_j, x_j)$ acts on it which is assumed to be known over all of its domain. On the other hand, the conventional Monte Carlo approximation drops $p(x_j)$ and discretizes ϕ in x_j.

Algorithm 4 Iterations converging to an approximate person-by-person optimal decentralized estimation strategy over a directed acyclic graph $\mathcal{G}$.

0) `(Initiate)` $l = 0$;

Choose $\gamma^0 = (\gamma_1^0, \gamma_2^0, ..., \gamma_N^0)$ such that $\gamma_j^0 \in \Gamma_j^{\mathcal{G}}$ for $j = 1, 2, ..., N$

1) `(Update)` $l = l + 1$;

`i) (Forward pass)`

`for` $j = 1, 2, ..., N$

$\tilde{P}_j^{\,l}(S_{x_j}) = \tilde{f}_j(\left\{\tilde{P}_{i\to j}^{l}(S_{x_i})\right\}_{i\in\pi(j)})$

$\left\{\tilde{P}_{j\to k}^{l}(S_{x_j})\right\}_{k\in\chi(j)} = \tilde{g}_j(\tilde{\phi}_j^{\,l-1}(S_{x_j}, \hat{x}_j), \tilde{P}^l{}_j(S_{x_j}))$

`ii) (Backward pass)`

`for` $j = N, N-1, ..., 1$

$\tilde{\phi}_j^{\,l}(S_{x_j}, \hat{x}_j) = \tilde{d}_j(\tilde{P}_j^{\,l}(S_{x_j}), \left\{\tilde{C}_{k\to j}^{l}(S_{x_j})\right\}_{k\in\chi(j)})$

$\left\{\tilde{C}_{j\to i}^{l}(S_{x_i})\right\}_{i\in\pi(j)} = \tilde{h}_j(\tilde{\phi}_j^{\,l}(S_{x_j}, \hat{x}_j), \left\{\tilde{P}_{i\to j}^{l}(S_{x_i})\right\}_{i\in\pi(j)}, \left\{\tilde{C}_{k\to j}^{l}(S_{x_j})\right\}_{k\in\chi(j)})$

2) `(Check) If` $\tilde{J}(\tilde{\gamma}^{l-1}) - \tilde{J}(\tilde{\gamma}^{l}) < \varepsilon$ `STOP, else GO TO 1;`

where $\omega_k^{(m)(p)} = p(y_k^{(p)}|x_k^{(m)})/p(y_k^{(p)})$. If Assumption 5 holds, the expression above turns to

$$\tilde{G}_j(\tilde{\gamma}_j^l) = \frac{1}{M}\sum_{m=1}^{M}\sum_{u_{\pi(j)}\in\mathcal{U}_{\pi(j)}} \tilde{P}_j^{l+1}(u_{\pi(j)}|x_j^{(m)})\times \left[\tilde{J}_{d|x_j^{(m)},u_{\pi(j)}} + \lambda\sum_{u_j\in\mathcal{U}_j} c_j^c(u_j, x_j^{(m)})\tilde{p}(u_j|x_j^{(m)}, u_{\pi(j)}; \mu_j^l)\right] \quad (3.52)$$

Note that $\{J(\gamma^l)|l = 0, 1, 2, ...\}$ obtained through Algorithm 2 is non-increasing whereas $\{\tilde{J}(\tilde{\gamma}^l)\}$, being a Monte Carlo approximation to the former, is not necessarily ehxibits this property. Let us defined an approximation error sequence $err[l] = J(\gamma^l) - \tilde{J}(\tilde{\gamma}^l)$. This sequence will be identically zero with probability one as $M, P \to \infty$. For finite M and P, it is possible to smooth the fluctuation of $err[l]$ through filtering and utilize the corresponding termination condition, e.g. check whether $\tilde{J}(\tilde{\gamma}^l) * h[l] < \varepsilon$ where $h[l]$ is the impulse response of a linear, time invariant filter and $*$ denotes convolution. Further investigation of robust termination of Algorithm 4 is beyond the scope of this thesis.

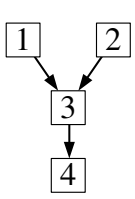
3.5 Examples

In this section we utilize the proposed scheme given by Algorithm 4 in example scenarios. First, we consider a random vector X with a Gaussian a-priori distribution. The observation noise processes are similarly assumed to be Gaussian. After presenting the results for this canonical distribution family in the example scenario, we consider a heavy tailed a-prior distribution; a multidimensional symmetric Laplacian, in particular. It is not straightforward to perform decentralized inference on such distributions which do not admit graphical model representations (e.g. a Markov random field representation) or closed form expressions for the canonical Bayesian estimation schemes. Algorithm 4, on the other hand, utilizing sample sets generated from marginal distributions produces results even for such cases.

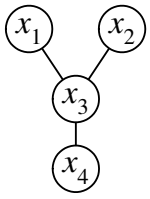
3.5.1 A Gaussian Example

We consider an example scenario in which a decentralized estimation network comprised of four platforms perform an estimation task. A Gaussian random field $X = \{X_1, X_2, X_3, X_4\}$ is of concern and platform j is associated with X_j. We assume the underlying communication structure described by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in Figure 3.1a. We note that $\mathcal{G}$ includes partitions of a star topology (induced by nodes $\{1, 2, 3\}$), and series topologies (induced by nodes $\{1, 3, 4\}$ and $\{2, 3, 4\}$. We assume that the bandwidth constraints render $\mathcal{U}_{1\to3} = \mathcal{U}_{2\to3} = \mathcal{U}_{3\to4} = \{0, 1, 2\}$. The online processing scheme operates as given in Section 3.2. Since nodes 1 and 2 are parentless, upon measuring y_1 and $y_2 \in \mathbb{R}$ induced by X_1 and X_2, they evaluate their local rules as $(u_{1\to3}, \hat{x}_1) = \gamma_1(y_1)$ and $(u_{2\to3}, \hat{x}_2) = \gamma(y_2)$ respectively. Upon receiving these messages and measuring $y_3 \in \mathbb{R}$ induced by X_3 node 3 evaluates its local rule $(u_{3\to4}, \hat{x}_3) = \gamma_3(y_3, u_{1\to3}, u_{2\to3})$, and similarly node 4 evaluates $\hat{x}_4 = \gamma_4(y_4, u_{3\to4})$. We note that the conditional independence and observation locality together with the polytree topology hold and the strategy $\gamma = (\gamma_1, ..., \gamma_4)$ is subject to design.

In addition we comply with Assumption 3 and select separable local costs providing Assumption 5 to hold. The cost function local to node j is in a separable form given by $c_j(u_j, \hat{x}_j, x_j) = c_j^d(x_j, \hat{x}_j) + \lambda c_j^c(u_j, x_j)$ where c_j^d and c_j^c penalize estimation errors and communication respectively. Therefore λ is a unit conversion coefficient admitting the interpretation



(a) Illustration of $\mathcal{G}$.



(b) Illustration of the MRF $\mathcal{G}_X$.

Figure 3.1: (a) Illustration of the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{X})$ where $\mathcal{V} = \{1, 2, 3, 4\}$ and $\mathcal{E} = \{(1, 3), (2, 3), (3, 4)\}$ which represents the decentralized estimation network that admits the online processing described in Section 3.2, (b) illustration of the Markov Random Field $\mathcal{G}_X$ subject to estimation by the decentralized estimation network in example.

of equivalent estimation penalty per unit cost of communication.

$$c_j^c(u_j, x_j) = \sum_{k \in \chi(j)} c_{j\to k}^c(u_{j\to k}, x_j)$$

where $c_{j\to k}^c(u_{j\to k})$ is the cost of transmitting the symbol $u_{j\to k}$ on the link $(j, k) \in \mathcal{E}$. It is selected as

$$c_{j\to k}^c(u_{j\to k}, x_j) = \begin{cases} 0, & \text{if } u_{j\to k} = 0 \\ 1, & \text{otherwise} \end{cases}$$

indicating the link use. Hence, $\mathcal{U}_{j\to k}$ together with $c_{j\to k}^c$ define a selective communication scheme where $u_{j\to k} = 0$ indicates no communications and otherwise transmission of a 1 bit message. The estimation error is penalized by $c_j^d(x_j, \hat{x}_j) = (x_j - \hat{x}_j)^2$. Hence the total cost of a strategy is $J(\gamma) = J_d(\gamma) + \lambda J_c(\gamma)$ where J_d is the MSE and J_c is the total link use rate.

The random field of concern is a multivariate Gaussian, i.e. $x \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_X)$, and Markov with respect to the graph $\mathcal{G}_X$ presented in Figure (3.1b). The covariance matrix is given by

$$\mathbf{C}_x = \begin{bmatrix} 2 & 1.125 & 1.5 & 1.125 \\ 1.125 & 2 & 1.5 & 1.125 \\ 1.5 & 1.5 & 2 & 1.5 \\ 1.125 & 1.125 & 1.5 & 2 \end{bmatrix} \tag{3.53}$$

which conforms with the Markov properties of $\mathcal{G}_X$. Altough the communication structure of the DE network is not related with the MRF representation of X and Algorithm 3 would produce results for any choice, for sake of simplicity we selected the graph in Figure 3.1b as the undirected counterpart of that in Figure 3.1a.

The noise processes n_j for $j \in \mathcal{V}$ are additive, mutually independent and given by $n_j \sim \mathcal{N}(0, 0.5)$, so that the observation likelihoods are $p(y_j|x_j) = \mathcal{N}(x_j, 0.5)$. Considering $\mathbf{C}_X$, each

sensor has an SNR of 6dB.

Since separable local cost functions are utilized, the person-by-person optimal rules are also splitted into estimation and communications functions given by Eq.(3.22) and (3.23) respectively. We initiate the local rules, i.e. ν_j^0 and μ_j^0 for $j \in \mathcal{V}$, as follows:

1. All the nodes, regardless of the incoming messages apply a myopic inference rule which is to make decisions regardless of incoming messsages as if they do not carry any information. We select this rule as the myopic MMSE estimation rule, i.e $E\{X_j|Y_j = y_j\}$ given by

$$\nu_j^0(y_j, u_{\pi(j)}) = \int_{-\infty}^{\infty} \mathrm{d}x_j \; x_j p(x_j|y_j) \tag{3.54}$$

2. All the nodes apply an initial communication rule as a quantization of the observation y_j, i.e.

$$\mu_i^0(y_i, u_{\pi(i)}) = \begin{cases} 1 \, , \, y_i < -2\sigma_n \\ 0 \, , \; -2\sigma_n \leqslant y_i \leqslant 2\sigma_n \\ 2 \, , \, y_i > 2\sigma_n \end{cases} \tag{3.55}$$

Considering $J(\gamma) = J_d(\gamma) + \lambda J_c(\gamma)$ and person-by-person optimal estimation strategies achieved by Algorithm 2, in principle, different values of λ would yield different performance points $(J_c(\gamma^*), J_d(\gamma^*))$. Moreover, after a certain value $\lambda = \lambda^*$, the communication cost λJ_c will dominate such that the decrease in the decision cost J_d with the contributions of the communicated symbols will not be enough to decrease J and symbol 0 will be the best choice. Moreover, the individual estimators will be the myopic rules, since myopic rules with no communications constitute a person by person optimal strategy. Hence, it is possible to interpret λ^* as the maximum price per bit that the system affords to decrease the expected estimation error. As we increase λ from 0 we obtain approximate points from the performance curve for Problem (P1) which lets us to quantify the tradeoff between the cost of estimation errors and communication.

In Figure (3.2) we present approximate pairs $(\tilde{J}_c, \tilde{J}_d)$ of the converged strategies $\tilde{\gamma}^*$ using Algorithm 4, for different choices of λ and $\left|\mathcal{U}_{i\to j}\right|$s, where J_c is the total link use rate and J_d is the total MSE. The upper and lower limits are MSEs corresponding to the myopic rule and the centralized optimal rule respectively. Note that, for the squared error cost, optimal

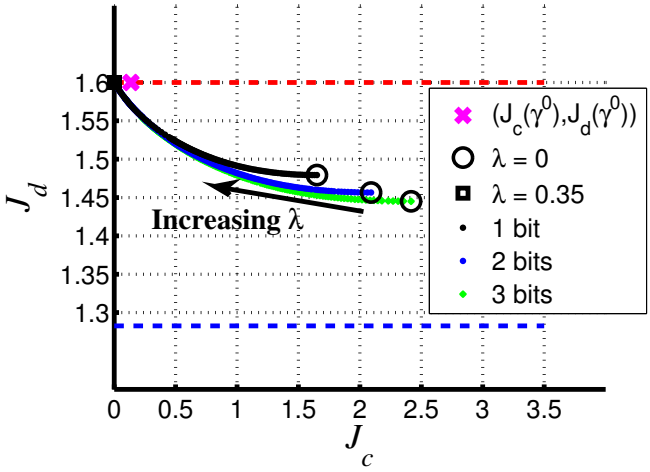


Figure 3.2: The example scenario involving a Gaussian joint $p(x, y)$ enables the exact computation of the MSE for the optimum centralized and myopic rules corresponding to the lower bound (blue dashed-line) and upper bound (red dashed-line) respectively. The estimation network in accordance with the directed acyclic graph $\mathcal{G}$ (Figure 3.1) is subject to optimization through Algorithm 4 starting with the initial rules given by Eq.s(3.54)and(3.55). The initial strategy achieves the performance $(J_c(\gamma^0), J_d(\gamma^0))$ given by the magenta cross. The pareto-optimal performance curves achieved for the approximate person-by-person optimal strategies while λ is increased from 0 with steps of 0.001 are approximated by $\{(\tilde{J}_c(\tilde{\gamma}^*_\lambda), \tilde{J}_d(\tilde{\gamma}^*_\lambda))\}$ where $\tilde{\gamma}^*_\lambda$ is the optimum strategy corresponding to λ. Results for $1, 2$ and 3 bits selective communication schemes are presented by black, blue and green dots.

centralized rule given by $E\{X|Y = y\}$ and causes a communication cost of $J_c = 3Q$ where Q is the number of bits used to represent a real number y_j before transmitting to the fusion center. Considering $(\tilde{J}_c, \tilde{J}_d)$ pairs for the 1-bit selective communication scheme, for $\lambda = 0$, the transmission has no cost, but the link use rate is well below 75% of the total 3 bits. This indicates that the information of receiving no messages is successfully maintained in this perspective. Moreover, the communication stops for $\lambda^* \approx 0.355$. Similarly, approximate points for 2-bits and 3-bits schemes indicate that, if λ is small enough, we can achieve less MSE for the same total communication load as we increase the link capacities.

3.5.2 A Heavy Tailed Example

Altough, being a Monte Carlo scheme, Algorithm 4 produces strategies that are approximating a person-by-person optimum, utilizing sample sets drawn from the marginal distributions of interest, it applies for arbitrary distributions provided that samples can be generated from them. In certain problem settings, even the centralized rule do not yield closed form expres-

sions. We consider such a scenario in which a random vector X is distributed by a heavy tailed prior $p(x)$, specifically a multivariate-symmetric Laplacian given by

$$p(x) = \frac{2}{(2\pi)^{d/2}|C_x|^{1/2}} \left(\frac{x^T C_x^{-1} x}{2}\right)^{1-d/2} K_{1-d/2}(\sqrt{2x^T C_x^{-1} x}) \tag{3.56}$$

where d is the dimension of x, C_x is the covariance matrix, and $K_\lambda(u)$ is the Bessel function of the second kind (also known as the Bessel function of the third kind) of order λ given by

$$K_\lambda(u) = \frac{(u/2)^\lambda \Gamma(1/2)}{\Gamma(\lambda + 1/2)} \int_1^\infty e^{ut}(t^2 - 1)^{\lambda - 1/2} \mathrm{d}t \tag{3.57}$$

for $u > 0$ and $\lambda \geq -1/2$ where Γ is the Gamma function given by $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$ [103]. Let us denote this distribution by $SL_d(\mathbf{C}_X)$. Unlike the Gaussian case, uncorrelatedness does not imply independence in the case of $SL_d(\mathbf{C}_X)$ and not being a member of the exponential family this distribution does not imply a Markov random field.

On the other hand, it is possible to generate samples from a multivariate symmetric Laplacian utilizing samples generated from a multivariate Gaussian of zero mean vector and the desired covariance matrix together with samples drawn from the unit univariate exponential distribution. Given $u \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_X)$ and $z \sim e^{-z}$, generate samples of X by $x = \sqrt{z}u$, then $x \sim SL_d(\mathbf{C}_\mathbf{x})$. Another implication is that it is possible to express $SL_d(\mathbf{C}_X)$ as

$$p(x) = \int_0^\infty \mathcal{N}(0, z\mathbf{C}_X) p(z) \mathrm{d}z \tag{3.58}$$

where $p(z) = e^{-z}$. This form, being a scaled sum of Gaussians, generalizes Gaussian mixtures and hence also referred to as a *scale mixture of Gaussians*. This family of distributions has been employed to model multiple variables that exhibit uncorrelatedness yet dependence such as the statistics of natural images (see e.g. [104][105]) and inference on such models exploit the underlying Gaussian model (see e.g. [106]).

It is possible to approximately model an arbitrary prior in terms of Gaussian mixtures. Our concern is to select a general model for a prior distribution to assess the performance of Algorithm 4. For the Gaussian mixtures, the Monte Carlo framework trivially follows the Gaussian case. We consider a generalization of Gaussian mixtures which turn out to be a multivariate-symmetric Laplacian $SL_d(\mathbf{C}_X)$. To the best knowledge of the author' s, this prior is also not suitable for decentralized inference in the sense that the exponential variable z requires centralized coordination.

For our case, we exploit the fact that j^{th} marginal distribution of $SL_d(\mathbf{C}_X)$ is given by $SL_1([\mathbf{C}_X]_{j,j})$ and it is straightforward to generate samples from these marginals [107]. We consider

$X = \{X_1, X_2, X_3, X_4\}$ such that $p_X(x) = SL_4(\mathbf{C}_X)$ where the underlying Gaussian is that has been considered in the previous section, i.e. $\mathbf{C}_X$ is given by Eq.(3.53). Similar to that in the previous section, we assume the underlying communication structure described by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in Figure 3.1a together with the link capacities, costs and observation likelihoods.

For the case, altough the marginal densities of observations yield closed form expressions, i.e.

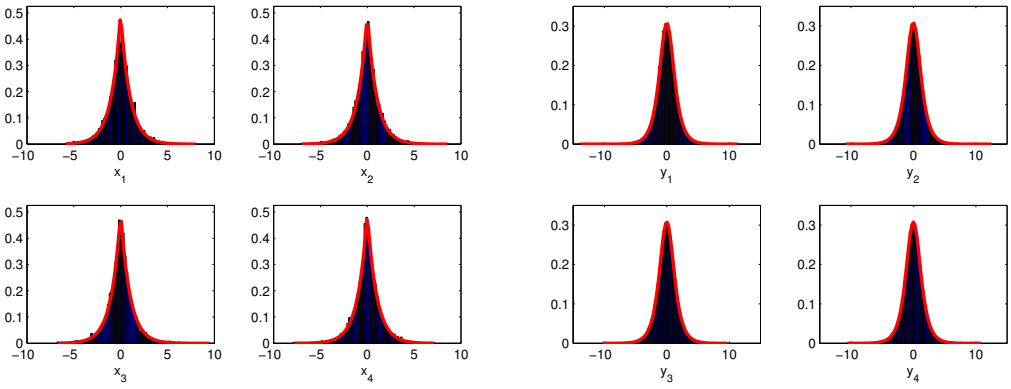
$$\begin{aligned} p(y_j) &= \int_{-\infty}^{\infty} p(x_j)p(y_j|x_j)\mathrm{d}x_j \\ &= \int_{-\infty}^{\infty} SL_1(\sigma_j^2)\mathcal{N}(0, \sigma_{n_j}^2)\mathrm{d}x_j \\ &= 0.1410\ \sqrt{\pi}e^{-y_j+1/4}\left(e^{2y_j} + 1 - \Phi\left(y_j + 1/2\right)e^{2y_j} + \Phi\left(y_j - 1/2\right)\right) \end{aligned} \tag{3.59}$$

for $j = 1, 2, 3, 4$ where Φ is the error function, it is not easy to sample from this density since it does not yield a distribution function in closed form. However, considering the mixture approximation $\sum_{x_j^{(m)} \in S_{x_j}} p(y_j|x_j^{(m)}) \approx p(y_j)$ where $S_{x_j} = \{x_j^{(1)}, x_j^{(2)}, ..., x_j^{(M)}\}$, it is possible to draw samples from approximately $p(y_j)$ by sampling from $p(y_j|x_j^{(m)})$ for $m = 1, 2, ..., M$ until the density *mixes*.

We generate 3000 samples from the prior marginals, i.e. $S_{x_j} = \{x_j^{(1)}, x_j^{(2)}, ..., x_j^{(M)}\}$ where $x_j^{(m)} \sim p(x_j)$ for $j = 1, 2, 3, 4$ and $M = 3000$, together with 45000 samples from the mixture density, i.e. $S_{y_j} = \{y_j^{(1)}, y_j^{(2)}, ..., y_j^{(P)}\}$ where $y_j^{(p)} \sim p(y_j)$ for $j = 1, 2, 3, 4$ and $P = 45000$. The histograms of the sample sets S_{x_j}s and S_{y_j}s together with the corresponding values of the marginal densities are presented in Figure 3.3.

We run Algorithm 4 for different choices of λ and 10 different sample sets each. The results are presented in Figure 3.4 in which for each value of λ an ellipsoid is fit to the performance points achieved for the sample sets using the algorithm in [108]. The upper and lower bounds are the mean squared error corresponding to the myopic and centralized rules respectively. Another intricacy is that the evaluation of the myopic and centralized MMSE rules as well as the corresponding MSE computations require numerical approximations for which we utilize Monte Carlo methods as well.

We observe that the variance of the approximation error is higher than that for the Gaussian case. This is in accordance with the expectation that heavy tailed distributions require utilization of larger sample sets. The framework we propose produces distributed solutions within a reasonable deviation in the problem settings which do not lead straightforward solutions even in the centralized case.



(a) Histograms of S_{x_j} for $j = 1, 2, 3, 4$.

(b) Histograms of S_{y_j} for $j = 1, 2, 3, 4$.

Figure 3.3: (a) Histogram and values of the corresponding marginal density (red dots) for S_{x_1} (upper-left), S_{x_2} (upper-right), S_{x_3} (lower-left) and S_{x_4} (lower-right), each constituted of 3000 samples. (b) Histogram and values of the corresponding marginal density (red dots) for S_{y_1} (upper-left), S_{y_2} (upper-right), S_{y_3} (lower-left) and S_{y_4} (lower-right), each constituted of 45000 samples, reveal that the generation scheme exhibits an appropiate *mixing*.

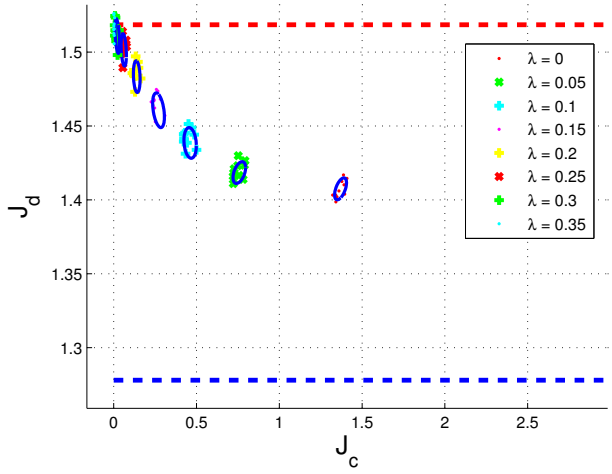


Figure 3.4: Approximate performance points corresponding to the strategies achieved using Algorithm 4 in the heavy tailed prior case. The results for the 10 sample sets for $\lambda = 0, 0.05, ..., 0.35$ are displayed. For each value of λ an ellipsoid is fit to the performance points for the sample sets.

3.6 Discussion

One of the main benefits of the framework we have introduced is that the strategy optimization is carried out in a manner exhibiting scalability in the number of nodes. In addition, the message passing structure of the original framework is maintained and since the particle representations involve samples from the marginals the sample sets can easily be generated. Hence, Algorithm 4 can be applied for problems involving a large number of variables and platforms. In addition, the estimation task is distributed through random variable-node associations and the nodes responsible for inference is a design choice. Therefore, while some nodes perform estimation, it is possible to select the remaining as fusion nodes providing information to the estimators.

In addition, the original discussion for decentralized detection over directed acyclic networks involves a channel model and considers unreliable communications as well [27]. Altough we proceed the discussion under error-free links assumption, the approximation framework we provide inherently integrates with the channel models associated with the directed acyclic graph.

Besides large network problems, another scenario that involves a directed structure is target tracking presented in Section 1.1.4. Multiple sensors receive measurements due to a target and at the final stage, the position of the target is subject to estimation. For the case, all the observations are induced by a common random variable (target position) and estimation is carried out by the node which is predicted to be closer to the target considering the track information. Therefore, a directed topology in which there is only one childless node that is selected to perform the estimation whereas the remaining contributes with messages yields a communication constrained target localization scheme (Figure 1.7). Specifically, it is reasonable to assume that in the network self-configuration stage, all nodes are equipped with the coverage information which renders a Voronoi tesellation. Then, for each patch, it is possible to solve the communication constrained problem under a uniform prior assumption. During tracking, the network switches its mode in accordance with the prediction of the target location and runs the corresponding strategy over the corresponding network. In order to fuse the prediction density with the estimation rule which is designed under a uniform prior assumption, it is possible to utilize Kernel methods. The benefit of providing estimation strategies through the framework we have introduced is that, equipped with a set of strategies achiev-

ing varying estimation accuracy and communication cost enables an energy-aware operation through sensor management schemes. In this particular scenario, the localization accuracy can be traded off for operational lifetime.

CHAPTER 4

DECENTRALIZED ESTIMATION NETWORKS CONSTRAINED BY UNDIRECTED GRAPHS

In this chapter, we discuss decentralized estimation networks with an underlying communication structure that admits an undirected graph model. Unlike the directed case in which the communication links are uni-directional rendering a directed acyclic topology, we are concerned with bidirectional links that allow transmissions among neighboring platforms. Such a model might be regarder to be a better match for the ad-hoc nature of wireless sensor networks considering that the underlying communications often utilize 1-hop transmissions (Section 1.1.1).

However, it is not straightforward to achieve a causal online processing that stops in finite steps under such a model. We consider the class of decentralized detection strategies over undirected graphs summarized in Section 2.7.2. The online processing presented in Section 2.7.2.1 is considered in [26] and further developed in [28] in the context of detection. For the estimation problem, we adopt this approach in which local rules performing in two stages are employed and an N dimensional random vector is of concern which takes values from a denumarable set $\mathcal{X}$. In the first stage, based on its observation each node evaluates its communication rule and outputs messages to neighbors. After collecting all the incoming messages from neighbors, each node proceeds with the second stage and based on its observation as well as the incoming messages from the neighbors evaluates its estimation rule. Such an online processing scheme prevents any deadlock and yields a tractable optimization scheme under a team theoretic treatment.

We note that, the local structure of networks constrained by directed graphs outputs the estimation together with the outgoing messages in single step, in general, based on the incoming

messages and the observation. The causal online scheme exhibiting no deadlock is guaranteed by the acyclic nature. For the case, the information flow renders nodes with more ancestors in a more convenient state from the viewpoint of estimation accuracy. The two stage local rules under an undirected graph underlines the nodes with more neighbors. However, the information horizon is limited with the neighbors in contrast to the directed case in which the incoming messages are due to the set of all ancestors.

Feedback schemes are more commonly considered for the detection problem (see e.g. [109]) rather than in a static estimation setting and with an emphasis on messages from finite sets. In [96], under a star topology, after collecting the incoming messages from the peripheral nodes, a fusion center makes a decision on which partition the variable of concern takes values and provides a feedback accordingly. This scheme is iterated until the interval of uncertainity is narrow enough. However, the variable representing the decision on the partition selection does not provide conditional independence for the observations and the fusion of messages are carried out using Monte Carlo approximations.

We are concerned with extending the topology as well as introducing the cost of communications within the problem setting. We introduce a similar framework that is provided in Chapter 3 noting that under reasonable assumptions, the analysis for the optimality conditions under a directed network applies to the unwrapped counterpart of the undirected graph. In Section 4.1 we overview the information structure exhibited under the two stage rules over an undirected graph. In Section 4.2 we present a detailed discussion on the estimation counterpart of the results summarized in Section 2.7.2.2. Then, similar to the Monte Carlo framework provided over directed graph, we introduce particle representations and approximate computational schemes for optimizing two stage local rules over undirected graphs such that a Bayesian cost composed of both penalty for estimation errors and cost of communications is decreased in Section 4.3. Then we provide applications on example scenarios and conclude.

4.1 Online Processing Modelled With Undirected Graphs

Consider the online processing regarding two stage decentralized strategies over undirected graphs presented in Section 2.7.2.1. Summarizing here for convenience; a set of platforms are

represented with the index set $\mathcal{V} = \{1, ..., N\}$, and with each $j \in \mathcal{V}$ a random variable X_j is associated that takes values from the set $\mathcal{X}_j$ which, unlike the detection case, is denumerable. $X = (X_1, ..., X_N)$ is the random field of concern where a realization x satisfies $x \in \mathcal{X}$ with $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_N$. Given a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected graph if it holds that $(i, j) \in \mathcal{E}$ implies $(j, i) \in \mathcal{E}$. Given $\mathcal{G}$, each edge $(i, j) \in \mathcal{E}$ corresponds to a communication link of capacity $\log_2(\left|\mathcal{U}_{i\to j}\right| - 1)$ bits such that $\mathcal{U}_{i\to j}$ is the set of admissible symbols with the symbol $0 \in \mathcal{U}_{i\to j}$ indicating no transmission.

Let $u_{ne(j)} \triangleq \{u_{i\to j} | i \in ne(j)\}$ denote the incoming messages to node j from neighbor nodes $ne(j)$, which takes values from $\mathcal{U}_{ne(j)} = \mathcal{U}_{ne_j^1\to j} \times ... \times \mathcal{U}_{ne_j^D\to j}$. Here $ne(j) = \{ne_j^1, ...,ne_j^D\}$. The outgoing messages from node j to neighbor nodes $ne(j)$ is given by $u_j \triangleq \{u_{j\to i} | i \in ne(j)\}$ and takes values from $\mathcal{U}_j$ which can be defined similarly with that for $\mathcal{U}_{ne(j)}$. The overall communication load is $u \triangleq \{u_{i\to j} | (i, j) \in \mathcal{E}\}$ and takes values from $\mathcal{U} = \mathcal{U}_1 \times ... \times \mathcal{U}_N$.

A causal online processing of measurements $\{y_j | j \in \mathcal{V}\} \in \mathcal{Y}$ where $\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_N$ takes place when each $j \in \mathcal{V}$, first performs its local communication rule $\mu_j : \mathcal{Y}_j \to \mathcal{U}_j$ based on only y_j , and as soon as $u_{ne(j)}$ are collected, proceeds with the local estimation rule $\nu_j : \mathcal{Y}_j \times \mathcal{U}_{ne(j)} \to \mathcal{X}_j$.

Let $\gamma_j = (\mu_j, \nu_j)$ and $\gamma = (\gamma_1, ..., \gamma_N)$ denote the local rule of node j and the strategy of the network respectively. Let $\mathcal{M}_j$ and $\mathcal{N}_j$ denote the set of all possible communication and estimation rules respectively local to node j. Then, $\Gamma_j = \mathcal{M}_j \times \mathcal{N}_j$ for $\gamma_j \in \Gamma_j$ and the set of possible strategies given $\mathcal{G}$ is $\Gamma^{\mathcal{G}} = \Gamma_1 \times ... \times \Gamma_N$.

Consider the random vector X that takes values from a denumerable set $\mathcal{X}$ and we assume that Assumption 1-3 and 5 given in Section 2.7.1 hold. Let $\Gamma^{\mathcal{G}}$ denote the set of all strategies that are composed of two stage local rules $\mathcal{G}$. The expected cost $J(\gamma) = E\{c(u, \hat{x}, x); \gamma\}$ given $\gamma \in \Gamma^{\mathcal{G}}$ is then

$$\begin{aligned} J(\gamma) &= E\{c^d(x, \hat{x}) + \lambda c^c(u, x); \gamma\} \\ &= J_d(\gamma) + \lambda J_c(\gamma) \end{aligned}$$

where the underlying distribution is rewritten here for convenience;

$$p(u, x, \hat{x}; \gamma) = \int_{y\in\mathcal{Y}} \mathrm{d}y \; p(u, \hat{x} | x, y; \gamma) p(x, y)$$

with the conditional imposed by the two-stage local rules over the undirected structure as

$$\begin{aligned} p(u,\hat{x}|x,y;\gamma) &= \prod_{j\in\mathcal{V}} p(u_j,\hat{x}_j|y_j,u_{ne(j)};\gamma_j) \\ &= \prod_{j\in\mathcal{V}} p(u_j|y_j;\mu_j)p(\hat{x}_j|y_j,u_j,u_{ne(j)};\nu_j) \\ &= \prod_{j\in\mathcal{V}} p(u_j|y_j;\mu_j)p(\hat{x}_j|y_j,u_{ne(j)};\nu_j) \end{aligned} \tag{4.1}$$

The last step above is under Assumption 1 and 5 with $\mu_j \in \mathcal{M}_j$ denoting the stage-one communication rule and $\nu_j \in \mathcal{N}_j$ denoting the stage-two estimation rule local to node j. It follows that, the local rules and the conditional densities determine each other with

$$p(u_j|y_j;\mu_j) = \delta_{u_j,\mu_j(y_j)} \tag{4.2}$$

and

$$p(\hat{x}_j|y_j,u_{ne(j)};\nu_j) = \delta(\hat{x}_j - \nu_j(y_j,u_{\pi(j)})) \tag{4.3}$$

where $\delta_{i,j}$ and $\delta(.)$ are the Kronecker' s and Dirac' s delta respectively.

Therefore, for any two-stage strategy over an undirected graph $\mathcal{G}$, i.e. $\gamma \in \Gamma^{\mathcal{G}}$, there corresponds an expected cost $J(\gamma) = J_d(\gamma) + \lambda J_c(\gamma)$ where $J_d(\gamma)$ and $J_c(\gamma)$ are penalties for estimation errors and communication load respectively. $J(\gamma)$ is computed by substituting the distributions above in Eq.(4.1) and unlike the detection case, integrations over $\mathcal{X}$ appear regarding the random variables X and $\hat{X}$. Similarly the problem definition (P1) given by Expression (2.22) is valid under the interpretation that the constraints of the feasible set describe the set of decentralized estimation strategies comprised of two stage local rules over an undirected graph.

4.2 Team Theoretic Solution Under Undirected Graph Constraints

The online processing modelled with an undirected graph exhibits a deadlock-free online processing that terminates in a finite number of steps through two stage local rules. In addition, scalability in the number of variables is provided if Assumptions 1 and 5 hold. The corresponding strategy space together with the Bayesian framework enables the utilization of the communication constrained problem definition (P1).

Considering the spaces of strategies that are comprised of two stage local rules over an undirected graph, finding the global solution to the constrained problem (P1) exhibit similar diffi-

culties as we have indicated in Chapter 3. Rather than the global best strategy, we consider a solution that satisfies the equilibrium conditions given in Section2.7.1.2.

The information structure of the directed case yield the conditions given by Proposition 3.3.1 provided that Assumption 1 holds which specializes to Proposition 3.3.2 if additionally Assumptions 1-4 are satisfied. On the other hand, considering decentralized strategies constrained by an undirected graph, Proposition 3.3.1 applies to the unwrapped directed counterpart under Assumption 1 and 5 [28]. Provided that Assumption 2 and 3 hold in addition, Proposition 3.3.2 applies to the unwrapped directed counterpart and the following holds.

Proposition 4.2.1 *Under Assumptions 1-3 and 5, $J(\gamma) = J_d(\gamma) + \lambda J_c(\gamma)$ holds and given a person by person optimal strategy $\gamma^* = (\gamma_1^*, ...\gamma_N^*)$ constituted of two stage local rules over an undirected graph and fixing all local rules other than the j^{th}, the j^{th} optimal rule reduces to local stage one communication rule given by*

$$\mu_j^*(y_j) = \arg\min_{u_j \in \mathcal{U}_j} \int_{\mathcal{X}_j} \mathrm{d}x_j p(x_j) p(y_j|x_j) [\lambda c_j^c(u_j, x_j) + C_j^*(u_j, x_j)] \tag{4.4}$$

for all $y_j \in \mathcal{Y}_j$ with nonzero probability and stage two estimation rule given by

$$v_j^*(y_j, u_{ne(j)}) = \arg\min_{\hat{x}_j \in \mathcal{X}_j} \int_{\mathcal{X}_j} \mathrm{d}x_j p(x_j) p(Y_j|x_j) P_j^*(u_{ne(j)}|x_j) c_j^d(\hat{x}_j, x_j) \tag{4.5}$$

for all $y_j \in \mathcal{Y}_j$ and for all $u_{ne(j)} \in \mathcal{U}_{ne(j)}$ with nonzero probability.

The incoming message likelihood is given by

$$P_j^*(u_{ne(j)}|x_j) = \int_{\mathcal{X}_{ne(j)}} \mathrm{d}x_{ne(j)} p(x_{ne(j)}|x_j) \prod_{i \in ne(j)} P_{i\to j}^*(u_{i\to j}|x_i) \tag{4.6}$$

with terms regarding influence of $i \in ne(j)$ on j given by

$$P_{i\to j}^*(u_{i\to j}|x_i) = \sum_{u_i \setminus u_{i\to j}} p(u_i|x_i; \mu_i^*) \tag{4.7}$$

for all $u_{i\to j} \in \mathcal{U}_{i\to j}$ where

$$p(u_i|x_i; \mu_i^*) = \int_{\mathcal{Y}_i} \mathrm{d}y_i p(y_i|x_i) p(u_i|y_i; \mu_i^*) \tag{4.8}$$

In addition for all $u_j \in \mathcal{U}_j$

$$C_j^*(u_j, x_j) = \sum_{i \in ne(j)} C_{i\to j}^*(u_{j\to i}, x_j) \tag{4.9}$$

holds with terms regarding the influence of j on $i \in ne(j)$ given by

$$C^*_{i\to j}(u_{j\to i}, x_j) = \int_{\mathcal{X}_{ne(i)\setminus j}} dx_{ne(i)\setminus j} \int_{\mathcal{X}_i} dx_i p(x_{ne(i)\setminus j}, x_i|x_j) \sum_{u_{ne(i)\setminus j}} \prod_{j'\in ne(i)\setminus j} P^*_{j'\to i}(u_{j'\to i}|x_{j'}) I^*_i(u_{ne(i)}, x_i; \gamma^*_i) \quad (4.10)$$

such that

$$I^*_i(u_{ne(i)}, x_i; \nu^*_i) = \int_{\mathcal{Y}_i} \mathrm{d}y_i \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i \, c^d_i(\hat{x}_i, x_i) p(\hat{x}_i|y_i, u_{ne(i)}; \nu^*_i) p(y_i|x_i) \quad (4.11)$$

Proof. Apply Proposition 3.3.2 on the unwrapped directed couterpart of the undirected graph $\mathcal{G}$ together with the two stage local rules and under Assumption 5. ■

The proposition above is an adoptation of Proposition 2.7.3 which considers decentralized detection networks constrained by undirected graphs. Given a person-by-person optimal strategy, we obtain communication and estimation rules local to node j in a variational form, based on the rules local to the remaining nodes. Considering $P^*_{i\to j}(u_{i\to j}|x_i)$ for $i \in ne(j)$, $P^*_j(u_{ne(j)}|x_j)$ is the likelihood of x_j given $u_{ne(i)}$. Eq.s(4.9)-(4.11) reveal that $C^*_j(u_j, x_j)$ is the total expected cost induced on the neighbors by u_j, i.e. $E\{c^d(\hat{x}_{ne(j)}, x_{ne(j)}) + \lambda c^c(u_{ne(j),x_{ne(j)}})|u_j, x_j\}$. Hence, we conclude that the j^{th} optimal communication rule selects the message that results with a minimum contribution to the overall cost and also noting that $p(x_j)p(y_j|x_j)P(u_{ne(j)}|x_j) \propto p(x_j|y_j, u_{ne(j)})$ holds under Assumptions 1-3 and 5, the optimal estimation rule selects $\hat{x}_j$ that yields minimum expected penalty given y_j and $u_{ne(j)}$.

Similar to the specification of Algorithm 3 by asserting Proposition 2.7.3 in Algorithm 1, it is possible to treat the right hand sides of Eq.s(4.6)-(4.11) as operators valid for any given set of local rules $\gamma_{\setminus j} \in \Gamma^{\mathcal{G}}_{\setminus j}$, not necessarily optimal and employ them in Algorithm 3. Doing that, it is possible to employ the iterative scheme that starting with an initial decentralized strategy comprised of two stage local rules over an undirected graph converges to a person by person optimal one in the context of decentralized estimation.

Finally, the objective value at the l^{th} step is easily found to be

$$J(\gamma^l) = \sum_{i\in\mathcal{V}} G^d_i(\nu^l_i) + \lambda \sum_{i\in\mathcal{V}} G^c_i(\mu^l_i) \quad (4.12)$$

where

$$G^d_i(\nu^l_i) = \sum_{u_{ne(i)}} \int_{\mathcal{X}_i} \mathrm{d}x_i p(x_i) P^{l+1}_i(u_{ne(i)}|x_i) I_i(u_{ne(i)}, x_i; \nu^l_i) \quad (4.13)$$

and

$$G_i^c(\mu_i^l) = \sum_{u_i} \int_{\mathcal{X}_i} \mathrm{d}x_i c_i^c(u_i, x_i) p(x_i) p(u_i|x_i; \mu_i^l) \tag{4.14}$$

in terms of the expressions discussed above.

Adopting the information structure proposed for decentralized detection networks constrained by undirected topologies for estimation, we have extended the canonical decentralized estimation schemes in a number of aspects. First of all, the framework we utilize provides results for any undirected graph, including the loopy ones, and hence extends the canonical star-topology setting. A collaborative processing is provided through associating random variable with platforms, which can be done in an arbitrary fashion, in principle. In addition, the online processing scheme we consider better matches the ad-hoc nature of the underlying network with bandwidth limited channels. Moreover, a penalty term due to the transmissions is introduced into the problem setting which reveals the tradeoff between the estimation accuracy and cost of communications.

4.3 Monte Carlo Approximations for Undirected Communication Constraints

In the detection setting, the set $\mathcal{X}$ from which the random variable of concern takes values from is of finite cardinality and there exists a mapping between the space of (person-by-person optimal) local rules and a finite dimensional Euclidean space. One appeal of Algorithm 3 in this context is that it is often the case that the steps of the algorithm involves no numerical approximations[1]. However, unlike the detection case, the person by person optimal rules given by Eqs.(4.4) and (4.5) do not yield any finite parameterization. In addition, it is not possible to carry out the steps of Algorithm 3 involving operators implied by Proposition 4.2.1 exactly, in general. We similarly resort particle representations and approximate computational schemes through Monte Carlo methods regarding Algorithm 3 (Section 2.7.2.2) in view of Proposition 4.2.1.

We proceed with progressive approximations and achieve an iterative procedure that requires sample sets generated only from the marginal distributions, i.e.

$$S_{x_j} \triangleq \{x_j^{(1)}, x_j^{(2)}, ..., x_j^{(M)}\} \text{ such that } x_j^{(m)} \sim p(x_j) \text{ for } m = 1, 2, ..., M_j \tag{4.15}$$

[1] An example in which some approximations would be useful arise when the observation likelihoods are non-parametric distributions such that $p(u_j|x_j; \mu_j) = \int_{\mu_j^{-1}(u_j)} \mathrm{d}y_j p(y_j|x_j)$ can not be explicitly found.

and

$$S_{y_j} \triangleq \{y_j^{(1)}, y_j^{(2)}, ..., y_j^{(P)}\} \text{ such that } y_j^{(p)} \sim p(y_j) \text{ for } p = 1, 2, ..., P_j \tag{4.16}$$

for $j \in \mathcal{V}$. We assume that $M_j = M$ and $P_j = P$ for all $j \in$ for brevity, altough this condition is not required, in general.

Each node $j \in \mathcal{V}$ starts with the knowledge of $p(x_{ne(j)}, x_j)$ and $p(y_j|x_j)$ together with an initial local rule $\gamma_j^0 \in \Gamma_j^{\mathcal{G}}$. In the resulting network, the platforms perform computations which correspond to approximations to person-by-person optimal rules preserving the scalability of online processing in the dimensionality of X. In addition, the resulting Monte Carlo optimization scheme scales with both the sample sizes and the number of nodes. The message passing structure in the `Update` step of Algorithm 3 together with the amenability for network self-organization is also preserved.

4.3.1 Approximating the Person-by-Person Optimal Local Rules

Consider the sufficent condition of person-by-person optimality condition for the j^{th} local rule which is given in Proposition 4.2.1. Suppose that all the integrands are known over their domains, i.e. for all $x_j \in \mathcal{X}_j$, $C_j^*(u_j, x_j)$ and $P_j^*(u_{ne(j)}|x_j)$ are known for all $u_j \in \mathcal{U}_j$ and $u_{ne(j)} \in \mathcal{U}_{ne(j)}$ respectively. It is not possible to evaluate the local rules exactly for general C_j^* and P_j^*.

We apply the conventional Monte Carlo method to the integrals appearing in the variational form of the communication and estimation rules in Eq.s (4.4) and (4.5) respectively. Given the usual sample set S_{x_j} as defined in Eq.(4.15) and considering Eq.(4.4) the Monte Carlo method yields

$$\tilde{\mu}_j^*(y_j) = \arg\min_{u_j \in \mathcal{U}_j} \sum_{x_j \in S_{x_j}} p(y_j|x_j)[\lambda c_j^c(u_j, x_j) + C_j^*(u_j, x_j)] \tag{4.17}$$

as an approximation to the stage one communication rule such that for all $y_j \in \mathcal{Y}_j$ with non-zero probability, i.e. $\tilde{\mu}_j^*(y_j) \approx \mu_j^*(y_j)$.

Considering the local estimation rule given by Eq.(4.5) the Monte Carlo method yields

$$\tilde{\nu}_j^*(y_j, u_{ne(j)}) = \arg\min_{\hat{x}_j \in \mathcal{X}_j} \sum_{x_j \in S_{x_j}} p(y_j|x_j) P_j^*(u_{ne(j)}|x_j) c_j^d(\hat{x}_j, x_j) \tag{4.18}$$

for all $y_j \in \mathcal{Y}_j$ and $u_{ne(j)} \in \mathcal{U}_{ne(j)}$ with non-zero probability such that $\tilde{\nu}_j^*(y_j, u_{ne(j)}) \approx \nu_j^*(y_j, u_{ne(j)})$.

It is reasonable to assume that we are able to evaluate at $x_j^{(m)} \in S_{x_j}$; $p(y_j|x_j^{(m)})$ for all $y_j \in \mathcal{Y}_j$, $c_j^c(u_j, x_j^{(m)})$ for all $u_j \in \mathcal{U}_j$ and $c_j^d(\hat{x}_j, x_j^{(m)})$ for all $\hat{x}_j \in \mathcal{X}_j$. The requirement regarding the incoming message likelihood and conditional expected cost is reduced to evaluating $P_j^*(u_{ne(j)}|x_j^{(m)})$ and $C_j^*(u_j, x_j^{(m)})$ for all $x_j^{(m)} \in S_{x_j}$ and for $u_j \in \mathcal{U}_j$ and $u_{ne(j)} \in \mathcal{U}_{ne(j)}$ respectively. Provided that they are known for the required values of their arguments, Eq.s(4.17) and (4.18) are 1-step approximations to the communication and estimation rules local to node j which can be denoted by $\tilde{\mu}_j^{*1}(y_j) \approx \mu_j^*(y_j)$ and $\tilde{\nu}_j^{*1}(y_j, u_{ne(j)}) \approx \nu_j^*(y_j, u_{ne(j)})$ respectively.

Example 4.3.1 *Consider the squared error penalty for the estimation error, i.e.* $c_j^d(\hat{x}_j, x_j) = (\hat{x}_j - x_j)^2$. *Then the approximate person-by-person optimal estimation rule yields a closed form solution, i.e.*

$$\tilde{\nu}_j^{*1}(y_j, u_{ne(j)}) = \frac{\sum_{m=1}^{M} x_j^{(m)} p(y_j|x_j^{(m)}) P_j^*(u_{ne(j)}|x_j^{(m)})}{\sum_{m=1}^{M} p(y_j|x_j^{(m)}) P_j^*(u_{ne(j)}|x_j^{(m)})}$$

4.3.2 Approximating the Incoming Message Likelihood and the Conditional Expected Cost

The 1-step approximations to the communication and estimation rules local to node j requires the evaluation of $C_j^*(u_j, x_j)$ and $P_j^*(u_{ne(j)}|x_j)$ at a finite discretization of their domains. However, the right hand sides of Eq.s(4.9) and (4.6) render it not possible, in general (consider, for example, arbitrary choices of $\gamma_{\setminus j} \in \Gamma_j^{\mathcal{G}}$). For convenience, suppose that $i \in ne(j)$, $C_{i\to j}^*(u_{j\to i}, x_j)$ is known at for all $u_{j\to i} \in \mathcal{U}_{j\to i}$ and $x_j = x_j^{(m)}$ where $x_j^{(m)} \in S_{x_j}$. Then using Eq.(4.9), C_j^* can be exactly evaluated at the required points.

However, the integral in the right hand side of Eq.(4.6) still necessitates an approximation for the evaluation of P_j^* at a single point of its domain. Suppose that for all $i \in ne(j)$, $P_{i\to j}^*(u_{i\to j}, x_i)_{m=1}^M$ is known for all $u_{i\to j} \in \mathcal{U}_{i\to j}$ and at $x_i = x_i^{(m)}$ where $x_i^{(m)} \in S_{x_i}$. Let us costruct a new sample set whose m^{th} element is the vector obtained by concatenating m^{th} elements from S_{x_i} for all $i \in ne(j)$, i.e. $S_{ne(j)} \triangleq \{x_{ne(j)}^{(m)} | x_{ne(j)}^{(m)} = (x_i^{(m)})_{i\in ne(j)}\}$. Note that $S_{ne(j)}$ is constructed by the usual sample sets of the neighboring nodes and $x_{ne(j)}^{(m)} \sim \prod_{i\in ne(j)} p(x_i)$. Then an Importance Sampling approximation to $P_j^*(u_{ne(j)}|x_j^{(m)})$ is possible utilizing the importance

sampling density $\prod_{i\in ne(j)} p(x_i)$ with the importance weights

$$\omega_j^{(m)(m')} = \frac{p(x_{ne(j)}^{(m')}|x_j^{(m)})}{\prod\limits_{i\in ne(j)} p(x_i^{(m')})}$$

as

$$\tilde{P}_j^*(u_{ne(j)}|x_j^{(m)}) = \frac{1}{\sum\limits_{m'=1}^{M} \omega_j^{(m)(m')}} \sum_{m'=1}^{M} \omega_j^{(m)(m')} \prod_{i\in ne(j)} P_{i\to j}^*(u_{i\to j}|x_i^{(m')}) \tag{4.19}$$

such that $\tilde{P}_j^*(u_{ne(j)}|x_j^{(m)}) \approx P_j^*(u_{ne(j)}|x_j^{(m)})$ for all $u_{ne(j)} \in \mathcal{U}_{ne(j)}$ and for all $x_j^{(m)} \in S_{x_j}$.

After replacing P_j^* with $\tilde{P}_j^*$ in Eq.(4.18), we obtain a further step approximation (2-steps approximation) for the person-by-person optimal estimation rule local to node j, i.e.

$$\tilde{\nu}_j^{*2}(y_j, u_{ne(j)}) \approx \tilde{\nu}_j^{*1}(y_j, u_{ne(j)}) \approx \nu_j^*(y_j, u_{ne(j)})$$

for all $y_j \in \mathcal{Y}_j$ and $u_{ne(j)} \in \mathcal{U}_{ne(j)}$ with non-zero probability provided that the node-to-node likelihood terms $P_{i\to j}^*$ for all $i \in ne(j)$ are known at the required points of its domain.

The assumption that the node-to-node conditional cost terms $C_{i\to j}^*$ are known at the required points yield exact evaluation of C_j^* and introduce no further approximation for the communication rule local to node j at this step.

4.3.3 Approximating the Node-to-Node Terms

Two-steps approximated communication and estimation rules local to node j are based on the assumption that the node-to-node likelihood term $P_{i\to j}^*(u_{i\to j}, x_i)$ for all neighbor nodes $i \in ne(j)$ are known for all $u_{i\to j} \in \mathcal{U}_{i\to j}$ and at the points from the usual sample sets, i.e. $x_i = x_i^{(m)}$ where $x_i^{(m)} \in S_{x_i}$, together with the node-to-node conditional cost term $C_{i\to j}^*(u_{j\to i}, x_j)$ is known at for all $u_{j\to i} \in \mathcal{U}_{j\to i}$ and at $x_j = x_j^{(m)}$ where $x_j^{(m)} \in S_{x_j}$. A further step is possible through replacing the node-to-note terms with their approximate values.

Consider Eq.(4.7) and note that $P_{i\to j}^*$ is a marginalization of $p(u_i|x_i;\mu_i^*)$ given by Eq.(4.8). Proposition 4.2.1 assumes that all the rules local to nodes other than j are known and fixed at the optimum including that $\mu_i = \mu_i^*$ for all $i \in ne(j)$. Also considering Eq.(4.2), it is possible to apply the method of Importance Sampling for the evaluation of $p(u_i|x_i;\mu_i^*)$ for all $u_i \in \mathcal{U}_i$ and at $x_i = x_i^{(m)}$ for $x_i^{(m)} \in S_{x_i}$ through the imortance sampling density $p(y_i)$ and utilizing the

sample set S_{y_i} together with the importance weights given by

$$\omega_i^{(m)(p)} = \frac{p(y_i^{(p)}|x_i^{(m)})}{p(y_i^{(p)})}$$

and obtain

$$\tilde{p}(u_i|x_i^{(m)};\mu_i^*) = \frac{1}{\sum_{p=1}^{P}\omega_i^{(m)(p)}}\sum_{p=1}^{P}\omega_i^{(m)(p)}\delta_{u_i,\mu_i^*(y_i^{(p)})}$$

for all $u_i \in \mathcal{U}_i$ and for all $x_i^{(m)} \in S_{x_i}$. In other words $\tilde{p}(u_i|x_i^{(m)};\mu_i^*) \approx p(u_i|x_i^{(m)};\mu_i^*)$ and after replacing the later with the former in Eq.s(4.7) we achieve $\tilde{P}^*_{i\to j}(u_{i\to j}|x_i^{(m)}) \approx P^*_{i\to j}(u_{i\to j}|x_i^{(m)})$. Similarly, replacing the later with the former in Eq.(4.19) a futher approximation is obtained to the 2-steps approximated rule. Let us denote the 3-steps approximated estimation rule by $\tilde{\nu}_j^{*3}(y_j, u_{ne(j)})$, then $\tilde{\nu}_j^{*3}(y_j, u_{ne(j)}) \approx \tilde{\nu}_j^{*2}(y_j, u_{ne(j)}) \approx \tilde{\nu}_j^{*1}(y_j, u_{ne(j)}) \approx \nu_j^*(y_j, u_{ne(j)})$ holds.

Next, we consider evaluating the remaining node-to-node term at the required points of its domain utilizing Eq.s(4.10) and (4.11). Consider $C^*_{i\to j}$ and suppose that for $i \in ne(j)$ $I_i^*(u_{ne(i)}, x_i; \nu_i^*)$ is known for all $u_{ne(i)} \in \mathcal{U}_{ne(i)}$ and $x_i \in \mathcal{X}_i$ (Proposition 4.2.1 states that $\gamma^*_{\setminus j}$ is determined). We also assume that for all $j' \in ne(i)\setminus j$, $P^*_{j'\to i}(u_{j'\to i}|x_{j'})$ is known for all $u_{j'\to i} \in \mathcal{U}_{j'\to i}$ and $x_{j'} \in \mathcal{X}_{j'}$. However, the right hand side of Eq.s(4.10) still does not yield a solution that can be practically carried out in general and we resort to Monte Carlo methods in order to approximately evaluate $C^*_{i\to j}$ at the required points of its domain.

Let us construct a new sample set by concatenating the m^{th} samples of the usual sample sets of neighbors of i other than j, i.e. $S_{x_{j'}}$ for $j' \in ne(i)\setminus j$ given by

$$S_{x_{ne(i)\setminus j}} \triangleq \{x^{(m)}_{ne(i)\setminus j} | x^{(m)}_{ne(i)\setminus j} = (x^{(m)}_{j'})_{j'\in ne(i)\setminus j}\}$$

We apply the same procedure with S_{x_i} and $S_{x_{ne(i)\setminus j}}$ yielding $S_{x_{i\cup ne(i)\setminus j}} = \{x^{(m)}_{i\cup ne(i)\setminus j}\}$ and observe that $x^{(m)}_{i\cup ne(i)\setminus j} \sim p(x_i)\prod_{j'\in ne(i)\setminus j} p(x_{j'})$ for all $x^{(m)}_{i\cup ne(i)\setminus j} \in S_{x_{i\cup ne(i)\setminus j}}$. Then, it is possible to utilize this sample set for an Importance Sampling approximation implying the importance density $p(x_i)\prod_{j'\in ne(i)\setminus j} p(x_{j'})$ together with the importance weights

$$\omega_i^{(m)(m')} = \frac{p(x^{(m')}_{ne(i)\setminus j}, x_i^{(m')}|x_j^{(m)})}{p(x_i^{(m')})\prod\limits_{j'\in ne(i)\setminus j} p(x^{(m')}_{j'})}$$

and obtain

$$\tilde{C}^*_{i\to j}(u_{j\to i}, x_j^{(m)}) = \sum_{u_{ne(i)\setminus j}} \frac{1}{\sum_{m'=1}^{M}\omega_i^{(m)(m')}}\sum_{m'=1}^{M}\omega_i^{(m)(m')}\times \prod_{j'\in ne(i)\setminus j} P^*_{j'\to i}(u_{j'\to i}|x^{(m')}_{j'})I_i^*(u_{ne(i)}, x_i^{(m')}; \nu_i^*) \quad (4.20)$$

Replacing $C^*_{i\to j}$ with $\tilde{C}^*_{i\to j}$ in the one step approximated local communication rule local to node j given by Eq.(4.17), we obtain $\tilde{\mu}^{*2}_j$ such that $\tilde{\mu}^{*2}_j(y_j) \approx \tilde{\mu}^{*1}_j(y_j)$ for all $y_j \in \mathcal{Y}_j$ with non-zero probability.

We have proposed an approximation for node-to-node likelihood terms. Finally we handle the evaluation of $I^*_i(u_{ne(i)}, x_i; \nu^*_i)$ at all $u_{ne(i)} \in \mathcal{U}_{ne(i)}$ and $x_i = x^{(m)}_i$ for all $x^{(m)}_i \in S_{x_i}$ required in Eq.(4.20). Note that substituting Eq.(4.3) in Eq.(4.11) yields

$$I^*_i(u_{ne(i)}, x_i; \nu^*_i) = \int_{\mathcal{Y}_i} \mathrm{d}y_i c^d_i(\nu^*_i(y_i, u_{ne(i)}), x_i) p(y_i|x_i)$$

for which the utilization of the sample set S_{y_i} implies an Importance Sampling approximation using the importance density $p(y_i)$ together with the importance weights

$$\omega^{(m)(p)}_i = \frac{p(y^{(p)}_i|x^{(m)}_i)}{p(y^{(p)}_i)}$$

given by

$$\tilde{I}^*_i(u_{ne(i)}, x^{(m)}_i; \nu^*_i) = \frac{1}{\sum_{p=1}^{P} \omega^{(m)(p)}_i} \sum_{p=1}^{P} \omega^{(m)(p)}_i c^d_i(\nu^*_i(y^{(p)}_i, u_{ne(i)}), x^{(m)}_i)$$

for all $u_{ne(i)} \in \mathcal{U}_{ne(i)}$ and $x^{(m)}_i \in S_{x_i}$ such that $\tilde{I}^*_i(u_{ne(i)}, x^{(m)}_i; \nu^*_i) \approx I^*_i(u_{ne(i)}, x^{(m)}_i; \nu^*_i)$. Replacing I^*_i with $\tilde{I}^*_i$ in Eq.(4.20) and Eq.(4.17), we obtain $\tilde{\mu}^{*3}_j$ such that $\tilde{\mu}^{*3}_j(y_j) \approx \tilde{\mu}^{*2}_j(y_j) \approx \tilde{\mu}^{*1}_j(y_j)$ for all $y_j \in \mathcal{Y}_j$ with non-zero probability.

4.3.4 Monte Carlo Optimization of Decentralized Estimation Networks Under Undirected Topologies

We have employed an information structure due to a two stage strategy over an undirected graph and, similar to the discussion in Chapter 3 for decentralized estimation networks constrained by directed acyclic graphs, provided a Monte Carlo framework presented in Section 4.3.1-4.3.3. Given a person-by-person optimal strategy $\gamma^* \in \Gamma^{\mathcal{G}}$ and keeping all the local rules other than the fixed at the optimal points, i.e. $\gamma_{\setminus j} = \gamma^*_{\setminus j}$, the proposed framework outputs an approximation to the j^{th} person-by-person optimal rule (regarding Proposition 4.2.1), i.e. $\tilde{\gamma}^*_j = (\tilde{\mu}^{*3}_j(y_j), \tilde{\nu}^{*3}_j(y_j, u_{ne(j)}))$ such that $\tilde{\mu}^{*3}_j(y_j) \approx \mu^*_j(y_j)$ and $\tilde{\nu}^{*3}_j(y_j, u_{ne(j)}) \approx \nu^*_j(y_j, u_{ne(j)})$ for all $y_j \in \mathcal{Y}_j$ and $u_{ne(j)} \in \mathcal{U}_{ne(j)}$ with nonzero probability.

It is possible to utilize the approximations for all local rules, i.e. γ^*_j for all $j \in \mathcal{V}$, and the node-to-node terms would require the usual sample sets utilized for 1-step approximation to

the local rules. In addition, the particle representations and approximate computations are valid for any set of two-stage local rules over an undirected graph, including those in an "approximating" form. Let us summarize the Monte Carlo framework with

$$\begin{aligned}
\tilde{\alpha}_j(S_{x_j}) &= \tilde{r}_j(\tilde{C}_{ne(j)\to j}) \\
\tilde{\beta}_j(S_{x_j}, \hat{x}_j) &= \tilde{q}_j(\tilde{P}_j(S_{x_j})) \\
\tilde{P}_j(S_{x_j}) &= \tilde{f}_j(\tilde{P}_{ne(j)\to j}) \\
\tilde{P}_{j\to ne(j)} &= \tilde{g}_j(\tilde{\alpha}_j(S_{x_j}) \\
\tilde{C}_{j\to ne(j)} &= \tilde{h}_j(\tilde{\beta}_j(S_{x_j}, \hat{x}_j), \tilde{P}_{ne(j)\to j})
\end{aligned}$$

where

$$\tilde{\alpha}_j(S_{x_j}) = \{(\lambda c_j^c(u_j, x_j) + \sum_{i\in ne(j)} \tilde{C}_{i\to j}(u_{j\to i}, x_j), u_j, x_j) | u_j \in \mathcal{U}_j, x_j \in S_{x_j}\}$$

$$\tilde{C}_{ne(j)\to j} = \{\tilde{C}_{i\to j}(S_{x_j}) | i \in ne(j)\}$$

$$\tilde{C}_{i\to j}(S_{x_j}) = \{(\tilde{C}_{i\to j}(u_{j\to i}, x_j), u_{j\to i}, x_j) | u_{j\to i} \in \mathcal{U}_{j\to i}, x_j \in S_{x_j}\}$$

$$\tilde{P}_j(S_{x_j}) = \{(\tilde{P}_j(u_{ne(j)}, x_j, u_{ne(j)}, x_j) | u_{ne(j)} \in \mathcal{U}_{ne(j)}, x_j \in S_{x_j}\}$$

$$\tilde{P}_{ne(j)\to j} = \{\tilde{P}_{i\to j}(S_{x_i}) | i \in j\}$$

$$\tilde{P}_{i\to j}(S_{x_i}) = \{(P_{i\to j}(u_{i\to j}, x_i), u_{i\to j}, x_i) | u_{i\to j} \in \mathcal{U}_{i\to j}, x_i \in \mathcal{X}_i\}$$

$$\tilde{P}_{j\to ne(j)} = \{\tilde{P}_{j\to i}(S_{x_j}) | i \in ne(j)\}$$

$$\tilde{C}_{j\to ne(j)} = \{\tilde{C}_{j\to i}(S_{x_i}) | i \in ne(j)\}$$

The Monte Carlo optimization scheme which is obtained through employing the framework in the `Update` step of Algorithm 3 is given by Algorithm 5. Finally, the objective value corresponding a strategy $\gamma \in \Gamma^{\mathcal{G}}$, i.e. $J(\gamma) = J_d(\gamma) + \lambda J_c(\gamma)$ given by Eq.s(4.12)-(4.14), can be computed approximately by

$$\tilde{J}(\tilde{\gamma}^l) = \sum_{i\in\mathcal{V}} \tilde{G}_i^d(\tilde{\nu}_i^l) + \lambda \sum_{i\in\mathcal{V}} \tilde{G}_i^c(\tilde{\mu}_i^l) \tag{4.21}$$

where

$$\tilde{G}_i^d(\tilde{\nu}_i^l) = \sum_{u_{ne(i)},m} \tilde{P}_i^{l+1}(u_{ne(i)}|x_i^{(m)}) \tilde{l}_i^l(u_{ne(i)}, x_i^{(m)}; \tilde{\nu}_i^l) \tag{4.22}$$

and

$$\tilde{G}_i^c(\tilde{\mu}_i^l) = \sum_{u_i,m} c_i^c(u_i, x_i^{(m)}) p(u_i|x_i^{(m)}; \tilde{\mu}_i^l) \tag{4.23}$$

Algorithm 5 Iterations converging to an approximate person-by-person optimal decentralized two-stage strategy for estimation over an undirected graph $\mathcal{G}$.

0) (Initiate) $l = 0$;

Choose $\gamma^0 = (\gamma_1^0, \gamma_2^0, ..., \gamma_N^0)$ such that $\gamma_j^0 \in \Gamma_j^{\mathcal{G}}$ for $j = 1, 2, ..., N$

1) (Update) $l = l + 1$;

i) for $j = 1, 2, ..., N$

$\tilde{P}^l_{j\to ne(j)} = \tilde{g}_j(\tilde{\alpha}_j^{l-1})$

ii) for $j = 1, 2, ..., N$

$\tilde{P}^l_j(S_{x_j}) = f_j(\tilde{P}^l_{ne(j)\to j})$

$\tilde{\beta}^l_j = \tilde{q}_j(P^l_j)$ % Update the stage-two rule

$\tilde{C}^l_{j\to ne(j)} = \tilde{h}_j(\tilde{\beta}_j, \tilde{P}^l_{ne(j)\to j})$

iii) for $j = 1, 2, ..., N$

$\tilde{\alpha}^l_j = \tilde{r}^l_j(\tilde{C}_{ne(j)\to j})$ % Update the stage-one rule

2) (Check) If $\tilde{J}(\tilde{\gamma}^{l-1}) - \tilde{J}(\tilde{\gamma}^l) < \varepsilon$ STOP, else GO TO 1;

Similar to the discussion presented in Section 3.4.4 for the directed case, in contrary to $\{J(\gamma^l)\}$, the sequence of approximated objectives, i.e. $\{\tilde{J}(\tilde{\gamma}^l)\}$, is not necessarily non-increasing and considering the error sequence $err[l] = J(\gamma^l) - \tilde{J}(\tilde{\gamma}^l)$, a more robust termination condition for the Check step of Algorithm 5, compared to $\tilde{J}(\tilde{\gamma}^{l-1}) - \tilde{J}(\tilde{\gamma}^l) < \varepsilon$ can be proposed. Further investigation of this issue is left out of the scope of the thesis.

4.4 Examples

4.4.1 A Gaussian Example

Consider a decentralized estimation network that performs a two stage strategy over the undirected graph $\mathcal{G} = (\{1, 2, 3, 4\}, \{(1, 3), (3, 1), (2, 3), (3, 2), (3, 4), (4, 3)\})$ (Figure 4.1) in which each edge represents a bidirectional communication link. The set of admissible symbols is given by $\mathcal{U}_{i\to j} = \{0, 1, 2\}$ for all $(i, j) \in \mathcal{E}$.

Similar to the example scenario presented in Section 3.5.1 the estimation task considers the Gaussian random field $X = \{X_1, X_2, X_3, X_4\}$ Markov with respect to the graph presented in Figure 3.1b and of identically zero mean vector. The covariance matrix is assumed to be that

given by Eq.(3.53). X_j is associated with the j^{th} platform. Contrary to the directed case, the online processing starts with each node evaluating its stage-one communication function on its measurement, i.e. $u_{1\to3} = \mu_1(y_1)$, $u_{2\to3} = \mu_2(y_2)$, $(u_{3\to1}, u_{3\to2}, u_{3\to4}) = \mu_3(y_3)$ and $u_{4\to3} = \mu_4(y_4)$ simultaneously in a parallel fashion. As soon as all the messages from the neigbors (or lack thereof) are received, stage two estimation rules are evaluated as $\hat{x}_1 = \nu_1(y_1, u_{3\to1})$,$\hat{x}_2 = \nu_2(y_2, u_{2\to3})$, $\hat{x}_3 = \nu_3(y_3, u_{1\to3}, u_{2\to3}, u_{4\to3})$ and $\hat{x}_4 = \nu_4(y_4, u_{3\to4})$. Subject to design is the strategy $\gamma = (\gamma_1, ..., \gamma_4)$ where $\gamma_j = (\mu_j, \nu_j)$.

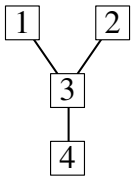


Illustration of $\mathcal{G}$.

Figure 4.1: Illustration of the undirected graph that the decentralized estimation strategy renders in the example scenario considered in Section 4.4.1.

The cost functions are selected as in Section 3.5.1 such that J_c is the total expected link use rate in bits and the estimation error penalty implies that J_d is the total mean squared error. Similarly, the noise processes $\{n_j\}_{j\in\mathcal{V}}$ are additive, mutually independent and zero mean Gaussian, i.e. $n_j \sim \mathcal{N}(0, \sigma_n^2)$ where $\sigma_n^2 = 0.5$, yielding an SNR of 6dB for each sensor.

For each platform j, the initial local estimation rule is the myopic mimimum MSE estimator which is based only on y_j, i.e. $\nu_j^0(y_j, u_{ne(j)}) = \int_{-\infty}^{\infty} \mathrm{d}x_j\, x_j p(x_j|y_j)$, and the communication rule is a threshold rule quantizing y_j as used in Section 3.5.1.

Considering Algorithm 3 presented in Section 2.7.2.2, the performance points (J_c, J_d) of the converged strategies vary with λ. For $\lambda \geq \lambda^*$, the strategy of no transmissions with myopic estimation rules achieves the minimum cost which is also a person-by-person optimal. Hence, λ^* admits the interpretation of being the maximum price per bit that the system affords to decrease the estimation penalty. We approximate the performance curve of solutions as we increase λ from 0 which is an approximate quantification for the tradeoff between the cost of estimation errors and communication.

In Figure 4.2 we present these pairs, i.e. $(\tilde{J}_c, \tilde{J}_d)$, for different choices of λ and $\left|\mathcal{U}_{i\to j}\right|$s. The upper and lower bounds are mean squared errors (MSEs) corresponding to the myopic rule

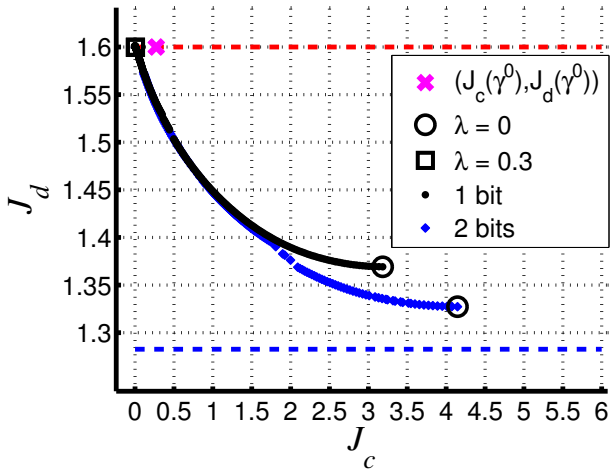


Figure 4.2: Approximated performance points corresponding to the two-stage decentralized estimation strategies output by Algorithm 5, over the undirected graph $\mathcal{G}$ in Figure 4.1 while λ is increased from 0 with steps of 0.001. The Gaussian scenario enables the exact computation of the mean squared error achieved by the conventional centralized and myopic MMSE estimators rendering the lower bound (red dashed line) and the upper bound (blue dashed line) respectively, for the person-by-person optimal strategies. The black and blue dots are the approximated performance sequences for the 1 bit and 2 bits selective communication schemes respectively.

and the centralized optimal rule[2]respectively. $(\tilde{J}_c, \tilde{J}_d)$ points for the 1-bit selective communication scheme reveal that altough the transmission has no cost for $\lambda = 0$, the total link use rate is only slightly higher than 50% of the total 6 bits indicating that the information from receiving no messages is successfully utilized. Moreover, the MSE performance is closer to that of the centralized scheme than the myopic scheme. The communication stops for $\lambda^* \approx 0.3$. Approximate performance points for 2-bits case present the decrease in MSE for the same network load as we increase the link capacities for small values of λ which is competetive with that of the centralized rule.

Comparing the approximated performance points of the directed and undirected strategies given in Figure 3.2 and 4.2 respectively for 1 bit and 2 bits selective communication schemes, we observe the benefits of bi-directional communications employed by the strategy over the undirected graph. For the directed case, nodes 1 and 2 are parentless and hence do not have means to exploit contributions from other platforms. Specifically all parentless nodes apply the initial rule, which has been selected as the myopic estimator. Therefore the nodes with

[2] For $c(x, \hat{x}) = (x - \hat{x})^T(x - \hat{x})$, the optimal centralized estimate is the mean vector of $p(x_1, ..., x_4|y_1, ..., y_4)$ which yields a minimum of $J_c = 3Q$ bits where Q is the number of bits used to represent y_j before transmission.

more ancestors are more likely to benefit the contribution of other nodes whereas for the undirected case, the nodes with more neighbors pose advantegous. The price paid is that the information horizon is limited with the observation of the neighbors whereas the local rules depend on a two-hop neighborhood due to the two stage mechanism necessary for causality.

4.4.2 A Heavy Tailed Example

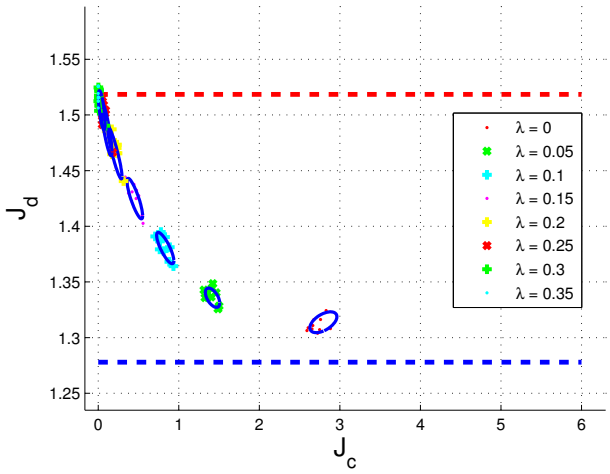


Figure 4.3: Approximated performance points corresponding to the two stage decentralized strategies achieved using Algorithm 5 in the heavy tailed prior case. The results for the 10 sample sets for $\lambda = 0, 0.05, ..., 0.35$ are displayed. For each value of λ an ellipsoid is fit to the performance points for the sample sets.

In order to present Algorithm 5 in an example involving a heavy tailed non-Gaussian prior distribution, we consider $X = \{X_1, X_2, X_3, X_4\}$ distributed with a multivariate-symmetric Laplacian density $SL_4(C_X)$ as considered in Section 3.5.2 for the directed case. The decentralized estimation strategy is as described for the Gaussian example in the previous section. For different values of λ, specifically $0.05, 0.1, ..., 0.35$ the Algorithm is run for 10 different sample sets $S_{x_1}, ..., S_{x_4}$ and $S_{y_1}, ..., S_{y_4}$ which are also employed for the directed case.

The approximated performance points are presented in Figure 4.3. Similarly, an ellipsoid is fit to the results for each value of λ. Similar to the Gaussian case, for small values of λ, a decentralized strategy with a comparable performance to the centralized rule is achieved. The benefits of the undirected topology is apparent comparing the results with that presented in Figure 3.4 for the directed case.

4.5 Discussion

We have adopted a decentralized inference strategy defined over an undirected graph $\mathcal{G}$ for the estimation problem. The local rules constituting the strategy act in two stages to provide a causal processing scheme as well as a tractable information structure through the unwrapped directed counterpart of the $\mathcal{G}$. Such an online processing scheme yields an iterative optimization procedure which starting from an initial strategy converges to a person-by-person one. Moreover, under reasonable conditions, the online processing exhibits scalability in the number of variables and the optimization procedure admits a message passing interpretation which also scales with the number of nodes. In the detection case, the expressions involved yields a reasonable computational burden, whereas for the estimation case they can not be practically carried out in general. In order to overcome this bottleneck, we have proposed a Monte Carlo framework including particle representations and approximate computational schemes.

On the other hand, the two stage strategy limits the information horizon of nodes with their neighbors. It is possible to extend this approach to multiple stages in order to widen the horizon but even for the detection case, issues regarding mathematical tractability arise and the iterative scheme requires approximations [28]. On the other hand, the estimation setting requires approximations inherently in the two-stage setting. For sake of brevity and considering the computational time required for the iterations, we kept the discussion on multi-stage strategies out of the scope of the thesis, to the best of the author' s, the framework we propose is valid in the multi-stage setting.

Another issue is that both in the directed and the undirected case, the inference task is distributed among platforms through random variable-node associations. Hence, not every node necessarily performs inference and a directed architecture in which only the child nodes carry out this task is possible. Moreover, the advantages of the undirected architectures due to the bi-directional transmissions they provide are presented. The problem they exhibit is due to the limited information horizon which is not a dominant factor on graphs with low radius. Therefore a hybrid network in which a directed strategy that assigns the inference tasks to the childless nodes and an undirected network that overlays these nodes is useful. A certain hybrid strategy guarantees that the person-by-person optimal hybrid network is achieved in terms of the optimal directed networks and the optimal overlay network [29]. The Monte Carlo framework we provide is valid for such a scheme in the estimation setting.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this thesis, motivated by networked sensing, we have studied the decentralized estimation problem under communication constraints. Networked platforms evaluate local functions, which constitute the global strategy, in order to estimate the value of a random vector in such a scenario where it is not feasible to collect observations at a center which is responsible for performing all the required computations. On the other hand, the so called "in-network" processing arises a certain communication load which consumes the major part of the limited energy budget. One aspect of the problem is to consider the strategy as the global function to be evaluated and associate its sub-tasks with the platforms such that the arising communication load is the lowest among a set of possible mappings. The graphical models approach to decentralized inference can be considering in accordance with such a treatment.

We are interested in design of global functions given the communication structure together with the cost of the transmissions arising. In particular, the tradeoff between the estimation accuracy and the communication cost is of concern to us in which schemes enabling gracefull degredation of the accuracy by decreasing the communication load is appealing. The conventional approach to decentralized estimation can be considered along these lines in which quantizers and fusion rules are sought while addressing a restricted set of the communication constraints due to lack of mathematical tracktability and scalability arising otherwise.

We adopt a recent framework for decentralized detection which captures a wide range of communication constraints and avoid restricting the problem definition. We consider two classes of decentralized strategies; i) decentralized strategies over directed acyclic graphs and ii) two-stage strategies over undirected graphs. The second reveals a representation in terms of the former and it is possible to investigate the hybrid strategies and multi-stage strategies

(see Chp.4 and 5 of [28]) in a similar fashion, through considering the directed "computation structure" they imply. Besides the intricacies related with the optimization of a given strategy in order to achieve an acceptable accuracy for a reasonable communication cost, additional challenges arise in the estimation setting. The expressions involved by the optimization procedures after a Team Theoretic investigation include integral operators that are impossible to evaluate exactly in practice, in general. Our contribution is a Monte Carlo framework that overcomes this difficulty through particle representations and approximate computational schemes in the same spirit with that of Non-parametric Belief Propagation which has proved useful for statistical inference under general continuous distributions.

The algorithms we provide has a number of advantages altough being approximate in nature. First of all they enable to introduce the prior information through the the Bayesian setting involved. Second, the cost of communications is captured within the Bayesian risk and also reflecting the network constraints to the feasible set, they allow a wide range of communication constrained problem to be modelled. The original non-approximate schemes exhibit scalability in the number of variables and platforms under certain assumptions, which we preserve in the estimation setting through utilizing the Monte Carlo methods adequately. In addition, samples from only marginal distributions are required, and are generated in a straightforward manner compared to samples from joint distributions. Moreover, we provide scalability in the sample set sizes. The framework is not limited with standard distributions such as Gaussians but can produce results for any set of distributions provided that samples can be generated from the marginals. Last but not the least, it enables the estimation accuracy gracefully degrade as the "price" of communication is incereased, also allowing a quantification of the tradeoff between the cost of communications and estimation accuracy through performance points of the converged strategies while the price of communications is varied.

5.1 Future Extensions

There is a number of issues left open throughout the thesis. First, it would be useful to employ the framework we provide in a number of problem settings and emprically observe the behaviour of the strategies obtained. One possible class of scenarios involve large networks since the Monte Carlo optimization schemes exhibit scalability. It is possible to consider the problem of random field estimation and compare the achieved performances of the result-

ing strategies with other schemes that can provide benchmarks. Another possible scenario involves target localization in a tracking scenario explained in Section 3.6. In addition, the framework can be applied to hybrid network structures and multi-stage strategies over undirected graphs. The later generalizes a finite step loopy messaging scheme and can employ a feedback mechanism. The fact that the memory requirement grows exponentially renders exact solutions infeasible even in the detection setting. Investigation of Monte Carlo approximations in this context remains an open issue.

One possible research direction is the regularization of the optimization schemes through Kernels. Such methods have been employed in the context of particle filtering in order to reduce the size of the sample sets. A side issue is that, through numerical approximations of the multi-dimensional error function, it is possible to obtain a different approximation scheme for the Gaussian case, not involving Monte Carlo methods.

Another aspect left open involves the investigation of error bounds and biases of the resulting strategies as well as robust termination conditions for the iterations. A starting point might be considering symmetric distributions together with Geweke' s conditions on consistent Importance Sampling approximations.

REFERENCES

[1] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, "Instrumenting the world with wireless sensor networks," in *Proceedings of the 2001 IEEE International Conference on Acoustic, Speech and Signal Processing*, May 2001.

[2] V. Raghunathan, C. Schurgers, S. Park, and M. B. Srivastava, "Energy-aware wireless microsensor networks," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 40–50, March 2002.

[3] G. J. Pottie and W. Kaiser, "Wireless integrated network sensors," *Communications of the ACM*, vol. 43, no. 5, pp. 51–58, May 2000.

[4] C.-Y. Chong and S. P. Kumar, "Sensor networks: Evolution, opportunities, and challenges," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1247–1256, August 2003.

[5] M. Çetin, L. Chen, J. W. Fisher III, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, "Distributed fusion in sensor networks: A graphical models perspective," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 42–55, July 2006.

[6] P. Santi, "Topology control in wireless ad hoc sensor networks," *ACM Computing Surveys*, vol. 87, no. 2, pp. 164–194, June 2005.

[7] M. Yarvis and W. Ye, *Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems*, M. Ilyas and I. Mahgoub, Eds. CRC Press, 2005, chp. 13, Tiered Architectures in Sensor Networks.

[8] W. Su, E. Çayırcı, and Özgür B. Akan, *Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems*, M. Ilyas and I. Mahgoub, Eds. CRC Press, 2005, chp. 16, Overview of Communication Protocols for Sensor Networks.

[9] B. Krishnamachari, *Networking Wireless Sensors*. Cambridge University Press, 2005.

[10] H. Karl and A. Willig, *Protocols and Architectures for Wireless Sensor Networks*. John Wiley &Sons, 2005.

[11] R. Min and A. Chandrakasan, "A framework for energy-scalable communication in high-density wireless networks," in *Proceedings of the 2002 International Symposium on Low Power Electronics and Design,*, August 2002, pp. 36–41.

[12] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, January 2000.

[13] P. S. Bernard and J. M. Wallace, *Turbulent Flow: Analysis, Measurement and Prediction*. John Wiley &Sons, 2002.

[14] H. Wackernagel, *Multivariate Geostatistics*, 3rd ed. Springer, 2003.

[15] O. Schabenberger and C. A. Gotway, *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC Press, 2005.

[16] F. Zhao and L. Guibas, *Wireless Sensor Networks: An Information Processing Approach*. Morgan Kaufmann, 2004.

[17] D. Li, K. D. Wong, Y. H. Hu, and A. M. Sayeed, "Detection, classification and tracking of targets," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 17–29, March 2002.

[18] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, February 2002.

[19] J. Liu, M. Chu, and J. E. Reich, "Multitarget tracking in distributed sensor networks," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 36–46, May 2007.

[20] J. Liu, J. Reich, and F. Zhao, "Collaborative in-network processing for target tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 4, pp. 378–391, January 2003.

[21] M. K. Chu, H. Haussecker, and F. Zhao, "Scalable information-driven sensor querying and routing for ad hoc heterogenous sensor networks," *International Journal of High Performance Computing and Applications*, vol. 16, no. 3, pp. 293–313, August 2002.

[22] F. Zhao, "Information driven dynamic sensor collaboration," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 61–72, March 2002.

[23] J. Liu, J. Liu, J. Reich, P. Cheung, and F. Zhao, "Distributed group management in sensor networks: Algorithms and applications to localization and tracking," *Telecommunication Systems Journal*, vol. 26, no. 2-4, pp. 235–251, 2004.

[24] M. K. Chu, S. K. Mitter, and F. Zhao, "Distributed multiple target tracking and data association in ad hoc sensor networks," in *Proceedings of the 6th International Conference on Information Fusion*. ICIF' 03, July 2003, pp. 447–454.

[25] M. K. Chu, "A hierarchical framework for constructing computationally efficient algorithms for distributed inference problems," Ph.D. dissertation, MIT, 2003.

[26] O. P. Kreidl and A. S. Willsky, "Decentralized detection in undirected network topologies," in *Proceedings of the IEEE 14th Workshop on Statistical Signal Processing*. SSP '07, August 2007, p. 650=654.

[27] ——, "An efficient message-passing algorithm for optimizing decentralized detection networks," MIT Laboratory for Information and Decision Systems, Tech. Rep. LIDS-P-2726, 2007, iEEE Transactions on Automatic Control, To appear.

[28] O. P. Kreidl, "Graphical models and message-passing algorithms for network-constrained decision problems," Ph.D. dissertation, MIT, 2008.

[29] O. P. Kreidl and A. S. Willsky, "Decentralized detection with long-distance communication," in *Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers*, October 2008, pp. 1352–1356.

[30] R. Diestel, *Graph Theory*. Springer-Verlag, 2005.

[31] S. L. Lauritzen, *Graphical Models*. Oxford University Press, 1996.

[32] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. McGraw-Hill, 2002.

[33] J. N. Tsitsiklis, *Advances in Statistical Signal Processing*. JAI Press, 1993, ch. Decentralized Detection, pp. 297–344.

[34] H. L. V. Trees, *Detection, Estimation and Modulation Theory, Part I*. John Wiley &Sons, 2001.

[35] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. Springer, 1994.

[36] S. M. Aji and R. J. McEliece, "The generalized distributive law," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 325–343, March 2000.

[37] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," *Artificial Intelligence*, vol. 42, pp. 393–405, 1990.

[38] R. Cowell, *Learning in Graphical Models*, M. I. Jordan, Ed. Kluwer Academic Publishers, 1998, "Introduction to Inference for Bayesian Networks".

[39] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, February 2001.

[40] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[41] P. Brémaud, *Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues*. Springer-Verlag, 1991.

[42] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*. Springer, 1999.

[43] F. R. Kschischang, "Codes defined on graphs," *IEEE Communications Magazine*, vol. 41, no. 4, pp. 118–125, August 2003.

[44] M. J. Wainwright, T. Jaakkola, and A. S. Willsky, "Tree consistency and bounds on the performance of the max-product algorithm and its generalizations," *Statististics and Computing*, vol. 14, no. 2, pp. 143 – 166, April 2004.

[45] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282– 2312, July 2005.

[46] A. T. Ihler, "Inference in sensor networks: Graphical models and particle methods," Ph.D. dissertation, MIT, 2005.

[47] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, December 1979.

[48] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Multi-target tracking using joint probabilistic data association," in *Proceedings of the 19th IEEE Conference on Decision and Control*, December 1980, pp. 807–812.

[49] S. Deb, M. Yeddanapudi, K. Pattipati, and Y. Bar-Shalom, "A generalized S-D assignment algorithm for multisensor-multitarget state estimation," *IEEE Transactions on Aerospace and Electronics Systems*, vol. 33, no. 2, pp. 523–538, April 1997.

[50] R. L. Popp, K. R. Pattipati, and Y. Bar-Shalom, "*m*-best *s*-d assignment algorithm with application to multitarget tracking," *IEEE Transaction on Aerospace and Electronic Systems*, vol. 37, no. 1, pp. 22–39, January 2001.

[51] C. Hue, J.-P. L. Cadre, and P. Pêrez, "Tracking multiple objects with particle filtering," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 791–812, July 2002.

[52] L. Chen, M. Wainwright, M. Çetin, and A. S. Willsky, "Data association based on optimization in graphical models with application to sensor networks," *Mathematical and Computer Modelling*, vol. 43, no. 9-10, pp. 1114–1135, May 2006, special Issue on Optimization and Control for Military Applications.

[53] A. T. Ihler, J. W. Fisher III, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for self-calibration in sensor networks," *IEEE Journal of Selected Areas in Communication*, vol. 23, no. 4, pp. 809–819, April 2005.

[54] C. Crick and A. Pfeffer, "Loopy belief propagation as a basis for communication in sensor networks," in *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*. UAI' 03, August 2003, pp. 159–166.

[55] M. A. Paskin and C. E. Guestrin, "Robust probabilistic inference in distributed systems," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. UAI' 04, July 2004, pp. 436–445.

[56] M. K. Chu, S. K. Mitter, and F. Zhao, "An information architecture for distributed inference on ad hoc sensor networks," in *Proceedings of 41st Annual Allerton Conference on Communication, Control and Computing*. Allerton, October 2003.

[57] V. Saligrama, M. Alanyalı, and O. Savaş, "Distributed detection in sensor networks with packet losses and finite capacity links," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4118–4132, November 2006.

[58] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, September 1949.

[59] N. Metropolis, A. W.Rosenbluth, M. N. Rosenbluth, A. H. Teller, and W. Teller, "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087 – 1092, June 1953.

[60] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, September 1962.

[61] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. Springer, 2001.

[62] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer, 2004.

[63] J. Geweke, "Bayesian inference in econometric models using monte carlo integration," *Econometrica*, vol. 57, no. 6, pp. 1317–1339, November 1989.

[64] A. Douced and W. Xiaodong, "Monte carlo methods for signal processing: a review in the statistical signal processing context," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 152–170, November 2005.

[65] J. C. Spall, "Estimation via markov chain monte carlo," *IEEE Control Systems Magazine*, vol. 23, no. 2, pp. 34–45, April 2003.

[66] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, May 2007.

[67] C. Hue, J.-P. L. Cadre, and P. Pérez, "Sequential monte carlo methods for multiple target tracking and data fusion," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 309–325, February 2002.

[68] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky, "Nonparametric belief propagation," MIT Laboratory for Information and Decision Systems, Tech. Rep. LIDS-P-2551, October 2002.

[69] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, "Visual hand tracking using nonparametric belief propagation," in *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop*, vol. 12. CVPR' 04, July 2004.

[70] Y. Yang, *Information processing and routing in wireless sensor networks*. World Scientific, 2006.

[71] A. Giridhar and P. Kumar, "Toward a theory of in-network computation in sensor networks," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 98–107, April 2006.

[72] E. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky, "Non-parametric belief propagation," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. CVPR' 03, June 2003.

[73] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural Computation*, vol. 12, no. 1, pp. 1–41, January 2000.

[74] M. Paskin, C. Guestrin, and J. McFadden, "A robust architecture for distributed inference in sensor networks," in *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks*. IPSN' 05, April 2005.

[75] L. Chen, M. Çetin, and A. S. Willsky, "Distributed data association for multi-target tacking in sensor networks," in *Proceedings of the 8th International Conference on Information Fusion*. ISIF' 05, July 2005.

[76] A. T. Ihler, J. W. Fisher III, and A. S. Willsky, "Loopy belief propagation: Convergence and effects of message errors," *Journal of Machine Learning Research*, vol. 6, pp. 905 – 936, May 2006.

[77] R. R. Tenney and J. Nils R. Sandell, "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-17, no. 4, pp. 501–510, July 1981.

[78] J. N. Tsitsiklis and M. Athans, "On the complexity of decentralized decision making and detection problems," *IEEE Transactions on Automatic Control*, vol. AC-30, no. 5, pp. 440–446, May 1985.

[79] F. A. Sadjadi, "Hypothesis testing in a distributed environment," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22, no. 2, pp. 134–137, March 1986.

[80] Z. Chair and P. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22, no. 1, pp. 98–101, January 1986.

[81] Z. B. Tang, K. R. Pattipati, and D. L. Kleinmann, "Optimization of detection networks: Part i–tandem structures," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 5, pp. 1044–1059, September/October 1991.

[82] Z.-B. Tang, K. Pattipati, and D. Kleinmann, "Optimization of detection networks: Part ii– tree structures," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 1, pp. 211–221, January/February 1993.

[83] A. Pete, K. Pattipati, and D. L. Kleinmann, "Optimization of detection network structures with multiple event structures," *IEEE Transactions on Automatic Control*, vol. 39, no. 8, pp. 1702–1707, August 1994.

[84] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part i–fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, January 1997.

[85] O. P. Kreidl and A. S. Willsky, "An efficient message-passing algorithm for optimizing decentralized detection networks," in *Proceedings of the IEEE 45th Conference on Decision and Control*, December 2006, pp. 6776–6783.

[86] Y.-C. Ho and K.-C. Chu, "Team decision theory and information structures in optimal control problems," *IEEE Transactions on Automatic Control*, vol. 29, no. 1, pp. 22 – 28, February 1972.

[87] Z.-Q. Luo, "Universal decentralized estimation in a bandwidth constrained sensor network," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 2210–2219, June 2005.

[88] Y. Wang and P. Ishwar, "Distributed field estimation with randomly deployed, noisy, binary sensors," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1177–1189, March 2009.

[89] J. A. Gubner, "Distributed estimation and quantization," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1456–1459, July 1993.

[90] W.-M. Lam and A. R. Reibman, "Design of quantizers for decentralized estimation systems," *IEEE Transactions on communications*, vol. 41, no. 11, pp. 1602–1605, November 1993.

[91] A. Ribeiro, I. D. S. ad Jin-Jun Xiao, G. B. Giannakis, and Z.-Q. Luo, *Wireless Sensor Networks: Signal Processing and Communication Perspectives*, A. Swami, Q. Zhao, Y.-W. Hong, and L. Tong, Eds. John Wiley &Sons, 2007.

[92] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks-part i: Gaussian case," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 1131 – 1143, March 2006.

[93] ——, "Bandwidth-constrained distributed estimation for wireless sensor networks-part ii: Unknown probability density function," *IEEE Transactions on Signal Processing*, vol. 54, no. 7, pp. 2784 – 2796, July 2006.

[94] P. Venkitasubramaniam, L. Tong, and A.Swami, "Quantization for maximin are in distributed estimation," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3596–3605, July 2007.

[95] Y. Wang, P. Ishwar, and V. Saligrama, "One-bit distributed sensing and coding for field estimation in sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4433–4445, September 2008.

[96] A. Dogandzic and K. Qiu, "Decentralized random-field estimation for sensor networks using quantized spatially correlated data and fusion center feedback," *IEEE Transactions on Signal Processing*, vol. 56, no. 12, pp. 6069–6085, December 2008.

[97] J. N. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Transaction on Automatic Control*, vol. 29, no. 1, pp. 42–50, January 1984.

[98] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, MIT, 2006.

[99] J. Max, "Quantization for minimum distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, March 1960.

[100] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. IT-28, no. 2, pp. 129–137, March 1982.

[101] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, November 1984.

[102] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, no. 140, pp. 398–409, June 1990.

[103] M. Abramowitz and I. A.Stegun, Eds., *Handbook of Mathematical Functions*. Dover Publications, Inc., 1965, page 376.

[104] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in Neural and Information Processing Systems*. Conference on Neural Information Processing Systems (NIPS), Nov. 29-Dec.2 1999, pp. 855–861.

[105] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Computational Harmonic Analysis*, vol. 11, pp. 89–123, 2001.

[106] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, May 2006.

[107] S. Kotz, T. J. Kozubowski, and K. Podgorski, *The Laplace Distribution and generalizations*. Birkhauser, 2001.

[108] M. Pilu, A. Fitzgibbon, and R. B. Fisher, "Ellipse-specific direct least-square fitting," in *Proceedings of the IEEE International Conference on Image Processing*, vol. 3. ICIP'96, September 1996, pp. 599–602.

[109] S. Alhakeem and P. K. Varshney, "Decentralized bayesian detection with feedback," *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, vol. 26, no. 4, pp. 503–513, July 1996.

APPENDIX A

Decentralized Strategies Under Directed Acyclic Graphs

A.1 Information Structure of Decentralized Strategies Under Directed Acyclic Graphs

Consider the decentralized inference scheme for which the online processing is modelled with a directed acyclic graph introduced in [27] summarized in Section 2.7.1.1 and utilized in Chapter 3. The Bayesian framework which assigns an expected cost $J(\gamma) = E\{c(u, x, \hat{x}); \gamma\}$for any given strategy $\gamma \in \Gamma^{\mathcal{G}}$ with the underlying distribution $p(u, x, \hat{x}; \gamma)$ is contributed by the strategy by the conditional distribution $p(u, \hat{x}|x, y; \gamma)$.

We first show that Eq.(2.24) and similarly Eq.(3.2) holds for any strategy under a directed acyclic graph for which the node labels are in accordance with the forward partial ordering. Consider $p(u, \hat{x}|x, y; \gamma)$ and separate the variables output by the local rule of node 1 which is a parentless node using the Bayes' rule and obtain

$$p(u, \hat{x}|x, y; \gamma) = p(u_1, \hat{x}_1|u_{\setminus 1}, \hat{x}_{\setminus 1}, x, y; \gamma_1, \gamma_{\setminus 1}) p(u_{\setminus 1}, \hat{x}_{\setminus 1}|x, y; \gamma_1, \gamma_{\setminus 1}) \tag{A.1}$$

The first term in the right hand side reduces to $p(u_1, \hat{x}_1|y_1; \gamma_1)$ since γ_1 operates only on $y_1 \in \mathcal{Y}_1$ and $u_{\setminus 1}$, $\hat{x}_{\setminus 1}$ together with $\gamma_{\setminus 1}$ has no bearing on the local rule γ_1 due to the directed acyclic structure of $\mathcal{G}$. Considering the second term, if γ_1 is determined, then given x, y the output u_1 is known and this term reduces to $p(u_{\setminus 1}, \hat{x}_{\setminus 1}|u_1, x, y; \gamma_{\setminus 1})$ Suppose that node k is the first node in the ordering with parents. Then, applying the procedure above until node k we achieve

$$p(u, \hat{x}|x, y; \gamma) = p(u_{\setminus\{1,\ldots,k-1\}}, \hat{x}_{\setminus\{1,\ldots,k-1\}}|u_{\{1,\ldots,k-1\}}, x, y; \gamma_{\setminus\{1,\ldots,k-1\}}) \prod_{i=1}^{k-1} p(u_i, \hat{x}_i|y_i; \gamma_i) \tag{A.2}$$

and it is possible to separate the contribution of the rule local to node k from the first term by

$$p(u_{\setminus\{1,\ldots,k-1\}}, \hat{x}_{\setminus\{1,\ldots,k-1\}} | u_{\{1,\ldots,k-1\}}, x, y; \gamma_{\setminus\{1,\ldots,k-1\}}) =$$
$$p(u_k, \hat{x}_k | u_{\setminus k}, \hat{x}_{\setminus\{1,\ldots,k\}}, x, y, ; \gamma_k, \gamma_{\setminus\{1,\ldots,k\}}) p(u_{\setminus\{1,\ldots,k\}}, \hat{x}_{\setminus\{1,\ldots,k\}} | u_{\{1,\ldots,k-1\}}, x, y; \gamma_k, \gamma_{\setminus\{1,\ldots,k\}}) \quad \text{(A.3)}$$

Since γ_k operates only on y_k and $u_{\pi(k)}$ and $u_{\pi(k)} \subseteq u_{\setminus k}$, the first term in the right hand side of the above equation reduces to $p(u_k, \hat{x}_k | y_k, u_{\pi(k)}; \gamma_k)$. Considering the second term, since when y is given and γ_k is determined u_k is known, we obtain $u_{\{1,\ldots,k-1\}} \cup u_k = u_{\{1,\ldots,k\}}$ and hence $p(u_{\setminus\{1,\ldots,k\}}, \hat{x}_{\{1,\ldots,k\}} | u_{\{1,\ldots,k\}}, x, y, \gamma_{\setminus\{1,\ldots,k\}})$. Recursively applying the same procedure to the second term that we obtain after separating the j^{th} local rule' s contribution and collecting the first terms within the multiplication, we end up with Eq.(2.24) which reveals the coupling of local rules to the global structure.

Provided that Assumption 1, i.e. the conditional independence, holds, then the underlying distribution to the Bayesian framework is obtained by

$$\begin{aligned} p(u, x, \hat{x}; \gamma) &= \int_{\mathcal{Y}} \mathrm{d}y \prod_{i=1}^{N} p(u_i, \hat{x}_i | y_i, u_{\pi(i)}; \gamma_i) \prod_{k=1}^{N} p(y_k | x) p(x) \\ &= p(x) \prod_{i=1}^{N} p(u_i, \hat{x}_i | x, u_{\pi(i)}; \gamma_i) \end{aligned} \quad \text{(A.4)}$$

which further implies that

$$p(u, \hat{x} | x; \gamma) = \prod_{i=1}^{N} p(u_i, \hat{x}_i | x, u_{\pi(i)}; \gamma_i) \quad \text{(A.5)}$$

An expression which is commonly encountered is the product of local rules except the j^{th}, i.e. $\prod_{i \neq j} p(u_i, \hat{x}_i | x, u_{\pi(i)}; \gamma_i)$. The conditional distribution that equals to this product is obtained by dividing both sides of Eq.(A.5) by the contribution of the j^{th} rule. Then we have

$$\begin{aligned} \prod_{i \neq j} p(u_i, \hat{x}_i | x, u_{\pi(i)}; \gamma_i) &= \frac{p(u, \hat{x} | x; \gamma)}{p(u_j, \hat{x}_j | x, u_{\pi(j)}; \gamma_j)} \\ &= \frac{p(u_{\setminus\pi(j)}, \hat{x} | x, u_{\pi(j)}; \gamma) p(u_{\pi(j)} | x; \gamma)}{p(u_j, \hat{x}_j | x, u_{\pi(j)}; \gamma_j)} \\ &= p(u_{\setminus j \cup \pi(j)}, \hat{x}_{\setminus j} | x, u_{\pi(j)}, u_j, \hat{x}_j; \gamma) p(u_{\pi(j)} | x; \gamma) \\ &= p(u_{\setminus j}, \hat{x}_{\setminus j} | x, u_j; \gamma_{\setminus j}) \end{aligned} \quad \text{(A.6)}$$

for which after applying the chain and Bayes' rule, we have substituted the conditional independence properties $u_{\pi(j)} \perp\!\!\!\perp (u_j, \hat{x}_j) | x; \gamma_{an(j)}$ where $an(j)$ are the set of ancestor nodes of j and $(u_{\setminus j}, \hat{x}_{\setminus j}) \perp\!\!\!\perp \hat{x}_j | x, u_j; \gamma_{\setminus j}$ due to the directed acyclic nature in the last step.

A.2 Proof of Proposition 3.3.2

We follow similar steps with that for the detection case in [27] whereas in our setting, X takes values from a denumarable set $\mathcal{X}$ and we do not utilize a channel mode, i.e. we assume that all links are error free.

Consider Eq.s(3.9) and (3.14) together with Eq.(A.6). After substituting the mathematical expression of the cost locality assumption, i.e. Eq.(2.34), in Eq.(3.14) we obtain

$$\begin{aligned}
&\theta_j^*(u_j, \hat{x}_j, x, u_{\pi(j)}) \\
&\quad = \sum_{i\in\mathcal{V}} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_i, x_i, \hat{x}_i) p(u_{\setminus j}, \hat{x}_{\setminus j}, x | u_j; \gamma^*_{\setminus j}) \\
&\quad = p(u_{\pi(j)}|x; \gamma^*_{\setminus j}) p(x) c(u_j, x_j, \hat{x}_j) + \sum_{i\in\mathcal{V}\setminus j} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_i, x_i, \hat{x}_i) p(u_{\setminus j}, \hat{x}_{\setminus j}, x | u_j; \gamma^*_{\setminus j})
\end{aligned}$$

and treat the summation over $i \in \mathcal{V}\setminus j$ in three groups: $de(j) \in \mathcal{V}\setminus j$ denoting the decendants of node j, $\pi(j) \in \mathcal{V}\setminus j$ denoting the parent of node j and $an(j)\setminus\pi(j) \in \mathcal{V}\setminus j$ denoting the ancestors of node j that are not its parents. Due to the directed acyclic nature, these sets are mutually exclusive. Hence

$$\begin{aligned}
&\sum_{i\in\mathcal{V}\setminus j} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_i, x_i, \hat{x}_i) p(u_{\setminus j}, \hat{x}_{\setminus j}, x | u_j; \gamma^*_{\setminus j}) \\
&\quad = \sum_{m\in de(j)} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_m, x_m, \hat{x}_m) p(u_{\setminus j}, \hat{x}_{\setminus j}, x | u_j; \gamma^*_{\setminus j}) \\
&\qquad\qquad + \sum_{k\in\pi(j)} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_k, x_k, \hat{x}_k) p(u_{\setminus j}, \hat{x}_{\setminus j}, x | u_j; \gamma^*_{\setminus j}) \\
&\qquad\qquad\quad + \sum_{n\in an(j)\setminus\pi(j)} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_n, x_n, \hat{x}_n) p(u_{\setminus j}, \hat{x}_{\setminus j}, x | u_j; \gamma^*_{\setminus j})
\end{aligned}$$

Consider the first group on the right hand side of the equation above. The following holds

$$\begin{aligned}
&\sum_{m\in de(j)} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_m, x_m, \hat{x}_m) p(u_{\setminus\{j\}\cup\pi(j)}, \hat{x}_{\setminus j} | x, u_j, u_{\pi(j)}; \gamma^*_{\setminus j}) p(u_{\pi(j)}|x; \gamma^*_{\setminus j}) p(x) \\
&\quad = \sum_{m\in de(j)} \sum_{u_m} \int_{\mathcal{X}_m} \mathrm{d}\hat{x}_m\, c(u_m, x_m, \hat{x}_m) \sum_{u_{\setminus\{j,m\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus\{j,m\}}} \mathrm{d}\hat{x}_{\setminus\{j,m\}}\, p(u_{\setminus\{j\}\cup\pi(j)}, \hat{x}_{\setminus j} | x, u_j, u_{\pi(j)}; \gamma^*_{\setminus j}) p(u_{\pi(j)}|x; \gamma^*_{\setminus j}) p(x) \\
&\quad = p(u_{\pi(j)}|x; \gamma^*_{\setminus j}) p(x) \sum_{m\in de(j)} \sum_{u_m} \int_{\mathcal{X}_m} \mathrm{d}\hat{x}_m\, c(u_m, x_m, \hat{x}_m) p(u_m, \hat{x}_m | x, u_j, u_{\pi(j)}; \gamma^*_{\setminus j}) \\
&\quad = p(u_{\pi(j)}|x; \gamma^*_{\setminus j}) p(x) \sum_{m\in de(j)} \sum_{u_m} \int_{\mathcal{X}_m} \mathrm{d}\hat{x}_m\, c(u_m, x_m, \hat{x}_m) p(u_m, \hat{x}_m | x, u_j; \gamma^*_{\setminus j})
\end{aligned}$$

where we apply the chain rule and rearrange the order of operators except that in the last step we assert the assumption that $\mathcal{G}$ is a polytree when $p(u_m, \hat{x}_m|x, u_j, u_{\pi(j)}; \gamma^*_{\setminus j})$ is reduced to $p(u_m, \hat{x}_m|x, u_j; \gamma^*_{\setminus j})$ for $m \in de(j)$. Since the polytree topology implies that there are no paths from any of the ancestors of node j to any of its decendats that does not pass through j, given u_j and having $\gamma^*_{\setminus j}$ determined, $u_{\pi(j)}$ has no bearing on $(u_m, \hat{x}_m)$ where $m \in de(j)$ which would not necessarily be the case if $\mathcal{G}$ were not a polytree.

Considering the summation over the second group, similar rearrangements are performed yielding

$$
\begin{aligned}
&\sum_{k\in\pi(j)} \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_k, x_k, \hat{x}_k) p(u_{\setminus j}, \hat{x}_{\setminus j}, x|u_j; \gamma^*_{\setminus j}) \\
&= \sum_{k\in\pi(j)} \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k\, c(u_k, x_k, \hat{x}_k) p(x) \sum_{u_{\setminus\{j\}\cup\pi(j)}} \int_{\mathcal{X}_{\setminus\{j,k\}}} \mathrm{d}\hat{x}_{\setminus\{j,k\}}\, p(u_{\setminus j}, \hat{x}_{\setminus j}|x, u_j; \gamma^*_{\setminus j}) \\
&= \sum_{k\in\pi(j)} \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k\, c(u_k, x_k, \hat{x}_k) p(x) p(u_{\pi(j)}, \hat{x}_k|x, u_j; \gamma^*_{\setminus j}) \\
&= \sum_{k\in\pi(j)} \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k\, c(u_k, x_k, \hat{x}_k) p(x) p(u_k, \hat{x}_k|x, u_j, u_{\pi(j)\setminus k}; \gamma^*_{\setminus j}) p(u_{\pi(j)\setminus k}|x, u_j; \gamma^*_{\setminus j}) \\
&= \sum_{k\in\pi(j)} \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k\, c(u_k, x_k, \hat{x}_k) p(x) p(u_k, \hat{x}_k|x; \gamma^*_{\setminus j}) p(u_{\pi(j)\setminus k}|x; \gamma^*_{\setminus j}) \\
&= p(x) p(u_{\pi(j)\setminus k}|x; \gamma^*_{\setminus j}) \sum_{k\in\pi(j)} \int_{\mathcal{X}_k} \mathrm{d}\hat{x}_k\, c(u_k, x_k, \hat{x}_k) p(u_k, \hat{x}_k|x; \gamma^*_{\setminus j})
\end{aligned}
$$

where in the first two steps, we rearrange operators and perform marginalization, in the third step we apply the chain rule. In the fourth step, the u_j and $u_{\pi(j)\setminus k}$ arguments of the conditional drops since due to the polytree topology no two parents of node j shares a common ascendant and these arguments are non-informative for $(u_k, \hat{x}_k)$ when $\gamma^*_k \in \gamma^*_{\setminus j}$ is determined. Also note that, at the last step, the terms contain no contribution of $(u_j, \hat{x}_j)$ and hence have no bearing on the optimization regarding the person-by-person optimal rule of node j.

A similar treatment of the third group yields

$$
\begin{aligned}
&\sum_{n\in an(j)\setminus\pi(j)}\sum_{u_{\setminus\{j\}\cup\pi(j)}}\int_{\mathcal{X}_{\setminus j}} \mathrm{d}\hat{x}_{\setminus j}\, c(u_n,x_n,\hat{x}_n)p(u_{\setminus j},\hat{x}_{\setminus j},x|u_j;\gamma^*_{\setminus j})\\
&= \sum_{n\in an(j)\setminus\pi(j)}\sum_{u_n}\int_{\mathcal{X}_n}\mathrm{d}\hat{x}_n\, c(u_n,x_n,\hat{x}_n)\sum_{u_{\setminus\{j,n\}\cup\pi(j)}}\int_{\mathcal{X}_{\setminus\{j,n\}}}\mathrm{d}\hat{x}_{\setminus\{j,n\}}\, p(u_{\setminus j},\hat{x}_{\setminus j},x|u_j;\gamma^*_{\setminus j})\\
&= \sum_{n\in an(j)\setminus\pi(j)}\sum_{u_n}\int_{\mathcal{X}_n}\mathrm{d}\hat{x}_n\, c(u_n,x_n,\hat{x}_n)p(x)p(u_n,\hat{x}_n|u_{\pi(j)},u_j,x;\gamma^*_{\setminus j})p(u_{\pi(j)}|u_j,x;\gamma^*_{\setminus j})\\
&= \sum_{n\in an(j)\setminus\pi(j)}\sum_{u_n}\int_{\mathcal{X}_n}\mathrm{d}\hat{x}_n\, c(u_n,x_n,\hat{x}_n)p(x)p(u_n,\hat{x}_n|x;\gamma^*_{\setminus j})p(u_{\pi(j)}|x;\gamma^*_{\setminus j})\\
&= p(x)p(u_{\pi(j)}|x;\gamma^*_{\setminus j})\sum_{n\in an(j)\setminus\pi(j)}\sum_{u_n}\int_{\mathcal{X}_n}\mathrm{d}\hat{x}_n\, c(u_n,x_n,\hat{x}_n)p(u_n,\hat{x}_n|x;\gamma^*_{\setminus j})
\end{aligned}
$$

revealing that it has no contribution on the optimization regarding the person-by-person optimal rule of node j either.

Therefore

$$
\begin{aligned}
&\theta^*_j(u_j,\hat{x}_j,x,u_{\pi(j)}) \propto\\
&\quad p(x)p(u_{\pi(j)}|x;\gamma^*_{\setminus j})\left[c(u_j,x_j,\hat{x}_j)+\sum_{m\in de(j)}\sum_{u_m}\int_{\mathcal{X}_m}\mathrm{d}\hat{x}_m\, c(u_m,x_m,\hat{x}_m)p(u_m,\hat{x}_m|x,u_j;\gamma^*_{\setminus j})\right]
\end{aligned}
$$

holds and under the measurement locality assumption, Eq.(3.14) easily yields

$$
\begin{aligned}
&\phi^*_j(u_j,\hat{x}_j,x_j,u_{\pi(j)}) \propto\\
&\quad p(x_j)p(u_{\pi(j)}|x_j;\gamma^*_{\setminus j})\left[c(u_j,x_j,\hat{x}_j)+\sum_{m\in de(j)}\sum_{u_m}\int_{\mathcal{X}_m}\mathrm{d}\hat{x}_m\, c(u_m,x_m,\hat{x}_m)p(u_m,\hat{x}_m|x_j,u_j;\gamma^*_{\setminus j})\right]
\end{aligned}
\tag{A.7}
$$

after marginalization.

Now that we have obtained the form in Eq.(3.15) it remains to show that $p(u_{\pi(j)}|x_j;\gamma^*_{\setminus j})$ is equal to $P^*_j(u_{\pi(j)}|x_j)$ given by the forward likelihood recursion Eq.s(3.16) and (3.17) together with that the summation over descendants is equal to $C^*_j(u_j,x_j)$ given by the induced cost recursion Eq.s(3.18) and (3.19).

We start with a general term $p(u_{\pi(j)}|x;\gamma)$ determined by the strategy γ and fist note that the directed acyclic nature together with the online processing in accordance with the forward ordering, $u_{\pi(j)}$ received from parents depend on $\gamma_{an(j)}$ and $x_{an(j)}$ yielding the equivalence $p(u_{\pi(j)}|x_{an(j)};\gamma^*_{an(j)}) \equiv p(u_{\pi(j)}|x;\gamma^*)$ (Figure A.1). In addition, starting with parentless nodes

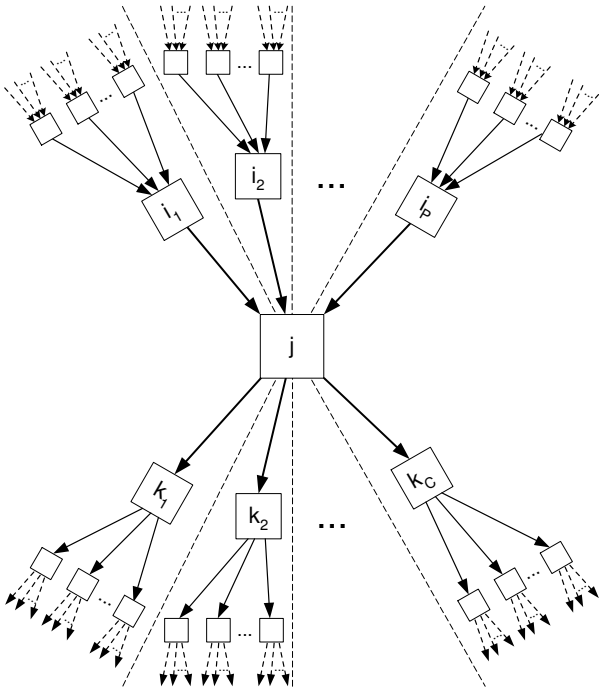


Figure A.1: A polytree from the viewpoint of node j.

for which $p(u_{\pi(j)}|x;\gamma^*) = 1$ the following recursion holds where we denote by $u_{\pi^2(j)}$ the set of incoming messages to parents of node j:

$$\begin{aligned} p(u_{\pi(j)}|x_{an(j)};\gamma^*_{an(j)}) &= \sum_{u_{\pi^2(j)}} \int_{\mathcal{X}_{\pi(j)}} \mathrm{d}\hat{x}_{\pi(j)}\, p(u_{\pi^2(j)}, u_{\pi(j)}, \hat{x}_{\pi(j)}|x_{an(j)};\gamma^*_{an(j)}) \\ &= \sum_{u_{\pi^2(j)}} \int_{\mathcal{X}_{\pi(j)}} \mathrm{d}\hat{x}_{\pi(j)} p(u_{\pi^2(j)}|x_{an(j)};\gamma^*_{an(j)}) p(u_{\pi(j)}, \hat{x}_{\pi(j)}|u_{\pi^2(j)}, x_{an(j)};\gamma^*_{an(j)}) \\ &= \sum_{u_{\pi^2(j)}} p(u_{\pi^2(j)}|x_{an(j)\setminus\pi(j)};\gamma^*_{an(j)\setminus\pi(j)}) \prod_{i\in\pi(j)} \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i \sum_{u_i\setminus u_{i\to j}} p(u_i, \hat{x}_i|x_i, u_{\pi(i)};\gamma^*_i) \end{aligned} \tag{A.8}$$

In addition, the polytree topology implies that no two parents of node j share a common ancestor and moreover the sets of ancestors of parents of node j are disjoint. Hence

$$p(u_{\pi^2(j)}|x_{an(j)\setminus\pi(j)};\gamma^*_{an(j)\setminus\pi(j)}) = \prod_{i'\in\pi(j)} p(u_{\pi(i')}|x_{an(i')};\gamma^*_{an(i')})$$

and after substituting in Eq.(A.8) we obtain

$$\begin{aligned} p(u_{\pi(j)}|x_{an(j)};\gamma^*_{an(j)}) &= \sum_{u_{\pi^2(j)}} \prod_{i'\in\pi(j)} p(u_{\pi(i')}|x_{an(i')};\gamma^*_{an(i')}) \prod_{i\in\pi(j)} \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i \sum_{u_i\setminus u_{i\to j}} p(u_i,\hat{x}_i|x_i,u_{\pi(i)};\gamma^*_i) \\ &= \sum_{u_{\pi^2(j)}} \prod_{i\in\pi(j)} p(u_{\pi(i)}|x_{an(i)};\gamma^*_{an(i)}) \sum_{u_i\setminus u_{i\to j}} \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i p(u_i,\hat{x}_i|x_i,u_{\pi(i)};\gamma^*_i) \\ &= \prod_{i\in\pi(j)} \sum_{u_{\pi(i)}} \sum_{u_i\setminus u_{i\to j}} p(u_{\pi(i)}|x_{an(i)};\gamma^*_{an(i)}) \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i p(u_i,\hat{x}_i|x_i,u_{\pi(i)};\gamma^*_i) \quad \text{(A.9)} \end{aligned}$$

Finally, in order to obtain $p(u_{\pi(j)}|x_j;\gamma^*_{an(j)})$ we multiply both sides of the above equation with $p(x_{an(j)}|x_j)$ and marginalize $X_{an(j)}$, i.e.

$$\begin{aligned} &p(u_{\pi(j)}|x_j;\gamma^*_{an(j)}) = \\ &= \int_{\mathcal{X}_{an(j)}} \mathrm{d}x_{an(j)}\, p(x_{an(j)}|x_j) \prod_{i\in\pi(j)} \sum_{u_{\pi(i)}} \sum_{u_i\setminus u_{i\to j}} p(u_{\pi(i)}|x_{an(i)};\gamma^*_{an(i)}) \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i p(u_i,\hat{x}_i|x_i,u_{\pi(i)};\gamma^*_i) \\ &= \int_{\mathcal{X}_{\pi(j)}} \mathrm{d}x_{\pi(j)} \int_{\mathcal{X}_{an(j)\setminus\pi(j)}} \mathrm{d}x_{an(j)\setminus\pi(j)}\, p(x_{\pi(j)}|x_j) p(x_{an(j)\setminus\pi(j)}|x_{\pi(j)},x_j) \prod_{i\in\pi(j)} p(u_{i\to j}|x_i,x_{an(i)};\gamma^*_i,\gamma^*_{an(i)}) \\ &= \int_{\mathcal{X}_{\pi(j)}} \mathrm{d}x_{\pi(j)}\, p(x_{\pi(j)}|x_j) \int_{\mathcal{X}_{an(j)\setminus\pi(j)}} \mathrm{d}x_{an(i)\setminus\pi(j)} \prod_{i'\in\pi(j)} p(x_{an(i')}|x_{\pi(j)},x_j,...) \prod_{i\in\pi(j)} p(u_{i\to j},\hat{x}_i|x_i,x_{an(i)};\gamma^*_i,\gamma^*_{an(i)}) \\ &= \int_{\mathcal{X}_{\pi(j)}} \mathrm{d}x_{\pi(j)}\, p(x_{\pi(j)}|x_j) \prod_{i\in\pi(j)} \int_{\mathcal{X}_{an(i)}} \mathrm{d}x_{an(i)} p(x_{an(i)}|x_i,...) p(u_{i\to j}|x_i,x_{an(i)};\gamma^*_i,\gamma^*_{an(i)}) \\ &= \int_{\mathcal{X}_{\pi(j)}} \mathrm{d}x_{\pi(j)}\, p(x_{\pi(j)}|x_j) \prod_{i\in\pi(j)} p(u_{i\to j}|x_i;\gamma^*_i,\gamma^*_{an(i)}) \\ &= \int_{\mathcal{X}_{\pi(j)}} \mathrm{d}x_{\pi(j)}\, p(x_{\pi(j)}|x_j) \prod_{i\in\pi(j)} \sum_{u_{\pi(i)}} \sum_{u_i\setminus u_{i\to j}} p(u_{\pi(i)}|x_i;\gamma^*_{an(i)}) \int_{\mathcal{X}_i} \mathrm{d}\hat{x}_i p(u_i,\hat{x}_i|x_i,u_{\pi(i)};\gamma^*_i) \end{aligned}$$

which is nothing but Eq.(3.17) substituted in Eq.s(3.16), where $P^*_j(u_{\pi(j)}|x_j)$ represents $p(u_{\pi(j)}|x_j;\gamma^*_{an(j)})$ and $P^*_{i\to j}(u_{j\to i}|x_i)$ is identified as $p(u_{i\to j}|x_i;\gamma^*_i,\gamma^*_{an(i)})$. In the first step above, we exploit the chain rule and in the next step, we substitute the disjointness of ancestors of parents of node j due to the polytree topology while factorizing $p(x_{an(j)\setminus\pi(j)}|x_{\pi(j)},x_j)$. To show that the factorization holds, let the parents of node j be $\pi(j) \triangleq \{i_1,...,i_P\}$. Then applying the chain rule consecutively we obtain

$$\begin{aligned} &p(x_{an(j)\setminus\pi(j)}|x_{\pi(j)},x_j) \\ &= p(x_{an(i_1)}|x_{\pi(j)},x_j) p(x_{an(j)\setminus\pi(j)\cup an(i_1)}|x_{\pi(j)},x_j,x_{an(i_1)}) \\ &= p(x_{an(i_1)}|x_{\pi(j)},x_j) p(x_{an(j)\setminus\pi(j)\cup an(i_1)\cup an(i_2)}|x_{\pi(j)},x_j,x_{an(i_1)},x_{an(i_2)}) p(x_{an(i_2)}|x_{\pi(j)},x_j,x_{an(i_1)}) \\ &... \\ &= p(x_{an(i_1)}|x_{\pi(j)},x_j) p(x_{an(i_2)}|x_{\pi(j)},x_j,x_{an(i_1)})...p(x_{an(i_P)}|x_{\pi(j)},x_j,x_{an(i_1)},...,x_{an(i_{P-1})}) \end{aligned}$$

Moreover $(u_i, \hat{x}_i)$ are independent from any fields of X given $(X_i, X_{an(i)})$ with γ_i^* and $\gamma_{an(i)}^*$ determined.

Similar steps show that the cost recursion given by Eq.s(3.18) and (3.19) hold, i.e. Eq.(3.19) substituted in Eq.s(3.18) is equal to summation over $m \in de(j)$ in Eq.(A.7). Consider

$$\begin{aligned}
&\sum_{m\in de(j)} \sum_{u_m} \int_{\mathcal{X}_m} \mathrm{d}\hat{x}_m \, c(u_m, x_m, \hat{x}_m) p(u_m, \hat{x}_m | x_j, u_j; \gamma_{\setminus j}^*) \\
&= \sum_{k\in\chi(j)} \left[\int_{\mathcal{X}_k} \sum_{u_k} c(u_k, x_k, \hat{x}_k) p(u_k, \hat{x}_k | x_j, u_j; \gamma_{\setminus j}^*) + \sum_{m\in de(k)} \mathrm{d}\hat{x}_m \sum_{u_m} c(u_m, x_m, \hat{x}_m) p(u_m, \hat{x}_m | x_j, u_j; \gamma_{\setminus j}^*) \right]
\end{aligned}$$

and let the summation over $m \in de(j)$ be denoted by $C_j^*(u_j, x_j)$. Then the expression above becomes

$$\begin{aligned}
&C_j^*(u_j, x_j) \\
&= \sum_{k\in\chi(j)} \left[\int_{\mathcal{X}_k} \mathrm{d}x_k \sum_{u_k} c(u_k, x_k, \hat{x}_k) p(u_k, \hat{x}_k | x_j, u_j; \gamma_{\setminus j}^*) + \int_{\mathcal{X}_k} \mathrm{d}x_k \sum_{u_k} C_k^*(u_k, x_k) p(u_k, \hat{x}_k | x_j, u_j; \gamma_{\setminus j}^*) \right] \\
&= \sum_{k\in\chi(j)} \int_{\mathcal{X}_k} \mathrm{d}x_k \sum_{u_k} \left[c(u_k, x_k, \hat{x}_k) + C_k^*(u_k, x_k) \right] p(u_k, \hat{x}_k | x_j, u_j; \gamma_{\setminus j}^*) \qquad \text{(A.10)}
\end{aligned}$$

where it is possible to extend the distribution $p(u_k, \hat{x}_k | x_j, u_j; \gamma_{\setminus j}^*)$ such that it is expressed in terms of the contributions of the rule local to node k, i.e.

$$\begin{aligned}
p(u_k, \hat{x}_k | x_j, u_j; \gamma_{\setminus j}^*) &= \int_{\mathcal{X}_{\pi(k)\setminus j}} \mathrm{d}x_{\pi(k)\setminus j} \int_{\mathcal{X}_k} \mathrm{d}x_k \sum_{u_{\pi(k)\setminus j}} p(x_{\pi(k)\setminus j}, x_k | x_j) p(u_{\pi(k)\setminus j} | x_{\pi(k)\setminus j}; \gamma_{\setminus j}^*) \times \\
&\qquad\qquad p(u_k, \hat{x}_k | x_j, x_{\pi(k)\setminus j}, x_k, u_j, u_{\pi(k)\setminus j}; \gamma_{\setminus j}^*) \\
&= \int_{\mathcal{X}_{\pi(k)\setminus j}} \mathrm{d}x_{\pi(k)\setminus j} \int_{\mathcal{X}_k} \mathrm{d}x_k \sum_{u_{\pi(k)\setminus j}} p(x_{\pi(k)\setminus j}, x_k | x_j) \prod_{m\in\pi(k)\setminus j} p(u_{m\to k} | x_m; \gamma_m^*, \gamma_{an(m)}^*) p(u_k, \hat{x}_k | x_k, u_{\pi(k)}; \gamma_k^*)
\end{aligned} \qquad \text{(A.11)}$$

where we identify $p(u_{m\to k} | x_m; \gamma_m^*, \gamma_{an(m)}^*)$ as $P_{m\to k}^*(u_{m\to k} | x_m)$ and substituted in Eq.(A.11) and Eq.(A.10) yields the cost recursion Eq.s(3.18) and (3.19). ■

VITA

Murat Üney is born in 1978 at Ankara and received his B.Sc. degree from Ankara University Department of Electronics Engineering in 1999. He completed his M.Sc. studies at Middle East Technical University (METU), Institute of Natural and Applied Sciences (INAS), Department of Electrical and Electronics Engineering in 2001. He was a research assistant at METU INAS between 1999 and 2002. His work experience include communications/DSP development engineering for the industry between 2002 and 2004. He has been a researcher for TÜBİTAK UEKAE İltaren Research Group between 2004 and 2007. At the same period he was continuing his Ph.D. studies in Signal Processing and Control Theory at METU under supervision of Prof. Kemal Leblebicioğlu. He has been with the Computer Vision and Pattern Analysis Laboratory, Signal and Information Processing Group of Sabancı University under supervision of Dr. Müjdat Çetin since 2007. His research interests reside in the intersection of signal processing, communications and control with an emphasis on probabilistic paradigms. In particular; signal processing theory, collaborative (statistical) signal processing, (probabilistic) graphical models, (approximate) statistical inference, machine learning, and sparse signal representations outline his current research efforts.

Parts of this thesis work published in conference proceedings and journals:

1. Murat Üney and Müjdat Çetin, "Monte Carlo Optimization Approach for Decentralized Estimation Networks Under Communication Constraints," *IEEE Transactions on Signal Processing*, in Preparation.
2. Murat Üney and Müjdat Çetin, "An Efficient Monte Carlo Approach for Optimizing Decentralized Estimation Networks Constrained by Undirected Topologies," in *Workshop on Statistical Signal Processing (SSP) 2009*. IEEE, 31 Aug.–3 Sept. 2009.
3. Murat Üney and Müjdat Çetin, "An Efficient Monte Carlo Approach for Optimizing Communication Constrained Decentralized Estimation Networks," in *The* 17^{th} *EUSIPCO*. EURASIP, 24 Aug.–28 Aug. 2009.
4. Murat Üney and Müjdat Çetin,"İletişim Kısıtları Altında Dağıtık Rasgele-Alan Kestirimi (Decentralized Random-Field Estimation Under Communication Constraints)," in *The* 17^{th} *Conference on*

Signal Processing, Communications, and their Applications (SIU 2009). IEEE, Antalya, Turkey, April 2009, (in Turkish, received the 2^{nd} place award in IEEE student paper competition).

5. Murat Üney and Müjdat Çetin,"Akustik Algılayıcı Ağlarında Çarpan Çizgeleri Kullanarak Hedef Konumlandırma (Target Localization in Acoustic Sensor Networks Using Factor Graphs)," in *The* 16^{th} *Conference on Signal Processing, Communications, and their Applications (SIU 2008)*. IEEE, Aydın, Turkey, April 2008, (in Turkish).
6. Murat Üney and Müjdat Çetin, "Graphical Model-based Approaches to Target Tracking in Sensor Networks: An Overview of Some Recent Work and Challenges," in *International Symposium on Image and Signal Processing and Analysis (ISPA 2007)*. IEEE, İstanbul, Turkey, September 2007.

Other publications:

1. Murat Üney and T. Engin Tuncer, "2-D Dizilerin Hatasız Ters-Evrişimi (Perfect Deconvolution of 2-D Signals)," in *National Symposium on Signal Processing, Communications and Its Applications (SIU 2002)*, Denizli, Turkey, 2002 (in Turkish).
2. Murat Üney and T. Engin Tuncer, "Kanal Yankı Giderici için Ters Dönüşümlü Dizilerin Kullanımı (Utilization of Invertible Pseudonoise Sequences for Fast Echo Cancellation)," in *National Symposium on Signal Processing, Communications and Its Applications (SIU 2001)*, Gazi Magusa, 2001 (in Turkish).